Master's Thesis
May 1, 2017

Parameter Estimation in High-Dimensions using Doubly Stochastic Approximation

Alec Koppel
email: akoppel@seas.upenn.edu
Dept. of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104

May 1, 2017
In collaboration with Aryan Mokhtari, Alejandro Ribeiro, Dept. of ESE, UPenn.

Abstract

We consider learning problems over training sets in which both, the number of training examples and the dimension of the feature vectors, are large. To solve these problems we propose the random parallel stochastic algorithm (RAPSA). We call the algorithm random parallel because it utilizes multiple parallel processors to operate on a randomly chosen subset of blocks

of the feature vector. We call the algorithm stochastic because processors choose training subsets uniformly at random. Algorithms that are parallel in either of these dimensions exist, but RAPSA is the first attempt at a methodology that is parallel in both the selection of blocks and the selection of elements of the training set. In RAPSA, processors utilize the randomly chosen functions to compute the stochastic gradient component associated with a randomly chosen block. The technical contribution of this paper is to show that this minimally coordinated algorithm converges to the optimal classifier when the training objective is convex. Moreover, we present an accelerated version of RAPSA (ARAPSA) that incorporates the objective function curvature information by premultiplying the descent direction by a Hessian approximation matrix. We further extend the results for asynchronous settings and show that if the processors perform their updates without any coordination the algorithms are still convergent to the optimal argument. RAPSA and its extensions are then numerically evaluated on a linear estimation problem and a binary image classification task using the MNIST handwritten digit dataset.[1]

A Introduction

Learning is often formulated as an optimization problem that finds a vector of parameters $\mathbf{x}^* \in \mathbb{R}^p$ that minimizes the average of a loss function across the elements of a training set. For a precise definition consider a training set with N elements and let $f_n : \mathbb{R}^p \to \mathbb{R}$ be a convex loss function associated with the nth element of the training set. The optimal parameter vector $\mathbf{x}^* \in \mathbb{R}^p$ is defined as the minimizer of the average cost $F(\mathbf{x}) := (1/N) \sum_{n=1}^{N} f_n(\mathbf{x})$,.

$$\mathbf{x}^* := \underset{\mathbf{x}}{\operatorname{argmin}}\, F(\mathbf{x}) := \underset{\mathbf{x}}{\operatorname{argmin}}\, \frac{1}{N} \sum_{n=1}^{N} f_n(\mathbf{x}). \tag{1}$$

Problems such as support vector machine classification, logistic and linear regression [4], matrix completion, and maximum likelihood estimation [5] can be put in the form of problem (1). In this paper, we are interested in large scale problems where both the number of features p and the number of elements N in the training set are very large – which arise, e.g., in text [6], image [7, 8], and genomic [9] processing.

When N and p are large, the parallel processing architecture in Figure 1 becomes of interest. In this architecture, the parameter vector $\mathbf{x}$ is divided into B blocks each of which contains $p_b \ll p$ features and a set of $I \ll B$ processors work in parallel on randomly chosen parameter blocks while using a stochastic subset of elements of the training set. In the schematic shown, Processor 1 fetches functions f_1 and f_n to operate on block $\mathbf{x}_b$ and Processor i fetches functions $f_{n'}$ and $f_{n''}$ to operate on block $\mathbf{x}_{b'}$. Other processors select other elements of the training set and other blocks with the majority of blocks remaining unchanged and the majority of functions remaining unused. The blocks chosen for update and the functions fetched for determination of block updates are selected independently at random in subsequent slots.

Problems that operate on blocks of the parameter vectors *or* subsets of the training set, but not on both, blocks *and* subsets, exist. Block coordinate descent (BCD) is the generic name for methods in which the variable space is divided in blocks that are processed separately. Early versions operate by cyclically updating all coordinates at each step [10–12], while more recent parallelized versions of coordinate descent have been developed to accelerate convergence of BCD [13–16]. Closer to

[1] This work has appeared as [1–3]

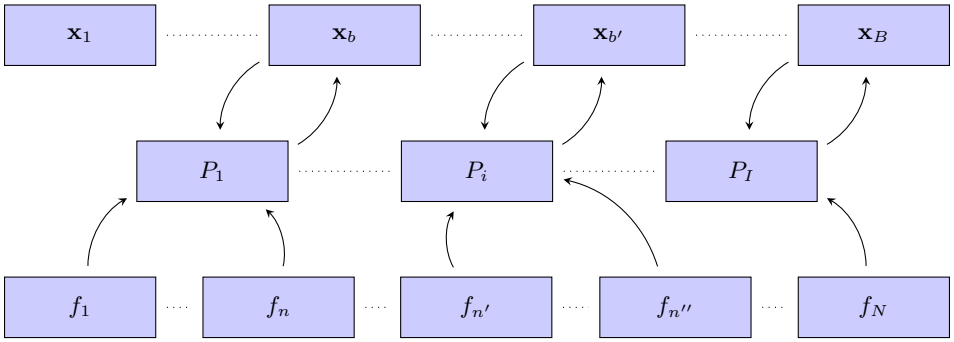


Figure 1: Random parallel stochastic algorithm (RAPSA). At each iteration, processor P_i picks a random block from the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_B\}$ and a random set of functions from the training set $\{f_1, \ldots, f_N\}$. The functions drawn are used to evaluate a stochastic gradient component associated with the chosen block. RAPSA is shown here to converge to the optimal argument $\mathbf{x}^*$ of (1).

the architecture in Figure 1, methods in which subsets of blocks are selected at random have also been proposed [15, 17–19]. BCD, serial, parallel, or random, can handle cases where the parameter dimension p is large but requires access to all N training samples at each iteration.

Parallel implementations of block coordinate methods have been developed initially in this setting for composite optimization problems [13]. A collection of parallel processors update randomly selected blocks concurrently at each step. Several variants that select blocks in order to maximize the descent at each step are proposed in [20–22]. The aforementioned works require that parallel processors operate on a common time index. In contrast, asynchronous parallel methods, originally proposed in [23], have been developed to solve optimization problems where processors are *not* required to operate with a common global clock. This work focused on solving a fixed point problem over a separable convex set, but the analysis is more restrictive than standard convexity assumptions. For a standard strongly convex optimization problem, in contrast, [17] establish linear convergence to the optimum. All of these works are developed for optimization problems with deterministic objectives.

To handle the case where the number of training examples N is very large, methods have been developed to only process a subset of sample points at a time. These methods are known by the generic name of stochastic approximation and rely on the use of stochastic gradients. In plain stochastic gradient descent (SGD), the gradient of the aggregate function is estimated by the gradient of a randomly chosen function f_n [5, 24, 25]. Since convergence of SGD is slow more often that not, various recent developments have been aimed at accelerating its convergence. These attempts include methodologies to reduce the variance of stochastic gradients [26–28] and the use of ideas from quasi-Newton optimization to handle difficult curvature profiles [29–32]. More pertinent to the work considered here are the use of cyclic block SGD updates [33] and the exploitation of sparsity properties of feature vectors to allow for parallel updates [34]. These methods are suitable when the number of elements in the training set N is large but don't allow for parallel feature processing unless parallelism is inherent to the problem's structure.

The random parallel stochastic algorithm (RAPSA) proposed in this paper represents the first effort at implementing the architecture in Fig. 1 that randomizes over both parameters and sample functions, and may be implemented in parallel. In RAPSA, the functions fetched by a processor are used to compute the stochastic gradient component associated with a randomly chosen block

(Section B). The processors do not coordinate in either choice except to avoid selection of the same block. Our main technical contribution is to show that RAPSA iterates converge to the optimal classifier $\mathbf{x}^*$ when using a sequence of decreasing stepsizes and to a neighborhood of the optimal classifier when using constant stepsizes (Section E). In the latter case, we further show that the rate of convergence to this optimality neighborhood is linear in expectation. These results are interesting because only a subset of features are updated per iteration and the functions used to update different blocks are, in general, different. We propose two extensions of RAPSA. Firstly, motivated by the improved performance results of quasi-Newton methods relative to gradient methods in online optimization, we propose an extension of RAPSA which incorporates approximate second-order information of the objective, called Accelerated RAPSA. We also consider an extension of RAPSA in which parallel processors are not required to operate on a common time index, which we call Asynchronous RAPSA. We further show how these extensions yield an accelerated doubly stochastic algorithm for an asynchronous system. We establish that the performance guarantees of RAPSA carry through to asynchronous computing architectures, even when the amount of asyncronicity is unknown, in contrast to [35], provided that it is bounded. We then numerically evaluate the proposed methods on a large-scale linear regression problem as well as the MNIST digit recognition problem (Section F).

B Random Parallel Stochastic Algorithm (RAPSA)

We consider a more general formulation of (1) in which the number N of functions f_n is not necessarily finite. Introduce then a random variable $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q$ that determine the choice of the random smooth convex function $f(\cdot, \boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}$. We consider the problem of minimizing the expectation of the random functions $F(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\theta}}[f(\mathbf{x}, \boldsymbol{\theta})]$,

$$\mathbf{x}^* := \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^p} F(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^p} \mathbb{E}_{\boldsymbol{\theta}} \left[f(\mathbf{x}, \boldsymbol{\theta}) \right]. \tag{2}$$

Problem (1) is a particular case of (2) in which each of the functions f_n is drawn with probability $1/N$. Observe that when $\boldsymbol{\theta} = (\mathbf{z}, \mathbf{y})$ with feature vector $\mathbf{z} \in \mathbb{R}^p$ and target variable $\mathbf{y} \in \mathbb{R}^q$ or $y \in \{0, 1\}$, the formulation in (2) encapsulates generic supervised learning problems such as regression or classification, respectively. We refer to $f(\cdot, \boldsymbol{\theta})$ as instantaneous functions and to $F(\mathbf{x})$ as the average function.

RAPSA utilizes I processors to update a random subset of blocks of the variable $\mathbf{x}$, with each of the blocks relying on a subset of randomly and independently chosen elements of the training set; see Figure 1. Formally, decompose the variable $\mathbf{x}$ into B blocks to write $\mathbf{x} = [\mathbf{x}_1; \dots; \mathbf{x}_B]$, where block b has length p_b so that we have $\mathbf{x}_b \in \mathbb{R}^{p_b}$. At iteration t, processor i selects a random index b_i^t for updating and a random subset $\boldsymbol{\Theta}_i^t$ of L instantaneous functions. It then uses these instantaneous functions to determine stochastic gradient components for the subset of variables $\mathbf{x}_b = \mathbf{x}_{b_i^t}$ as an average of the components of the gradients of the functions $f(\mathbf{x}^t, \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t$,

$$\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\theta}), \qquad b = b_i^t. \tag{3}$$

Note that L can be interpreted as the size of mini-batch for gradient approximation. The stochastic gradient block in (3) is then modulated by a possibly time varying stepsize γ^t and used by processor

Algorithm 1 Random Parallel Stochastic Algorithm (RAPSA)

for $t = 0, 1, 2, \dots$ **do**
  loop in parallel, processors $i = 1, \dots, I$ execute:
    Select block $b_i^t \in \{1, \dots, B\}$ uniformly at random from set of blocks
    Choose training subset $\mathbf{\Theta}_i^t$ for block $\mathbf{x}_b$,
    Compute stochastic gradient : $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \mathbf{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\theta}), \quad b = b_i^t$ [cf. (3)]
    Update the coordinates b_i^t of the decision variable $\mathbf{x}_b^{t+1} = \mathbf{x}_b^t - \gamma^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$
  end loop; Transmit updated blocks $i \in \mathcal{I}^t \subset \{1, \dots, B\}$ to shared memory
end for

i to update the block $\mathbf{x}_b = \mathbf{x}_{b_i^t}$

$$\mathbf{x}_b^{t+1} = \mathbf{x}_b^t - \gamma^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t) \qquad b = b_i^t. \tag{4}$$

RAPSA is defined by the joint implementation of (3) and (4) across all I processors, and is summarized in Algorithm 1. We would like to emphasize that the number of updated blocks which is equivalent to the number of processors I is not necessary equal to the total number of blocks B. In other words, we may update only a subset of coordinates $I/B < 1$ at each iteration. We define $r := I/B$ as the ratio of the updated blocks to the total number of blocks which is smaller than 1.

The selection of blocks is coordinated so that no processors operate in the same block. The selection of elements of the training set is uncoordinated across processors. The fact that at any point in time a random subset of blocks is being updated utilizing a random subset of elements of the training set means that RAPSA requires almost no coordination between processors. The contribution of this paper is to show that this very lean algorithm converges to the optimal argument $\mathbf{x}^*$ as we show in Section E.

C Accelerated Random Parallel Stochastic Algorithm (ARAPSA)

As we mentioned in Section B, RAPSA operates on first-order information which may lead to slow convergence in ill-conditioned problems. We introduce Accelerated RAPSA (ARAPSA) as a parallel doubly stochastic algorithm that incorporates second-order information of the objective by separately approximating the function curvature for each block. We do this by implementing the oLBFGS algorithm for different blocks of the variable $\mathbf{x}$. For related approaches, see, for instance, [36–40]. Define $\hat{\mathbf{B}}_b^t$ as an approximation for the Hessian inverse of the objective function that corresponds to the block b with the corresponding variable $\mathbf{x}_b$. If we consider b_i^t as the block that processor i chooses at step t, then the update of ARAPSA is defined as multiplication of the descent direction of RAPSA by $\hat{\mathbf{B}}_b^t$, i.e.

$$\mathbf{x}_b^{t+1} = \mathbf{x}_b^t - \gamma^t \hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t) \qquad b = b_i^t. \tag{5}$$

Subsequently, we define the $\hat{\mathbf{d}}_b^t := \hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$. We next detail how to properly specify the block approximate Hessian $\hat{\mathbf{B}}_b^t$ so that it behaves in a manner comparable to the true Hessian. To

Algorithm 2 Computation of the ARAPSA step $\hat{\mathbf{d}}_b^t = \hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ for block $\mathbf{x}_b$.

function $\hat{\mathbf{d}}_b^t = \mathbf{q}^\tau = \text{ARAPSA Step}\Big(\hat{\mathbf{B}}_b^{t,0},\ \mathbf{p}^0 = \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t),\ \{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1}\Big)$
for $u = 0, 1, \ldots, \tau - 1$ **do** {Loop to compute constants α^u and sequence $\mathbf{p}^u$}
  Compute and store scalar $\alpha^u = \hat{\rho}_b^{t-u-1} (\mathbf{v}_b^{t-u-1})^T \mathbf{p}^u$
  Update sequence vector $\mathbf{p}^{u+1} = \mathbf{p}^u - \alpha^u \hat{\mathbf{r}}_b^{t-u-1}$.
end for
Multiply $\mathbf{p}^\tau$ by initial matrix: $\mathbf{q}^0 = \hat{\mathbf{B}}_b^{t,0} \mathbf{p}^\tau$
for $u = 0, 1, \ldots, \tau - 1$ **do** {Loop to compute constants β_u and sequence $\mathbf{q}_u$}
  Compute scalar $\beta^u = \hat{\rho}_b^{t-\tau+u} (\hat{\mathbf{r}}_b^{t-\tau+u})^T \mathbf{q}^u$
  Update sequence vector $\mathbf{q}^{u+1} = \mathbf{q}^u + (\alpha^{\tau-u-1} - \beta^u) \mathbf{v}_b^{t-\tau+u}$
end for {return $\hat{\mathbf{d}}_b^t = \mathbf{q}^\tau$}

do so, define for each block coordinate $\mathbf{x}_b$ at step t the variable variation $\mathbf{v}_b^t$ and the stochastic gradient variation $\hat{\mathbf{r}}_b^t$ as

$$\mathbf{v}_b^t \ = \ \mathbf{x}_b^{t+1} - \mathbf{x}_b^t, \qquad \hat{\mathbf{r}}_b^t \ = \ \nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \mathbf{\Theta}_i^t) - \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t). \tag{6}$$

Observe that the stochastic gradient variation $\hat{\mathbf{r}}_b^t$ is defined as the difference of stochastic gradients at times $t+1$ and t corresponding to the block $\mathbf{x}_b$ for a common set of realizations $\mathbf{\Theta}_i^t$. The term $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ is the same as the stochastic gradient used at time t in (5), while $\nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \mathbf{\Theta}_i^t)$ is computed only to determine the stochastic gradient variation $\hat{\mathbf{r}}_b^t$. An alternative and perhaps more natural definition for the stochastic gradient variation is $\nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \mathbf{\Theta}_i^{t+1}) - \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$. However, as pointed out in [29], this formulation is insufficient for establishing the convergence of stochastic quasi-Newton methods. We proceed to developing a block-coordinate quasi-Newton method by first noting an important property of the true Hessian, and design our approximate scheme to satisfy this property. In particular, observe that the true Hessian inverse $(\mathbf{H}_b^t)^{-1}$ corresponding to block $\mathbf{x}_b$ satisfies the block secant condition, stated as $(\mathbf{H}_b^t)^{-1} \hat{\mathbf{r}}_b^t = \mathbf{v}_b^t$ when the iterates $\mathbf{x}_b^t$ and $\mathbf{x}_b^{t+1}$ are close to each other. The secant condition may be interpreted as stating that the stochastic gradient of a quadratic approximation of the objective function evaluated at the next iteration agrees with the stochastic gradient at the current iteration. We select a Hessian inverse approximation matrix associated with block $\mathbf{x}_b$ such that it satisfies the secant condition $\hat{\mathbf{B}}_b^{t+1} \hat{\mathbf{r}}_b^t = \mathbf{v}_b^t$, and thus behaves in a comparable manner to the true block Hessian.

