SAGA with Arbitrary Sampling

Xun Qian [1] **Zheng Qu** [2] **Peter Richárik** [1 3]

Abstract

We study the problem of minimizing the average of a very large number of smooth functions, which is of key importance in training supervised learning models. One of the most celebrated methods in this context is the SAGA algorithm of Defazio et al. (2014). Despite years of research on the topic, a general-purpose version of SAGA—one that would include arbitrary importance sampling and minibatching schemes—does not exist. We remedy this situation and propose a general and flexible variant of SAGA following the *arbitrary sampling* paradigm. We perform an iteration complexity analysis of the method, largely possible due to the construction of new stochastic Lyapunov functions. We establish linear convergence rates in the smooth and strongly convex regime, and under a quadratic functional growth condition (i.e., in a regime not assuming strong convexity). Our rates match those of the primal-dual method Quartz (Qu et al., 2015) for which an arbitrary sampling analysis is available, which makes a significant step towards closing the gap in our understanding of complexity of primal and dual methods for finite sum problems.

1. Introduction

We consider a convex composite optimization problem

$$\min_{x\in\mathbb{R}^d} P(x) := \left(\sum_{i=1}^{n} \lambda_i f_i(x)\right) + \psi(x), \qquad (1)$$

where $f := \sum_i \lambda_i f_i$ is a conic combination (with coefficients $\lambda_1, \ldots, \lambda_n > 0$) of a very large number of smooth convex functions $f_i : \mathbb{R}^d \to \mathbb{R}$, and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function. We do not assume ψ to be smooth. In particular, ψ can be the indicator function of a nonempty closed convex set, turning problem (1) into a constrained minimization of function f. We are interested in the regime where $n \gg d$, although all our theoretical results hold without this assumption.

In a typical setup in the literature, $\lambda_i = 1/n$ for all $i \in [n] := \{1, 2, \ldots, n\}$, $f_i(x)$ corresponds to the loss of a supervised machine learning model x on example i from a training dataset of size n, and f represents the average loss (i.e., empirical risk). Problems of the form (1) are often called "finite-sum" or regularized empirical risk minimization (ERM) problems, and are of immense importance in supervised learning, essentially forming the dominant training paradigm (Shalev-Shwartz & Ben-David, 2014).

1.1. Variance-reduced methods

One of the most successful methods for solving ERM problems is stochastic gradient descent (SGD) (Robbins & Monro, 1951; Nemirovski et al., 2009) and its many variants, including those with minibatches (Takáč et al., 2013), importance sampling (Needell et al., 2015; Zhao & Zhang, 2015) and momentum (Loizou & Richtárik, 2017a;b).

One of the most interesting developments in recent years concerns *variance-reduced* variants of SGD. The first method in this category is the celebrated[1] stochastic average gradient (SAG) method of Schmidt et al. (2017). Many additional variance-reduced methods were proposed since, including SDCA (Richtárik & Takáč, 2014; Shalev-Shwartz & Zhang, 2013; Shalev-Shwartz, 2016), SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013; Xiao & Zhang, 2014b), S2GD (Konečný & Richtárik, 2017; Konečný et al., 2016), MISO (Mairal, 2015), JacSketch (Gower et al., 2018) and SAGD (Bibi et al., 2018).

1.2. SAGA: the known and the unknown

Since the SAG gradient estimator is not unbiased, SAG is notoriously hard to analyze. Soon after SAG was proposed, the SAGA method (Defazio et al., 2014) was developed, obtained by replacing the biased SAG estimator by a simi-

*Equal contribution [1]King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia [2]University of Hong Kong, Hong Kong [3]Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia. Correspondence to: Peter Richtárik <peter.richtarik@kaust.edu.sa>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]Schmidt et al. (2017) received the 2018 Lagrange Prize in continuous optimization for their work on SAG.

lar, but unbiased, SAGA estimator. This method admits a simpler analysis, retaining the favourable convergence properties of SAG. SAGA is one of the early and most successful variance-reduced methods for (1).

Better understanding of the behaviour of SAGA remains one of the open challenges in the literature. Consider problem (1) with $\lambda_i = 1/n$ for all i. Assume, for simplicity, that each f_i is L_i-smooth and f is μ-strongly convex. In this regime, the iteration complexity of SAGA with uniform sampling probabilities is $\mathcal{O}((n+\frac{L_{\max}}{\mu})\log\frac{1}{\epsilon})$, where $L_{\max} := \max_i L_i$, which was established already by Defazio et al. (2014). Schmidt et al. (2015) conjectured that there exist nonuniform sampling probabilities for which the complexity improves to $\mathcal{O}((n+\frac{\bar{L}}{\mu})\log\frac{1}{\epsilon})$, where $\bar{L} := \sum_i L_i/n$. However, the "correct" importance sampling strategy leading to this result was not discovered until recently in the work of Gower et al. (2018), where the conjecture was resolved in the affirmative. One of the key difficulties in the analysis was the construction of a suitable stochastic Lyapunov function controlling the iterative process. Likewise, until recently, very little was known about the minibatch performance of SAGA, even for the simplest *uniform* minibatch strategies. Notable advancements in this area were made by Gower et al. (2018), who have the currently best rates for SAGA with standard uniform minibatch strategies and the first importance sampling results for a block variant of SAGA.

1.3. Contributions

SAGA with arbitrary sampling. We study the performance of SAGA under fully general data sampling strategies known in the literature as *arbitrary sampling*, generalizing all previous results, and obtaining an array of new theoretically and practically useful samplings. We call our general method SAGA-AS. Our theorems are expressed in terms of new Lyapunov functions, the constructions of which was essential to our success.

In the arbitrary sampling paradigm, first proposed by Richtárik & Takáč (2016) in the context of randomized coordinate descent methods, one considers all (proper) random set valued mappings S (called "samplings") with values being subsets of $[n]$. A sampling is uniquely determined by assigning a probability to all 2^n subsets of $[n]$. A sampling is called proper[2] if probability of each $i \in [n]$ being sampled is positive; that is, if $p_i := \mathbb{P}(i \in S) > 0$ for all i. The term "arbitrary sampling" refers to an arbitrary proper sampling.

Smooth case. We perform an iteration complexity analysis in the smooth case ($\psi \equiv 0$), assuming f is μ-strongly convex. Our analysis generalizes the results of Defazio et al. (2014) and Gower et al. (2018) to arbitrary sampling. The JacSketch method Gower et al. (2018) and its analysis rely on the notion of a bias-correcting random variable. Unfortunately, such a random variable does not exist for SAGA-AS. We overcome this obstacle by proposing a *bias-correcting random vector (BCRV)* which, as we show, always exists. While Gower et al. (2018); Bibi et al. (2018) consider particular suboptimal choices, *we are able to find the BCRV which minimizes the iteration complexity bound*. Unlike all known and new variants of SAGA considered in (Gower et al., 2018), *SAGA-AS does not arise as a special case of JacSketch*. Our linear rates for SAGA-AS are the same as those for the primal-dual stochastic fixed point method Quartz (Qu et al., 2015) (the first arbitrary sampling based method for (1)) in the regime when Quartz is applicable, which is the case when an explicit strongly convex regularizer is present. In contrast, we do not need an explicit regularizer, which means that *SAGA-AS can utilize the strong convexity of f fully*, even if the strong convexity parameter μ is not known. While the importance sampling results in (Gower et al., 2018) require each f_i to be strongly convex, we impose this requirement on f only.

Nonsmooth case. We perform an iteration complexity analysis in the general nonsmooth case. When the regularizer ψ is strongly convex, which is the same setting as that considered in (Qu et al., 2015), our iteration complexity bounds are essentially the same as that of Quartz. However, we also prove linear convergence results, with the same rates, under a quadratic functional growth condition (which does not imply strong convexity) (Necoara et al., 2018). These are *the first linear convergence result for any variant of SAGA without strong convexity.* Moreover, to the best of our knowledge, *SAGA-AS is the only variance-reduced method which achieves linear convergence without any a priori knowledge of the error bound condition number.*

Our arbitrary sampling rates are summarized in Table 1.

1.4. Brief review of arbitrary sampling results

The arbitrary sampling paradigm was proposed by Richtárik & Takáč (2016), where a randomized coordinate descent method with arbitrary sampling of subsets of coordinates was analyzed for unconstrained minimization of a strongly convex function. Subsequently, the primal-dual method Quartz with arbitrary sampling of dual variables was studied in Qu et al. (2015) for solving (1) in the case when ψ is strongly convex (and $\lambda_i = \frac{1}{n}$ for all i). An accelerated randomized coordinate descent method with arbitrary sampling called ALPHA was proposed by Qu & Richtárik (2016a) in the context of minimizing the sum of a smooth convex function and a convex block-separable regularizer. A key concept in the analysis of all known methods in the arbi-

[2] It does not make sense to consider samplings S that are not proper. Indeed, if $p_i = 0$ for some i, a method based on S will lose access to f_i and, consequently, ability to solve (1).

Regime	Arbitrary sampling	Thm
Smooth $\psi \equiv 0$ f_i is L_i-smooth, f is μ-strongly convex	$\max\left\{\max\limits_{1\le i\le n}\left\{\frac{1}{p_i}+\frac{4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}{\mu}\right\}, \frac{2\mathcal{B}(1+1/\mathcal{B})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$	3.3
Nonsmooth P satisfies μ-growth condition (19) and Assumption 4.3 $f_i(x)=\phi_i(\mathbf{A}_i^\top x)$, ϕ_i is $1/\gamma$-smooth, f is L-smooth	$\left(2+\max\left\{\frac{6L}{\mu}, 3\max\limits_{1\le i\le n}\left\{\frac{1}{p_i}+\frac{4v_i\lambda_i}{p_i\mu\gamma}\right\}\right\}\right)\log\left(\frac{1}{\epsilon}\right)$	4.4
Nonsmooth ψ is μ-strongly convex $f_i(x)=\phi_i(\mathbf{A}_i^\top x)$, ϕ_i is $1/\gamma$-smooth	$\max\limits_{1\le i\le n}\left\{1+\frac{1}{p_i}+\frac{3v_i\lambda_i}{p_i\mu\gamma}\right\}\log\left(\frac{1}{\epsilon}\right)$	4.5

Table 1. Iteration complexity results for SAGA-AS. We have $p_i := \mathbb{P}(i \in S)$, where S is a sampling of subsets of $[n]$ utilized by SAGA-AS. The key complexity parameters $\mathcal{A}_i$, $\mathcal{B}$, and v_i are defined in the sections containing the theorems.

trary sampling paradigm is the notion of *expected separable overapproximation (ESO)*, introduced by Richtárik & Takáč (2016), and in the context of arbitrary sampling studied in depth by Qu & Richtárik (2016b). A stochastic primal-dual hybrid gradient algorithm (aka Chambolle-Pock) with arbitrary sampling of dual variables was studied by Chambolle et al. (2017). Recently, an accelerated coordinate descent method with arbitrary sampling for minimizing a smooth and strongly convex function was studied by Hanzely & Richtárik (2018). Finally, the first arbitrary sampling analysis in a nonconvex setting was performed by Horváth & Richtárik (2018), which is also the first work in which the optimal sampling out of class of all samplings of a given minibatch size was identified.

2. The Algorithm

Let $F : \mathbb{R}^d \to \mathbb{R}^n$ and $\mathbf{G} : \mathbb{R}^d \to \mathbb{R}^{d\times n}$ be defined by $F(x) := (f_1(x), \cdots, f_n(x))^\top \in \mathbb{R}^n$ and $\mathbf{G}(x) := [\nabla f_1(x), \cdots, \nabla f_n(x)] \in \mathbb{R}^{d\times n}$. We refer to $\mathbf{G}(x)$ as the Jacobian of F at x.

2.1. JacSketch

Gower et al. (2018) propose a new family of variance reduced SGD methods—called JacSketch—which progressively build a variance-reduced estimator of the gradient via the utilization of a new technique they call *Jacobian sketching*. As shown in (Gower et al., 2018), state-of-the-art variants SAGA can be obtained as a special case of JacSketch. However, *SAGA-AS does not arise as a special case of JacSketch*. In fact, the generic analysis provided in (Gower et al., 2018) (Thm 3.6) is too coarse and does not lead to good bounds for any variants of SAGA with importance sampling. On the other hand, the analysis of Gower et al. (2018) which does do well for importance sampling does not generalize to arbitrary sampling, regularized objectives or regimes without strong convexity.

In this section we provide a brief review of the JacSketch method, establishing some useful notation along the way, with the goal of pointing out the moment of departure from JacSketch construction which leads to SAGA-AS. The iterations of JacSketch have the form

$$x^{k+1} = x^k - \alpha g^k, \tag{2}$$

where $\alpha > 0$ is a fixed step size, and g^k is an variance-reduced unbiased estimator of the gradient $\nabla f(x^k)$ built iteratively through a process involving Jacobian sketching, sketch-and-project and bias-correction.

Starting from an arbitrary matrix $\mathbf{J}^0 \in \mathbb{R}^{d\times n}$, at each $k \ge 0$ of JacSketch an estimator $\mathbf{J}^k \in \mathbb{R}^{d\times n}$ of the true Jacobian $\mathbf{G}(x^k)$ is constructed using a sketch-and-project iteration:

$$\begin{aligned}\mathbf{J}^{k+1} = \quad & \arg\min_{\mathbf{J}\in\mathbb{R}^{d\times n}} \|\mathbf{J}-\mathbf{J}^k\| \\ & \text{subject to } \mathbf{J}\mathbf{S}_k = \mathbf{G}(x^k)\mathbf{S}_k.\end{aligned} \tag{3}$$

Above, $\|\mathbf{X}\| := \sqrt{\mathrm{Tr}(\mathbf{X}\mathbf{X}^\top)}$ is the Frobenius norm[3], and $\mathbf{S}_k \in \mathbb{R}^{n\times\tau}$ is a random matrix drawn from some ensamble of matrices $\mathcal{D}$ in an i.i.d. fashion in each iteration. The solution to (3) is the closest matrix to $\mathbf{J}^k$ consistent with true Jacobian in its action onto $\mathbf{S}_k$. Intuitively, the higher τ is, the more accurate $\mathbf{J}^{k+1}$ will be as an estimator of the true Jacobian $\mathbf{G}(x^k)$. However, in order to control the cost of computing $\mathbf{J}^{k+1}$, in practical applications one chooses $\tau \ll n$. In the case of standard SAGA, for instance, $\mathbf{S}_k$ is a random standard unit basis vector in $\mathbb{R}^n$ chosen uniformly at random (and hence $\tau = 1$).

