SAAGs: Biased Stochastic Variance Reduction Methods for Large-scale Learning

Vinod Kumar Chauhan[*] · **Anuj Sharma** · **Kalpana Dahiya**

Abstract Stochastic approximation is one of the effective approach to deal with the large-scale machine learning problems and the recent research has focused on reduction of variance, caused by the noisy approximations of the gradients. In this paper, we have proposed novel variants of SAAG-I and II (Stochastic Average Adjusted Gradient) Chauhan et el. (2017) [6], called SAAG-III and IV, respectively. Unlike SAAG-I, starting point is set to average of previous epoch in SAAG-III, and unlike SAAG-II, the snap point and starting point are set to average and last iterate of previous epoch in SAAG-IV, respectively. To determine the step size, we have used Stochastic Backtracking-Armijo line Search (SBAS) which performs line search only on selected mini-batch of data points. Since backtracking line search is not suitable for large-scale problems and the constants used to find the step size, like Lipschitz constant, are not always available so SBAS could be very effective in such cases. We have extended SAAGs (I, II, III and IV) to solve non-smooth problems and designed two update rules for smooth and non-smooth problems. Moreover, our theoretical results have proved linear convergence

[*]corresponding author

Vinod Kumar Chauhan
Computer Science & Applications
Panjab University Chandigarh, INDIA
E-mail: jmdvinodjmd@gmail.com, vkumar@pu.ac.in
Homepage: https://sites.google.com/site/jmdvinodjmd

Anuj Sharma
Computer Science & Applications
Panjab University Chandigarh, INDIA
E-mail: anujs@pu.ac.in
Homepage: https://sites.google.com/view/anujsharma

Kalpana Dahiya
University Institute of Engineering and Technology
Panjab University Chandigarh, INDIA
E-mail: kalpanas@pu.ac.in

of SAAG-IV for all the four combinations of smoothness and strong-convexity, in expectation. Finally, our experimental studies have proved the efficacy of proposed methods against the state-of-art techniques.

Keywords stochastic gradient descent · stochastic optimization · variance reduction · strongly covex · smooth and non-smooth · SGD · large-scale learning

1 Introduction

Large-scale machine learning problems have large number of data points or large number of features in each data point, or both are large. This leads to high per-iteration complexity of the iterative learning algorithms, which results in slow training of models. Thus, large-scale learning or learning on the big data is one of the major challenge today in machine learning [6, 35]. To tackle this large-scale learning challenge, recently research has focused on stochastic optimization approach [9], coordinate descent approach [26], proximal algorithms [20], parallel and distributed algorithms [29] (as discussed in [7]), and momentum acceleration algorithms [1]. Stochastic approximation leads to variance in the values of deterministic gradient and noisy gradients which are calculated using stochastic approximation, and affects the convergence of learning algorithm. There are several approaches to deal with stochastic noise but most important of these (as discussed in [10]) are: (a) using mini-batching [8], (b) decreasing learning rates [24], (c) variance reduction [18], and (d) importance sampling [10]. To deal with the large-scale learning problems, we use mini-batching and variance reduction in this paper.

1.1 Optimization Problem

In this paper, we consider composite convex optimization problem, as given below:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = f(w) + g(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) + g(w), \right\} \tag{1}$$

where $f(w)$ is a finite average of component functions $f_i(w) : \mathbb{R}^d \rightarrow \mathbb{R}$, $i = 1, 2, ..., n$, are convex and smooth, $g(w) : \mathbb{R}^d \rightarrow \mathbb{R}$ is a relatively simple convex but possibly non-differentiable function (also referred to as regularizer and sometimes as proximal function). This kind of optimization problems can be found in operation research, data science, signal processing, statistics and machine learning etc. For example, regularized Empirical Risk Minimization (ERM) problem is a common problem in machine learning, which is average of losses on the training dataset. In ERM, component function $f_i(w)$ denotes value of loss function at one data point, e.g., in binary classification, it can be logistic loss, i.e., $f_i(w) = \log\left(1 + \exp\left(-y_i w^T x_i\right)\right)$, where $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is collection of training data points, and hinge-loss, i.e., $f_i(w) = \max\left(0, 1 - y_i w^T x_i\right)$; for regression problem, it can be least squares, i.e., $f_i(w) = 1/2(w^T x_i - y_i)^2$. The regularizer can be $\lambda_1 \|w\|_1$ (l_1-regularizer), $\lambda_2/2\|w\|^2$ (l_2-regularizer) and

$\lambda_1 \|w\|_1 + \lambda_2/2\|w\|^2$ (elastic net regularizer), where λ_1 and λ_2 are regularization coefficients. Thus, problems like logistic regression, SVM, ridge regression and lasso etc. fall under ERM.

1.2 Solution Techniques for Optimization Problem

The simple first order method to solve problem (1) is given by Cauchy in his seminal work in 1847, known as Gradient Descent (GD) method [5], is given below for $(k+1)^{th}$ iteration as:

$$w_{k+1} = w_k - \eta_k \left[\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w_k) + \nabla g(w_k) \right], \tag{2}$$

where η_k is the learning rate (also known as step size in optimization). For non-smooth regularizer, i.e., when $g(w)$ is non-smooth then typically, proximal step is calculated after the gradient step, and method is called Proximal Gradient Descent (PGD), as given below:

$$w_{k+1} = Prox_{\eta_k}^g (z_k) = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2\eta_k} \|w - z_k\|^2 + g(w) \right\}, \tag{3}$$

where $z_k = w_k - \eta_k/n \sum_{i=1}^{n} \nabla f_i(w_k)$. GD converges linearly for strongly-convex and smooth problem, and (3) converges at a rate of $O(1/k)$ for non-strongly convex differentiable problems, where k is the number of iterations. The per-iteration complexity of GD and PGD methods is $O(nd)$. Since for large-scale learning problems, the values of n (number of data points) and/or d (number of features in each data point) is very large, so the per-iteration complexity of these methods is very large. Each iteration becomes computationally expensive and might even be infeasible to process for a limited capacity machine, which leads to slow training of models in machine learning. Thus, the challenge is to develop efficient algorithms to deal with the large-scale learning problems [6, 35].

To tackle this challenge, stochastic approximation is one of the popular approach, first introduced by Robbins and Monro, in their seminal work back in 1951, which makes each iteration independent of number of data points [14, 22]. Based on this approach, we have Stochastic Gradient Descent (SGD) method [2], as given below, to solve problem (1) for the smooth case:

$$w_{k+1} = w_k - \eta_k \left[\nabla f_{i_k}(w_k) + \nabla g(w_k) \right], \tag{4}$$

where i_k is selected uniformly randomly from {1,2,...,n} and $\eta_k \propto 1/\sqrt{k}$. The per-iteration complexity of SGD is $O(d)$ and it is very effective to deal with problems with large number of data points. Since $E\left[\nabla f_{i_k}(w_k) \right] = \nabla f(w_k)$, $\nabla f_{i_k}(w_k)$ is an unbiased estimator of $\nabla f(w_k)$, but the variance in these two values need decreasing learning rates. This leads to slow convergence of learning algorithms. Recently, a lot of research is going on to reduce the variance between $\nabla f_{i_k}(w_k)$ and $\nabla f(w_k)$. First variance reduction method introduced by [18], called SAG (Stochastic Average Gradient), some other common and latest methods are SVRG (Stochastic Variance Reduced Gradient) [13], Prox-SVRG [27], S2GD (Semi-Stochastic Gradient Descent) [15, 30], SAGA [11], Katyusha [1], VR-SGD (Variance Reduced

Gradient Descent) [12], SAAG-I and II [6] etc. Like, GD, these methods utilize full gradient and like, SGD, these methods calculate gradient for one or few data points during each iteration. Thus, just like GD, these methods converge linearly for strongly convex and smooth problems and like, SGD, they calculate gradient using one or few data points and have low per-iteration complexity. Thus, the variance reduction methods enjoy best of GD and SGD. Please refer to, [4] for a review on optimization methods for solving large-scale machine learning problems. In this paper, we have proposed new variants of SAAG-I and II, called SAAG-III and IV, as variance reduction techniques.

1.3 Research Contributions

The research contributions of this paper are summarized below:

(a) Novel variants of SAAG-I and II are proposed, called SAAG-III and SAAG-IV, respectively. Unlike SAAG-I, for SAAG-III the starting point is set to the average of iterates of previous epoch except for the first one, $w_0^{s+1} = 1/m \sum_{i=1}^m w_i^s$, where m is number of inner iterations. Unlike SAAG-II, for SAAG-IV, the starting point and snap point are set to the last iterate and average of previous epoch except for the first one, $w_0^{s+1} = w_m^s$ and $\tilde{w}_0^s = 1/m \sum_{i=1}^m w_i^s$.
(b) SAAG-I and II, including SAAG-III and IV, are extended to solve problems with non-smooth regularizer by introducing two different update rules for smooth and non-smooth cases (see Section 3 and 4, for details).
(c) Theoretical results prove linear convergence of SAAG-IV for all the four combinations of smoothness and strong-convexity in expectation.
(d) Finally, empirical results prove the efficacy of proposed methods against state-of-art methods in terms of convergence and accuracy against training time, number of epochs and number of gradient evaluations.

2 Notations and Related Work

This section discusses notations used in the paper and related work.

2.1 Notations

The training dataset is represented as $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where n is the number of data points and $x_i \in \mathbb{R}^d$, d is the number of features. w denotes the parameter vector and λ_i, $i = 1, 2$ denotes the regularization parameter. $\|.\|$ denotes Euclidean norm, also called l_2-norm, and $\|.\|_1$ denotes l_1-norm. L and μ are used to denote L-smoothness and μ-strong convexity of problem, respectively. η_k^s denotes the learning rate, $s \in \{1, 2, ..., S\}$ denotes epoch number and S is the total number of epochs. b denotes the mini-batch size, $k \in \{1, 2, ..., m\}$ denotes the inner iterations, s.t., $n = mb$. The value of loss function at (x_i, y_i) is denoted by component function f_i. $w^* = \arg\min_w F(w)$ and $F(w^*)$ is the optimal objective function value, sometimes denoted as F^*.

2.2 Related Work

The emerging technologies and the availability of different data sources have lead to rapid expansion of data in all science and engineering domains. On one side, this massive data has potential to uncover more fine-grained patterns, take timely and accurate decisions, and on other side it creates a lot of challenges to make sense of it, like, slow training and scalability of models, because of inability of traditional technologies to process this huge data. The term "Big Data" was coined to highlight the data explosion and need of new technologies to process this massive data. Big data is a vast subject in itself. Big data can be characterized, mainly using three Vs: Volume, Velocity and Variety but recently a lot of other Vs have been used. When one deals with 'volume' aspect of big data in machine learning, it is called the large-scale machine learning problem or big data problem. The large-scale learning problems have large number of data points or large number of features in each data point, or both, which lead to large per-iteration complexity of learning algorithms, and ultimately lead to slow training of machine learning models. Thus, one of the major challenge before machine learning is to develop efficient and scalable learning algorithms [6, 7, 35].

To solve problem (1) for smooth regularizer, a simple method is GD, given by [5], and it converges linearly for strongly-convex and smooth problems. For non-smooth regularizer, typically, proximal step is applied to GD step, called PGD method which converges at a rate of $O(1/k^2)$ for non-strongly convex problems. The per-iteration complexity of GD and PGD is $O(nd)$ which is very large for large-scale learning problems and results in slow training of models. Stochastic approximation is one of the approach to tackle this challenge. It was first introduced by Robbins and Monro [22] and is very effective to deal with problems with large number of data points because each iteration uses one (or few) data points, like in SGD [2, 32]. In SGD, each iteration is n times faster than GD, as their per-iteration complexities are $O(d)$ and $O(nd)$, respectively. SGD need decreasing learning rates, i.e., $\eta_k \propto 1/\sqrt{k}$ for k^{th} iteration, because of variance in gradients, so it converges slower than GD, with sub-linear convergence rate $O(1/T)$ even for strongly convex problem [21]. There are several approaches to deal with stochastic noise but most important of these (as discussed in [10]) are: (a) using mini-batching [30], (b) decreasing learning rates[24], (c) variance reduction [18], and (d) importance sampling [10].

Variance reduction techniques, first introduced by [18], called SAG, converges linearly, like, GD for strongly convex and smooth problems, and uses one randomly selected data point, like SGD during each iteration. SAG enjoys benefits of both GD and SGD, as it converges linearly for strongly convex and smooth case like, GD but it has per-iteration complexity of SGD. Later, a lot of variance reduction methods were proposed, like, SVRG [13], SAGA [11], S2GD [15], SDCA [23], SPDC [33], Katyusha [1], Catalyst [19], SAAG-I, II [6] and VR-SGD [12] etc. These variance reduction methods can use constant learning rate and can be divided into three categories (as discussed in [12]): (a) primal methods which can be applied to primal optimization problem, like, SAG, SAGA, SVRG etc., (b) dual methods which can be applied to dual problems, like, SDCA, and (c) primal-dual methods which involve primal and dual variable both, like, SPDC etc.

In this paper, we have proposed novel variants of SAAG-I and II, named as SAAG-III and SAAG-IV, respectively. Unlike SAAG-I, for SAAG-III the start-

ing point is set to the average of iterates of previous epoch except for the first one, $w_0^{s+1} = 1/m \sum_{i=1}^{m} w_i^s$, where m is number of inner iterations. Unlike SAAG-II, for SAAG-IV, the starting point and snap point are set to the last iterate and average of previous epoch except for the first one, $w_0^{s+1} = w_m^s$ and $\tilde{w}_0^s = 1/m \sum_{i=1}^{m} w_i^s$. [6] proposed Batch Block Optimization Framework (BBOF) to tackle the big data (large-scale learning) challenge in machine learning, along with two variance reduction methods SAAG-I and II. BBOF is based on best of stochastic approximation (SA) and best of coordinate descent (CD) (another approach which is very effective to deal with large-scale learning problems especially problems with high dimensions). Techniques based on best features of SA and CD approaches are also used in [25, 28, 31, 34], and [31] calls it doubly stochastic since both data points and coordinate are sampled during each iteration. It is observed that for ERM, it is difficult to get the advantage of BBOF in practice because results with CD or SGD are faster than BBOF setting as BBOF needs extra computations while sampling and updating the block of coordinates. When one block of coordinates is updated and as we move to another block, the partial gradients need dot product of parameter vector (w) and data points (like, in logistic regression). Since each update changes w so for each block update needs to calculate the dot product. On other hand, if all coordinates are updated at a time, like in SGD, that would need to calculate dot product only once. Although, Gauss-Seidel update of parameters helps in faster convergence but the overall gain is less because of extra computational load. Moreover, SAAG-I and II have been proposed to work in BBOF (mini-batch and block-coordinate) setting as well as mini-batch (and considering all coordinates). Since BBOF is not very helpful for ERM so the SAAG-III and IV are proposed for mini-batch setting only. SAAGs (I, II, III and IV) can be extended to stochastic setting (consider one data point during each iteration) but SAAG-I and II are unstable for stochastic setting, and SAAG-III and IV, could not beat the existing methods in stochastic setting. SAAGs has been extended to deal with smooth and non-smooth regularizers, as we have used two different update rules, like [12] (see Section 3 for details).