The oLBFGS Hessian inverse update rule maintains the secant condition at each iteration by using information of the last $\tau \geq 1$ pairs of variable and stochastic gradient variations $\{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1}$. To state the update rule of oLBFGS for revising the Hessian inverse approximation matrices of the blocks, define a matrix as $\hat{\mathbf{B}}_b^{t,0} := \eta_b^t \mathbf{I}$ for each block b and t, where the constant η_b^t for $t > 0$ is given by

$$\eta_b^t := \frac{(\mathbf{v}_b^{t-1})^T \hat{\mathbf{r}}_b^{t-1}}{\|\hat{\mathbf{r}}_b^{t-1}\|^2}, \tag{7}$$

while the initial value is $\eta_b^t = 1$. The matrix $\hat{\mathbf{B}}_b^{t,0}$ is the initial approximate for the Hessian inverse associated with block $\mathbf{x}_b$. The approximate matrix $\hat{\mathbf{B}}_b^t$ is computed by updating the initial matrix $\hat{\mathbf{B}}_b^{t,0}$ using the last τ pairs of curvature information $\{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1}$. We define the approximate Hessian inverse $\hat{\mathbf{B}}_b^t = \hat{\mathbf{B}}_b^{t,\tau}$ corresponding to block $\mathbf{x}_b$ at step t as the outcome of τ recursive

applications of the update

$$\hat{\mathbf{B}}_b^{t,u+1} = (\hat{\mathbf{Z}}_b^{t-\tau+u})^T \, \hat{\mathbf{B}}_b^{t,u} \, (\hat{\mathbf{Z}}_b^{t-\tau+u}) + \hat{\rho}_b^{t-\tau+u} \, (\mathbf{v}_b^{t-\tau+u}) \, (\mathbf{v}_b^{t-\tau+u})^T, \tag{8}$$

where the matrices $\hat{\mathbf{Z}}_b^{t-\tau+u}$ and the constants $\hat{\rho}_b^{t-\tau+u}$ in (8) for $u = 0, \ldots, \tau - 1$ are defined as

$$\hat{\rho}_b^{t-\tau+u} \;=\; \frac{1}{(\mathbf{v}_b^{t-\tau+u})^T \hat{\mathbf{r}}_b^{t-\tau+u}} \quad \text{and} \quad \hat{\mathbf{Z}}_b^{t-\tau+u} \;=\; \mathbf{I} - \hat{\rho}_b^{t-\tau+u} \hat{\mathbf{r}}_b^{t-\tau+u} (\mathbf{v}_b^{t-\tau+u})^T. \tag{9}$$

The block-wise oLBFGS update defined by (6) - (9) is summarized in Algorithm 2. The computation cost of $\hat{\mathbf{B}}_b^t$ in (8) is in the order of $O(p_b^2)$, however, for the update in (5) the descent direction $\hat{\mathbf{d}}_b^t := \hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$ is required. [41] introduce an efficient implementation of product $\hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$ that requires computation complexity of order $O(\tau p_b)$. We use the same idea for computing the descent direction of ARAPSA for each block – more details are provided below. Therefore, the computation complexity of updating each block for ARAPSA is in the order of $O(\tau p_b)$, while RAPSA requires $O(p_b)$ operations. On the other hand, ARAPSA accelerates the convergence of RAPSA by incorporating the second order information of the objective function for the block updates, as may be observed in the numerical analyses provided in Section F.

For reference, ARAPSA is also summarized in algorithmic form in Algorithm 3. Steps 2 and 3 are devoted to assigning random blocks to the processors. In Step 2 a subset of available blocks $\mathcal{I}^t$ is chosen. These blocks are assigned to different processors in Step 3. In Step 5 processors compute the partial stochastic gradient corresponding to their assigned blocks $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$ using the acquired samples in Step 4. Steps 6 and 7 are devoted to the computation of the ARAPSA descent direction $\hat{\mathbf{d}}_i^t$. In Step 6 the approximate Hessian inverse $\hat{\mathbf{B}}_b^{t,0}$ for block $\mathbf{x}_b$ is initialized as $\hat{\mathbf{B}}_b^{t,0} = \eta_b^t \mathbf{I}$ which is a scaled identity matrix using the expression for η_b^t in (7) for $t > 0$. The initial value of η_b^t is $\eta_b^0 = 1$. In Step 7 we use Algorithm 2 for efficient computation of the descent direction $\hat{\mathbf{d}}_b^t = \hat{\mathbf{B}}_b^t \, \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$. The descent direction $\hat{\mathbf{d}}_b^t$ is used to update the block $\mathbf{x}_b^t$ with stepsize γ^t in Step 8. Step 9 determines the value of the partial stochastic gradient $\nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \boldsymbol{\Theta}_i^t)$ which is required for the computation of stochastic gradient variation $\hat{\mathbf{r}}_b^t$. In Step 10 the variable variation $\mathbf{v}_b^t$ and stochastic gradient variation $\hat{\mathbf{r}}_b^t$ associated with block $\mathbf{x}_b$ are computed to be used in the next iteration.

D Asynchronous Architectures

Up to this point, the RAPSA method dictates that distinct parallel processors select blocks $b_i^t \in \{1, \ldots, B\}$ uniformly at random at each time step t as in Figure 1. However, the requirement that each processor operates on a common time index is burdensome for parallel operations on large computing clusters, as it means that nodes must wait for the processor which has the longest computation time at each step before proceeding. Remarkably, we are able to extend the methods developed in Sections B and C to the case where the parallel processors need not to operate on a common time index (lock-free) and establish that their performance guarantees carry through, so long as the degree of their asynchronicity is bounded in a certain sense. In doing so, we alleviate the computational bottleneck in the parallel architecture, allowing processors to continue processing data as soon as their local task is complete.

Algorithm 3 Accelerated Random Parallel Stochastic Algorithm (ARAPSA)

for $t = 0, 1, 2, \ldots$ **do**
  loop in parallel, processors $i = 1, \ldots, I$ execute:
    Select block b_i^t uniformly at random from set of blocks $\{1, \ldots, B\}$
    Choose a set of realizations $\boldsymbol{\Theta}_i^t$ for the block $\mathbf{x}_b$
    Compute stochastic gradient : $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\theta})$ [cf. (3)]
    Compute the initial Hessian inverse approximation: $\hat{\mathbf{B}}_b^{t,0} = \eta_b^t \mathbf{I}$
    Compute descent direction: $\hat{\mathbf{d}}_b^t = \text{ARAPSA Step} \left(\hat{\mathbf{B}}_b^{t,0},\ \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t),\ \{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1} \right)$
    Update the coordinates of the decision variable $\mathbf{x}_b^{t+1} = \mathbf{x}_b^t - \gamma^t\ \hat{\mathbf{d}}_b^t$
    Compute *updated* stochastic gradient: $\nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \boldsymbol{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \boldsymbol{\theta})$ [cf. (3)]
    Update variations $\mathbf{v}_b^t = \mathbf{x}_b^{t+1} - \mathbf{x}_b^t$ and $\hat{\mathbf{r}}_i^t = \nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+1}, \boldsymbol{\Theta}_i^t) - \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$ [cf.(6)]
  end loop; Transmit updated blocks $i \in \mathcal{I}^t \subset \{1, \ldots, B\}$ to shared memory
end for

Algorithm 4 Asynchronous RAPSA at processor i

while $t < T$ **do**
  Processor $i \in \{1, \ldots, I\}$ **at time index** t **executes the following steps**:
    Select block b_i^t uniformly at random from set of blocks $\{1, \ldots, B\}$
    Choose a set of realizations $\boldsymbol{\Theta}_i^t$ for the block $\mathbf{x}_b$, $b = b_i^t$
    Compute stochastic gradient : $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\theta})$ [cf. (3)]
    Update the coordinates of the decision variable $\mathbf{x}_b^{t+\tau+1} = \mathbf{x}_b^{t+\tau} - \gamma^{t+\tau} \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$
  Send updated parameters $\mathbf{x}_b^{t+1}$ associated with block $b = b_i^t$ to shared memory
  If another processor is also operating on block b_i^t at time t, randomly overwrite
end while

D.1 Asynchronous RAPSA

Consider the case where each node operates asynchronously. In this case, at an instantaneous time index t, only one processor executes an update, as all others are assumed to be busy. If two processors complete their prior task concurrently, then they draw the same time index at the next available slot, in which case the tie is broken at random. Suppose processor i selects block $b_i^t \in \{1, \ldots, B\}$ at time t. Then it grabs the associated component of the decision variable $\mathbf{x}_b^t$ and computes the stochastic gradient $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$ associated with the samples $\boldsymbol{\Theta}_i^t$. This process may take time and during this process other processors may overwrite the variable $\mathbf{x}_b$. Consider the case that the process time of computing stochastic gradient or equivalently the descent direction is τ. Thus, when processor i updates the block b using the evaluated stochastic gradient $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$, it performs the update

$$\mathbf{x}_b^{t+\tau+1} = \mathbf{x}_b^{t+\tau} - \gamma^{t+\tau}\, \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t) \qquad b = b_i^t. \tag{10}$$

Algorithm 5 Asynchronous Accelerated RAPSA at processor i

while $t < T$ **do**
  Processor $i \in \{1, \ldots, I\}$ **at time index** t **executes the following steps**:
    Select block b_i^t uniformly at random from set of blocks $\{1, \ldots, B\}$
    Choose a set of realizations $\boldsymbol{\Theta}_i^t$ for the block $\mathbf{x}_b$, $b = b_i^t$
    Compute stochastic gradient : $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\theta})$ [cf. (3)]
    Compute the initial Hessian inverse approximation: $\hat{\mathbf{B}}_b^{t,0} = \eta_b^t \mathbf{I}$
    Compute descent direction: $\hat{\mathbf{d}}_b^t = \text{ARAPSA Step}\left(\hat{\mathbf{B}}_b^{t,0},\ \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t),\ \{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1}\right)$
    Update the coordinates of the decision variable $\mathbf{x}_b^{t+\tau+1} = \mathbf{x}_b^{t+\tau} - \gamma^{t+\tau}\ \hat{\mathbf{d}}_b^t$
    Compute *updated* stochastic gradient: $\nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+\tau+1}, \boldsymbol{\Theta}_i^t) = \frac{1}{L} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i^t} \nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+\tau+1}, \boldsymbol{\theta})$ [cf. (3)]
    Update variations $\mathbf{v}_b^t = \mathbf{x}_b^{t+\tau+1} - \mathbf{x}_b^t$ and $\hat{\mathbf{r}}_b^t = \nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+\tau+1}, \boldsymbol{\Theta}_i^t) - \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \boldsymbol{\Theta}_i^t)$ [cf.(12)]
    Overwrite the oldest pairs of $\mathbf{v}_b$ and $\hat{\mathbf{r}}_b$ in local memory by $\mathbf{v}_b^t$ and $\hat{\mathbf{r}}_b^t$, respectively.
  Send updated parameters $\mathbf{x}_b^{t+1}$**,** $\{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1}$ **to shared memory.**
  If another processor is operating on block b_i^t, choose to overwrite with probability $1/2$.
end while

Thus, the descent direction evaluated based on the available information at step t is used to update the variable at time $t + \tau$. Asynchronous RAPSA is summarized in Algorithm 4. Note that the delay comes from asynchronous implementation of the algorithm and the fact that other processors are able to modify the variable $\mathbf{x}_b$ during the time that processor i computes its descent direction. We assume the the random time τ that each processor requires to compute its descent direction is bounded above by a constant Δ, i.e., $\tau \leq \Delta$ – see Assumption 4.

Despite the minimal coordination of the asynchronous random parallel stochastic algorithm in (10), we may establish the same performance guarantees as that of RAPSA in Section B. These analytical properties are investigated at length in Section E.

Remark 1 *One may raise the concern that there could be instances that two processors or more work on a same block. Although, this event is not very likely since $I << B$, there is a positive chance that it might happen. This is true since the available processor picks the block that it wants to operate on uniformly at random from the set $\{1, \ldots, B\}$. We show that this event does not cause any issues and the algorithm can eventually converge to the optimal argument even if more than one processor work on a specific block at the same time – see Section E.2. Functionally, this means that if one block is worked on concurrently by two processors, the memory coordination requires that the result of one of the two processors is written to memory with probability $1/2$. This random overwrite rule applies to the case that three or more processors are operating on the same block as well. In this case, the result of one of the conflicting processors is written to memory with probability $1/C$ where C is the number of conflicting processors.*

D.2 Asynchronous ARAPSA

In this section, we study the asynchronous implementation of accelerated RAPSA (ARAPSA). The main difference between the synchronous of implementation ARAPSA in Section C and the asynchronous version is in the update of the variable $\mathbf{x}_b^t$ corresponding to the block b. Consider the case that processor i finishes its previous task at time t, chooses the block $b = b_i^t$, and reads the variable $\mathbf{x}_b^t$. Then, it computes the stochastic gradient $f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ using the set of random variables $\mathbf{\Theta}_i^t$. Further, processor i computes the descent direction $\hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ using the last τ sets of curvature information $\{\mathbf{v}_b^u, \hat{\mathbf{r}}_b^u\}_{u=t-\tau}^{t-1}$ as shown in Algorithm 1. If we assume that the required time to compute the descent direction $\hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ is τ', processor i updates the variable $\mathbf{x}_b^{t+\tau'}$ as

$$\mathbf{x}_b^{t+\tau'+1} = \mathbf{x}_b^{t+\tau'} - \gamma^{t+\tau'} \hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t) \qquad b = b_i^t. \tag{11}$$

Note that the update in (11) is different from the synchronous version in (5) in the time index of the variable that is updated using the available information at time t. In other words, in the synchronous implementation the descent direction $\hat{\mathbf{B}}_b^t \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ is used to update the variable $\mathbf{x}_b^t$ with the same time index, while this descent direction is executed to update the variable $\mathbf{x}_b^{t+\tau'}$ in asynchronous ARAPSA.

Note that the definitions of the variable variation $\mathbf{v}_b^t$ and the stochastic gradient variation $\hat{\mathbf{r}}_b^t$ are different in asynchronous setting and they are given by

$$\mathbf{v}_b^t \;=\; \mathbf{x}_b^{t+\tau'+1} - \mathbf{x}_b^t, \qquad \hat{\mathbf{r}}_b^t \;=\; \nabla_{\mathbf{x}_b} f(\mathbf{x}^{t+\tau'+1}, \mathbf{\Theta}_i^t) - \nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t). \tag{12}$$

This modification comes from the fact that the stochastic gradient $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ is already evaluated for the descent direction in (11). Thus, we define the stochastic gradient variation by computing the difference of the stochastic gradient $\nabla_{\mathbf{x}_b} f(\mathbf{x}^t, \mathbf{\Theta}_i^t)$ and the stochastic gradient associated with the same random set $\mathbf{\Theta}_i^t$ evaluated at the most recent iterate which is $\mathbf{x}_b^{t+\tau'+1}$. Likewise, the variable variation is redefined as the difference $\mathbf{x}_b^{t+\tau'+1} - \mathbf{x}_b^t$. The steps of asynchronous ARAPSA are summarized in Algorithm 5.

E Convergence Analysis

We show in this section that the sequence of objective function values $F(\mathbf{x}^t)$ generated by RAPSA approaches the optimal objective function value $F(\mathbf{x}^*)$. We further show that the convergence guarantees for synchronous RAPSA generalize to the asynchronous setting. In establishing this result we define the set $\mathcal{S}^t$ corresponding to the components of the vector $\mathbf{x}$ associated with the blocks selected at step t defined by indexing set $\mathcal{I}^t \subset \{1, \ldots, B\}$. Note that components of the set $\mathcal{S}^t$ are chosen uniformly at random from the set of blocks $\{\mathbf{x}_1, \ldots, \mathbf{x}_B\}$. With this definition, due to convenience for analyzing the proposed methods, we rewrite the time evolution of the RAPSA iterates (Algorithm 1) as

$$\mathbf{x}_i^{t+1} \;=\; \mathbf{x}_i^t - \gamma^t \nabla_{\mathbf{x}_i} f(\mathbf{x}^t, \mathbf{\Theta}_i^t) \qquad \text{for all } \mathbf{x}_i \in \mathcal{S}^t, \tag{13}$$

while the rest of the blocks remain unchanged, i.e., $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t$ for $\mathbf{x}_i \notin \mathcal{S}^t$. Since the number of updated blocks is equal to the number of processors, the ratio of updated blocks is $r := |\mathcal{I}^t|/B = I/B$. To prove convergence of RAPSA, we require the following assumptions.

