Data Dependent Convergence For Consensus Stochastic Optimization

Avleen Bijral, Anand D. Sarwate, *Senior Member, IEEE,* Nathan Srebro

***Abstract*—We study a distributed consensus-based stochastic gradient descent (SGD) algorithm and show that the rate of convergence involves the spectral properties of two matrices: the standard spectral gap of a weight matrix from the network topology and a new term depending on the spectral norm of the sample covariance matrix of the data. This data-dependent convergence rate shows that distributed SGD algorithms perform better on datasets with small spectral norm. Our analysis method also allows us to find data-dependent convergence rates as we limit the amount of communication. Spreading a fixed amount of data across more nodes slows convergence; for asymptotically growing data sets we show that adding more machines can help when minimizing twice-differentiable losses.**

I. Introduction

Decentralized optimization algorithms for statistical computation and machine learning on large data sets try to trade off efficiency (in terms of estimation error) and speed (from parallelization). From an empirical perspective, it is often unclear when these methods will work for a particular data set, and to what degree additional communication can improve performance. For example, in high-dimensional problems communication can be costly. We would therefore like to know when limiting communication is feasible or beneficial. The theoretical analysis of distributed optimization methods has focused on providing strong *data-independent* convergence rates under analytic assumptions on the objective function such as convexity and smoothness. In this paper we show how the tradeoff between efficiency and speed is affected by the data distribution itself. We study a class of distributed optimization algorithms and prove an upper bound on the error that depends on the spectral norm of the data covariance. By tuning the frequency with which nodes communicate, we obtain a bound that depends on data distribution, network size and topology, and amount of communication. This allows us to interpolate between regimes where communication is cheap (e.g. shared memory systems) and those where it is not (clusters and sensor networks).

We study the problem of minimizing a regularized convex function [1] of the form

$$\begin{aligned} J(\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{w}^\top \mathbf{x}_i; y_i) + \frac{\mu}{2}\left\|\mathbf{w}\right\|^2 \qquad (1) \\ &= \mathbb{E}_{\mathbf{x}\sim\hat{\mathcal{P}}}\left[\ell(\mathbf{w}^\top \mathbf{x}; y)\right] + \frac{\mu}{2}\left\|\mathbf{w}\right\|^2, \end{aligned}$$

where $\ell(\cdot)$ is convex and Lipschitz and the expectation is with respect to the empirical distribution $\hat{\mathcal{P}}$ corresponding to a given data set with N total data points $\{(\mathbf{x}_i, y_i)\}$. For the purposes of this work we will assume $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. This regularized empirical risk minimization formulation encompasses algorithms such as support vector machine classification, ridge regression, logistic regression, and others [2]. For example $\mathbf{x}$ could represent d pixels in a grayscale image and y a binary label indicating whether the image is of the digit 7: $\mathbf{w}^\top \mathbf{x}$ gives a confidence value about whether the image is a 7 or not. We would like to solve this problem using a network of m processors connected via a communication network (represented mathematically by a graph indicating which nodes can communicate directly with each other). The system would distribute these N points across the m nodes, inducing local objective functions $J_j(\mathbf{w})$ approximating (1).

In such a computational model, nodes can perform local computations and send messages to each other to jointly minimize (1). The strategy we analyze is what is referred to as distributed primal averaging [3]: each node in the network processes points sequentially, performing a stochastic gradient descent (SGD) update locally and averaging the current iterate values of their neighbors after each gradient step. This can also be thought of as a distributed consensus-based version of Pegasos [4] when the loss function is the hinge loss. We consider a general topology with m nodes attempting to minimize a global objective function $J(\mathbf{w})$ that decomposes into a sum of m local objectives: $J(\mathbf{w}) = \sum_{i=1}^{m} J_i(\mathbf{w})$. This is a model for optimization in systems such as data centers, distributed control systems, and sensor networks.

Main Results. Our goal in this paper is to characterize how how the spectral norm $\rho^2 = \sigma_1(\mathbb{E}_{\hat{\mathcal{P}}}[\mathbf{x}\mathbf{x}^\top])$ of the sample covariance of the data affects the rate of convergence when solving (1) and how limiting communication can also affect the rate. Elucidating this dependence can help guide empirical practice by providing insight into when these methods will work well. We prove an upper bound on the suboptimality gap for distributed primal averaging that depends on ρ^2 as well as

Manuscript received May 17, 2013; revised January 9, 2014; second revision . Date of current version August 30, 2016. The work of A.D. Sarwate is supported in part by the US National Science Foundation under awards CCF-1218331, CCF-1440033, and CCF-1453432.

A. Bijral and N. Srebro are with Toyota Technological Institute at Chicago 6045 S. Kenwood Ave, Chicago, IL 60637 USA, email: abijral@ttic.edu, nati@ttic.edu

Anand D. Sarwate is with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, 84 Brett Road, Piscataway, NJ 08854, USA. Email: anand.sarwate@rutgers.edu

Communicated by xxxxxxxxxx.

Digital Object Identifier xxxxxxxxxxxx

the mixing time of the weight matrix associated to the algorithm. Our result shows that networks of size $m < 1/\rho^2$ gain from parallelization. To understand the communication-limited regime, we extend our analysis to intermittent communication. We show that by using a mini-batching strategy we can offset the penalty of infrequent communication by communicating after a mini-batch (sub)gradient step. Finally, in an asymptotic regime with infinite data at every node we show that for twice-differentiable loss functions this network effect disappears and that we gain from additional parallelization.

Related Work. Several authors have proposed distributed algorithms involving nodes computing local gradient steps and averaging iterates, gradients, or other functions of their neighbors [3], [5], [6]. By alternating local updates and consensus with neighbors, estimates at the nodes converge to the optimizer of $J(\cdot)$. In these works no assumption is made on the local objective functions and they can be arbitrary. Consequently the convergence guarantees do not reflect the setting when the data is homogenous (for e.g. when data has the same distribution), specifically error increases as we add more machines. This is counterintuitive, especially in the large scale regime, since this suggests that despite homogeneity the methods perform worse than the centralized setting (all data on one node).