The projection subproblem (3) has the explicit solution (see Lemma B.1 in (Gower et al., 2018)):

$$\mathbf{J}^{k+1} = \mathbf{J}^k + (\mathbf{G}(x^k) - \mathbf{J}^k)\mathbf{\Pi}_{\mathbf{S}_k},$$

where $\mathbf{\Pi}_{\mathbf{S}} := \mathbf{S}(\mathbf{S}^\top\mathbf{S})^\dagger\mathbf{S}^\top$ is an orthogonal projection matrix (onto the column space of $\mathbf{S}$), and $\dagger$ denotes the Moore-Penrose pseudoinverse. Since $\mathbf{J}^{k+1}$ is constructed to be an approximation of $\mathbf{G}(x^k)$, and since $\nabla f(x^k) = \mathbf{G}(x^k)\lambda$, where $\lambda := (\lambda_1, \ldots, \lambda_n)^\top$, it makes sense to estimate the

[3] Gower et al. (2018) consider a *weighted* Frobenius norm, but this is not useful for our purposes.

gradient via $\nabla f(x^k) \approx \mathbf{J}^{k+1}\lambda$. However, this gradient estimator is not unbiased, which poses dramatic challenges for complexity analysis. Indeed, the celebrated SAG method, with its infamously technical analysis, uses precisely this estimator in the special case when $\mathbf{S}_k$ is chosen to be a standard basis vector in $\mathbb{R}^n$ sampled uniformly at random. Fortunately, as show in (Gower et al., 2018), an unbiased estimator can be constructed by taking a random linear combination of $\mathbf{J}^{k+1}\lambda$ and $\mathbf{J}^k\lambda$:

$$\begin{aligned} g^k &= (1-\theta_{\mathbf{S}_k})\mathbf{J}^k\lambda + \theta_{\mathbf{S}_k}\mathbf{J}^{k+1}\lambda \\ &= \mathbf{J}^k\lambda + \theta_{\mathbf{S}_k}(\mathbf{G}(x^k)-\mathbf{J}^k)\mathbf{\Pi}_{\mathbf{S}_k}\lambda, \end{aligned} \tag{4}$$

where $\theta = \theta_{\mathbf{S}} \in \mathbb{R}$ is a bias-correcting random variable (dependent on $\mathbf{S}$), defined as any random variable for which $\mathbb{E}[\theta_{\mathbf{S}}\mathbf{\Pi}_{\mathbf{S}}\lambda] = \lambda$. Under this condition, g^k becomes an unbiased estimator of $\nabla f(x^k)$. The JacSketch method is obtained by alternating optimization steps (2) (producing iterates x^k) with sketch-and-project steps (producing $\mathbf{J}^k$).

2.2. Bias-correcting random vector

In order to construct SAGA-AS, we take a departure here and consider a *bias-correcting random vector* $(\theta^1_{\mathbf{S}}, \cdots, \theta^n_{\mathbf{S}})^\top \in \mathbb{R}^n$ instead. From now on it will be useful to think of $\theta_{\mathbf{S}}$ as an $n \times n$ diagonal matrix, with the vector $(\theta^1_{\mathbf{S}}, \cdots, \theta^n_{\mathbf{S}})$ embedded in its diagonal. In contrast to (4), we propose to construct g^k via $g^k = \mathbf{J}^k\lambda + (\mathbf{G}(x^k) - \mathbf{J}^k)\theta_{\mathbf{S}_k}\mathbf{\Pi}_{\mathbf{S}_k}\lambda$. It is easy to see that under the following assumption, g^k will be an unbiased estimator of $\nabla f(x^k)$.

Assumption 2.1 (Bias-correcting random vector)**.** We say that the diagonal random matrix $\theta_{\mathbf{S}} \in \mathbb{R}^{n\times n}$ is a bias-correcting random vector if

$$\mathbb{E}[\theta_{\mathbf{S}}\mathbf{\Pi}_{\mathbf{S}}\lambda] = \lambda. \tag{5}$$

2.3. Choosing distribution $\mathcal{D}$

In order to complete the description of SAGA-AS, we need to specify the distribution $\mathcal{D}$. We choose $\mathcal{D}$ to be a distribution over random column submatrices of the $n \times n$ identity matrix $\mathbf{I}$. Such a distribution is uniquely characterized by a random subset of the columns of $\mathbf{I} \in \mathbb{R}^{n\times n}$, i.e., a random subset of $[n]$. This leads us to the notion of a *sampling*, already outlined in the introduction.

Definition 2.2 (Sampling)**.** A *sampling* S is a random set-valued mapping with values being the subsets of $[n]$. It is uniquely characterized by the choice of probabilities $p_C := \mathbb{P}[S = C]$ associated with every subset C of $[n]$. Given a sampling S, we let $p_i := \mathbb{P}[i \in S] = \sum_{C:i\in C} p_C$. We say that a sampling S is i) *proper* if $p_i > 0$ for all i, ii) *serial* if $|S| = 1$ with probability 1, iii) τ*-nice* if it selects from all subsets of $[n]$ of cardinality τ uniformly at random, and iv) *independent*[a] if it is formed as follows: for each $i \in [n]$ we flip a biased coin, independently, with probability of success $p_i > 0$; if we are successful, we include i in S.

[a]Independent samplings were also considered for nonconvex ERM problems in (Horváth & Richtárik, 2018) and for accelerated coordinate descent in (Hanzely & Richtárik, 2018).

Given a proper sampling S, we sample matrices $\mathbf{S} \sim \mathcal{D}$ as follows: i) Draw a random set S, ii) Define $\mathbf{S} = \mathbf{I}_{:S} \in \mathbb{R}^{n\times|S|}$ (random column submatrix of $\mathbf{I}$ corresponding to columns $i \in S$). For $h = (h_1, \cdots, h_n)^\top \in \mathbb{R}^n$ and sampling S define vector $h_S \in \mathbb{R}^n$ as follows:

$$(h_S)_i := h_i 1_{(i\in S)}, \text{ where } 1_{(i\in S)} := \begin{cases} 1, & \text{if } i \in S \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to observe (see Lemma 4.7 in (Gower et al., 2018)) that for $\mathbf{S} = \mathbf{I}_{:S}$ we have the identity

$$\mathbf{\Pi}_{\mathbf{S}} = \mathbf{\Pi}_{\mathbf{I}_{:S}} = \mathbf{I}_S := \mathrm{Diag}(e_S). \tag{6}$$

To simplify notation, we will write θ_S instead of $\theta_{\mathbf{S}} = \theta_{\mathbf{I}_{:S}}$.

Lemma 2.3. Let S be a proper sampling and define $\mathcal{D}$ by setting $\mathbf{S} = \mathbf{I}_{:S}$. Then condition (5) is equivalent to

$$\mathbb{E}[\theta^i_S 1_{(i\in S)}] \equiv \textstyle\sum_{C\subseteq[n]:i\in C} p_C \theta^i_C = 1, \quad \forall i \in [n]. \tag{7}$$

This condition is satisfied by the *default vector* $\theta^i_S \equiv \frac{1}{p_i}$.

In general, there is an infinity of bias-correcting random vectors characterized by (7). In SAGA-AS we reserve the freedom to choose any of these vectors.

2.4. SAGA-AS

By putting all of the development above together, we have arrived at SAGA-AS (Algorithm 1). Note that since we consider problem (1) with a regularizer ψ, the optimization step involves a proximal operator, defined as

$$\mathrm{prox}^{\psi}_{\alpha}(x) := \arg\min \left\{ \tfrac{1}{2\alpha}\|x-y\|^2 + \psi(y) \right\}, \ \alpha > 0.$$

To shed more light onto the key steps of SAGA-AS, note that an alternative way of writing the Jacobian update is $\mathbf{J}^{k+1}_{:i} = \nabla f_i(x^k)$ for $i \in S_k$ and $\mathbf{J}^{k+1}_{:i} = \mathbf{J}^k_{:i}$ for $i \notin S_k$. The gradient estimate can be alternatively written as

$$\begin{aligned} g^k &= \textstyle\sum_{i=1}^n \lambda_i \mathbf{J}^k_{:i} + \sum_{i\in S_k} \lambda_i \theta^i_{S_k} \left[\nabla f_i(x^k) - \mathbf{J}^k_{:i}\right] \\ &= \textstyle\sum_{i\notin S_k} \lambda_i \mathbf{J}^k_{:i} + \sum_{i\in S_k} \lambda_i \left[\theta^i_{S_k}\nabla f_i(x^k) + (1-\theta^i_{S_k})\mathbf{J}^k_{:i}\right]. \end{aligned}$$

3. Analysis in the Smooth Case

In this section we consider problem (1) in the smooth case; i.e., we let $\psi \equiv 0$. We make the following assumption on S.

Algorithm 1 SAGA with Arbitrary Sampling (SAGA-AS)

Parameters: Arbitrary proper sampling S; bias-correcting random vector θ_S; stepsize $\alpha > 0$
Initialization: Choose $x^0 \in \mathbb{R}^d$, $\mathbf{J}^0 \in \mathbb{R}^{d\times n}$
for $k = 0, 1, 2, \ldots$ **do**
  Sample a fresh set $S_k \sim S \subseteq [n]$
  $\mathbf{J}^{k+1} = \mathbf{J}^k + (\mathbf{G}(x^k) - \mathbf{J}^k)\mathbf{I}_{S_k}$
  $g^k = \mathbf{J}^k\lambda + (\mathbf{G}(x^k) - \mathbf{J}^k)\theta_{S_k}\mathbf{I}_{S_k}\lambda$
  $x^{k+1} = \text{prox}^{\psi}_{\alpha}\left(x^k - \alpha g^k\right)$
end for

Assumption 3.1. There exists constants $\mathcal{A}_i \geq 0$ for $1 \leq i \leq n$ and $0 \leq \mathcal{B} \leq 1$ such that for any matrix $\mathbf{M} \in \mathbb{R}^{d\times n}$,

$$\mathbb{E}[\|\mathbf{M}\theta_S\mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] \leq \sum_i \mathcal{A}_i\lambda_i^2\|\mathbf{M}_{:i}\|^2 + \mathcal{B}\|\mathbf{M}\lambda\|^2. \quad (8)$$

3.1. Main result

Given an arbitrary proper sampling S, and bias-correcting random vector θ_S, for each $i \in [n]$ define

$$\beta_i := \textstyle\sum_{C\subseteq[n]:i\in C} p_C|C|(\theta_C^i)^2, \quad i \in [n], \quad (9)$$

where $|C|$ is the cardinality of the set C. As we shall see, these quantities play a key importance in our complexity result, presented next.

Lemma 3.2. (i) For an arbitrary proper sampling S, the Assumption 3.1 is satisfied by $\mathcal{A}_i = \beta_i$ and $\mathcal{B} = 0$. (ii) For τ-nice sampling S with $\theta_S^i = \frac{1}{p_i}$, Assumption 3.1 is satisfied by $\mathcal{A}_i = \frac{n}{\tau}\cdot\frac{n-\tau}{n-1}$ and $\mathcal{B} = \frac{n(\tau-1)}{\tau(n-1)}$. (iii) For independent sampling S with $\theta_S^i = \frac{1}{p_i}$, Assumption 3.1 is satisfied by $\mathcal{A}_i = (\frac{1}{p_i} - 1)$ and $\mathcal{B} = 1$.

Theorem 3.3. Let S be an arbitrary proper sampling, and let θ_S be a bias-correcting random vector satisfying (7). Let f be μ-strongly convex and L-smooth, f_i be convex and L_i-smooth. Let $\{x^k, \mathbf{J}^k\}$ be the iterates produced by Algorithm 1. Consider the stochastic Lyapunov function

$$\Psi^k := \|x^k - x^*\|^2 + 2\alpha\textstyle\sum_{i=1}^n \sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^k - \nabla f_i(x^*)\|^2,$$

where $\sigma_i = \frac{1}{4(1+\mathcal{B})L_i\mathcal{A}_ip_i\lambda_i}$ for all i. If stepsize α satisfies

$$\alpha \leq \min\left\{\min_{1\leq i\leq n}\frac{p_i}{\mu+4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_ip_i}, \frac{\mathcal{B}^{-1}}{2(1+1/\mathcal{B})L}\right\} \quad (10)$$

then $\mathbb{E}[\Psi^k] \leq (1-\mu\alpha)^k\mathbb{E}[\Psi^0]$. This implies that if we choose α equal to the upper bound in (10), then $\mathbb{E}[\Psi^k] \leq \epsilon\cdot\mathbb{E}[\Psi^0]$ as long as $k \geq \max\left\{\max_i\left\{\frac{1}{p_i} + \frac{4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}{\mu}\right\}, \frac{2\mathcal{B}(1+1/\mathcal{B})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$.

If μ is unknown and we choose

$$\alpha \leq \min\left\{\min_{1\leq i\leq n}\frac{p_i}{8(1+\mathcal{B})L_i\mathcal{A}_i\lambda_ip_i}, \frac{\mathcal{B}^{-1}}{2(1+1/\mathcal{B})L}\right\}, \quad (11)$$

then $\mathbb{E}[\Psi^k] \leq (1-\min\{\mu\alpha, \frac{p_i}{2}\})^k\mathbb{E}[\Psi^0]$. This implies that if we choose α equal to the upper bound in (11), then we can get $\mathbb{E}[\Psi^k] \leq \epsilon\cdot\mathbb{E}[\Psi^0]$ as long as $k \geq \max\left\{\max_i\left\{\frac{2}{p_i}, \frac{8(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}{\mu}\right\}, \frac{2\mathcal{B}(1+1/\mathcal{B})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$.

Our result involves a novel *stochastic* Lyapunov function Ψ^k, different from that in (Gower et al., 2018).

3.2. Optimal bias-correcting random vector

Note that for an arbitrary proper sampling the complexity bound gets better as β_i get smaller. Having said that, even for a fixed sampling S, the choice of β_i is not unique, Indeed, this is because β_i depends on the choice of θ_S. In view of Lemma 2.3, we have many choices for this random vector. Let $\Theta(S)$ be the collection of all bias-correcting random vectors associated with sampling S. In our next result we will compute the bias-correcting random vector θ_S which leads to the minimal complexity parameters β_i. In the rest of the paper, let $\mathbb{E}^i[\cdot] := \mathbb{E}[\cdot \mid i \in S]$.

Lemma 3.4. Let S be a proper sampling. Then
(i) $\min_{\theta\in\Theta(S)}\beta_i = \frac{1}{\sum_{C:i\in C}p_C/|C|} = \frac{1}{p_i\mathbb{E}^i[1/|S|]}$ for all i, and the minimum is obtained at $\theta \in \Theta(S)$ given by $\theta_C^i = \frac{1}{|C|\sum_{C:i\in C}p_C/|C|} = \frac{1}{p_i|C|\mathbb{E}^i[1/|S|]}$ for all $C : i \in C$;
(ii) $\frac{1}{\mathbb{E}^i[1/|S|]} \leq \mathbb{E}^i[|S|]$, for all i.

3.3. Importance Sampling for Minibatches

In this part we construct an importance sampling for minibatches. This is in general a daunting task, and only a handful of papers exist on this topic. In particular, Csiba & Richtárik (2018) and Hanzely & Richtárik (2019) focused on coordinate descent methods, and Gower et al. (2018) considered minibatch SAGA with importance sampling over subsets of $[n]$ forming a partition.

Let $\tau := \mathbb{E}[|S|]$ be the expected minibatch size, and $\bar{L} := \sum_{i\in[n]}L_i\lambda_i$. We consider the independent sampling with $\theta_S^i = \frac{1}{p_i}$. From Lemma 3.2 and Thm 3.3, the iteration complexity becomes

$$\max\left\{\max_{1\leq i\leq n}\left\{\frac{1}{p_i} + \frac{8L_i\mathcal{A}_i\lambda_i}{\mu}\right\}, \frac{4\bar{L}}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right),$$

where $\mathcal{A}_i = \frac{1}{p_i} - 1$. Hence

$$\frac{1}{p_i} + \frac{8L_i\mathcal{A}_i\lambda_i}{\mu} = \frac{\mu+8L_i\lambda_i(1-p_i)}{\mu p_i}. \quad (12)$$

Let $q_i = \frac{(\mu+8L_i\lambda_i)\tau}{\sum_{i\in[n]}(\mu+8L_i\lambda_i)}$, and $T = \{i \mid q_i > 1\}$. Next, we

discuss two cases:

Case 1. Suppose $T = \emptyset$. In this case, we choose $p_i = q_i$. From (12), we have

$$\frac{1}{p_i} + \frac{8L_i\mathcal{A}_i\lambda_i}{\mu} \leq \frac{\mu+8L_i\lambda_i}{\mu p_i} = \frac{\sum_{i\in[n]}(\mu+8L_i\lambda_i)}{\mu\tau} = \frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}.$$

Theorefore, the iteration complexity has the following upper bound: $\max\left\{\frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}, \frac{4L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$.

