Stochastic Large-scale Machine Learning Algorithms with Distributed Features and Observations

Biyi Fang BIYIFANG2021@U.NORTHWESTERN.EDU
Department of Engineering Science and Applied Mathematics
Northwestern University
Evanston, IL 60208, USA

Diego Klabjan D-KLABJAN@NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

Editor:

Abstract

As the size of modern datasets exceeds the disk and memory capacities of a single computer, machine learning practitioners have resorted to parallel and distributed computing. Given that optimization is one of the pillars of machine learning and predictive modeling, distributed optimization methods have recently garnered ample attention, in particular when either observations or features are distributed, but not both. We propose a general stochastic algorithm where observations, features, and gradient components can be sampled in a double distributed setting, i.e., with both features and observations distributed. Very technical analyses establish convergence properties of the algorithm under different conditions on the learning rate (diminishing to zero or constant). Computational experiments in Spark demonstrate a superior performance of our algorithm versus a benchmark in early iterations of the algorithm, which is due to the stochastic components of the algorithm.
Keywords. optimization, machine learning, stochasticity, convexity, large scale

1. Introduction

As technology advances, collecting and analyzing large-scale and real time data has become widely used in a variety of fields. Large-scale machine learning can not only present a useful summary of a dataset but it can also make predictions. In the era of big data, large scale datasets have become more accessible. "Large" usually refers to both the number of observations and high feature dimension. For example, an English Wikipedia dataset can have 11 million documents (observations) and over several hundred of thousands unique word types (features). This demands a sophisticated large-scale machine learning system able to take advantages of all available information from such datasets and not only random samples. On the other hand, as large scale data is becoming more accessible, storing the whole dataset on a single server is often impossible due to the inadequate disk and memory capacities. Consequently, it is considerable to store and analyze them distributively. Often the data collection process by design distributes observations and features. There is a large

amount of literature dealing with optimization problems subject to a dataset with either distributed observations or distributed features. Nevertheless, very limited contribution has been made to the case where both the observations and features are in a distributed environment.

In this paper, we propose an algorithm, namely SODDA (StOchastic Doubly Distributed Algorithm), designed for a doubly distributed dataset and inspired by the work Harikandeh et al. (2015) and Nathan and Klabjan (2017). The algorithm is aimed to solve a series of optimization problems which can be formulated as the minimization of a finite sum of convex functions plus a convex regularization term if necessary. SODDA is a primal method building on the previous RAndom Distributed Stochastic Algorithm (RADiSA) Nathan and Klabjan (2017). SODDA first further splits the partitions (a partition is a set of features and observations stored locally) with respect to features into sub-partitions with no overlap; then in each iteration, randomly chooses sub-partitions associated with different blocks of features; lastly, similar to stochastic gradient descent (SGD), updates in parallel each sub-block of the current local solution by using observations from the randomly selected sub-partition of local observations and the local sub-block of features, coupled with the Stochastic Variance-Reduced Gradient (SVRG). One generalization of SVRG utilized in SODDA is that SODDA does not require a full solution update; instead, it allows each sub-block of the current local solution to be updated individually and assembled at the end of each iteration. Although, we might amplify the error by approximately computing the gradient, SODDA reduces the communication cost significantly. Another technique aiming to cut down the communication cost is estimating the full gradient needed as part of the SVRG component, which is a big distinction between SODDA and RADiSA. RADiSA requires the full gradient in each outer iteration, which is computationally demanding, especially when the solution is far from an optimal solution. SODDA has three stochastic components: the first two are that it randomly selects blocks of local features and subsets of local observations to execute the estimated gradient, and the third component that randomly chooses further sub-blocks of local features to record the approximated gradient, which contributes to a reduction of the number of gradient coordinate computations required in early iterations. In other words, only random coordinates of the gradient are computed.

In this paper, we not only propose a more computationally efficient method, SODDA, when compared to RADiSA Nathan and Klabjan (2017), but also present a complete technical proof of convergence. For a smooth and strongly convex function, we prove that SODDA enjoys at least a sublinear convergence rate and a linear convergence rate for a diminishing learning rate and a constant learning rate, respectively. Furthermore, we prove that SODDA iterates converge to an optimal solution when using a constant learning rate selected from a certain interval. Moreover, the convergence property of RADiSA, which is not provided in Nathan and Klabjan (2017), is implied directly from SODDA. In summary, we make the following five contributions.

- We provide a better scalable stochastic doubly distributed method, i.e. SODDA, for doubly distributed datasets. This algorithm does not require the calculation of a full gradient, thus it is a less computationally intensive methodology for doubly distributed setting problems.

- We provide a proof of a sublinear convergence result for smooth and strongly-convex loss functions when using a sequence of decreasing learning rates.
- We show that SODDA iterates converge with linear rate to a neighborhood of an optimal solution when using an arbitrary constant learning rate.
- We further argue that SODDA iterates converge to an optimal solution in the strongly-convex case when using any constant learning rate in a specified interval.
- We present numerical results showing that SODDA outperforms RADiSA-avg, which is the best doubly distributed optimization algorithm in Nathan and Klabjan (2017), on all instances considered in early iterations. More precisely, SODDA finds good quality solutions faster than RADiSA-avg.

The paper is organized as follow. In the next section, we review several related works in distributed optimization. In Section 3, we state the formal optimization problem and standard assumptions underlying our analyses, followed by the exposition of the SODDA algorithm. In Section 4, we show the convergence analyses of SODDA with respect to both a decreasing learning rate and a constant learning rate. In Section 5, we present experimental results comparing SODDA with RADiSA-avg.

2. Related Work

There are a large number of extensions of the plain stochastic gradient descent algorithm related to distributed datasets, however, a full retrospection of this immense literature exceeds the scope of this work. In this section, we state several approaches which are most related to our new method and interpret the relationships among them.

In plain SGD, the gradient of the aggregate function is approximated by one randomly picked function Robbins and Monro (1951). It saves a heavy load of computation when compared with gradient descent, whereas more often than not, the convergence happens to be slow. Recently, a large variety of approaches have been proposed targeted on accelerating the convergence rate and dealing with observations in the distributed setting.

SGD for distributed observations: One attempt that works for datasets with distributed observations is parallelizing it by means of mini-batch SGD. Both the synchronous version Chen et al. (2016) and the asynchronous version Tsitsiklis et al. (1986) basically work in the following way: the parameter server performs parameter updates after all worker nodes send their own gradients based on local information in parallel, and then broadcasts the updated parameters to all worker nodes afterwards. In the synchronous approach, the master node needs to wait until all gradients are collected but in the asynchronous approach, the master node performs updates whenever it is needed. An alternative method which introduces the concept of variance reduced is CentralVR De and Goldstein (2016), where the master node not only needs to spread parameters but also the full gradient after every certain number of iterations, and each worker node would involve the full gradient as a corrector when computing their own gradients. As a consequence, the variance in the estimation of the stochastic gradient could be reduced and a larger learning rate is allowed to accomplish faster convergence and higher accuracy.

SGD for distributed features: Another attempt for distributed features is parallelization via features. Block successive upper bound minimization (BSUM)Hong et al. (2015) is one of the methods working for datasets with distributed features, where the master node spreads all parameters and each worker node conducts parameter updates on a randomly chosen and non-overlapping subset of the feature vector. Distributed Block Coordinate Descent Mareček et al. (2015) is another approach designed for datasets with distributed features. The parameters associated with these feature blocks are partitioned accordingly. In the algorithm, each processor randomly chooses a certain number of blocks out of those stored locally, performs the corresponding parameter updates in parallel, and then transmits to other processors. However, it is impossible to avoid communication when computing the gradient of all parameters unless there are extra assumptions on the objective loss function, which does not usually hold. An alternative approach is Communication-Efficient Distributed Dual Coordinate Ascent (CoCoA) Jaggi et al. (2014), which is a primal-dual method also working for data with distributed features. By exploiting the fact that the associated blocks of dual variables work in different processors without overlap, the algorithm aggregates the parallel updates efficiently from the different processors without much conflict and reduces the necessary communication dramatically. A faster converging method extended from the aforementioned approach is CoCoA$^+$ Ma et al. (2015), which allows a larger learning rate for parameter updates by introducing a more generalized local CoCoA subproblem at each processor.

SGD for distributed observations and features: Given datasets with distributed observations and features, all the methods mentioned so far are not applicable. One of the algorithms fitting the bill is a block distributed ADMM Parikh and Boyd (2014), which is the block splitting variant of ADMM. Nonetheless, the convergence rate of ADMM-based methods is slow. Random parallel stochastic algorithm (RAPSA)Mokhtari et al. (2016) is another algorithm which utilizes multiple parallel units to operate on a randomly chosen subset of blocks of the feature vector. It needs to access the whole feature vector to perform a parameter update while SODDA only needs a subset of the feature. Decentralized double stochastic averaging gradient algorithm (DSA) Mokhtari and Ribeiro (2016) is an alternative method designed for doubly distributed datasets, whereas the global cost function that DSA optimizes is a linear combination of the local objective functions which only contain local parameters, compared to the loss function of SODDA which contains global parameters. A faster converging and more pertinent algorithm is RADiSA Nathan and Klabjan (2017), which is also focusing on settings where both the observations and features of the problem at hand are stored in a distributed fashion. RADiSA conducts parameter updates based on stochastic partial gradients, which are calculated from randomly selected local observations and a randomly assigned sub-block of local features in parallel. RADiSA is a special case of SODDA, since the full gradient required by it is replaced by an approximated gradient which only uses partial observations and features. Consequently, SODDA provides a faster convergence than RADiSA without sacrificing too much accuracy. Meanwhile, we present technical convergence analyses for SODDA under different types of learning rates which imply the convergence of RADiSA.

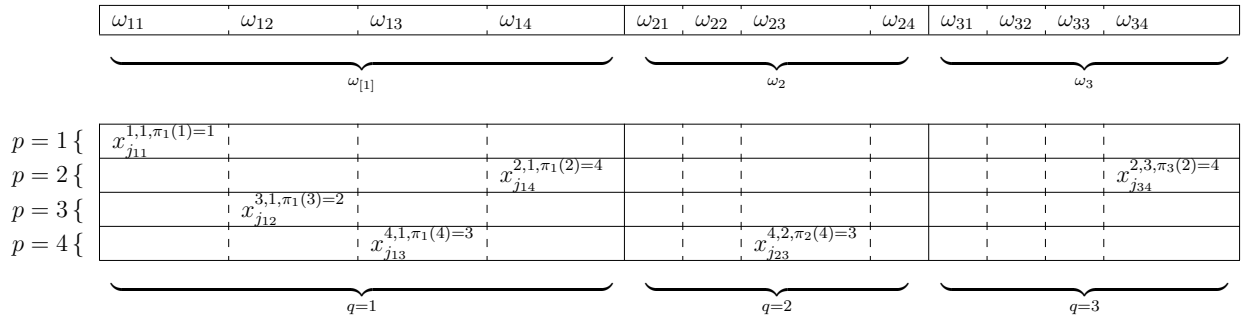


Figure 1: $Q = 3$, $P = 4$

3. Algorithm

We consider the problem of optimizing a finite but large sum of smooth functions, i.e., given a training set $\{(x_i, y_i)\}_{i=1}^N$, where each $x_i^T \in \mathbb{R}^d$ is associated with a corresponding label y_i,

$$\min_{\omega \in \mathbb{R}^d} F(\omega) := \frac{1}{N} \sum_{i=1}^{N} \bar{f}(x_i\omega, y_i) = \frac{1}{N} \sum_{i=1}^{N} f_i(x_i\omega). \tag{1}$$

Several machine learning loss functions fit this model, e.g. least square, logistic regression, and hinge loss.

In SODDA, we assume that the training set $\{(x_i, y_i)\}_{i=1}^N$ is distributed across both observations and features. More specifically, the features and observations are split into Q and P partitions, respectively. We denote the matrix corresponding to the p'th observation partition and its q'th feature partition as $x^{p,q} \in \mathbb{R}^{\frac{N}{P} \times \frac{d}{Q}}$. Note that the partitions consisting of same features share the common block of parameters $\omega_{[q]}$. In Figure 1, there are 12 partitions. Parameters $\omega_{[1]}$ correspond to all parameters under $q = 1$. In order to efficiently parallelize the computation, optimization of each partition can be done concurrently by considering only local observations and features. This strategy poses a big challenge in how to combine the parameters. For example, $\omega_{[1]}$ are modified by all processors working on $x^{1,1}$, $x^{2,1}$, $x^{3,1}$, $x^{4,1}$. It is unclear how to combine them (averaging them is a possible strategy however this would not yield convergence, see e.g. Weimer et al. (2010), Zinkevich et al. (2010)). To circumvent this, we further artificially subdivide the features.

To this end, we define a function $\pi_q(p) : \{1, 2, \cdots, P\} \to \{1, 2, \cdots, P\}$ in the corresponding q'th feature partition, where P is the number of partitions for observations. A submatrix of the training set from the partition $x^{p,q}$ corresponding to block $\omega_{q,\pi_q(p)}$ is denoted by $x^{p,q,\pi_q(p)}$, see Figure 1. Each processor operates on random observations from partition $x^{p,q}$ and all feature in $x^{p,q,\pi_q(p)}$. Note that $\omega_{[q]} = \left(\omega_{q,\pi_q(p)}\right)_{p=1}^{P}$. Let us define $n = N/P$, $m = d/Q$, $\tilde{m} = d/QP$, and $j_{q,\pi_q(p)}$ be randomly chosen from $\{1, 2, \cdots, n\}$ associated with sub-block $x^{p,q,\pi_q(p)}$. In Figure 1, where $P = 3$ and $Q = 4$, $x_{j_{12}}^{3,1,\pi_1(3)=2} \in \mathrm{R}^{\tilde{m}}$ represents a random observation j_{12} with a subset of features selected from the 2nd sub-block of the block corresponding to the observation partition 3 and feature partition 1 given $\pi_1(3) = 2$. Similarly, $x_{j_{23}}^{4,2,\pi_2(4)=3}$ symbolizes a random observation j_{23} with a subset of features selected from the 3rd sub-block of the block $x^{4,2}$.

Next, we introduce notation for the partial gradient. For any $\mathcal{C} \subseteq \{1, \cdots, d\}$ and any j, let us denote $\bar{\nabla}_{w_\mathcal{C}} f_j(\cdot) \in \mathbb{R}^d$ as the vector defined by

$$\left(\bar{\nabla}_{\omega_\mathcal{C}} f_j(\cdot)\right)_k = \begin{cases} 0, & k \notin \mathcal{C} \\ \left(\nabla f_j(\cdot)\right)_k, & k \in \mathcal{C}. \end{cases}$$

We need this notation since we sample gradient components. The loss function using this notation becomes

$$F(\omega) = \frac{1}{N} \sum_{k=1}^{P} \sum_{j=1}^{n} f_j^k \left(\sum_{q=1}^{Q} \sum_{p=1}^{P} x_j^{k,q,\pi_q(p)} \omega_{q,\pi_q(p)} \right),$$

where f_j^k is $\bar{f}$ associated with observation j in observation partition k.

Given the fact that the data is doubly distributed, SODDA further divides the features $x^{\cdot,q}$ into P subsets along all observations, i.e. $x^{\cdot,q} = [x^{\cdot,q,1}, \cdots, x^{\cdot,q,P}]$. In each iteration, SODDA first computes an approximation of the full gradient at the current parameter vector. Then, after randomly choosing a sub-matrix from each matrix $x^{q,p}$ as long as there is no overlap with respect to ω, each processor is assigned a sub-matrix of the local dataset and updates its local parameter by employing generalized SVRG. In the end of each iteration, SODDA concatenates all partial parameters which becomes the incumbent parameter vector for the next iteration.

The entire algorithm is exhibited in Algorithm 1. Steps 1- 3 initiate all the parameters. Steps 5-7 give the subsets of features and observations used to compute the partial gradient of the current iterate $\tilde{\omega}$ in step 8. Since the dataset is doubly distributed, the algorithm computes an estimate of the exact full gradient so as to reduce the communication cost in step 8. Additionally, we use the term no feature sampling to address the case when $\mathcal{b}^t = d$; in other words, the whole feature vector is employed to compute the gradient μ^t. To this end, we have three random components. The first one is the common one to sample observations. The second one is to compute only a random subset of subgradient coordinates, and the third one is to evaluate these components not at the exact $x\omega$ but only on the subset of the underlying inner product summation terms. Step 10 determines how sub-blocks are selected in each block of the dataset associated with $\omega_{[q]}$, for each q. The definition of $(\pi_q)_{q=1}^{Q}$ guarantees that one, and only one sub-block is selected with respect to $\omega_{q,\pi_q(p)}$, i.e. $x^{p,q,\pi_q(p)}$. Then, for each sub-block $x^{p,q,\pi_q(p)}$, after randomly picking an observation $x^{p,q,\pi_q(p)}_{j_{q,\pi_q(p)}}$ in the selected sub-block in step 15, each block of parameter updates is given in step 16. Instead of using the full vector, we estimate the stochastic gradient by using local features and narrow down the variance by involving the approximated full gradient. Finally, at the end of each iteration, after each processor finishes its own task, step 19 aggregates all the updated partial solutions, i.e. $\omega = \left[\omega_{[1]}, \omega_{[2]}, \cdots, \omega_{[Q]}\right]$, where partial parameters $\omega_{[q]}$ represent the concatenation of the local parameters $\omega_{q,\pi_q(p)}$, for $p = 1, 2, \cdots, P$.

Algorithm 1 SODDA

Inputs:
batch size B, learning rate γ_t, sequence $\left\{\mathcal{b}^t, c^t, \mathcal{d}^t\right\}_{t=0}^{\infty}$ where $c^t \leq \mathcal{b}^t \leq d$, $\mathcal{d}^t \leq N$ for every t
Data:
$x^{p,q,k} \in \mathrm{R}^{n\times\tilde{m}}$ for $p,k = 1,\cdots,P$ and $q = 1,\cdots,Q$, $\omega_{q,\pi_q(p)} \in \mathrm{R}^{\tilde{m}}$
Initialize:
$w^0 \leftarrow 0$
for t $= 0,1,2,\cdots$ **do**
  $\mathcal{B}^t$=$\mathcal{b}^t$ elements uniformly at random sampled without replacement from all features
  $\mathcal{C}^t$=c^t elements uniformly at random sampled without replacement from $\mathcal{B}^t$
  $\mathcal{D}^t$=$\mathcal{d}^t$ elements uniformly at random sampled without replacement from all observations
  $\mu^t = \frac{1}{\mathcal{d}^t}\sum_{j\in\mathcal{D}^t} \bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j^{\mathcal{B}^t}\omega_{\mathcal{B}^t}^t)$
  for $q = 1,2,\cdots,Q$, **do**
    select function $(\pi_q)_{q=1}^{Q}$
  end for
  for $p = 1,2,\cdots,P$ and $q = 1,2,\cdots,Q$, **do in parallel**
    $\bar{\omega}_{q,\pi_q(p)}^{(0)} = \omega_{q,\pi_q(p)}^t$
    for $i = 0,\cdots,B-1$ **do**
      randomly pick $j_{q,\pi_q(p)} \in \{1,\cdots,n\}$
      $\bar{\omega}_{q,\pi_q(p)}^{(i+1)} = \bar{\omega}_{q,\pi_q(p)}^{(i)} - \gamma_{t+1}\left[\nabla_{\omega_{q,\pi_q(p)}} f_{j_{q,\pi_q(p)}}^{\pi_q(p)}(x_{j_{q,\pi_q(p)}}^{p,q,\pi_q(p)}\bar{\omega}_{q,\pi_q(p)}^{(i)}) - \nabla_{\omega_{q,\pi_q(p)}} f_{j_{q,\pi_q(p)}}^{\pi_q(p)}(x_{j_{q,\pi_q(p)}}^{p,q,\pi_q(p)} w_{q,\pi_q(p)}^t) + \mu_{q,\pi_q(p)}^t\right]$
    end for
  end for
  $\omega^{t+1} = \left[\omega_{[1]},\omega_{[2]},\cdots,\omega_{[Q]}\right]$, where $\omega_{[q]} = \left[\bar{\omega}_{q1}^{(B)},\bar{\omega}_{q2}^{(B)},\cdots,\bar{\omega}_{qP}^{(B)}\right]$
end for

4. Analysis

In this section, we prove that the sequence of the loss function values $F(\omega^t)$ generated by SODDA approaches the optimal loss function value $F(\omega^*)$. We assume the existence and the uniqueness of the minimizer ω^* that achieves the optimal loss function value. Meanwhile, we require the following standard assumptions.

Assumption 1:

- Functions $f_i(x_i\omega)$ are differentiable with respect to ω for every $i = 1,\cdots,N$.
- For every $i = 1,\cdots,N$, the norm of the gradient $\nabla f_i(x_i\omega)$ is bounded for all ω, more precisely, there exists a constant M_1, such that, for any ω,

$$\|\nabla f_i(x_i\omega)\| \leq M_1.$$

Assumption 2:

- The expectation function $F(\omega)$ is strongly convex with parameter $\xi > 0$.