3 SAAG-I, II and Proximal Extensions

Originally, [6] proposed SAAG-I and II for smooth problems, we have extended SAAG-I and II to non-smooth problems. Unlike proximal methods which use single update rule for both smooth and non-smooth problems, we have used two different update rules and introduced proximal step for non-smooth problem. For mini-batch B_k of size b, epoch s and inner iteration k, SAAG-I and II are given below:

SAAG-I:

$$
\begin{aligned}
&\text{smooth} \qquad w_{k+1}^s = w_k^s - \eta_k^s \left[\sum_{i=1}^{n} \tau_i^{s,k} + \nabla g(w_k^s) \right], \\
&\text{non-smooth} \;\; w_{k+1}^s = Prox_{\eta_k^s}^g \left(w_k^s - \eta_k^s \sum_{i=1}^{n} \tau_i^{s,k} \right),
\end{aligned}
\tag{5}
$$

where $\tau_i^{s,k} = \begin{cases} \dfrac{1}{b} \nabla f_i(w_k^s) & \text{if } i \in B_k \\ \dfrac{1}{l} \tilde{\tau}_i^{s,k} & \text{otherwise} \end{cases}$ and $\tilde{\tau}_i^{s,k} = \begin{cases} \nabla f_i(w_k^s) & \text{if } i \in B_k \\ \tilde{\tau}_i^{s-1,k} & \text{otherwise.} \end{cases}$

SAAG-II:

$$\text{smooth} \qquad w_{k+1}^s = w_k^s - \eta_k^s \left[\tilde{\nabla} f_{B_k}(w_k^s) + \nabla g(w_k^s) \right],$$
$$\text{non-smooth} \quad w_{k+1}^s = Prox_{\eta_k^s}^g \left(w_k^s - \eta_k^s \tilde{\nabla} f_{B_k}(w_k^s) \right), \qquad (6)$$

where $\tilde{\nabla} f_{B_k}(w_k^s) = 1/b \sum_{i \in B_k} \nabla f_i(w_k^s) - 1/n \sum_{i \in B_k} \nabla f_i(\tilde{w}^s) + \tilde{\mu}^s$ and $\tilde{\mu}^s = 1/n \sum_{i=1}^n \nabla f_i(\tilde{w}^s)$. Unlike SVRG and VRSGD, and like SAG, SAAGs are biased gradient estimators because the expectation of gradient estimator is not equal to full gradient, i.e., $\mathbb{E}[\tilde{\nabla} f_{B_k}(w_k^s)] = \nabla f(w_k^s) + (m-1)/m \nabla f(\tilde{w}^{s-1})$, as detailed in Lemma 6.

SAAG-I algorithm, represented by Algorithm 1, divides the dataset into m mini-batches of equal size (say) b and takes input S, number of epochs. During each inner iteration, it randomly selects one mini-batch of data points from $[n]$, calculates gradient over mini-batch, updates the total gradient value and performs stochastic backtracking-Armijo line search (SBAS) over B_k. Then parameters are updated using Option I for smooth regularizer and using Option II for non-smooth regularizer. Inner iterations are run m times where $n = mb$ and then last iterate is used as the starting point for next epoch.

SAAG-II algorithm, represented by Algorithm 2, takes input as number of epochs

---

Algorithm 1 SAAG-I

Inputs: $m = $ #mini-batches and $S = $ max #*epochs*.
Initialize: w_0^1
1: **for** $s = 1, 2, ..., S$ **do**
2:     **for** $k = 0, 1, ..., (m-1)$ **do**
3:         Randomly select one mini-batch B_k from [n].
4:         Update the gradient values, $\tau_i^{s,k}$ and $\tilde{\tau}_i^{s,k}$.
5:         Calculate η_k^s using stochastic backtracking line search on B_k.
6:         Option I (smooth): $w_{k+1}^s = w_k^s - \eta_k^s \left[\sum_{i=1}^n \tau_i^{s,k} + \nabla g(w_k^s) \right]$,
7:         Option II (non-smooth): $w_{k+1}^s = Prox_{\eta_k^s}^g \left(w_k^s - \eta_k^s \sum_{i=1}^n \tau_i^{s,k} \right)$,

        where $\tau_i^{s,k} = \begin{cases} \dfrac{1}{b} \nabla f_i(w_k^s) & \text{if } i \in B_k \\ \dfrac{1}{l} \tilde{\tau}_i^{s,k} & \text{otherwise} \end{cases}$ and $\tilde{\tau}_i^{s,k} = \begin{cases} \nabla f_i(w_k^s) & \text{if } i \in B_k \\ \tilde{\tau}_i^{s-1,k} & \text{otherwise}. \end{cases}$

8:     **end for**
9: **end for**
10: **Output:** w_m^S

(S) and number of mini-batches (m) of equal size (say) b. It initializes $w_0^1 = \tilde{w}^0$. During each inner iteration, it randomly selects one mini-batch B_k, calculates two gradients over B_k at last iterate and snap-point, updates $\tilde{\nabla} f_{B_k}(w_k^s)$ and performs stochastic backtracking-Armijo line search (SBAS) over B_k. Then parameters are updated using Option I for smooth regularizer and using Option II for non-smooth regularizer. After inner iterations, it uses the last iterate to set the snap point and the starting point for the next epoch.

Algorithm 2 SAAG-II

Inputs: $m = $#mini-batches and $S = $max #*epochs*.
Initialize: w_0^1

1: **for** $s = 1, 2, ..., S$ **do**
2:     $\tilde{w}^s = w_0^s$
3:     $\tilde{\mu}^s = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w}^s)$ // calculate full gradient
4:     **for** $k = 0, 1, ..., (m-1)$ **do**
5:         Randomly select one mini-batch B_k from [n].
6:         Calculate $\tilde{\nabla} f_{B_k}(w_k^s) = \frac{1}{b} \sum_{i \in B_k} \nabla f_i(w_k^s) - \frac{1}{n} \sum_{i \in B_k} \nabla f_i(\tilde{w}^s) + \tilde{\mu}^s$.
7:         Calculate η_k^s using stochastic backtracking-Armijo line search on B_k.
8:             Option I (smooth): $w_{k+1}^s = w_k^s - \eta_k^s \left[\tilde{\nabla} f_{B_k}(w_k^s) + \nabla g(w_k^s) \right]$
9:             Option II (non-smooth): $w_{k+1}^s = Prox_{\eta_k^s}^g \left(w_k^s - \eta_k^s \tilde{\nabla} f_{B_k}(w_k^s) \right)$
10:    **end for**
11:    $w_0^{s+1} = w_m^s$
12: **end for**
13: **Output:** w_m^S

---

4 SAAG-III and IV Algorithms

SAAG-III algorithm, represented by Algorithm 3, divides the dataset into m mini-batches of equal size (say) b and takes input S, number of epochs. During each inner iteration, it randomly selects one mini-batch of data points B_k from $[n]$, calculates gradient over mini-batch, updates the total gradient value and performs stochastic backtracking-Armijo line search (SBAS) over B_k. Then parameters are updated using Option I for smooth regularizer and using Option II for non-smooth regularizer. Inner iterations are run m times where $n = mb$ and then iterate average is calculated and used as the starting point for next epoch, $w_0^{s+1} = 1/m \sum_{i=1}^m w_i^s$. SAAG-IV algorithm, represented by Algorithm 4, takes input as number of epochs

---

Algorithm 3 SAAG-III

Inputs: $m = $#mini-batches and $S = $max #*epochs*.
Initialize: w_0^1

1: **for** $s = 1, 2, ..., S$ **do**
2:     **for** $k = 0, 1, ..., (m-1)$ **do**
3:         Randomly select one mini-batch B_k from [n].
4:         Update the gradient values, $\tau_i^{s,k}$ and $\tilde{\tau}_i^{s,k}$.
5:         Calculate η_k^s using stochastic backtracking line search on B_k.
6:         Option I (smooth): $w_{k+1}^s = w_k^s - \eta_k^s \left[\sum_{i=1}^n \tau_i^{s,k} + \nabla g(w_k^s) \right]$,
7:         Option II (non-smooth): $w_{k+1}^s = Prox_{\eta_k^s}^g \left(w_k^s - \eta_k^s \sum_{i=1}^n \tau_i^{s,k} \right)$,

            where $\tau_i^{s,k} = \begin{cases} \frac{1}{b} \nabla f_i(w_k^s) & \text{if } i \in B_k \\ \frac{1}{l} \tilde{\tau}_i^{s,k} & \text{otherwise} \end{cases}$ and $\tilde{\tau}_i^{s,k} = \begin{cases} \nabla f_i(w_k^s) & \text{if } i \in B_k \\ \tilde{\tau}_i^{s-1,k} & \text{otherwise.} \end{cases}$

8:     **end for**
9:     $w_0^{s+1} = 1/m \sum_{i=1}^m w_i^s$ // iterate averaging
10: **end for**
11: **Output:** w_m^S

Algorithm 4 SAAG-IV

Inputs: $m = $ #mini-batches and $S = $max #*epochs*.
Initialize: $w_0^1 = \tilde{w}^0$

1: **for** $s = 1, 2, ..., S$ **do**
2:     $\tilde{\mu}^s = \dfrac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{w}^{s-1})$ // calculate full gradient
3:     **for** $k = 0, 1, ..., (m-1)$ **do**
4:        Randomly select one mini-batch B_k from [n].
5:        Calculate $\tilde{\nabla} f_{B_k}(w_k^s) = \dfrac{1}{b} \sum_{i \in B_k} \nabla f_i(w_k^s) - \dfrac{1}{n} \sum_{i \in B_k} \nabla f_i(\tilde{w}^{s-1}) + \tilde{\mu}^s$.
6:        Calculate η_k^s using stochastic backtracking-Armijo line search on B_k.
7:        Option I (smooth): $w_{k+1}^s = w_k^s - \eta_k^s \left[\tilde{\nabla} f_{B_k}(w_k^s) + \nabla g(w_k^s) \right]$
8:        Option II (non-smooth): $w_{k+1}^s = Prox_{\eta_k^s}^g \left(w_k^s - \eta_k^s \tilde{\nabla} f_{B_k}(w_k^s) \right)$
9:     **end for**
10:    $\tilde{w}^s = 1/m \sum_{i=1}^{m} w_i^s$ // iterate averaging
11:    $w_0^{s+1} = w_m^s$ //initialize starting point
12: **end for**
13: **Output:** w_m^S

(S) and number of mini-batches (m) of equal size (say) b. It initializes $w_0^1 = \tilde{w}^0$. During each inner iteration, it randomly selects one mini-batch B_k, calculates two gradients over B_k at last iterate and snap-point, updates $\tilde{\nabla} f_{B_k}(w_k^s)$ and performs stochastic backtracking-Armijo line search (SBAS) over B_k. Then parameters are updated using Option I for smooth regularizer and using Option II for non-smooth regularizer. After inner iterations, it calculates average to set the snap point, and uses last iterate as the starting point for the new epoch, as $\tilde{w}^s = 1/m \sum_{i=1}^{m} w_i^s$ and $w_0^{s+1} = w_m^s$, respectively.

The comparative study of SAAGs is represented by Figure 1 for smooth problem (l_2-regularized logistic regression), which compares accuracy and suboptimality against training time (in seconds), gradients/n and epochs. The results are reported on Adult dataset with mini-batch of 32 data points. It is clear from all the six criteria plots that results for SAAG-III and IV are very stable than SAAG-I and II, respectively, because of averaging of iterates. SAAG-IV performs better than SAAG-II and SAAG-III performs closely but stably than SAAG-I. Moreover, SAAG-I and SAAG-II stabilize with increase in mini-batch size but the performance of methods decreases with mini-batch size (see, Appendix for effect of mini-batch sizes on SAAGs).

5 Analysis

In general, SAAG-IV gives better results for large-scale learning problems as compared to SAAG-III as shown by empirical results presented in Fig. 3, 4 with news20, rcv1 datasets and results in Appendix A.3 Effect of mini-batch size. So, in this section, we have provided convergence rates of SAAG-IV considering all cases of smoothness with strong convexity. Moreover, analysis of SAAG-III represents a typical case due to the biased nature of gradient estimator and the fact that the full gradient is incrementally maintained rather than being calculated at a fix point, like in SAAG-IV. So, analysis of SAAG-III is left open. The convergence
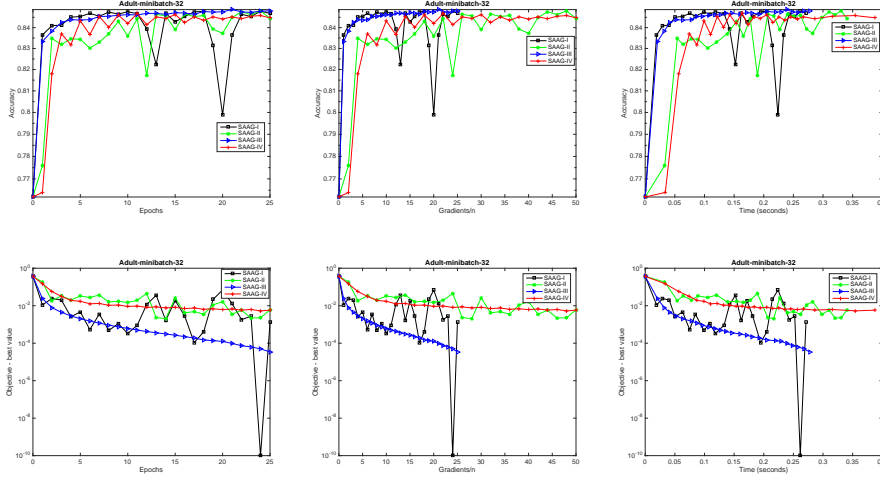
Fig. 1: Comparison of SAAG-I, II, III and IV on smooth problem using Adult dataset with mini-batch of 32 data points. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

rates for all the different combinations of smoothness and strong convexity are given below:

Theorem 1. *Under the assumptions of Lipschitz continuity with smooth regularizer, the convergence of SAAG-IV is given below:*

$$\mathbb{E}\left[f(\tilde{w}^s) - f^*\right] \leq C^s \left[f(\tilde{w}^0) - f^*\right] + V, \tag{7}$$

where, $C = \left[\dfrac{4\alpha(b)}{(\beta - 1 - 4\alpha(b))}\dfrac{c}{m} + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta - 1 - 4\alpha(b))}\right] < 1$ *and* V *is constant.*

Theorem 2. *Under the assumptions of Lipschitz continuity and strong convexity with smooth regularizer, the convergence of SAAG-IV method is given below:*

$$\mathbb{E}\left[f(\tilde{w}^s) - f^*\right] \leq C^s \left[f(\tilde{w}^0) - f^*\right] + V, \tag{8}$$

where,

$$C = \left[\frac{cL\beta}{m\mu} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta - 1)} - \frac{c(m-1)}{m^2} + \frac{4\alpha(b)}{(\beta - 1)}\right]$$

$$\left(1 - \frac{4\alpha(b)}{(\beta - 1)} - c\left(\frac{m-1}{m^2} - \frac{4\alpha(b)}{(\beta - 1)}\frac{1}{m}\right)\right)^{-1} < 1 \ and \ V \ is \ constant.$$

Theorem 3. *Under the assumptions of Lipschitz continuity with non-smooth regularizer, the convergence of SAAG-IV is given below:*

$$\mathbb{E}\left[F(\tilde{w}^s) - F(w^*)\right] \leq C^s \left[F(\tilde{w}^0) - F(w^*)\right] + V, \tag{9}$$

where,

$$C = \left(\frac{4\alpha(b)}{(\beta - 1 - 4\alpha(b))} \frac{c}{m} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta - 1 - 4\alpha(b))} \right) < 1 \ and \ V \ is \ constant.$$

Theorem 4. *Under the assumptions of Lipschitz continuity and strong convexity with non-smooth regularizer, the convergence of SAAG-IV is given below:*

$$\mathbb{E}\left[F(\tilde{w}^s) - F^*\right] \leq C^s \left[F(\tilde{w}^0) - F^*\right] + V, \tag{10}$$

where,

$$C = \left(\frac{Lc\beta}{m\mu} + \frac{4c\alpha(b)}{m(\beta - 1)} - \frac{c(m-1)}{m^2} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta - 1)} \right)$$

$$\left(1 - \frac{4\alpha(b)}{(\beta - 1)} - \frac{c(m-1)}{m^2} + \frac{4c\alpha(b)}{m(\beta - 1)} \right)^{-1} < 1 \ and \ V \ is \ constant.$$

All the proofs are given in the Appendix B and all these results prove linear convergence (as per definition of convergence) of SAAG-IV for all the four combinations of smoothness and strong-convexity with some initial errors due to the constant terms in the results. SAAGs are based on intuitions from practice [6] and they try to give more importance to the latest gradient values than the older gradient values, which make them biased techniques and results into this extra constant term. This constant term signifies that SAAGs converge to a region close to the solution, which is very practical because all the machine learning algorithms are used to solve the problems approximately and we never find an exact solution for the problem [3], because of computational difficulty. Moreover, the constant term pops up due to the mini-batched gradient value at optimal point, i.e., $\frac{1}{|B_k|}\sum_{i \in B_k} \nabla f_i(w^*)$. If the size of the mini-batch increases and eventually becomes equal to the dataset then this constant becomes equal to full gradient and vanishes, i.e., $\frac{1}{n}\sum_{i=1}^n \nabla f_i(w^*) = 0$.

The linear convergence for all combinations of strong convexity and smoothness of the regularizer, is the maximum rate exhibited by the first order methods without curvature information. SAG, SVRG, SAGA and VR-SGD also exhibit linear convergence for the strong convexity and smooth problem but except VR-SGD, they don't cover all the cases, e.g., SVRG does not cover the non-strongly convex cases. However, the theoretical results provided by VR-SGD, prove linear convergence for strongly convex cases, like our results, but VR-SGD provides only $O(1/T)$ convergence for non-strongly convex cases, unlike our linear convergence results.

6 Experimental Results

In this section, we have presented the experimental results[1]. SAAG-III and IV are compared against the most widely used variance reduction method, SVRG and one of the latest method VR-SGD which has been proved to outperform existing

---

[1] experimental results can be reproduced using the code available at link: https://drive.google.com/open?id=1Rdp_pmHLQAA9OBxBtHzz6FCduCypAzhd

techniques. The results have been reported in terms of suboptimality and accuracy against time, epochs and gradients/n. The SAAGs can be applied to strongly and non-strongly convex problems with smooth or non-smooth regularizers. But the results have been reported with strongly convex problems with and without smoothness because problems can be easily converted to strongly convex problems by adding l_2-regularization.

6.1 Experimental Setup

The experiments are reported using six different criteria which plot suboptimality (objective minus best value) versus epochs (where one epoch refers to one pass through the dataset), suboptimality versus gradients/n, suboptimality versus time, accuracy versus time, accuracy versus epochs and accuracy versus gradients/n. The x-axis and y-axis data are represented in linear and log scale, respectively. The experiments use the following binary datasets: rcv1 (#data - 20, 242, #features - 47, 236), news20 (#data - 19, 996, #features - 1, 355, 191), real-sim (#data - 72,309, #features - 20, 958) and Adult (also called as a9a, #data - 32,561 and #features - 123), which are available from the LibSVM website[2]. All the datasets are divided into 80% and 20% as training and test dataset, respectively. The value of regularization parameter is set as $\lambda = 1 * 10^{-5}$(including λ_1, λ_2) for all the algorithms. The parameters for, Stochastic Backtracking-Armijo line Search (SBAS), are set as: $\alpha = 0.1, \beta = 0.5$ and learning rate is initialized as, $\eta = 1.0$. The inner iterations are set as, $m = n/b$ (as used in [12]). Moreover, in SBAS, algorithms looks for maximum 10 iterations and after that it returns the current value of learning rate if it reduces the objective value otherwise it returns 0.0. This is done to avoid sticking in the algorithm because of stochastic line search. All the experiments have been conducted on MacBook Air (8 GB 1600 MHz DDR3, 1.6 GHz Intel Core i5 and 256GB SSD) using MEX files.

6.2 Results with Smooth Problem

The results are reported with l_2-regularized logistic regression problem as given below:

$$\min_w F(w) = \frac{1}{n} \sum_{i=1}^n \log\left(1 + \exp\left(-y_i w^T x_i\right)\right) + \frac{\lambda}{2}\|w\|^2. \tag{11}$$

Figure 2 represents the comparative study of SAAG-III, IV, SVRG and VR-SGD on real-sim dataset. As it is clear from the first row of the figure, SAAG-III and IV give better accuracy and attain the results faster than other the other methods. From the second row of the figure, it is clear that SAAGs converges faster than SVRG and VR-SGD. Moreover, SAAG-III performs better than SAAG-IV and VR-SGD performs slightly better than SVRG as established in [12]. Figure 3 reports results with news20 dataset and as depicted in the figure the results are similar to real-sim dataset Fig. (2). SAAGs give better accuracy and converge faster than SVRG and VR-SGD methods, but SAAG-IV gives best results. This is because as the mini-batch size or the dataset size increases, SAAG-II and SAAG-IV perform better (as reported in [6]).

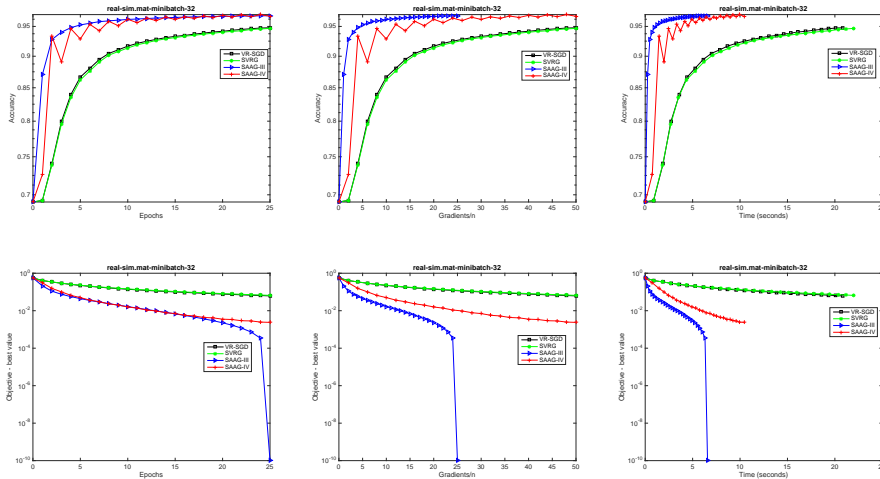[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Fig. 2: Comparison of SAAG-III, IV, SVRG and VR-SGD on smooth problem using real-sim dataset with mini-batch of 32 data points. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

6.3 Results with non-smooth Problem

The results are reported with elastic-net-regularized logistic regression problem (non-smooth regularizer) as given below:

$$\min_{w} F(w) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right) + \frac{\lambda_1}{2}\|w\|^2 + \lambda_2\|w\|_1, \qquad (12)$$

where $f(w) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right) + \frac{\lambda_1}{2}\|w\|^2$ and $g(w) = \lambda_2\|w\|_1$.

Figure 4 represents the comparative study of SAAG-III, IV, SVRG and VR-SGD on rcv1 dataset. As it is clear from the figure, for all the six criteria plots, SAAG-III and IV outperform SVRG and VR-SGD, and provide better accuracy and faster convergence. SAAG-IV gives best results in terms of suboptimality but in terms of accuracy, SAAG-III and IV have close performance except for accuracy versus gradients/n, where SAAG-III gives better results because SAAG-III calculates gradients at last iterate only unlike SAAG-IV which calculates gradients at snap point and last iterate. Figure 5 reports results with Adult dataset, and as it is clear from plots, SAAG-III outperforms all the methods in all the six criteria plots. Moreover, SAAG-IV lags behind because the dataset and/or mini-batch size is small. Some results are also given with SVM in Appendix A.1

6.4 Effect of Regularization Constant

Figure 6, studies the effect of regularization coefficient on SAAG-III, IV, SVRG and VR-SGD for smooth problem (l_2-regularized logistic regression) using rcv1
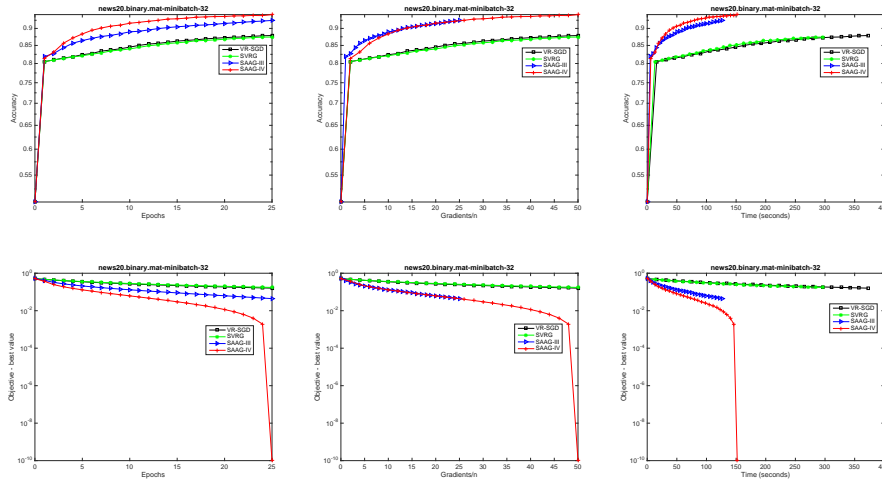
Fig. 3: Comparison of SAAG-III, IV, SVRG and VR-SGD on smooth problem using news20 dataset with mini-batch of 32 data points. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

dataset and considering regularization coefficient values in $\{10^{-3}, 10^{-5}, 10^{-7}\}$. As it is clear from the plots, all the methods are affected by the large (10^{-3}) regularization coefficient value and have low accuracy but for sufficiently small values methods don't have much effect. For suboptimality plots, all have slower convergence for large regularization coefficient but then convergence improves with the decrease in regularization, because decreasing the regularization, increases the over-fitting. The results for non-smooth problem are similar, so they are given in the Appendix A.5.

6.5 Effect of mini-batch size

Figure 7, studies the effect of mini-batch size on SAAG-III, IV, SVRG and VR-SGD for smooth problem (l_2-regularized logistic regression) using rcv1 dataset and considers mini-batch values in $\{32, 64, 128\}$ data points. As it is clear from the plots, except for the results with training time, the performance of SAAG-III, IV and SVRG fall with increase in mini-batch size but for VR-SGD performance first improves slightly and then falls slightly. For results with training time the performance of SAAG-III and IV falls with mini-batch size for suboptimality and remains almost same for accuracy. But the performance of VR-SGD and SVRG improves because VR-SGD and SVRG train quickly for large mini-batches. Similar results are obtained for study of effect of mini-batch size on non-smooth problem so the results are given in the Appendix A.3.
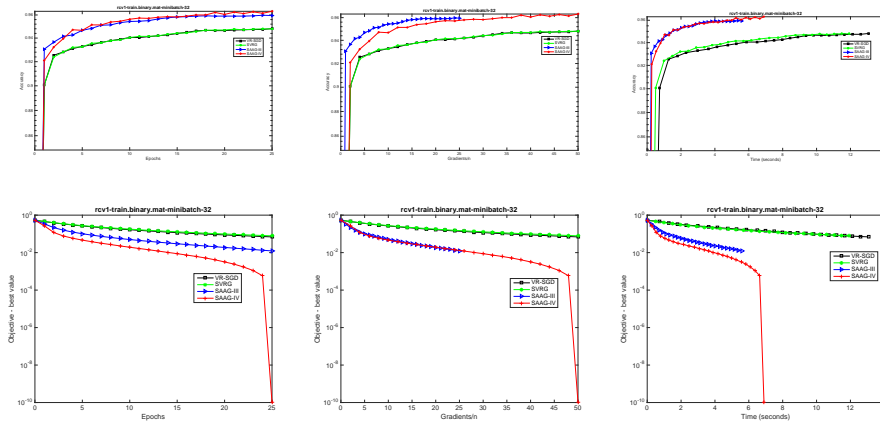
Fig. 4: Comparison of SAAG-III, IV, SVRG and VR-SGD on non-smooth problem using rcv1 dataset with mini-batch of 32 data points. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