Assumption 1 *The instantaneous objective functions $f(\mathbf{x}, \boldsymbol{\theta})$ are differentiable and the average function $F(\mathbf{x})$ is strongly convex with parameter $m > 0$.*

Assumption 2 *The average objective function gradients $\nabla F(\mathbf{x})$ are Lipschitz continuous with respect to the Euclidian norm with parameter M, i.e., for all $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^p$, it holds that*

$$\|\nabla F(\mathbf{x}) - \nabla F(\hat{\mathbf{x}})\| \ \leq \ M \ \|\mathbf{x} - \hat{\mathbf{x}}\|. \tag{14}$$

Assumption 3 The second moment of the norm of the stochastic gradient is bounded for all $\mathbf{x}$, i.e., there exists a constant K such that for all variables $\mathbf{x}$, it holds

$$\mathbb{E}_{\boldsymbol{\theta}}\big[\|\nabla f(\mathbf{x}^t, \boldsymbol{\theta}^t)\|^2 \,\big|\, \mathbf{x}^t\big] \leq K. \tag{15}$$

Notice that Assumption 1 only enforces strong convexity of the average function F, while the instantaneous functions f_i may not be even convex. Further, notice that since the instantaneous functions f_i are differentiable the average function F is also differentiable. The Lipschitz continuity of the average function gradients ∇F is customary in proving objective function convergence for descent algorithms. The restriction imposed by Assumption 3 is a standard condition in stochastic approximation literature [24], its intent being to limit the variance of the stochastic gradients [42].

E.1 Convergence of RAPSA

We turn our attention to the random parallel stochastic algorithm defined in (3)-(4) in Section B, establishing performances guarantees in both the diminishing and constant algorithm step-size regimes. Our first result comes in the form of a expected descent lemma that relates the expected difference of subsequent iterates to the gradient of the average function.

Lemma 1 *Consider the random parallel stochastic algorithm defined in* (3)-(4)*. Recall the definitions of the set of updated blocks $\mathcal{I}^t$ which are randomly chosen from the total B blocks. Define $\mathcal{F}^t$ as a sigma algebra that measures the history of the system up until time t. Then, the expected value of the difference $\mathbf{x}^{t+1} - \mathbf{x}^t$ with respect to the random set $\mathcal{I}^t$ given $\mathcal{F}^t$ is*

$$\mathbb{E}_{\mathcal{I}^t}\big[\mathbf{x}^{t+1} - \mathbf{x}^t \mid \mathcal{F}^t\big] = -r\gamma^t \ \nabla f(\mathbf{x}^t, \boldsymbol{\Theta}^t). \tag{16}$$

Moreover, the expected value of the squared norm $\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$ with respect to the random set $\mathcal{S}^t$ given $\mathcal{F}^t$ can be simplified as

$$\mathbb{E}_{\mathcal{I}^t}\big[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \mid \mathcal{F}^t\big] = r(\gamma^t)^2 \ \left\|\nabla f(\mathbf{x}^t, \boldsymbol{\Theta}^t)\right\|^2. \tag{17}$$

Proof: See Appendix A.1. ■

Notice that in the regular stochastic gradient descent method the difference of two consecutive iterates $\mathbf{x}^{t+1} - \mathbf{x}^t$ is equal to the stochastic gradient $\nabla f(\mathbf{x}^t, \boldsymbol{\Theta}^t)$ times the stepsize γ^t. Based on the first result in Lemma 1, the expected value of stochastic gradients with respect to the random set of blocks $\mathcal{I}^t$ is the same as the one for SGD except that it is multiplied by the fraction of updated blocks r. Expression in (17) shows the same relation for the expected value of the squared difference $\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$. These relationships confirm that in expectation RAPSA behaves as SGD which allows us to establish the global convergence of RAPSA.

Proposition 1 *Consider the random parallel stochastic algorithm defined in* (3)-(4)*. If Assumptions 1-3 hold, then the objective function error sequence $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ satisfies*

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*) \mid \mathcal{F}^t\right] \leq \left(1 - 2mr\gamma^t\right)\left(F(\mathbf{x}^t) - F(\mathbf{x}^*)\right) + \frac{rMK(\gamma^t)^2}{2}. \tag{18}$$

Proof: See Appendix A.2. ■

Proposition 1 leads to a supermartingale relationship for the sequence of objective function errors $F(\mathbf{x}^t) - F(\mathbf{x}^*)$. In the following theorem we show that if the sequence of stepsize satisfies standard stochastic approximation diminishing step-size rules (non-summable and squared summable), the sequence of objective function errors $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ converges to null almost surely. Considering the strong convexity assumption this result implies almost sure convergence of the sequence $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ to null.

Theorem 1 *Consider the random parallel stochastic algorithm defined in* (3)-(4) *(Algorithm 1). If Assumptions 1-3 hold true and the sequence of stepsizes are non-summable* $\sum_{t=0}^{\infty} \gamma^t = \infty$ *and square summable* $\sum_{t=0}^{\infty} (\gamma^t)^2 < \infty$*, then sequence of the variables* $\mathbf{x}^t$ *generated by RAPSA converges almost surely to the optimal argument* $\mathbf{x}^*$*,*

$$\lim_{t \to \infty} \|\mathbf{x}^t - \mathbf{x}^*\|^2 \;=\; 0 \qquad a.s. \tag{19}$$

Moreover, if stepsize is defined as $\gamma^t := \gamma^0 T^0/(t + T^0)$ *and the stepsize parameters are chosen such that* $2mr\gamma^0 T^0 > 1$*, then the expected average function error* $\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right]$ *converges to null at least with a sublinear convergence rate of order* $\mathcal{O}(1/t)$*,*

$$\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] \;\leq\; \frac{C}{t + T^0}, \tag{20}$$

where the constant C *is defined as*

$$C = \max\left\{\frac{rMK(\gamma^0 T^0)^2}{4mr\gamma^0 T^0 - 2},\; T^0(F(\mathbf{x}^0) - F(\mathbf{x}^*))\right\}. \tag{21}$$

Proof: See Appendix A.3. ■

The result in Theorem 1 shows that when the sequence of stepsize is diminishing as $\gamma^t = \gamma^0 T^0/(t + T^0)$, the average objective function value $F(\mathbf{x}^t)$ sequence converges to the optimal objective value $F(\mathbf{x}^*)$ with probability 1. Further, the rate of convergence in expectation is at least in the order of $\mathcal{O}(1/t)$. [2] Diminishing stepsizes are useful when exact convergence is required, however, for the case that we are interested in a specific accuracy ϵ the more efficient choice is using a constant stepsize. In the following theorem we study the convergence properties of RAPSA for a constant stepsize $\gamma^t = \gamma$.

Theorem 2 *Consider the random parallel stochastic algorithm defined in* (3)-(4) *(Algorithm 1). If Assumptions 1-3 hold true and the stepsize is constant* $\gamma^t = \gamma$*, then a subsequence of the variables* $\mathbf{x}^t$ *generated by RAPSA converges almost surely to a neighborhood of the optimal argument* $\mathbf{x}^*$ *as*

$$\liminf_{t \to \infty} \; F(\mathbf{x}^t) - F(\mathbf{x}^*) \;\leq\; \frac{\gamma MK}{4m} \qquad a.s. \tag{22}$$

Moreover, if the constant stepsize γ *is chosen such that* $2mr\gamma < 1$ *then the expected average function value error* $\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right]$ *converges linearly to an error bound as*

$$\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] \leq (1 - 2m\gamma r)^t \left(F(\mathbf{x}^0) - F(\mathbf{x}^*)\right) + \frac{\gamma MK}{4m}. \tag{23}$$

[2] The expectation on the left hand side of (32), and throughout the subsequent convergence rate analysis, is taken with respect to the full algorithm history $\mathcal{F}_0$, which all realizations of both $\mathbf{\Theta}_t$ and $\mathcal{I}_t$ for all $t \geq 0$.

Proof: See Appendix A.4. ■

Notice that according to the result in (23) there exits a trade-off between accuracy and speed of convergence. Decreasing the constant stepsize γ leads to a smaller error bound $\gamma MK/4m$ and a more accurate convergence, while the linear convergence constant $(1-2m\gamma r)$ increases and the convergence rate becomes slower. Further, note that the error of convergence $\gamma MK/4m$ is independent of the ratio of updated blocks r, while the constant of linear convergence $1-2m\gamma r$ depends on r. Therefore, updating a fraction of the blocks at each iteration decreases the speed of convergence for RAPSA relative to SGD that updates all of the blocks, however, both of the algorithms reach the same accuracy.

To achieve accuracy ϵ the sum of two terms in the right hand side of (23) should be smaller than ϵ. Let's consider ϕ as a positive constant that is strictly smaller than 1, i.e., $0<\phi<1$. Then, we want to have

$$\frac{\gamma MK}{4m} \leq \phi\epsilon, \quad (1-2m\gamma r)^t \left(F(\mathbf{x}^0)-F(\mathbf{x}^*)\right) \leq (1-\phi)\epsilon. \tag{24}$$

Therefore, to satisfy the first condition in (24) we set the stepsize as $\gamma = 4m\phi\epsilon/MK$. Apply this substitution into the second inequality in (24) and consider the inequality $a+\ln(1-a)<0$ for $0<a<1$, to obtain that

$$t \geq \frac{MK}{8m^2 r\phi\epsilon} \ln\left(\frac{F(\mathbf{x}^0)-F(\mathbf{x}^*)}{(1-\phi)\epsilon}\right). \tag{25}$$

The lower bound in (25) shows the minimum number of required iterations for RAPSA to achieve accuracy ϵ.

E.2 Convergence of Asynchronous RAPSA

In this section, we study the convergence of Asynchronous RAPSA (Algorithm 4) developed in Section D and we characterize the effect of delay in the asynchronous implementation. To do so, the following condition on the delay τ is required.

Assumption 4 *The random variable τ which is the delay between reading and writing for processors does not exceed the constant Δ, i.e.,*

$$\tau \leq \Delta. \tag{26}$$

The condition in Assumption 4 implies that processors can finish their tasks in a time that is bounded by the constant Δ. This assumption is typical in the analysis of asynchronous algorithms.

To establish the convergence properties of asynchronous RAPSA recall the set $\mathcal{S}^t$ containing the blocks that are updated at step t with associated indices $\mathcal{I}^t \subset \{1,\ldots,B\}$. Therefore, the update of asynchronous RAPSA can be written as

$$\mathbf{x}_i^{t+1} \;=\; \mathbf{x}_i^t - \gamma^t \, \nabla_{\mathbf{x}_i} f(\mathbf{x}^{t-\tau}, \mathbf{\Theta}_i^{t-\tau}) \qquad \text{for all } \mathbf{x}_i \in \mathcal{S}^t, \tag{27}$$

and the rest of the blocks remain unchanged, i.e., $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t$ for $\mathbf{x}_i \notin \mathcal{S}^t$.

Note that the random set $\mathcal{I}^t$ and the associated block set $\mathcal{S}^t$ are chosen at time $t-\tau$ in practice; however, for the sake of analysis we can assume that these sets are chosen at time t. In other words, we can assume that at step $t-\tau$ processor i computes the full (for all blocks) stochastic gradient $\nabla f(\mathbf{x}^{t-\tau}, \mathbf{\Theta}_i^{t-\tau})$ and after finishing this task at time t, it chooses uniformly at random the block that it wants to update. Thus, the block $\mathbf{x}_i$ in (27) is chosen at step t. This new interpretation of the

update of asynchronous RAPSA is only important for the convergence analysis of the algorithm and we use it in the proof of following lemma which is similar to the result in Lemma 1 for synchronous RAPSA.

Lemma 2 *Consider the asynchronous random parallel stochastic algorithm (Algorithm 4) defined in* (10)*. Recall the definitions of the set of updated blocks $\mathcal{I}^t$ which are randomly chosen from the total B blocks. Define $\mathcal{F}^t$ as a sigma algebra that measures the history of the system up until time t. Then, the expected value of the difference $\mathbf{x}^{t+1}-\mathbf{x}^t$ with respect to the random set $\mathcal{I}^t$ given $\mathcal{F}^t$ is*

$$\mathbb{E}_{\mathcal{I}^t}\left[\mathbf{x}^{t+1}-\mathbf{x}^t \mid \mathcal{F}^t\right] = -\frac{\gamma^t}{B}\ \nabla f(\mathbf{x}^{t-\tau}, \mathbf{\Theta}^{t-\tau}). \tag{28}$$

Moreover, the expected value of the squared norm $\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2$ with respect to the random set $\mathcal{S}^t$ given $\mathcal{F}^t$ satisfies the identity

$$\mathbb{E}_{\mathcal{I}^t}\left[\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2 \mid \mathcal{F}^t\right] = \frac{(\gamma^t)^2}{B}\ \left\|\nabla f(\mathbf{x}^{t-\tau}, \mathbf{\Theta}^{t-\tau})\right\|^2. \tag{29}$$

Proof: See Appendex B.1. ■

The results in Lemma 2 is a natural extension of the results in Lemma 1 for the lock-free setting, since in the asynchronous scheme only one of the blocks is updated at each iteration and the ratio r can be simplified as $1/B$. We use the result in Lemma 2 to characterize the decrement in the expected sub-optimality in the following proposition.

Proposition 2 *Consider the asynchronous random parallel stochastic algorithm defined in* (10) *(Algorithm 4) . If Assumptions 1-3 hold, then the objective function error sequence $F(\mathbf{x}^t)-F(\mathbf{x}^*)$ satisfies*

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{x}^{t+1})-F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right]\\ &\quad\leq \left(1-\frac{2m\gamma^t}{B}\left[1-\frac{\rho M}{2}\right]\right)\mathbb{E}\left[F(\mathbf{x}^t)-F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right]+\frac{MK(\gamma^t)^2}{2B}+\frac{\tau^2 MK\gamma^t(\gamma^{t-\tau})^2}{2\rho B^2}.\end{aligned} \tag{30}$$

Proof: See Appendix B.2. ■

We proceed to use the result in Proposition 2 to prove that the sequence of iterates generated by asynchronous RAPSA converges to the optimal argument $\mathbf{x}^*$ defined by (2).

Theorem 3 *Consider the asynchronous RAPSA defined in* (10) *(Algorithm 4) . If Assumptions 1-3 hold true and the sequence of stepsizes are non-summable $\sum_{t=0}^{\infty}\gamma^t=\infty$ and square summable $\sum_{t=0}^{\infty}(\gamma^t)^2<\infty$, then sequence of the variables $\mathbf{x}^t$ generated by RAPSA converges almost surely to the optimal argument $\mathbf{x}^*$,*

$$\liminf_{t\to\infty}\|\mathbf{x}^t-\mathbf{x}^*\|^2 \ = \ 0 \qquad a.s. \tag{31}$$

Moreover, if stepsize is defined as $\gamma^t:=\gamma^0T^0/(t+T^0)$ and the stepsize parameters are chosen such that $2mr\gamma^0T^0>1$, then the expected average function error $\mathbb{E}\left[F(\mathbf{x}^t)-F(\mathbf{x}^*)\right]$ converges to null at least with a sublinear convergence rate of order $\mathcal{O}(1/t)$,

$$\mathbb{E}\left[F(\mathbf{x}^t)-F(\mathbf{x}^*)\right] \ \leq \ \frac{C}{t+T^0}, \tag{32}$$

where the constant C is defined as

$$C = \max\left\{\frac{rMK(\gamma^0 T^0)^2}{4mr\gamma^0 T^0 - 2},\ T^0(F(\mathbf{x}^0) - F(\mathbf{x}^*))\right\}. \tag{33}$$

Proof: See Appendix B.3. ■

Theorem 3 establishes that the RAPSA algorithm when run on a lock-free computing architecture, still yields convergence to the optimal argument $\mathbf{x}^*$ defined by (2). Moreover, the expected objective error sequence converges to null as $\mathcal{O}(1/t)$. These results, which correspond to the diminishing step-size regime, are comparable to the performance guarantees (Theorem 1) previously established for RAPSA on a synchronous computing cluster, meaning that the algorithm performance does not degrade significantly when implemented on an asynchronous system. This issue is explored numerically in Section F.

F Numerical analysis

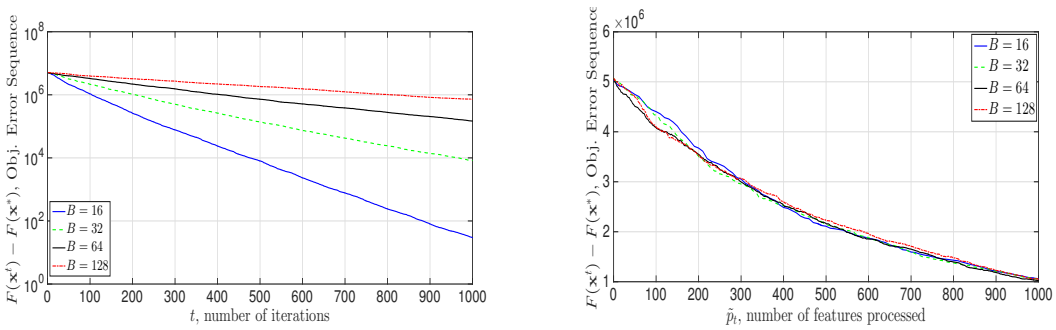
In this section we study the numerical performance of the doubly stochastic approximation algorithms developed in Sections B-D by first considering a linear regression problem. We then use RAPSA to develop a visual classifier to distinguish between distinct hand-written digits.