Case 2. Suppose $T \neq \emptyset$. In this case, we choose $p_i = 1$ for $i \in T$ and $q_i \leq p_i \leq 1$ for $i \notin T$ such that $\sum_{i\in[n]} p_i = \tau$. Notice that $p_i = 1$ means $\mathcal{A}_i = 0$. Hence, for $i \in T$, from (12), we have $\frac{1}{p_i} + \frac{8L_i\mathcal{A}_i\lambda_i}{\mu} = \frac{\mu+8L_i\lambda_i(1-p_i)}{\mu p_i} = 1$. For $i \notin T$, from (12), we have

$$\frac{1}{p_i} + \frac{8L_i\mathcal{A}_i\lambda_i}{\mu} \leq \frac{\mu+8L_i\lambda_i}{\mu p_i} \leq \frac{\sum_{i\in[n]}(\mu+8L_i\lambda_i)}{\mu\tau} = \frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}.$$

Theorefore, the iteration complexity also has the following upper bound: $\max\left\{\frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}, \frac{4L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$. To summarize the above two cases, by choosing $\min\{q_i, 1\} \leq p_i \leq 1$ such that $\sum_{i\in[n]} p_i = \tau$, the iteration complexity admits the following upper bound:

$$\max\left\{\frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}, \frac{4L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right). \tag{13}$$

It should be noticed that in practice, we can just choose $p_i = \min\{q_i, 1\}$ for convenience, and then (13) also holds, but with $\mathbb{E}[|S|] = \sum_{i\in[n]} p_i \leq \tau$.

Linear speedup. When $\tau \leq \frac{n\mu+8\bar{L}}{4L}$, (13) becomes $\left(\frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}\right)\log\left(\frac{1}{\epsilon}\right)$, which yields linear speedup with respect to τ. When $\tau \geq \frac{n\mu+8\bar{L}}{4L}$, (13) becomes $\frac{4L}{\mu}\log\left(\frac{1}{\epsilon}\right)$.

3.4. SAGA-AS vs Quartz

In this section, we compare our results for SAGA-AS with known complexity results for the primal-dual method Quartz of Qu et al. (2015). We do this because this was the first and (with the exception of the dfSDCA method of Csiba & Richtárik (2015)) remains the only SGD-type method for solving (1) which was analyzed in the arbitrary sampling paradigm. Prior to this work we have conjectured that SAGA-AS would attain the same complexity as Quartz. As we shall show, this is indeed the case.

The problem studied in (Qu et al., 2015) is

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \phi_i(\mathbf{A}_i^\top x) + \psi(x), \tag{14}$$

where $\mathbf{A}_i \in \mathbb{R}^{d\times m}$, $\phi_i : \mathbb{R}^m \to \mathbb{R}$ is $1/\gamma$-smooth and convex, $\psi : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex function. When ψ is also smooth, problem (14) can be written in the form of problem (1) with $\lambda_i = 1/n$, and

$$f_i(x) = \phi_i(\mathbf{A}_i^\top x) + \psi(x), \tag{15}$$

Quartz guarantees the duality gap to be less than ϵ in expectation using at most

$$\mathcal{O}\left\{\max_i\left(\frac{1}{p_i} + \frac{v_i}{p_i\mu\gamma n}\right)\log\left(\frac{1}{\epsilon}\right)\right\} \tag{16}$$

iterations, where the parameters $v_1, ..., v_n$ are assumed to satisfy the following expected separable overapproximation (ESO) inequality, which needs to hold for all $h_i \in \mathbb{R}^m$:

$$\mathbb{E}_S\left[\left\|\sum_{i\in S}\mathbf{A}_i h_i\right\|^2\right] \leq \sum_{i=1}^n p_i v_i \|h_i\|^2. \tag{17}$$

If in addition ψ is L_ψ-smooth, then f_i in (15) is smooth with $L_i \leq \frac{\lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)}{\gamma} + L_\psi$. We now consider several particular samplings:

Serial samplings. By Lemma 5 in (Qu et al., 2015), $v_i = \lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)$. Hence, the bound (16) becomes

$$\mathcal{O}\left\{\max_i\left(\frac{1}{p_i} + \frac{\lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)}{p_i\mu\gamma n}\right)\log\left(\frac{1}{\epsilon}\right)\right\}.$$

By choosing $\theta_S^i = 1/p_i$ (this is both the default choice mentioned in Lemma 2.3 and the optimal choice in view of Lemma 3.4), $\mathcal{A}_i = \beta_i = 1/p_i$, $\mathcal{B} = 0$, and our iteration complexity bound in Thm 3.3 becomes

$$\max\left\{\max_i\left\{\frac{1}{p_i} + \frac{4L_\psi}{n\mu p_i} + \frac{4\lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)}{p_i\mu\gamma n}\right\}, \frac{2L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right).$$

We can see that as long as $L_\psi/\mu = \mathcal{O}(n)$, the two bounds are essentially the same.

Parallel (τ-nice) sampling. By Lemma 6 in (Qu et al., 2015), $v_i = \lambda_{\max}\left(\sum_{j=1}^d\left(1 + \frac{(\omega_j-1)(\tau-1)}{n-1}\right)\mathbf{A}_{ji}^\top\mathbf{A}_{ji}\right)$, where $\mathbf{A}_{ji}$ is the j-th row of $\mathbf{A}_i$, and for each $1 \leq j \leq d$, ω_j is the number of nonzero blocks in the j-th row of $\mathbf{A}$, i.e., $\omega_j := |\{i \in [n] : \mathbf{A}_{ji} \neq 0\}|$. In the dense case (i.e., $\omega_j = n$), $v_i = \tau\lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)$. Hence, (16) becomes

$$\mathcal{O}\left\{\max_i\left(\frac{n}{\tau} + \frac{\lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)}{\mu\gamma}\right)\log\left(\frac{1}{\epsilon}\right)\right\}.$$

By choosing $\theta_S^i = 1/p_i = n/\tau$, from Lemma 3.2 and Thm 3.3, the iteration complexity becomes

$$\max\left\{\frac{n}{\tau} + \frac{n-\tau}{n-1}\frac{4(1+\mathcal{B})\max_i nL_i\lambda_i}{\mu\tau}, \frac{2(1+\mathcal{B})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right),$$

where $\mathcal{B} = \frac{n(\tau-1)}{\tau(n-1)} \approx 1$ and $L_i \leq \frac{\lambda_{\max}(\mathbf{A}_i^\top\mathbf{A}_i)}{\gamma} + L_\psi$. We can see the bounds would be better than Quartz. However, if $\omega_j \ll n$, then Quartz may enjoy a tighter bound.

The parameters $v_1, \cdots, v_n$ used in (Qu et al., 2015) allow one to exploit the sparsity of the data matrix $\mathbf{A} = (\mathbf{A}_1, \cdots, \mathbf{A}_n)$ and achieve almost linear speedup when $\mathbf{A}$ is sparse or has favourable spectral properties. In the next section, we study further SAGA-AS in the case when the objective function is of the form (14), and obtain results which, like Quartz, are able to improve with data sparsity.

4. Analysis in the Composite Case

We now consider the general problem (1) with $\psi \neq 0$. In order to be able to take advantage of data sparsity, we assume that functions f_i take the form

$$f_i(x) \equiv \phi_i(\mathbf{A}_i^\top x). \tag{18}$$

Then clearly $\nabla f_i(x) = \mathbf{A}_i \nabla \phi_i(\mathbf{A}_i^\top x)$. Thus if SAGA-AS starts with $\mathbf{J}^0 = \begin{pmatrix} \mathbf{A}_1\alpha_1^0 & \mathbf{A}_2\alpha_2^0 & \cdots & \mathbf{A}_n\alpha_n^0 \end{pmatrix}$, for some $\alpha_i^0 \in \mathbb{R}^m$, $i \in [n]$, then we always have $\mathbf{J}^k = \begin{pmatrix} \mathbf{A}_1\alpha_1^k & \mathbf{A}_2\alpha_2^k & \cdots & \mathbf{A}_n\alpha_n^k \end{pmatrix}$, for some $\alpha_i^k \in \mathbb{R}^m$, $i \in [n]$. We assume that the set of minimizers $\mathcal{X}^* := \arg\min\{P(x) : x \in \mathbb{R}^d\}$, is nonempty, and let $P^* = P(x^*)$ for $x^* \in \mathcal{X}^*$. Further, denote $[x]^* = \arg\min\{\|x - y\| : y \in \mathcal{X}^*\}$; the closest optimal solution from x. Further, for any $M > 0$ define $\mathcal{X}(M)$ to be the set of points with objective value bounded by $P^* + M$, i.e., $\mathcal{X}(M) := \{x \in \mathrm{dom}(\psi) : P(x) \leq P^* + M)\}$. We make several further assumptions:

Assumption 4.1 (Smoothness)**.** Each $\phi_i : \mathbb{R}^m \to \mathbb{R}$ is $1/\gamma$-smooth and convex, i.e., $0 \leq \langle \nabla\phi_i(a) - \nabla\phi_i(b), a - b\rangle \leq \|a-b\|^2/\gamma, \forall a, b \in \mathbb{R}^m$.

Assumption 4.2 (Quadratic functional growth condition; see (Necoara et al., 2018))**.** For any $M > 0$, there is $\mu > 0$ such that for any $x \in \mathcal{X}(M)$

$$P(x) - P^* \geq \tfrac{\mu}{2}\|x - [x]^*\|^2. \tag{19}$$

Assumption 4.3 (Nullspace consistency)**.** For any $x^*, y^* \in \mathcal{X}^*$ we have $\mathbf{A}_i^\top x^* = \mathbf{A}_i^\top y^*, \ \forall i \in [n]$.

We shall need a slightly stronger condition than Assumption 4.2: there is a constant $\mu > 0$ such that

$$P(x^k) - P^* \geq \tfrac{\mu}{2}\|x^k - [x^k]^*\|^2, w.p.1, \ \ \forall k \geq 1, \tag{20}$$

for the sequence $\{x^k\}$ produced by Algorithm 1. For (20) to be true, we only need (19) to hold for any $x \in \mathrm{dom}(\psi)$. This can be done by adding sufficiently large box constraint to problem (1) to make $\mathrm{dom}(\psi)$ compact without changing the optimal solution set.

4.1. Linear convergence under quadratic functional growth condition

Our first complexity result states a linear convergence rate of SAGA-AS under the quadratic functional growth condition. Note that the Lyapunov function is *not stochastic* (i.e., $\mathbb{E}[\Psi^k] \mid x^k, \alpha_i^k]$ is not random).

Theorem 4.4. Assume f is L-smooth. Let $v \in \mathbb{R}^n_+$ be a positive vector satisfying (17) for a proper sampling S. Let $\{x^k, \mathbf{J}^k\}$ be the iterates produced by Algorithm 1 with $\theta_S^i = 1/p_i$ for all i and S. Let any $x^* \in \mathcal{X}^*$. Consider the Lyapunov function $\Psi^k := \|x^k - [x^k]^*\|^2 + \alpha\sum_{i=1}^n \sigma_i p_i^{-1} v_i \lambda_i^2 \|\alpha_i^k - \nabla\phi_i(\mathbf{A}_i^\top x^*)\|^2$, where $\sigma_i = \frac{\gamma}{2v_i\lambda_i}$ for all i. Then there is a constant $\mu > 0$ such that the following is true. If stepsize α satisfies

$$\alpha \leq \min\left\{\tfrac{2}{3}\min_{1\leq i\leq n} \tfrac{p_i}{\mu + 4v_i\lambda_i/\gamma}, \tfrac{1}{3L}\right\}, \tag{21}$$

then $\mathbb{E}[\Psi^k] \leq \left(\frac{1+\alpha\mu/2}{1+\alpha\mu}\right)^k \mathbb{E}[\Psi^0]$. This implies that if we choose α equal to the upper bound in (21), then $\mathbb{E}[\Psi^k] \leq \epsilon \cdot \mathbb{E}[\Psi^0]$ when

$$k \geq \left(2 + \max\left\{\tfrac{6L}{\mu}, 3\max_i\left(\tfrac{1}{p_i} + \tfrac{4v_i\lambda_i}{p_i\mu\gamma}\right)\right\}\right)\log\left(\tfrac{1}{\epsilon}\right).$$

If μ is unknow and we choose

$$\alpha \leq \min\left\{\min_{1\leq i\leq n} \tfrac{p_i}{12v_i\lambda_i/\gamma}, \tfrac{1}{3L}\right\}, \tag{22}$$

then $\mathbb{E}[\Psi^k] \leq (1 - \min\{\frac{\alpha\mu}{2(1+\alpha\mu)}, \frac{p_i}{2}\})^k \mathbb{E}[\Psi^0]$. This implies that if we choose α equal to the upper bound in (22), then $\mathbb{E}[\Psi^k] \leq \epsilon \cdot \mathbb{E}[\Psi^0]$ when

$$k \geq \left(2 + \max\left\{\tfrac{6L}{\mu}, \max_i\left\{\tfrac{24v_i\lambda_i}{\mu p_i\gamma}, \tfrac{2}{p_i}\right\}\right\}\right)\log\left(\tfrac{1}{\epsilon}\right).$$

Non-strongly convex problems in the form of (1) and (18) which satisfy Assumptions 4.1, 4.2 and 4.3 include the case when each ϕ_i is strongly convex and ψ is polyhedral (Necoara et al., 2018). In particular, Thm 4.4 applies to the following logistic regression problem that we use in the experiments ($\lambda_1 \geq 0$ and $\lambda_2 \geq 0$)

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^n \log(1 + e^{b_i A_i^\top x}) + \lambda_1\|x\|_1 + \tfrac{\lambda_2}{2}\|x\|^2. \tag{23}$$

4.2. Linear convergence for strongly convex regularizer

For the problem studied in (Qu et al., 2015) where the regularizer ψ is μ-strongly convex (and hence Assumption (4.2) holds and the minimizer is unique), we obtain the following refinement of Thm 4.4.