7 Conclusion

We have proposed novel variants of SAAG-I and II, called SAAG-III and IV, respectively, by using average of iterates for SAAG-III as a starting point, and average of iterates and last iterate for SAAG-IV as the snap point and starting point, respectively, for new epoch, except the first one. SAAGs (I, II, III and IV), are also extended to solve non-smooth problems by using two different update rules and introducing proximal step for non-smooth problem. Theoretical results proved linear convergence of SAAG-IV for all the four combinations of smoothness and strong-convexity with some initial errors, in expectation. The empirical results proved the efficacy of proposed methods against existing variance reduction methods in terms of, accuracy and suboptimality, against training time, epochs and gradients/n.
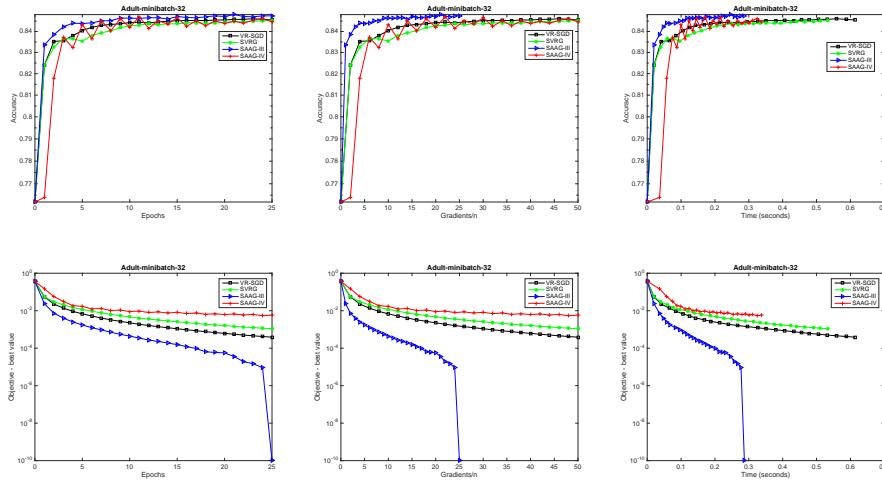
Fig. 5: Comparison of SAAG-III, IV, SVRG and VR-SGD on non-smooth problem using Adult dataset with mini-batch of 32 data points. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.
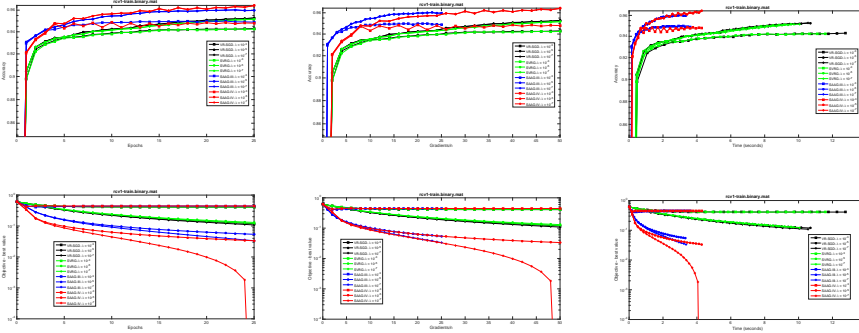
Fig. 6: Study of effect of regularization coefficient on SAAG-III, IV, SVRG and VR-SGD for smooth problem using rcv1 dataset and considering regularization coefficient values in $\{10^{-3}, 10^{-5}, 10^{-7}\}$. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

Fig. 7: Study of effect of mini-batch size on SAAG-III, IV, SVRG and VR-SGD for smooth problem, using rcv1 dataset with mini-batch sizes of 32, 64 and 128. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.
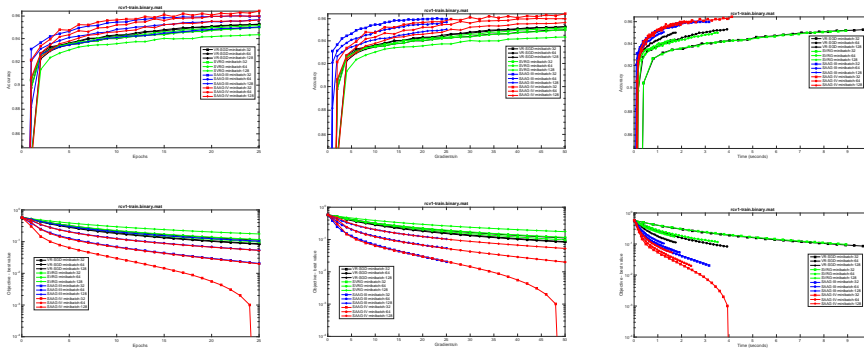
A More Experiments

A.1 Results with Support Vector Machine (SVM)

This subsection compares SAAGs against SVRG and VR-SGD on SVM problem with mushroom and gisette datasets. Methods use stochastic backtracking line search method to find the step size. Fig. 8 presents the results and compares the suboptimality against the training time (in seconds). Results are similar to experiments with logistic regression but are not that smooth. SAAGs outperform other methods on mushroom dataset (first row) and gisette dataset (second row) for suboptimality against training time and accuracy against time but all methods give almost similar results on accuracy versus training time for mushroom dataset. SAAG-IV outperforms other method and SAAG-III sometimes lags behind VR-SGD method. It is also observed that results with logistic regression are better than the results with the SVM problem. The optimization problem for SVM is given below:

$$\min_{w} F(w) = \frac{1}{n} \sum_{i=1}^{n} \max \left(0, 1 - y_i w^T x_i \right)^2 + \frac{\lambda}{2} \|w\|^2, \tag{13}$$

where λ is the regularization coefficient (also penalty parameter) which balances the trade off between margin size and error [7].
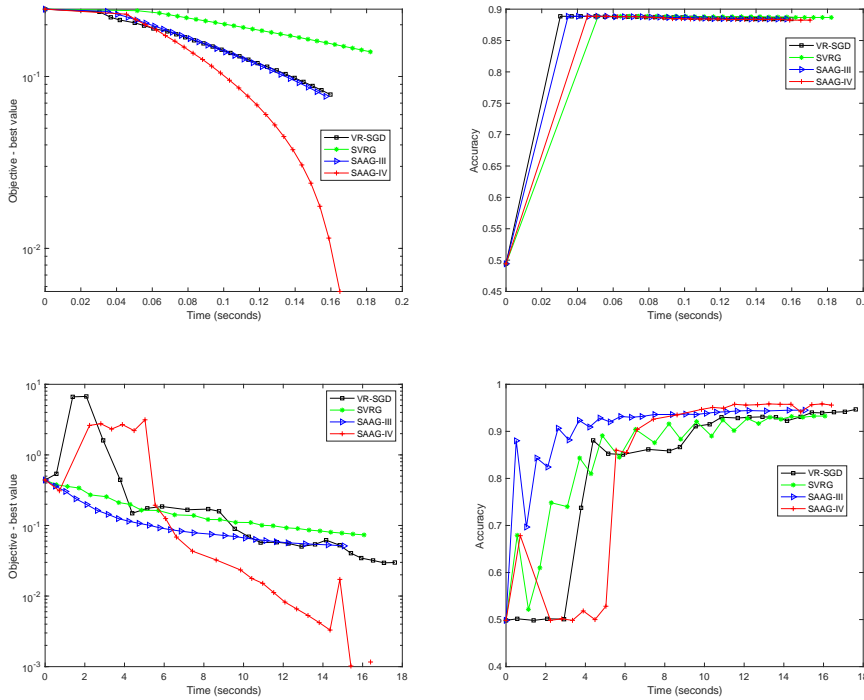

Fig. 8: Results with SVM using mini-batch of 1000 data points on mushroom (first row) and gisette (second row) datasets.

A.2 Comparison of SAAGs (I, II, III and IV) for non-smooth problem

Comparison of SAAGs for non-smooth problem is depicted in Figure 9 using Adult dataset with mini-batch of 32 data points. As it is clear from the figure, just like the smooth problem, results with SAAG-III and IV are stable and better or equal to SAAG-I and II.
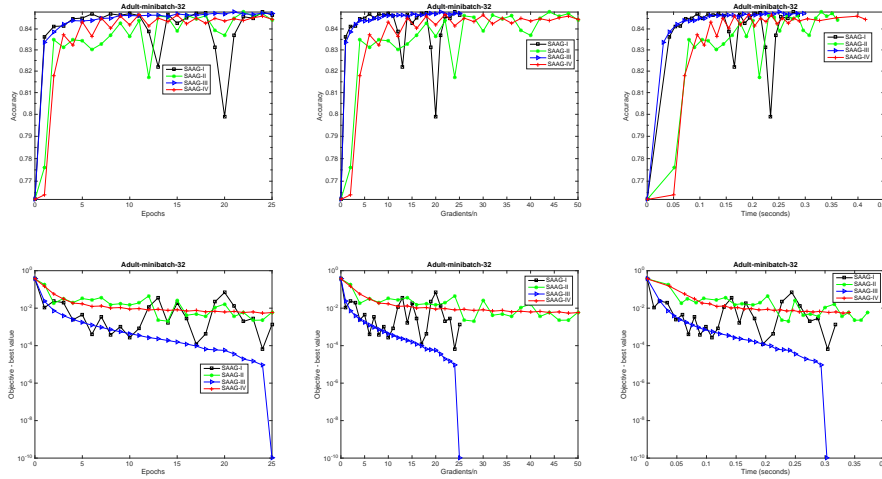

Fig. 9: Comparison of SAAG-I, II, III and IV on non-smooth problem (elastic-net-regularized logistic regression) using Adult dataset with mini-batch size of 32 data points. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

A.3 Effect of mini-batch size on SAAG-III, IV, SVRG and VR-SGD for non-smooth problem

Effect of mini-batch size on SAAG-III, IV, SVRG and VR-SGD for non-smooth problem is depicted in Figure 10 using rcv1 binary dataset with mini-batch of 32, 64 and 128 data points. Similar to smooth problem, proposed methods outperform SVRG and VR-SGD methods. SAAG-IV gives the best result in terms of time and epochs but in terms of gradients/n, SAAG-III gives best results.

A.4 Effect of mini-batch size on SAAGs (I, II, III, IV) for smooth problem

Effect of mini-batch size on SAAGs (I, II, III, IV) for smooth problem is depicted in Figure 11 using Adult dataset with mini-batch sizes of 32, 64 and 128 data points. The results are similar to non-smooth problem.

A.5 Effect of regularization coefficient for non-smooth problem

Figure 12 depicts effect of regularization coefficient on SAAG-III, IV, SVRG and VR-SGD for non-smooth problem using rcv1 dataset. It considers regularization coefficient values as 10^{-3},

Fig. 10: Study of effect of mini-batch size on SAAG-III, IV, SVRG and VR-SGD for non-smooth problem, using rcv1 dataset with mini-batch sizes of 32, 64 and 128. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.
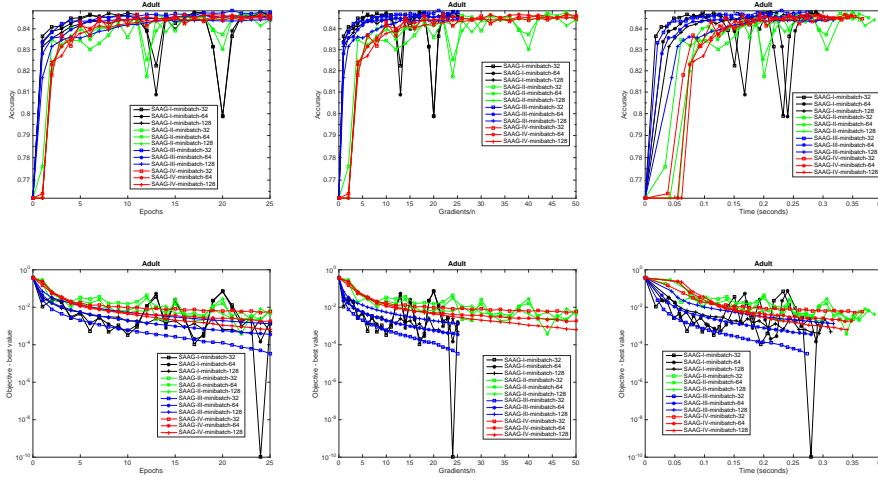
Fig. 11: Study of effect of mini-batch size on SAAGs (I, II, III, IV) for smooth problem, using Adult dataset with mini-batch sizes of 32, 64 and 128. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.

10^{-5} and 10^{-7}. The results are similar to smooth problem. As it is clear from the figure, for larger values, 10^{-3}, all the methods do not perform well but once the coefficient is sufficiently small, it does not make much difference, and in all the cases our proposed methods outperform SVRG and VR-SGD.