We provide a first analysis of a consensus based stochastic gradient method in the homogenous setting and demonstrate that there exist regimes where we benefit from having more machines in any network. To mitigate the effect of limited communication, we propose and analyze a mini-batched extension to reduce communication costs. We interpret this as an intermediate regime between full communication and one-shot communication [7], [8]. Finally, we show that for twice-differentiable losses, having more machines always helps (via a variance reduction) in the infinite data regime, using results of Bianchi et al. [9].

In contrast to our stochastic gradient based results, data dependence via the Hessian of the objective has also been demonstrated in parallel coordinate descent based approaches of Liu et al. [10] and the Shotgun algorithm of Bradley et al. [11]. The assumptions differ from us in that the objective function is assumed to be smooth [10] or $\mathcal{L}_1$ regularized [11]. Most importantly, our results hold for arbitrary networks of compute nodes, while the coordinate descent based results hold only for networks where all nodes communicate with a central aggregator (sometimes referred to as a master-slave architecture, or a star network), which can be used to model shared-memory systems.

Another interesting line of work is the impact of delay on convergence in distributed optimization [12]. These results show that delays in the gradient computation for a star network are asymptotically negligible when optimizing smooth loss functions. We study general network topologies but with intermittent, rather than delayed communication. Our result suggest that certain datasets are more tolerant of skipped communication rounds, based on the spectral norm of their covariance.

We take an approach similar to that of Takáč et al. [13], who developed a spectral-norm based analysis of mini-batching for non-smooth functions. We decompose the iterate in terms of the data points encountered in the sample path [14]. This differs from analysis based on smoothness considerations alone [8], [12], [14], [15] and gives practical insight into how communication (full or intermittent) impacts the performance of these algorithms. Note that our work is fundamentally different in that these other works either assume a centralized setting [8], [14], [15] or implicitly assume a specific network topology (e.g. [7] uses a star topology). For the main results we only assume strong convexity while the existing guarantees for the cited methods depend on a variety of regularity and smoothness conditions.

Limitations. In the stochastic convex optimization (see for e.g. [16]) setting the quantity of interest is the population objective corresponding to problem 1. When minimizing this population objective our results suggest that adding more machines worsens convergence (See Theorem 1). For finite data our convergence results satisfy the intuition that adding more nodes in an arbitrary network will hurt convergence. The finite homogenous setting is most relevant in settings such as data centers, where the processors hold data which essentially looks the same. In some sensor network applications such as global calibration, this assumption may also hold, but for spatially varying phenomena the task may be different and our model may no longer be as relevant [17].

In this paper we focus on a simple and well-studied protocol [3] and for simplicity study a homogenous setting where the data is split evenly across nodes. However, our analysis approach and insights may yield data-dependent bounds for other more complex algorithms such as distributed dual averaging [5]. More sophisticated gradient averaging schemes such as that of Mokhtari and Ribeiro [18] can exploit dependence across iterations [19], [20] to improve the convergence rate; analyzing the impact of the data distribution is considerably more complex in these algorithms.

In the infinite or large scale data setting, common in machine learning applications, this is counterintuitive since when each node has infinite data, any distributed scheme including one on arbitrary networks shouldn't perform worse than the centralized scheme (all data on one node). Thus our analysis is limited in that it doesn't unify the stochastic optimization and the consensus setting in a completely satisfactory manner. To partially remedy this we explore distributed primal averaging for smooth strongly convex objectives in the asymptotic regime and show that one can gain from adding more machines in any network.

We believe that our results are first step towards understanding data-dependent bounds for distributed stochastic optimization in settings common to machine learning. Our analysis coincides with phenomenon seen in practice: for data sets with small ρ, distributing the computation across many machines is beneficial, but for data with larger ρ more machines is not necessarily better. We provide upper bounds on the gap between the iterates and the optimal solution: these bounds do not immediately yield parameters for practical use, but our work does suggest that taking into account the data dependence can improve the empirical performance of these methods.

II. Model

We will use boldface for vectors. Let $[k] = \{1, 2, \ldots, k\}$. Unless otherwise specified, the norm $\|\cdot\|$ is the standard Euclidean norm. The spectral norm of a matrix A is defined to be the largest singular value $\sigma_1(A)$ of the matrix A, or equivalently the square root of the largest eigenvalue of $A^\top A$. For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, we will denote the neighbors of a vertex $i \in \mathcal{V}$ by $\mathcal{N}(i) \subseteq \mathcal{V}$.

Data model. Let $\mathcal{P}$ be a distribution on $\mathbb{R}^{d+1}$ such that for $(\mathbf{x}, y) \sim \mathcal{P}$, we have $\|\mathbf{x}\| \leq 1$ almost surely. Let $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be i.i.d sample of d-dimensional vectors from $\mathcal{P}$ and let $\hat{\mathcal{P}}$ be the empirical distribution of S. Let $\hat{\mathbf{\Sigma}} = \mathbb{E}_{\mathbf{x}\sim\hat{\mathcal{P}}}[\mathbf{x}\mathbf{x}^\top]$ be the sample second-moment matrix of S. Our goal is to express the performance of our algorithms in terms of $\rho = \sigma_1(\hat{\mathbf{\Sigma}})$, the spectral norm of $\hat{\mathbf{\Sigma}}$. The spectral norm ρ can vary significantly across different data sets. For example, for sparse data sets ρ is often small. This can also happen if the data happens to lie in low-dimensional subspace (smaller than the ambient dimension d).