Assumption 3:

- The loss gradients $\nabla f_i(x_i\omega)$ are Lipschitz continuous with respect to the Euclidian norm with parameter $L \geq 1$, i.e., for all ω, $\hat{\omega} \in \mathbb{R}^d$ and any i, it holds

$$\|\nabla f_i(x_i\omega) - \nabla f_i(x_i\hat{\omega})\| \leq L\|\omega - \hat{\omega}\|.$$

The restriction imposed by Assumption 1 provides an upper bound to the first gradient of each data point, which is a standard condition in stochastic approximation literature Robbins and Monro (1951). Its intent is to limit the variance of the stochastic gradients Nemirovski et al. (2009). In Assumption 2, only the expected loss function $F(\omega)$ is enforced to be strongly convex, whereas the individual loss functions f_i could even be non-convex. Notice that in Assumption 3, since each individual function ∇f_i is imposed to be Lipschitz-continuous with constant L with respect to ω, both the gradient of the expected loss function $\nabla F(\omega)$ and the individual function ∇f_i are L-Lipschitz continuous with respect to ω and $\omega_{q,\pi_q(p)}$ for any $q \in \{1, 2, \cdots, Q\}$ and any $p \in \{1, 2, \cdots, P\}$. Note that if f_i's are ϵ-Lipschitz continuous for some $0 < \epsilon \leq 1$, we can take $L = 1$. Moreover, these assumptions hold for several widely used machine learning loss functions, i.e. hinge, square, logistic loss.

Under these standard assumptions, by finding a relationship for the sequence of the loss function errors $F(\omega^t) - F(\omega^*)$ and employing the supermartingale convergence argument, which is a standard technique for analyzing stochastic optimization problems (see e.g. textbooks Benveniste et al. (2012), Bertsekas and Tsitsiklis (1989), Borkar (2008)), we prove that the sequence of the loss function values $F(\omega^t)$ converges to the optimal function value $F(\omega^*)$ almost surely when using the standard diminishing learning rate, i.e. non-summable and squared summable. Consequently, the sequence of ω^t enjoys the almost sure convergence to ω^* when taking Assumption 2 into consideration.

Theorem 1 *If there is no feature sampling in step 5 and Assumptions 1-3 hold, and the sequence of learning rates are non-summable* $\sum_{t=1}^{\infty} \gamma_t = \infty$ *and square summable* $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$*, and the sequence* $(c^t, \mathscr{d}^t)_{t=0}^{\infty}$ *is selected so that* $c^t \leq d$ *and* $\mathscr{d}^t \leq N$*, then the sequence of parameters* ω^t *generated by SODDA converges almost surely to the optimal solution* ω^**, that is*

$$\lim_{t \to \infty} \left\| \omega^t - \omega^* \right\| = 0 \quad \text{a.s.} \tag{2}$$

Proof See Appendix C. ■

Based on the fact that the exact form for the update step in expectation is not available in SODDA, i.e., we do not explicitly know $\omega^{t+1} - \omega^t$ any technique that relies on such an explicit formula is inappropriate. This expectation is given by steps 9-18 in the algorithm. The main challenges are coming from evaluating individual function gradient $\nabla f(x_{qp}\omega_{q,\pi_q(p)})$ and grouping different sub-blocks all together. The way we deal with it, which is borrowed from Bertsekas and Tsitsiklis (2000), is grouping the first two partial gradients together and treating the last partial gradient $\mu^t_{q,\pi_q(p)}$ as a corrector.

The very technical proof follows the following steps. In steps 14-17, SODDA utilizes local features from a random observation to update the corresponding subset of parameters, and involves the information from the estimated full gradient to reduce unreasonable fluctuation. Therefore, in association with the conditional Jensen's inequality and properties of Lipschitz continuity, the norm of the difference and the square norm of the difference of the first two terms in the bracket of the updating procedure in step 16 are able to be bounded by a function involving γ_t. Thus, representing ω^{t+1} as a function of ω^t and applying strong convexity of F, coupled with all the bounds derived before, yield a supermartingale relationship for the

sequence of loss function errors $F(\omega^t) - F(\omega^*)$. Combined with the property that γ_t is non-summable but square summable, (7) is achieved by applying the supermartingale convergence theorem.

Theorem 1 asserts the almost sure convergence of the iterates generated by SODDA with non-summable and squared summable learning rate. Furthermore, given $\gamma_t = \frac{1}{t}$ and B big enough, the following theorem states that the loss function $F(\omega^t)$ converges to the optimal value $F(\omega^*)$ with probability 1 and the rate of convergence in expectation is at least in the order of $\mathcal{O}(\frac{1}{t})$.

Theorem 2 *Under Assumptions 1-3 with no feature sampling, if the learning rate is defined as $\gamma_t := \frac{1}{t}$ for $t = 1, 2, \cdots$, and the batch size is chosen such that $B \geq \frac{d}{2\xi}$, and the sequence $(c^t, \mathscr{d}^t)_{t=0}^{\infty}$ satisfies $c^t \leq d$ and $\mathscr{d}^t \leq N$, then there exists a positive constant C_1 such that the expected loss function errors $\mathbb{E}[F(\omega^t) - F(\omega^*)]$ of SODDA converge to 0 at least with a sublinear convergence rate of order $\mathcal{O}(1/t)$, i.e.*

$$\mathbb{E}[F(\omega^t) - F(\omega^*)] \leq \frac{Q}{1+t}, \tag{3}$$

where constant Q is defined as

$$Q = \max\left\{F(\omega^0) - F(\omega^*), \cdots, ([\lambda]+2)\,\mathbb{E}[F(\omega^{[\lambda]+1}) - F(\omega^*)], \frac{C_1}{\lambda - 1}\right\}, \tag{4}$$

with $\lambda = \frac{2\xi B}{d}$.

Proof See Appendix C ■

Given the specific relationship between the learning rate γ_t and the iterator t, i.e. $\gamma_t = 1/t$, applying the supermartingale convergence theorem and performing induction on an upper bound of $\mathbb{E}\left[F(\omega^t) - F(\omega^*)\right]$ allow us to establish at least sublinear convergence of SODDA.

A diminishing learning rate is beneficial if the exact convergence is required. If we are only interested in a specific accuracy, it is more efficient to choose a constant learning rate. In the following theorem, we employ a similar argument used in proving Theorem 1 and Theorem 2 except that B and γ are linked by a condition. Again by providing a supermartingale relationship for the sequence of the loss function errors $F(\omega^t) - F(\omega^*)$, we are able to study the convergence properties generated by SODDA for a constant learning rate γ.

Theorem 3 *If there is no feature sampling in step 5 and Assumptions 1-3 hold, and the learning rate is constant $\gamma_t = \gamma$ such that $BL\gamma QP \leq 1$, which also implies that $\gamma \leq 1$, and the sequence $(c^t, \mathscr{d}^t)_{t=0}^{\infty}$ satisfies $c^t \leq d$ and $\mathscr{d}^t \leq N$, then there exists a positive constant C_2 such that the sequence of parameters ω^t generated by SODDA converges almost surely to a neighborhood of the optimal solution ω^*, that is*

$$\liminf_{t\to\infty} F(\omega^t) - F(\omega^*) \leq \frac{C_2 d B^3 \gamma}{2\xi} \quad \text{a.s.} \tag{5}$$

Moreover, if the constant learning rate γ is chosen such that $\gamma < \min\left\{\frac{d}{2\xi B}, \frac{1}{BLQP}, 1\right\}$, then the expected loss function errors $\mathbb{E}\left[F(\omega^t) - F(\omega^*)\right]$ converge linearly to an error bound as

$$\mathbb{E}\left[F(\omega^t) - F(\omega^*)\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right)^t \left(F(\omega^0) - F(\omega^*)\right) + \frac{C_2 d B^3 \gamma}{2\xi}. \tag{6}$$

Proof See Appendix D. ■

Note that $BL\gamma QP \leq 1$ trades off γ and B, i.e. the larger B is, the smaller γ must be. The major difficulties are similar but not identical to those of Theorem 1. We apply a similar idea but treating both B and $\gamma_t = \gamma$ as variables to obtain an upper bound in closed form for the difference of the first two partial gradients in step 16. Then Lipschitz continuity of $\nabla F(\omega)$ leads to a supermartingale relationship for the sequence of the loss function errors $F(\omega^t) - F(\omega^*)$. As a consequence, claims in (10) and (11) follow according to the supermartingale convergence theorem. The only distinction between Theorem 1 and Theorem 3 is caused by the property of the learning rate. The error exists in each iteration, which is a function of the learning rate γ_t, however, in Theorem 5, the error function goes to 0 as the number of iterations increases, which is not the case when the learning rate is a constant. Therefore, we can only ensure a relatively high-quality solution.

In order to allow feature sampling we have to control the growth of ω^t.

4.1 Analyses with Feature Sampling

To this end, we require the following assumptions together with Assumptions 2 and 3. In this subsection, we also do not require $\mathcal{b}^t = d$, i.e., step 5 in Algorithm 1 now requires sampling.

Assumption 4:

- There exists a constant M_2, such that

$$\left\|\omega^t\right\| \leq \frac{M_2}{2},$$

  for any t.

Assumption 5:

- The sample variance of the norms of the gradients is bounded by G^2 for all ω^t, i.e.

$$\frac{1}{N-1}\sum_{j=1}^{N}\left(\left\|\nabla f_j(x_j w^t)\right\|^2 - \left\|\nabla F(\omega^t)\right\|^2\right) \leq G^2.$$

The restriction in Assumption 4 is reasonable and also has been used inworkHarikandeh et al. (2015). Assumption 5 is also standard, see e.g. Harikandeh et al. (2015). Moreover, these assumptions hold for several widely used machine learning loss functions, i.e. hinge, square, logistic loss. Notice that without Assumption 4, we can not further assume the boundness of the sample variance of the norms of the gradients in Assumption 5. Next, we present an example showing that SODDA does not converge under Assumptions 2 and 3.

Theorem 4 *There is a convex loss function and $\mathcal{P}$ where SODDA does not converge when only given Assumptions 2 and 3, and any $\gamma_t \leq K$ for every t and constant K depending on input data.*

Proof See Appendix E ■

The main role of Assumption 4 is to maintain a reasonable error generated by the stochastic partial gradients in steps 16. Then, under these standard assumptions and applying the similar trick as in the proof of Theorem 1, we argue that the sequence of ω^t enjoys the almost sure convergence to ω^*.

Theorem 5 *If Assumptions 2-5 hold, and the sequence of learning rates are non-summable $\sum_{t=1}^{\infty} \gamma_t = \infty$ and square summable $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$, and the sequences $(\mathcal{b}^t, c^t, \mathcal{d}^t)_{t=0}^{\infty}$ are selected so that $\mathcal{b}^t \in \left[\max\left\{c^t, \frac{d}{1+\frac{4d\eta\gamma_{t+1}^2}{c^t M_2^2 L^2}}\right\}, d\right]$ for some constant $\eta \geq 0$, $c^t \leq d$ and $\mathcal{d}^t \leq N$, then the sequence of parameters ω^t generated by SODDA converges almost surely to the optimal solution ω^*, that is*

$$\lim_{t\to\infty} \left\|\omega^t - \omega^*\right\| = 0 \quad \text{a.s.} \tag{7}$$

Proof See Appendix F. ■

The proof of Theorem 5 is very similar to the proof of Theorem 1. The only difference is caused by feature sampling. The main challenges are how to pick a suitable $\mathcal{b}^t$ and how to narrow down the error generated by $\mathcal{b}^t$. Then, given an appropriate $\mathcal{b}^t$, the norm of the estimator of the full gradient μ^t and its square are bounded by a function containing the full gradient at ω^t and the learning rate γ_t. Meanwhile, η is a positive constant which controls the divergence of the approximate full gradient from the exact full gradient in step 8, i.e. when η is 0, the whole feature vector is used as $\mathcal{b}^t = d$. The rest of the proof is identical to the proof of Theorem 1.

Theorem 5 asserts the almost sure convergence of the iterates generated by SODDA with non-summable and squared summable learning rate. Furthermore, given $\gamma_t = \frac{1}{t}$ and B big enough, the following theorem states that the loss function $F(\omega^t)$ converges to the optimal value $F(\omega^*)$ with probability 1 and the rate of convergence in expectation is at least in the order of $\mathcal{O}(\frac{1}{t})$.

Theorem 6 *Under Assumptions 2-5, if the learning rate is defined as $\gamma_t := \frac{1}{t}$ for $t = 1, 2, \cdots$, and the batch size is chosen such that $B \geq \frac{d}{2\xi}$, and the sequence $(\mathcal{b}^t, c^t, \mathcal{d}^t)_{t=0}^{\infty}$ satisfies the same conditions as in Theorem 5, then there exists a positive constant C_3 such that the expected loss function errors $\mathbb{E}[F(\omega^t) - F(\omega^*)]$ of SODDA converges to 0 at least with a sublinear convergence rate of order $\mathcal{O}(1/t)$, i.e.*

$$\mathbb{E}[F(\omega^t) - F(\omega^*)] \leq \frac{Q}{1+t}, \tag{8}$$

where constant Q is defined as

$$Q = \max\left\{F(\omega^0) - F(\omega^*), \cdots, ([\lambda]+2)\,\mathbb{E}[F(\omega^{[\lambda]+1}) - F(\omega^*)], \frac{C_3}{\lambda - 1}\right\}, \tag{9}$$

with $\lambda = \frac{2\xi B}{d}$.

Proof See Appendix F. ■

A diminishing learning rate is beneficial if the exact convergence is required. If we are only interested in a specific accuracy, it is more efficient to choose a constant learning rate.

Theorem 7 *If Assumptions 2-5 hold, and the learning rate is constant $\gamma_t = \gamma$ such that $BL\gamma QP \leq 1$, which also implies that $\gamma \leq 1$, and the sequence $(\mathcal{b}^t, c^t, \mathcal{d}^t)_{t=0}^{\infty}$ satisfies the same conditions as in Theorem 5, then there exists a positive constant C_4 such that the sequence of parameters ω^t generated by SODDA converges almost surely to a neighborhood of the optimal solution ω^*, that is*

$$\liminf_{t\to\infty} F(\omega^t) - F(\omega^*) \leq \frac{C_4 d B^3 \gamma}{2\xi} \quad \text{a.s.} \tag{10}$$

Moreover, if the constant learning rate γ is chosen such that $\gamma < \min\left\{\frac{d}{2\xi B}, \frac{1}{BLQP}, 1\right\}$, then the expected loss function errors $\mathbb{E}\left[F(\omega^t) - F(\omega^*)\right]$ converges linearly to an error bound as

$$\mathbb{E}\left[F(\omega^t) - F(\omega^*)\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right)^t \left(F(\omega^0) - F(\omega^*)\right) + \frac{C_4 d B^3 \gamma}{2\xi}. \tag{11}$$

Proof See Appendix G. ■

Theorem 7 guarantees that SODDA finds good quality solutions when using an appropriate learning rate γ and batch size B. Notice that although methods of type SVRG achieve linear convergence to the exact solution in expectation under a constant step size, SVRG performs the exact full gradient after every certain number of iterations which would trigger emergence of communication under the doubly distributed setting and is unnecessary especially in early iterations. In addition, based on (11), there is a trade-off between the accuracy and the convergence rate. Although reducing the learning rate γ or batch size B narrows down the error bound $\frac{C_4 d B^3 \gamma}{2\xi}$ and contributes significantly to a more accurate convergence, the constant convergence rate $1 - \frac{2\xi B}{d}\gamma$ suffers greatly since it increases and gets closer to 1, which leads to a slower convergence rate.

To address this problem, in the following theorem, by considering not only the loss function errors $F(\omega^t) - F(\omega^*)$ but also the errors $\|\omega^t - \omega^*\|^2$, we prove that the sequence of the loss function values $F(\omega^t)$ generated by SODDA converges to the optimal value $F(\omega^*)$ for any constant learning rate selected from a certain region. In addition, we are able to further assert that the sequence of ω^t converges to ω^* when taking Assumption 2 into account. Furthermore, since we employ an approximation of the exact full gradient for the sake of the

efficiency of the algorithm in Theorem 7, the algorithm converges only to a neighborhood of an optimal solution under a constant learning rate. In the following theorem, if we are allowed to employ the exact full gradient in expectation, then the algorithm in Theorem 8 converges to the exact solution in expectation under a constant step size.

Theorem 8 *If Assumptions 2-5 hold, and the learning rate $\gamma_t = \gamma$ is a constant such that $\gamma \in (0, \min\left\{1, \frac{1}{BLQP}, \gamma_1, \gamma_2\right\})$, where both γ_1 and γ_2 are positive constants specified in Appendix H, and the sequence $(\mathcal{b}^t, c^t, \mathcal{d}^t)_{t=0}^{\infty} = (d, c^t, N)_{t=0}^{\infty}$ for arbitrary positive $c^t \leq d$, then the sequence of parameters ω^t generated by SODDA converges to ω^*, that is*

$$\lim_{t\to\infty} \left\|\omega^t - \omega^*\right\| = 0. \tag{12}$$

Proof See Appendix H. ■

The Lyapunov analysis, which is a common strategy to deal with a constant learning rate (see e.g. Schmidt et al. (2017)), fails for our algorithm due to the analogous reasons as those for Theorem 5. The success of the Lyapunov analysis heavily relies on the number of negative terms available when computing the loss function errors $F(\omega^t) - F(\omega^*)$ and the errors $\omega^t - \omega^*$. Unfortunately, the doubly distributed data setting results in lack of information in each iteration in step 16, which leads to a scarcity of negative terms to ensure the decrease of the loss function value.

Our steps to study the convergence analysis are as follows. We first establish either exact forms or upper bounds for all terms involving gradients. Then, from the update rule, we find a criteria for the constant learning rate γ so as to make the errors $\omega^t - \omega^*$ at least not increase as the number of iterations increases. In addition, given Lipschitz continuity of $\nabla F(\omega)$, we find a recursive formula regarding the loss function error, which provides another constraint for γ such that the loss function error vanishes as t increases. Finally, the convergence of SODDA and the existence of γ are guaranteed by two cubic inequality constraints aforementioned.

5. Numerical Study

In this section, we compare the SODDA method with RADiSA-avg Nathan and Klabjan (2017), which is the best known optimization algorithm for solving problem (1) with doubly distributed data. All the algorithms are implemented in Scala with Spark 2.0. The experiments are conducted in a Hadoop cluster with 4 nodes, each containing 8 Intel Xeon 2.2GHz cores. We conduct experiments on three different-size synthetic datasets that are larger than the datasets in Nathan and Klabjan (2017) and two datasets used in Wongchaisuwat and Klabjan (2018) extracted from SemMed Database. For all of these datasets, we train one of the most popular classification models: binary classification hinge loss support vector machines (SVM), and set the learning rate $\gamma_t = \frac{1}{(1+\sqrt{t-1})}$, which is also employed in Nathan and Klabjan (2017). Furthermore, we set the feature partition number $Q = 3$ and observation partition number $P = 5$, which is also one of the cases studied in Nathan and Klabjan (2017). We do not compare different learning rates and Q, P since these have been extensively studied in Nathan and Klabjan (2017).

5.1 SVM with Synthetic data

We first compare SODDA with RADiSA-avg Nathan and Klabjan (2017) using synthetic data. The datasets for these experiments are generated based on a standard procedure introduced in Zhang et al. (2012), which is also used in Nathan and Klabjan (2017): the x_i's and z are sampled from the uniform distribution in $[-1, 1]$, and $y_i := \text{sgn}(x_i z)$ with probability 0.01 of flipping the sign. In addition, all the data is in the dense format and the features are standardized to have unit variance. The size of each partition from the small-size dataset is $50,000 \times 6,000$, the one from the mid-size data is $60,000 \times 7,000$ and the one from the large-size data is $60,000 \times 9,000$. The information about these three datasets is listed in Table 1.

data size	small	medium	large
$P \times Q$	5×3	5×3	5×3
size of each partition	$50,000 \times 6,000$	$60,000 \times 7,000$	$60,000 \times 9,000$
Number of Spark executors used	18	25	25

Table 1: Synthetic datasets for numerical experiments

First, we conduct $\mathcal{b}^t, c^t, \mathcal{d}^t$ subsequence related experiments. We justify the value of $(\mathcal{b}^t, c^t, \mathcal{d}^t)$ from the small-size dataset, since the other two datasets would take more computational time. We study the impact of $(\mathcal{b}^t, c^t, \mathcal{d}^t)$ to the performance of SODDA by varying one of the three parameters $(\mathcal{b}^t, c^t, \mathcal{d}^t)$ while keeping the other two parameters fixed.

The most important results are presented in Figure 2. In Figure 2(a), we study the cases where the number of total observations used to estimate the full gradient in step 8 varies from 60% to 90% with $\mathcal{b}^t = c^t = 100\%$. In Figure 2(b), we consider the cases when c^t varies from 40% to 80% given that every feature is involved to compute the approximated full gradient, i.e. $\mathcal{b}^t = 100\%$. Figure 2(c) represents the cases where only partial features are used in step 8 but everything available is fully used, i.e. $\mathcal{b}^t = c^t$. In Figures 2(d)-(f), we study three different $\mathcal{b}^t$ choices and for each one we vary c^t. Figure 2(g) is an extension of Figure 2(d) showing the long-time performance under the corresponding set of parameters.