Theorem 4.5. Let ψ be μ-strongly convex. Let $v \in \mathbb{R}^n_+$ be a positive vector satisfying (17) for a proper sampling S. Let $\{x^k, \mathbf{J}^k\}$ be the iterates produced by Algorithm 1 with $\theta_S^i = 1/p_i$ for all i and S. Consider the same Lyapunov function as in Thm 4.4, but set $\sigma_i = \frac{2\gamma}{3v_i\lambda_i}$ for all i. If stepsize α satisfies

$$\alpha \leq \min_{1\leq i\leq n} \tfrac{p_i}{\mu + 3v_i\lambda_i/\gamma}, \tag{24}$$

then $\mathbb{E}[\Psi^k] \leq (1 + \alpha\mu)^{-k}\mathbb{E}[\Psi^0]$. So, if we choose α equal to the upper bound in (24), then $\mathbb{E}[\Psi^k] \leq$

$\epsilon \cdot \mathbb{E}[\Psi^0]$ when $k \geq \max_i \left\{1 + \frac{1}{p_i} + \frac{3v_i\lambda_i}{p_i\mu\gamma}\right\} \log\left(\frac{1}{\epsilon}\right)$. If μ is unknown and we choose $\sigma_i = \frac{\gamma}{(1+\alpha\mu)v_i\lambda_i}$ for all i and $\alpha \leq \min_{1\leq i\leq n} \frac{p_i\gamma}{4v_i\lambda_i}$, then $\mathbb{E}[\Psi^k] \leq \left(1 - \min\{\frac{\alpha\mu}{1+\alpha\mu}, \frac{p_i}{2}\}\right)^k \mathbb{E}[\Psi^0]$. So, if we choose α equal to the upper bound, then $\mathbb{E}[\Psi^k] \leq \epsilon \cdot \mathbb{E}[\Psi^0]$ when $k \geq \max_i \left\{1 + \frac{4v_i\lambda_i}{p_i\mu\gamma}, \frac{2}{p_i}\right\} \log\left(\frac{1}{\epsilon}\right)$.

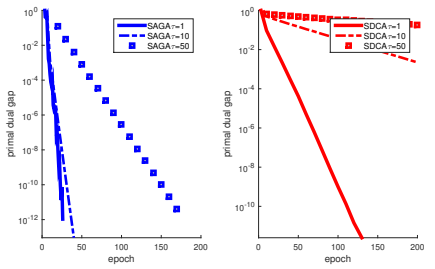
Note that up to some small constants, the rate provided by Thm 4.5 is the same as that of Quartz. Hence, the analysis for special samplings provided in Section 3.4 applies, and we conclude that *SAGA-AS is also able to accelerate on sparse data.*

5. Experiments

We tested SAGA-AS to solve the logistic regression problem (23) on 3 different datasets: w8a, a9a and ijcnn1[4]. The experiments presented in Section 5.1 and 5.2 are tested for $\lambda_1 = 0$ and $\lambda_2 = 1e-5$, which is of the same order as the number of samples in the three datasets. In Section 5.3 we test on the unregularized problem with $\lambda_1 = \lambda_2 = 0$. In all the plots, the x-axis records the number of pass of the dataset. More experiments can be found in the Suppl.

5.1. Batch sampling

Here we compare SAGA-AS with SDCA for τ-nice sampling S with $\tau \in \{1, 10, 50\}$. Note that SDCA with τ-nice sampling works the same both in theory and in practice as Quartz with τ-nice sampling. We report in Figure 1 the results obtained for the dataset ijcnn1. When we increase τ by 50, the number of epochs of SAGA-AS only increased by less than 6. This indicates a considerable speedup if parallel computation is used in the implementation.



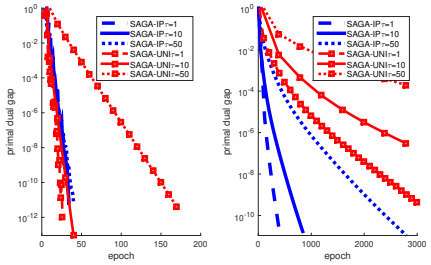
(a) SAGA; ijcnn1 (b) SDCA; ijcnn1

Figure 1. mini-batch SAGA V.S. mini-batch SDCA

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

5.2. Importance sampling

We compare uniform sampling SAGA (SAGA-UNI) with importance sampling SAGA (SAGA-IP), as described in Section 3.3 , on three values of $\tau \in \{1, 10, 50\}$. The results for the datasets w8a and ijcnn1 are shown in Figure 2. For the dataset ijcnn1, mini-batch with importance sampling almost achieves linear speedup as the number of epochs does not increase with τ. For the dataset w8a, mini-batch with importance sampling can even need less number of epochs than serial uniform sampling. Note that we adopt the importance sampling strategy described in (Hanzely & Richtárik, 2019) and the actual running time is the same as uniform sampling.

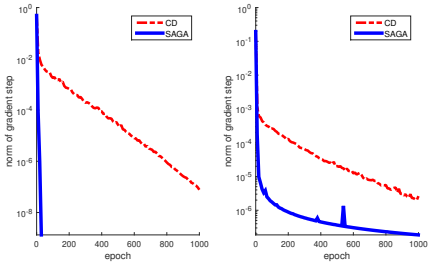


(a) ijcnn1 (b) w8a

Figure 2. importance sampling V.S. uniform sampling

5.3. Comparison with coordinate descent

We consider the un-regularized logistic regression problem (23) with $\lambda_1 = \lambda_2 = 0$. In this case, Thm 4.4 applies and we expect to have linear convergence of SAGA without any knowledge on the constant μ satsifying Assumption (4.2). This makes SAGA comparable with descent methods such as gradient method and coordinate descent (CD) method. However, comparing with their deterministic counterparts, the speedup provided by CD can be at most of order d while the speedup by SAGA can be of order n. Thus SAGA is much preferable than CD when n is larger than d. We provide numerical evidence in Figure 3.



(a) ijcnn1 (b) a9a

Figure 3. SAGA V.S. CD

References

Bibi, A., Sailanbayev, A., Ghanem, B., Gower, R. M., and Richtárik, P. Improving SAGA via a probabilistic interpolation with gradient descent. *arXiv: 1806.05633*, 2018.

Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schönlieb, C. B. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2017.

Csiba, D. and Richtárik, P. Primal method for ERM with flexible mini-batching schemes and non-convex losses. *arXiv:1506.02227*, 2015.

Csiba, D. and Richtárik, P. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27): 1–21, 2018.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 1646–1654. Curran Associates, Inc., 2014.

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arXiv Preprint arXiv: 1805.02632*, 2018.

Hanzely, F. and Richtárik, P. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. *arXiv Preprint arXiv: 1809.09354*, 2018.

Hanzely, F. and Richtárik, P. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Horváth, S. and Richtárik, P. Nonconvex variance reduced optimization with arbitrary sampling. *arXiv:1809.04146*, 2018.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Konečný, J. and Richtárik, P. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, pp. 1–14, 2017. URL http://arxiv.org/abs/1312.1666.

Konečný, J., Lu, J., Richtárik, P., and Takáč, M. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10 (2):242–255, 2016.

Loizou, N. and Richtárik, P. Linearly convergent stochastic heavy ball method for minimizing generalization error. In *NIPS Workshop on Optimization for Machine Learning*, 2017a.

Loizou, N. and Richtárik, P. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv:1712.09677*, 2017b.

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015. URL http://jmlr.org/papers/volume16/mokhtari15a/mokhtari15a.pdf.

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pp. 1–39, 2018. doi: https://doi.org/10.1007/s10107-018-1232-1.

Needell, D., Srebro, N., and Ward, R. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 2015.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Qu, Z. and Richtárik, P. Coordinate descent with arbitrary sampling I: Algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016a.

Qu, Z. and Richtárik, P. Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884, 2016b.

Qu, Z., Richtárik, P., and Zhang, T. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 865–873. Curran Associates, Inc., 2015.

Richtárik, P. and Takáč, M. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, 2016.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(2): 1–38, 2014.

Richtárik, P. and Takáč, M. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

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

Schmidt, M., Babanezhad, R., Ahmed, M. O., Defazio, A., Clifton, A., and Sarkar, A. Non-uniform stochastic average gradient method for training conditional random fields. In *18th International Conference on Artificial Intelligence and Statistics*, 2015.

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

Shalev-Shwartz, S. SDCA without duality, regularization, and individual convexity. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 747–754, 2016.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014.

Shalev-Shwartz, S. and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. Mini-batch primal and dual methods for SVMs. In *30th International Conference on Machine Learning*, pp. 537–552, 2013.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014a. doi: 10.1137/140961791.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014b.

Zhao, P. and Zhang, T. Stochastic optimization with importance sampling. *The 32nd International Conference on Machine Learning*, 37:1–9, 2015.

SUPPLEMENTARY MATERIAL
SAGA with Arbitrary Sampling

A. Importance Sampling for Minibatches: Partition Sampling

In this section, we consider SAGA-AS with partition sampling for the problem

$$x^* = \arg\min_{x\in\mathbb{R}^d} \left[f(x) := \sum_{i=1}^n \frac{1}{n} f_i(x) \right]. \tag{25}$$

First we give the definition of partition of $[n]$ and partition sampling.

Definition A.1. A partition $\mathcal{G}$ of $[n]$ is a set consisted of the subsets of $[n]$ such that $\cup_{C\in\mathcal{G}} C = [n]$ and $C_i \cap C_j = \emptyset$ for any $C_i, C_j \in \mathcal{G}$ with $i \neq j$. A partition sampling S is a sampling such that $p_C = \mathbb{P}[S = C] > 0$ for all $C \in \mathcal{G}$ and $\sum_{C\in\mathcal{G}} p_C = 1$.

From (7) of Lemma 2.3, for partition sampling, we have $\theta_C^i = \frac{1}{p_C}$ if $i \in C$ for all $C \in \mathcal{G}$. Hence, SAGA-AS for problem (25) with partition sampling becomes Algorithm 2 (SAGA-PS):

Algorithm 2 SAGA with Partition Sampling for problem (25)–SAGA-PS

Parameters: Partition sampling S and stepsize $\alpha > 0$
Initialization: Choose $x^0 \in \mathbb{R}^d$, $\mathbf{J}^0 \in \mathbb{R}^{d\times n}$
for $k = 0, 1, 2, ...$ **do**
  Sample a fresh set $S_k \sim S$
  $\mathbf{J}^{k+1} = \mathbf{J}^k + \left(\mathbf{G}(x^k) - \mathbf{J}^k\right) \mathbf{\Pi}_{\mathbf{I}_{S_k}}$
  $g^k = \mathbf{J}^k\lambda + \frac{1}{np_{S_k}} \left(\mathbf{G}(x^k) - \mathbf{J}^k\right) \mathbf{\Pi}_{\mathbf{I}_{S_k}} e$
  $x^{k+1} = x^k - \alpha g^k$
end for

Next we will give the iteration complexity of SAGA-PS by reformulation. For any partition sampling S, let $f_C(x) = \frac{1}{|C|}\sum_{i\in C} f_i(x)$ for $C \in \mathcal{G}$, and let f_C be L_C-smooth. In problem (25), $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{n}\sum_{C\in\mathcal{G}} |C| f_C(x) = \sum_{C\in\mathcal{G}} \frac{|C|}{n} f_C(x)$. Let $m = |\mathcal{G}|$ and without loss of generality, we denote $\mathcal{G} = \{C_1, C_2, ..., C_m\}$. Let $\lambda_i = \frac{|C_i|}{n}$ for $1 \leq i \leq m$. Then we can see the minibatch SAGA with partition sampling for problem (25) (Algorithm 2) can be regarded as Algorithm 1 for problem (1) with the sampling: $\mathbb{P}[\{i\}] = p_{C_i}$ and $\theta^j_{\{i\}} = \frac{1}{p_{C_i}}$ for all $i, j \in [m]$. Hence, by applying Thm 3.3, we can obtain the following theorem.

Theorem A.2. Let S be any partition sampling with partition $\mathcal{G} = \{C_1, C_2, ..., C_m\}$. Let f in problem (25) be μ-strongly convex and f_C be convex and L_C-smooth. Let $\{x^k, \mathbf{J}^k\}$ be the iterates produced by Algorithm 2.
Consider the stochastic Lyapunov function

$$\Psi_S^k := \|x^k - x^*\|^2 + 2\alpha\sigma_S \left\| \frac{1}{np_S} (\mathbf{J}^k - \mathbf{G}(x^k))\Pi_{\mathbf{I}_S} e \right\|^2,$$

where $\sigma_S = \frac{n}{4L_S|S|}$ is a stochastic Lyapunov constant. If stepsize α satisfies

$$\alpha \leq \min_{C\in\mathcal{G}} \frac{p_C}{\mu + \frac{4L_C|C|}{n}}, \tag{26}$$

then $\mathbb{E}[\Psi_S^k] \leq (1-\mu\alpha)^k \mathbb{E}[\Psi_S^0]$. This implies that if we choose α equal to the upper bound in (26), then

$$k \geq \max_{C\in\mathcal{G}} \left\{ \frac{1}{p_C} + \frac{4L_C|C|}{\mu n p_C} \right\} \log\left(\frac{1}{\epsilon}\right) \Rightarrow \mathbb{E}[\Psi_S^k] \leq \epsilon \cdot \mathbb{E}[\Psi_S^0].$$

Thm A.2 contains Thm 5.2 with τ-partition sampling in (Gower et al., 2018) as a special case, and with a little weaker condition: instead of demanding f_C be μ-strongly convex, we only need f be μ-strongly convex.

A.1. Partition sampling

From Thm A.2 we can propose importance partition sampling for Algorithm 2. For a partition sampling S with the partition $\mathcal{G} = \{C_1, C_2, ..., C_m\}$, where $m = |\mathcal{G}|$, the iteration complexity of Algorithm 2 is given by

$$\max_{C\in\mathcal{G}} \left\{ \frac{1}{p_C} + \frac{4L_C|C|}{\mu n p_C} \right\} \log\left(\frac{1}{\epsilon}\right).$$

We can minimize the complexity bound in p_C by choosing

$$p_C = \frac{\mu n + 4L_C|C|}{\sum_{C\in\mathcal{G}}(\mu n + 4L_C|C|)}. \tag{27}$$

With these optimal probabilities, the stepsize bound is $\alpha \leq \frac{n}{\sum_{C\in\mathcal{G}}(\mu n + 4L_C|C|)}$, and by choosing the maximum allowed stepsize the resulting complexity becomes

$$\left(|\mathcal{G}| + \frac{4\sum_{C\in\mathcal{G}} L_C|C|}{\mu n} \right) \log\left(\frac{1}{\epsilon}\right). \tag{28}$$

B. Proofs of Lemmas 2.3 , 3.2, and 3.4

B.1. Proof of Lemma 2.3

From (6), we know (5) in Assumption 2.1 is actually

$$\mathbb{E}[\theta_S \mathrm{Diag}(e_S)\lambda] = \lambda,$$

which is equivalent to

$$\mathbb{E}[\theta_S^i 1_{i\in S}\lambda_i] = \lambda_i,$$

for all $i \in [n]$. Then for all $i \in [n]$, we have

$$\begin{aligned}\mathbb{E}[\theta_S^i 1_{i\in S}\lambda_i] &= \sum_{C\subseteq[n]} p_C \theta_C^i 1_{i\in C}\lambda_i \\ &= \sum_{C\subseteq[n]:i\in C} p_C \theta_C^i \lambda_i \\ &= \lambda_i.\end{aligned}$$

Since $\lambda_i > 0$ for all $i \in [n]$, we have that (5) is equivalent to

$$\sum_{C\subseteq[n]:i\in C} p_C \theta_C^i = 1, \quad \forall i \in [n].$$

B.2. Proof of Lemma 3.2(i)

Applying Lemma C.1 with $z = |C|$, we can obtain

$$\begin{aligned}\mathbb{E}[\|\mathbf{M}\theta_S \mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] &= \sum_C p_C \|\mathbf{M}\theta_S \mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2 \\ &\le \sum_C p_C |C| \sum_{i\in C} \|\theta_C^i \lambda_i \mathbf{M}_{:i}\|^2 \\ &= \sum_{i=1}^n \lambda_i^2 \sum_{C:i\in C} p_C |C| (\theta_C^i)^2 \|\mathbf{M}_{:i}\|^2 \\ &= \sum_{i=1}^n \beta_i \lambda_i^2 \|\mathbf{M}_{:i}\|^2.\end{aligned}$$