Problem. Our problem is to minimize a particular instance of (1) where the expectation is over a finite collection of data points:

$$\mathbf{w}^* \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{w}} J(\mathbf{w}) \tag{2}$$

We will take a model in which m individual nodes or processors operate in a discrete-time fashion to compute $\mathbf{w}^*$. In each time-slot (or iteration) they can perform local computations and pass messages for a fixed number T of iterations. Let $\hat{\mathbf{w}}_j(t)$ be the estimate of $\mathbf{w}^*$ at node $j \in [m]$ in the t-th iteration. We bound the expected gap (over the data distribution) at iteration T between $J(\mathbf{w}^*)$ and the value $J(\hat{\mathbf{w}}_i(T))$ of the global objective $J(\hat{\mathbf{w}}_j(T))$ at the output $\hat{\mathbf{w}}_j(T)$ of each node j in our distributed network. It would also be interesting to study the gap between the iterates and the minimizer of the population objective in (1); we defer this for future work. We will denote the subgradient set of $J(\mathbf{w})$ by $\partial J(\mathbf{w})$ and a subgradient of $J(\mathbf{w})$ by $\nabla J(\mathbf{w}) \in \partial J(\mathbf{w})$.

In our analysis we will make the following assumptions about the individual functions $\ell(\mathbf{w}^\top \mathbf{x})$: (a) The loss functions $\{\ell(\cdot)\}$ are convex, and (b) The loss functions $\{\ell(\cdot; y)\}$ are L-Lipschitz for some $L > 0$ and all y. Note that $J(\mathbf{w})$ is μ-strongly convex due to the ℓ_2-regularization. Our analysis will not depend on the the response y except through the Lipschitz bound L so we will omit the explicit dependence on y to simplify the notation in the future.

Network Model. We consider a model in which minimization in (2) must be carried out by m computational devices, or nodes. These nodes are arranged in a network whose topology is given by a graph $\mathcal{G}$ – an edge (i, j) in the graph means nodes i and j can communicate. A matrix $\mathbf{P}$ is called graph conformant if $P_{ij} > 0$ only if the edge (i, j) is in the graph. We will consider algorithms which use a doubly stochastic and graph conformant sequence of matrices $\mathbf{P}(t)$.

Sampling Model. We assume the N data points are divided evenly and uniformly at random among the m nodes, and define $n \overset{\text{def}}{=} N/m$ to be the number of points at each node. This is a necessary assumption since our bounds are data dependent and depend on subsampling bounds of spectral norm of certain random submatrices. However our data independent bound holds for arbitrary splits. Let S_i be the subset of n points at node i. The local stochastic gradient procedure consists of each node $i \in [m]$ sampling from S_i with replacement. This is an approximation to the local objective function

$$J_i(\mathbf{w}) = \sum_{j \in S_i} \frac{\ell(\mathbf{w}^\top \mathbf{x}_{i,j})}{n} + \frac{\mu}{2} \|\mathbf{w}\|^2 . \tag{3}$$

Algorithm. In the subsequent sections we analyze the distributed version (Algorithm 1) of standard SGD. This algorithm is not new [3], [6] and has been analyzed extensively in the literature. The step-size $\eta_t = 1/(\mu t)$ is commonly used for large scale strongly convex machine learning problems like SVMs (e.g. [4]) and ridge regression: to avoid an extra parameter in the bounds, we take this setting. More discussion on the optimality of the step-size for general strongly convex functions can be found in Rakhlin and Shamir [21]. In Algorithm 1 node i samples a point uniformly with replacement from a local pool of n points and then updates its iterate by computing a weighted sum with its neighbors followed by a local subgradient step. The selection is uniform to guarantee that the subgradient is an unbiased estimate of a true subgradient of the local objective $J_i(\mathbf{w})$, and greatly simplifies the analysis. Different choices of $\mathbf{P}(t)$ will allow us to understand the effect of limiting communication in this distributed optimization algorithm.

Algorithm 1 Consensus-based Strongly Convex Optimization

```
Input: $\{\mathbf{x}_{i,j}\}$,where $i \in [m]$ and $j \in [n]$ and $N = mn$,
matrix sequence $\mathbf{P}(t)$, $\mu > 0$, $T \geq 1$

{Each $i \in [m]$ executes}
Initialize: set $\mathbf{w}_i(1) = \mathbf{0} \in \mathbb{R}^d$.
for $t = 1$ to $T$ do
    Sample $\mathbf{x}_{i,t}$ uniformly with replacement from $S_i$.
    Compute $\mathbf{g}_i(t) \in \partial\ell(\mathbf{w}_i(t)^\top \mathbf{x}_{i,t})\mathbf{x}_{i,t} + \mu\mathbf{w}_i(t)$
    $\mathbf{w}_i(t+1) = \sum_{j=1}^{m} \mathbf{w}_j(t) P_{ij}(t) - \eta_t \mathbf{g}_i(t)$
end for
Output: $\hat{\mathbf{w}}_i(T) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_i(t)$ for any $i \in [m]$.
```

Expectations and probabilities. There are two sources of stochasticity in our model: the first in the split of data points to the individual nodes, and the second in sampling the points during the gradient descent procedure. We assume that the split is done uniformly at random, which implies that the expected covariance matrix at each node is the same as the population covariance matrix $\hat{\mathbf{\Sigma}}$. Conditioned on the split, we assume that the sampling at each node is uniformly at random from the data point at that node, which makes the stochastic subgradient an unbiased estimate of the subgradient of the local objective function. Let $\mathcal{F}_t$ be the sigma algebra generated by the random point selections of the algorithm up to time t, so that the iterates $\{\mathbf{w}_i(t) : i \in [m]\}$ are measurable with respect to $\mathcal{F}_t$.