F.1 Linear Regression

We consider a setting in which observations $\mathbf{z}_n \in \mathbb{R}^q$ are collected which are noisy linear transformations $\mathbf{z}_n = \mathbf{H}_n\mathbf{x} + \mathbf{w}_n$ of a signal $\mathbf{x} \in \mathbb{R}^p$ which we would like to estimate, and $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_q)$ is a Gaussian random variable. For a finite set of samples N, the optimal $\mathbf{x}^*$ is computed as the least squares estimate $\mathbf{x}^* := \text{argmin}_{\mathbf{x}\in\mathbb{R}^p}(1/N)\sum_{n=1}^N \|\mathbf{H}_n\mathbf{x} - \mathbf{z}_n\|^2$. We run RAPSA on LMMSE estimation problem instances where $q = 1$, $p = 1024$, and $N = 10^4$ samples are given. The observation matrices $\mathbf{H}_n \in \mathbb{R}^{q\times p}$, when stacked over all n (an $N \times p$ matrix), are generated from a matrix normal distribution whose mean is a tri-diagonal matrix. The main diagonal is 2, while the super and sub-diagonals are all set to $-1/2$. Moreover, the true signal has entries chosen uniformly at random from the fractions $\mathbf{x} \in \{1, \ldots, p\}/p$. Additionally, the noise variance perturbing the observations is set to $\sigma^2 = 10^{-2}$. We assume that the number of processors $I = 16$ is fixed and each processor is in charge of 1 block. We consider different number of blocks $B = \{16, 32, 64, 128\}$. Note that when the number of blocks is B, there are $p/B = 1024/B$ coordinates in each block.

Results for RAPSA We first consider the algorithm performance of RAPSA (Algorithm 1) when using a constant step-size $\gamma^t = \gamma = 10^{-2}$. The size of mini-batch is set as $L = 10$ in the subsequent experiments. To determine the advantages of incomplete randomized parallel processing, we vary the number of coordinates updated at each iteration. In the case that $B = 16$, $B = 32$, $B = 64$, and $B = 128$, in which case the number of updated coordinates per iteration are 1024, 512, 256, and 128, respectively. Notice that the case that $B = 16$ can be interpreted as parallel SGD, which is mathematically equivalent to Hogwild! [34], since all the coordinates are updated per iteration, while in other cases $B > 16$ only a subset of 1024 coordinates are updated.

Fig. 2(a) illustrates the convergence path of RAPSA's objective error sequence defined as $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ with $F(\mathbf{x}) = (1/N)\sum_{n=1}^N \|\mathbf{H}_n\mathbf{x} - \mathbf{z}_n\|^2$ as compared with the number of iterations t. In terms of iteration t, we observe that the algorithm performance is best when the number of processors equals the number of blocks, corresponding to parallelized stochastic gradient method.



(a) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t

(b) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. feature $\tilde{p}_t$

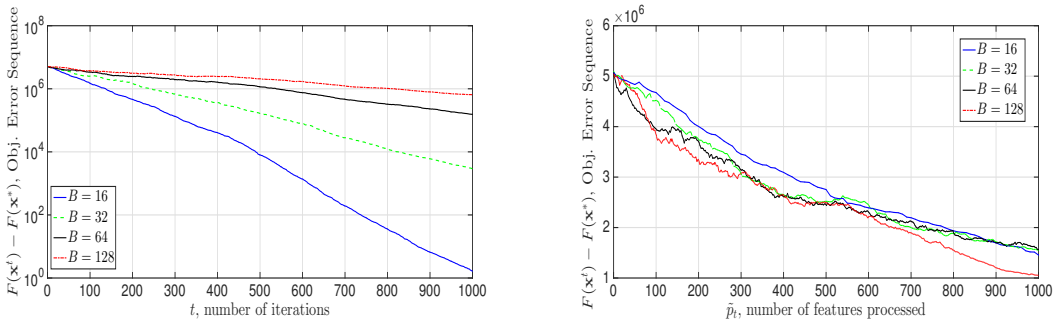
Figure 2: RAPSA on a linear regression (quadratic minimization) problem with signal dimension $p = 1024$ for $N = 10^3$ iterations with mini-batch size $L = 10$ for different number of blocks $B = \{16, 32, 64, 128\}$ initialized as $10^4 \times \mathbf{1}$. We use constant step-size $\gamma^t = \gamma = 10^{-2}$. Convergence is in terms of number of iterations is best when the number of blocks updated per iteration is equal to the number of processors ($B = 16$, corresponding to parallelized SGD), but comparable across the different cases in terms of number of features processed. This shows that there is no price payed in terms of convergence speed for reducing the computation complexity per iteration.

However, comparing algorithm performance over iteration t across varying numbers of blocks updates is unfair. If RAPSA is run on a problem for which $B = 32$, then at iteration t it has only processed *half* the data that parallel SGD, i.e., $B = 16$, has processed by the same iteration. Thus for completeness we also consider the algorithm performance in terms of number of features processed $\tilde{p}_t$ which is given by $\tilde{p}_t = ptI/B$.

In Fig. 2(b), we display the convergence of the excess mean square error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ in terms of number of features processed $\tilde{p}_t$. In doing so, we may clearly observe the advantages of updating fewer features/coordinates per iteration. Specifically, the different algorithms converge in a nearly identical manner, but RAPSA with $I << B$ may be implemented without any complexity bottleneck in the dimension of the decision variable p (also the dimension of the feature space).

We observe a comparable trend when we run RAPSA with a hybrid step-size scheme $\gamma^t = \min(\epsilon, \epsilon\tilde{T}_0/t)$ which is a constant $\epsilon = 10^{-1.5}$ for the first $\tilde{T}_0 = 400$ iterations, after which it diminishes as $O(1/t)$. We again observe in Figure 3(a) that convergence is fastest in terms of excess mean square error versus iteration t when all blocks are updated at each step. However, for this step-size selection, we see that updating fewer blocks per step is *faster* in terms of number of features processed. This result shows that updating fewer coordinates per iteration yields convergence gains in terms of number of features processed. This advantage comes from the advantage of Gauss-Seidel style block selection schemes in block coordinate methods as compared with Jacobi schemes. In particular, it's well understood that for problems settings with specific conditioning, cyclic block updates are superior to parallel schemes, and one may respectively interpret RAPSA as compared to parallel SGD as executing variants of cyclic or parallel block selection schemes. We note that the magnitude of this gain is dependent on the condition number of the Hessian of the expected objective $F(\mathbf{x})$.

Results for Accelerated RAPSA We now study the benefits of incorporating approximate second-order information about the objective $F(\mathbf{x})$ into the algorithm in the form of ARAPSA (Algorithm 3). We first run ARAPSA for the linear regression problem outlined above when using a constant step-size $\gamma^t = \gamma = 10^{-2}$ with fixed mini-batch size $L = 10$. Moreover, we again vary the



(a) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t

(b) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. feature $\tilde{p}_t$

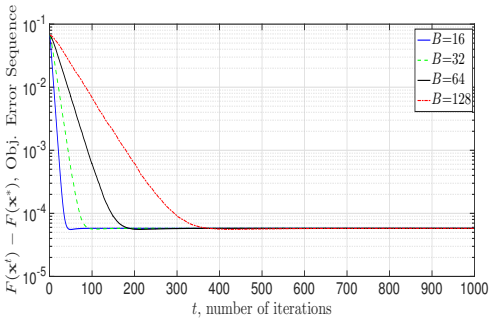
Figure 3: RAPSA on a linear regression problem with signal dimension $p = 1024$ for $N = 10^3$ iterations with mini-batch size $L = 10$ for different number of blocks $B = \{16, 32, 64, 128\}$ using initialization $\mathbf{x}_0 = 10^4 \times \mathbf{1}$. We use hybrid step-size $\gamma^t = \min(10^{-1.5}, 10^{-1.5}\tilde{T}_0/t)$ with annealing rate $\tilde{T}_0 = 400$. Convergence is faster with smaller B which corresponds to the proportion of blocks updated per iteration r closer to 1 in terms of number of iterations. Contrarily, in terms of number of features processed $B = 128$ has the best performance and $B = 16$ has the worst performance. This shows that updating less features/coordinates per iterations can lead to faster convergence in terms of number of processed features.

number of blocks as $B = 16$, $B = 32$, $B = 64$, and $B = 128$, corresponding to updating all, half, one-quarter, and one-eighth of the elements of vector $\mathbf{x}$ per iteration, respectively.

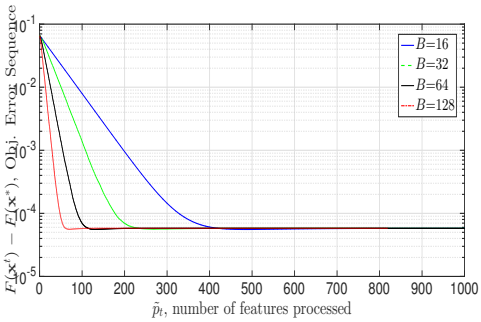
Fig. 4(a) displays the convergence path of ARAPSA's excess mean-square error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ versus the number of iterations t. We observe that parallelized oL-BFGS ($I = B$) converges fastest in terms of iteration index t. On the contrary, in Figure 4(b), we may clearly observe that larger B, which corresponds to using *fewer* elements of $\mathbf{x}$ per step, converges faster in terms of number of features processed. The Gauss-Seidel effect is more substantial for ARAPSA as compared with RAPSA due to the fact that the argmin of the instantaneous objective computed in block coordinate descent is better approximated by its second-order Taylor-expansion (ARAPSA, Algorithm 3) as compared with its linearization (RAPSA, Algorithm 1).

We now consider the performance of ARAPSA when a hybrid algorithm step-size is used, i.e. $\gamma^t = \min(10^{-1.5}, 10^{-1.5}\tilde{T}_0/t)$ with attenuation threshold $\tilde{T}_0 = 400$. The results of this numerical experiment are given in Figure 5. We observe that the performance gains of ARAPSA as compared to parallelized oL-BFGS apparent in the constant step-size scheme are more substantial in the hybrid setting. That is, in Figure 5(a) we again see that parallelized oL-BFGS is best in terms of iteration index t – to achieve the benchmark $F(\mathbf{x}^t) - F(\mathbf{x}^*) \leq 10^{-4}$, the algorithm requires $t = 100$, $t = 221$, $t = 412$, and $t > 1000$ iterations for $B = 16$, $B = 32$, $B = 64$, and $B = 128$, respectively. However, in terms of $\tilde{p}_t$, the number of elements of $\mathbf{x}$ processed, to reach the benchmark $F(\mathbf{x}^t) - F(\mathbf{x}^*) \leq 0.1$, we require $\tilde{p}_t > 1000$, $\tilde{p}_t = 570$, $\tilde{p}_t = 281$, and $\tilde{p}_t = 203$, respectively, for $B = 16$, $B = 32$, $B = 64$, and $B = 128$.

Comparison of RAPSA and ARAPSA We turn to numerically analyzing the performance of Accelerated RAPSA and RAPSA on the linear estimation problem for the case that parameter vectors $\mathbf{x} \in \mathbb{R}^p$ are $p = 500$ dimensional for $N = 10^4$ iterations in the constant step-size case $\gamma = 10^{-2}$. Both algorithms are initialized as $\mathbf{x}_0 = 10^3 \times \mathbf{1}$ with mini-batch size $L = 10$, and ARAPSA uses the curvature memory level $\tau = 10$. The number of processors is fixed again as $I = 16$, and the number of blocks is $B = 64$, meaning that $r = 1/4$ of the elements of $\mathbf{x}$ are

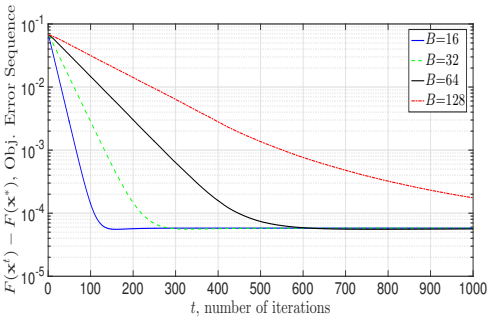


(a) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t

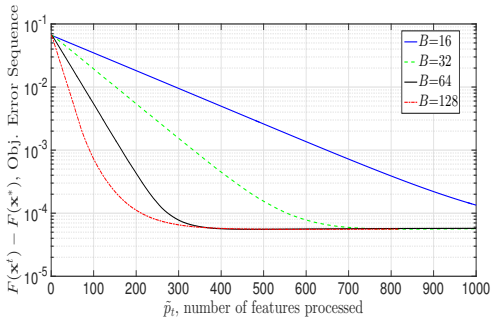


(b) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. feature $\tilde{p}_t$

Figure 4: ARAPSA on a linear regression problem with signal dimension $p = 1024$ for $N = 10^3$ iterations with mini-batch size $L = 10$ for different number of blocks $B = \{16, 32, 64, 128\}$. We use constant step-size $\gamma^t = \gamma = 10^{-1}$ using initialization $10^4 \times \mathbf{1}$. Convergence is comparable across the different cases in terms of number of iterations, but in terms of number of features processed $B = 128$ has the best performance and $B = 16$ (corresponding to parallelized oL-BFGS) converges slowest. We observe that using fewer coordinates per iterations leads to faster convergence in terms of number of processed elements of $\mathbf{x}$.

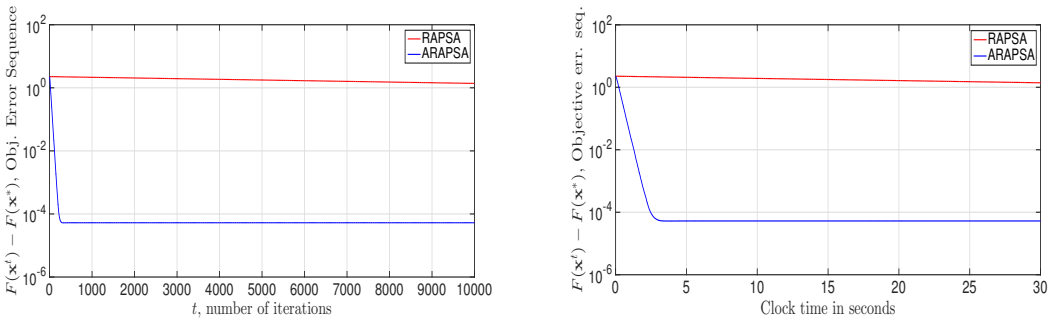


(a) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t



(b) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. feature $\tilde{p}_t$

Figure 5: ARAPSA on a linear regression problem with signal dimension $p = 1024$ for $N = 10^4$ iterations with mini-batch size $L = 10$ for different number of blocks $B = \{16, 32, 64, 128\}$. We use hybrid step-size $\gamma^t = \min(10^{-1.5}, 10^{-1.5}\tilde{T}_0/t)$ with annealing rate $\tilde{T}_0 = 400$. Convergence is comparable across the different cases in terms of number of iterations, but in terms of number of features processed $B = 128$ has the best performance and $B = 16$ has the worst performance. This shows that updating less features/coordinates per iterations leads to faster convergence in terms of number of processed features.



(a) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t

(b) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. clock time (s)

Figure 6: A numerical comparison of RAPSA and ARAPSA on the linear estimation problem stated at the beginning of Section F.1 for $N = 10^4$ iterations with signal dimension $p = 500$ with constant step-size $\gamma = 10^{-2}$ when there are $I = 16$ processors and $B = 64$ blocks, meaning that one quarter of the elements of $\mathbf{x}$ are updated per iteration. Observe that the rate of convergence for ARAPSA is empirically orders of magnitude higher than RAPSA.

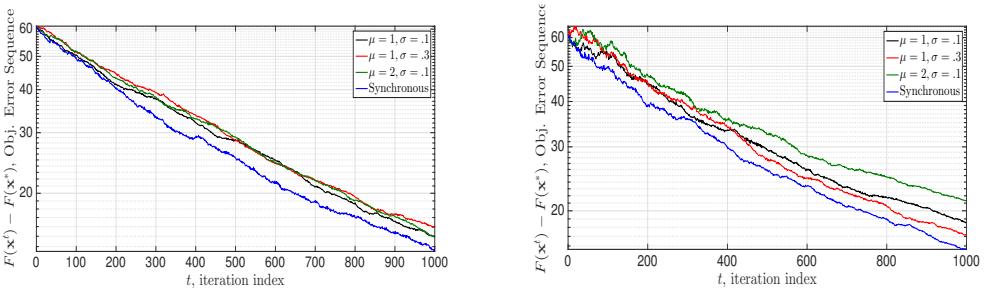
operated on at each iteration.