Fig. 12: Study of effect of regularization coefficient on SAAG-III, IV, SVRG and VR-SGD for non-smooth problem using rcv1 datase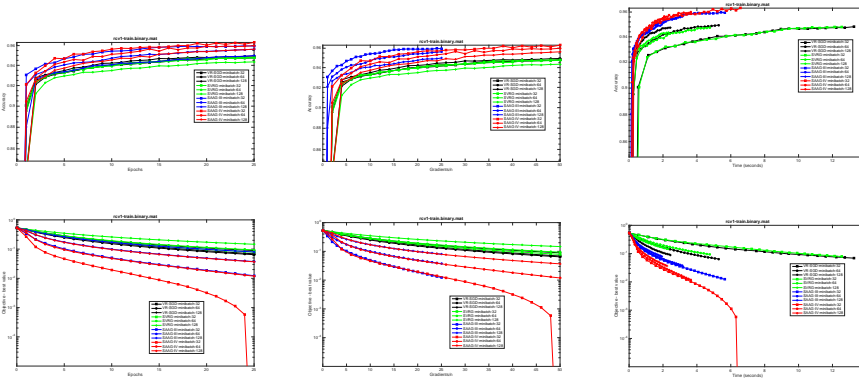t and taking regularization coefficient values as 10^{-3}, 10^{-5} and 10^{-7}. First row compares accuracy against epochs, gradients/n and time, and second row compares suboptimality against epochs, gradients/n and time.
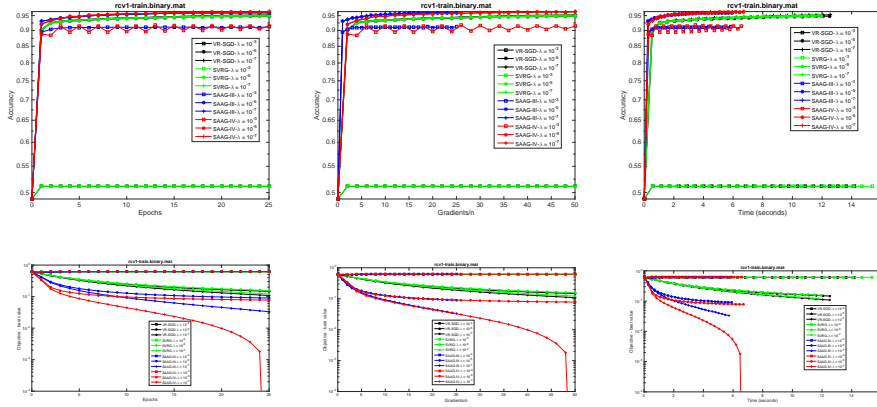
B Proofs

Following assumptions are considered in the paper:

Assumption 1 (Smoothness). *Suppose function $f_i : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that gradient ∇f_i, $\forall i$ is L-Lipschitz-continuous, where $L > 0$ is Lipschitz constant, then, we have,*

$$\|\nabla f_i(y) - \nabla f_i(x)\| \leq L\|y - x\|, \tag{14}$$

$$and, \quad f_i(y) \leq f_i(x) + \nabla f_i(x)^T(y - x) + \frac{L}{2}\|y - x\|^2. \tag{15}$$

Assumption 2 (Strong Convexity). *Suppose function $F : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex function for $\mu > 0$ and F^* is the optimal value of F, then, we have,*

$$F(y) \geq F(x) + \nabla F(x)^T(y - x) + \frac{\mu}{2}\|y - x\|^2, \tag{16}$$

$$and, \ F(x) - F^* \leq \frac{1}{2\mu}\|\nabla F(x)\|^2 \tag{17}$$

Assumption 3 (Assumption 3 in [12]). *For all $s = 1, 2, ..., S$, the following inequality holds*

$$\mathbb{E}\left[F(w_0^s) - F(w^*)\right] \leq c\mathbb{E}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] \tag{18}$$

where $0 < c \ll m$ is a constant.

We derive our proofs by taking motivation from [12] and [27]. Before providing the proofs, we provide certain lemmas, as given below:

Lemma 1 (3-Point Property [17]). *Let $\hat{z}$ be the optimal solution of the following problem: $\min\limits_{z \in \mathbb{R}^d} \ \frac{\tau}{2}\|z - z_0\|^2 + r(z)$, where $\tau \geq 0$ and $r(z)$ is a convex function (but possibly non-differentiable). Then for any $z \in \mathbb{R}^d$, then the following inequality holds,*

$$\frac{\tau}{2}\|\hat{z} - z_0\|^2 + r(\hat{z}) \leq r(z) + \frac{\tau}{2}\left(\|z - z_0\|^2 - \|z - \hat{z}\|^2\right) \tag{19}$$

Lemma 2 (Theorem 4 in [16]). *For non-smooth problems, taking $\tilde{\nabla}'_{s,k} = \frac{1}{b}\sum_{i\in B_k}\nabla f_i(w_k^s) - \frac{1}{b}\sum_{i\in B_k}\nabla f_i(\tilde{w}^{s-1}) + \frac{1}{n}\sum_{i=1}^n f_i(\tilde{w}^{s-1})$, we have $\mathbb{E}\left[\tilde{\nabla}'_{s,k}\right] = \nabla f(w_k^s)$ and the variance satisfies following inequality,*

$$\mathbb{E}\left[\|\tilde{\nabla}'_{s,k} - \nabla f(w_k^s)\|^2\right] \leq 4L\alpha(b)\left[F(w_k^s) - F(w^*) + F(\tilde{w}^{s-1}) - F(w^*)\right], \qquad (20)$$

where $\alpha(b) = (n-b)/(b(n-1))$.

Following the Lemma 2 for non-smooth problems, one can easily prove the following results for the smooth problems,

Lemma 3. *For smooth problems, taking $\tilde{\nabla}'_{s,k} = \frac{1}{b}\sum_{i\in B_k}\nabla f_i(w_k^s) - \frac{1}{b}\sum_{i\in B_k}\nabla f_i(\tilde{w}^{s-1}) + \frac{1}{n}\sum_{i=1}^n f_i(\tilde{w}^{s-1})$, we have $\mathbb{E}\left[\tilde{\nabla}'_{s,k}\right] = \nabla f(w_k^s)$ and the variance satisfies following inequality,*

$$\mathbb{E}\left[\|\tilde{\nabla}'_{s,k} - \nabla f(w_k^s)\|^2\right] \leq 4L\alpha(b)\left[f(w_k^s) - f^* + f(\tilde{w}^{s-1}) - f^*\right], \qquad (21)$$

where $\alpha(b) = (n-b)/(b(n-1))$.

Lemma 4 (Extension of Lemma 3.4 in [27] to mini-batches). *Under Assumption 1 for smooth regularizer, we have*

$$\mathbb{E}\left[\|\nabla_{B_k}f(w_k^s) - \nabla_{B_k}f(w^*)\|^2\right] \leq 2L\left[f(w_k^s) - f(w^*)\right] \qquad (22)$$

Proof. Given any $k = 0, 1, ..., (m-1)$, consider the function,

$$\phi_{B_k}(w) = f_{B_k}(w) - f_{B_k}(w^*) - \nabla_{B_k}f(w^*)^T(w - w^*)$$

It is straightforward to check that $\nabla\phi_{B_k}(w^*) = 0$, hence $\min_w \phi_{B_k}(w) = \phi_{B_k}(w^*) = 0$. Since $\phi_{B_k}(w)$ is Lipschitz continuous so we have,

$$\frac{1}{2L}\|\nabla\phi_{B_k}(w)\|^2 \leq \phi_{B_k}(w) - \min_w \phi_{B_k}(w) = \phi_{B_k}(w) - \phi_{B_k}(w^*) = \phi_{B_k}(w)$$

$$\implies \|\nabla f_{B_k}(w) - \nabla f_{B_k}(w^*)\|^2 \leq 2L\left[f_{B_k}(w) - f_{B_k}(w^*) - \nabla_{B_k}f(w^*)^T(w - w^*)\right]$$

Taking expectation, we have

$$\mathbb{E}[\|\nabla f_{B_k}(w) - \nabla f_{B_k}(w^*)\|^2] \leq 2L\left[f(w) - f(w^*) - \nabla f(w^*)^T(w - w^*)\right] \qquad (23)$$

By optimality, $\nabla f(w^*) = 0$, we have

$$\mathbb{E}[\|\nabla f_{B_k}(w) - \nabla f_{B_k}(w^*)\|^2] \leq 2L\left[f(w) - f(w^*)\right]$$

This proves the required lemma. $\qquad\qquad\square$

Lemma 5 (Extension of Lemma 3.4 in [27] to mini-batches). *Under Assumption 1 for non-smooth regularizer, we have*

$$\mathbb{E}\left[\|\nabla_{B_k}f(w_k^s) - \nabla_{B_k}f(w^*)\|^2\right] \leq 2L\left[F(w_k^s) - F(w^*)\right] \qquad (24)$$

Proof. From inequality 23, we have

$$\mathbb{E}[\|\nabla f_{B_k}(w) - \nabla f_{B_k}(w^*)\|^2] \leq 2L\left[f(w) - f(w^*) - \nabla f(w^*)^T(w - w^*)\right] \qquad (25)$$

By optimality, there exist $\xi \in \partial g(w^*)$, such that, $\nabla F(w^*) = \nabla f(w^*) + \xi = 0$, we have

$$\begin{aligned}\mathbb{E}[\|\nabla f_{B_k}(w) - \nabla f_{B_k}(w^*)\|^2] &\leq 2L\left[f(w) - f(w^*) + \xi^T(w - w^*)\right]\\ &\leq 2L\left[f(w) - f(w^*) + g(w) - g(w^*)\right]\\ &\leq 2L\left[F(w) - F(w^*)\right]\end{aligned} \qquad (26)$$

second inequality follows from the convexity of g. This proves the required lemma. $\qquad\square$

Lemma 6 (Variance Bound for smooth problem). *Under the Assumption 1 and taking* $\nabla_{B_k} f(w_k^s) = \frac{1}{b} \sum_{i \in B_k} \nabla f_i(w_k^s)$, $\nabla_{B_k'} f(\tilde{w}^{s-1}) = \frac{1}{n} \sum_{i \in B_k} \nabla f_i(\tilde{w}^{s-1})$, $\tilde{\mu}^s = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{w}^{s-1})$ *and the gradient estimator,* $\tilde{\nabla}_{s,k} = \nabla_{B_k} f(w_k^s) - \nabla_{B_k'} f(\tilde{w}^{s-1}) + \tilde{\mu}^s$, *then the variance satisfies the following inequality*[3],

$$\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right] \le 8L\alpha(b)\left[f(w_k^s) - f^*\right] + \frac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2}\left[f(\tilde{w}^{s-1}) - f^*\right] + R' \tag{27}$$

where $\alpha(b) = (n-b)/(b(n-1))$ *and* R' *is a constant.*

Proof. First the expectation of estimator is given by

$$\begin{aligned}
\mathbb{E}\left[\tilde{\nabla}_{s,k}\right] &= \mathbb{E}\left[\nabla_{B_k} f(w_k^s) - \nabla_{B_k'} f(\tilde{w}^{s-1}) + \tilde{\mu}^s\right] \\
&= \nabla f(w_k^s) - \frac{b}{n}\nabla f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) \\
&= \nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1}),
\end{aligned} \tag{28}$$

second equality follows as $n = mb$. Now the variance bound is calculated as follows,

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right] \\
&= \mathbb{E}\left[\|\nabla_{B_k} f(w_k^s) - \nabla_{B_k'} f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) - \nabla f(w_k^s)\|^2\right] \\
&= \mathbb{E}\left[\|\nabla_{B_k} f(w_k^s) - \nabla_{B_k} f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) - \nabla f(w_k^s) + \frac{m-1}{m}\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right] \\
&\le 2\mathbb{E}\left[\|\nabla_{B_k} f(w_k^s) - \nabla_{B_k} f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) - \nabla f(w_k^s)\|^2\right] + \frac{2(m-1)^2}{m^2}\mathbb{E}\left[\|\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right] \\
&\le 8L\alpha(b)\left[f(w_k^s) - f^* + f(\tilde{w}^{s-1}) - f^*\right] + \frac{2(m-1)^2}{m^2}\mathbb{E}\left[\|\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right]
\end{aligned} \tag{29}$$

inequality follows from, $\|a+b\|^2 \le 2\left(\|a\|^2 + \|b\|^2\right)$ for $a, b \in \mathbb{R}^d$ and applying the Lemma 3.

$$\begin{aligned}
\text{Now,} \quad &\frac{2(m-1)^2}{m^2}\mathbb{E}\left[\|\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right] \\
&\le \frac{2(m-1)^2}{m^2}\left[2\mathbb{E}\|\nabla_{B_k} f(\tilde{w}^{s-1}) - \nabla_{B_k} f(w^*)\|^2 + 2\mathbb{E}\|\nabla_{B_k} f(w^*)\|^2\right] \\
&\le \frac{8L(m-1)^2}{m^2}\left[f(\tilde{w}^{s-1}) - f(w^*)\right] + R'
\end{aligned} \tag{30}$$

first inequality follows from, $\|a+b\|^2 \le 2\left(\|a\|^2 + \|b\|^2\right)$ for $a, b \in \mathbb{R}^d$, second inequality follows from Lemma 4 and assuming $\mathbb{E}\|\nabla_{B_k} f(w^*)\|^2 \le R$, $\forall k$ and where taking $R' = \frac{2(m-1)^2}{m^2} * R$. Now, substituting the above inequality in 29, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right] \\
&\le 8L\alpha(b)\left[f(w_k^s) - f^* + f(\tilde{w}^{s-1}) - f^*\right] + \frac{8L(m-1)^2}{m^2}\left[f(\tilde{w}^{s-1}) - f(w^*)\right] + R' \\
&= 8L\alpha(b)\left[f(w_k^s) - f^*\right] + \frac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2}\left[f(\tilde{w}^{s-1}) - f^*\right] + R',
\end{aligned} \tag{31}$$

This proves the required lemma. □

---

[3] For the simplification of proof, we take $f(w) = F(w)$, i.e., $f_i(w) = f_i(w) + g(w)$ $\forall i$ and then $g(w) \equiv 0$

Lemma 7 (Variance Bound for non-smooth problem). *Under Assumption 1 and taking notations as in Lemma 6, the variance bound satisfies the following inequality,*

$$\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right] \le 8L\alpha(b)\left[F(w_k^s) - F^*\right] + \frac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2}\left[F(\tilde{w}^{s-1}) - F^*\right] + R',$$
(32)

where $\alpha(b) = (n-b)/(b(n-1))$ and R' is constant.