In these plots, we observe that every set of parameters $(\mathcal{b}^t, c^t, \mathcal{d}^t)$ with the small-size dataset outperforms RADiSA-avg in early iterations, however, the benefits peak at certain points. More precisely, from Figure 2(a), we discover that the marginal benefit grows up dramatically when $\mathcal{d}^t$ increases from 60% to 80% and slows down from 80% to 90%, thus, $\mathcal{d}^t = 85\%$ seems to be most beneficial. When it comes to c^t, we observe that although the value of c^t does not influence the accuracy of the solution, a higher value of c^t leads to a faster convergence speed to a good quality solution in Figure 2(b). Thus, we set $c^t = 80\%$ as a good value. From Figures 2(c)-(g), we observe that the value of $\mathcal{b}^t$ affects the accuracy of the solution significantly, therefore, we set $\mathcal{b}^t = 85\%$ after taking both the accuracy of the solution and the computational time into consideration.

In these figures, we observe that SODDA always outperforms RADiSA-avg in early iterations on the small-size dataset, and there is a trade-off between the accuracy of the loss function value and the sampling sizes used in the algorithm. More precisely, using less data leads to a faster convergence speed but a less accurate solution, while using more data contributes to a more accurate solution but requires more time.

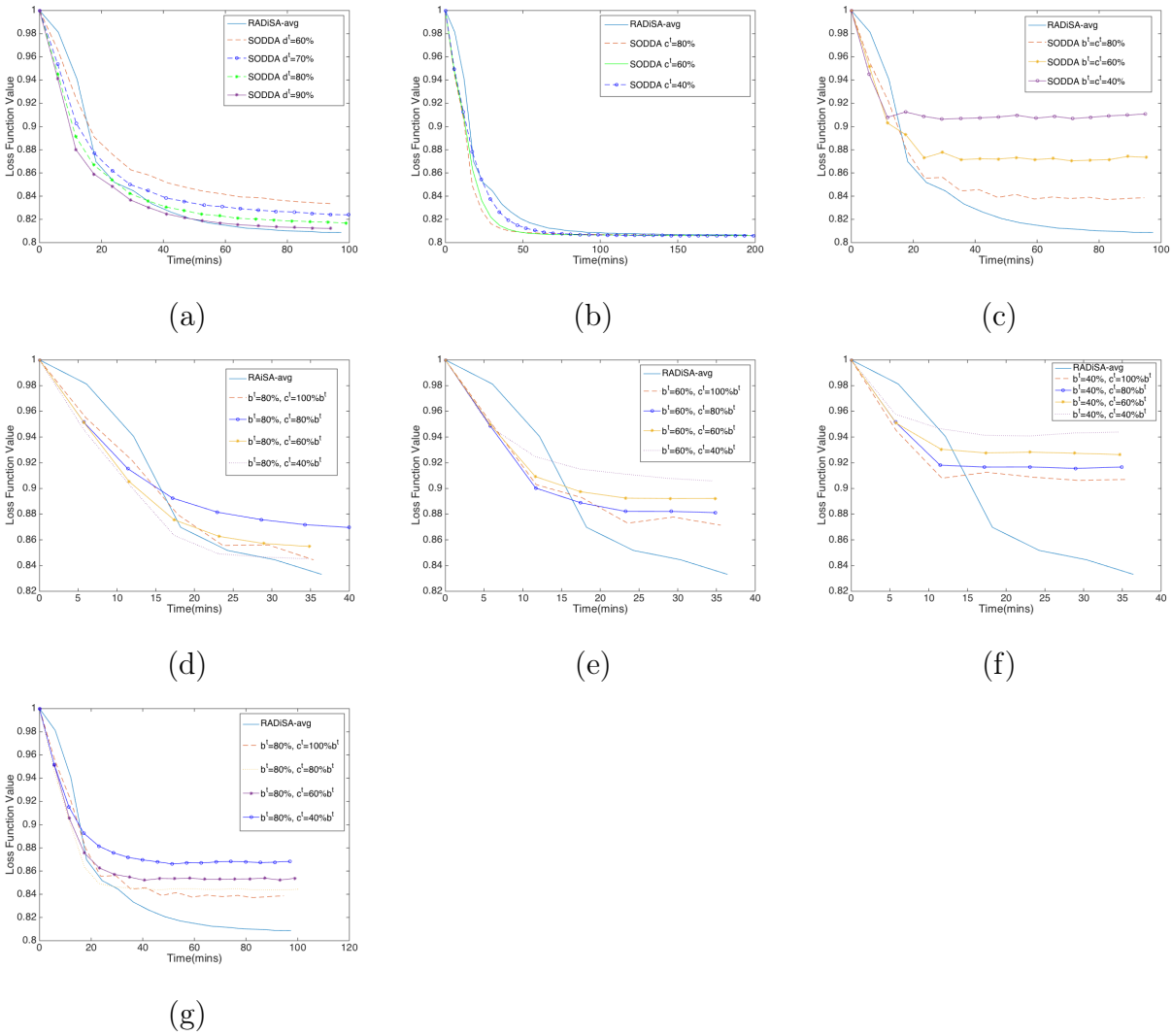
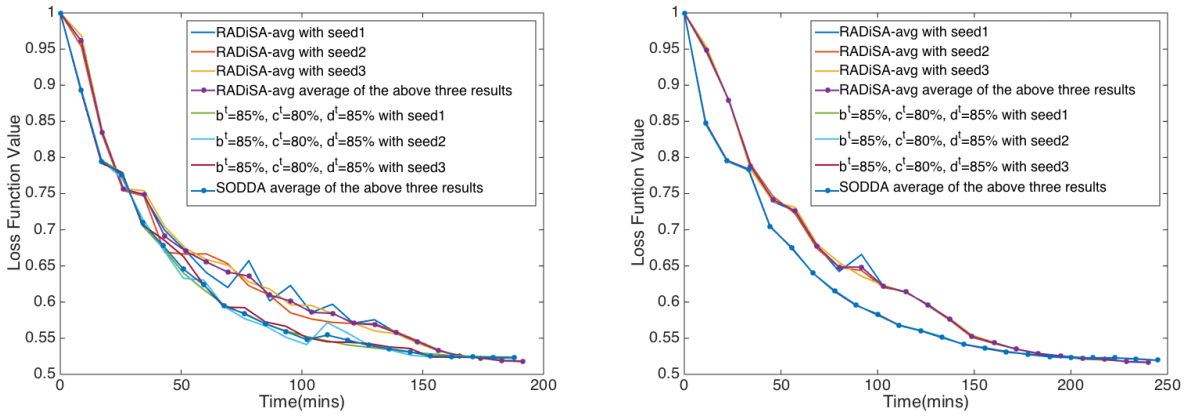


Figure 2: Comparison of SODDA and RADiSA-avg on small-size dataset

After specifying the values of $\mathscr{b}^t$, c^t and $\mathscr{d}^t$ to be $(85\%, 80\%, 85\%)$, we test both SODDA and RADiSA-avg on the mid- and large-size datasets with three different seeds. The results are presented in Figure 3. As we can observe, SODDA always exhibits a stronger and faster convergence than RADiSA-avg. It is interesting that as the size of the dataset increases, the intersection time of SODDA and RADiSA-avg comes later, which gives SODDA more advantages over RADiSA-avg when dealing with large datasets.

In the SODDA algorithm, we randomly choose a subset of observations and a block of features to estimate the full gradient in step 8. Moreover, both SODDA and RADiSA-avg utilize an observation randomly selected from a randomly chosen sub-matrix in the update step, where SODDA employs a sub-block of the approximated full gradient as a corrector but RADiSA-avg employs the exact full gradient. In order to eliminate the uncertainty about the choice of seeds, we conduct experiments on the large-size dataset under the same set of parameter $(\mathscr{b}^t, c^t, \mathscr{d}^t)$ with different seeds. Table 2 summarizes the influence of the change of the seed on the large-size dataset. For 10 different seeds, we run 40 iterations for each.



(a) (b)

Figure 3: Comparison of SODDA and RADiSA-avg for three different seeds on the mid- and large-size datasets

The first two columns present the average of the difference of the maximum objective value and the average function value across the 10 seeds, and the average of the difference of the average function value and the minimum objective value across the 10 seeds, respectively. Similarly, the remaining terms are defined as the maximum of the difference of the maximum objective value and the average function value, and the maximum of the difference of the average function value and the minimum objective value. As we can see in Table 2, the perturbation caused by the change of the seed is negligible especially when compared to the objective function value, which is a positive characteristic. Thus, in the remaining experiments, we no longer need to consider the impact of the randomness caused by either SODDA or RADiSA-avg.

	avg(max-avg)	avg(avg-min)	max(max-avg)	max(avg-min)
SODDA	0.4600×10^{-4}	0.0251×10^{-4}	0.2500×10^{-3}	3.0000×10^{-3}
RADiSA-avg	1.6373×10^{-4}	1.2606×10^{-4}	1.8000×10^{-3}	2.3500×10^{-3}

Table 2: Variation of SODDA and RADiSA-avg by using different seeds

5.2 SVM with SemMed Database

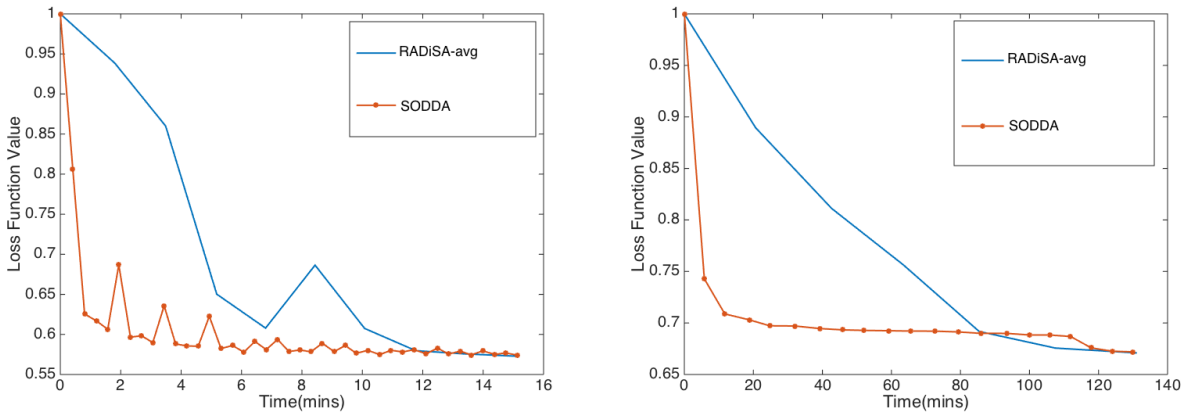
In the last set of experiments, we study the performances of SODDA with the $(\mathscr{b}^t, c^t, \mathscr{d}^t)$ selected in the previous section and RADiSA-avg on the Semantic MEDLINE Database Kilicoglu et al. (2012) with SemRep, a semantic interpreter of biomedical text Rindflesch and Fiszman (2003) as an extraction tool to construct the knowledge graph (KG). Like the preprocessing done in Wongchaisuwat and Klabjan (2018), we apply the inference method, which is called the Path Ranking Algorithm (PRA) Lao and Cohen (2010), to KG constructed from SemRep. The model under consideration is still linear SVM, and all the datasets considered are in the sparse format. The first dataset DIAG-neg10 is based on relationship

"DIAGNOSES," while LOC-neg5 is created in a similar manner based on "LOCATION OF." The data is summarized in Table 3.

Figure 4 illustrates the convergence paths of the objective loss function $F(\omega)$ generated by SODDA and RADiSA-avg versus time. We observe that using SODDA is much better than RADiSA with respect to not only the running time but also the loss reduction in early iterations. Comparing Figure 4(a) with Figure 4(b), we discover that the superior behavior of RADiSA over RADiSA-avg is more apparent and robust when applied to larger datasets, which is expected since it is more beneficial for datasets with larger size to perform partial computation instead of full computation of gradients in step 8.

Dataset	Observations (N)	Features (d)	Size of each partition ($n \times m$)
DIAG-neg10	425,185	26,946	$85,037 \times 8,982$
LOC-neg5	5,638,696	26,966	$1,127,740 \times 8,989$

Table 3: Datasets extracted from SemMed database



(a) DIAG-neg10 (b) LOC-neg5

Figure 4: Comparison of SODDA and RADiSA-avg on SemMed database

5.3 Key Findings

From the first set of experiments conducted on different synthetic datasets in the dense format, we justify a good set of parameters $(\mathscr{b}^t, c^t, \mathscr{d}^t) = (85\%, 80\%, 85\%)$ and eliminate the potential impact of the randomness involved in SODDA and RADiSA to the performance of the convergence. Furthermore, we discover that SODDA always exhibits a stronger and faster convergence than RADiSA-avg for every dataset considered and parameter values chosen. In the second set of experiments, we observe the same dominance of SODDA when compared to RADiSA-avg on sparse datasets.

In conclusion, SODDA provides a faster, stronger and more robust convergence than RADiSA-avg for both dense and sparse datasets.

References

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice Hall Englewood Cliffs, 1989.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Baptism's 91 Witnesses, 2008.

Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous SGD. *arXiv preprint arXiv:1604.00981*, 2016.

Soham De and Tom Goldstein. Efficient distributed SGD with variance reduction. In *2016 IEEE 16th International Conference on Data Mining*, pages 111–120. IEEE, 2016.

Reza Harikandeh, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt, Jakub Konečnỳ, and Scott Sallinen. Stop wasting my gradients: Practical SVRG. In *Advances in Neural Information Processing Systems*, pages 2251–2259, 2015.

Mingyi Hong, Meisam Razaviyayn, Zhi-Quan Luo, and Jong-Shi Pang. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine*, 33(1):57–77, 2015.

Martin Jaggi, Virginia Smith, Martin Takác, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 3068–3076, 2014.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012.

Ni Lao and William W Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81(1):53–67, 2010.

Chenxin Ma, Virginia Smith, Martin Jaggi, Michael I Jordan, Peter Richtárik, and Martin Takáč. Adding vs. averaging in distributed primal-dual optimization. *arXiv preprint arXiv:1502.03508*, 2015.

Jakub Mareček, Peter Richtárik, and Martin Takáč. Distributed block coordinate descent for minimizing partially separable functions. In *Numerical Analysis and Optimization*, pages 261–288. Springer, 2015.

Aryan Mokhtari and Alejandro Ribeiro. DSA: Decentralized double stochastic averaging gradient algorithm. *The Journal of Machine Learning Research*, 17(1):2165–2199, 2016.

Aryan Mokhtari, Alec Koppel, and Alejandro Ribeiro. A class of parallel doubly stochastic algorithms for large-scale learning. *arXiv preprint arXiv:1606.04991*, 2016.

Alexandros Nathan and Diego Klabjan. Optimization for large-scale machine learning with distributed features and observations. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 132–146, 2017.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Neal Parikh and Stephen Boyd. Block splitting for distributed optimization. *Mathematical Programming Computation*, 6(1):77–102, 2014.

Thomas C Rindflesch and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

Markus Weimer, Sriram Rao, and Martin Zinkevich. A convenient framework for efficient parallel multipass algorithms. In *LCCC: NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*, 2010.

Papis Wongchaisuwat and Diego Klabjan. Truth Validation with Evidence. *arXiv preprint arXiv:1802.05786*, 2018.

Caoxie Zhang, Honglak Lee, and Kang Shin. Efficient distributed linear classification algorithms via the alternating direction method of multipliers. In *Artificial Intelligence and Statistics*, pages 1398–1406, 2012.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.

6. Appendix

A Problem Set-up

We study the optimization problem of minimizing

$$\min_{\omega\in\mathbb{R}^d} F(\omega) := \frac{1}{N}\sum_{i=1}^{N} f_i(x_i\omega) = \frac{1}{N}\sum_{k=1}^{P}\sum_{j=1}^{n} f_j^k\left(\sum_{q=1}^{Q}\sum_{p=1}^{P} x_j^{p,q,k}\omega_{q,\pi_q(p)}\right),$$

where the features and the observations of the data $\{(x_i, y_i)\}_{i=1}^{N}$ are split into Q and P partitions respectively, and each feature partition is further separated into P smaller divisions. We have

$$\begin{aligned} &n = N/P, \quad m = d/Q, \quad \tilde{m} = d/QP, \\ &\omega = \left(\omega_{11}, \omega_{12}, \cdots, \omega_{1P}, \omega_{21}, \cdots, \omega_{2P}, \cdots, \omega_{QP}\right). \end{aligned}$$

B Notation

Recall that in steps 9- 18, the inner loop of SODDA performs iterations on each parameter subset $\omega_{q,\pi_q(p)}$ (for $i \geq 0$):

$$\begin{aligned} \bar{\omega}_{q,\pi_q(p)}^{(i+1)} = \bar{\omega}_{q,\pi_q(p)}^{(i)} \\ -\gamma_{t+1}\left[\nabla_{\omega_{q,\pi_q(p)}} f^p_{j_{q,\pi_q(p)}}(x^{p,q,\pi_q(p)}_{j_{q,\pi_q(p)}}\bar{\omega}^{(i)}_{q,\pi_q(p)}) - \nabla_{\omega_{q,\pi_q(p)}} f^p_{j_{q,\pi_q(p)}}(x^{p,q,\pi_q(p)}_{j_{q,\pi_q(p)}}\tilde{w}_{q,\pi_q(p)}) + \mu^t_{q,\pi_q(p)}\right], \end{aligned}$$

where $j_{q,\pi_q(p)}$ is a randomly selected observation in sub-block $x^{p,q,\pi_q(p)}$. It is convenient to use the notation

$$v^{t,i} = \begin{bmatrix} \nabla_{\omega_{11}} f^{\pi_1^{-1}(1)}_{j_{11}}\left(x^{\pi_1^{-1}(1),1,1}_{j_{11}}\bar{\omega}^{t,i-1}_{11}\right) - \nabla_{\omega_{11}} f^{\pi_1^{-1}(1)}_{j_{11}}\left(x^{\pi_1^{-1}(1),1,1}_{j_{11}}\tilde{\omega}_{11}\right) \\ \nabla_{\omega_{12}} f^{\pi_1^{-1}(2)}_{j_{12}}\left(x^{\pi_1^{-1}(2),1,2}_{j_{12}}\bar{\omega}^{t,i-1}_{12}\right) - \nabla_{\omega_{12}} f^{\pi_1(2)}_{\pi_1^{-1}(2)}\left(x^{\pi_1^{-1}(2),1,2}_{j_{12}}\tilde{\omega}_{12}\right) \\ \vdots \\ \nabla_{\omega_{1P}} f^{\pi_1^{-1}(P)}_{j_{1P}}\left(x^{\pi_1^{-1}(P),1,P}_{j_{1P}}\bar{\omega}^{t,i-1}_{1P}\right) - \nabla_{\omega_{1P}} f^{\pi_1^{-1}(P)}_{j_{1P}}\left(x^{\pi_1^{-1}(P),1,P}_{j_{1P}}\tilde{\omega}_{1P}\right) \\ \nabla_{\omega_{21}} f^{\pi_2^{-1}(1)}_{j_{21}}\left(x^{\pi_2^{-1}(1),2,1}_{j_{21}}\bar{\omega}^{t,i-1}_{21}\right) - \nabla_{\omega_{21}} f^{\pi_2^{-1}(1)}_{j_{21}}\left(x^{\pi_2^{-1}(1),2,1}_{j_{21}}\tilde{\omega}_{21}\right) \\ \vdots \\ \nabla_{\omega_{2P}} f^{\pi_2^{-1}(P)}_{j_{2P}}\left(x^{\pi_2^{-1}(P),2,P}_{j_{2P}}\bar{\omega}^{t,i-1}_{2P}\right) - \nabla_{\omega_{2P}} f^{\pi_2^{-1}(P)}_{j_{2P}}\left(x^{\pi_2^{-1}(P),2,P}_{j_{2P}}\tilde{\omega}_{2P}\right) \\ \vdots \\ \nabla_{\omega_{QP}} f^{\pi_Q^{-1}(P)}_{j_{QP}}\left(x^{\pi_Q^{-1}(P),Q,P}_{j_{QP}}\bar{\omega}^{t,i-1}_{QP}\right) - \nabla_{\omega_{QP}} f^{\pi_Q^{-1}(P)}_{j_{QP}}\left(x^{\pi_Q^{-1}(P),Q,P}_{j_{QP}}\tilde{\omega}_{QP}\right) \end{bmatrix} \in \mathbb{R}^d,$$

where $\pi_q^{-1}(p) \in \{1, 2, \cdots, P\}$ is the inverse function of $\pi_q(p)$, for all q and p. With this notation, we can integrate all subsets $\omega_{q,\pi_q(p)}$ and simplify the inner loop of the SODDA as follows

$$
\begin{aligned}
&\omega^t \\
&\quad \bar{\omega}^{t,0} = \omega^t \\
&\quad \bar{\omega}^{t,1} = \bar{\omega}^{t,0} - \gamma_{t+1}\left(\mu^t + v^{t,1}\right) \\
&\quad \bar{\omega}^{t,2} = \bar{\omega}^{t,1} - \gamma_{t+1}\left(\mu^t + v^{t,2}\right) \\
&\quad \vdots \\
&\quad \bar{\omega}^{t,B} = \bar{\omega}^{t,B-1} - \gamma_{t+1}\left(\mu^t + v^{t,B}\right) \\
&\omega^{t+1} = \bar{\omega}^{t,B}
\end{aligned}
.
$$

In what follows corresponding assumptions hold. Lastly, we define $\mathcal{F}^t$ as the sigma algebra that measures the history of the algorithm up until iteration t.