The results of this numerical evaluation are given in Figure 6. We plot the objective error sequence versus iteration t in Figure 6(a). Observe that ARAPSA converges to within 10^{-4} of the optimum by $t = 300$ iterations in terms of $F(\mathbf{x}^t) - F(\mathbf{x}^*)$, whereas RAPSA, while descending slowly, approaches within 10 of the optimum by $t = 10^4$ iterations. The performance advantages of ARAPSA as compared to RAPSA are also apparent in Figure 6(b), which readjusts the results of Figure 6(a) to be in terms of *actual* elapsed time. We see that despite the higher complexity of ARAPSA per iteration, its empirical performance results in extremely fast convergence on linear estimation problems. That is, in about 3 seconds, the algorithm converges to within 10^{-4} of the optimal estimator in terms of objective function evaluation.

Results for Asynchronous RAPSA We turn to studying the empirical performance of the asynchronous variant of RAPSA (Algorithm 4) proposed in Section D.1. The model we use for asynchronicity is modeled after a random delay phenomenon in physical communication systems in which each local server has a distinct clock which is not locked to the others. Each processor's clock begins at time $t_0^i = t_0$ for all processors $i = 1, \ldots, I$ and selects subsequent times as $t_k = t_{k-1} + w_k^i$, where $w_k^i \sim \mathcal{N}(\mu, \sigma^2)$ is a normal random variable with mean μ and variance σ^2. The variance in this model effectively controls the amount of variability between the clocks of distinct processors.

We run Asynchronous RAPSA for the linear estimation problem when the parameter vector $\mathbf{x}$ is $p = 500$ dimensional for $N = 10^3$ iterations with no mini-batching $L = 1$ for both the case that the algorithm step-size is diminishing and constant step-size regimes for the case that the noise distribution perturbing the collected observations has variance $\sigma^2 = 10^{-2}$, and the observation matrix is as discussed at the outset of Section F.1. Further, the algorithm is initialized as $\mathbf{x}_0 = 10^3\mathbf{1}$. We run the algorithm for a few different instantiations of asynchronicity, that is, $w_k^i \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 1$ or $\mu = 2$, and $\sigma = .1$ or $\sigma = .3$.

The results of this numerical experiment are given in Figure 7 for both the constant and diminishing step-size schemes. We see that the performance of the asynchronous parallel scheme is comparable across different levels of variability among the local clocks of each processor. In particular, in Figure 7(a) which corresponds to the case where the algorithm is run with constant step-size



(a) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t.

(b) Excess Error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t.

Figure 7: Asynchronous RAPSA (Algorithm 4) on the linear estimation problem in the constant ($\gamma = 10^4$, left) and diminishing ($\gamma_t = 10^6/(t + 250)$, right) step-size schemes with no mini-batching $L = 1$ for a binary training subset of size $N = 10^3$ with no regularization $\lambda = 0$ when the algorithm is initialized as $\mathbf{x}_0 = 10^3 \times \mathbf{1}$. Varying the asynchronicity distribution has little effect, but we find that convergence behavior is slower than its synchronized counterpart, as expected.

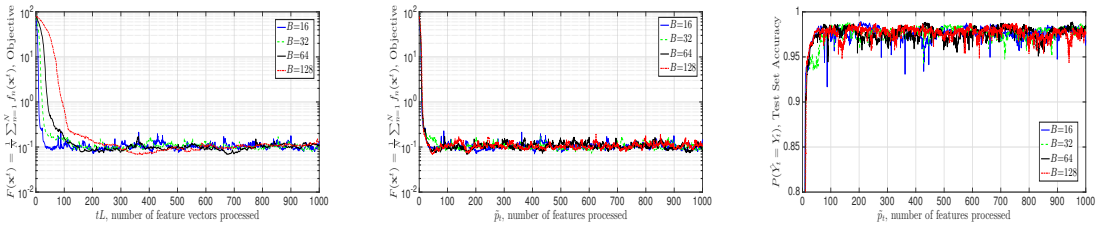
$\gamma = 10^{-2}$, we observe comparable performance in terms of the objective function error sequence $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ with iteration t – across the varying levels of asynchrony we have $F(\mathbf{x}^t) - F(\mathbf{x}^*) \leq 10$ by $t = 10^3$. This trend may also be observed in the diminishing step-size scheme $\gamma^t = 1/t$ which is given in Figure 7(b). That is, the distance to the optimal objective is nearly identical across differing levels of asynchronicity. In both cases, the synchronized algorithm performs better than its asynchronous counterpart.

F.2 Hand-Written Digit Recognition

We now make use of RAPSA for visual classification of written digits. To do so, let $\mathbf{z} \in \mathbb{R}^p$ be a feature vector encoding pixel intensities (elements of the unit interval $[0, 1]$ with smaller values being closer to black) of an image and let $y \in \{-1, 1\}$ be an indicator variable of whether the image contains the digit 0 or 8, in which case the binary indicator is respectively $y = -1$ or $y = 1$. We model the task of learning a hand-written digit detector as a logistic regression problem, where one aims to train a classifier $\mathbf{x} \in \mathbb{R}^p$ to determine the relationship between feature vectors $\mathbf{z}_n \in \mathbb{R}^p$ and their associated labels $y_n \in \{-1, 1\}$ for $n = 1, \ldots, N$. The instantaneous function f_n in (1) for this setting is the λ-regularized negative log-likelihood of a generalized linear model of the odds ratio of whether the label is $y_n = 1$ or $y_n = -1$. The empirical risk minimization associated with training set $\mathcal{T} = \{(\mathbf{z}_n, y_n)\}_{n=1}^N$ is to find $\mathbf{x}^*$ as the maximum a posteriori estimate

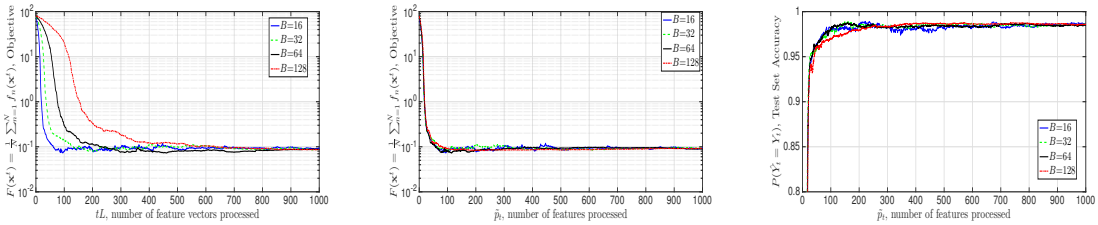
$$\mathbf{x}^* := \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^p} \frac{\lambda}{2} \|\mathbf{x}\|^2 + \frac{1}{N} \sum_{n=1}^{N} \log(1 + \exp(-y_n \mathbf{x}^T \mathbf{z}_n)) \,, \tag{34}$$

where the regularization term $(\lambda/2)\|\mathbf{x}\|^2$ encodes a prior belief on the joint distribution of $(\mathbf{z}, y)$ and helps to avoid overfitting. We use the MNIST dataset [43], in which feature vectors $\mathbf{z}_n \in \mathbb{R}^p$ are $p = 28^2 = 784$ pixel images whose values are recorded as intensities, or elements of the unit interval $[0, 1]$. Considered here is the subset associated with digits 0 and 8, a training set $\mathcal{T} = \{\mathbf{z}_n, y_n\}_{n=1}^N$ with $N = 1.76 \times 10^4$ sample points.



(a) Objective $F(\mathbf{x}^t)$ vs. iteration t. (b) Objective $F(\mathbf{x}^t)$ vs. feature $\tilde{p}_t$. (c) Test Set Accuracy vs. feature $\tilde{p}_t$

Figure 8: RAPSA on MNIST data with constant step-size $\gamma^t = \gamma = 10^{-.5}$ with no mini-batching $L = 1$. Algorithm performance is best in terms of number of iterations t when all blocks are used per step (parallelized SGD), but in terms of number of features processed, the methods perform comparably. Thus RAPSA performs as well as SGD while breaking the complexity bottleneck in p, the dimension of decision variable $\mathbf{x}$.



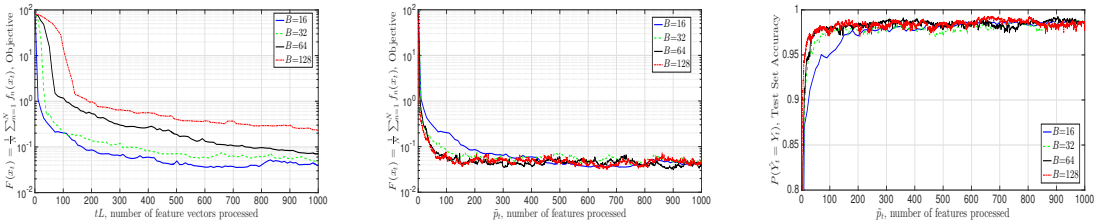
(a) Objective $F(\mathbf{x}^t)$ vs. iteration t. (b) Objective $F(\mathbf{x}^t)$ vs. feature $\tilde{p}_t$. (c) Test Set Accuracy vs. feature $\tilde{p}_t$

Figure 9: RAPSA on MNIST data with hybrid step-size $\gamma^t = \min(10^{-3/4}, 10^{-3/4}\tilde{T}_0/t)$, with $\tilde{T}_0 = 300$ and no mini-batching $L = 1$. As with the constant step-size selection, we observe that updating all blocks per iteration is best in terms of t, but in terms of elements of $\mathbf{x}$ updated, algorithm performance is nearly identical, meaning that no price is payed for breaking the complexity bottleneck in p.

Results for RAPSA We run RAPSA on this training subset for the cases that $B = 16$, $B = 32$, $B = 64$, and $B = 128$, which are associated with updating p, $p/2$, $p/4$, and $p/8$ features per iteration. We consider the use of RAPSA with both constant and hybrid step-size selections. In Figure 8, we display the results when we select a constant learning rate $\gamma^t = \gamma = 10^{-.5} = 0.316$. In Figure 8(a) we plot the objective $F(\mathbf{x}^t)$ versus iteration t, and observe that algorithm performance improves with using more elements of $\mathbf{x}$ per iteration. That is, using all p coordinates of $\mathbf{x}$ achieves superior convergence with respect to iteration t. However, as previously noted, iteration index t is an unfair comparator for objective convergence since the four different setting process different number of features per iteration. In Figure 8(b), we instead consider $F(\mathbf{x}^t)$ versus the number of coordinates of $\mathbf{x}$, denoted as $\tilde{p}_t$, that algorithm performance is comparable across the different selections of B. This demonstrates that RAPSA breaks the computational bottleneck in p while suffering no reduction in convergence speed with respect to $\tilde{p}_t$.

We consider further the classification accuracy on a test subset of size $\tilde{N} = 5.88 \times 10^3$, the results of which are shown in Fig. 9(c). We see that the result for classification accuracy on a test set is consistent with the results for the convergence of the objective function value, and asymptotically reach approximately 98% across the different instances of RAPSA.

In Figure 9 we show the result of running RAPSA for this logistic regression problem with



(a) Objective $F(\mathbf{x}^t)$ vs. iteration t. (b) Objective $F(\mathbf{x}^t)$ vs. feature $\tilde{p}_t$. (c) Test Set Accuracy vs. feature $\tilde{p}_t$

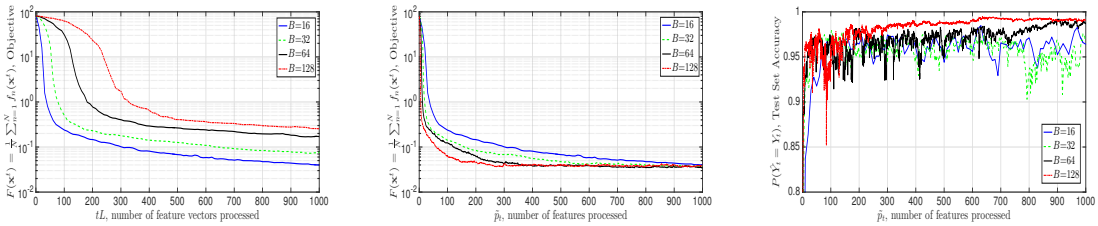
Figure 10: ARAPSA on MNIST data with constant step-size $\gamma^t = \gamma = 10^{-2}$ and mini-batch size $L = 10$, curvature memory $\tau = 10$, and regularizer $\lambda = 7.5 \times 10^{-3}$. Algorithm performance is comparable across different numbers of decision variable coordinates updated per iteration t, but in terms of number of features processed, ARAPSA performance best when using the least information per update.

hybrid step-size $\gamma^t = \min(10^{-3/4}, 10^{-3/4}\tilde{T}_0/t)$, with $\tilde{T}_0 = 300$ and no mini-batching $L = 1$. In Fig. 9(a), which displays the objective $F(\mathbf{x}^t)$ versus iteration t, that using full stochastic gradients is better than only updating *some* of the coordinates in terms of the number of iterations t. In particular, to reach the objective benchmark $F(\mathbf{x}^t) \leq 10^{-1}$, we have to run RAPSA $t = 74$, $t = 156$, and $t = 217$, and $t = 631$ iterations, for the cases that $B = 16$, $B = 32$, $B = 64$, and $B = 128$. We illustrate the objective $F(\mathbf{x}^t)$ vs. feature $\tilde{p}_t$ in Fig. 9(b). Here we recover the advantages of randomized incomplete parallel processing: updating fewer blocks per iteration yields comparable algorithm performance.

We additionally display the algorithm's achieved test-set accuracy on a test subset of size $\tilde{N} = 5.88 \times 10^3$ in Fig. 9(c) under the hybrid step-size regime. We again see that after a burn-in period, the classifier achieves the highly accurate asymptotic error rate of between $1 - 2\%$ across the different instantiations of RAPSA. We note that the test set accuracy achieved by the hybrid scheme is superior to the constant step-size setting.

Results for Accelerated RAPSA We now run Accelerated RAPSA (Algorithm 3) as stated in Section C for this problem setting for the entire MNIST binary training subset associated with digits 0 and 8, with mini-batch size $L = 10$ and the level of curvature information set as $\tau = 10$. We further select regularizer $\lambda = 1/\sqrt{N} = 7.5 \times 10^{-3}$, and consider both constant and hybrid step-size regimes. As before, we study the advantages of incomplete randomized parallel processing by varying the number of blocks $B \in \{16, 32, 64, 128\}$ on an architecture with a fixed number $|\mathcal{I}_t| = I = 16$ of processors. This setup is associated with using all p entries of vector $\mathbf{x}$ at each iteration as compared with $1/2$, $1/4$, and $1/8$ of its entries.

Figures 10 the results of this algorithm run when a constant step-size $\gamma = 10^{-2}$ is used. Observe in Figure 10(a) that the algorithm achieves convergence across the differing numbers of blocks B in terms of iteration t, with faster learning rates achieved with smaller B. In particular, to reach the benchmark $F(\mathbf{x}^t) \leq 10^{-1}$, we require $t = 145$, $t = 311$, and $t = 701$ iterations for $B = 16$, $B = 32$, and $B = 64$, respectively, whereas the case $B = 128$ does not achieve this benchmark by $t = 10^3$. This trend is inverted, however, in Figure 10(b), which displays the objective $F(\mathbf{x}^t)$ with $\tilde{p}_t$ the number of coordinates of $\mathbf{x}$ on which the algorithm operates per step. Observe that using *fewer* entries of $\mathbf{x}$ per iteration is better in terms of number of features processed $\tilde{p}_t$. Furthermore, ARAPSA achieves comparable accuracy on a test set of images, approximately near 98% across different selections of B, as is displayed in Figure 10(c).



(a) Objective $F(\mathbf{x}^t)$ vs. iteration t. (b) Objective $F(\mathbf{x}^t)$ vs. feature $\tilde{p}_t$. (c) Test Set Accuracy vs. feature $\tilde{p}_t$

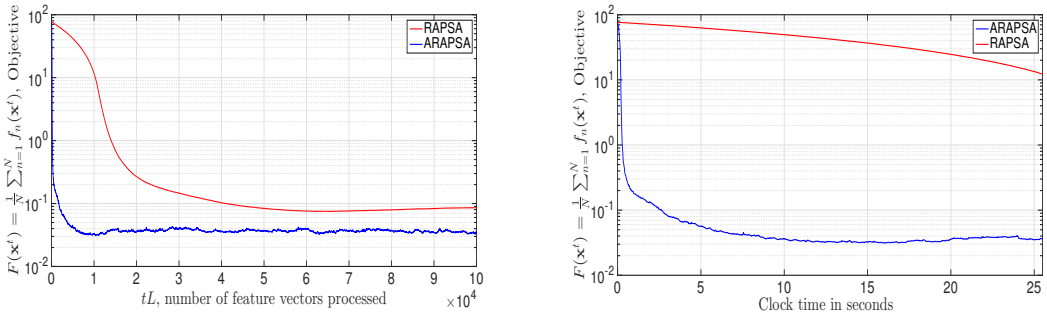
Figure 11: ARAPSA on MNIST data with hybrid step-size $\gamma^t = \min(10^{-1}, 10^{-1}\tilde{T}_0/t)$, with $\tilde{T}_0 = 500$, mini-batch size $L = 10$, curvature memory $\tau = 10$, and regularizer $\lambda = 7.5 \times 10^{-3}$. Algorithm performance is comparable across different numbers of decision variable coordinates updated per iteration t, but in terms of number of features processed, RAPSA performance best when using the least information per update.