III. Convergence and Implications

Methods like Algorithm 1, also referred to as primal averaging, have been analyzed previously [3], [6], [22]. In these works it is shown that the convergence properties depend on the structure of the underlying network via the second largest eigenvalue of $\mathbf{P}$. We consider in this section the case when $\mathbf{P}(t) = \mathbf{P}$ for all t where $\mathbf{P}$ is a fixed Markov matrix. This corresponds to a synchronous setting where communication occurs at every iteration.

We analyze the use of the step-size $\eta_t = 1/(\mu t)$ in Algorithm 1 and show that the convergence depends on the spectral norm $\rho^2 = \sigma_1(\hat{\mathbf{\Sigma}})$ of the sample covariance matrix.

Theorem 1: Fix a Markov matrix $\mathbf{P}$ and let $\rho^2 = \sigma_1(\hat{\mathbf{\Sigma}})$ denote the spectral norm of the covariance matrix of the data distribution. Consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex, $\mathbf{P}(t) = \mathbf{P}$ for all t, and $\eta_t = 1/(\mu t)$. Let $\lambda_2(\mathbf{P})$ denote the second largest eigenvalue of $\mathbf{P}$. Then if the number of samples on each machine n satisfies

$$n > \frac{4}{3\rho^2} \log(d) \tag{4}$$

and the number of iterations T satisfies

$$T > 2e \log(1/\sqrt{\lambda_2(\mathbf{P})}) \tag{5}$$

$$\frac{T}{\log(T)} > \max\left(\frac{4}{3\rho^2}\log(d), \frac{\left(\frac{8}{5}\right)^{\frac{1}{4}}\sqrt{m/\rho}}{\log(1/\lambda_2(\mathbf{P}))}\right), \tag{6}$$

then the expected error for each node i satisfies

$$\begin{aligned} &\mathbb{E}\left[J(\hat{\mathbf{w}}_i(T)) - J(\mathbf{w}^*)\right] \leq \\ &\left(\frac{1}{m} + \frac{100\sqrt{m\rho^2}\cdot \log T}{1-\sqrt{\lambda_2(\mathbf{P})}}\right)\cdot\frac{L^2}{\mu}\cdot\frac{\log T}{T}. \end{aligned} \tag{7}$$

Remark 1: Theorem 1 indicates that the number of machines should be chosen as a function of ρ. We can identify three sub-cases of interest:

Case (a): $m \leq \frac{1}{\rho^{2/3}}$: In this regime since $1/m > \sqrt{m\rho^2}$ (ignoring the constants and the $\log T$ term) we always benefit from adding more machines.

Case (b): $\frac{1}{\rho^{2/3}} < m \leq \frac{1}{\rho^2}$: The result tells us that there is no degradation in the error and the bound improves by a factor $\sqrt{m}\rho$. Sparse data sets generally have a smaller value of ρ^2 (as seen in Takáč et al. [13]); Theorem 1 suggests that for such data sets we can use a larger number of machines without losing performance. However the requirements on the number of iterations also increases. This provides additional perspective on the observation by Takáč et al [13] that sparse datasets are more amenable to parallelization via mini-batching. The same holds for our type of parallelization as well.

Case (c): $m > \frac{1}{\rho^2}$: In this case we pay a penalty $\sqrt{m\rho^2} \geq 1$ suggesting that for datasets with large ρ we should expect to lose performance even with relatively fewer machines.

Note that $m > 1$ is implicit in the condition $T > 2e\log(1/\sqrt{\lambda_2}))$ since $\lambda_2 = 0$ for $m = 1$. This excludes the single node Pegasos [13] case. Additionally in the case of general strongly convex losses (not necessarily dependent on $\mathbf{w}^\top\mathbf{x}$) we can obtain a convergence rate of $\mathcal{O}(\log^2(T)/T)$. We do not provide the proof here.

Remark 2: The lower bound on the number of iterations 6 can be considerably improved by instead looking at the intrinsic dimension of the data since for several real datasets the intrinsic dimension can be much smaller than the dimension of the ambient space. However this requires us to assume a lower bound on the norm of the data samples, which is a less natural assumption.

IV. Stochastic Communication

In this section we generalize our analysis in Theorem 1 to handle time-varying and stochastic communication matrices $\mathbf{P}(t)$. In particular, we study the case where the matrices are chosen i.i.d. over time. Any strategy that doesn't involve communicating at every step will incur a larger gap between the local node estimates and their average. We call this the network error. Our goal is to show how knowing ρ^2 can help us balance the network error and optimality gap.

First we bound the network error for the case of stochastic time varying communication matrices $P(t)$ and then a simple extension leads to a generalized version of Theorem 1.

Lemma 2: Let $\{\mathbf{P}(t)\}$ be a i.i.d sequence of doubly stochastic Markov matrices and consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex. We have the following inequality for the expected squared error between the iterate $\mathbf{w}_i(t)$ at node i at time t and the average $\bar{\mathbf{w}}(t)$ defined in Algorithm 1:

$$\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]} \leq \frac{2L}{\mu}\cdot\frac{\sqrt{m}}{b}\cdot\frac{\log(2bet^2)}{t}. \tag{8}$$

where $b = \log\left(1/\lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right)\right)$.

Armed with Lemma 2 we prove the following theorem for Algorithm 1 in the case of stochastic communication.