We also introduce $f \in \hat{\mathcal{O}}(g)$ if there exists a constant $C > 0$ such that $f(x) \leq C \cdot g(x)$ for every $x \geq 0$.

C Diminishing Learning Rate Convergence without Feature Sampling

Lemma 1 *Let $\Phi = \{\phi_1, \cdots, \phi_R\}$ be a set of random vectors measurable with respect to $\sigma-$algebra $\mathcal{H}$, let $g : \Phi \to \mathbb{R}^k$ be a measurable function, and let $\mathcal{b}$ be an integer such that $1 \leq \mathcal{b} \leq R$. Let $\mathcal{B}$ be a set of size $\mathcal{b}$ uniformly and randomly selected vectors from Φ without replacement. Given two constants w_1 and w_2, we have*

$$
\mathbb{E}\left[w_1 \sum_{i \in \mathcal{B}} g(\phi_i) + w_2 \sum_{i \notin \mathcal{B}} g(\phi_i) \middle| \mathcal{H} \right] = \left(\frac{\mathcal{b}}{R} w_1 + \frac{R - \mathcal{b}}{R} w_2 \right) \sum_{i=1}^{R} g(\phi_i).
$$

Proof Using the definition of the expectation we obtain

$$
\mathbb{E}\left[w_1 \sum_{i \in \mathcal{B}} g(\phi_i) + w_2 \sum_{i \notin \mathcal{B}} g(\phi_i) \middle| \mathcal{H} \right] = \sum_{\mathcal{B}} \frac{1}{\binom{R}{\mathcal{b}}} \left[w_1 \sum_{i \in \mathcal{B}} g(\phi_i) + w_2 \sum_{i \notin \mathcal{B}} g(\phi_i) \right],
$$

where the first summation indicates summation over all subsets of $\mathcal{B}$ of cardinality $\mathcal{b}$. Thus, the expected value of $w_1 \sum_{i \in \mathcal{B}} g(\phi_i) + w_2 \sum_{i \notin \mathcal{B}} g(\phi_i)$ with respect to $\mathcal{B}$ and conditioning on $\mathcal{H}$ is

$$
\begin{aligned}
\mathbb{E}\left[w_1 \sum_{i \in \mathcal{B}} g(\phi_i) + w_2 \sum_{i \notin \mathcal{B}} g(\phi_i) \middle| \mathcal{H} \right] &= \left(\frac{\mathcal{b}}{R} w_1 + \left(1 - \frac{\mathcal{b}}{R}\right) w_2 \right) \sum_{i=1}^{R} g(\phi_i) \\
&= \left(\frac{\mathcal{b}}{R} w_1 + \frac{R - \mathcal{b}}{R} w_2 \right) \sum_{i=1}^{R} g(\phi_i),
\end{aligned}
$$

since each i is selected with probability $\frac{\mathcal{b}}{R}$. ■

Lemma 2 *If Assumption 1 holds, then $\|\nabla F(\omega^t)\|$ and $\sum_{j=1}^{N} \|\nabla f_j(x_j \omega^t)\|^2$ for any t satisfy*

$$
\left\|\nabla F(\omega^t)\right\| \leq M_1, \tag{13}
$$

$$
\sum_{j=1}^{N} \left\|\nabla f_j(x_j \omega^t)\right\|^2 \leq N M_1^2. \tag{14}
$$

Proof Using Assumptions 1 we obtain

$$\left\|\bigtriangledown F(\omega^t)\right\| = \left\|\frac{1}{N}\sum_{i=1}^{N}\bigtriangledown f_j(x_j\omega^t)\right\| \leq \frac{1}{N}\sum_{i=1}^{N}\left\|\bigtriangledown f_j(x_j\omega^t)\right\| \leq \frac{1}{N}\sum_{i=1}^{N} M_1 = M_1.$$

Similarly, for any ω^t we have

$$\sum_{j=1}^{N}\left\|\bigtriangledown f_j(x_j w^t)\right\|^2 \leq \sum_{i=1}^{N} M_1^2 = NM_1^2.$$

This completes the proof of the lemma. ■

We assume that ω^* is the unique optimal solution to (1). Under these standard assumptions and the previous results, our first proposition argues a supermartingale relationship for the sequence of the loss function errors $F(\omega^t) - F(\omega^*)$.

Proposition 1 *If Assumptions 1-3 hold, and the sequence of learning rates satisfies $\gamma_t \leq 1$ for all t, and the sequences $(c^t, \mathscr{l}^t)_{t=0}^{\infty}$ are selected so that $c^t \leq d$ and $\mathscr{l}^t \leq N$, then the loss function error sequence $F(\omega^t) - F(\omega^*)$ generated by SODDA satisfies*

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)|\mathscr{F}^t\right] \leq (1 - \frac{2\xi B}{d}\gamma_{t+1})[F(\omega^t) - F(\omega^*)] + C_1\gamma_{t+1}^2, \tag{15}$$

where C_1 is a positive constant.

Proof We write $\bigtriangledown F(\omega^t) = (\bigtriangledown F(\omega^t)_{11}, \cdots, \bigtriangledown F(\omega^t)_{1P}, \bigtriangledown F(\omega^t)_{21}, \cdots, \bigtriangledown F(\omega^t)_{QP})$ and $e^t = (e_{11}^t, \cdots, e_{1P}^t, e_{21}^t, \cdots, e_{QP}^t)$. In order to simplify the notation, we also denote $\pi = (\pi_q)_{q=1}^{Q}$ and $j_{q,\pi_q(p)}^{(i)}$, the index drawn in step 10 of the algorithm for given $\pi_q(p)$, where everything computed is at iteration t.

Claim 1 *For any t we have*

$$\mathbb{E}\left[\left.\frac{1}{\mathscr{l}^t}\sum_{j\in\mathscr{D}^t}\bar{\bigtriangledown}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right|\mathscr{F}^t\right] = \frac{c^t}{d}\bigtriangledown F(\omega^t), \tag{16}$$

$$\mathbb{E}\left[\left.\left\|\frac{1}{\mathscr{l}^t}\sum_{j\in\mathscr{D}^t}\bar{\bigtriangledown}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\right|\mathscr{F}^t\right] \leq \frac{c^t M_1^2}{d}. \tag{17}$$

Proof Applying Lemma 1 with $w_1 = 1$, $w_2 = 0$, $\Phi = \left\{\bar{\bigtriangledown}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\}_{j=1}^{N}$, $g(z) = z$, $\mathscr{H} = \sigma(\mathscr{F}^t, \mathcal{C}^t)$, $\mathscr{B} = \mathscr{D}^t$ and the law of iterated expectation imply

$$\mathbb{E}\left[\mathbb{E}\left[\left.\frac{1}{\mathscr{l}^t}\sum_{j\in\mathscr{D}^t}\bar{\bigtriangledown}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right|\mathscr{F}^t, \mathcal{C}^t\right]\right|\mathscr{F}^t\right] = \frac{1}{\mathscr{l}^t}\cdot\frac{\mathscr{l}^t}{N}\sum_{j=1}^{N}\mathbb{E}\left[\left.\bar{\bigtriangledown}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right|\mathscr{F}^t\right].$$

For each j, we in turn have

$$\mathbb{E}\left[\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\middle|\mathcal{F}^t\right] = \frac{1}{\binom{d}{c^t}}\sum_{\mathcal{C}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)$$
$$= \frac{1}{\binom{d}{c^t}}\cdot\binom{d-1}{c^t-1}\nabla f_j(x_j\omega^t) = \frac{c^t}{d}\nabla f_j(x_j\omega^t).$$

This yields

$$\mathbb{E}\left[\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\middle|\mathcal{F}^t\right] = \frac{1}{N}\sum_{j=1}^{N}\frac{c^t}{d}\nabla f_j(x_j\omega^t) = \frac{c^t}{Nd}\sum_{j=1}^{N}\nabla f_j(x_j\omega^t). \tag{18}$$

By substituting the definition of $\nabla F(\omega^t)$ into (18) claim (16) follows.

Let us proceed to find an upper bound for the expected value of $\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2$ given $\mathcal{F}^t$. Applying the law of iterated expectation and Lemma 1 with $w_1 = 1$, $w_2 = 0$, $\Phi = \left\{\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\}_{j=1}^N$, $g(z) = \|z\|^2$, $\mathcal{H} = \sigma(\mathcal{F}^t, c^t)$ and $\mathcal{B} = \mathcal{D}^t$ give

$$\mathbb{E}\left[\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right] \le \frac{1}{\mathscr{d}^t}\mathbb{E}\left[\sum_{j\in\mathcal{D}^t}\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right] \tag{19}$$
$$= \frac{1}{\mathscr{d}^t}\mathbb{E}\left[\mathbb{E}\left[\sum_{j\in\mathcal{D}^t}\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t,\mathcal{C}^t\right]\middle|\mathcal{F}^t\right] = \frac{1}{\mathscr{d}^t}\cdot\frac{\mathscr{d}^t}{N}\mathbb{E}\left[\sum_{j=1}^{N}\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right]$$
$$= \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right],$$

which in turn yields

$$\mathbb{E}\left[\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right] \le \frac{1}{N}\sum_{j=1}^{N}\frac{c^t}{d}\mathbb{E}\left[\left\|\nabla f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right]$$
$$= \frac{c^t}{dN}\sum_{j=1}^{N}\mathbb{E}\left[\left\|\nabla f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right], \tag{20}$$

where we apply Lemma 1 for each j again with $w_1 = 1$, $w_2 = 0$, $\Phi = \left\{(\nabla f_j(x_j\omega^t))_i\right\}_{i=1}^d$, $g(z) = z^2$, $\mathcal{H} = \mathcal{F}^t$ and $\mathcal{B} = \mathcal{C}^t$. By inserting (14) from Lemma 2 into (20) the claim in (17) follows. ■

For $i = 1, 2, \cdots, B$, the expected value of $\|v^{t,i}\|^2$ given all the preceding information is

$$\mathbb{E}\left[\left\|v^{t,1}\right\|^2 |\mathcal{F}^t, \mathcal{B}^t, \mathcal{C}^t, \mathcal{D}^t, \pi\right] = 0 \tag{21}$$

$$\begin{aligned}
&\mathbb{E}\left[\left\|v^{t,i}\right\|^2 |\mathcal{F}^t, \mathcal{B}^t, \mathcal{C}^t, \mathcal{D}^t, \pi, j_{11}^{(1)}, j_{12}^{(1)}, \cdots, j_{QP}^{(1)}, j_{11}^{(2)}, j_{12}^{(2)}, \cdots, j_{QP}^{(2)}, \cdots, j_{11}^{(i-2)}, j_{12}^{(i-2)}, \cdots, j_{QP}^{(i-2)}\right] \\
&= \sum_{j_{QP}^{(i-1)}=1}^{n} \cdots \sum_{j_{11}^{(i-1)}=1}^{n} \frac{1}{n^{QP}} \left\| \begin{pmatrix} \nabla_{\omega_{11}} f_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1)} \left(x_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1),1,1} \bar{\omega}_{11}^{t,i-1} \right) - \nabla_{\omega_{11}} f_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1)} \left(x_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1),1,1} \omega_{11}^{t} \right) \\ \vdots \\ \nabla_{\omega_{1P}} f_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P)} \left(x_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P),1,P} \bar{\omega}_{1P}^{t,i-1} \right) - \nabla_{\omega_{1P}} f_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P)} \left(x_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P),1,P} \omega_{1P}^{t} \right) \\ \vdots \\ \nabla_{\omega_{QP}} f_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P)} \left(x_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P),Q,P} \bar{\omega}_{QP}^{t,i-1} \right) - \nabla_{\omega_{QP}} f_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P)} \left(x_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P),Q,P} \omega_{QP}^{t} \right) \end{pmatrix} \right\|^2,
\end{aligned} \tag{22}$$

for $i = 2, 3, \cdots, B$.

We prove a bound of the expected value of $\|v^{t,i}\|^2$ given $\mathcal{F}^t$ by induction for $i \in \{1, 2, \cdots, B\}$.

Claim 2 *For $i = 1, 2, \cdots, B$, we have*

$$\mathbb{E}\left[\left\|v^{t,i}\right\|^2 |\mathcal{F}^t\right] = \hat{\mathcal{O}}(\gamma_{t+1}^2). \tag{23}$$

Proof The claim holds for $i = 1$ due to (21).

For $i = 1, 2, \cdots, k-1$, we assume that

$$\mathbb{E}\left[\left\|v^{t,i}\right\|^2 |\mathcal{F}^t\right] = \hat{\mathcal{O}}(\gamma_{t+1}^2). \tag{24}$$

Now consider $v^{t,k}$. Let us show that the expected value of $\left\|v^{t,k}\right\|^2$ is bounded. By using (22) we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|v^{t,k}\right\|^2 |\mathcal{F}^t, \mathcal{B}^t, \mathcal{C}^t, \mathcal{D}^t, \pi, j_{11}^{(1)}, j_{12}^{(1)}, \cdots, j_{QP}^{(1)}, j_{11}^{(2)}, j_{12}^{(2)}, \cdots, j_{QP}^{(2)}, \cdots, j_{11}^{(k-2)}, j_{12}^{(k-2)}, \cdots, j_{QP}^{(k-2)}\right] \\
&\leq \sum_{j_{QP}^{(k-1)}=1}^{n} \cdots \sum_{j_{11}^{(k-1)}=1}^{n} \sum_{q=1}^{Q} \sum_{p=1}^{P} \frac{1}{n^{QP}} \left\| \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \bar{\omega}_{qp}^{t,k-1} \right) - \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \omega_{qp}^{t} \right) \right\|^2 \\
&= \sum_{q=1}^{Q} \sum_{p=1}^{P} \left(\frac{1}{n} \sum_{j_{q,\pi_q(p)}^{(k-1)}=1}^{n} \left\| \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \bar{\omega}_{qp}^{t,k-1} \right) - \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \omega_{qp}^{t} \right) \right\|^2 \right) \\
&\leq \sum_{q=1}^{Q} \sum_{p=1}^{P} \left(\frac{1}{n} \cdot nL^2 \left\| \bar{\omega}_{qp}^{t,k-1} - \bar{\omega}_{qp}^{t,0} \right\|^2 \right) = \sum_{q=1}^{Q} \sum_{p=1}^{P} \left(L^2 \left\| \bar{\omega}_{qp}^{t,k-1} - \bar{\omega}_{qp}^{t,0} \right\|^2 \right) \\
&= \sum_{q=1}^{Q} \sum_{p=1}^{P} L^2 \left\| \gamma_{t+1} \left[(k-1)\mu_{qp}^{t} + v_{qp}^{t,1} + \cdots + v_{qp}^{t,k-1} \right] \right\|^2.
\end{aligned} \tag{25}$$

Applying the definition of μ^t yields

$$\mathbb{E}\left[\left\|\mu^t\right\|^2|\mathcal{F}^t\right] = \mathbb{E}\left[\left\|\frac{1}{d^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right] \leq \frac{c^t M_1^2}{d} = \hat{\mathcal{O}}(1). \tag{26}$$

The second inequality holds due to (17) in Claim 1. By using the law of iterated expectation, (24), (25) and (26) we get

$$\begin{aligned}
\mathbb{E}\left[\left\|v^{t,k}\right\|^2|\mathcal{F}^t\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|v^{t,k}\right\|^2|\mathcal{F}^t,,\mathcal{B}^t,\mathcal{C}^t,\mathcal{D}^t,\pi,j_{11}^{(1)}\cdots,j_{QP}^{(k-2)}\right]|\mathcal{F}^t\right] \\
&\leq \mathbb{E}\left[\sum_{q=1}^{Q}\sum_{p=1}^{P}L^2\left\|\gamma_{t+1}\left[(k-1)\mu_{qp}^t+\sum_{i=1}^{k-1}v_{qp}^{t,i}\right]\right\|^2\middle|\mathcal{F}^t\right] \\
&\leq kL^2\gamma_{t+1}^2\sum_{q=1}^{Q}\sum_{p=1}^{P}\left(\mathbb{E}\left[\left\|(k-1)\mu_{qp}^t\right\|^2\middle|\mathcal{F}^t\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\left\|v_{qp}^{t,i}\right\|^2|\mathcal{F}^t\right]\right) \\
&\leq kL^2\gamma_{t+1}^2QP\left((k-1)^2\,\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\left\|v^{t,i}\right\|^2|\mathcal{F}^t\right]\right) \\
&= kL^2\gamma_{t+1}^2QP\left[(k-1)^2\hat{\mathcal{O}}(1)+(k-1)\hat{\mathcal{O}}(\gamma_{t+1}^2)\right] = \hat{\mathcal{O}}(\gamma_{t+1}^2).
\end{aligned} \tag{27}$$

This completes the proof of the claim. ■

By using the conditional Jensen's inequality and (23) we get

$$\mathbb{E}[\left\|v^{t,i}\right\| \mid \mathcal{F}^t] = \mathbb{E}[\sqrt{\|v^{t,i}\|^2} \mid \mathcal{F}^t] \leq \sqrt{\mathbb{E}[\|v^{t,i}\|^2 \mid \mathcal{F}^t]} = \hat{\mathcal{O}}(\gamma_{t+1}). \tag{28}$$

By summing up all increments in iteration t, we obtain

$$\omega^{t+1} = \omega^t - \gamma_{t+1}\left[B\mu^t + v^{t,1} + v^{t,2} + \cdots + v^{t,B}\right].$$

Then, the expected value of the difference $\omega^{t+1}-\omega^t$ given $\mathcal{F}^t$ is

$$\begin{aligned}
\mathbb{E}\left[\omega^{t+1}-\omega^t|\mathcal{F}^t\right] &= -\gamma_{t+1}\,\mathbb{E}\left[B\mu^t + v^{t,1} + \cdots + v^{t,B}|\mathcal{F}^t\right] \\
&= -\gamma_{t+1}B\frac{c^t}{d}\nabla F(\omega^t) - \gamma_{t+1}\sum_{i=1}^{B}\mathbb{E}[v^{t,i} \mid \mathcal{F}^t],
\end{aligned} \tag{29}$$

by using (16) in Claim 1. Moreover, the expected value of the squared norm $\|\omega^{t+1}-\omega^t\|^2$ given $\mathcal{F}^t$ is

$$\begin{aligned}
\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right] &= \mathbb{E}\left[\left\|\gamma_{t+1}\left[B\mu^t + v^{t,1} + v^{t,2} + \cdots + v^{t,B}\right]\right\|^2|\mathcal{F}^t\right] \\
&\leq \gamma_{t+1}^2(B+1)\left\{B^2\,\mathbb{E}\left[\left\|\mu^t\right\|^2|\mathcal{F}^t\right]+\sum_{i=1}^{B}\mathbb{E}\left[\left\|v^{t,i}\right\|^2|\mathcal{F}^t\right]\right\} \\
&= \hat{\mathcal{O}}(\gamma_{t+1}^2)\left\{B^2\cdot\hat{\mathcal{O}}(1)+B\cdot\hat{\mathcal{O}}(\gamma_{t+1}^2)\right\} = \hat{\mathcal{O}}(\gamma_{t+1}^2),
\end{aligned} \tag{30}$$

due to (17), (23) and (26). From

$$-\nabla F(\omega^t)\cdot v^{t,i} \le \left\|\nabla F(\omega^t)\right\| \left\|v^{t,i}\right\|,$$

for every $\nabla F(\omega^t)$ and $v^{t,i}$, by using (13) and (28) we obtain

$$\begin{aligned}-\gamma_{t+1}\nabla F(\omega^t)\cdot \mathbb{E}\left[v^{t,i}|\mathcal{F}^t\right] &= \gamma_{t+1}\,\mathbb{E}\left[-\nabla F(\omega^t)\cdot v^{t,i}|\mathcal{F}^t\right] \le \gamma_{t+1}\,\mathbb{E}\left[\left\|\nabla F(\omega^t)\right\|\cdot\left\|v^{t,i}\right\||\mathcal{F}^t\right]\\ &= \gamma_{t+1}\left\|\nabla F(\omega^t)\right\|\cdot\mathbb{E}\left[\left\|v^{t,i}\right\||\mathcal{F}^t\right] = \hat{\mathcal{O}}(\gamma_{t+1}^2),\end{aligned} \tag{31}$$

since $\mathbb{E}\left[XY|\mathcal{H}\right] = X\,\mathbb{E}\left[Y|\mathcal{H}\right]$ if X is $\mathcal{H}-$measurable.