We now run Accelerated RAPSA when the learning rate is hand-tuned to optimize performance via a hybrid scheme $\gamma^t = \min(10^{-1}, 10^{-1}\tilde{T}_0/t)$, with attenuation threshold $\tilde{T}_0 = 500$. The results of this experiment are given in Figure 11. In particular, in Figure 11(a) we plot the objective $F(\mathbf{x}^t)$ with iteration t when the number of blocks B is varied. We see that parallelized oL-BFGS ($I = B$ so that $r = 1$) performs best in terms of t: to achieve the threshold condition $F(\mathbf{x}^t) \leq 10^{-1}$, we require $t = 278$, $t = 522$ iterations for $B = 16$ and $B = 32$, respectively, whereas the cases $B = 64$ and $B = 128$ do not achieve this benchmark by $t = 10^3$. However, the instance of ARAPSA with the fastest and most accurate convergence uses the *least* coordinates of $\mathbf{x}$ when we compare the objective with $\tilde{p}_t$, as may be observed in Figure 11(b). This trend is corroborated in Figure 11(c), where we observe that ARAPSA with $B = 128$ achieves 99% test-set accuracy the fastest, followed by $B = 64$, $B = 32$, and $B = 16$.

Comparison of RAPSA and ARAPSA We now compare the performance of RAPSA and its accelerated variant on the MNIST digit recognition problem for a binary subset of the training data consisting of $N = 10^5$ samples. We run both algorithms on an $I = 16$ processor simulated architecture with $B = 64$ blocks, such that $r = 1/4$ of the elements of $\mathbf{x}$ are operated upon at each step. We consider the constant algorithm step-size scheme $\gamma = 10^{-2}$ with mini-batch size $L = 10$.

The results of this online training procedure are given in (12), where we plot the objective optimality gap $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ versus the number of feature vectors processed tL (Figure 12(a)) and actual elapsed time (Figure 12(b)). We see ARAPSA achieves superior convergence behavior with respect to RAPSA in terms of number of feature vectors processed: to achieve the benchmark $F(\mathbf{x}^t) - F(\mathbf{x}^*) \leq 10^{-1}$, ARAPSA requires fewer than $tL = 200$ feature vectors, whereas RAPSA requires $tL = 4 \times 10^4$ feature vectors. This relationship is corroborated in Figure 12(b), where we see that within a couple seconds ARAPSA converges to within 10^{-1}, whereas after *five* times as long, RAPSA does not achieve this benchmark.

Results for Asynchronous RAPSA We now evaluate the empirical performance of the asynchronous variant of RAPSA (Algorithm 4) proposed in Section D.1 on the logistic regression formulation of the MNIST digit recognition problem. The model we use for asynchronicity is the one outlined in Section F.1, that is, each local processor has a distinct local clock which is not required coincide with others, begins at time $t_0^i = t_0$ for all processors $i = 1, \ldots, I$, and then selects subsequent times as $t_k = t_{k-1} + w_k^i$. Here $w_k^i \sim \mathcal{N}(\mu, \sigma^2)$ is a normal random variable with mean μ and variance σ^2 which controls the amount of variability between the clocks of distinct processors.



(a) $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. iteration t

(b) $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ vs. clock time (s)

Figure 12: A comparison of RAPSA and ARAPSA on the MNIST digit recognition problem for a binary training subset of size $N = 10^3$ with mini-batch size $L = 10$ in the constant step-size scheme $\gamma = 10^{-2}$. The objective optimality gap $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ is shown with respect to the number of feature vectors processed tL (left) and actual elapsed time (right). While the performance difference between RAPSA and ARAPSA is not as large as in the linear estimation problem, we still observe that ARAPSA substantially accelerates the convergence of RAPSA for a standard machine learning problem.

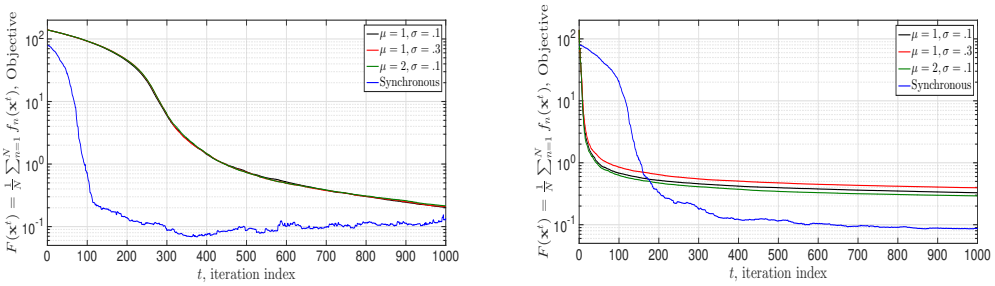
We run the algorithm with no regularization $\lambda = 0$ or mini-batching $L = 1$ and initialization $\mathbf{x}_0 = \mathbf{1}$.

The results of this numerical setup are given in Figure 13. We consider the expected risk $F(\mathbf{x}^t)$ in both both the constant ($\gamma = 10^{-2}$, Figure 13(a)) and diminishing ($\gamma^t = 1/t$, Figure 13(b)) algorithm step-size schemes. We see that the level of asynchronicity does not significantly impact the performance in either scheme, and that the convergence guarantees established in Theorem 3 hold true in practice. We again observe that the version of RAPSA with synchronized computations converges at a faster rate than Asynchronous RAPSA.

G Conclusions

We proposed the random parallel stochastic algorithm (RAPSA) proposed as a doubly stochastic approximation algorithm capable of optimization problems associated with learning problems in which both the number of predictive parameters and sample size are huge-scale. RAPSA is doubly stochastic since each processors utilizes a random set of functions to compute the stochastic gradient associated with a randomly chosen sets of variable coordinates. We showed the proposed algorithm converges to the optimal solution sublinearly when the step-size is diminishing. Moreover, linear convergence to a neighborhood of the optimal solution can be achieved using a constant step-size. We further introduced accelerated and asynchronous variants of RAPSA, and presented convergence guarantees for asynchronous RAPSA.

A detailed numerical comparison between RAPSA and parallel SGD for learning a linear estimator and a logistic regressor is provided. The numerical results showcase the advantage of RAPSA with respect to parallel SGD. Further empirical results illustrate the advantages of ARAPSA with respect to parallel oL-BFGS, and that implementing the algorithm on a lock-free parallel computing cluster does not substantially degrade empirical performance.



(a) Objective $F(\mathbf{x}^t)$ vs. iteration t.

(b) Objective $F(\mathbf{x}^t)$ vs. iteration t.

Figure 13: Asynchronous RAPSA on MNIST data in the constant ($\gamma = 10^{-2}$, left) and diminishing ($\gamma^t = 1/t$, right) step-size schemes with no mini-batching $L = 1$ for a binary training subset of size $N = 10^3$ with no regularization $\lambda = 0$ when the algorithm is initialized as $\mathbf{x}_0 = \mathbf{1}$. The variability in local processor clocks does not significantly impact performance in both the diminishing and constant step-size settings; however, the synchronous algorithm converges at a faster rate.

A Proof of Results Leading to Theorems 1 and 2

A.1 Proof of Lemma 1

Recall that the components of vector $\mathbf{x}^{t+1}$ are equal to the components of $\mathbf{x}^t$ for the coordinates that are not updated at step t, i.e., $i \notin \mathcal{I}^t$. For the updated coordinates $i \in \mathcal{I}^t$ we know that $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \gamma^t \nabla_{\mathbf{x}_i^t} f(\mathbf{x}^t, \boldsymbol{\theta}^t)$. Therefore, $B - I$ blocks of the vector $\mathbf{x}^{t+1} - \mathbf{x}^t$ are 0 and the remaining I randomly chosen blocks are given by $-\gamma^t \nabla_{\mathbf{x}_i^t} f(\mathbf{x}^t, \boldsymbol{\theta}^t)$. Notice that there are $\binom{B}{I}$ different ways for picking I blocks out of the whole B blocks. Therefore, the probability of each combination of blocks is $1/\binom{B}{I}$. Further, each block appears in $\binom{B-1}{I-1}$ of the combinations. Therefore, the expected value can be written as

$$\mathbb{E}_{\mathcal{I}^t}\left[\mathbf{x}^{t+1} - \mathbf{x}^t \mid \mathcal{F}^t\right] = \frac{\binom{B-1}{I-1}}{\binom{m}{I}} \left(-\gamma^t \nabla f(\mathbf{x}^t, \boldsymbol{\Theta}^t)\right). \tag{35}$$

Observe that simplifying the ratio in the right hand sides of (35) leads to

$$\frac{\binom{B-1}{I-1}}{\binom{B}{I}} = \frac{\frac{(B-1)!}{(I-1)! \times (B-I)!}}{\frac{p!}{I! \times (B-I)!}} = \frac{I}{B} = r. \tag{36}$$

Substituting the simplification in (36) into (35) follows the claim in (16). To prove the claim in (17) we can use the same argument that we used in proving (16) to show that

$$\mathbb{E}_{\mathcal{I}^t}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^t\|^2 \mid \mathcal{F}^t\right] = \frac{\binom{B-1}{I-1}}{\binom{B}{I}} (\gamma^t)^2 \left\|\nabla f(\mathbf{x}^t, \boldsymbol{\Theta}^t)\right\|^2. \tag{37}$$

By substituting the simplification in (36) into (37) the claim in (17) follows.

A.2 Proof of Proposition 1

By considering the Taylor's expansion of $F(\mathbf{x}^{t+1})$ near the point $\mathbf{x}^t$ and observing the Lipschitz continuity of gradients ∇F with constant M we obtain that the average objective function $F(\mathbf{x}^{t+1})$ is bounded above by

$$F(\mathbf{x}^{t+1}) \leq F(\mathbf{x}^t) + \nabla F(\mathbf{x}^t)^T(\mathbf{x}^{t+1} - \mathbf{x}^t) + \frac{M}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2. \tag{38}$$

Compute the expectation of the both sides of (38) with respect to the random set $\mathcal{I}^t$ given the observed set of information $\mathcal{F}^t$. Substitute $\mathbb{E}_{\mathcal{I}^t}\left[\mathbf{x}^{t+1} - \mathbf{x}^t \mid \mathcal{F}^t\right]$ and $\mathbb{E}_{\mathcal{I}^t}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \mid \mathcal{F}^t\right]$ with their simplifications in (16) and (17), respectively, to write

$$\mathbb{E}_{\mathcal{I}^t}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - r\gamma^t \ \nabla F(\mathbf{x}^t)^T \nabla f(\mathbf{x}^t, \mathbf{\Theta}^t) + \frac{rM(\gamma^t)^2}{2} \ \left\|\nabla f(\mathbf{x}^t, \mathbf{\Theta}^t)\right\|^2. \tag{39}$$

Notice that the stochastic gradient $\nabla f(\mathbf{x}^t, \mathbf{\Theta}^t)$ is an unbiased estimate of the average function gradient $\nabla F(\mathbf{x}^t)$. Therefore, we obtain $\mathbb{E}_{\mathbf{\Theta}^t}\left[\nabla f(\mathbf{x}^t, \mathbf{\Theta}^t) \mid \mathcal{F}^t\right] = \nabla F(\mathbf{x}^t)$. Observing this relation and considering the assumption in (15), the expected value of (39) with respect to the set of realizations $\mathbf{\Theta}^t$ can be written as

$$\mathbb{E}_{\mathcal{I}^t, \mathbf{\Theta}^t}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - r\gamma^t \ \left\|\nabla F(\mathbf{x}^t)\right\|^2 + \frac{rM(\gamma^t)^2 K}{2}. \tag{40}$$

Subtracting the optimal objective function value $F(\mathbf{x}^*)$ form the both sides of (40) implies that

$$\mathbb{E}_{\mathcal{I}^t, \mathbf{\Theta}^t}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - F(\mathbf{x}^*) - r\gamma^t \ \left\|\nabla F(\mathbf{x}^t)\right\|^2 + \frac{rM(\gamma^t)^2 K}{2}. \tag{41}$$

We proceed to find a lower bound for the gradient norm $\|\nabla F(\mathbf{x}^t)\|$ in terms of the objective value error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$. Assumption 1 states that the average objective function F is strongly convex with constant $m > 0$. Therefore, for any $\mathbf{y}, \mathbf{z} \in \mathbb{R}^p$ we can write

$$F(\mathbf{y}) \geq F(\mathbf{z}) + \nabla F(\mathbf{z})^T(\mathbf{y} - \mathbf{z}) + \frac{m}{2}\|\mathbf{y} - \mathbf{z}\|^2. \tag{42}$$

For fixed $\mathbf{z}$, the right hand side of (42) is a quadratic function of $\mathbf{y}$ whose minimum argument we can find by setting its gradient to zero. Doing this yields the minimizing argument $\hat{\mathbf{y}} = \mathbf{z} - (1/m)\nabla F(\mathbf{z})$ implying that for all $\mathbf{y}$ we must have

$$\begin{aligned} F(\mathbf{y}) &\geq F(\mathbf{w}) + \nabla F(\mathbf{z})^T(\hat{\mathbf{y}} - \mathbf{z}) + \frac{m}{2}\|\hat{\mathbf{y}} - \mathbf{z}\|^2 \\ &= F(\mathbf{z}) - \frac{1}{2m}\|\nabla F(\mathbf{z})\|^2. \end{aligned} \tag{43}$$

Observe that the bound in (43) holds true for all $\mathbf{y}$ and $\mathbf{z}$. Setting values $\mathbf{y} = \mathbf{x}^*$ and $\mathbf{z} = \mathbf{x}^t$ in (43) and rearranging the terms yields a lower bound for the squared gradient norm $\|\nabla F(\mathbf{x}^t)\|^2$ as

$$\|\nabla F(\mathbf{x}^t)\|^2 \geq 2m(F(\mathbf{x}^t) - F(\mathbf{x}^*)). \tag{44}$$

Substituting the lower bound in (44) by the norm of gradient square $\|\nabla F(\mathbf{x}^t)\|^2$ in (41) follows the claim in (18).

A.3 Proof of Theorem 1

We use the relationship in (18) to build a supermartingale sequence. To do so, define the stochastic process α^t as

$$\alpha^t := F(\mathbf{x}^t) - F(\mathbf{x}^*) + \frac{rMK}{2}\sum_{u=t}^{\infty}(\gamma^u)^2. \tag{45}$$

Note that α^t is well-defined because $\sum_{u=t}^{\infty}(\gamma^u)^2 \leq \sum_{u=0}^{\infty}(\gamma^u)^2 < \infty$ is summable. Further define the sequence β_t with values

$$\beta^t := \ 2m\gamma^t r(F(\mathbf{x}^t) - F(\mathbf{x}^*)). \tag{46}$$

The definitions of sequences α^t and β^t in (45) and (46), respectively, and the inequality in (18) imply that the expected value α^{t+1} given $\mathcal{F}^t$ can be written as

$$\mathbb{E}\left[\alpha^{t+1} \,\middle|\, \mathcal{F}^t\right] \ \leq \ \alpha^t - \beta^t. \tag{47}$$

Since the sequences α^t and β^t are nonnegative it follows from (47) that they satisfy the conditions of the supermartingale convergence theorem – see e.g. Theorem E7.4 of [44]. Therefore, we obtain that: (i) The sequence α^t converges almost surely to a limit. (ii) The sum $\sum_{t=0}^{\infty}\beta^t < \infty$ is almost surely finite. The latter result yields

$$\sum_{t=0}^{\infty} 2m\gamma^t r(F(\mathbf{x}^t) - F(\mathbf{x}^*)) < \infty. \qquad \text{a.s.} \tag{48}$$

Since the sequence of step sizes is non-summable there exits a subsequence of sequence $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ which is converging to null. This observation is equivalent to almost sure convergence of $\liminf F(\mathbf{x}^t) - F(\mathbf{x}^*)$ to null

$$\liminf_{t\to\infty} F(\mathbf{x}^t) - F(\mathbf{x}^*) = 0. \qquad \text{a.s.} \tag{49}$$

Based on the martingale convergence theorem for the sequences α^t and β^t in relation (47), the sequence α^t almost surely converges to a limit. Consider the definition of α^t in (45). Observe that the sum $\sum_{u=t}^{\infty}(\gamma^u)^2$ is deterministic and its limit is null. Therefore, the sequence of the objective function value error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ almost surely converges to a limit. This observation in association with the result in (49) implies that the whole sequence of $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ converges almost surely to null,

$$\lim_{t\to\infty} \ F(\mathbf{x}^t) - F(\mathbf{x}^*) = 0. \qquad \text{a.s.} \tag{50}$$

The last step is to prove almost sure convergence of the sequence $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ to null, as a result of the limit in (50). To do so, we follow by proving a lower bound for the objective function value error $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ in terms of the squared norm error $\|\mathbf{x}^t - \mathbf{x}^*\|^2$. According to the strong convexity assumption, we can write the following inequality

$$F(\mathbf{x}^t) \geq F(\mathbf{x}^*) + \nabla F(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*) + \frac{m}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2. \tag{51}$$

Observe that the gradient of the optimal point is the null vector, i.e., $\nabla F(\mathbf{x}^*) = \mathbf{0}$. This observation and rearranging the terms in (51) imply that

$$F(\mathbf{x}^t) - F(\mathbf{x}^*) \geq \frac{m}{2}\|\mathbf{x}^t - \mathbf{x}^*\|^2. \tag{52}$$

The upper bound in (52) for the squared norm $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ in association with the fact that the sequence $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ almost surely converges to null, leads to the conclusion that the sequence $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ almost surely converges to zero. Hence, the claim in (19) is valid.