Theorem 3: Let $\{\mathbf{P}(t)\}$ be an i.i.d sequence of doubly stochastic matrices and $\rho^2 = \sigma_1(\hat{\mathbf{\Sigma}})$ denote the spectral norm of the sample covariance matrix. Consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex, and $\eta_t = 1/(\mu t)$. Then if the number of samples on each machine n satisfies

$$n > \frac{4}{3\rho^2}\log(d) \tag{9}$$

and the number of iterations T satisfies

$$T > 2e\log(1/\sqrt{\lambda_2(\mathbb{E}\left[\mathbf{P}^2(t)\right])}) \tag{10}$$

and

$$\frac{T}{\log(T)} > \max\left(\frac{4}{3\rho^2}\log(d), \cdot\frac{\sqrt{(8/5)m/\rho^2}}{\log(1/\lambda_2(\mathbb{E}\left[\mathbf{P}^2(t)\right]))}\right), \tag{11}$$

then the expected error for the output of each node i satisfies

$$\begin{aligned} &\mathbb{E}\left[J(\hat{\mathbf{w}}_i(T)) - J(\mathbf{w}^*)\right] \\ &\leq \left(\frac{1}{m} + \frac{100\sqrt{m\rho^2}\cdot\log T}{1-\sqrt{\lambda_2(\mathbb{E}\left[\mathbf{P}^2(t)\right])}}\right)\cdot\frac{L^2}{\mu}\cdot\frac{\log T}{T}. \end{aligned} \tag{12}$$

Remark: This result generalizes the conclusions of Theorem 1 to the case of stochastic communication schemes. Thus allowing for the data dependent interpretations of convergence in a more general setting.

V. Limiting Communication

As an application of the stochastic communication scenario of Theorem 3 we now analyze the effect of reducing the communication overhead of Algorithm 1. This reduction can improve the overall running time ("wall time") of the algorithm because communication latency can hinder the convergence of many algorithms in practice [23]. A natural way of limiting communication is to communicate only a fraction ν of the T total iterations; at other times nodes simply perform local gradient steps.

We consider a sequence of i.i.d random matrices $\{\mathbf{P}(t)\}$ for Algorithm 1 where $\mathbf{P}(t) \in \{\mathbf{I}, \mathbf{P}\}$ with probabilities $1-\nu$ and ν, respectively, where $\mathbf{I}$ is the identity matrix (implying no communication since $P_{ij}(t) = 0$ for $i \neq j$) and, as in the previous section, $\mathbf{P}$ is a fixed doubly stochastic matrix respecting the graph constraints. For this model the expected number of times communication takes place is simply νT. Note that now we have an additional randomization due to the Bernoulli distribution over the doubly stochastic matrices. Analyzing a matrix $\mathbf{P}(t)$ that depends on the current value of the iterates is considerably more complicated.

A straightforward application of Theorem 3 reveals that the optimization error is proportional to $\frac{1}{\nu}$ and decays as $\mathcal{O}(\frac{1}{\nu} \cdot \frac{\log^2(T)}{T})$. However, this ignores the effect of the local communication-free iterations.

A mini-batch approach. To account for local communication free iterations we modify the intermittent communication scheme to follow a deterministic schedule of communication every $1/\nu$ steps. However, instead of taking single gradient steps between communication rounds, each node gathers the (sub)gradients and then takes an aggregate gradient step. That is, after the t-th round of communication, the node samples a batch $\mathcal{I}_t$ of indices sampled with replacement from its local data set with $|\mathcal{I}_t| = 1/\nu$. We can think of this as the base algorithm with a better gradient estimate at each step. The update rule is now

$$\mathbf{w}_i(t+1) = \sum_{j \in \mathcal{N}_i} \mathbf{w}_j(t) P_{ij}(t) - \eta_t \nu \sum_{i \in \mathcal{I}_i} \mathbf{g}_i(t). \tag{13}$$

We define $\mathbf{g}_i^{1/\nu}(t) = \sum_{i \in \mathcal{I}_i} \mathbf{g}_i(t)$. Now the iteration count is over the communication steps and $\mathbf{g}_i^{1/\nu}(t)$ is the aggregated mini-batch (sub)gradient of size $1/\nu$. Note that this is analogous to the random scheme above but the analysis is more tractable.

Theorem 4: Fix a Markov matrix $\mathbf{P}$ and let $\rho^2 = \sigma_1(\hat{\mathbf{\Sigma}})$ denote the spectral norm of the covariance matrix of the data distribution. Consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex, $\mathbf{P}(t) = \mathbf{P}$ for all t, and $\eta_t = 1/(\mu t)$ for scheme (13). Let $\lambda_2(\mathbf{P})$ denote the second largest eigenvalue of $\mathbf{P}$. Then if the number of samples on each machine n satisfies

$$n > \frac{4}{3\rho^2} \log(d) \tag{14}$$

and

$$\begin{aligned} T &> \frac{2e}{\nu} \log(1/\sqrt{\lambda_2(\mathbf{P})}) \\ \frac{T}{\log(\nu T)} &> \max\left(\frac{4}{3\nu\rho^2}\log(d), \frac{\left(\frac{8}{5}\right)^{\frac{1}{4}} \sqrt{m/\rho^2}}{\log(1/\lambda_2)}\right) \\ \frac{1}{\nu} &> \frac{4}{3\rho^2} \cdot \log(d) \end{aligned} \tag{15}$$

and then the expected error for each node i satisfies

$$\mathbb{E}\left[J(\hat{\mathbf{w}}_i(T)) - J(\mathbf{w}^*)\right] \leq \left(\frac{1}{m} + 200\sqrt{5}\frac{\sqrt{m\rho^4}\log(\nu T)}{1-\sqrt{\lambda_2}}\right) \cdot \frac{L^2}{\mu} \cdot \frac{\log(\nu T)}{T}. \tag{16}$$

where ν is the frequency of communication and where $\lambda_2 = \lambda_2(\mathbf{P})$.

Remark: Theorem 4 suggests that if the inverse frequency of communication is large enough then we can obtain a sharper bound on the error by a factor of ρ. This is significantly better than a $\mathcal{O}(\sqrt{m\rho^2} \cdot \frac{\log \nu T}{\nu T})$ baseline guarantee from a direct application of Theorem 1 when the number of iterations is νT.