B.3. Proof of Lemma 3.2(ii)

Denote $\mathbf{P}_{ij} = \mathbb{P}[i \in S \ \& j \in S]$, then we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{M}\theta_S \mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] &= \mathbb{E}[\|\sum_{i\in S} \frac{\lambda_i}{p_i}\mathbf{M}_{:i}\|^2] \\ &= \mathbb{E}\left[\sum_{i,j\in S} \left\langle \frac{\lambda_i}{p_i}\mathbf{M}_{:i}, \frac{\lambda_j}{p_j}\mathbf{M}_{:j}\right\rangle\right] \\ &= \sum_C p_C \sum_{i,j\in C} \left\langle \frac{\lambda_i}{p_i}\mathbf{M}_{:i}, \frac{\lambda_j}{p_j}\mathbf{M}_{:j}\right\rangle \\ &= \sum_{i,j=1}^n \sum_{C:i,j\in C} p_C \left\langle \frac{\lambda_i}{p_i}\mathbf{M}_{:i}, \frac{\lambda_j}{p_j}\mathbf{M}_{:j}\right\rangle \\ &= \sum_{i,j=1}^n \frac{\mathbf{P}_{ij}}{p_i p_j} \langle \lambda_i \mathbf{M}_{:i}, \lambda_j \mathbf{M}_{:j}\rangle\end{aligned}$$

For τ-nice sampling, we have $\mathbf{P}_{ij} = \frac{\tau(\tau-1)}{n(n-1)}$ for $i \neq j$, and $p_i = \frac{\tau}{n}$. Hence

$$\begin{aligned}
\mathbb{E}[\|\mathbf{M}\theta_S\mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] &= \sum_{i\neq j}\frac{n(\tau-1)}{\tau(n-1)}\langle\lambda_i\mathbf{M}_{:i},\lambda_j\mathbf{M}_{:j}\rangle + \sum_{i=1}^n\frac{n}{\tau}\lambda_i^2\|\mathbf{M}_{:i}\|^2 \\
&= \sum_{i,j=1}^n\frac{n(\tau-1)}{\tau(n-1)}\langle\lambda_i\mathbf{M}_{:i},\lambda_j\mathbf{M}_{:j}\rangle + \sum_{i=1}^n\frac{n(n-\tau)}{\tau(n-1)}\lambda_i^2\|\mathbf{M}_{:i}\|^2 \\
&= \sum_{i=1}^n\frac{n(n-\tau)}{\tau(n-1)}\lambda_i^2\|\mathbf{M}_{:i}\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\|\mathbf{M}\lambda\|^2.
\end{aligned}$$

B.4. Proof of Lemma 3.2(iii)

For independent sampling S, we have $\mathbf{P}_{ij} = p_ip_j$ if $i \neq j$. Then

$$\begin{aligned}
\mathbb{E}[\|\mathbf{M}\theta_S\mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] &= \sum_{i,j=1}^n\frac{\mathbf{P}_{ij}}{p_ip_j}\langle\lambda_i\mathbf{M}_{:i},\lambda_j\mathbf{M}_{:j}\rangle \\
&= \sum_{i\neq j}\frac{\mathbf{P}_{ij}}{p_ip_j}\langle\lambda_i\mathbf{M}_{:i},\lambda_j\mathbf{M}_{:j}\rangle + \sum_{i=1}^n\frac{1}{p_i}\lambda_i^2\|\mathbf{M}_{:i}\|^2 \\
&= \sum_{i,j=1}^n\langle\lambda_i\mathbf{M}_{:i},\lambda_j\mathbf{M}_{:j}\rangle + \sum_{i=1}^n(\frac{1}{p_i}-1)\lambda_i^2\|\mathbf{M}_{:i}\|^2 \\
&= \sum_{i=1}^n(\frac{1}{p_i}-1)\lambda_i^2\|\mathbf{M}_{:i}\|^2 + \|\mathbf{M}\lambda\|^2.
\end{aligned}$$

B.5. Sparse matrix

For each $1 \leq j \leq d$, let ω_j be the number of nonzero elements in the j-th row of $\mathbf{M}$, i.e., $\omega_j := |\{i \in [n] : \mathbf{M}_{ji} \neq 0\}|$. Then

$$\|\mathbf{M}_{j:}\lambda\|^2 = \left(\sum_{i=1}^n\mathbf{M}_{ji}\lambda_i\right)^2 \leq \omega_j\sum_{i=1}^n\left(\mathbf{M}_{ji}\lambda_i\right)^2.$$

and thus

$$\|\mathbf{M}\lambda\|^2 = \sum_{j=1}^d\|\mathbf{M}_{j:}\lambda\|^2 \leq \sum_{j=1}^d\left(\omega_j\sum_{i=1}^n\left(\mathbf{M}_{ji}\lambda_i\right)^2\right) = \sum_{i=1}^n\left(\sum_{j=1}^d\omega_j\left(\mathbf{M}_{ji}\right)^2\right)\lambda_i^2.$$

For τ-nice sampling, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{M}\theta_S\mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] &\leq \sum_{i=1}^n\frac{n(n-\tau)}{\tau(n-1)}\lambda_i^2\|\mathbf{M}_{:i}\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\sum_{i=1}^n\left(\sum_{j=1}^d\omega_j\left(\mathbf{M}_{ji}\right)^2\right)\lambda_i^2 \\
&= \sum_{i=1}^n\left(\frac{n(n-\tau)}{\tau(n-1)} + \frac{n(\tau-1)}{\tau(n-1)}\frac{\sum_{j=1}^d\omega_j\left(\mathbf{M}_{ji}\right)^2}{\|\mathbf{M}_{:i}\|^2}\right)\lambda_i^2\|\mathbf{M}_{:i}\|^2 \\
&= \sum_{i=1}^n\frac{n}{\tau}\left(1 + \frac{(\tau-1)}{(n-1)}\frac{\sum_{j=1}^d(\omega_j-1)\left(\mathbf{M}_{ji}\right)^2}{\|\mathbf{M}_{:i}\|^2}\right)\lambda_i^2\|\mathbf{M}_{:i}\|^2.
\end{aligned}$$

For independent sampling, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{M}\theta_S\mathbf{\Pi}_{\mathbf{I}_S}\lambda\|^2] &\leq \sum_{i=1}^{n}\left(\frac{1}{p_i}-1\right)\lambda_i^2\|\mathbf{M}_{:i}\|^2+\sum_{i=1}^{n}\left(\sum_{j=1}^{d}\omega_j\left(\mathbf{M}_{ji}\right)^2\right)\lambda_i^2\\ &=\sum_{i=1}^{n}\left(\frac{1}{p_i}-1+\frac{\sum_{j=1}^{d}\omega_j\left(\mathbf{M}_{ji}\right)^2}{\|\mathbf{M}_{:i}\|^2}\right)\lambda_i^2\|\mathbf{M}_{:i}\|^2.\end{aligned}$$

B.6. Proof of Lemma 3.4(i)

From Lemma 2.3, the problem $\min_{\theta\in\Theta(S)}\beta_i$ is equivalent to the following linearly constrained convex problem:

$$\begin{array}{ll}\min & \beta_i=\sum_{C:i\in C}p_C|C|(\theta_C^i)^2\\ \text{s.t.} & \sum_{C:i\in C}p_C\theta_C^i=1.\end{array}\tag{29}$$

The KKT system of problem (29) is

$$\begin{cases}2p_C|C|\theta_C^i+\tilde{\lambda}p_C=0,\forall C:i\in C\\ \sum_{C:i\in C}p_C\theta_C^i=1,\end{cases}\tag{30}$$

where $\tilde{\lambda}\in\mathbb{R}$ is the Lagrangian dual variable. By solving system (30), we obtain the optimal solution

$$\theta_C^i=\frac{1}{|C|\sum_{C:i\in C}\frac{p_C}{|C|}}=\frac{1}{p_i|C|\mathbb{E}^i[\frac{1}{|S|}]},$$

for all $C:i\in C$, and the minimum of β_i is

$$\begin{aligned}\beta_i &= \sum_{C:i\in C}p_C|C|(\theta_C^i)^2\\ &= \sum_{C:i\in C}p_C|C|\frac{1}{(p_i|C|\mathbb{E}^i[\frac{1}{|S|}])^2}\\ &= \frac{1}{p_i}\cdot\frac{1}{(\mathbb{E}^i[\frac{1}{|S|}])^2}\sum_{C:i\in C}\frac{p_C}{p_i}\frac{1}{|C|}\\ &= \frac{1}{p_i}\cdot\frac{1}{\mathbb{E}^i[\frac{1}{|S|}]}.\end{aligned}$$

B.7. Proof of Lemma 3.4(ii)

From (7) of Lemma 2.3, we can choose $\theta_C^i=\frac{1}{p_i}$ for all i and C. Hence

$$\min_{\theta\in\Theta(S)}\beta_i\leq\sum_{C:i\in C}p_C|C|\frac{1}{p_i^2}=\frac{1}{p_i}\mathbb{E}^i[|S|],$$

which implies

$$\frac{1}{\mathbb{E}^i[\frac{1}{|S|}]}\leq\mathbb{E}^i[|S|].$$

C. Smooth Case: Proof of Theorem 3.3

C.1. Lemmas

The following inequality is a direct consequence of convexity of $x \mapsto \|x\|^2$.

Lemma C.1. Let $a^i \in \mathbb{R}^d$ for $1 \leq i \leq z$ with $z \geq 1$. Then

$$\left\|\sum_{i=1}^{z} a^i\right\|^2 \leq z\sum_{i=1}^{z}\|a^i\|^2.$$

Lemma C.2. Let σ_i be any non-negative constant for $1 \leq i \leq n$. Then

$$\mathbb{E}\left[\sum_{i=1}^{n}\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k+1}-\nabla f_i(x^*)\|^2\right] \leq \mathbb{E}\left[\sum_{i=1}^{n}(1-p_i)\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k}-\nabla f_i(x^*)\|^2\right] + \mathbb{E}\left[\sum_{i=1}^{n}\sigma_i p_i\mathcal{A}_i\lambda_i^2\left\|\nabla f_i(x^k)-\nabla f_i(x^*)\right\|^2\right].$$

Proof. First notice that

$$\mathbf{J}_{:i}^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{if } i \in S^k \\ \mathbf{J}_{:i}^{k}. & \text{if } i \notin S^k \end{cases}$$

Then by taking conditional expectation on $\mathbf{J}^k$ and x^k, we have

$$\begin{aligned}
& \mathbb{E}\left[\sum_{i=1}^{n}\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k+1}-\nabla f_i(x^*)\|^2 \mid \mathbf{J}^k, x^k\right] \\
= & \sum_{i=1}^{n}\mathbb{E}\left[\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k+1}-\nabla f_i(x^*)\|^2 \mid \mathbf{J}^k, x^k\right] \\
= & \sum_{i=1}^{n}\sum_{C} p_C\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k+1}|_{S^k=C}-\nabla f_i(x^*)\|^2 \\
= & \sum_{i=1}^{n}\left(\sum_{C:i\in C} p_C\sigma_i\mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2 + \sum_{C:i\notin C} p_C\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k}-\nabla f_i(x^*)\|^2\right) \\
= & \sum_{i=1}^{n}(1-p_i)\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k}-\nabla f_i(x^*)\|^2 + \sum_{i=1}^{n}\sigma_i p_i\mathcal{A}_i\lambda_i^2\left\|\nabla f_i(x^k)-\nabla f_i(x^*)\right\|^2.
\end{aligned}$$

Taking expectations again and applying the tower property, we obtain the result.

□

In the next lemma we bound the second moment of the gradient estimate g^k.

Lemma C.3. The second moment of the gradient estimate is bounded by

$$\mathbb{E}\left[\|g^k\|^2 \mid \mathbf{J}^k, x^k\right] \leq 2\sum_{i=1}^{n}\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k}-\nabla f_i(x^*)\|^2 + 2\sum_{i=1}^{n}\mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2 + 2\mathcal{B}\|\nabla f(x^k)-\nabla f(x^*)\|^2.$$

Proof. Recall that

$$\begin{aligned}
g^k &= \mathbf{J}^k\lambda + (\mathbf{G}(x^k)-\mathbf{J}^k)\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda \\
&= \mathbf{J}^k\lambda - (\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda + (\mathbf{G}(x^k)-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda.
\end{aligned}$$

By applying Lemma C.1 with $z = 2$, we get

$$\left\|g^k\right\|^2 \leq 2\left\|(\mathbf{G}(x^k)-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 + 2\left\|(\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda - \mathbf{J}^k\lambda\right\|^2,$$

which implies that

$$\begin{aligned}\mathbb{E}\left[\left\|g^k\right\|^2 \mid \mathbf{J}^k, x^k\right] &\leq 2\mathbb{E}\left[\left\|(\mathbf{G}(x^k)-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right] \\ &\quad +2\mathbb{E}\left[\left\|(\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda-\mathbf{J}^k\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right].\end{aligned}$$

For $\mathbb{E}\left[\left\|(\mathbf{G}(x^k)-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right]$, by Assumption 3.1 we get

$$\begin{aligned}\mathbb{E}\left[\left\|(\mathbf{G}(x^k)-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right] &\leq \sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2+\mathcal{B}\|(\mathbf{G}(x^k)-\mathbf{G}(x^*))\lambda\|^2 \\ &= \sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2+\mathcal{B}\|\nabla f(x^k)-\nabla f(x^*)\|^2.\end{aligned}$$

For $\mathbb{E}\left[\left\|(\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda-\mathbf{J}^k\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right]$, since

$$\mathbb{E}\left[(\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda \mid \mathbf{J}^k, x^k\right]=\mathbf{J}^k\lambda,$$

and $\mathbb{E}[\|X-\mathbb{E}[X]\|^2]=\mathbb{E}[\|X\|^2]-\|\mathbb{E}[X]\|^2$, we have that

$$\begin{aligned}\mathbb{E}\left[\left\|(\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda-\mathbf{J}^k\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right] &= \mathbb{E}\left[\left\|(\mathbf{J}^k-\mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right]-\|\mathbf{J}^k\lambda\|^2 \\ &\overset{(8)}{\leq} \sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\mathbf{J}^k_{:i}-\nabla f_i(x^*)\|^2+(\mathcal{B}-1)\|\mathbf{J}^k\lambda\|^2 \\ &\leq \sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\mathbf{J}^k_{:i}-\nabla f_i(x^*)\|^2.\end{aligned}$$

Finally, we arrive at the result

$$\mathbb{E}\left[\|g^k\|^2 \mid \mathbf{J}^k, x^k\right] \leq 2\sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\mathbf{J}^k_{:i}-\nabla f_i(x^*)\|^2+2\sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2+2\mathcal{B}\|\nabla f(x^k)-\nabla f(x^*)\|^2.$$

□

C.2. Proof of Theorem 3.3

Having established the above lemmas, we are ready to proceed to the proof of pour main theorem covering the smooth case (Thm 3.3).