The next step is to study the convergence rate of RAPSA in expectation. In this step we assume that the diminishing stepsize is defined as $\gamma^t = \gamma^0 T^0/(t + T^0)$. Recall the inequality in (18). Substitute γ^t by $\gamma^0 T^0/(t + T^0)$ and compute the expected value of (18) given $\mathcal{F}^0$ to obtain

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*)\right] \leq \left(1 - \frac{2mr\gamma^0 T^0}{(t + T^0)}\right) \mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] + \frac{rMK(\gamma^0 T^0)^2}{2(t + T^0)^2}. \tag{53}$$

We use the following lemma to show that the result in (53) implies sublinear convergence of the sequence of expected objective value error $\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right]$.

Lemma 3 *Let $c > 1$, $b > 0$ and $t^0 > 0$ be given constants and $u_t \geq 0$ be a nonnegative sequence that satisfies*

$$u^{t+1} \leq \left(1 - \frac{c}{t + t^0}\right) u^t + \frac{b}{(t + t^0)^2}\ , \tag{54}$$

for all times $t \geq 0$. The sequence u^t is then bounded as

$$u^t \leq \frac{Q}{t + t^0}, \tag{55}$$

for all times $t \geq 0$, where the constant Q is defined as $Q := \max\{b/(c-1),\ t^0 u^0\}$.

Proof: See [42]. ■

Lemma 3 shows that if a sequence u^t satisfies the condition in (54) then the sequence u^t converges to null at least with the rate of $\mathcal{O}(1/t)$. By assigning values $t^0 = T^0$, $u^t = \mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right]$, $c = 2mr\gamma^0 T^0$, and $b = rMK(\gamma^0 T^0)^2/2$, the relation in (53) implies that the inequality in (54) is satisfied for the case that $2mr\gamma^0 T^0 > 1$. Therefore, the result in (55) holds and we can conclude that

$$\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] \leq \frac{C}{t + T^0}, \tag{56}$$

where the constant C is defined as

$$C = \max\left\{\frac{rMK(\gamma^0 T^0)^2}{4rm\gamma^0 T^0 - 2},\ T^0(F(\mathbf{x}^0) - F(\mathbf{x}^*))\right\}. \tag{57}$$

A.4 Proof of Theorem 2

To prove the claim in (22) we use the relationship in (18) (Proposition 1) to construct a supermartingale. Define the stochastic process α^t with values

$$\alpha^t := \big(F(\mathbf{x}^t) - F(\mathbf{x}^*)\big) \times \mathbf{1}\left\{\min_{u \leq t} F(\mathbf{x}^u) - F(\mathbf{x}^*) > \frac{\gamma MK}{4m}\right\} \tag{58}$$

The process α^t tracks the optimality gap $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ until the gap becomes smaller than $\gamma MK/2m$ for the first time at which point it becomes $\alpha^t = 0$. Notice that the stochastic process α^t is always non-negative, i.e., $\alpha^t \geq 0$. Likewise, we define the stochastic process β^t as

$$\beta^t := 2\gamma mr\left(F(\mathbf{x}^t) - F(\mathbf{x}^*) - \frac{\gamma MK}{4m}\right) \times \mathbf{1}\left\{\min_{u \leq t}\ F(\mathbf{x}^u) - F(\mathbf{x}^*) > \frac{\gamma MK}{4m}\right\}, \tag{59}$$

which follows $2\gamma mr\left(F(\mathbf{x}^t)-F(\mathbf{x}^*)-\gamma MK/4m\right)$ until the time that the optimality gap $F(\mathbf{x}^t)-F(\mathbf{x}^*)$ becomes smaller than $\gamma MK/2m$ for the first time. After this moment the stochastic process β^t becomes null. According to the definition of β^t in (59), the stochastic process satisfies $\beta^t \geq 0$ for all $t \geq 0$. Based on the relationship (18) and the definitions of stochastic processes α^t and β^t in (58) and (59) we obtain that for all times $t \geq 0$

$$\mathbb{E}\left[\alpha^{t+1} \mid \mathcal{F}^t\right] \leq \alpha^t - \beta^t. \tag{60}$$

To check the validity of (60) we first consider the case that $\min_{u\leq t}\ F(\mathbf{x}^u)-F(\mathbf{x}^*) > \gamma MK/4m$ holds. In this scenario we can simply the stochastic processes in (58) and (59) as $\alpha^t = F(\mathbf{x}^t)-F(\mathbf{x}^*)$ and $\beta^t = 2\gamma mr\left(F(\mathbf{x}^t)-F(\mathbf{x}^*)-\gamma MK/4m\right)$. Therefore, according to the inequality in (18) the result in (60) is valid. The second scenario that we check is $\min_{u\leq t}\ F(\mathbf{x}^u)-F(\mathbf{x}^*) \leq \gamma MK/4m$. Based on the definitions of stochastic processes α^t and β^t, both of these two sequences are equal to 0. Further, notice that when $\alpha^t = 0$, it follows that $\alpha^{t+1} = 0$. Hence, the relationship in (60) is true.

Given the relation in (60) and non-negativity of stochastic processes α^t and β^t we obtain that α^t is a supermartingale. The supermartingale convergence theorem yields: i) The sequence α^t converges to a limit almost surely. ii) The sum $\sum_{t=1}^{\infty}\beta^t$ is finite almost surely. The latter result implies that the sequence β^t is converging to null almost surely, i.e.,

$$\lim_{t\to\infty}\beta^t \ = \ 0 \quad \text{a.s.} \tag{61}$$

Based on the definition of β^t in (59), the limit in (61) is true if one of the following events holds: i) The indicator function is null after for large t. ii) The limit $\lim_{t\to\infty}\left(F(\mathbf{x}^t)-F(\mathbf{x}^*)-\gamma MK/4m\right) = 0$ holds true. From any of these two events we it is implied that

$$\liminf_{t\to\infty}\ F(\mathbf{x}^t)-F(\mathbf{x}^*) \ \leq \ \frac{\gamma MK}{4m} \quad \text{a.s.} \tag{62}$$

Therefore, the claim in (22) is valid. The result in (62) shows the objective function value sequence $F(\mathbf{x}^t)$ almost sure converges to a neighborhood of the optimal objective function value $F(\mathbf{x}^*)$.

We proceed to prove the result in (23). Compute the expected value of (18) given $\mathcal{F}^0$ and set $\gamma^t = \gamma$ to obtain

$$\mathbb{E}\left[F(\mathbf{x}^{t+1})-F(\mathbf{x}^*)\right] \leq (1-2m\gamma r)\,\mathbb{E}\left[F(\mathbf{x}^t)-F(\mathbf{x}^*)\right] + \frac{rMK\gamma^2}{2}. \tag{63}$$

Notice that the expression in (63) provides an upper bound for the expected value of objective function error $\mathbb{E}\left[F(\mathbf{x}^{t+1})-F(\mathbf{x}^*)\right]$ in terms of its previous value $\mathbb{E}\left[F(\mathbf{x}^t)-F(\mathbf{x}^*)\right]$ and an error term. Rewriting the relation in (63) for step $t-1$ leads to

$$\mathbb{E}\left[F(\mathbf{x}^{t})-F(\mathbf{x}^*)\right] \leq (1-2m\gamma r)\,\mathbb{E}\left[F(\mathbf{x}^{t-1})-F(\mathbf{x}^*)\right] + \frac{rMK\gamma^2}{2}. \tag{64}$$

Substituting the upper bound in (64) for the expectation $\mathbb{E}\left[F(\mathbf{x}^t)-F(\mathbf{x}^*)\right]$ in (63) follows an upper bound for the expected error $\mathbb{E}\left[F(\mathbf{x}^{t+1})-F(\mathbf{x}^*)\right]$ as

$$\mathbb{E}\left[F(\mathbf{x}^{t+1})-F(\mathbf{x}^*)\right] \leq (1-2m\gamma r)^2\,\mathbb{E}\left[F(\mathbf{x}^{t-1})-F(\mathbf{x}^*)\right] + \frac{rMK\gamma^2}{2}\left(1+(1-2mr\gamma)\right). \tag{65}$$

By recursively applying the steps in (64)-(65) we can bound the expected objective function error $\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*)\right]$ in terms of the initial objective function error $F(\mathbf{x}^0) - F(\mathbf{x}^*)$ and the accumulation of the errors as

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*)\right] \leq (1 - 2m\gamma r)^{t+1} \left(F(\mathbf{x}^0) - F(\mathbf{x}^*)\right) + \frac{rMK\gamma^2}{2} \sum_{u=0}^{t} (1 - 2mr\gamma)^u. \tag{66}$$

Substituting t by $t-1$ and simplifying the sum in the right hand side of (66) yields

$$\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] \leq (1 - 2m\gamma r)^{t} \left(F(\mathbf{x}^0) - F(\mathbf{x}^*)\right) + \frac{MK\gamma}{4m} \left[1 - (1 - 2mr\gamma)^t\right]. \tag{67}$$

Observing that the term $1 - (1 - 2mr\gamma)^t$ in the right hand side of (67) is strictly smaller than 1 for the stepsize $\gamma < 1/(2mr)$, the claim in (23) follows.

B Proofs Leading up to Theorem 3

B.1 Proof of Lemma 2

Proof: Recall that the components of vector $\mathbf{x}^{t+1}$ are equal to the components of $\mathbf{x}^t$ for the coordinates that are not updated at step t, i.e., $i \notin \mathcal{I}^t$. For the updated coordinates $i \in \mathcal{I}^t$ we know that $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \gamma^t \nabla_{\mathbf{x}_i^t} f(\mathbf{x}^{t-\tau}, \boldsymbol{\theta}^{t-\tau})$. Therefore, $B-1$ blocks of the vector $\mathbf{x}^{t+1} - \mathbf{x}^t$ are 0 and only one block is given by $-\gamma^t \nabla_{\mathbf{x}_i} f(\mathbf{x}^{t-\tau}, \boldsymbol{\theta}^{t-\tau})$. Since the corresponding processor picks its block uniformly at random from the B sets of blocks we obtain that the expected value of the difference $\mathbf{x}^{t+1} - \mathbf{x}^t$ with respect to the index of the block at time t is given by

$$\mathbb{E}_{\mathcal{I}^t}\left[\mathbf{x}^{t+1} - \mathbf{x}^t \mid \mathcal{F}^t\right] = \frac{1}{B}\left(-\gamma^t \nabla f(\mathbf{x}^{t-\tau}, \boldsymbol{\Theta}^{t-\tau})\right). \tag{68}$$

Substituting the simplification in (68) in place of (35) in the proof of Lemma 1 and simplifying the resulting expression yields the claim in (28). To prove the claim in (29) we can use the same argument that we used in proving (28) to show that

$$\mathbb{E}_{\mathcal{I}^t}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^t\|^2 \mid \mathcal{F}^t\right] = \frac{(\gamma^t)^2}{B} \left\|\nabla f(\mathbf{x}^{t-\tau}, \boldsymbol{\Theta}^{t-\tau})\right\|^2, \tag{69}$$

which completes the proof. ■

B.2 Proof of Proposition 2

By considering the Taylor's expansion of $F(\mathbf{x}^{t+1})$ near the point $\mathbf{x}^t$ and observing the Lipschitz continuity of gradients ∇F with constant M we obtain that the average objective function $F(\mathbf{x}^{t+1})$ is bounded above by

$$F(\mathbf{x}^{t+1}) \leq F(\mathbf{x}^t) + \nabla F(\mathbf{x}^t)^T(\mathbf{x}^{t+1} - \mathbf{x}^t) + \frac{M}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2. \tag{70}$$

Compute the expectation of the both sides of (70) with respect to the random indexing set $\mathcal{I}^t \subset \{1, \ldots, B\}$ associated with chosen blocks given the observed set of information $\mathcal{F}^t$. Substitute $\mathbb{E}_{\mathcal{I}^t}\left[\mathbf{x}^{t+1} - \mathbf{x}^t \mid \mathcal{F}^t\right]$ and $\mathbb{E}_{\mathcal{I}^t}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \mid \mathcal{F}^t\right]$ with their simplifications in (28) and (29), respectively, to write

$$\mathbb{E}_{\mathcal{I}^t}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - \frac{\gamma^t}{B}\, \nabla F(\mathbf{x}^t)^T \nabla f(\mathbf{x}^{t-\tau}, \boldsymbol{\Theta}^{t-\tau}) + \frac{M(\gamma^t)^2}{2B}\, \left\|\nabla f(\mathbf{x}^{t-\tau}, \boldsymbol{\Theta}^{t-\tau})\right\|^2. \quad (71)$$

Notice that the stochastic gradient $\nabla f(\mathbf{x}^{t-\tau}, \boldsymbol{\Theta}^{t-\tau})$ is an unbiased estimate of the average function gradient $\nabla F(\mathbf{x}^{t-\tau})$. Therefore, we obtain $\mathbb{E}\left[\nabla f(\mathbf{x}^{t-\tau}, \boldsymbol{\Theta}^{t-\tau}) \mid \mathcal{F}^t\right] = \nabla F(\mathbf{x}^{t-\tau})$. Observing this relation and considering the assumption in (15), the expected value of (71) given the sigma algebra $\mathcal{F}^t$ can be written as

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - \frac{\gamma^t}{B}\, \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau}) + \frac{M(\gamma^t)^2 K}{2B}. \quad (72)$$

By adding and subtracting the term $(\gamma^t/B)\|\nabla F(\mathbf{x}^t)\|^2$ to the right hand side of (72) we obtain

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - \frac{\gamma^t}{B}\, \|\nabla F(\mathbf{x}^t)\|^2 + \frac{\gamma^t}{B}\left(\|\nabla F(\mathbf{x}^t)\|^2 - \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau})\right) + \frac{M(\gamma^t)^2 K}{2B}. \quad (73)$$

Observe that the third term on the right-hand side of (73) is the directional error due to the presence of delays from asynchronicity. We proceed to find an upper bound for the expression $\|\nabla F(\mathbf{x}^t)\|^2 - \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau})$, which means that the error due to delay may be mitigated. To do so, notice that we can write

$$\begin{aligned} \|\nabla F(\mathbf{x}^t)\|^2 - \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau}) &= \nabla F(\mathbf{x}^t)^T (\nabla F(\mathbf{x}^t) - \nabla F(\mathbf{x}^{t-\tau})) \\ &\leq \|\nabla F(\mathbf{x}^t)\| \|\nabla F(\mathbf{x}^t) - \nabla F(\mathbf{x}^{t-\tau})\|, \end{aligned} \quad (74)$$

where for the inequality we have used the Cauchy–Schwarz inequality. Apply the fact that the gradient of the objective function is M-Lipschitz continuous, which implies that $\|\nabla F(\mathbf{x}^t) - \nabla F(\mathbf{x}^{t-\tau})\| \leq L\|\mathbf{x}^t - \mathbf{x}^{t-\tau}\|$. Substituting the upper bound $L\|\mathbf{x}^t - \mathbf{x}^{t-\tau}\|$ for $\|\nabla F(\mathbf{x}^t) - \nabla F(\mathbf{x}^{t-\tau})\|$ into (74) we obtain

$$\|\nabla F(\mathbf{x}^t)\|^2 - \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau}) \leq L\|\nabla F(\mathbf{x}^t)\| \|\mathbf{x}^t - \mathbf{x}^{t-\tau}\|. \quad (75)$$

The difference norm $\|\mathbf{x}^t - \mathbf{x}^{t-\tau}\|$ is equivalent to $\|\sum_{s=t-\tau}^{t-1} (\mathbf{x}^{s+1} - \mathbf{x}^s)\|$ which can be bounded above by $\sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|$ by the triangle inequality. Therefore,

$$\|\nabla F(\mathbf{x}^t)\|^2 - \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau}) \leq L\|\nabla F(\mathbf{x}^t)\| \sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|. \quad (76)$$