Additionally the result suggests that if we communicate on a mini batch(where batch size $b = 1/\nu$) that is large enough we can improve Theorem 1, specifically now we get a $1/m$ improvement when $m \leq 1/\rho^{4/3}$.

VI. Asymptotic Regime

In this section we explore the sub-optimality of distributed primal averaging when $T \to \infty$ for the case of smooth strongly convex objectives. As discussed before the results of Section III suggest that we never gain from adding more machines in any network. Now we investigate the behaviour of Algorithm 1 in the asymptotic regime and show that the network effect disappears and we do indeed gain from more machines in any network.

Our analysis depends on the asymptotic normality of a variation of Algorithm 1 [9, Theorem 5]. The main differences between Algorithm 1 and the consensus algorithm of Bianchi et al. [9] is that we average the iterates before making the local update.

We make the following assumptions for the analysis in this section: (1) The loss function differentials $\{\partial\left(\ell(\cdot)\right)\}$ are differentiable and G-Lipschitz for some $G > 0$, (2) the stochastic gradients are of the form $\mathbf{g}_i(t) = \nabla J(\mathbf{w}_i(t)) + \boldsymbol{\xi}_t$ where $\mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top] = \mathbf{C}$, and (3) there exists $p > 0$ such that $\mathbb{E}\left[\|\boldsymbol{\xi}_t\|^{2+p}\right] < \infty$. Our results hold for all smooth strongly convex objectives not necessarily dependent on $\mathbf{w}^\top \mathbf{x}$.

Lemma 5: Fix a Markov matrix $\mathbf{P}$. Consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex and twice differentiable, $\mathbf{P}(t) = \mathbf{P}$ for all t, and $\eta_t = 1/(\lambda t)$. then

TABLE I
DATA SETS AND PARAMETERS FOR EXPERIMENTS

data set	training	test	dim.	λ	ρ^2
RCV1	$781,265$	$23,149$	$47,236$	10^{-4}	0.01
Covertype	$522,911$	$58,001$	$47,236$	10^{-6}	0.21

the expected error for each node i satisfies for a arbitrary split of N samples into m nodes

$$\limsup_{T\to\infty} T \cdot \mathbb{E}\left[J\left(\sum_{j=1}^{m} P_{ij}\mathbf{w}_j(T)\right) - J(\mathbf{w}^*)\right] \leq \sum_{j\in\mathcal{N}(i)} (P_{ij})^2 \cdot \mathrm{Tr}\left(\mathbf{H}\right) \cdot \frac{G}{\mu} \tag{17}$$

where $\mathbf{H}$ is the solution to the equation

$$\nabla J^2(\mathbf{w}^*)\mathbf{H} + \mathbf{H}\nabla J^2(\mathbf{w}^*)^T = \mathbf{C}. \tag{18}$$

Remark: This result shows that asymptotically the network effect from Theorem 3 disappears and that additional nodes can speed convergence.

An application of Lemma 5 to problem (1) gives us the following result for the specialized case of a k-regular graph with constant weight matrix $\mathbf{P}$.

Theorem 6: Consider Algorithm 1 when the objective $J(\mathbf{w})$ has the form 1 , $\mathbf{P}(t) = \mathbf{P}$ and corresponds to a k-regular graph with uniform weights for all t, and $\eta_t = 1/(\lambda t)$. then the expected error for each node i satisfies

$$\limsup_{T\to\infty} T \cdot \mathbb{E}\left[J\left(\sum_{j=1}^{m} P_{ij}\mathbf{w}_j(T)\right) - J(\mathbf{w}^*)\right] \leq \frac{25\rho L^2}{k} \cdot \mathrm{Tr}\left(\nabla^2 J(\mathbf{w}^*)^{-1}\right) \cdot \frac{G}{\mu} \tag{19}$$

where the expectation is with respect to the history of the sampled gradients as well as the uniform random splits of N data points across m machines.

Remark: For objective (1) we obtain a $1/k$ variance reduction and the network effect disappears.

VII. EXPERIMENTS

Our goals in our experimental evaluation are to validate the theoretical dependence of the convergence rate on ρ^2 and to see if the conclusions hold when the assumptions we make in the analysis are violated. Note that all our experiments are based on simulations on a multicore computer. Implementation in a truly distributed computation system raises many other considerations such because the "wall time" depends on many factors such as slow nodes and communication latency [23].

A. Data sets, tasks, and parameter settings

The data sets used in our experiments are summarized in Table VII-A. Covertype is the forest covertype dataset [24] used in [4] obtained from the UC Irvine Machine Learning Repository [25], and rcv1 is from the Reuters collection [26] obtained from libsvm collection [27]. The RCV1 data set has a small value of $\hat{\rho}^2$, whereas Covertype has a larger value. In all the experiments we looked at ℓ_2-regularized classification objectives for problem (1). Each plot is averaged over 5 runs.

The data consists of pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. In all experiments we optimize the ℓ_2-regularized empirical hinge loss where $\ell(\mathbf{w}^\top \mathbf{x}) = (1 - \mathbf{w}^\top \mathbf{x} y)^+$. The values of the regularization parameter μ are chosen from to be the same as those in Shalev-Shwarz et al. [4].

We simulated networks of compute nodes of varying size (m) arranged in a k-regular graph with $k = \lfloor 0.25m \rfloor$ or a fixed degree (not dependent on m). Note that the dependence of the convergence rate of procedures like Algorithm 1 on the properties of the underlying network has been investigated before and we refer the reader to Agarwal and Duchi [12] for more details. In this paper we experiment only with k-regular graphs. The weights on the Markov matrix $\mathbf{P}$ are set by using the max-degree Markov chain (see [28]). One can also optimize for the fastest mixing Markov chain ([28], [29]). Each node is randomly assigned $n = \lfloor N/m \rfloor$ points.