Proof.

$$\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right]$$

$$= \mathbb{E}\left[\|\nabla_{B_k} f(w_k^s) - \nabla_{B_k'} f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) - \nabla f(w_k^s)\|^2\right]$$

$$= \mathbb{E}\left[\|\nabla_{B_k} f(w_k^s) - \nabla_{B_k} f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) - \nabla f(w_k^s) + \frac{m-1}{m}\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right]$$

$$\le 2\mathbb{E}\left[\|\nabla_{B_k} f(w_k^s) - \nabla_{B_k} f(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}) - \nabla f(w_k^s)\|^2\right] + \frac{2(m-1)^2}{m^2}\mathbb{E}\left[\|\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right]$$

$$\le 8L\alpha(b)\left[F(w_k^s) - F(w^*) + F(\tilde{w}^{s-1}) - F(w^*)\right] + \frac{2(m-1)^2}{m^2}\mathbb{E}\left[\|\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right]$$
(33)

inequality follows from, $\|a+b\|^2 \le 2\left(\|a\|^2 + \|b\|^2\right)$ for $a, b \in \mathbb{R}^d$ and applying the Lemma 2.

$$\text{Now, } \frac{2(m-1)^2}{m^2}\mathbb{E}\left[\|\nabla_{B_k} f(\tilde{w}^{s-1})\|^2\right]$$

$$\le \frac{2(m-1)^2}{m^2}\left[2\mathbb{E}\|\nabla_{B_k} f(\tilde{w}^{s-1}) - \nabla_{B_k} f(w^*)\|^2 + 2\mathbb{E}\|\nabla_{B_k} f(w^*)\|^2\right]$$
(34)

$$\le \frac{8L(m-1)^2}{m^2}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R'$$

first inequality follows from, $\|a+b\|^2 \le 2\left(\|a\|^2 + \|b\|^2\right)$ for $a, b \in \mathbb{R}^d$, second inequality follows from Lemma 5 and assuming $\|\nabla_{B_k} f(w^*)\|^2 \le R$, $\forall k$ and taking $R' = \frac{2(m-1)^2}{m^2} * R$. Now, substituting the above inequality in 33, we have

$$\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right]$$

$$\le 8L\alpha(b)\left[F(w_k^s) - F^* + F(\tilde{w}^{s-1}) - F^*\right] + \frac{8L(m-1)^2}{m^2}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R'$$
(35)

$$= 8L\alpha(b)\left[F(w_k^s) - F^*\right] + \frac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2}\left[F(\tilde{w}^{s-1}) - F^*\right] + R',$$

This proves the required lemma.   □

Proof of Theorem 1 (Non-strongly convex and smooth problem with SAAG-IV)

Proof. By smoothness, we have,

$$f(w_{k+1}^s) \le f(w_k^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L}{2}\|w_{k+1}^s - w_k^s\|^2$$

$$= f(w_k^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 - \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2$$

$$= f(w_k^s) + <\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + <\nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s>$$

$$- \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2,$$
(36)

where β is appropriately chosen positive value. Now, separately simplifying the terms, we have

$$
\mathbb{E}\left[f(w_k^s)+ <\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s > +\frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2\right]
$$

$$
= f(w_k^s) + \mathbb{E}\left[< \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s >\right] + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$
= f(w_k^s)+ < \nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s > +\frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$
= f(w_k^s)+ < \nabla f(w_k^s), w_{k+1}^s - w_k^s > +\frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + \frac{m-1}{m} < \nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s >
$$

$$
\leq f(w_k^s)+ < \nabla f(w_k^s), w^* - w_k^s > +\frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2
$$

$$
+ \frac{m-1}{m} < \nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s >
$$

$$
= f(w_k^s)+ < \nabla f(w_k^s), w^* - w_k^s > +\frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2
$$

$$
+ \frac{m-1}{m} \left[< \nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w^* > - < \nabla f(\tilde{w}^{s-1}), w_k^s - w^* >\right]
$$

$$
\leq f(w^*) + \frac{L\beta}{2} \left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right]
$$

$$
+ \frac{m-1}{m} \left[\frac{1}{2\delta}\|\nabla f(\tilde{w}^{s-1})\|^2 + \frac{\delta}{2}\|w_{k+1}^s - w^*\|^2 - \left[\frac{1}{2\delta}\|\nabla f(\tilde{w}^{s-1})\|^2 + \frac{\delta}{2}\|w_k^s - w^*\|^2\right]\right]
$$

$$
= f(w^*) + \frac{L\beta}{2} \left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{\delta(m-1)}{2m} \left[\|w_{k+1}^s - w^*\|^2 - \|w_k^s - w^*\|^2\right],
$$

$$
= f(w^*) + \left(\frac{L\beta}{2} - \frac{\delta(m-1)}{2m}\right) \left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right],
$$

$$
= f(w^*),
$$

(37)

second equality follows from, $\mathbb{E}\left[\tilde{\nabla}_{s,k}\right] = \nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1})$, first inequality follows from Lemma 1, second inequality follows from the convexity, i.e., $f(w^*) \geq f(w_k^s)+ < \nabla f(w_k^s), w^* - w_k^s >$ and Young's inequality, i.e., $x^T y \leq 1/(2\delta)\|x\|^2 + \delta/2\|y\|^2$ for $\delta > 0$, and last equality follows by choosing $\delta = \frac{mL\beta}{(m-1)}$.

$$
\text{and, } \mathbb{E}\left[< \nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s > -\frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]
$$

$$
\leq \mathbb{E}\left[\frac{1}{2L(\beta-1)}\|\nabla f(w_k^s) - \tilde{\nabla}_{s,k}\|^2 + \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2 - \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]
$$

$$
= \frac{1}{2L(\beta-1)}\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right]
$$

$$
\leq \frac{1}{2L(\beta-1)} \left[8L\alpha(b) [f(w_k^s) - f^*] + \frac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2} [f(\tilde{w}^{s-1}) - f^*] + R'\right]
$$

$$
= \frac{4\alpha(b)}{(\beta-1)} [f(w_k^s) - f^*] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)} [f(\tilde{w}^{s-1}) - f^*] + R''
$$

(38)

first inequality follows from Young's inequality, second inequality follows from Lemma 6 and $R'' = R'/(2L(\beta-1))$.

Now, substituting the values into 36 from inequalities 37 and 38, and taking expectation w.r.t.

mini-batches, we have

$$\mathbb{E}\left[f(w_{k+1}^s)\right] \le f(w^*) + \frac{4\alpha(b)}{(\beta-1)}\left[f(w_k^s) - f^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$$

$$\mathbb{E}\left[f(w_{k+1}^s) - f(w^*)\right] \le \frac{4\alpha(b)}{(\beta-1)}\left[f(w_k^s) - f^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$$

$$(39)$$

Taking sum over $k = 0, 1, ..., (m-1)$ and dividing by m, we have

$$\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\left[f(w_{k+1}^s) - f^*\right]$$

$$\le \frac{1}{m}\sum_{k=0}^{m-1}\left[\frac{4\alpha(b)}{(\beta-1)}\left[f(w_k^s) - f^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''\right]$$

$$\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[f(w_k^s) - f^*\right] \qquad (40)$$

$$\le \frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\sum_{k=1}^{m}\left[f(w_k^s) - f^*\right] + \frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left[f(w_0^s) - f^* - \{f(w_m^s) - f^*\}\right]$$

$$+ \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$$

Subtracting $\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\sum_{k=1}^{m}\left[f(w_k^s) - f^*\right]$ from both sides, we have

$$\left(1 - \frac{4\alpha(b)}{(\beta-1)}\right)\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[f(w_k^s) - f^*\right]$$

$$\le \frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left[f(w_0^s) - f^* - \{f(w_m^s) - f^*\}\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$$

$$(41)$$

Since $f(w_m^s) - f^* \ge 0$ so dropping this term and using Assumption 3, we have

$$\left(1 - \frac{4\alpha(b)}{(\beta-1)}\right)\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[f(w_k^s) - f^*\right]$$

$$\le \frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left[f(w_0^s) - f^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$$

$$\le \frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left[c\left(f(\tilde{w}^{s-1}) - f^*\right)\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$$

$$= \left[\frac{4\alpha(b)}{(\beta-1)}\frac{c}{m} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\right]\left[\left(f(\tilde{w}^{s-1}) - f^*\right)\right] + R'',$$

$$(42)$$

Dividing both sides by $\left(1 - \frac{4\alpha(b)}{(\beta-1)}\right)$, and since $\tilde{w}^s = 1/m\sum_{k=1}^{m}w_k^s$ so by convexity, $f(\tilde{w}^s) \le 1/m\sum_{k=1}^{m}f(w_k^s)$, we have

$$\mathbb{E}\left[f(\tilde{w}^s) - f^*\right] \le \left[\frac{4\alpha(b)}{(\beta-1-4\alpha(b))}\frac{c}{m} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1-4\alpha(b))}\right]\left[f(\tilde{w}^{s-1}) - f^*\right] + R''',$$

$$(43)$$

where $R''' = R''(\beta-1)/(\beta-1-4\alpha(b))$. Now, applying this inequality recursively, we have

$$\mathbb{E}\left[f(\tilde{w}^s) - f^*\right] \le C^s\left[f(\tilde{w}^0) - f^*\right] + R'''', \qquad (44)$$

inequality follows for $R'''' = R'''/(1-C)$, since $\sum_{i=0}^{k} r^i \leq \sum_{i=0}^{\infty} r^i = \dfrac{1}{1-r}$, $\quad \|r\| < 1$ and

$C = \left[\dfrac{4\alpha(b)}{(\beta - 1 - 4\alpha(b))} \dfrac{c}{m} + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta - 1 - 4\alpha(b))} \right]$. For certain choice of β, one can easily prove that $C < 1$. This proves linear convergence with some initial error.

$\square$

Proof of Theorem 2 (Strongly convex and smooth problem with SAAG-IV)

Proof. By smoothness, we have,

$$
\begin{aligned}
f(w_{k+1}^s) &\leq f(w_k^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L}{2}\|w_{k+1}^s - w_k^s\|^2 \\
&= f(w_k^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 - \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2 \\
&= f(w_k^s) + <\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + <\nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> \\
&\quad - \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2
\end{aligned}
$$

$$(45)$$

where β is appropriately chosen positive value. Now, separately simplifying the terms, we have

$$
\begin{aligned}
&\mathbb{E}\left[f(w_k^s) + <\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 \right] \\
&= f(w_k^s) + \mathbb{E}\left[<\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> \right] + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 \\
&= f(w_k^s) + <\nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 \\
&= f(w_k^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + \frac{m-1}{m} <\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s> \\
&\leq f(w_k^s) + <\nabla f(w_k^s), w^* - w_k^s> + \frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2 \\
&\quad + \frac{m-1}{m} <\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s> \\
&= f(w_k^s) + <\nabla f(w_k^s), w^* - w_k^s> + \frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2 \\
&\quad + \frac{m-1}{m} \left[<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - \tilde{w}^{s-1}> - <\nabla f(\tilde{w}^{s-1}), w_k^s - \tilde{w}^{s-1}> \right] \\
&\leq f(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2 \right] \\
&\quad + \frac{m-1}{m} \left[f(w_{k+1}^s) - f(\tilde{w}^{s-1}) - \left(f(w_k^s) - f(\tilde{w}^{s-1}) \right) \right] \\
&= f(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2 \right] + \frac{m-1}{m} \left[f(w_{k+1}^s) - f(w_k^s) \right],
\end{aligned}
$$

$$(46)$$

second equality follows from, $\mathbb{E}\left[\tilde{\nabla}_{s,k} \right] = \nabla f(w_k^s) + \dfrac{m-1}{m}\nabla f(\tilde{w}^{s-1})$, first inequality follows from Lemma 1 and second inequality follows from the convexity, i.e., $f(x) \geq f(y) + <$

$\nabla f(y), x - y >.$

and, $\mathbb{E}\left[< \nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s > - \dfrac{L(\beta - 1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]$

$\leq \mathbb{E}\left[\dfrac{1}{2L(\beta - 1)}\|\nabla f(w_k^s) - \tilde{\nabla}_{s,k}\|^2 + \dfrac{L(\beta - 1)}{2}\|w_{k+1}^s - w_k^s\|^2 - \dfrac{L(\beta - 1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]$

$\leq \dfrac{1}{2L(\beta - 1)}\left[8L\alpha(b)\left[f(w_k^s) - f^*\right] + \dfrac{8L\left(\alpha(b)m^2 + (m - 1)^2\right)}{m^2}\left[f(\tilde{w}^{s-1}) - f^*\right] + R'\right]$

$= \dfrac{4\alpha(b)}{(\beta - 1)}\left[f(w_k^s) - f^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m - 1)^2\right)}{m^2(\beta - 1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$

$\hspace{12cm}(47)$

first inequality follows from Young's inequality and second inequality follows from Lemma 6 and $R'' = R'/(2L(\beta - 1))$.

Now, substituting the values into 45 from inequalities 46 and 47, and taking expectation w.r.t. mini-batches, we have

$\mathbb{E}\left[f(w_{k+1}^s)\right]$

$\leq f(w^*) + \dfrac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \dfrac{m - 1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right]$

$+ \dfrac{4\alpha(b)}{(\beta - 1)}\left[f(w_k^s) - f^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m - 1)^2\right)}{m^2(\beta - 1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R'',$

$\mathbb{E}\left[f(w_{k+1}^s) - f(w^*)\right]$

$= \dfrac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \dfrac{m - 1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right]$

$+ \dfrac{4\alpha(b)}{(\beta - 1)}\left[f(w_k^s) - f^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m - 1)^2\right)}{m^2(\beta - 1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$

$\hspace{12cm}(48)$

Taking sum over $k = 0, 1, ..., (m - 1)$ and dividing by m, we have

$\dfrac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\left[f(w_{k+1}^s) - f(w^*)\right]$

$\leq \dfrac{1}{m}\sum_{k=0}^{m-1}\left\{\dfrac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \dfrac{m - 1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right]\right\}$

$+ \dfrac{1}{m}\sum_{k=0}^{m-1}\left\{\dfrac{4\alpha(b)}{(\beta - 1)}\left[f(w_k^s) - f^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m - 1)^2\right)}{m^2(\beta - 1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''\right\}$

$\dfrac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[f(w_k^s) - f(w^*)\right]$

$\hspace{12cm}(49)$

$\leq \dfrac{L\beta}{2m}\left[\|w^* - w_0^s\|^2 - \|w^* - w_m^s\|^2\right] + \dfrac{m - 1}{m^2}\left[f(w_m^s) - f(w_0^s)\right]$

$+ \dfrac{4\alpha(b)}{(\beta - 1)}\dfrac{1}{m}\left\{\sum_{k=1}^{m}\left[f(w_k^s) - f^*\right] + f(w_0^s) - f^* - (f(w_m^s) - f^*)\right\}$

$+ \dfrac{4\left(\alpha(b)m^2 + (m - 1)^2\right)}{m^2(\beta - 1)}\left[f(\tilde{w}^{s-1}) - f^*\right] + R''$

Subtracting $\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\sum_{k=1}^{m}\left[f(w_k^s)-f^*\right]$ from both sides, we have

$$
\begin{aligned}
&\left(1-\frac{4\alpha(b)}{(\beta-1)}\right)\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[f(w_k^s)-f^*\right] \\
&\leq \frac{L\beta}{2m}\left[\|w^*-w_0^s\|^2-\|w^*-w_m^s\|^2\right]-\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)[f(w_0^s)-f(w_m^s)] \\
&\quad+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1})-f^*\right]+R'' \\
&\leq \frac{L\beta}{2m}\|w^*-w_0^s\|^2-\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\left[f(w_0^s)-f^*-\{f(w_m^s)-f^*\}\right] \\
&\quad+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1})-f^*\right]+R'' \\
&\leq \frac{L\beta}{2m}\frac{2}{\mu}\left(f(w_0^s)-f^*\right)-\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\left[c\left[f(\tilde{w}^{s-1})-f^*\right]-c\left[f(\tilde{w}^s)-f^*\right]\right] \\
&\quad+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1})-f^*\right]+R'' \\
&\leq \frac{L\beta}{2m}\frac{2}{\mu}c\left[f(\tilde{w}^{s-1})-f^*\right]-\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\left[c\left[f(\tilde{w}^{s-1})-f^*\right]-c\left[f(\tilde{w}^s)-f^*\right]\right] \\
&\quad+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1})-f^*\right]+R''
\end{aligned}
\tag{50}
$$

second inequality follows by dropping, $\|w^*-w_m^s\|^2>0$, third inequality follows from the strong convexity, i.e., $\|w_0^s-w^*\|^2\leq 2/\mu\left(f(w_0^s)-f^*\right)$ and application of Assumption 3 twice, and fourth inequality follows from Assumption 3.