For convex F we have

$$F(\omega^{t+1}) \le F(\omega^t) + \nabla F(\omega^t)^T\left(\omega^{t+1}-\omega^t\right) + \frac{L}{2}\left\|\omega^{t+1}-\omega^t\right\|^2,$$

which in turn yields

$$\begin{aligned}&\mathbb{E}\left[F(\omega^{t+1})|\mathcal{F}^t\right] \le F(\omega^t) + \nabla F(\omega^t)^T\,\mathbb{E}\left[\left(\omega^{t+1}-\omega^t\right)|\mathcal{F}^t\right] + \frac{L}{2}\,\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right]\\ &= F(\omega^t) + \nabla F(\omega^t)^T\left\{-\gamma_{t+1}B\frac{c^t}{d}\nabla F(\omega^t) - \gamma_{t+1}\sum_{i=1}^{B}\mathbb{E}[v^{t,i}\mid\mathcal{F}^t]\right\} + \frac{L}{2}\,\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right]\\ &= F(\omega^t) - \gamma_{t+1}\frac{c^tB}{d}\left\|\nabla F(\omega^t)\right\|^2 - \gamma_{t+1}\nabla F(\omega^t)^T\sum_{i=1}^{B}\mathbb{E}\left[v^{t,i}|\mathcal{F}^t\right] + \frac{L}{2}\,\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right]\\ &\le F(\omega^t) - \gamma_{t+1}\frac{c^tB}{d}\left\|\nabla F(\omega^t)\right\|^2 + \hat{\mathcal{O}}(\gamma_{t+1}^2) \le F(\omega^t) - \gamma_{t+1}\frac{c^tB}{d}\left\|\nabla F(\omega^t)\right\|^2 + C_1\gamma_{t+1}^2,\end{aligned} \tag{32}$$

where C_1 is a positive constant and we use (29), (30) and (31). Subtracting the optimal objective function $F(\omega^*)$ to the both sides of (32) and using the fact that $c^t \ge 1$ imply that

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)|\mathcal{F}^t\right] \le F(\omega^t) - F(\omega^*) - \gamma_{t+1}\frac{B}{d}\left\|\nabla F(\omega^t)\right\|^2 + C_1\gamma_{t+1}^2. \tag{33}$$

We proceed to find a lower bound of $\left\|\nabla F(\omega^t)\right\|^2$ in terms of $F(\omega^t) - F(\omega^*)$. Assumption 2 implies that, for any $y, z \in \mathbb{R}^m$

$$F(y) \ge F(z) + \nabla F(z)^T(y-z) + \frac{\xi}{2}\left\|y-z\right\|^2. \tag{34}$$

For fixed z, the right hand side of (34) is a quadratic function of y and it gets its minimum at $\hat{y} = z - \frac{1}{\xi}\nabla F(z)$. Therefore

$$F(y) \ge F(z) + \nabla F(z)^T(\hat{y}-z) + \frac{\xi}{2}\left\|\hat{y}-z\right\|^2 = F(z) - \frac{1}{2\xi}\left\|\nabla F(z)\right\|^2, \tag{35}$$

for any $y, z \in \mathbb{R}^d$. Setting $y = \omega^*$ and $z = \omega^t$ in (35) gives

$$\left\|\nabla F(\omega^t)\right\|^2 \ge 2\xi\left(F(\omega^t) - F(\omega^*)\right). \tag{36}$$

Substituting the lower bound in (36) by the norm of gradient square $\|\nabla F(\omega^t)\|^2$ in (33) yields the proposition in (15). ■

Proposition 1 represents a supermartingale relationship for the sequence of the loss function errors $F(\omega^t) - F(\omega^*)$. In the following theorem, by employing the supermartingale convergence argument, we show that if the sequence of learning rates satisfy the standard stochastic approximation diminishing learning rate rule (non-summable and squared summable), the sequence of loss function errors $F(\omega^t) - F(\omega^*)$ converges to 0 almost surely. Combining with strong convexity of $F(\omega)$ in Assumption 2, this result implies that $\|\omega^t - \omega^*\|$ converges to 0 almost surely.

Proof of Theorem 1

Proof We use the relationship in (15) to build a supermartingale sequence. First, let us define

$$\alpha^t := F(\omega^t) - F(\omega^*) + \sum_{u=t}^{\infty} C_1 \gamma_{u+1}^2, \tag{37}$$

$$\beta^t := \frac{2\xi B}{d} \gamma_{t+1} \left(F(\omega^t) - F(\omega^*) \right). \tag{38}$$

Note that α^t is well-defined since $\sum_{u=t}^{\infty} \gamma_{u+1}^2 < \sum_{u=1}^{\infty} \gamma_u^2 < \infty$. The definition of α^t and β^t in (37) and (38), and the inequality in (15) imply the expected value of α^{t+1} given $\mathcal{F}^t$ is

$$\mathbb{E}\left[\alpha^{t+1} | \mathcal{F}^t\right] \leq \alpha^t - \beta^t. \tag{39}$$

Since α^t and β^t are nonnegative and due to (39), they satisfy the conditions of the supermartingale convergence theorem. Thus, we conclude that

$$\text{(i)} \quad \alpha^t \text{ converges to a limit a.s., and} \tag{40}$$

$$\text{(ii)} \quad \sum_{t=1}^{\infty} \beta^t < \infty. \quad \text{a.s.} \tag{41}$$

Property (41) yields

$$\sum_{t=0}^{\infty} \frac{2c^t \xi B}{d} \gamma_{t+1} \left(F(\omega^t) - F(\omega^*) \right) < \infty. \quad \text{a.s.}$$

Since $\sum_{t=0}^{\infty} \gamma_{t+1} = \infty$, there exists a subsequence of $F(\omega^t) - F(\omega^*)$ which converges to 0, i.e.

$$\liminf_{t \to \infty} F(\omega^t) - F(\omega^*) = 0. \quad \text{a.s.} \tag{42}$$

Since $\sum_{u=t}^{\infty} C_1 \gamma_{u+1}^2$ is deterministic and due to (40), $F(\omega^t) - F(\omega^*)$ converges to a limit almost surely. In association with (42) we conclude

$$\lim_{t \to \infty} F(\omega^t) - F(\omega^*) = 0. \quad \text{a.s.} \tag{43}$$

We proceed to show the almost convergence of $\|\omega^t - \omega^*\|^2$. Using (34) again and setting $y = \omega^t$ and $z = \omega^*$ implies

$$F(\omega^t) \geq F(\omega^*) + \nabla F(\omega^*)^T(\omega^t - \omega^*) + \frac{\xi}{2}\left\|\omega^t - \omega^*\right\|^2. \tag{44}$$

Since the gradient of the optimal solution is 0, i.e.$\nabla F(\omega^*) = 0$, (44) can be rearranged as

$$F(\omega^t) - F(\omega^*) \geq \frac{\xi}{2}\left\|\omega^t - \omega^*\right\|^2.$$

Observing that the upper bound of $\left\|\omega^t - \omega^*\right\|^2$ converges to 0 almost surely by (43), we conclude that the sequence $\|\omega^t - \omega^*\|^2$ converges to zero almost surely. Hence, the claim in (2) is valid. ■

Proof of Theorem 2

Proof Replacing γ_{t+1} by $\frac{1}{t+1}$ and computing the expected value of (15) given $\mathcal{F}^0$ by using the law of iterated expectation we obtain

$$\mathbb{E}[F(\omega^{t+1}) - F(\omega^*)] \leq \left(1 - \frac{2\xi B}{(t+1)d}\right)\mathbb{E}[F(\omega^t) - F(\omega^*)] + \frac{C_1}{(t+1)^2}. \tag{45}$$

Let us define

$$\begin{aligned} a_t &:= \mathbb{E}[F(\omega^{t+1}) - F(\omega^*)] \\ \lambda &:= \frac{2\xi B}{d} \\ \beta &:= C_1. \end{aligned}$$

Note that β is positive. Based on the relationship in (45), we obtain

$$a_{t+1} \leq \left(1 - \frac{\lambda}{t+1}\right)a_t + \frac{\beta}{(t+1)^2}. \tag{46}$$

for all times $t \geq 0$. Now, we proceed to show

$$a_t \leq \frac{Q}{t+1}, \tag{47}$$

where $Q = \max\left\{a_0, 2a_1, \cdots, ([\lambda]+1)a_{[\lambda]}, ([\lambda]+2)\,a_{[\lambda]+1}, \frac{\beta}{\lambda - 1}\right\}$. The definition of Q implies that the relationship in (47) holds for $t = 1, 2, \cdots, [\lambda]$. The remaining cases are shown by induction.

When $t = [\lambda] + 1$, the definition of Q implies

$$a_{[\lambda]+1} \leq \frac{Q}{[\lambda]+2}.$$

When $t = k - 1$, we assume that the relationship in (47) holds. Considering the case when $t = k$ and using (46) implies

$$a_{k+1} \leq \left(1 - \frac{\lambda}{k+1}\right) a_k + \frac{\beta}{(k+1)^2} \leq \left(1 - \frac{\lambda}{k+1}\right) \frac{Q}{k+1} + \frac{\beta}{(k+1)^2}.$$

In order to satisfy (47), we require

$$\left(1 - \frac{\lambda}{k+1}\right) \frac{Q}{k+1} + \frac{\beta}{(k+1)^2} \leq \frac{Q}{k+2}.$$

Elementary algebraic manipulation shows that this is equivalent to

$$\beta(k+2) \leq Q\left[\lambda(k+2) - (k+1)\right]$$

and in turn

$$\frac{\beta(k+2)}{\lambda(k+2) - (k+1)} = \frac{\beta}{\lambda - \frac{k+1}{k+2}} \leq Q,$$

where we require $\lambda \geq 1$. The definition of Q, i.e. $Q \geq \frac{\beta}{\lambda - 1}$ and the relationship that $\lambda - \frac{k+1}{k+2} > \lambda - 1$ imply that

$$\frac{\beta}{\lambda - \frac{k+1}{k+2}} < \frac{\beta}{\lambda - 1} \leq Q,$$

and thus (47) holds for $t = k$. Thus, if $B \geq \frac{d}{2\xi}$, for any time $t \geq 0$, the result in (3) holds where the constant Q is defined based on (4). ■

Corollary 1 *If Assumptions 1-3 hold and the sequence of learning rates are non-summable* $\sum_{t=1}^{\infty} \gamma_t = \infty$ *and square summable* $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$*, then the sequence of parameters* ω^t *generated by* RADiSA *converges almost surely to the optimal solution* ω^**, that is*

$$\lim_{t \to \infty} \left\|\omega^t - \omega^*\right\|^2 = 0 \quad \text{a.s.} \tag{48}$$

Moreover, if learning rate is defined as $\gamma_t := \frac{1}{t}$ *for* $t = 1, 2, \cdots$ *and the batch size is chosen such that* $B \geq \frac{1}{2\xi}$*, then the expected loss function errors* $\mathbb{E}[F(\omega^t) - F(\omega^*)]$ *of* RADiSA *converges to 0 at least with a sublinear convergence rate of order* $\mathcal{O}(1/t)$*, i.e.*

$$\mathbb{E}[F(\omega^t) - F(\omega^*)] \leq \frac{Q}{1+t}, \tag{49}$$

where constant Q *is defined in (4) with some positive constant* C_1' *taking the place of* C_1 *and* $c^t = d$*.*

Proof RADiSA is a special case of SODDA with $c^t = d$, $\not{a}^t = N$. ■

D Constant Learning Rate without Feature Sampling

Proposition 2 *If Assumptions 1-3 hold, and the learning rate is constant $\gamma_t = \gamma$ such that $BL\gamma QP \leq 1$ and $\gamma \leq 1$, and the sequences $(c^t, \mathscr{d}^t)_{t=0}^{\infty}$ satisfy the same conditions as in Theorem 1, then the loss function error sequence $F(\omega^t) - F(\omega^*)$ generated by SODDA satisfies*

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)\middle|\mathscr{F}^t\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right)\left[F(\omega^t) - F(\omega^*)\right] + C_2 B^4\gamma^2, \tag{50}$$

where C_2 is a positive constant.

Proof For $i = 1, 2, \cdots, B$, the expected value of $\|v^{t,i}\|$ given all the preceding information is

$$\mathbb{E}\left[\left\|v^{t,1}\right\| | \mathscr{F}^t, \mathscr{B}^t, \mathscr{C}^t, \mathscr{D}^t, \pi\right] = 0 \tag{51}$$

$$\begin{aligned}
&\mathbb{E}\left[\left\|v^{t,i}\right\| | \mathscr{F}^t, \mathscr{B}^t, \mathscr{C}^t, \mathscr{D}^t, \pi, j_{11}^{(1)}, j_{12}^{(1)}, \cdots, j_{QP}^{(1)}, j_{11}^{(2)}, j_{12}^{(2)}, \cdots, j_{QP}^{(2)}, \cdots, j_{11}^{(i-2)}, j_{12}^{(i-2)}, \cdots, j_{QP}^{(i-2)}\right] \\
&= \sum_{j_{QP}^{(i-1)}=1}^{n} \cdots \sum_{j_{11}^{(i-1)}=1}^{n} \frac{1}{n^{QP}} \left\| \begin{pmatrix} \nabla_{\omega_{11}} f_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1)}\left(x_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1),1,1}\bar{\omega}_{11}^{t,i-1}\right) - \nabla_{\omega_{11}} f_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1)}\left(x_{j_{11}^{(i-1)}}^{\pi_1^{-1}(1),1,1}\omega_{11}^{t}\right) \\ \vdots \\ \nabla_{\omega_{1P}} f_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P)}\left(x_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P),1,P}\bar{\omega}_{1P}^{t,i-1}\right) - \nabla_{\omega_{1P}} f_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P)}\left(x_{j_{1P}^{(i-1)}}^{\pi_1^{-1}(P),1,P}\omega_{1P}^{t}\right) \\ \vdots \\ \nabla_{\omega_{QP}} f_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P)}\left(x_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P),Q,P}\bar{\omega}_{QP}^{t,i-1}\right) - \nabla_{\omega_{QP}} f_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P)}\left(x_{j_{QP}^{(i-1)}}^{\pi_Q^{-1}(P),Q,P}\omega_{QP}^{t}\right) \end{pmatrix} \right\|,
\end{aligned} \tag{52}$$

for $i = 2, 3, \cdots, B$.

We prove a bound of the expected value of $\sum_{i=1}^{B}\|v^{t,i}\|$ and $\sum_{i=1}^{B}\|v^{t,i}\|^2$ given $\mathscr{F}^t$ by induction.

Claim 3 *For any t, if $BL\gamma QP \leq 1$ and $\gamma \leq 1$, we have*

$$\sum_{i=1}^{B}\mathbb{E}\left[\left\|v^{t,i}\right\|\middle|\mathscr{F}^t\right] = \hat{\mathcal{O}}(B^3\gamma) \tag{53}$$

$$\sum_{i=1}^{B}\mathbb{E}\left[\left\|v^{t,i}\right\|^2\middle|\mathscr{F}^t\right] = \hat{\mathcal{O}}(B^4\gamma^2) + \hat{\mathcal{O}}(B^7\gamma^4). \tag{54}$$

Proof By using (51) we have

$$\mathbb{E}\left[\left\|v^{t,1}\right\|\middle|\mathscr{F}^t\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|v^{t,1}\right\| | \mathscr{F}^t, \mathscr{B}^t, \mathscr{C}^t, \mathscr{D}^t, \pi\right]\middle|\mathscr{F}^t\right] = 0. \tag{55}$$

For $i = 2, 3, \cdots, B$, using (52) gives

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v^{t,i}\right\| |\mathcal{F}^t, \mathcal{B}^t, \mathcal{C}^t, \mathcal{D}^t, \pi, j_{11}^{(1)}, j_{12}^{(1)}, \cdots, j_{QP}^{(1)}, j_{11}^{(2)}, j_{12}^{(2)}, \cdots, j_{QP}^{(2)}, \cdots, j_{11}^{(i-2)}, j_{12}^{(i-2)}, \cdots, j_{QP}^{(i-2)}\right] \\
&\leq \sum_{j_{QP}^{(i-1)}=1}^{n} \cdots \sum_{j_{11}^{(i-1)}=1}^{n} \sum_{q=1}^{Q}\sum_{p=1}^{P} \frac{1}{n^{QP}} \left\| \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \bar{\omega}_{qp}^{t,k-1} \right) \right. \\
&\qquad\qquad \left. - \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \omega_{qp}^{t} \right) \right\| \\
&= \sum_{q=1}^{Q}\sum_{p=1}^{P} \left(\frac{1}{n} \sum_{j_{q,\pi_q(p)}^{(i-1)}=1}^{n} \left\| \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \bar{\omega}_{qp}^{t,k-1} \right) - \nabla_{\omega_{qp}} f_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p)} \left(x_{j_{qp}^{(k-1)}}^{\pi_q^{-1}(p),q,p} \omega_{qp}^{t} \right) \right\| \right) \\
&\leq \sum_{q=1}^{Q}\sum_{p=1}^{P} \left(\frac{1}{n} \cdot nL \left\| \bar{\omega}_{qp}^{t,i-1} - \bar{\omega}_{qp}^{t,0} \right\| \right) = \sum_{q=1}^{Q}\sum_{p=1}^{P} \left(L \left\| \bar{\omega}_{qp}^{t,i-1} - \bar{\omega}_{qp}^{t,0} \right\| \right) \\
&= \sum_{q=1}^{Q}\sum_{p=1}^{P} L \left\| \gamma \left[(i-1)\mu_{qp}^{t} + v_{qp}^{t,1} + \cdots + v_{qp}^{t,i-1} \right] \right\|. \qquad (56)
\end{aligned}
$$

By using the law of iterated expectation and (56) we get

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v^{t,i}\right\| |\mathcal{F}^t\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|v^{t,i}\right\| |\mathcal{F}^t, , \mathcal{B}^t, \mathcal{C}^t, \mathcal{D}^t, \pi, j_{11}^{(1)} \cdots, j_{QP}^{(i-2)}\right] |\mathcal{F}^t\right] \\
&\leq \mathbb{E}\left[\sum_{q=1}^{Q}\sum_{p=1}^{P} L \left\| \gamma \left[(i-1)\mu_{qp}^{t} + \sum_{j=1}^{i-1} v_{qp}^{t,j} \right] \right\| \middle| \mathcal{F}^t \right] \\
&\leq L\gamma \sum_{q=1}^{Q}\sum_{p=1}^{P} \left(\mathbb{E}\left[\left\| (i-1)\mu_{qp}^{t} \right\| \middle| \mathcal{F}^t \right] + \sum_{j=1}^{i-1} \mathbb{E}\left[\left\| v_{qp}^{t,j} \right\| |\mathcal{F}^t \right] \right) \qquad (57) \\
&\leq L\gamma QP \left((i-1)\,\mathbb{E}\left[\left\| \mu^{t} \right\| \middle| \mathcal{F}^t \right] + \sum_{j=1}^{i-1} \mathbb{E}\left[\left\| v^{t,j} \right\| |\mathcal{F}^t \right] \right).
\end{aligned}
$$

Let us define

$$
\begin{aligned}
a_i &:= \mathbb{E}\left[\left\| v^{t,i} \right\| \middle| \mathcal{F}^t \right] \\
\nu &= L\gamma QP \\
D_1 &:= \mathbb{E}\left[\left\| \mu^{t} \right\| \middle| \mathcal{F}^t \right].
\end{aligned}
$$

Then the recursive formula becomes

$$
\begin{aligned}
a_1 &= 0 \\
a_i &\leq \nu \left((i-1)D_1 + \sum_{j=1}^{i-1} a_j \right), \qquad (58)
\end{aligned}
$$

for $i = 2, 3, \cdots, B$. Let us define $\bar{a}_i$ as

$$\bar{a}_i = \begin{cases} 0, & i = 1 \\ \nu \left((i-1)D_1 + \sum_{j=1}^{i-1} \bar{a}_j \right), & i \neq 1. \end{cases} \tag{59}$$

Now, let us show that $a_i \leq \bar{a}_i$ for $i = 1, 2, \cdots, B$ by induction. When $i = 1$, applying the definitions of a_i and $\bar{a}_i$ yields $a_1 = \bar{a}_1$. Assume that when $i = 1, 2, \cdots, k-1$, $a_i \leq \bar{a}_i$ holds true. Now, consider $\bar{a}_k$. Since $\nu, D_1, a_i \geq 0$ for any i, by using (58) and (59) we have

$$\bar{a}_k = \nu \left((k-1)D_1 + \sum_{j=1}^{k-1} \bar{a}_j \right) \geq \nu \left((k-1)D_1 + \sum_{j=1}^{k-1} a_j \right) \geq a_k.$$

Therefore, $S_l \leq \bar{S}_l$, where we define $S_l = \sum_{i=1}^{l} a_i$ and $\bar{S}_l = \sum_{i=1}^{l} \bar{a}_i$. Summing up all the recursive equations for $\bar{a}_i$ in (59) up to l implies

$$\bar{S}_l = \frac{l(l-1)}{2} \nu D_1 + \nu \left(\bar{S}_1 + \cdots + \bar{S}_{l-1} \right) \tag{60}$$

and

$$\bar{S}_{l+1} - \bar{S}_l = l\nu D_1 + \nu \bar{S}_l,$$

which in turn yields

$$\frac{\bar{S}_{l+1}}{(1+\nu)^{l+1}} - \frac{\bar{S}_l}{(1+\nu)^l} = \frac{l\nu D_1}{(1+\nu)^{l+1}}.$$

By summing up all increments for $l = 1, \cdots, B-1$, we obtain

$$\frac{\bar{S}_B}{(1+\nu)^B} = \frac{\bar{S}_B}{(1+\nu)^B} - \frac{\bar{S}_1}{(1+\nu)} = \frac{\nu D_1}{(1+\nu)} \sum_{l=1}^{B} \frac{l}{(1+\nu)^l} = \frac{D_1 \left[(1+\nu)^B - 1 - \nu B \right]}{\nu (1+\nu)^B},$$

which in turn yields

$$\sum_{i=1}^{B} \mathbb{E} \left[\left\| v^{t,i} \right\| \middle| \mathcal{F}^t \right] = \sum_{i=1}^{B} a_i = S_B \leq \bar{S}_B = \frac{D_1 \left[(1+\nu)^B - 1 - \nu B \right]}{\nu}.$$

Since $\binom{B}{l} = \frac{B!}{l!(B-l)!} \leq B^l$, we obtain

$$(1+\nu)^B = \sum_{l=0}^{B} \binom{B}{l} \nu^l \leq \sum_{l=0}^{B} (B\nu)^l,$$

which in turn yields

$$\sum_{i=1}^{B} \mathbb{E} \left[\left\| v^{t,i} \right\| \middle| \mathcal{F}^t \right] \leq \frac{D_1 \sum_{l=2}^{B} (B\nu)^l}{\nu}. \tag{61}$$

Substituting the definitions of ν and C back into (61) gives

$$\sum_{i=1}^{B} \mathbb{E}\left[\left\|v^{t,i}\right\| \middle| \mathcal{F}^t\right] \leq \frac{\mathbb{E}\left[\left\|\mu^t\right\| \middle| \mathcal{F}^t\right] \sum_{l=2}^{B} (BL\gamma QP)^l}{L\gamma QP}. \tag{62}$$

By using the conditional Jensen's inequality and (26) we get

$$\mathbb{E}\left[\left\|\mu^t\right\| \middle| \mathcal{F}^t\right] = \mathbb{E}\left[\sqrt{\left\|\mu^t\right\|^2} \middle| \mathcal{F}^t\right] \leq \sqrt{\mathbb{E}\left[\left\|\mu^t\right\|^2 \middle| \mathcal{F}^t\right]} = \sqrt{\hat{\mathcal{O}}(1)} = \hat{\mathcal{O}}(1). \tag{63}$$

The last equality holds due to the property that $\gamma \leq 1$. Moreover, since $BL\gamma QP \leq 1$, we have

$$\sum_{l=2}^{B} (BL\gamma QP)^l = B \cdot \hat{\mathcal{O}}(B^2\gamma^2) = \hat{\mathcal{O}}(B^3\gamma^2). \tag{64}$$

Substituting (106) and (64) in (62) implies the claim in (53).