B. Intermittent Communication

In this experiment we show the objective function for RCV1 and Covertype as we change the frequency of communication (Figure 1), communicating after every $1, 10, 50$ and 500 iterations. Indeed as predicted we see that the dataset with the larger ρ^2 appears to be affected more by intermittent communication. This indicates that network bandwidth can be conserved for datasets with a smaller ρ^2.

C. Comparison of Different Schemes

We compare the three different schemes proposed in this paper. On a network of $m = 64$ machines we plot the performance of the mini batch extension of Algorithm 1 with batch size 128 against the intermittent scheme that communicates after every 128 iterations and also the standard version of the algorithm. In Figure 3 we see that as predicted in Theorem 4 the mini batch scheme proposed in (13) does better than the vanilla and the intermittent scheme.

D. Infinite Data

To provide some empirical evidence of Lemma 5 we generate a very large ($N = 10^7$) synthetic dataset from a multivariate Normal distribution and created a simple binary classification task using a random hyperplane. As we can see in figure 2 for the SVM problem and a k-regular network we continue to gain as we add more machines and then eventually we stabilize but never lose from more machines. We only show the first few thousand iterations for clarity.

E. Diminishing Communication

To test if our conclusions apply when the i.i.d assumption for the matrices $\mathbf{P}(t)$ does not hold we simulate a diminishing communication regime. Such a scheme can be useful when the nodes are already close to the optimal solution and communicating their respective iterate is wasteful. Intuitively it is in the beginning the nodes should communicate more

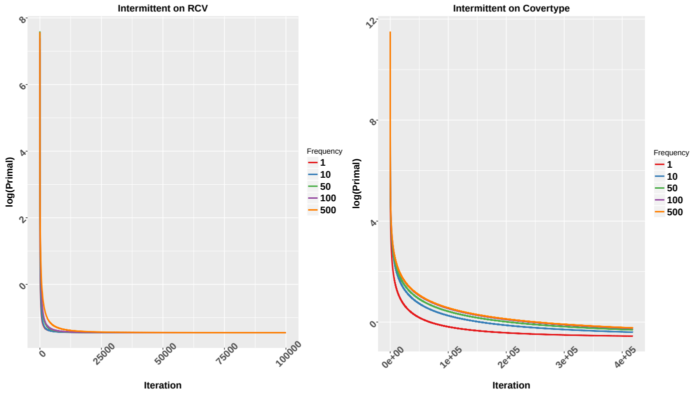

Fig. 1. Performance of Algorithm 1 with intermittent communication scheme on datasets with very different ρ^2. The algorithm works better for smaller ρ^2 and there is less decay in performance for `RCV1` as we decrease the number of communication rounds as opposed to `Covertype` ($\rho^2 = 0.01$ vs $\rho^2 = 0.21$).

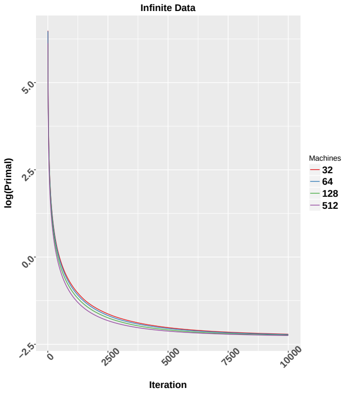

Fig. 2. No network effect with increasing benefit of adding more machines in the case of infinite data.

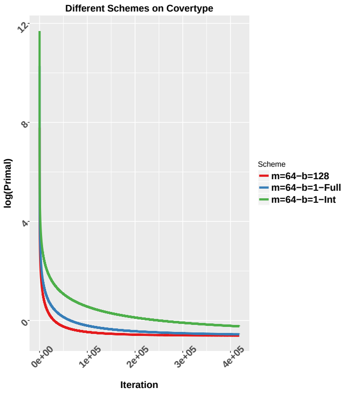

Fig. 3. Comparison of three different schemes a) Algorithm 1 with Mini-Batching b) Standard c) Intermittent with $b = (1/\nu) = 128$. As predicted the mini-batch scheme performs much better than the others.

frequently. To formalize the intuition we propose the following communication model

$$\mathbf{P}(t) = \begin{cases} \mathbf{P} & \text{w.p. } Ct^{-p} \\ \mathbf{I} & \text{w.p. } 1 - Ct^{-p} \end{cases} \tag{20}$$

where $C, p > 0$. Thus the sequence of matrices are not identically distributed and the conclusions of Theorem 3 do not apply.

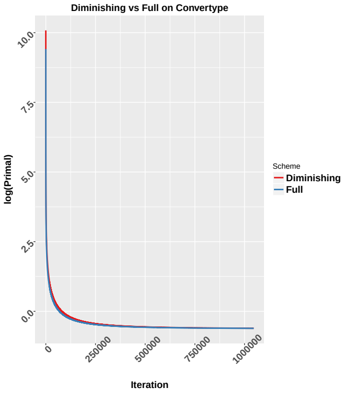

Fig. 4. The performance on Covertype with a full and a diminishing communication scheme is similar.

However in Figure 4 ($C = 1, p = 0.5$) we see that on a network of $m = 128$ nodes the performance for the diminishing regime is similar to the full communication case and we can hypothesize that our results also hold for non i.i.d communication matrices.

VIII. Discussion and Implications

In this paper we described a distributed primal-averaging stochastic gradient descent algorithm and analyzed its performance in terms of the spectral norm ρ^2 of the data covariance matrix under a homogenous assumption. In the consensus problem this setting has not been analyzed before and existing work corresponds to weaker results when this assumption holds.