Substitute the upper bound in (76) for $\|\nabla F(\mathbf{x}^t)\|^2 - \nabla F(\mathbf{x}^t)^T \nabla F(\mathbf{x}^{t-\tau})$ into (73) to obtain

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - \frac{\gamma^t}{B}\, \|\nabla F(\mathbf{x}^t)\|^2 + \frac{L\gamma^t}{B}\|\nabla F(\mathbf{x}^t)\| \sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\| + \frac{M(\gamma^t)^2 K}{2B}. \quad (77)$$

Note that for any positive scalars a, b, and ρ the inequality $ab \leq (\rho/2)a^2 + (1/2\rho)b^2$ holds. If we set $a := \|\nabla F(\mathbf{x}^t)\|$ and $b := \sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|$ we obtain that

$$\begin{aligned}\|\nabla F(\mathbf{x}^t)\| \sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\| &\leq \frac{\rho}{2}\|\nabla F(\mathbf{x}^t)\|^2 + \frac{1}{2\rho}\left[\sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|\right]^2 \\ &\leq \frac{\rho}{2}\|\nabla F(\mathbf{x}^t)\|^2 + \frac{\tau}{2\rho}\sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|^2, \end{aligned} \tag{78}$$

where the last inequality is an application of the triangle inequality to the second term on the right-hand side of the first line in (78). Now substituting the upper bound in (78) into (77) yields

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^t\right] \leq F(\mathbf{x}^t) - \left(\frac{\gamma^t}{B} - \frac{\rho L \gamma^t}{2B}\right)\|\nabla F(\mathbf{x}^t)\|^2 + \frac{\tau L \gamma^t}{2\rho B}\sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|^2 + \frac{M(\gamma^t)^2 K}{2B}. \tag{79}$$

Compute the expected value of the both sides of (79) given the sigma-algebra $\mathcal{F}^{t-1}$ to obtain

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^{t-1}\right] \leq \mathbb{E}\left[F(\mathbf{x}^t) \mid \mathcal{F}^{t-1}\right] &- \left(\frac{\gamma^t}{B} - \frac{\rho L \gamma^t}{2B}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}^t)\|^2 \mid \mathcal{F}^{t-1}\right] \\ &+ \frac{\tau L \gamma^t}{2\rho B}\mathbb{E}\left[\sum_{s=t-\tau}^{t-1} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|^2 \mid \mathcal{F}^{t-1}\right] + \frac{M(\gamma^t)^2 K}{2B}\ , \end{aligned} \tag{80}$$

which can be simplified as

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^{t-1}\right] \leq \mathbb{E}\left[F(\mathbf{x}^t) \mid \mathcal{F}^{t-1}\right] &- \left(\frac{\gamma^t}{B} - \frac{\rho L \gamma^t}{2B}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}^t)\|^2 \mid \mathcal{F}^{t-1}\right] \\ &+ \frac{\tau L \gamma^t}{2\rho B}\mathbb{E}\left[\sum_{s=t-\tau}^{t-2} \|\mathbf{x}^{s+1} - \mathbf{x}^s\|^2 \mid \mathcal{F}^{t-1}\right] + \frac{\tau L \gamma^t (\gamma^{t-1})^2 K}{2\rho B^2} + \frac{M(\gamma^t)^2 K}{2B}. \end{aligned} \tag{81}$$

Do the same up to $t - \tau$ to get

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}^{t+1}) \mid \mathcal{F}^{t-\tau}\right] \leq \mathbb{E}\left[F(\mathbf{x}^t) \mid \mathcal{F}^{t-\tau}\right] &- \left(\frac{\gamma^t}{B} - \frac{\rho L \gamma^t}{2B}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}^t)\|^2 \mid \mathcal{F}^{t-\tau}\right] \\ &+ \frac{\tau L \gamma^t K}{2\rho B^2}\sum_{s=t-\tau}^{t-1} (\gamma^s)^2 + \frac{M(\gamma^t)^2 K}{2B}. \end{aligned} \tag{82}$$

Notice that the sequence of stepsizes γ^t is decreasing, thus the sum $\sum_{s=t-\tau}^{t-1}(\gamma^s)^2$ in (82) can be bounded above by $\tau(\gamma^{t-\tau})^2$. Applying this substutition and subtracting the optimal objective function value $F(\mathbf{x}^*)$ from both sides of the implied expression lead to

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right] \leq \mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right] &- \left(\frac{\gamma^t}{B} - \frac{\rho L \gamma^t}{2B}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}^t)\|^2 \mid \mathcal{F}^{t-\tau}\right] \\ &+ \frac{\tau^2 L \gamma^t K (\gamma^{t-\tau})^2}{2\rho B^2} + \frac{M(\gamma^t)^2 K}{2B}. \end{aligned} \tag{83}$$

We make use of the fact that the average function $F(\mathbf{x})$ is m-strongly convex in applying the relation $\|\nabla F(\mathbf{x}^t)\|^2 \geq 2m(F(\mathbf{x}^t) - F(\mathbf{x}^*))$ to the expression (84). Therefore,

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right] \leq \mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right] - 2m\left(\frac{\gamma^t}{B} - \frac{\rho L \gamma^t}{2B}\right)\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*) \mid \mathcal{F}^{t-\tau}\right] + \frac{\tau^2 L \gamma^t K (\gamma^{t-\tau})^2}{2\rho B^2} + \frac{M(\gamma^t)^2 K}{2B}, \tag{84}$$

as stated in Proposition 2.

B.3 Proof of Theorem 3

Proof: We use the result in Proposition 2 to define a martingale difference sequence with delay. Begin by defining the non-negative stochastic processes α^t, β^t, and ζ^t for $t \geq 0$ as

$$\alpha^t := F(\mathbf{x}^t) - F(\mathbf{x}^*), \qquad \beta^t := \frac{2m\gamma^t}{B}\left[1 - \frac{\rho M}{2}\right](F(\mathbf{x}^t) - F(\mathbf{x}^*)),$$
$$\zeta^t := \frac{MK(\gamma^t)^2}{2B} + \frac{\tau^2 MK\gamma^t(\gamma^{t-\tau})^2}{2\rho B^2}. \tag{85}$$

According to the definitions in (85) and the inequality in (30) we can write

$$\mathbb{E}\left[\alpha^{t+1} \mid \mathcal{F}^{t-\tau}\right] \leq \mathbb{E}\left[\alpha^t \mid \mathcal{F}^{t-\tau}\right] - \mathbb{E}\left[\beta^t \mid \mathcal{F}^{t-\tau}\right] + \zeta^t. \tag{86}$$

Computing the expected value of both sides of (86) with respect to the initial sigma algebra $\mathbb{E}\left[\cdot \mid \mathcal{F}^0\right] = \mathbb{E}\left[\cdot\right]$ yields

$$\mathbb{E}\left[\alpha^{t+1}\right] \leq \mathbb{E}\left[\alpha^t\right] - \mathbb{E}\left[\beta^t\right] + \zeta^t. \tag{87}$$

Sum both sides of (87) from $t = 0$ to $t = \infty$ and consider the fact that ζ^t is summable and the sequence α^t is non-negative. Thus, we obtain that the series $\sum_{t=0}^{\infty}\mathbb{E}\left[\beta^t\right] < \infty$ is finite. By using Monotone Convergence Theorem, we pull the expectation outside the summand to obtain that $\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t\right] < \infty$. If we define $Y_n := \sum_{t=0}^{n}\beta^t$, we obtain that $Y_n \geq 0$ and $Y_n \leq Y_{n+1}$. Thus, from the result $\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t\right] < \infty$ we can conclude that $\sum_{t=0}^{\infty}\beta^t < \infty$ with probability 1. Now considering the definition of β^t in (85) and the non-summability of the stepsizes $\sum_{t=0}^{\infty}\gamma^t = \infty$, we obtain that a subsequence of the sequence $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ almost surely converges to zero, i..e. the liminf of the sequence $F(\mathbf{x}^t) - F(\mathbf{x}^*)$ is zero,

$$\liminf_{t\to\infty} \; F(\mathbf{x}^t) - F(\mathbf{x}^*) = 0, \quad \text{a.s.} \tag{88}$$

The next step is to study the convergence rate of asynchronous RAPSA in expectation. By setting $\gamma^t = \gamma^0 T^0/(t + T^0)$ in (30) and computing the expected value given the initial sigma algebra $\mathcal{F}^0$ we obtain

$$\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*)\right] \tag{89}$$
$$\leq \left(1 - \frac{2m\gamma^0 T^0}{B(t+T^0)}\left[1 - \frac{\rho M}{2}\right]\right)\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] + \frac{MK(\gamma^0 T^0)^2}{2B(t+T^0)^2} + \frac{\tau^2 MK(\gamma^0 T^0)^3}{2\rho B^2(t+T^0)(t-\tau+T^0)^2}.$$

Observe that it is not hard to check that if $t \geq 2\tau + 1$, then the inequality $(t - \tau + T^0)^2 > t + T^0$ holds and we can substitute $1/((t - \tau + T^0)^2)$ in (89) by the upper bound $1/(t + T^0)$. Applying this substitution yields

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*)\right] \\ &\quad \leq \left(1 - \frac{2m\gamma^0 T^0}{B(t+T^0)}\left[1 - \frac{\rho M}{2}\right]\right)\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] + \frac{MK(\gamma^0 T^0)^2}{2B(t+T^0)^2} + \frac{\tau^2 MK(\gamma^0 T^0)^3}{2\rho B^2(t+T^0)^2}. \qquad (90)\end{aligned}$$

We use the result in Lemma 3 to show sublinear convergence of the sequence of expected objective value error $\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right]$.

Lemma 3 shows that if a sequence u^t satisfies the condition in (54) then the sequence u^t converges to null at least with the rate of $\mathcal{O}(1/t)$. By assigning values $t^0 = T^0$, $u^t = \mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right]$, $c = (2m\gamma^0 T^0/B)(1 - \rho M/2)$, and $b = MK(\gamma^0 T^0)^2/2B + (\tau^2 MK(\gamma^0 T^0)^3)(2\rho B^2)$, the relation in (53) implies that the inequality in (54) is satisfied for the case that $2mr\gamma^0 T^0 > 1$. Therefore, the result in (55) holds and we can conclude that

$$\mathbb{E}\left[F(\mathbf{x}^t) - F(\mathbf{x}^*)\right] \leq \frac{C}{t + T^0}, \qquad (91)$$

where the constant C is defined as

$$C = \max\left\{\frac{MK(\gamma^0 T^0)^2/2B + (\tau^2 MK(\gamma^0 T^0)^3)(2\rho B^2)}{(2m\gamma^0 T^0/B)(1 - \rho M/2) - 1},\ T^0(F(\mathbf{x}^0) - F(\mathbf{x}^*))\right\}. \qquad (92)$$

■

References

[1] A. Mokhtari, A. Koppel, and A. Ribeiro, "Doubly random parallel stochastic methods for large scale learning," in *2016 American Control Conference (ACC)*, July 2016, pp. 4847–4852.

[2] A. Koppel, A. Mokhtari, and A. Ribeiro, "Doubly stochastic algorithms for large-scale optimization," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 1705–1709.

[3] A. Mokhtari, A. Koppel, and A. Ribeiro, "A class of parallel doubly stochastic algorithms for large-scale learning," *Journal of Machine Learning Research (submitted)*, 2016.

[4] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online kernel learning via sparse projections in function space," *Journal of Machine Learning Research (submitted)*, 2016. [Online]. Available: https://fling.seas.upenn.edu/~akoppel/assets/papers/kernel_report.pdf

[5] A. Koppel, J. Fink, G. Warnell, E. Stump, and A. Ribeiro, "Online learning for characterizing unknown environments in ground robotic vehicle models," in *Proc. Int. Conf. Intelligent Robots and Systems*.

[6] G. Sampson, R. Haigh, and E. Atwell, "Natural language analysis by stochastic optimization: A progress report on project april," *J. Exp. Theor. Artif. Intell.*, vol. 1, no. 4, pp. 271–287, Oct. 1990. [Online]. Available: http://dx.doi.org/10.1080/09528138908953710

[7] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[8] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "D4l: Decentralized dynamic discriminative dictionary learning," *IEEE Trans. Signal and Info. Process. over Networks*, vol. (to appear), June 2017, available at http://www.seas.upenn.edu/ aribeiro/wiki.

[9] M. Taşan, G. Musso, T. Hao, M. Vidal, C. A. MacRae, and F. P. Roth, "selecting causal genes from genome-wide association studies via functionally coherent subnetworks," *Nature methods*, 2014.

[10] Z. Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *Journal of Optimization Theory and Applications*, vol. 72, no. 1, pp. 7–35, 1992. [Online]. Available: http://dx.doi.org/10.1007/BF00939948

[11] P. Tseng and C. O. L. Mangasarian, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim Theory Appl*, pp. 475–494, 2001.

[12] Y. Xu and W. Yin, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *arXiv preprint arXiv:1410.1386*, 2014.

[13] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Mathematical Programming*, pp. 1–52, 2015.

[14] Z. Lu and L. Xiao, "On the complexity analysis of randomized block-coordinate descent methods," *Mathematical Programming*, pp. 1–28, 2013.

[15] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[16] A. Beck and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.

[17] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 285–322, 2015.

[18] Y. Yang, G. Scutari, and D. P. Palomar, "Parallel stochastic decomposition algorithms for multi-agent systems," in *Signal Processing Advances in Wireless Communications (SPAWC), 2013 IEEE 14th Workshop on*, June 2013, pp. 180–184.

[19] Z. Lu and L. Xiao, "On the complexity analysis of randomized block-coordinate descent methods," *Mathematical Programming*, vol. 152, no. 1-2, pp. 615–642, 2015.

[20] C. Scherrer, A. Tewari, M. Halappanavar, and D. Haglin, "Feature clustering for accelerating parallel coordinate descent," in *Advances in Neural Information Processing Systems*, 2012, pp. 28–36.

[21] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *Signal Processing, IEEE Transactions on*, vol. 63, no. 7, pp. 1874–1889, 2015.

[22] S. Shalev-Shwartz and T. Zhang, "Accelerated mini-batch stochastic dual coordinate ascent," in *Advances in Neural Information Processing Systems*, 2013, pp. 378–385.

[23] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 23.

[24] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177729586

[25] A. Koppel, B. Sadler, and A. Ribeiro, "Proximity without consensus in online multi-agent optimization," *IEEE Transactions on Signal Processing*, 2017.

[26] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *arXiv preprint arXiv:1309.2388*, 2013.

[27] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, 2013, pp. 315–323.

[28] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[29] N. N. Schraudolph, J. Yu, and S. Günter, "A stochastic quasi-newton method for online convex optimization," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 436–443.

[30] A. Bordes, L. Bottou, and P. Gallinari, "Sgd-qn: Careful quasi-newton stochastic gradient descent," *The Journal of Machine Learning Research*, vol. 10, pp. 1737–1754, 2009.

[31] A. Mokhtari and A. Ribeiro, "Res: Regularized stochastic bfgs algorithm," *Signal Processing, IEEE Transactions on*, vol. 62, no. 23, pp. 6089–6104, 2014.

[32] ——, "Global convergence of online limited memory bfgs," *Journal of Machine Learning Research*, vol. 16, pp. 3151–3181, 2015. [Online]. Available: http://jmlr.org/papers/v16/mokhtari15a.html

[33] Y. Xu and W. Yin, "Block stochastic gradient iteration for convex and nonconvex optimization," *SIAM Journal on Optimization*, vol. 25, no. 3, pp. 1686–1716, 2015.

[34] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2011, pp. 693–701.

[35] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2737–2745. [Online]. Available: http://papers.nips.cc/paper/5751-asynchronous-parallel-stochastic-gradient-for-nonconvex-optimization.pdf

[36] C. G. Broyden, J. E. D. Jr., Wang, and J. J. More, "On the local and superlinear convergence of quasi-newton methods," *IMA J. Appl. Math*, vol. 12, no. 3, pp. 223–245, June 1973.

[37] R. H. Byrd, J. Nocedal, and Y. Yuan, "Global convergence of a class of quasi-newton methods on convex problems," *SIAM J. Numer. Anal.*, vol. 24, no. 5, pp. 1171–1190, October 1987.

[38] J. J. E. Dennis and J. J. More, "A characterization of super linear convergence and its application to quasi-newton methods," *Mathematics of computation*, vol. 28, no. 126, pp. 549–560, 1974.

[39] D. H. Li and M. Fukushima, "A modified bfgs method and its global convergence in nonconvex minimization," *Journal of Computational and Applied Mathematics*, vol. 129, no. 1, pp. 15–35, 2001.

[40] A. Simonetto, A. Koppel, A. Mokhtari, G. Leus, and A. Ribeiro, "Decentralized Prediction-Correction Methods for Networked Time-Varying Convex Optimization," 2013, arXiv:1602.01716.

[41] L. Dong C. and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, no. 45(1-3), pp. 503–528, 1989.

[42] A. Nemirovski, A. Juditsky, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[43] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits." [Online]. Available: http://yann.lecun.com/exdb/mnist/

[44] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Englewood Cliffs: NJ: Prentice-Hall, 1995.