Since $\tilde{w}^s=1/m\sum_{k=1}^{m}w_k^s$ so by convexity using, $f(\tilde{w}^s)\leq 1/m\sum_{k=1}^{m}f(w_k^s)$, we have

$$
\begin{aligned}
&\left(1-\frac{4\alpha(b)}{(\beta-1)}\right)\mathbb{E}\left[f(\tilde{w}^s)-f^*\right] \\
&\leq \frac{L\beta}{2m}\frac{2}{\mu}c\left[f(\tilde{w}^{s-1})-f^*\right]-\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\left[c\left[f(\tilde{w}^{s-1})-f^*\right]-c\left[f(\tilde{w}^s)-f^*\right]\right] \\
&\quad+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[f(\tilde{w}^{s-1})-f^*\right]+R''
\end{aligned}
\tag{51}
$$

Subtracting, $c\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\mathbb{E}\left[f(\tilde{w}^s)-f^*\right]$ both sides, we have

$$
\begin{aligned}
&\left(1-\frac{4\alpha(b)}{(\beta-1)}-c\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\right)\mathbb{E}\left[f(\tilde{w}^s)-f^*\right] \\
&\leq \left[\frac{cL\beta}{m\mu}+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}-\frac{c(m-1)}{m^2}+\frac{4\alpha(b)}{(\beta-1)}\right]\left[f(\tilde{w}^{s-1})-f^*\right]+R''
\end{aligned}
\tag{52}
$$

Dividing both sides by $\left(1-\frac{4\alpha(b)}{(\beta-1)}-c\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\right)$, we have

$$
\mathbb{E}\left[f(\tilde{w}^s)-f^*\right]\leq C\left[f(\tilde{w}^{s-1})-f^*\right]+R'''
\tag{53}
$$

where
$$
C=\left[\frac{cL\beta}{m\mu}+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}-\frac{c(m-1)}{m^2}+\frac{4\alpha(b)}{(\beta-1)}\right]\left(1-\frac{4\alpha(b)}{(\beta-1)}-c\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\right)^{-1}
$$
and $R'''=R''\left(1-\frac{4\alpha(b)}{(\beta-1)}-c\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\right)\right)^{-1}$.

Now, recursively applying the inequality, we have

$$\mathbb{E}\left[f(\tilde{w}^s) - f^*\right] \leq C^s \left[f(\tilde{w}^0) - f^*\right] + R'''', \tag{54}$$

inequality follows for $R'''' = R'''/(1 - C)$, since $\sum_{i=0}^{k} r^i \leq \sum_{i=0}^{\infty} r^i = \dfrac{1}{1-r}, \quad \|r\| < 1$. For certain choice of β, one can easily prove that $C < 1$. This proves linear convergence with some initial error. $\qquad\square$

Proof of Theorem 3 (Non-strongly convex and non-smooth problem with SAAG-IV)

Proof. By smoothness, we have,

$$f(w_{k+1}^s) \leq f(w_k^s) + \;<\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L}{2}\|w_{k+1}^s - w_k^s\|^2 \tag{55}$$

Now, $F(w_{k+1}^s) = f(w_{k+1}^s) + g(w_{k+1}^s)$

$$\leq f(w_k^s) + g(w_{k+1}^s) + \;<\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L}{2}\|w_{k+1}^s - w_k^s\|^2$$

$$= f(w_k^s) + g(w_{k+1}^s) + \;<\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 - \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2$$

$$= f(w_k^s) + g(w_{k+1}^s) + \;<\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + \;<\nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s>$$

$$- \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2 \tag{56}$$

where β is appropriately chosen positive value. Now, separately simplifying the terms, we have

$$\mathbb{E}\left[f(w_k^s) + g(w_{k+1}^s) + \;<\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2\right]$$

$$= f(w_k^s) + g(w_{k+1}^s) + \mathbb{E}\left[<\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s>\right] + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2$$

$$= f(w_k^s) + g(w_{k+1}^s) + \;<\nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2$$

$$= f(w_k^s) + g(w_{k+1}^s) + \;<\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + \frac{m-1}{m}<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s>$$

$$\leq f(w_k^s) + g(w^*) + \;<\nabla f(w_k^s), w^* - w_k^s> + \frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2$$

$$+ \frac{m-1}{m}<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s>$$

$$= f(w_k^s) + g(w^*) + \;<\nabla f(w_k^s), w^* - w_k^s> + \frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2$$

$$+ \frac{m-1}{m}\left[<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w^*> - <\nabla f(\tilde{w}^{s-1}), w_k^s - w^*>\right]$$

$$\leq f(w^*) + g(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right]$$

$$+ \frac{m-1}{m}\left[\frac{1}{2\delta}\|\nabla f(\tilde{w}^{s-1})\|^2 + \frac{\delta}{2}\|w_{k+1}^s - w^*\|^2 - \left[\frac{1}{2\delta}\|\nabla f(\tilde{w}^{s-1})\|^2 + \frac{\delta}{2}\|w_k^s - w^*\|^2\right]\right]$$

$$= F(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{\delta(m-1)}{2m}\left[\|w_{k+1}^s - w^*\|^2 - \|w_k^s - w^*\|^2\right],$$

$$= F(w^*) + \left(\frac{L\beta}{2} - \frac{\delta(m-1)}{2m}\right)\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right],$$

$$= F(w^*), \tag{57}$$

second equality follows from, $\mathbb{E}\left[\tilde{\nabla}_{s,k}\right] = \nabla f(w_k^s) + \dfrac{m-1}{m}\nabla f(\tilde{w}^{s-1})$, first inequality follows from Lemma 1, second inequality follows from the convexity, i.e., $f(w^*) \geq f(w_k^s) + < \nabla f(w_k^s), w^* - w_k^s >$ and Young's inequality, i.e., $x^T y \leq 1/(2\delta)\|x\|^2 + \delta/2\|y\|^2$ for $\delta > 0$, and last equality follows by choosing $\delta = \dfrac{mL\beta}{(m-1)}$.

And, $\mathbb{E}\left[< \nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s > - \dfrac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]$

$$\leq \mathbb{E}\left[\dfrac{1}{2L(\beta-1)}\|\nabla f(w_k^s) - \tilde{\nabla}_{s,k}\|^2 + \dfrac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2 - \dfrac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]$$

$$= \dfrac{1}{2L(\beta-1)}\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right]$$

$$\leq \dfrac{1}{2L(\beta-1)}\left[8L\alpha(b)\left[F(w_k^s) - F^*\right] + \dfrac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2}\left[F(\tilde{w}^{s-1}) - F^*\right] + R'\right]$$

$$= \dfrac{4\alpha(b)}{(\beta-1)}\left[F(w_k^s) - F^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R''$$

(58)

first inequality follows from Young's inequality and second inequality follows from Lemma 7 and $R'' = R'/(2L(\beta-1))$. Now, substituting the values into 56 from inequalities 57 and 58, and taking expectation w.r.t. mini-batches, we have

$$\mathbb{E}\left[F(w_{k+1}^s)\right] \leq F(w^*) + \dfrac{4\alpha(b)}{(\beta-1)}\left[F(w_k^s) - F^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R''$$

$$\mathbb{E}\left[F(w_{k+1}^s) - F(w^*)\right] \leq \dfrac{4\alpha(b)}{(\beta-1)}\left[F(w_k^s) - F^*\right] + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R''$$

(59)

Taking sum over $k = 0, 1, ..., (m-1)$ and dividing by m, we have

$$\dfrac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\left[F(w_{k+1}^s) - F(w^*)\right]$$

$$\leq \dfrac{4\alpha(b)}{(\beta-1)}\dfrac{1}{m}\sum_{k=0}^{m-1}\left[F(w_k^s) - F(w^*)\right] + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''$$

$$\dfrac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[F(w_k^s) - F(w^*)\right]$$

$$\leq \dfrac{4\alpha(b)}{(\beta-1)}\dfrac{1}{m}\left\{\sum_{k=1}^{m}\left[F(w_k^s) - F(w^*)\right] + F(w_0^s) - F(w^*) - \{F(w_m^s) - F(w^*)\}\right\}$$

$$+ \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''$$

(60)

Subtracting $\dfrac{4\alpha(b)}{(\beta-1)}\dfrac{1}{m}\sum_{k=1}^{m}\left[F(w_k^s) - F(w^*)\right]$ from both sides, we have

$$\left(1 - \dfrac{4\alpha(b)}{(\beta-1)}\right)\dfrac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[F(w_k^s) - F(w^*)\right]$$

$$\leq \dfrac{4\alpha(b)}{(\beta-1)}\dfrac{1}{m}\left[F(w_0^s) - F(w^*) - \{F(w_m^s) - F(w^*)\}\right]$$

$$+ \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''$$

(61)

Since $F(w_m^s) - F(w^*) \geq 0$ so dropping this term and using Assumption 3, we have

$$
\left(1 - \frac{4\alpha(b)}{(\beta-1)}\right) \frac{1}{m} \sum_{k=1}^{m} \mathbb{E}\left[F(w_k^s) - F(w^*)\right]
$$

$$
\leq \frac{4\alpha(b)}{(\beta-1)} \frac{1}{m} \left[F(w_0^s) - F(w^*)\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)} \left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''
$$

$$
\leq \frac{4\alpha(b)}{(\beta-1)} \frac{1}{m} \left[c\left(F(\tilde{w}^{s-1}) - F(w^*)\right)\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)} \left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''
$$

$$
= \left(\frac{4\alpha(b)}{(\beta-1)} \frac{c}{m} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\right) \left[\left(F(\tilde{w}^{s-1}) - F(w^*)\right)\right] + R'',
$$

$$(62)$$

Dividing both sides by $\left(1 - \dfrac{4\alpha(b)}{(\beta-1)}\right)$, and since $\tilde{w}^s = 1/m \sum_{k=1}^{m} w_k^s$ so by convexity, $F(\tilde{w}^s) \leq 1/m \sum_{k=1}^{m} F(w_k^s)$, we have

$$
\mathbb{E}\left[F(\tilde{w}^s) - F^*\right] \leq \left(\frac{4\alpha(b)}{(\beta-1-4\alpha(b))} \frac{c}{m} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1-4\alpha(b))}\right) \left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''',
$$

$$(63)$$

where $R''' = R'' \left(1 - \dfrac{4\alpha(b)}{(\beta-1)}\right)^{-1}$. Now, applying above inequality recursively, we have

$$
\mathbb{E}\left[F(\tilde{w}^s) - F(w^*)\right] \leq C^s \left[F(\tilde{w}^0) - F(w^*)\right] + R'''',
$$

$$(64)$$

inequality follows for $R'''' = R'''/(1-C)$, since $\sum_{i=0}^{k} r^i \leq \sum_{i=0}^{\infty} r^i = \dfrac{1}{1-r}, \quad \|r\| < 1$ and $C = \left(\dfrac{4\alpha(b)}{(\beta-1-4\alpha(b))} \dfrac{c}{m} + \dfrac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1-4\alpha(b))}\right)$. For certain choice of β, one can easily prove that $C < 1$. This proves linear convergence with some initial error. $\qquad\square$

Proof of Theorem 4 (Strongly convex and non-smooth problem with SAAG-IV)

Proof. By smoothness, we have,

$$
f(w_{k+1}^s) \leq f(w_k^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L}{2}\|w_{k+1}^s - w_k^s\|^2
$$

Now,

$$
F(w_{k+1}^s) = f(w_{k+1}^s) + g(w_{k+1}^s)
$$

$$
\leq f(w_k^s) + g(w_{k+1}^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$
= f(w_k^s) + g(w_{k+1}^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 - \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$
= f(w_k^s) + g(w_{k+1}^s) + <\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + <\nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s>
$$

$$
- \frac{L(\beta-1)}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$(65)$$

where β is appropriately chosen positive value. Now, separately simplifying the terms, we have

$$
\mathbb{E}\left[f(w_k^s) + g(w_{k+1}^s) + <\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2\right]
$$

$$
= f(w_k^s) + g(w_{k+1}^s) + \mathbb{E}\left[<\tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s>\right] + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$
= f(w_k^s) + g(w_{k+1}^s) + <\nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2
$$

$$
= f(w_k^s) + g(w_{k+1}^s) + <\nabla f(w_k^s), w_{k+1}^s - w_k^s> + \frac{L\beta}{2}\|w_{k+1}^s - w_k^s\|^2 + \frac{m-1}{m}<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s>
$$

$$
\leq f(w_k^s) + g(w^*) + <\nabla f(w_k^s), w^* - w_k^s> + \frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2
$$

$$
+ \frac{m-1}{m}<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - w_k^s>
$$

$$
= f(w_k^s) + g(w^*) + <\nabla f(w_k^s), w^* - w_k^s> + \frac{L\beta}{2}\|w^* - w_k^s\|^2 - \frac{L\beta}{2}\|w^* - w_{k+1}^s\|^2
$$

$$
+ \frac{m-1}{m}\left[<\nabla f(\tilde{w}^{s-1}), w_{k+1}^s - \tilde{w}^{s-1}> - <\nabla f(\tilde{w}^{s-1}), w_k^s - \tilde{w}^{s-1}>\right]
$$

$$
\leq f(w^*) + g(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right]
$$

$$
+ \frac{m-1}{m}\left[f(w_{k+1}^s) - f(\tilde{w}^{s-1}) - \left(f(w_k^s) - f(\tilde{w}^{s-1})\right)\right]
$$

$$
= f(w^*) + g(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{m-1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right],
$$

$$
= F(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{m-1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right],
$$

$$
(66)
$$

second equality follows from, $\mathbb{E}\left[\tilde{\nabla}_{s,k}\right] = \nabla f(w_k^s) + \frac{m-1}{m}\nabla f(\tilde{w}^{s-1})$, first inequality follows from Lemma 1 and second inequality follows from the convexity, i.e., $f(x) \geq f(y) + <\nabla f(y), x - y>$.