Let $\mathbb{E}_k[\cdot]$ denote the expectation conditional on $\mathbf{J}^k$ and x^k. First, from Assumption 2.1, it is evident that

$$\mathbb{E}_k[g^k]=\mathbf{J}^k\lambda+\nabla f(x^k)-\mathbf{J}^k\lambda=\nabla f(x^k). \tag{31}$$

Then we can obtain

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1}-x^*\|^2\right] &= \mathbb{E}_k\left[\|x^k-x^*-\alpha g^k\|^2\right] \\ &\overset{(31)}{=} \|x^k-x^*\|^2-2\alpha\langle\nabla f(x^k),x^k-x^*\rangle+\alpha^2\mathbb{E}_k[\|g^k\|^2] \\ &\leq (1-\mu\alpha)\|x^k-x^*\|^2-2\alpha(f(x^k)-f(x^*))+\alpha^2\mathbb{E}_k[\|g^k\|^2] \\ &\overset{\text{Lemma C.3}}{\leq} (1-\mu\alpha)\|x^k-x^*\|^2+2\alpha^2\sum_{i=1}^n\left[\mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2\right]-2\alpha(f(x^k)-f(x^*)) \\ &\quad +2\alpha^2\sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\mathbf{J}^k_{:i}-\nabla f_i(x^*)\|^2+2\alpha^2\mathcal{B}\|\nabla f(x^k)-\nabla f(x^*)\|^2.\end{aligned}$$

Taking expectation again and applying the tower property, we obtain

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1}-x^*\|^2\right] &\leq (1-\mu\alpha)\mathbb{E}\left[\|x^k-x^*\|^2\right]+2\alpha^2\mathbb{E}\left[\sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2\right]-2\alpha\mathbb{E}\left[(f(x^k)-f(x^*))\right]\\ &\quad +2\alpha^2\mathbb{E}\left[\sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^k-\nabla f_i(x^*)\|^2\right]+2\alpha^2\mathcal{B}\mathbb{E}\left[\|\nabla f(x^k)-\nabla f(x^*)\|^2\right].\end{aligned}$$

Therefore, for the stochastic Lyapunov function Ψ^{k+1}, we have

$$\begin{aligned}\mathbb{E}[\Psi^{k+1}] &= \mathbb{E}\left[\|x^{k+1}-x^*\|^2+2\alpha\sum_{i=1}^n\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^{k+1}-\nabla f_i(x^*)\|^2\right]\\ &\overset{\text{Lemma C.2}}{\leq} \mathbb{E}\left[(1-\mu\alpha)\|x^k-x^*\|^2\right]+2\alpha\mathbb{E}\left[\sum_{i=1}^n\left((\alpha\mathcal{A}_i\lambda_i^2+\sigma_i p_i\mathcal{A}_i\lambda_i^2)\left\|\nabla f_i(x^k)-\nabla f_i(x^*)\right\|^2\right)\right]\\ &\quad +2\alpha\mathbb{E}\left[\sum_{i=1}^n(1-p_i+\frac{\alpha}{\sigma_i})\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^k-\nabla f_i(x^*)\|^2\right]+2\alpha^2\mathcal{B}\mathbb{E}\left[\|\nabla f(x)-\nabla f(x^*)\|^2\right]-2\alpha\mathbb{E}\left[(f(x^k)-f(x^*))\right]\\ &\leq \mathbb{E}\left[(1-\mu\alpha)\|x^k-x^*\|^2\right]+2\alpha\mathbb{E}\left[\sum_{i=1}^n(1-p_i+\frac{\alpha}{\sigma_i})\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^k-\nabla f_i(x^*)\|^2\right]-2\alpha\mathbb{E}\left[(f(x^k)-f(x^*))\right]\\ &\quad +4\alpha\max_i\{L_i(\alpha\mathcal{A}_i\lambda_i+\sigma_i p_i\mathcal{A}_i\lambda_i)\}\mathbb{E}\left[(f(x^k)-f(x^*))\right]+4\alpha^2\mathcal{B}L\mathbb{E}\left[(f(x^k)-f(x^*))\right]\\ &= \mathbb{E}\left[(1-\mu\alpha)\|x^k-x^*\|^2\right]+2\alpha\mathbb{E}\left[\sum_{i=1}^n(1-p_i+\frac{\alpha}{\sigma_i})\sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^k-\nabla f_i(x^*)\|^2\right]\\ &\quad +\alpha\left(4\max_i\{L_i(\alpha\mathcal{A}_i\lambda_i+\sigma_i p_i\mathcal{A}_i\lambda_i)\}-\frac{2}{1+\mathcal{B}}\right)\mathbb{E}\left[(f(x^k)-f(x^*))\right]\\ &\quad +\alpha\left(4\alpha\mathcal{B}L-\frac{2\mathcal{B}}{1+\mathcal{B}}\right)\mathbb{E}\left[(f(x^k)-f(x^*))\right],\end{aligned}$$

where the last inequality we use f_i is L_i-smooth and f is L-smooth.

In order to guarantee that $\mathbb{E}\Psi^{k+1}\leq(1-\mu\alpha)\mathbb{E}\Psi^k$, α should be chosen such that

$$4\alpha\mathcal{B}L-\frac{2\mathcal{B}}{1+\mathcal{B}}\leq 0,\quad\Rightarrow\quad\alpha\leq\frac{1}{2\mathcal{B}(1+1/\mathcal{B})L},$$

$$L_i(\alpha\mathcal{A}_i\lambda_i+\sigma_i p_i\mathcal{A}_i\lambda_i)\leq\frac{1}{2(1+\mathcal{B})},\quad\Rightarrow\quad\alpha\leq\frac{1}{2(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}-\sigma_i p_i,$$

and

$$(1-p_i)\sigma_i+\alpha\leq(1-\mu\alpha)\sigma_i,\quad\Rightarrow\quad\alpha\leq\frac{\sigma_i p_i}{\mu\sigma_i+1},$$

for all $1\leq i\leq n$. Since $\sigma_i=\frac{1}{4(1+\mathcal{B})L_i\mathcal{A}_i p_i\lambda_i}$, if α satisfies

$$\alpha\leq\min\left\{\min_{1\leq i\leq n}\frac{p_i}{\mu+4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i p_i},\ \frac{1}{2\mathcal{B}(1+1/\mathcal{B})L}\right\}$$

then we have the recursion $\mathbb{E}\Psi^{k+1}\leq(1-\mu\alpha)\mathbb{E}\Psi^k$.

In the case where μ is unknown, we can choose

$$\alpha\leq\min\left\{\min_{1\leq i\leq n}\frac{p_i}{8(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i p_i},\ \frac{1}{2\mathcal{B}(1+1/\mathcal{B})L}\right\}.$$

Then

$$4\alpha\mathcal{B}L - \frac{2\mathcal{B}}{1+\mathcal{B}} \leq 0, \quad L_i(\alpha\mathcal{A}_i\lambda_i + \sigma_i p_i \mathcal{A}_i\lambda_i) \leq \frac{1}{2(1+\mathcal{B})},$$

and

$$1 - p_i + \frac{\alpha}{\sigma_i} \leq 1 - \frac{p_i}{2},$$

for all $1 \leq i \leq n$. Therefore, we have

$$\mathbb{E}\Psi^{k+1} \leq \left(1 - \min\left\{\mu\alpha, \frac{p_i}{2}\right\}\right)\mathbb{E}\Psi^k.$$

C.3. Proof of Theorem A.2

In problem (25),

$$f(x) = \sum_{C\in\mathcal{G}} \frac{|C|}{n} f_C(x) = \sum_{i=1}^{m} \frac{|C_i|}{n} f_{C_i}(x).$$

Hence by choosing $\tilde{f}_i = f_{C_i}$ and $\tilde{\lambda}_i = \frac{|C_i|}{n}$ for $1 \leq i \leq m$, problem (25) has the same form as problem (1), and the Lipschitz smothness constant $\tilde{L}_i$ of $\tilde{f}_i$ is L_{C_i}.

From the partition sampling S, we can construct a sampling $\tilde{S}$ from S as follows: $\tilde{S}(S) = \{i\}$ if $S = C_i$. It is obvious that $\mathbb{P}[\tilde{S} = \{i\}] = p_{C_i}$ for all $1 \leq i \leq m$. For sampling $\tilde{S}$, we choose $\theta_{\tilde{S}}$ be such that $\theta_{\{i\}} = \frac{1}{p_{C_i}}\mathbf{I}$. For the sampling $\tilde{S}$ and $\theta_{\tilde{S}}$, the corresponding $\tilde{\beta}_i = \frac{1}{p_{C_i}}$ from (9).

Let $\tilde{S}_k(S_k) = \{i\}$ if $S_k = C_i$ for $k \geq 0$. Then $\tilde{S}_k \sim \tilde{S}$ is equivalent to $S_k \sim S$. Let $\{\tilde{x}^k, \tilde{\mathbf{J}}^k\}$ be produced by Algorithm 1 with $\tilde{\lambda}_i$, $\tilde{f}_i$, $\tilde{S}$, and $\theta_{\tilde{S}}$. Then it is easy to see that

$$\begin{cases} \tilde{x}^k = x^k, \\ \tilde{\mathbf{J}}^k_{:i} = \frac{1}{|C_i|}\mathbf{J}^k\Pi_{\mathbf{I}_{C_i}}e, \ \forall 1 \leq i \leq m. \end{cases} \tag{32}$$

Therefore, by applying Thm 3.3 to $\{\tilde{x}^k, \tilde{\mathbf{J}}^k\}$, we can get the results.

D. Nonsmooth Case: Proofs of Theorems 4.4 and 4.5

D.1. Lemmas

Lemma D.1. Let $\mathbb{E}_k[\cdot]$ denote the expectation conditional on $\mathbf{J}^k$ and x^k. First, from Assumption 2.1, it is evident that

$$\mathbb{E}_k[g^k] = \mathbf{J}^k\lambda + \nabla f(x^k) - \mathbf{J}^k\lambda = \nabla f(x^k). \tag{33}$$

Lemma D.2. Under Assumption 4.1 and Assumption 4.3, for any $x^*, y^* \in \mathcal{X}^*$ and $x \in \mathbb{R}^d$, we have,

$$\langle \nabla f(x) - \nabla f(y^*), x - y^* \rangle \geq \gamma \sum_{i=1}^{n} \lambda_i \left\| \nabla \phi_i(\mathbf{A}_i^\top x) - \nabla \phi_i(\mathbf{A}_i^\top x^*) \right\|^2.$$

Proof. Since ϕ_i is $1/\gamma$-smooth, we have

$$\langle \nabla \phi_i(\tilde{x}) - \nabla \phi_i(\tilde{y}), \tilde{x} - \tilde{y} \rangle \geq \gamma \left\| \nabla \phi_i(\tilde{x}) - \nabla \phi_i(\tilde{y}) \right\|^2, \ \ \forall i = 1, \cdots, n.$$

Let $\tilde{x} = \mathbf{A}_i^\top x$, and $\tilde{y} = \mathbf{A}_i^\top y^*$ in the above inequlity. Then we get

$$\langle \nabla \phi_i(\mathbf{A}_i^\top x) - \nabla \phi_i(\mathbf{A}_i^\top y^*), \mathbf{A}_i(x - y^*) \rangle \geq \gamma \left\| \nabla \phi_i(\mathbf{A}_i^\top x) - \nabla \phi_i(\mathbf{A}_i^\top y^*) \right\|^2, \ \ \forall i = 1, \cdots, n,$$

which is actually

$$\langle \nabla f_i(x) - \nabla f_i(y^*), x - y^* \rangle \geq \gamma \left\| \nabla \phi_i(\mathbf{A}_i^\top x) - \nabla \phi_i(\mathbf{A}_i^\top y^*) \right\|^2, \ \ \forall i = 1, \cdots, n,$$

Summing over i we get

$$\langle \nabla f(x) - \nabla f(y^*), x - y^* \rangle \geq \gamma \sum_{i=1}^{n} \lambda_i \left\| \nabla \phi_i(\mathbf{A}_i^\top x) - \nabla \phi_i(\mathbf{A}_i^\top y^*) \right\|^2$$

The statement then follows directly from Assumption 4.3. □

Lemma D.3. Under Assumption 4.2, if f is L-smooth and $\alpha \leq \frac{1}{3L}$, then there is $\mu > 0$ such that for all $k \geq 0$,

$$\mathbb{E}_k\left[\|x^{k+1} - [x^{k+1}]^*\|^2\right] \leq \frac{1}{1+\mu\alpha}\mathbb{E}_k\left[\|x^k - [x^k]^*\|^2\right] + \frac{2\alpha^2}{1+\mu\alpha}\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right]$$

[a]

[a] This result was mainly proved in (Xiao & Zhang, 2014a). For completeness, we include a proof.

Proof. Since

$$x^{k+1} = \arg\min\left\{\frac{1}{2\alpha}\|y - x^k + \alpha g^k\|^2 + \psi(y)\right\},$$

we know $\alpha^{-1}(x^k - x^{k+1}) - g^k \in \partial\psi(x^{k+1})$. Using the convexity of f and ψ, we obtain

$$P^* \geq f(x^k) + \langle \nabla f(x^k), [x^k]^* - x^k \rangle + \psi(x^{k+1}) + \langle \alpha^{-1}(x^k - x^{k+1}) - g^k, [x^k]^* - x^{k+1} \rangle. \tag{34}$$

Next we bound $f(x^{k+1})$ by $f(x^k)$. Since f is L-smooth, we have

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2. \tag{35}$$

Combining (34) and (35) we get

$$\begin{aligned}
f(x^{k+1})+\psi(x^{k+1})-P^* &\leq \langle \nabla f(x^k), x^{k+1}-[x^k]^*\rangle + \frac{L}{2}\|x^{k+1}-x^k\|^2 \\
&\quad -\langle \alpha^{-1}(x^k-x^{k+1})-g^k, [x^k]^*-x^{k+1}\rangle \\
&= \frac{L}{2}\|x^{k+1}-x^k\|^2 - \frac{1}{\alpha}\langle x^k-x^{k+1}, [x^k]^*-x^{k+1}\rangle + \langle \nabla f(x^k)-g^k, [x^k]^*-x^{k+1}\rangle \\
&= \left(\frac{L}{2}-\frac{1}{2\alpha}\right)\|x^{k+1}-x^k\|^2 - \frac{1}{2\alpha}\|x^{k+1}-x^k\|^2 \\
&\quad -\frac{1}{\alpha}\langle x^k-x^{k+1}, [x^k]^*-x^k\rangle + \langle \nabla f(x^k)-g^k, [x^k]^*-x^{k+1}\rangle \\
&\leq -\frac{1}{2\alpha}\|x^{k+1}-x^k\|^2 - \frac{1}{\alpha}\langle x^k-x^{k+1}, [x^k]^*-x^k\rangle + \langle \nabla f(x^k)-g^k, [x^k]^*-x^{k+1}\rangle \\
&= -\frac{1}{2\alpha}\|x^{k+1}-[x^k]^*\|^2 + \frac{1}{2\alpha}\|x^k-[x^k]^*\|^2 + \langle \nabla f(x^k)-g^k, [x^k]^*-x^{k+1}\rangle.
\end{aligned}$$

In view of (20), we have

$$(\mu\alpha+1)\|x^{k+1}-[x^{k+1}]^*\|^2 \leq \|x^k-[x^k]^*\|^2 + 2\alpha\langle \nabla f(x^k)-g^k, [x^k]^*-x^{k+1}\rangle.$$