Now, let us proceed to find an upper bound for $\sum_{i=1}^{B} \mathbb{E}\left[\left\|v^{t,i}\right\|^2 \middle| \mathcal{F}^t\right]$. From (27), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|v^{t,1}\right\|^2 \middle| \mathcal{F}^t\right] = 0 \\
&\mathbb{E}\left[\left\|v^{t,i}\right\|^2 \middle| \mathcal{F}^t\right] \leq iL^2\gamma^2 QP \left((i-1)^2\, \mathbb{E}\left[\left\|\mu^t\right\|^2 \middle| \mathcal{F}^t\right] + \sum_{j=1}^{i-1} \mathbb{E}\left[\left\|v^{t,j}\right\|^2 \middle| \mathcal{F}^t\right] \right).
\end{aligned}$$

Let us define

$$\begin{aligned}
b_i &:= \mathbb{E}\left[\left\|v^{t,i}\right\|^2 \middle| \mathcal{F}^t\right] \\
\theta &:= L^2\gamma^2 QP \\
D_2 &:= \mathbb{E}\left[\left\|\mu^t\right\|^2 \middle| \mathcal{F}^t\right].
\end{aligned}$$

Then the recursive formula becomes

$$\begin{aligned}
b_1 &= 0 \\
b_i &\leq i\theta \left((i-1)^2 D_2 + \sum_{j=1}^{i-1} b_j \right),
\end{aligned} \tag{65}$$

for $i = 2, 3, \cdots, B$. Let us define $\bar{b}_i$ as

$$\bar{b}_i = \begin{cases} 0, & i = 1 \\ i\theta \left((i-1)^2 D_2 + \sum_{j=1}^{i-1} \bar{b}_j \right), & i \neq 1. \end{cases} \tag{66}$$

As before we derive $b_i \leq \bar{b}_i$ for $i = 1, 2, \cdots, B$. Therefore, $\mathcal{S}_l \leq \bar{\mathcal{S}}_l$, where we define $\mathcal{S}_l = \sum_{i=1}^{l} b_i$ and $\bar{\mathcal{S}}_l = \sum_{i=1}^{l} \bar{b}_i$. Summing up all the recursive equations for $\bar{b}_i$ in (66) up to l implies

$$\bar{\mathcal{S}}_l = \theta D_2 \sum_{i=1}^{l-1} (i+1)i^2 + \theta \sum_{i=1}^{l-1} (i+1)\bar{\mathcal{S}}_i, \tag{67}$$

and

$$\bar{\mathcal{S}}_{l+1} - \bar{\mathcal{S}}_l = \theta D_2 (l+1) l^2 + \theta (l+1) \bar{\mathcal{S}}_l,$$

which in turn yields

$$\frac{\bar{\mathcal{S}}_{l+1}}{\Pi_{i=1}^{l+1}(1+i\theta)} - \frac{\bar{\mathcal{S}}_l}{\Pi_{i=1}^{l}(1+i\theta)} = \frac{\theta D_2 (l+1) l^2}{\Pi_{i=1}^{l+1}(1+i\theta)}.$$

By summing up all increments for $l = 1, 2, \cdots, B-1$, we obtain

$$\frac{\bar{\mathcal{S}}_B}{\Pi_{i=1}^{B}(1+i\theta)} = \frac{\bar{\mathcal{S}}_B}{\Pi_{i=1}^{B}(1+i\theta)} - \frac{\bar{\mathcal{S}}_1}{(1+\theta)} = \theta D_2 \left[\sum_{l=1}^{B-1} \frac{(l+1) l^2}{\Pi_{i=1}^{l+1}(1+i\theta)} \right],$$

which in turn yields

$$\begin{aligned} \sum_{i=1}^{B} \mathbb{E}\left[\left\| v^{t,i} \right\|^2 \middle| \mathcal{F}^t \right] = \sum_{i=1}^{B} b_i = \mathcal{S}_l \leq \bar{\mathcal{S}}_B &= \theta D_2 \left[\sum_{l=1}^{B-1} \left((l+1) l^2 \Pi_{i=l+2}^{B} (1+i\theta) \right) \right] \\ &\leq \theta D_2 B (B-1)^2 \left[\sum_{l=1}^{B-1} \left(\Pi_{i=l+2}^{B} (1+i\theta) \right) \right], \end{aligned} \tag{68}$$

where we denote $\Pi_{i=B}^{B+1}(1+i\theta) = 1$. Since

$$\sum_{l=1}^{B-1} \left(\Pi_{i=l+2}^{B} (1+i\theta) \right) \leq \sum_{l=1}^{B-1} \left(\Pi_{i=1}^{B} (1+i\theta) \right) \leq \sum_{l=1}^{B-1} (1+B\theta)^B = (B-1)(1+B\theta)^B,$$

it in turn yields

$$\sum_{i=1}^{B} \mathbb{E}\left[\left\| v^{t,i} \right\|^2 \middle| \mathcal{F}^t \right] \leq \theta D_2 B (B-1)^3 (1+B\theta)^B. \tag{69}$$

Substituting the definitions of θ and D_2 back into (69) gives

$$\sum_{i=1}^{B} \mathbb{E}\left[\left\| v^{t,i} \right\|^2 \middle| \mathcal{F}^t \right] \leq L^2 \gamma^2 Q P \, \mathbb{E}\left[\left\| \mu^t \right\|^2 \middle| \mathcal{F}^t \right] B^4 (B L^2 \gamma^2 Q P + 1)^B. \tag{70}$$

Since $BL\gamma QP \leq 1$ and $QP \geq 1$, we conclude

$$B^2 L^2 \gamma^2 Q P \leq (BL\gamma QP)^2 \leq 1,$$

which in turn yields

$$\begin{aligned} (1 + BL^2\gamma^2 QP)^B &= 1 + \sum_{i=1}^{B} \binom{B}{i} (BL^2\gamma^2 QP)^i \leq 1 + \sum_{i=1}^{B} B^i (BL^2\gamma^2 QP)^i \\ &= 1 + \sum_{i=1}^{B} (B^2 L^2 \gamma^2 Q P)^i = 1 + \sum_{i=1}^{B} \hat{\mathcal{O}}(B^2 L^2 \gamma^2 Q P) \\ &= 1 + \hat{\mathcal{O}}(B^3 L^2 \gamma^2 Q P). \end{aligned} \tag{71}$$

Combining (26) and (71) gives

$$\sum_{i=1}^{B} \mathbb{E}\left[\left\| v^{t,i} \right\|^2 \middle| \mathcal{F}^t \right] = \hat{\mathcal{O}}(B^4\gamma^2)(1 + \hat{\mathcal{O}}(B^3L^2\gamma^2QP)) = \hat{\mathcal{O}}(B^4\gamma^2) + \hat{\mathcal{O}}(B^7\gamma^4).$$

This completes the proof of the claim. ■

By using (13), (29), (30), (31) and Lipschitz continuity of $\nabla F(\omega)$ we have

$$\begin{aligned}
&\mathbb{E}\left[F(\omega^{t+1})|\mathcal{F}^t\right] \leq F(\omega^t) + \nabla F(\omega^t)^T \mathbb{E}\left[\left(\omega^{t+1} - \omega^t\right)|\mathcal{F}^t\right] + \frac{L}{2}\mathbb{E}\left[\left\|\omega^{t+1} - \omega^t\right\|^2 |\mathcal{F}^t\right] \\
&= F(\omega^t) + \nabla F(\omega^t)^T \left\{ -\gamma B \frac{c^t}{d} \nabla F(\omega^t) - \gamma \sum_{i=1}^{B} \mathbb{E}[v^{t,i} \mid \mathcal{F}^t] \right\} \\
&\qquad + \frac{L}{2}\gamma^2(B+1)\left\{ B^2 \mathbb{E}\left[\left\|\mu^t\right\|^2 |\mathcal{F}^t\right] + \sum_{i=1}^{B} \mathbb{E}\left[\left\|v^{t,i}\right\|^2 |\mathcal{F}^t\right] \right\} \\
&\leq F(\omega^t) - \gamma \frac{B}{d} \left\|\nabla F(\omega^t)\right\|^2 + \gamma \left\|\nabla F(\omega^t)\right\| \sum_{i=1}^{B} \mathbb{E}\left[\left\|v^{t,i}\right\| |\mathcal{F}^t\right] \\
&\qquad + \frac{L}{2}\gamma^2(B+1)\left\{ B^2 \mathbb{E}\left[\left\|\mu^t\right\|^2 |\mathcal{F}^t\right] + \sum_{i=1}^{B} \mathbb{E}\left[\left\|v^{t,i}\right\|^2 |\mathcal{F}^t\right] \right\} \\
&\leq F(\omega^t) - \gamma \frac{B}{d} \left\|\nabla F(\omega^t)\right\|^2 + \hat{\mathcal{O}}(B^3\gamma^2) + \hat{\mathcal{O}}(B\gamma^2)\left\{ \hat{\mathcal{O}}(B^2) + \hat{\mathcal{O}}(B^4\gamma^2) + \hat{\mathcal{O}}(B^7\gamma^4) \right\} \\
&= F(\omega^t) - \gamma \frac{B}{d} \left\|\nabla F(\omega^t)\right\|^2 + \hat{\mathcal{O}}(B^3\gamma^2) + \hat{\mathcal{O}}(B^5\gamma^4) + \hat{\mathcal{O}}(B^8\gamma^6).
\end{aligned}$$

Since $LQP \geq 1$ and $BL\gamma QP \leq 1$, the above equation becomes

$$\mathbb{E}\left[F(\omega^{t+1})|\mathcal{F}^t\right] \leq F(\omega^t) - \gamma \frac{B}{d} \left\|\nabla F(\omega^t)\right\|^2 + \hat{\mathcal{O}}(B^4\gamma^2). \tag{72}$$

Subtracting $F(\omega^*)$ from both sides of (72) and applying (36) yields the claim in (50), where C_2 is a positive constant. ■

Proof of Theorem 3

Proof We use the relationship in (50) to construct a supermartingale sequence. Define the stochastic processes α^t and β^t as

$$\alpha^t := \left[F(\omega^t) - F(\omega^*)\right] \times \mathbb{1}_{\left\{\min_{u \leq t} F(\omega^u) - F(\omega^*) > \frac{C_2 dB^3 \gamma}{2\xi}\right\}} \tag{73}$$

$$\beta^t := \frac{2\xi B}{d}\gamma \left[F(\omega^t) - F(\omega^*) - \frac{C_2 dB^3 \gamma}{2\xi}\right] \times \mathbb{1}_{\left\{\min_{u \leq t} F(\omega^u) - F(\omega^*) > \frac{C_2 dB^3 \gamma}{2\xi}\right\}}. \tag{74}$$

The process α^t tracks the optimality gap $F(\omega^t) - F(\omega^*)$ until the gap becomes smaller than $\frac{C_2 dB^3 \gamma}{2\xi}$ for the first time. Notice that the stochastic process α^t is never negative. Likewise,

the same properties hold for β^t. Based on the relationship in (50) and the definitions of stochastic processes α^t and β^t in (73) and (74), we obtain that for all $t \geq 0$

$$\mathbb{E}\left[\alpha^{t+1} \middle| \mathcal{F}^t\right] \leq \alpha^t - \beta^t. \tag{75}$$

Given the relationship in (75) and non-negativity of stochastic processes α^t and β^t we obtain that α^t is supermartingale. The supermartingale convergence theorem yields

$$\begin{aligned} &\text{(i)} \quad \alpha^t \text{ converges to a limit a.s., and} \\ &\text{(ii)} \quad \sum_{t=1}^{\infty} \beta^t < \infty. \quad \text{a.s.} \end{aligned} \tag{76}$$

Property (76) implies that the sequence β^t is converging to 0 almost surely, i.e.,

$$\lim_{t \to \infty} \beta^t = 0 \quad \text{a.s.} \tag{77}$$

Based on the definition of β^t in (74), the limit in (77) is true if one of the following events holds:

$$\begin{aligned} &\text{(i)} \quad \text{the indicator function is 0 after large } t, \\ &\text{(ii)} \quad \lim_{t \to \infty} F(\omega^t) - F(\omega^*) - \frac{C_2 d B^3 \gamma}{2\xi} = 0. \end{aligned}$$

From either one of these two events we conclude that

$$\liminf_{t \to \infty} F(\omega^t) - F(\omega^*) \leq \frac{C_2 d B^3 \gamma}{2\xi} \quad \text{a.s.} \tag{78}$$

Therefore, the claim in (5) is valid. The result in (78) shows that the loss function value sequence $F(\omega^t)$ almost sure converges to a neighborhood of the optimal loss function value $F(\omega^*)$.

We proceed to prove the result in (6). We compute the expected value of (50) given $\mathcal{F}^0$ to obtain

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right) \mathbb{E}\left[F(\omega^t) - F(\omega^*)\right] + C_2 B^4 \gamma^2. \tag{79}$$

Rewriting the relationship in (79) for step $t-1$ gives

$$\mathbb{E}\left[F(\omega^t) - F(\omega^*)\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right) \mathbb{E}\left[F(\omega^{t-1}) - F(\omega^*)\right] + C_2 B^4 \gamma^2. \tag{80}$$

Substituting the upper bound in (80) for the expectation of $F(\omega^t) - F(\omega^*)$ in (79) implies

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right)^2 \mathbb{E}\left[F(\omega^{t-1}) - F(\omega^*)\right] + C_2 B^4 \gamma^2 \left(1 + \left(1 - \frac{2\xi B}{d}\gamma\right)\right). \tag{81}$$

By recursively applying steps (80) and (81) we can bound the expectation of $F(\omega^{t+1})-F(\omega^*)$ in terms of the initial loss function error $F(\omega^0)-F(\omega^*)$ as

$$\mathbb{E}\left[F(\omega^{t+1})-F(\omega^*)\right] \leq \left(1-\frac{2\xi B}{d}\gamma\right)^{t+1}\left[F(\omega^0)-F(\omega^*)\right] + C_2 B^4\gamma^2 \sum_{u=0}^{t}\left(1-\frac{2\xi B}{d}\gamma\right)^{u}. \tag{82}$$

Substituting t by $t-1$ and simplifying the sum in the right-hand side of (82) yields

$$\mathbb{E}\left[F(\omega^{t})-F(\omega^*)\right] \leq \left(1-\frac{2\xi B}{d}\gamma\right)^{t}\left[F(\omega^0)-F(\omega^*)\right] + \frac{C_2 d B^3\gamma}{2\xi}\left[1-\left(1-\frac{2\xi B}{d}\gamma\right)^{t}\right]. \tag{83}$$

Since $\gamma < \frac{d}{2\xi B}$, the term $1-\left(1-\frac{2\xi B}{d}\gamma\right)^t$ in the right-hand side of (83) is strictly smaller than 1 and the claim in (6) follows. ■

E Counter Example without Assumption 4 and 5

Proof of Theorem 4

Proof We consider the setting where there are two samples $[A|b] = \left[\begin{array}{cc|c} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \end{array}\right]$, which are split into four partitions, and the parameter vector is specified as $\omega = [\omega_1, \omega_2]$. Then, applying MSE and linear regression yields

$$F([\omega_1,\omega_2]) = \frac{1}{2}\left\|A\omega - b\right\|_2^2. \tag{84}$$

Consequently, the gradient is

$$\bigtriangledown F = A^T A\omega - A^T b,$$

and the Hessian $H = A^T A$. Note that Assumption 2 holds. For Assumption 3, we can select $L = \max_{i,j}|H_{ij}|$. Notice that

$$\left\|A^T A\right\| = \|A\|^2 \geq \left(\frac{1}{\sqrt{2}}\|A\|_F\right)^2 = \frac{a_{11}^2+a_{12}^2+a_{21}^2+a_{22}^2}{2}. \tag{85}$$

Let us consider the inner loop in steps 12-18. The approximate individual loss functions using each sample are

$$f_1(\omega_1) = \frac{1}{2}(\omega_1 a_{11} - b_1)^2, \quad f_1(\omega_2) = \frac{1}{2}(\omega_2 a_{12} - b_1)^2,$$
$$f_2(\omega_1) = \frac{1}{2}(\omega_1 a_{21} - b_2)^2, \quad f_2(\omega_2) = \frac{1}{2}(\omega_2 a_{22} - b_2)^2,$$

which in turn yields

$$\begin{aligned}
&\bigtriangledown_{\omega_1} f_1(\omega_1 a_{11}) = a_{11}^2\omega_1 - a_{11}b_1, \quad \bigtriangledown_{\omega_2} f_1(\omega_2 a_{12}) = a_{12}^2\omega_2 - a_{12}b_1,\\
&\bigtriangledown_{\omega_1} f_2(\omega_1 a_{21}) = a_{21}^2\omega_1 - a_{21}b_2, \quad \bigtriangledown_{\omega_2} f_2(\omega_2 a_{22}) = a_{22}^2\omega_2 - a_{22}b_2.
\end{aligned}$$

The update formulas in the algorithm for the first sample and ω_1 read

$$\begin{aligned}
\bar{\omega}_1^{(i+1)} &= \bar{\omega}_1^{(i)} - \gamma_{t+1}\left(a_{11}^2\bar{\omega}_1^{(i)} - a_{11}^2\omega_1^t + \mu_1^t\right)\\
\bar{\omega}_1^{(i+1)} &= \left(1 - a_{11}^2\gamma_{t+1}\right)\bar{\omega}_1^{(i)} + \gamma_{t+1}\left(a_{11}^2\omega_1^t - \mu_1^t\right)\\
\bar{\omega}_1^{(i+1)} - \left(\omega_1^t - \frac{\mu_1^t}{a_{11}^2}\right) &= \left(1 - a_{11}^2\gamma_{t+1}\right)\left(\bar{\omega}_1^{(i)} - \left(\omega_1^t - \frac{\mu_1^t}{a_{11}^2}\right)\right).
\end{aligned}$$

Therefore,

$$\bar{\omega}_1^{(i+1)} = \left(1 - a_{11}^2\gamma_{t+1}\right)^{i+1}\left(\bar{\omega}_1^{(0)} - \left(\omega_1^t - \frac{\mu_1^t}{a_{11}^2}\right)\right) + \left(\omega_1^t - \frac{\mu_1^t}{a_{11}^2}\right).$$

Since $\bar{\omega}_1^{(0)} = \omega_1^t$ due to step 13, we obtain

$$\bar{\omega}_1^{(i+1)} = \left(1 - a_{11}^2\gamma_{t+1}\right)^{i+1}\frac{\mu_1^t}{a_{11}^2} + \left(\omega_1^t - \frac{\mu_1^t}{a_{11}^2}\right).$$

If $\gamma_t \leq \frac{1}{\min\left\{a_{11}^2, a_{12}^2, a_{21}^2, a_{22}^2\right\}}$, then $a_{11}^2\gamma_t < 1$, which in turn yields

$$\lim_{i\to\infty}\bar{\omega}_1^{(i+1)} = \omega_1^t - \frac{\mu_1^t}{a_{11}^2}.$$