And, $\mathbb{E}\left[<\nabla f(w_k^s) - \tilde{\nabla}_{s,k}, w_{k+1}^s - w_k^s> - \frac{L(\beta - 1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]$

$$
\leq \mathbb{E}\left[\frac{1}{2L(\beta - 1)}\|\nabla f(w_k^s) - \tilde{\nabla}_{s,k}\|^2 + \frac{L(\beta - 1)}{2}\|w_{k+1}^s - w_k^s\|^2 - \frac{L(\beta - 1)}{2}\|w_{k+1}^s - w_k^s\|^2\right]
$$

$$
= \frac{1}{2L(\beta - 1)}\mathbb{E}\left[\|\tilde{\nabla}_{s,k} - \nabla f(w_k^s)\|^2\right]
$$

$$
\leq \frac{1}{2L(\beta - 1)}\left[8L\alpha(b)\left[F(w_k^s) - F^*\right] + \frac{8L\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2}\left[F(\tilde{w}^{s-1}) - F^*\right] + R'\right]
$$

$$
= \frac{4\alpha(b)}{(\beta - 1)}\left[F(w_k^s) - F^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta - 1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R'',
$$

$$
(67)
$$

first inequality follows from Young's inequality and second inequality follows from Lemma 7 and $R'' = R'/(2L(\beta - 1))$. Now, substituting the values into 65 from inequalities 66 and 67,

and taking expectation w.r.t. mini-batches, we have

$$
\begin{aligned}
&\mathbb{E}\left[F(w_{k+1}^s)\right]\\
&\leq F(w^*) + \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{m-1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right]\\
&\quad + \frac{4\alpha(b)}{(\beta-1)}\left[F(w_k^s) - F^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R''\\
&\mathbb{E}\left[F(w_{k+1}^s) - F(w^*)\right]\\
&\leq \frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{m-1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right]\\
&\quad + \frac{4\alpha(b)}{(\beta-1)}\left[F(w_k^s) - F^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R''
\end{aligned}
\tag{68}
$$

Taking sum over $k = 0, 1, ..., (m-1)$ and dividing by m, we have

$$
\begin{aligned}
&\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\left[F(w_{k+1}^s) - F(w^*)\right]\\
&\leq \frac{1}{m}\sum_{k=0}^{m-1}\left\{\frac{L\beta}{2}\left[\|w^* - w_k^s\|^2 - \|w^* - w_{k+1}^s\|^2\right] + \frac{m-1}{m}\left[f(w_{k+1}^s) - f(w_k^s)\right]\right\}\\
&\quad + \frac{1}{m}\sum_{k=0}^{m-1}\left\{\frac{4\alpha(b)}{(\beta-1)}\left[F(w_k^s) - F^*\right] + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F^*\right] + R''\right\}\\
&\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[F(w_k^s) - F(w^*)\right]\\
&\leq \frac{L\beta}{2m}\left[\|w^* - w_0^s\|^2 - \|w^* - w_m^s\|^2\right] + \frac{m-1}{m^2}\left[f(w_m^s) - f(w_0^s)\right]\\
&\quad + \frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left\{\sum_{k=1}^{m}\left[F(w_k^s) - F(w^*)\right] + F(w_0^s) - F(w^*) - \left(F(w_m^s) - F(w^*)\right)\right\}\\
&\quad + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1}) - F(w^*)\right] + R''
\end{aligned}
\tag{69}
$$

Subtracting, $\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\sum_{k=1}^{m}\left[F(w_k^s)-F(w^*)\right]$ from both sides, we have

$$
\left(1-\frac{4\alpha(b)}{(\beta-1)}\right)\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[F(w_k^s)-F(w^*)\right]
$$

$$
\leq \frac{L\beta}{2m}\left[\|w^*-w_0^s\|^2-\|w^*-w_m^s\|^2\right]+\frac{m-1}{m^2}\left[f(w_m^s)-f(w_0^s)\right]
$$

$$
+\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left\{F(w_0^s)-F(w^*)-(F(w_m^s)-F(w^*))\right\}
$$

$$
+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{''}
$$

$$
\leq \frac{L\beta}{2m}\left[\|w^*-w_0^s\|^2\right]+\frac{m-1}{m^2}\left[F(w_m^s)-F(w_0^s)\right]+\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left\{F(w_0^s)-F(w^*)-(F(w_m^s)-F(w^*))\right\}
$$

$$
+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{'''}
$$

$$
\leq \frac{L\beta}{2m}\frac{2}{\mu}\left[F(w_0^s)-F(w^*)\right]+\frac{m-1}{m^2}\left[F(w_m^s)-F(w_0^s)\right]+\frac{4\alpha(b)}{(\beta-1)}\frac{1}{m}\left\{F(w_0^s)-F(w^*)-(F(w_m^s)-F(w^*))\right\}
$$

$$
+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{'''}
$$

$$
= \left(\frac{L\beta}{m\mu}+\frac{4\alpha(b)}{m(\beta-1)}-\frac{m-1}{m^2}\right)\left[F(w_0^s)-F(w^*)\right]+\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{m(\beta-1)}\right)\left[F(w_m^s)-F(w^*)\right]
$$

$$
+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{'''}
$$

$$
\leq \left(\frac{L\beta}{m\mu}+\frac{4\alpha(b)}{m(\beta-1)}-\frac{m-1}{m^2}\right)c\left[F(\tilde{w}^{s-1})-F(w^*)\right]+\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{m(\beta-1)}\right)c\left[F(\tilde{w}^{s})-F(w^*)\right]
$$

$$
+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{'''}
$$

$$
\leq \left(\frac{Lc\beta}{m\mu}+\frac{4c\alpha(b)}{m(\beta-1)}-\frac{c(m-1)}{m^2}+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\right)\left[F(\tilde{w}^{s-1})-F(w^*)\right]
$$

$$
+\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{m(\beta-1)}\right)c\left[F(\tilde{w}^{s})-F(w^*)\right]+R^{'''}
$$

$$
\tag{70}
$$

second inequality follows from dropping, $\|w^*-w_m^s\|^2\geq 0$ and converting, $f(w_m^s)-f(w_0^s)$ to $F(w_m^s)-F(w_0^s)$ by introducing some constant, third inequality follows from the strong convexity, i.e., $\|w_0^s-w^*\|^2\leq 2/\mu\left(f(w_0^s)-f^*\right)$, fourth inequality follows from Assumption 3 and $R^{'''}=R^{''}+(m-1)g(w_0^s)/m^2$.

Since $\tilde{w}^s=1/m\sum_{k=1}^{m}w_k^s$ so by convexity using, $f(\tilde{w}^s)\leq 1/m\sum_{k=1}^{m}f(w_k^s)$, and subtracting $\left(\frac{m-1}{m^2}-\frac{4\alpha(b)}{m(\beta-1)}\right)c\left[F(\tilde{w}^s)-F(w^*)\right]$ from both sides, we have

$$
\left(1-\frac{4\alpha(b)}{(\beta-1)}-\frac{c(m-1)}{m^2}+\frac{4c\alpha(b)}{m(\beta-1)}\right)\mathbb{E}\left[F(\tilde{w}^s)-F(w^*)\right]
$$

$$
\leq \left(\frac{Lc\beta}{m\mu}+\frac{4c\alpha(b)}{m(\beta-1)}-\frac{c(m-1)}{m^2}+\frac{4\left(\alpha(b)m^2+(m-1)^2\right)}{m^2(\beta-1)}\right)\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{'''}
$$

$$
\tag{71}
$$

Dividing both sides by $\left(1-\frac{4\alpha(b)}{(\beta-1)}-\frac{c(m-1)}{m^2}+\frac{4c\alpha(b)}{m(\beta-1)}\right)$, we have

$$
\mathbb{E}\left[F(\tilde{w}^s)-F(w^*)\right]\leq C\left[F(\tilde{w}^{s-1})-F(w^*)\right]+R^{''''}
$$

$$
\tag{72}
$$

where
$$C = \left(\frac{Lc\beta}{m\mu} + \frac{4c\alpha(b)}{m(\beta-1)} - \frac{c(m-1)}{m^2} + \frac{4\left(\alpha(b)m^2 + (m-1)^2\right)}{m^2(\beta-1)} \right) \left(1 - \frac{4\alpha(b)}{(\beta-1)} - \frac{c(m-1)}{m^2} + \frac{4c\alpha(b)}{m(\beta-1)} \right)^{-1}$$
and $R'''' = R''' \left(1 - \frac{4\alpha(b)}{(\beta-1)} - \frac{c(m-1)}{m^2} + \frac{4c\alpha(b)}{m(\beta-1)} \right)^{-1}$. Now, applying this inequality recursively, we have

$$\mathbb{E}\left[F(\tilde{w}^s) - F(w^*)\right] \le C^s \left[F(\tilde{w}^0) - F(w^*)\right] + R''''', \tag{73}$$

inequality follows for $R''''' = R''''/(1-C)$, since $\sum_{i=0}^{k} r^i \le \sum_{i=0}^{\infty} r^i = \frac{1}{1-r}, \quad \|r\| < 1$. For certain choice of β, one can easily prove that $C < 1$. This proves linear convergence with some initial error. $\qquad\square$

References

1. Allen-Zhu Z (2017) Katyusha: The First Direct Acceleration of Stochastic Gradient Methods. Journal of Machine Learning Research (to appear) Full version available at `http://arxiv.org/abs/1603.05953`
2. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Lechevallier Y, Saporta G (eds) Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), Springer, Paris, France, pp 177–187, URL `http://leon.bottou.org/papers/bottou-2010`
3. Bottou L, Bousquet O (2007) The tradeoffs of large scale learning. In: Proceedings of the 20th International Conference on Neural Information Processing Systems, Curran Associates Inc., USA, NIPS'07, pp 161–168, URL `http://dl.acm.org/citation.cfm?id=2981562.2981583`
4. Bottou L, Curtis FE, Nocedal J (2016) Optimization Methods for Large-Scale Machine Learning URL `http://arxiv.org/abs/1606.04838`
5. Cauchy AL (1847) Méthode générale pour la résolution des systèmes d'équations simultanées. Compte Rendu des S'eances de L'Acad'emie des Sciences XXV S'erie A(25):536–538
6. Chauhan VK, Dahiya K, Sharma A (2017) Mini-batch block-coordinate based stochastic average adjusted gradient methods to solve big data problems. In: Proceedings of the Ninth Asian Conference on Machine Learning, PMLR, vol 77, pp 49–64, URL `http://proceedings.mlr.press/v77/chauhan17a.html`
7. Chauhan VK, Dahiya K, Sharma A (2018) Problem formulations and solvers in linear svm: a review. Artificial Intelligence Review DOI 10.1007/s10462-018-9614-6, URL `https://doi.org/10.1007/s10462-018-9614-6`
8. Chauhan VK, Sharma A, Dahiya K (2018) Faster learning by reduction of data access time. Applied Intelligence 48(12):4715–4729, DOI 10.1007/s10489-018-1235-x
9. Chauhan VK, Sharma A, Dahiya K (2018) Stochastic Trust Region Inexact Newton Method for Large-scale Machine Learning. arXiv URL `http://arxiv.org/abs/1812.10426`, 1812.10426
10. Csiba D, Richt P (2016) Importance Sampling for Minibatches pp 1–19, arXiv:1602.02283v1
11. Defazio A, Bach F, Lacoste-Julien S (2014) Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, NIPS'14, pp 1646–1654
12. Fanhua S, Zhou K, Cheng J, Tsang IW, Zhang L, Tao D (2018) Vr-sgd: A simple stochastic variance reduction method for machine learning. arXiv URL `https://arxiv.org/abs/1802.09932`
13. Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp 315–323

14. Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23:462–466
15. Konečný J, Richtárik P (2013) Semi-Stochastic Gradient Descent Methods 1:19, URL http://arxiv.org/abs/1312.1666, 1312.1666
16. Konečný J, Liu J, Richtárik P, Takáč M (2016) Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting. IEEE J Sel Top Signal Process 10(2):242–255
17. Lan G (2012) An optimal method for stochastic composite optimization. Math Program 133(1):365–397
18. Le Roux N, Schmidt M, Bach F (2012) A Stochastic Gradient Method with an Exponential Convergence Rate for Strongly-Convex Optimization with Finite Training Sets. Tech. rep., INRIA
19. Lin H, Mairal J, Harchaoui Z (2015) A universal catalyst for first-order optimization. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15, pp 3384–3392
20. Parikh N, Boyd S (2014) Proximal algorithms. Found Trends Optim 1(3):127–239
21. Rakhlin A, Shamir O, Sridharan K (2012) Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML'12, pp 1571–1578
22. Robbins H, Monro S (1951) A stochastic approximation method vol-22:pp. 400–407
23. Shalev-Shwartz S, Zhang T (2013) Stochastic dual coordinate ascent methods for regularized loss. J Mach Learn Res 14(1):567–599
24. Shalev-Shwartz S, Singer Y, Srebro N (2007) Pegasos: Primal estimated sub-gradient solver for svm. In: Proceedings of the 24th International Conference on Machine Learning, ACM, New York, NY, USA, ICML '07, pp 807–814
25. Wang H, Banerjee A (2014) Randomized Block Coordinate Descent for Online and Stochastic Optimization. arXiv preprint (1):1–19, 1407.0107
26. Wright SJ (2015) Coordinate descent algorithms. Math Program 151(1):3–34
27. Xiao L, Zhang T (2014) A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization 24(4):2057–2075
28. Xu Y, Yin W (2015) Block stochastic gradient iteration for convex and nonconvex optimization. SIAM Journal on Optimization 25(3):1686–1716
29. Yang Y, Chen J, Zhu J (2016) Distributing the stochastic gradient sampler for large-scale lda. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '16, pp 1975–1984
30. Yang Z, Wang C, Zhang Z, Li J (2018) Random barzilaiborwein step size for mini-batch algorithms. Eng Appl Artif Intell 72:124 – 135
31. Zebang Shen TMCZ Hui Qian (2017) Accelerated doubly stochastic gradient algorithm for large-scale empirical risk minimization. In: IJCAI
32. Zhang T (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the Twenty-first International Conference on Machine Learning, ACM, New York, NY, USA, ICML '04, pp 116–
33. Zhang Y, Xiao L (2015) Stochastic primal-dual coordinate method for regularized empirical risk minimization. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pp 353–361
34. Zhao T, Yu M, Wang Y, Arora R, Liu H (2014) Accelerated Mini-batch Randomized Block Coordinate Descent Method. Advances in Neural Information Processing Systems pp 3329–3337
35. Zhou L, Pan S, Wang J, Vasilakos AV (2017) Machine learning on big data: Opportunities and challenges. Neurocomputing 237:350 – 361, DOI https://doi.org/10.1016/j.neucom.2017.01.026