Taking conditional expectation on both side, we get

$$\begin{aligned}
\mathbb{E}_k\left[\|x^{k+1}-[x^{k+1}]^*\|^2\right] &\leq \frac{1}{\mu\alpha+1}\mathbb{E}_k\left[\|x^k-[x^k]^*\|^2\right] + \frac{2\alpha}{\mu\alpha+1}\mathbb{E}_k\left[\langle \nabla f(x^k)-g^k, [x^k]^*-x^{k+1}\rangle\right] \\
&\overset{(33)}{=} \frac{1}{\mu\alpha+1}\mathbb{E}_k\left[\|x^k-[x^k]^*\|^2\right] + \frac{2\alpha}{\mu\alpha+1}\mathbb{E}_k\left[\langle \nabla f(x^k)-g^k, \bar{x}^{k+1}-x^{k+1}\rangle\right],
\end{aligned}$$

where $\bar{x}^{k+1}=\operatorname{prox}_\alpha^\psi(x^k-\alpha\nabla f(x^k))$ is determined by x^k. Recall that $x^{k+1}=\operatorname{prox}_\alpha^\psi(x^k-\alpha g^k)$. We use the non-expansiveness of the proximal mapping:

$$\|\bar{x}^{k+1}-x^{k+1}\| \leq \alpha\|\nabla f(x^k)-g^k\|.$$

Then the statement follows by the Cauchy-Schwarz inequality. □

Lemma D.4.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \sigma_i p_i^{-1} v_i \lambda_i^2 \left\|\alpha_i^{k+1}-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right] &\leq \mathbb{E}\left[\sum_{i=1}^n \sigma_i v_i \lambda_i^2 \left\|\nabla\phi(\mathbf{A}_i^\top x^k)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right] \\
&\quad +\mathbb{E}\left[\sum_{i=1}^n \sigma_i p_i^{-1} v_i \lambda_i^2 (1-p_i)\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right]
\end{aligned}$$

Proof. First notice that

$$\alpha_i^{k+1} = \begin{cases} \nabla\phi(\mathbf{A}_i^\top x^k), & \text{if } i\in S^k \\ \alpha_i^k, & \text{if } i\notin S^k \end{cases}$$

Then by taking conditional expectation on α^k, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n \sigma_i p_i^{-1} v_i \lambda_i^2 \left\|\alpha_i^{k+1}-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 \mid \alpha^k\right] \\
&= \sum_{i=1}^n \sigma_i p_i^{-1} v_i \lambda_i^2 \mathbb{E}\left[\left\|\alpha_i^{k+1}-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 | \alpha^k\right] \\
&= \sum_{i=1}^n \sigma_i p_i^{-1} v_i \lambda_i^2 \left(p_i\|\nabla\phi_i(\mathbf{A}_i^\top x^k)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\|^2 + (1-p_i)\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right).
\end{aligned}$$

Taking expectations again and applying the tower property, we obtain the result.

□

Lemma D.5. For any $x^*, y^* \in \mathcal{X}^*$,

$$\mathbb{E}\left[\|g^k - \nabla f(y^*)\|^2 \mid \mathbf{J}^k, x^k\right] \leq 2\sum_{i=1}^n p_i^{-1} v_i \lambda_i^2 \left\|\nabla\phi_i(\mathbf{A}_i^\top x^k) - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 + 2\sum_{i=1}^n p_i^{-1} v_i \lambda_i^2 \left\|\alpha_i^k - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2.$$

Proof. Recall that

$$\begin{aligned} g^k &= \mathbf{J}^k\lambda + \left(\mathbf{G}(x^k) - \mathbf{J}^k\right)\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda \\ &= \mathbf{J}^k\lambda - \left(\mathbf{J}^k - \mathbf{G}(x^*)\right)\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda + (\mathbf{G}(x^k) - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda. \end{aligned}$$

By applying Lemma C.1 with $z = 2$, we get

$$\|g^k - \nabla f(y^*)\|^2 \leq 2\left\|(\mathbf{G}(x^k) - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 + 2\left\|(\mathbf{J}^k - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda - \mathbf{J}^k\lambda + \nabla f(y^*)\right\|^2,$$

which implies that

$$\begin{aligned} \mathbb{E}\left[\|g^k - \nabla f(y^*)\|^2 \mid \mathbf{J}^k, x^k\right] &\leq 2\mathbb{E}\left[\left\|(\mathbf{G}(x^k) - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right] \\ &\quad + 2\mathbb{E}\left[\left\|(\mathbf{J}^k - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda - \mathbf{J}^k\lambda + \nabla f(y^*)\right\|^2 \mid \mathbf{J}^k, x^k\right] \\ &\leq 2\mathbb{E}\left[\left\|(\mathbf{G}(x^k) - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right] \\ &\quad + 2\mathbb{E}\left[\left\|(\mathbf{J}^k - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda\right\|^2 \mid \mathbf{J}^k, x^k\right], \end{aligned}$$

where the last inequality follows from

$$\mathbb{E}\left[(\mathbf{J}^k - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda \mid \mathbf{J}^k, x^k\right] = \mathbf{J}^k\lambda - \nabla f(x^*) \overset{Assumption\ 4.3}{=} \mathbf{J}^k\lambda - \nabla f(y^*)$$

and $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$. Hence,

$$\begin{aligned} \mathbb{E}\left[\|g^k - \nabla f(x^*)\|^2 \mid \mathbf{J}^k, x^k\right] &\leq 2\mathbb{E}\left[\left\|\sum_{i\in S_k} p_i^{-1}\lambda_i\mathbf{A}_i\left(\phi_i'(\mathbf{A}_i^\top x^k) - \phi_i'(\mathbf{A}_i^\top x^*)\right)\right\|^2 \mid \mathbf{J}^k, x^k\right] \\ &\quad + 2\mathbb{E}\left[\left\|\sum_{i\in S_k} p_i^{-1}\lambda_i\mathbf{A}_i\left(\alpha^k - \phi_i'(\mathbf{A}_i^\top x^*)\right)\right\|^2 \mid \mathbf{J}^k, x^k\right] \\ &\overset{(17)}{\leq} 2\sum_{i=1}^n p_i^{-1} v_i \lambda_i^2 \left\|\phi_i'(\mathbf{A}_i^\top x^k) - \phi_i'(\mathbf{A}_i^\top x^*)\right\|^2 + 2\sum_{i=1}^n p_i^{-1} v_i \lambda_i^2 \left\|\alpha_i^k - \phi_i'(\mathbf{A}_i^\top x^*)\right\|^2. \end{aligned}$$

□

Lemma D.6. For any $x^* \in \mathcal{X}^*$,

$$\mathbb{E}\left[\|g^k - \nabla f(x^k)\|^2 \mid \mathbf{J}^k, x^k\right] \leq 2\sum_{i=1}^n p_i^{-1} v_i \lambda_i^2 \left\|\nabla\phi_i(\mathbf{A}_i^\top x^k) - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 + 2\sum_{i=1}^n p_i^{-1} v_i \lambda_i^2 \left\|\alpha_i^k - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2.$$

Proof.

$$\begin{aligned} g^k - \nabla f(x^k) &= \mathbf{J}^k\lambda - (\mathbf{J}^k - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda + (\mathbf{G}(x^k) - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda - \nabla f(x^k) \\ &= \mathbf{J}^k\lambda - (\mathbf{J}^k - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda - \nabla f(x^*) \\ &\quad +(\mathbf{G}(x^k) - \mathbf{G}(x^*))\theta_{S_k}\mathbf{\Pi}_{\mathbf{I}_{S_k}}\lambda - \nabla f(x^k) + \nabla f(x^*). \end{aligned}$$

The rest of the proof is the same as in Lemma D.5. □

D.2. Proof of Theorem 4.4

For any $x^* \in \mathcal{X}^*$, we have

$$x^* = \text{prox}_\alpha^\psi(x^* - \alpha\nabla f(x^*)).$$

Therefore,

$$\begin{aligned} & \mathbb{E}_k\left[\left\|x^{k+1} - [x^{k+1}]^*\right\|^2\right] \\ \leq\ & \mathbb{E}_k\left[\left\|x^{k+1} - [x^k]^*\right\|^2\right] \\ =\ & \mathbb{E}_k\left[\left\|\text{prox}_\alpha^\psi(x^k - \alpha g^k) - [x^k]^*\right\|^2\right] \\ =\ & \mathbb{E}_k\left[\left\|\text{prox}_\alpha^\psi(x^k - \alpha g^k) - \text{prox}_\alpha^\psi([x^k]^* - \alpha\nabla f([x^k]^*))\right\|^2\right] \\ \leq\ & \mathbb{E}_k\left[\left\|x^k - \alpha g^k - ([x^k]^* - \alpha\nabla f([x^k]^*))\right\|^2\right] \\ \overset{(33)}{=}\ & \left\|x^k - [x^k]^*\right\|^2 - 2\alpha\langle\nabla f(x^k) - \nabla f([x^k]^*), x^k - [x^k]^*\rangle + \alpha^2\mathbb{E}_k\left[\|g^k - \nabla f([x^k]^*)\|^2\right] \end{aligned}$$

Now we apply Lemma D.3. For any $\beta \in [0,1]$, we have

$$\begin{aligned} & \mathbb{E}_k\left[\left\|x^{k+1} - [x^{k+1}]^*\right\|^2\right] \\ \leq\ & \left(\frac{\beta}{1+\alpha\mu} + 1 - \beta\right)\left\|x^k - [x^k]^*\right\|^2 - 2\alpha(1-\beta)\langle\nabla f(x^k) - \nabla f([x^k]^*), x^k - [x^k]^*\rangle \\ & +\alpha^2(1-\beta)\mathbb{E}_k\left[\|g^k - \nabla f([x^k]^*)\|^2\right] + \frac{2\alpha^2\beta}{1+\mu\alpha}\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] \end{aligned}$$

Plugging in Lemma D.2, Lemma D.5 and Lemma D.6 we obtain:

$$\begin{aligned} & \mathbb{E}_k\left[\left\|x^{k+1} - [x^{k+1}]^*\right\|^2\right] \\ \leq\ & \left(\frac{\beta}{1+\alpha\mu} + 1 - \beta\right)\left\|x^k - [x^k]^*\right\|^2 - 2\gamma\alpha(1-\beta)\sum_{i=1}^n \lambda_i\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k) - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 \\ & + \left(\alpha^2(1-\beta) + \frac{2\alpha^2\beta}{1+\mu\alpha}\right)\left(2\sum_{i=1}^n p_i^{-1}v_i\lambda_i^2\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k) - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 + 2\sum_{i=1}^n p_i^{-1}v_i\lambda_i^2\left\|\alpha_i^k - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right) \\ =\ & \left(\frac{\beta}{1+\alpha\mu} + 1 - \beta\right)\left\|x^k - [x^k]^*\right\|^2 + \sum_{i=1}^n 2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta) + \frac{2\alpha^2\beta}{1+\mu\alpha}\right)\left\|\alpha_i^k - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 \\ & - \sum_{i=1}^n\left(2\gamma\lambda_i\alpha(1-\beta) - 2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta) + \frac{2\alpha^2\beta}{1+\mu\alpha}\right)\right)\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k) - \nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 \end{aligned}$$

Taking expectation again and applying the tower property, we obtain

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1}-[x^{k+1}]^*\|^2\right] &\leq \left(\frac{\beta}{1+\alpha\mu}+1-\beta\right)\mathbb{E}\left[\left\|x^k-[x^k]^*\right\|^2\right] \\ &\quad+\sum_{i=1}^n 2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta)+\frac{2\alpha^2\beta}{1+\mu\alpha}\right)\mathbb{E}\left[\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right] \\ &\quad-\sum_{i=1}^n\left(2\gamma\lambda_i\alpha(1-\beta)-2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta)+\frac{2\alpha^2\beta}{1+\mu\alpha}\right)\right)\mathbb{E}\left[\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right]\end{aligned}$$

Therefore, for the stochastic Lyapunov function Ψ^{k+1}, we have in view of Lemma D.4,

$$\begin{aligned}\mathbb{E}\left[\Psi^{k+1}\right] &\leq \left(\frac{\beta}{1+\alpha\mu}+1-\beta\right)\mathbb{E}\left[\left\|x^k-[x^k]^*\right\|^2\right] \\ &\quad+\sum_{i=1}^n 2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta)+\frac{2\alpha^2\beta}{1+\mu\alpha}+\frac{\alpha\sigma_i(1-p_i)}{2}\right)\mathbb{E}\left[\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right] \\ &\quad-\sum_{i=1}^n\left(2\gamma\lambda_i\alpha(1-\beta)-2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta)+\frac{2\alpha^2\beta}{1+\mu\alpha}\right)-\alpha\sigma_i v_i\lambda_i^2\right)\mathbb{E}\left[\left\|\nabla\phi_i(\mathbf{A}_i^\top x)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right]\end{aligned}$$

In order to guarantee that $\mathbb{E}[\Psi^{k+1}]\leq\left(\frac{\beta}{1+\alpha\mu}+1-\beta\right)\mathbb{E}[\Psi^k]$, α and β should be chosen such that

$$2p_i^{-1}v_i\lambda_i^2\left(\alpha^2(1-\beta)+\frac{2\alpha^2\beta}{1+\mu\alpha}\right)+\alpha\sigma_i v_i\lambda_i^2\leq 2\gamma\lambda_i\alpha(1-\beta),$$

and

$$\alpha^2(1-\beta)+\frac{2\alpha^2\beta}{1+\mu\alpha}+\frac{\alpha\sigma_i(1-p_i)}{2}\leq\frac{\alpha\sigma_i}{2}\left(\frac{\beta}{1+\alpha\mu}+1-\beta\right).$$

for all $1\leq i\leq n$. Now we let $\beta=\frac{1}{2}$ and $\delta\geq\alpha\left(\frac{1}{2}+\frac{1}{1+\mu\alpha}\right)$ so that

$$1-\delta\mu\leq\frac{1}{2(1+\alpha\mu)}+\frac{1}{2}.$$

Then the above inequalities can be satisfied if

$$p_i^{-1}v_i\lambda_i\delta+\frac{\sigma_i v_i\lambda_i}{2}\leq\frac{\gamma}{2},\quad\Rightarrow\quad\delta\leq\frac{\gamma p_i}{2v_i\lambda_i}-\frac{\sigma_i p_i}{2},\tag{36}$$

and

$$\delta+\frac{\sigma_i(1-p_i)}{2}\leq\frac{\sigma_i}{2}(1-\delta\mu),\quad\Rightarrow\quad\delta\leq\frac{\sigma_i p_i}{\mu\sigma_i+2},\tag{37}$$

for all $1\leq i\leq n$. Since $\sigma_i=\frac{\gamma}{2v_i\lambda_i}$, if δ satisfies

$$\delta\leq\min_{1\leq i\leq n}\frac{p_i}{\mu+4v_i\lambda_i/\gamma},$$

then (36) and (37) hold. This means if we choose α such that

$$\alpha\left(\frac{1}{2}+\frac{1}{1+\mu\alpha}\right)\leq\min_{1\leq i\leq n}\frac{p_i}{\mu+4v_i\lambda_i/\gamma},$$

then

$$\mathbb{E}[\Psi^{k+1}]\leq\left(\frac{1}{2(1+\alpha\mu)}+\frac{1}{2}\right)\mathbb{E}[\Psi^k].$$

In particular, we can choose

$$\alpha \leq \frac{2}{3}\min_{1\leq i\leq n}\frac{p_i}{\mu+4v_i\lambda_i/\gamma}.$$

Notice that when we apply Lemma D.3, α should satisfy $\alpha \leq \frac{1}{3L}$. Hence,

$$\alpha \leq \min\left\{\frac{2}{3}\min_{1\leq i\leq n}\frac{p_i}{\mu+4v_i\lambda_i/\gamma}, \frac{1}{3L}\right\}.$$

From $\mathbb{E}[\Psi^k] \leq (1-\frac{\alpha\mu}{2(1+\alpha\mu)})^k\mathbb{E}[\Psi^0]$, we know if

$$k \geq \left(2+\max\left\{\frac{6L}{\mu}, 3\max_i\left(\frac{1}{p_i}+\frac{4v_i\lambda_i}{\mu p_i\gamma}\right)\right\}\right)\log\left(\frac{1}{\epsilon}\right)$$

Then $\mathbb{E}[\Psi^k] \leq \epsilon\mathbb{E}[\Psi^0]$.