Thus, if the number of iterations B for the inner loop is big enough, then, approximately, $\omega_1^{t+1} = \omega_1^t - \frac{\mu_1^t}{a_{11}^2}$. Similarly, when using the same data point to update ω_2, we obtain

$$\omega_2^{t+1} = \omega_2^t - \frac{\mu_2^t}{a_{12}^2}.$$

When using $([a_{21}, a_{22}], b_2)$ to update ω_1 and ω_2, we obtain

$$\omega_1^{t+1} = \omega_1^t - \frac{\mu_1^t}{a_{21}^2}, \quad \omega_2^{t+1} = \omega_2^t - \frac{\mu_2^t}{a_{22}^2}.$$

Therefore, the inner loop mimics gradient descent with constant learning rate. Then, the explicit update formula for ω^t is

$$\omega^{t+1} = \omega^t - \eta \bigtriangledown F(\omega^t) = (I - \eta A^T A)\omega^t + \eta A^T b, \tag{86}$$

where the second equality holds due to the definition of $\bigtriangledown F$, and $\eta = \begin{bmatrix} \frac{1}{a_{11}^2} & 0 \\ 0 & \frac{1}{a_{12}^2} \end{bmatrix}$ when using the first sample and $\eta = \begin{bmatrix} \frac{1}{a_{21}^2} & 0 \\ 0 & \frac{1}{a_{22}^2} \end{bmatrix}$ when using the second sample. Thus, if $|a_{11}| = |a_{12}| =$

$\min\{|a_{11}|, |a_{12}|, |a_{21}|, |a_{22}|\}$ and $\max\{|a_{21}|, |a_{22}|\} > |a_{11}|$, then, when using $([a_{11}, a_{12}], b_1)$ to update ω_1, the corresponding $\left\|I - \eta A^T A\right\| > 1$ since $\left\|\eta A^T A\right\| \geq \frac{\left\|A^T A\right\|}{\|\eta^{-1}\|} > \left|\frac{a_{11}^2+a_{12}^2+a_{21}^2+a_{22}^2}{2a_{11}^2}\right| = 2$, which in turn yields the divergence of the algorithm. Similarly, the same conclusion applies to $([a_{21}, a_{22}], b_2)$ when $|a_{21}| = |a_{22}| = \min\{|a_{11}|, |a_{12}|, |a_{21}|, |a_{22}|\}$ and $\max\{|a_{11}|, |a_{12}|\} > |a_{21}|$. Some possible values of A and b are in Table 4.

a_{11}	a_{12}	b_1	a_{21}	a_{22}	b_2	optimal ω^*	ω^{100}
1	1	1	2	3	0	$[3, -1]$	$[3.651 \times 10^{55}, -6.811 \times 10^{56}]$
2	1	1	1	1	0	$[1, -1]$	$[54.606, -29.148]$
1	2	1	1	3	0	$[3, -1]$	$[-4.414 \times 10^{11}, -8.765 \times 10^{11}]$
1	2	1	2	3	1	$[-27, 17]$	$[4.973 \times 10^{29}, -3,455 \times 10^{30}]$
1	4	1	2	3	0	$[-\frac{3}{5}, \frac{2}{5}]$	$[-177.419, 2976.815]$

Table 4: Counter Examples

■

F Diminishing Learning Rate Convergence with Feature Sampling

We assume that ω^* is the unique optimal solution to (1), and also that $\|\omega^*\| \leq \frac{M_2}{2}$. By using Assumption 4, we conclude that the distance between any $\omega \in \Omega$ and ω^* is bounded, i.e.

$$\left\|\omega^t - w^*\right\| \leq M_2. \tag{87}$$

The second moment of ω^t is also bounded for all t, i.e.

$$\left\|\omega^t\right\|^2 \leq \frac{M_2^2}{4}, \tag{88}$$

for any t. Let us define $\mu^t = \frac{1}{\mathscr{d}^t}\sum_{j\in\mathscr{D}^t} \bar{\nabla}_{\omega_{\mathscr{C}^t}} f_j(x_j\omega^t) + e^t$, where e^t is defined as

$$e^t := \frac{1}{\mathscr{d}^t}\sum_{j\in\mathscr{D}^t} \bar{\nabla}_{\omega_{\mathscr{C}^t}} f_j(x_j^{\mathscr{B}^t}\omega_{\mathscr{B}^t}^t) - \frac{1}{\mathscr{d}^t}\sum_{j\in\mathscr{D}^t} \bar{\nabla}_{\omega_{\mathscr{C}^t}} f_j(x_j\omega^t). \tag{89}$$

Lemma 3 *If Assumptions 3-5 hold, then* $\|\nabla F(\omega^t)\|$ *and* $\sum_{j=1}^N \|\nabla f_j(x_j\omega^t)\|^2$ *for any* t *satisfy*

$$\left\|\nabla F(\omega^t)\right\| \leq M_2 L, \tag{90}$$

$$\sum_{j=1}^N \left\|\nabla f_j(x_j\omega^t)\right\|^2 \leq (N-1)G^2 + NM_2^2L^2. \tag{91}$$

Proof Using the fact that ω^* is the optimal solution and Assumptions 3 and 4 we obtain

$$\left\|\nabla F(\omega^t)\right\| = \left\|\nabla F(\omega^t) - \nabla F(\omega^*)\right\| \leq L\left\|\omega^t - \omega^*\right\| \leq LM_2.$$

The last inequality holds due to (87). Assumption 5 implies that, for any ω^t we have

$$\frac{1}{N-1}\sum_{j=1}^{N}\left(\left\|\nabla f_j(x_j\omega^t)\right\|^2-\left\|\nabla F(\omega^t)\right\|^2\right)\leq G^2,$$

which in turn yields

$$\frac{1}{N-1}\sum_{j=1}^{N}\left\|\nabla f_j(x_j\omega^t)\right\|^2\leq G^2+\frac{N}{N-1}\left\|\nabla F(\omega^t)\right\|^2,$$

and

$$\sum_{j=1}^{N}\left\|\nabla f_j(x_jw^t)\right\|^2\leq (N-1)G^2+N\left\|\nabla F(\omega^t)\right\|^2. \tag{92}$$

Inserting (90) into (92) gives (91). ■

Claim 4 *If Assumptions 3 and 4 hold and for some constant $\eta\geq 0$ and the learning rate γ_{t+1}, we have*

$$\mathcal{b}^t\in\left[\max\left\{c^t,\frac{d}{1+\frac{4d\eta\gamma_{t+1}^2}{c^tM_2^2L^2}}\right\},d\right] \tag{93}$$

for every t, then the expected value of $\left\|e^t\right\|^2$ conditioned on $\mathcal{F}^t$ generated by SODDA is bounded by $\eta\gamma_{t+1}^2$, that is

$$\mathbb{E}\left[\left\|e^t\right\|^2|\mathcal{F}^t\right]\leq\eta\gamma_{t+1}^2,$$

where η is a constant unrelated to B.

Proof By using (89) we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|e^t\right\|^2|\mathcal{F}^t\right]&=\mathbb{E}\left[\left\|\frac{1}{\mathcal{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j^{\mathcal{B}^t}\omega_{\mathcal{B}^t}^t)-\frac{1}{\mathcal{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right]\\
&=\frac{1}{(\mathcal{d}^t)^2}\mathbb{E}\left[\left\|\sum_{j\in\mathcal{D}^t}\left(\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j^{\mathcal{B}^t}\omega_{\mathcal{B}^t}^t)-\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j\omega^t)\right)\right\|^2\middle|\mathcal{F}^t\right]\\
&\leq\frac{1}{\mathcal{d}^t}\mathbb{E}\left[\sum_{j\in\mathcal{D}^t}\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j^{\mathcal{B}^t}\omega_{\mathcal{B}^t}^t)-\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right]\\
&=\frac{1}{\mathcal{d}^t}\mathbb{E}\left[\sum_{j\in\mathcal{D}^t}\mathbb{E}\left[\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j^{\mathcal{B}^t}\omega_{\mathcal{B}^t}^t)-\bar{\nabla}_{\omega_{\mathcal{C}^t}}f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t,\mathcal{B}^t,\mathcal{D}^t\right]\middle|\mathcal{F}^t\right].
\end{aligned}\tag{94}$$

Applying Lemma 1 with $w_1 = 1$, $w_2 = 0$,$\Phi = \left\{ \left(\bar{\nabla}_{\omega_{\mathcal{B}^t}} f_j(x_j^{\mathcal{B}^t} \omega_{\mathcal{B}^t}^t) - \bar{\nabla}_{\omega_{\mathcal{B}^t}} f_j(x_j \omega^t) \right)_i \right\}_{i=1}^{\mathcal{b}^t}$, $g(z) = z^2$, $\mathcal{H} = \sigma(\mathcal{F}^t, \mathcal{B}^t, \mathcal{D}^t)$ and $\mathcal{B} = \mathcal{C}^t$ to (94) yields

$$\begin{aligned}
\mathbb{E}\left[\left\| e^t \right\|^2 | \mathcal{F}^t \right] &\leq \frac{c^t}{\mathcal{b}^t \mathcal{d}^t} \mathbb{E}\left[\sum_{j \in \mathcal{D}^t} \left\| \bar{\nabla}_{\omega_{\mathcal{B}^t}} f_j(x_j^{\mathcal{B}^t} \omega_{\mathcal{B}^t}^t) - \bar{\nabla}_{\omega_{\mathcal{B}^t}} f_j(x_j \omega^t) \right\|^2 \middle| \mathcal{F}^t \right] \\
&= \frac{c^t}{\mathcal{b}^t \mathcal{d}^t} \mathbb{E}\left[\sum_{j \in \mathcal{D}^t} \left\| \bar{\nabla}_{\omega_{\mathcal{B}^t}} f_j(x_j \omega_{(\mathcal{B}^t,0)}^t) - \bar{\nabla}_{\omega_{\mathcal{B}^t}} f_j(x_j \omega^t) \right\|^2 \middle| \mathcal{F}^t \right] \\
&\leq \frac{L^2 c^t}{\mathcal{b}^t \mathcal{d}^t} \mathbb{E}\left[\sum_{j \in \mathcal{D}^t} \left\| \omega_{(\mathcal{B}^t,0)}^t - \omega^t \right\|^2 \middle| \mathcal{F}^t \right] \\
&= \frac{L^2 c^t}{\mathcal{b}^t \mathcal{d}^t} \mathbb{E}\left[\mathbb{E}\left[\sum_{j \in \mathcal{D}^t} \left\| \omega_{[d] \setminus \mathcal{B}^t}^t \right\|^2 \middle| \mathcal{F}^t, \mathcal{B}^t \right] \middle| \mathcal{F}^t \right] \\
&= \frac{L^2 c^t}{\mathcal{b}^t \mathcal{d}^t} \mathbb{E}\left[\mathbb{E}\left[\mathcal{d}^t \left\| \omega_{[d] \setminus \mathcal{B}^t}^t \right\|^2 \middle| \mathcal{F}^t, \mathcal{B}^t \right] \middle| \mathcal{F}^t \right] \\
&= \frac{L^2 c^t}{\mathcal{b}^t} \mathbb{E}\left[\left\| \omega_{[d] \setminus \mathcal{B}^t}^t \right\|^2 \middle| \mathcal{F}^t \right].
\end{aligned}$$

The second inequality uses Assumption 3 and we use $[d] = \{1, \cdots, d\}$. Now, let us use Lemma 1 again with $w_1 = 0$, $w_2 = 1$, $\Phi = \{(\omega^t)_i\}_{i=1}^d$, $g(z) = z^2$, $\mathcal{H} = \mathcal{F}^t$ and $\mathcal{B} = \mathcal{B}^t$ to get

$$\begin{aligned}
\mathbb{E}\left[\left\| e^t \right\|^2 | \mathcal{F}^t \right] &\leq \frac{L^2 c^t}{\mathcal{b}^t} \frac{d - \mathcal{b}^t}{d} \left\| \omega^t \right\|^2 \\
&\leq \frac{L^2 c^t (d - \mathcal{b}^t)}{d \mathcal{b}^t} \frac{M_2^2}{4}.
\end{aligned}$$

The last inequality uses (88). In order to bound the expected value of $\|e^t\|^2$ by $\eta \gamma_{t+1}^2$, we require

$$\frac{L^2 c^t (d - \mathcal{b}^t)}{d \mathcal{b}^t} \frac{M_2^2}{4} \leq \eta \gamma_{t+1}^2,$$

which is equivalent to

$$\mathcal{b}^t \geq \frac{d}{1 + \frac{4 d \eta \gamma_{t+1}^2}{c^t M_2^2 L^2}}.$$

Meanwhile, in order to make $\nabla_{\omega_{\mathcal{C}^t}} f_j(x_j^{\mathcal{B}^t} \omega_{\mathcal{B}^t}^t)$ well defined, we require $\mathcal{b}^t \geq c^t$. ■

Next, similar to Proposition 1, we present the following proposition.

Proposition 3 *If Assumptions 2-5 hold, and the sequence of learning rates satisfies $\gamma_t \leq 1$ for all t, and the sequences $(\mathcal{b}^t, c^t, \mathcal{d}^t)_{t=0}^{\infty}$ are selected so that (93) for some constant $\eta \geq 0$,*

$c^t \leq d$ and $\mathscr{d}^t \leq N$, then the loss function error sequence $F(\omega^t) - F(\omega^*)$ generated by SODDA satisfies

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)|\mathcal{F}^t\right] \leq (1 - \frac{2\xi B}{d}\gamma_{t+1})[F(\omega^t) - F(\omega^*)] + C_3\gamma_{t+1}^2, \tag{95}$$

where C_3 is a positive constant.

Proof Since the proof is very similar to the proof of Proposition 1, we point out the differences rather than present all the details.

Claim 5 *For any t we have*

$$\mathbb{E}\left[\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right] \leq \frac{c^t}{Nd}\left[(N-1)G^2 + NM_2^2L^2\right]. \tag{96}$$

Proof Applying the law of iterated expectation and Lemma 1 with $w_1 = 1$, $w_2 = 0$, $\Phi = \left\{\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\}_{j=1}^N$, $g(z) = \|z\|^2$, $\mathcal{H} = \sigma(\mathcal{F}^t, c^t)$ and $\mathcal{B} = \mathcal{D}^t$ give

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right] &\leq \frac{1}{\mathscr{d}^t}\mathbb{E}\left[\sum_{j\in\mathcal{D}^t}\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right] \\
&= \frac{1}{\mathscr{d}^t}\mathbb{E}\left[\mathbb{E}\left[\sum_{j\in\mathcal{D}^t}\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t, \mathcal{C}^t\right] \middle| \mathcal{F}^t\right] = \frac{1}{\mathscr{d}^t}\cdot\frac{\mathscr{d}^t}{N}\mathbb{E}\left[\sum_{j=1}^N\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right] \\
&= \frac{1}{N}\sum_{j=1}^N\mathbb{E}\left[\left\|\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right],
\end{aligned}$$

which in turn yields

$$\mathbb{E}\left[\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathcal{C}^t}} f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right] \leq \frac{1}{N}\sum_{j=1}^N\frac{c^t}{d}\mathbb{E}\left[\left\|\nabla f_j(x_j\omega^t)\right\|^2 \middle| \mathcal{F}^t\right] = \frac{c^t}{Nd}\sum_{j=1}^N\left\|\nabla f_j(x_j\omega^t)\right\|^2, \tag{97}$$

where we apply Lemma 1 for each j again with $w_1 = 1$, $w_2 = 0$, $\Phi = \left\{(\nabla f_j(x_j\omega^t))_i\right\}_{i=1}^d$, $g(z) = z^2$, $\mathcal{H} = \mathcal{F}^t$ and $\mathcal{B} = \mathcal{C}^t$. By inserting (91) from Lemma 3 into (97) the claim in (96) follows. ■

From Claim 4 we conclude

$$\mathbb{E}\left[\left\|e^t\right\|^2|\mathcal{F}^t\right] = \hat{\mathcal{O}}(\gamma_{t+1}^2). \tag{98}$$

By using the conditional Jensen's inequality and (98) we get

$$\mathbb{E}\left[\left\|e^t\right\||\mathcal{F}^t\right] = \mathbb{E}\left[\sqrt{\|e^t\|^2}|\mathcal{F}^t\right] \leq \sqrt{\mathbb{E}\left[\|e^t\|^2|\mathcal{F}^t\right]} \leq \sqrt{\eta}\gamma_{t+1} = \hat{\mathcal{O}}(\gamma_{t+1}). \tag{99}$$

Consequently, applying the definition of μ^t yields

$$\begin{aligned}\mathbb{E}\left[\left\|\mu^t\right\|^2|\mathcal{F}^t\right] &\leq 2\left\{\mathbb{E}\left[\left\|e^t\right\|^2\middle|\mathcal{F}^t\right]+\mathbb{E}\left[\left\|\frac{1}{\mathscr{d}^t}\sum_{j\in\mathcal{D}^t}\bar{\nabla}_{\omega_{\mathscr{C}^t}}f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right]\right\}\\ &= \hat{\mathcal{O}}(\gamma_{t+1}^2)+\hat{\mathcal{O}}(1).\end{aligned} \tag{100}$$

The second equality holds due to (96) in Claim 5 and (98). Then, the conclusion of (27) holds since

$$\begin{aligned}&\mathbb{E}\left[\left\|v^{t,k}\right\|^2|\mathcal{F}^t\right]=\mathbb{E}\left[\mathbb{E}\left[\left\|v^{t,k}\right\|^2|\mathcal{F}^t,,\mathcal{B}^t,\mathcal{C}^t,\mathcal{D}^t,\pi,j_{11}^{(1)}\cdots,j_{QP}^{(k-2)}\right]|\mathcal{F}^t\right]\\ &\leq\mathbb{E}\left[\sum_{q=1}^{Q}\sum_{p=1}^{P}L^2\left\|\gamma_{t+1}\left[(k-1)\mu_{qp}^t+\sum_{i=1}^{k-1}v_{qp}^{t,i}\right]\right\|^2\middle|\mathcal{F}^t\right]\\ &\leq kL^2\gamma_{t+1}^2\sum_{q=1}^{Q}\sum_{p=1}^{P}\left(\mathbb{E}\left[\left\|(k-1)\mu_{qp}^t\right\|^2\middle|\mathcal{F}^t\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\left\|v_{qp}^{t,i}\right\|^2|\mathcal{F}^t\right]\right)\\ &\leq kL^2\gamma_{t+1}^2QP\left((k-1)^2\,\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\left\|v^{t,i}\right\|^2|\mathcal{F}^t\right]\right)\\ &= kL^2\gamma_{t+1}^2QP\left[2(k-1)^2\left(\hat{\mathcal{O}}(\gamma_{t+1}^2)+\hat{\mathcal{O}}(1)\right)+\hat{\mathcal{O}}(\gamma_{t+1}^2)\right]=\hat{\mathcal{O}}(\gamma_{t+1}^2).\end{aligned} \tag{101}$$

Thus, we maintain the same claims as those in Claim 2. Then, (29) is revised to be

$$\begin{aligned}\mathbb{E}\left[\omega^{t+1}-\omega^t|\mathcal{F}^t\right] &= -\gamma_{t+1}\,\mathbb{E}\left[B\mu^t+v^{t,1}+\cdots+v^{t,B}|\mathcal{F}^t\right]\\ &= -\gamma_{t+1}B\left(\mathbb{E}[e^t\mid\mathcal{F}^t]+\frac{c^t}{d}\nabla F(\omega^t)\right)-\gamma_{t+1}\sum_{i=1}^{B}\mathbb{E}[v^{t,i}\mid\mathcal{F}^t],\end{aligned} \tag{102}$$

by using (16) in Claim 1. Moreover, the expected value of the squared norm $\|\omega^{t+1}-\omega^t\|^2$ given $\mathcal{F}^t$ is

$$\begin{aligned}\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right] &= \mathbb{E}\left[\left\|\gamma_{t+1}\left[B\mu^t+v^{t,1}+v^{t,2}+\cdots+v^{t,B}\right]\right\|^2|\mathcal{F}^t\right]\\ &\leq\gamma_{t+1}^2(B+1)\left\{B^2\,\mathbb{E}\left[\left\|\mu^t\right\|^2|\mathcal{F}^t\right]+\sum_{i=1}^{B}\mathbb{E}\left[\left\|v^{t,i}\right\|^2|\mathcal{F}^t\right]\right\}\\ &= \hat{\mathcal{O}}(\gamma_{t+1}^2)\left\{\hat{\mathcal{O}}(\gamma_{t+1}^2)+\hat{\mathcal{O}}(1)+\hat{\mathcal{O}}(\gamma_{t+1}^2)\right\}=\hat{\mathcal{O}}(\gamma_{t+1}^2),\end{aligned} \tag{103}$$

due to Claim 4, (96), (100) and (101), which is identical to (30). Similarly, applying (90) and (99) yields

$$\begin{aligned}-\gamma_{t+1}B\nabla F(\omega^t)^T\,\mathbb{E}\left[e^t|\mathcal{F}^t\right] &\leq \gamma_{t+1}B\left\|\nabla F(\omega^t)\right\|\mathbb{E}\left[\left\|e^t\right\||\mathcal{F}^t\right]\\ &= \hat{\mathcal{O}}(\gamma_{t+1})\cdot\hat{\mathcal{O}}(\gamma_{t+1})=\hat{\mathcal{O}}(\gamma_{t+1}^2).\end{aligned} \tag{104}$$

Therefore, (32) remains the same as follows

$$\begin{aligned}
&\mathbb{E}\left[F(\omega^{t+1})|\mathcal{F}^t\right] \leq F(\omega^t) + \triangledown F(\omega^t)^T \mathbb{E}\left[\left(\omega^{t+1}-\omega^t\right)|\mathcal{F}^t\right] + \frac{L}{2}\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right] \\
&= F(\omega^t) + \triangledown F(\omega^t)^T \left\{-\gamma_{t+1}B\left(\mathbb{E}[e^t \mid \mathcal{F}^t] + \frac{c^t}{d}\triangledown F(\omega^t)\right) - \gamma_{t+1}\sum_{i=1}^{B}\mathbb{E}[v^{t,i} \mid \mathcal{F}^t]\right\} \\
&\qquad + \frac{L}{2}\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right] \\
&= F(\omega^t) - \gamma_{t+1}\frac{c^tB}{d}\left\|\triangledown F(\omega^t)\right\|^2 - \gamma_{t+1}B\triangledown F(\omega^t)^T\mathbb{E}\left[e^t|\mathcal{F}^t\right] - \gamma_{t+1}\triangledown F(\omega^t)^T\sum_{i=1}^{B}\mathbb{E}\left[v^{t,i}|\mathcal{F}^t\right] \\
&\qquad + \frac{L}{2}\mathbb{E}\left[\left\|\omega^{t+1}-\omega^t\right\|^2|\mathcal{F}^t\right] \\
&\leq F(\omega^t) - \gamma_{t+1}\frac{c^tB}{d}\left\|\triangledown F(\omega^t)\right\|^2 + \hat{\mathcal{O}}(\gamma_{t+1}^2) \leq F(\omega^t) - \gamma_{t+1}\frac{c^tB}{d}\left\|\triangledown F(\omega^t)\right\|^2 + C_3\gamma_{t+1}^2,
\end{aligned}$$

where C_3 is a positive constant and we use (31), (102), (103) and (104). The remaining steps are identical with those in Proposition 1. ■

The proof of Theorem 5 is the same as the proof of Theorem 1, and the proof of Theorem 6 is the same as the proof of Theorem 2.