If μ is unknown, we can choose

$$\delta \leq \min_{1\leq i\leq n}\frac{p_i}{8v_i\lambda_i/\gamma}.$$

Then

$$p_i^{-1}v_i\lambda_i\delta + \frac{\sigma_i v_i\lambda_i}{2} \leq \frac{3}{8}\gamma < \frac{\gamma}{2},$$

and

$$1-p_i+\frac{2\delta}{\sigma_i} \leq 1-\frac{p_i}{2},$$

for all $1\leq i\leq n$. This means if we choose α such that

$$\alpha\left(\frac{1}{2}+\frac{1}{1+\mu\alpha}\right) \leq \min_{1\leq i\leq n}\frac{p_i}{8v_i\lambda_i/\gamma},$$

then

$$\mathbb{E}[\Psi^{k+1}] \leq \left(1-\min\left\{\frac{\alpha\mu}{2(1+\alpha\mu)}, \frac{p_i}{2}\right\}\right)\mathbb{E}[\Psi^k].$$

In particular, we can choose

$$\alpha \leq \min_{1\leq i\leq n}\frac{p_i}{12v_i\lambda_i/\gamma}.$$

Notice that when we apply Lemma D.3, α should satisfy $\alpha \leq \frac{1}{3L}$. Hence,

$$\alpha \leq \min\left\{\min_{1\leq i\leq n}\frac{p_i}{12v_i\lambda_i/\gamma}, \frac{1}{3L}\right\}.$$

From $\mathbb{E}[\Psi^k] \leq (1-\min\{\frac{\alpha\mu}{2(1+\alpha\mu)}, \frac{p_i}{2}\})^k\mathbb{E}[\Psi^0]$, we know if

$$k \geq \left(2+\max\left\{\frac{6L}{\mu}, \max_i\left\{\frac{24v_i\lambda_i}{\mu p_i\gamma}, \frac{2}{p_i}\right\}\right\}\right)\log\left(\frac{1}{\epsilon}\right)$$

Then $\mathbb{E}[\Psi^k] \leq \epsilon\mathbb{E}[\Psi^0]$.

D.3. Proof of Theorem 4.5

First notice that, if ψ is μ-strongly convex, then P is μ-strongly convex which implies that the optimal solution of problem (1) is unique. Let $\mathcal{X}^* = \{x^*\}$. For the x^*, we have

$$x^* = \text{prox}_\alpha^\psi(x^* - \alpha\nabla f(x^*)).$$

Therefore,

$$\begin{aligned}
& \mathbb{E}_k\left[\left\|x^{k+1}-x^*\right\|^2\right] \\
= \quad & \mathbb{E}_k\left[\left\|\mathrm{prox}_\alpha^\psi(x^k-\alpha g^k)-x^*\right\|^2\right] \\
= \quad & \mathbb{E}_k\left[\left\|\mathrm{prox}_\alpha^\psi(x^k-\alpha g^k)-\mathrm{prox}_\alpha^\psi(x^*-\alpha\nabla f(x^*))\right\|^2\right] \\
\leq \quad & \frac{1}{1+\alpha\mu}\mathbb{E}_k\left[\left\|x^k-\alpha g^k-(x^*-\alpha\nabla f(x^*))\right\|^2\right] \\
\overset{(33)}{=} \quad & \frac{1}{1+\alpha\mu}\left\|x^k-x^*\right\|^2-\frac{2\alpha}{1+\alpha\mu}\langle\nabla f(x^k)-\nabla f(x^*),x^k-x^*\rangle+\frac{\alpha^2}{1+\alpha\mu}\mathbb{E}_k\left[\|g^k-\nabla f(x^*)\|^2\right] \\
\overset{\text{Lemma D.2 and D.5}}{\leq} \quad & \frac{1}{1+\alpha\mu}\|x^k-x^*\|^2+\frac{2\alpha}{1+\alpha\mu}\sum_{i=1}^n\left(\frac{\alpha v_i\lambda_i^2}{p_i}-\gamma\lambda_i\right)\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2 \\
& \quad +\frac{2\alpha^2}{1+\alpha\mu}\sum_{i=1}^n p_i^{-1}v_i\lambda_i^2\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2
\end{aligned}$$

Taking expectation again and applying the tower property, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1}-x^*\|^2\right] \quad \leq \quad & \frac{1}{1+\alpha\mu}\mathbb{E}[\|x^k-x^*\|^2]+\frac{2\alpha}{1+\alpha\mu}\mathbb{E}\left[\sum_{i=1}^n\left(\frac{\alpha v_i\lambda_i^2}{p_i}-\gamma\lambda_i\right)\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right] \\
& +\frac{2\alpha^2}{1+\alpha\mu}\mathbb{E}\left[\sum_{i=1}^n p_i^{-1}v_i\lambda_i^2\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right]
\end{aligned}$$

Therefore, for the stochastic Lyapunov function Ψ^{k+1}, we have

$$\begin{aligned}
\mathbb{E}\left[\Psi^{k+1}\right] \quad \overset{\text{Lemma } D.4}{\leq} \quad & \frac{1}{1+\alpha\mu}\mathbb{E}[\|x^k-x^*\|^2]+\frac{\alpha}{1+\alpha\mu}\mathbb{E}\left[\sum_{i=1}^n p_i^{-1}v_i\lambda_i^2\left(2\alpha+(1+\alpha\mu)\sigma_i(1-p_i)\right)\left\|\alpha_i^k-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right] \\
& \frac{2\alpha}{1+\alpha\mu}\mathbb{E}\left[\sum_{i=1}^n\lambda_i\left(\frac{\alpha v_i\lambda_i}{p_i}-\gamma+\frac{1+\alpha\mu}{2}\sigma_i v_i\lambda_i\right)\left\|\nabla\phi_i(\mathbf{A}_i^\top x^k)-\nabla\phi_i(\mathbf{A}_i^\top x^*)\right\|^2\right]
\end{aligned}$$

In order to guarantee that $\mathbb{E}[\Psi^{k+1}]\leq\left(\frac{1}{1+\alpha\mu}\right)\mathbb{E}[\Psi^k]$, α should be choosen such that

$$\frac{\alpha v_i\lambda_i}{p_i}-\gamma+\frac{1+\alpha\mu}{2}\sigma_i v_i\lambda_i\leq 0, \quad \Rightarrow \quad \alpha\leq\frac{p_i\gamma}{v_i\lambda_i}-\frac{1+\alpha\mu}{2}p_i\sigma_i,$$

and

$$2\alpha+(1+\alpha\mu)\sigma_i(1-p_i)\leq\sigma_i, \quad \Rightarrow \quad \alpha\leq\frac{\sigma_i p_i}{2+\mu\sigma_i(1-p_i)},$$

for all $1\leq i\leq n$. Assume $\alpha\leq 1/\mu$. Then the above inequalities can be satisfied if

$$\alpha\leq\frac{p_i\gamma}{v_i\lambda_i}-p_i\sigma_i,$$

and

$$\alpha\leq\frac{\sigma_i p_i}{2+\mu\sigma_i},$$

for all i. Since $\sigma_i=\frac{2\gamma}{3v_i\lambda_i}$, if α is chosen as

$$\alpha\leq\min_{1\leq i\leq n}\frac{p_i}{\mu+3v_i\lambda_i/\gamma},$$

which is actually satisfies $\alpha \leq 1/\mu$, then we have the recursion $\mathbb{E}[\Psi^{k+1}] \leq \left(\frac{1}{1+\alpha\mu}\right)\mathbb{E}[\Psi^k]$.

In the case where μ is unknown, we can choose $\sigma_i = \frac{\gamma}{(1+\alpha\mu)v_i\lambda_i}$ and

$$\alpha \leq \min_{1\leq i\leq n} \frac{p_i\gamma}{4v_i\lambda_i}.$$

Then

$$\frac{\alpha v_i\lambda_i}{p_i} - \gamma + \frac{1+\alpha\mu}{2}\sigma_i v_i\lambda_i \leq -\frac{\gamma}{4} < 1,$$

and

$$1 - p_i + \frac{2\alpha}{(1+\alpha\mu)\sigma_i} \leq 1 - \frac{p_i}{2},$$

for all $1 \leq i \leq n$. Therefore, we have the recursion

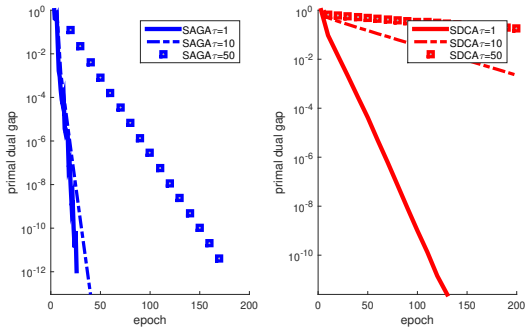
$$\mathbb{E}[\Psi^{k+1}] \leq \left(1 - \min\left\{\frac{\alpha\mu}{1+\alpha\mu}, \frac{p_i}{2}\right\}\right)\mathbb{E}[\Psi^k]$$

E. Extra Experiments

We include in this section more experimental results.

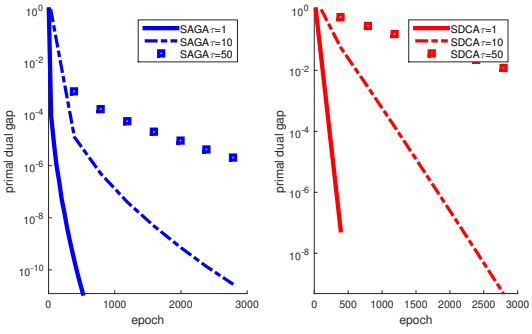
E.1. Batch sampling

Here we compare SAGA-AS with SDCA, for the case when S is a τ-nice sampling for three different values of $\tau \in \{1, 10, 50\}$. Note that SDCA with τ-nice sampling works the same both in theory and in practice as Quartz with τ-nice sampling. We report in Figure 4, Figure 5 and Figure 6 the results obtained for the dataset ijcnn1, a9a and w8a. Note that for ijcnn1, when we increase τ by 50, the number of epochs of SAGA-AS only increased by less than 6. This indicates a considerable speedup if parallel computation can be included in the implementation of mini-batch case.



(a) SAGA; ijcnn1 (b) SDCA; ijcnn1

Figure 4. mini-batch SAGA V.S. mini-batch SDCA: ijcnn1

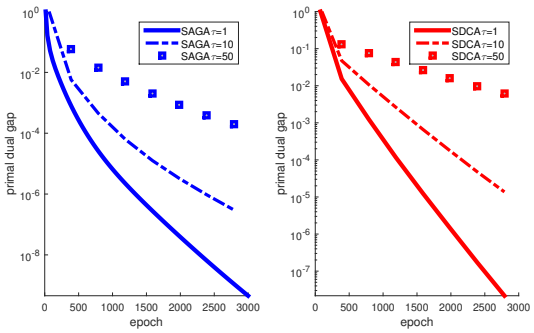


(a) SAGA; a9a (b) SDCA; a9a

Figure 5. mini-batch SAGA V.S. mini-batch SDCA: a9a

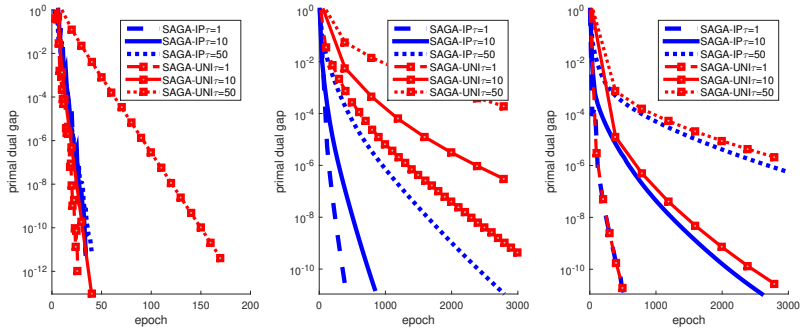
E.2. Importance sampling

We compare uniform sampling SAGA (SAGA-UNI) with importance sampling SAGA (SAGA-IP), as described in Section 3.3 , on three values of $\tau \in \{1, 10, 50\}$. The results for the datasets w8a, ijcnn1 and a9a are shown in Figure 7. For the dataset ijcnn1, mini-batch with importance sampling almost achieves linear speedup as the number of epochs does not increase with τ. For the dataset w8a, mini-batch with importance sampling can even need less number of epochs than serial uniform sampling. For the dataset a9a, importance sampling slightly but consistently improves over uniform sampling. Note that we adopt the importance sampling strategy described in (Hanzely & Richtárik, 2019) and the actual running time is the same as uniform sampling.



(a) SAGA; w8a (b) SDCA; w8a

Figure 6. mini-batch SAGA V.S. mini-batch SDCA: w8a

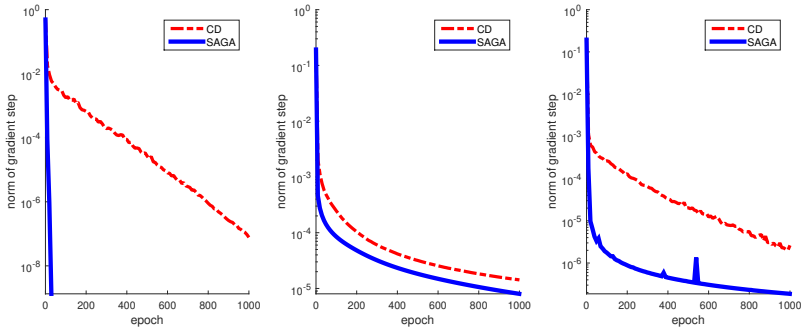


(a) ijcnn1 (b) w8a (c) a9a

Figure 7. importance sampling V.S. uniform sampling

E.3. Comparison with Coordinate Descent

We consider the un-regularized logistic regression problem (23) with $\lambda_1 = \lambda_2 = 0$. In this case, Thm 4.4 applies and we expect to have linear convergence of SAGA without any knowledge on the constant μ satsifying Assumption (4.2), see Remark **??**. This makes SAGA comparable with descent methods such as gradient method and coordinate descent (CD) method. However, comparing with their deterministic counterparts, the speedup provided by CD can be at most of order d while the speedup by SAGA can be of order n. Thus SAGA is much preferable than CD when n is larger than d. We provide numerical evidence in Figure 8.



(a) ijcnn1 (b) w8a (c) a9a

Figure 8. SAGA V.S. CD