G Constant Learning Rate with Feature Sampling

Proposition 4 *If Assumptions 2-5 hold, and the learning rate is constant $\gamma_t = \gamma$ such that $BL\gamma QP \leq 1$ and $\gamma \leq 1$, and the sequences $(\mathcal{b}^t, c^t, \mathcal{d}^t)_{t=0}^{\infty}$ satisfy the same conditions as in Theorem 5, then the loss function error sequence $F(\omega^t) - F(\omega^*)$ generated by SODDA satisfies*

$$\mathbb{E}\left[F(\omega^{t+1}) - F(\omega^*)\middle|\mathcal{F}^t\right] \leq \left(1 - \frac{2\xi B}{d}\gamma\right)\left[F(\omega^t) - F(\omega^*)\right] + C_4B^4\gamma^2, \tag{105}$$

where C_4 is a positive constant.

Proof The proof is the same as the Proof of Proposition 2 since

$$\mathbb{E}\left[\left\|\mu^t\right\|\middle|\mathcal{F}^t\right] = \mathbb{E}\left[\sqrt{\left\|\mu^t\right\|^2}\middle|\mathcal{F}^t\right] \leq \sqrt{\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right]} = \sqrt{\hat{\mathcal{O}}(1) + \hat{\mathcal{O}}(\gamma^2)} = \hat{\mathcal{O}}(1), \tag{106}$$

due to (100). ■

The proof of Theorem 7 is the same as the proof of Theorem 3.

H Convergence to Optimality of Constant Learning Rate with Feature Sampling

Proof of Theorem 8

Proof Given $\mathcal{b}^t = d$ and $\mathcal{d}^t = N$, applying Claim 1 implies

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\middle|\mathcal{F}^t\right] = \frac{c^t}{N}\nabla F(\omega^t), \tag{107}$$

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right] &= \frac{1}{\binom{d}{c^t}}\sum_{\mathcal{C}^t}\left\|\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\right\|^2 \\ &= \frac{1}{\binom{d}{c^t}}\binom{d-1}{c^t-1}\left\|\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\right\|^2 = \frac{c^t}{d}\left\|\nabla F(\omega^t)\right\|^2,\end{aligned} \tag{108}$$

which in turn gives an upper bound to $\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\right\|\middle|\mathcal{F}^t\right]$; that is

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\right\|\middle|\mathcal{F}^t\right] \leq \sqrt{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^{N}\bar{\nabla}_{\mathcal{C}^t} f_j(x_j\omega^t)\right\|^2\middle|\mathcal{F}^t\right]} = \sqrt{\frac{c^t}{d}}\left\|\nabla F(\omega^t)\right\|. \tag{109}$$

On the one hand, given $BL\gamma QP \leq 1$, we find that

$$\sum_{l=2}^{B}(BL\gamma QP) \leq (B-1)(BL\gamma QP)^2 = (B-1)B^2L^2\gamma^2Q^2P^2. \tag{110}$$

By combining this expression with (62), we conclude that

$$\mathbb{E}\left[\left\|\sum_{i=1}^{B}v^{t,i}\right\|\middle|\mathcal{F}^t\right] \leq \sum_{i=1}^{B}\mathbb{E}\left[\left\|v^{t,i}\right\|\middle|\mathcal{F}^t\right] \leq \mathbb{E}\left[\left\|\mu^t\right\|\middle|\mathcal{F}^t\right](B-1)B^2L\gamma QP. \tag{111}$$

On the other hand, given $\gamma \leq 1$, from expression (71), we find that

$$(1+BL^2\gamma^2QP)^B \leq 1+\sum_{i=1}^{B}(B^2L^2\gamma^2QP)^i = 1+B^3L^2QP. \tag{112}$$

By applying (112) to (70), we deduce that

$$\mathbb{E}\left[\left\|\sum_{i=1}^{B}v^{t,i}\right\|^2\middle|\mathcal{F}^t\right] \leq B\sum_{i=1}^{B}\mathbb{E}\left[\left\|v^{t,i}\right\|^2\middle|\mathcal{F}^t\right] \leq B^5(1+B^3L^2QP)L^2\gamma^2QP\,\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right]. \tag{113}$$

Second, let us evaluate the error after each iteration. By summing up all increments in iteration t, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\omega^{t+1}-\omega^*\right\|^2\middle|\mathcal{F}^t\right] &= \mathbb{E}\left[\left\|\omega^t-\gamma\left(B\mu^t+\sum_{i=1}^{B}v^{t,i}\right)-\omega^*\right\|^2\middle|\mathcal{F}^t\right] \\
&= \left\|\omega^t-\omega^*\right\|^2-2\left\langle\mathbb{E}\left[\gamma\left(B\mu^t+\sum_{i=1}^{B}v^{t,i}\right)\middle|\mathcal{F}^t\right],\omega^t-\omega^*\right\rangle+\mathbb{E}\left[\left\|\gamma\left(B\mu^t+\sum_{i=1}^{B}v^{t,i}\right)\right\|^2\middle|\mathcal{F}^t\right] \\
&\leq \left\|\omega^t-\omega^*\right\|^2-2\gamma B\left\langle\mathbb{E}\left[\mu^t\middle|\mathcal{F}^t\right],\omega^t-\omega^*\right\rangle+2\gamma\,\mathbb{E}\left[\left\|\sum_{i=1}^{B}v^{t,i}\right\|\middle|\mathcal{F}^t\right]\left\|\omega^t-\omega^*\right\| \\
&\quad+2\gamma^2B^2\,\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right]+2\gamma^2\,\mathbb{E}\left[\left\|\sum_{i=1}^{B}v^{t,i}\right\|^2\middle|\mathcal{F}^t\right].
\end{aligned} \tag{114}$$

Based on (107), (108), (109), (111) and (113), (114) can be further simplified as

$$\begin{aligned}
\mathbb{E}\left[\left\|\omega^{t+1}-\omega^*\right\|^2\middle|\mathcal{F}^t\right] &\leq \left\|\omega^t-\omega^*\right\|^2-\frac{2\gamma Bc^t}{d}\left\langle\nabla F(\omega^t),\omega^t-\omega^*\right\rangle+\frac{2\gamma^2B^2c^t}{d}\left\|\nabla F(\omega^t)\right\|^2 \\
&\quad+2(B-1)B^2L\gamma^2QP\,\mathbb{E}\left[\left\|\mu^t\right\|\middle|\mathcal{F}^t\right]\left\|\omega^t-\omega^*\right\|+2B^5(1+B^3L^2QP)L^2\gamma^4QP\,\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right] \\
&\leq \left\|\omega^t-\omega^*\right\|^2-\frac{2\gamma Bc^t}{d}\left\langle\nabla F(\omega^t),\omega^t-\omega^*\right\rangle+\frac{2\gamma^2B^2c^t}{d}\left\|\nabla F(\omega^t)\right\|^2 \\
&\quad+2(B-1)B^2L\gamma^2QP\sqrt{\frac{c^t}{d}}\left\|\nabla F(\omega^t)\right\|\left\|\omega^t-\omega^*\right\|+2B^5(1+B^3L^2QP)L^2\gamma^4QP\frac{c^t}{d}\left\|\nabla F(\omega^t)\right\|^2.
\end{aligned} \tag{115}$$

Recalling the Lipschitz continuity of the gradient of the objective function stated in Assumption 3, we have that [Nesterov (2013),Theorem 2.1.5]

$$\frac{1}{L}\left\|\nabla F(\omega^t)\right\|^2=\frac{1}{L}\left\|\nabla F(\omega^t)-\nabla F(\omega^*)\right\|^2\leq\left\langle\nabla F(\omega^t)-\nabla F(\omega^*),\omega^t-\omega^*\right\rangle. \tag{116}$$

Because $F(\omega)$ is strongly convex by Assumption 2, we have

$$\frac{1}{\xi}\left\|\nabla F(\omega^t)\right\|=\frac{1}{\xi}\left\|\nabla F(\omega^t)-\nabla F(\omega^*)\right\|\geq\left\|\omega^t-\omega^*\right\|. \tag{117}$$

By using (116) and (117), 115 can be reformulated as

$$\begin{aligned}
\mathbb{E}\left[\left\|\omega^{t+1}-\omega^*\right\|^2\middle|\mathcal{F}^t\right] &\leq \left\|\omega^t-\omega^*\right\|^2-\frac{2\gamma Bc^t}{Ld}\left\|\nabla F(\omega^t)\right\|^2+\frac{2\gamma^2B^2c^t}{d}\left\|\nabla F(\omega^t)\right\|^2\\
&\quad+\frac{2(B-1)B^2L\gamma^2QP}{\xi}\sqrt{\frac{c^t}{d}}\left\|\nabla F(\omega^t)\right\|^2+2B^5(1+B^3L^2QP)L^2\gamma^4QP\frac{c^t}{d}\left\|\nabla F(\omega^t)\right\|^2\\
&=\left\|\omega^t-\omega^*\right\|^2+\left(-\frac{2\gamma Bc^t}{Ld}+\frac{2\gamma^2B^2c^t}{d}+\frac{2(B-1)B^2L\gamma^2QP}{\xi}\sqrt{\frac{c^t}{d}}\right.\\
&\quad\left.+2B^5(1+B^3L^2QP)L^2\gamma^4QP\frac{c^t}{d}\right)\left\|\nabla F(\omega^t)\right\|^2\\
&=\left\|\omega^t-\omega^*\right\|^2+A(t)\left\|\nabla F(\omega^t)\right\|^2,
\end{aligned}\tag{118}$$

with

$$A(t)=-\frac{2\gamma Bc^t}{Ld}+\frac{2\gamma^2B^2c^t}{d}+\frac{2(B-1)B^2L\gamma^2QP}{\xi}\sqrt{\frac{c^t}{d}}+2B^5(1+B^3L^2QP)L^2\gamma^4QP\frac{c^t}{d}.$$

Therefore, $\mathbb{E}\left[\left\|\omega^{t+1}-\omega^*\right\|^2\middle|\mathcal{F}^t\right]\leq\left\|\omega^t-\omega^*\right\|^2$ as long as $A(t)\leq 0$ for all t, $BL\gamma QP\leq 1$ and $\gamma\leq 1$.

In view of Assumption 3 we have

$$\begin{aligned}
\mathbb{E}\left[F(\omega^{t+1})\middle|\mathcal{F}^t\right] &\leq \mathbb{E}\left[F(\omega^t)+\left\langle\nabla F(\omega^t),\omega^{t+1}-\omega^t\right\rangle+\frac{L}{2}\left\|\omega^{t+1}-\omega^t\right\|^2\middle|\mathcal{F}^t\right]\\
&=F(\omega^t)+\mathbb{E}\left[\left\langle\nabla F(\omega^t),-\gamma B\mu^t-\gamma\sum_{i=1}^{t}v^{t,i}\right\rangle\middle|\mathcal{F}^t\right]+\frac{L}{2}\mathbb{E}\left[\left\|\gamma B\mu^t+\gamma\sum_{i=1}^{B}v^{t,i}\right\|^2\middle|\mathcal{F}^t\right]\\
&\leq F(\omega^t)-\gamma B\left\langle\nabla F(\omega^t),\mathbb{E}\left[\mu^t\middle|\mathcal{F}^t\right]\right\rangle+\gamma\left\|\nabla F(\omega^t)\right\|\mathbb{E}\left[\left\|\sum_{i=1}^{B}v^{t,i}\right\|\middle|\mathcal{F}^t\right]+L\gamma^2B^2\,\mathbb{E}\left[\left\|\mu^t\right\|^2\middle|\mathcal{F}^t\right]\\
&\quad+L\gamma^2\,\mathbb{E}\left[\left\|\sum_{i=1}^{B}v^{t,i}\right\|^2\middle|\mathcal{F}^t\right].
\end{aligned}\tag{119}$$

Substituting (107), (108), (111) and (113) implies

$$\begin{aligned}
\mathbb{E}\left[F(\omega^{t+1})\middle|\mathcal{F}^t\right] &\leq F(\omega^t) - \frac{\gamma B c^t}{d}\left\|\nabla F(\omega^t)\right\|^2 + (B-1)B^2L\gamma^2QP\sqrt{\frac{c^t}{d}}\left\|\nabla F(\omega^t)\right\|^2 \\
&\quad + \frac{L\gamma^2B^2c^t}{d}\left\|\nabla F(\omega^t)\right\|^2 + B^5(1+B^3L^2QP)L^3\gamma^4QP\frac{c^t}{d}\left\|\nabla F(\omega^t)\right\|^2 \\
&= F(\omega^t) + \left(-\frac{\gamma B c^t}{d} + (B-1)B^2L\gamma^2QP\sqrt{\frac{c^t}{d}} + \frac{L\gamma^2B^2c^t}{d}\right. \\
&\quad \left. + B^5(1+B^3L^2QP)L^3\gamma^4QP\frac{c^t}{d}\right)\left\|\nabla F(\omega^t)\right\|^2 \\
&= F(\omega^t) + B(t)\left\|\nabla F(\omega^t)\right\|^2.
\end{aligned} \tag{120}$$

A similar requirement is needed in (120) as that in (118), i.e. $B(t) < 0$ for all t.

Let us denote $\triangle_t = F(\omega^t) - F(\omega^*)$. Then if $A(t) \leq 0$ for all t, we obtain

$$\triangle_t = F(\omega^t) - F(\omega^*) \leq \left\langle \nabla F(\omega^t), \omega^t - \omega^* \right\rangle \leq \left\|\omega^0 - \omega^*\right\| \left|\nabla F(\omega^t)\right\|.$$

Thus, if $B(t) < 0$ for all t and by using the law of iterated expectation, (120) becomes

$$\mathbb{E}\left[\triangle_{t+1}\right] \leq \mathbb{E}\left[\triangle_k\right] + \min_t B(t)\,\mathbb{E}\left[\left\|\nabla F(\omega^t)\right\|^2\right] \leq \mathbb{E}\left[\triangle_k\right] + \frac{\min_t B(t)}{\left\|\omega^0 - \omega^*\right\|^2}\mathbb{E}\left[\triangle_k^2\right],$$

which in turn yields

$$\frac{1}{\mathbb{E}\left[\triangle_{t+1}\right]} \geq \frac{1}{\mathbb{E}\left[\triangle_t\right]} - \frac{\min_t B(t)}{\left\|\omega^0 - \omega^*\right\|^2}\frac{\mathbb{E}\left[\triangle_t\right]}{\mathbb{E}\left[\triangle_{t+1}\right]} \geq \frac{1}{\mathbb{E}\left[\triangle_t\right]} - \frac{\min_t B(t)}{\left\|\omega^0 - \omega^*\right\|^2}.$$

Summing up these inequalities, we get

$$\frac{1}{\mathbb{E}\left[\triangle_{t+1}\right]} \geq \frac{1}{\triangle_0} - \frac{\min_t B(t)}{\left\|\omega^0 - \omega^*\right\|^2}(t+1),$$

which in turn yields

$$\lim_{t\to\infty}\mathbb{E}\left[\triangle_{t+1}\right] = 0. \tag{121}$$

Hence, (12) follows from (121) and similar reasoning as Theorem 5, provided $A(t) \leq 0$, $B(t) < 0$ for all t and $BL\gamma QP \leq 1$, i.e.

$$\frac{2\gamma Bc^t}{Ld} \geq \frac{2\gamma^2B^2c^t}{d} + \frac{2(B-1)B^2L\gamma^2QP}{\xi}\sqrt{\frac{c^t}{d}} + 2B^5(1+B^3L^2QP)L^2\gamma^4QP\frac{c^t}{d} \tag{122}$$

$$\frac{\gamma Bc^t}{d} > (B-1)B^2L\gamma^2QP\sqrt{\frac{c^t}{d}} + \frac{L\gamma^2B^2c^t}{d} + B^5(1+B^3L^2QP)L^3\gamma^4QP\frac{c^t}{d} \tag{123}$$

$$BL\gamma QP \leq 1 \tag{124}$$

$$\gamma \leq 1. \tag{125}$$

By multiplying (122) and (123) by $\frac{1}{2\gamma B}$ and $\frac{1}{\gamma B}$, respectively, we have that

$$\frac{c^t}{Ld} \geq \frac{\gamma B c^t}{d} + \frac{(B-1)BL\gamma QP}{\xi}\sqrt{\frac{c^t}{d}} + B^4(1+B^3L^2QP)L^2\gamma^3 QP\frac{c^t}{d}, \tag{126}$$

$$\frac{c^t}{d} > (B-1)BL\gamma QP\sqrt{\frac{c^t}{d}} + \frac{L\gamma B c^t}{d} + B^4(1+B^3L^2QP)L^3\gamma^3 QP\frac{c^t}{d}. \tag{127}$$

Note that if we are able to find a constant learning rate γ satisfying

$$\bar{A}_1 = \frac{\min_t c^t}{Ld} \geq \gamma\left[\left(B + \frac{(B-1)BLQP}{\xi}\right) + B^4(1+B^3L^2QP)L^2\gamma^2 QP\right] = \bar{B}_1\gamma + \bar{C}_1\gamma^3 \tag{128}$$

$$\bar{A}_2 = \frac{\min_t c^t}{d} > \gamma\left[((B-1)BLQP + LB) + B^4(1+B^3L^2QP)L^3\gamma^2 QP\right] = \bar{B}_2\gamma + \bar{C}_2\gamma^3, \tag{129}$$

with

$$\begin{aligned}
\bar{A}_1 &= \frac{\min_t c^t}{Ld}\\
\bar{B}_1 &= B + \frac{(B-1)BLQP}{\xi}\\
\bar{C}_1 &= B^4(1+B^3L^2QP)L^2QP\\
\bar{A}_2 &= \frac{\min_t c^t}{d}\\
\bar{B}_2 &= (B-1)BLQP + LB\\
\bar{C}_2 &= B^4(1+B^3L^2QP)L^3QP,
\end{aligned}$$

then the same constant learning rate γ is also valid for (126) and (127). Observing that the right-hand sides of both (128) and (129) have the same form, i.e. they are both cubic equations with 0 being the only real root. Solving (128) and (129) show that

$$\gamma \in (0, \min\{\gamma_1, \gamma_2\}),$$

where

$$\begin{aligned}
\gamma_1 &= -2\sqrt{\frac{\bar{B}_1}{3\bar{C}_1}}\sinh\left(\frac{1}{3}\operatorname{arcsinh}\left(-\frac{3\bar{A}_1}{2\bar{B}_1}\sqrt{\frac{3\bar{C}_1}{\bar{B}_1}}\right)\right)\\
\gamma_2 &= -2\sqrt{\frac{\bar{B}_2}{3\bar{C}_2}}\sinh\left(\frac{1}{3}\operatorname{arcsinh}\left(-\frac{3\bar{A}_2}{2\bar{B}_2}\sqrt{\frac{3\bar{C}_2}{\bar{B}_2}}\right)\right).
\end{aligned}$$

Combining (124), (125) with the above equation, finally, the constant learning rate is required to be

$$\gamma \in \left(0, \min\left\{1, \frac{1}{BLQP}, \gamma_1, \gamma_2\right\}\right).$$

■