For certain strongly convex objectives we showed that the objective value gap between any node's iterate and the optimum centralized estimate decreases as $O(\log^2(T)/T)$; crucially, the constant depended on ρ^2 and the spectral gap of the network matrix. This dependence on ρ^2 also appears in the analysis of mini-batching [13]. We showed how limiting communication can improve the total runtime and reduce network costs by extending our analysis with a similar data dependent bound. Moreover we show that in the asymptotic regime the network penalty disappears. Our analysis suggests that distribution-dependent bounds can help us understand how data properties can mediate the tradeoff between computation and communication in distributed optimization. In a sense, data distributions with smaller ρ^2 are easier to optimize over in a distributed setting. This set of distributions includes sparse data sets, an important class for applications.

In the future we will extend data dependent guarantees to serial algorithms as well as the *average-at-end* scheme [7], [8]. Extending our fixed batch-size to random size can help us understand the benefit of communication-free iterations. Finally, we can also study the impact of asynchrony and more general time-varying topologies.

Appendix A
Proof of Theorem 1

Theorem 1 provides a bound on the suboptimality gap for the output $\hat{\mathbf{w}}_i(T)$ of Algorithm 1 at node i, which is the average of that node's iterates. In the analysis we relate this local average to the average iterate across nodes at time t:

$$\bar{\mathbf{w}}(t) = \sum_{i=1}^{m} \frac{\mathbf{w}_i(t)}{m}. \tag{21}$$

We will also consider the average of $\bar{\mathbf{w}}(t)$ over time.

The proof consists of three main steps. First, we establish the following inequality for the objective error:

$$\begin{aligned}
&\mathbb{E}\left[J(\bar{\mathbf{w}}(t)) - J(\mathbf{w}^*)\right] \leq \\
&\quad \frac{(\eta_t^{-1} - \mu)}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}^*\|^2\right] \\
&\quad - \frac{\eta_t^{-1}}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t+1) - \mathbf{w}^*\|^2\right] \\
&\quad + \frac{\eta_t}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{m} \frac{\mathbf{g}_i(t)}{m}\right\|^2\right] \\
&\quad + \sum_{i=1}^{m} \sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]} \\
&\quad \cdot \sqrt{\mathbb{E}\left[(\|\nabla J_i(\mathbf{w}_i(t))\| + \|\nabla J_i(\bar{\mathbf{w}}(t))\|)^2\right]}/m,
\end{aligned} \tag{22}$$

where $\bar{\mathbf{w}}(t)$ is the average of the iterates at all nodes and the expectation is with respect to $\mathcal{F}_t$ while conditioned on the sample split across nodes. All expectations, except when explicitly stated, will be conditioned on this split.

Second, we bound $\mathbb{E}\left[\|\nabla J(\mathbf{w}_i(t))\|^2\right]$ and $\frac{\eta_t}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{m} \frac{\mathbf{g}_i(t)}{m}\right\|^2\right]$ in terms of the spectral norm of the covariance matrix of the distribution $\mathcal{P}$ by additionally taking expectation with respect to the sample S. Finally, we bound the network error $\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]$ in term of the network size m and a spectral property of the matrix $\mathbf{P}$. Combining the bounds using inequality (22) and applying the definition of subgradients yields the result of Theorem 1.

A. Spectral Norm of Random Submatrices

In this section we establish a Lemma pertaining to the spectral norm of submatrices that is central to our results. Specifically we prove the following inequality, which follows by applying the Matrix Bernstein inequality of Tropp [30].

Lemma 7: Let $\mathcal{P}$ be a distribution on $\mathbb{R}^d$ with second moment matrix $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}}[\mathbf{Z}\mathbf{Z}^\top]$ such that $\|\mathbf{Z}_k\| \leq 1$ almost surely. Let $\zeta^2 = \sigma_1(\mathbf{\Sigma})$. Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_K$ be an i.i.d. sample from $\mathcal{P}$ and let $\mathbf{Q}_K = \sum_{k=1}^{K} \mathbf{Z}_k\mathbf{Z}_k^\top$ be the empirical second moment matrix of the data. Then for $K > \frac{4}{3\zeta^2}\log(d)$,

$$\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_K)}{K}\right] \leq 5\zeta^2. \tag{23}$$

Thus when $\mathcal{P}$ is the empirical distribution we get that $\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_K)}{K}\right] \leq 5\zeta^2$.
Remark: We can replace the ambient dimension d in the requirement for K by an intrinsic dimensionality term but this requires a lower bound on the norm of any data point in the sample.

Proof: Let $\mathbf{Z}$ be the $d \times K$ matrix whose columns are $\{\mathbf{Z}_k\}$. Define $\mathbf{X}_k = \mathbf{Z}_k\mathbf{Z}_k^\top - \mathbf{\Sigma}$. Then $\mathbb{E}[\mathbf{X}_k] = \mathbf{0}$ and

$$\lambda_{\max}(\mathbf{X}_k) = \lambda_{\max}\left(\mathbf{Z}_k\mathbf{Z}_k^\top - \mathbf{\Sigma}\right) \leq \|\mathbf{Z}_k\|^2 \leq 1,$$

because $\mathbf{\Sigma}$ is positive semidefinite and $\|x_i\| \leq 1$ for all i. Furthermore,

$$\begin{aligned}
\sigma_1\left(\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{X}_k^2\right]\right) &= K\sigma_1\left(\mathbb{E}\left[\mathbf{Z}_k\mathbf{Z}_k^\top\mathbf{Z}_k\mathbf{Z}_k^\top\right] - \mathbf{\Sigma}^2\right) \\
&\leq K\sigma_1\left(\mathbb{E}\left[\|Y_k\|^2\mathbf{Z}_k\mathbf{Z}_k^\top\right]\right) + K\sigma_1\left(\mathbf{\Sigma}\right)^2 \\
&\leq K(\zeta^2 + \zeta^4) \\
&\leq 2K\zeta^2,
\end{aligned}$$

since $\rho \leq 1$.

Applying the Matrix Bernstein inequality of Tropp [30, Theorem 6.1]:

$$\mathbb{P}\left(\sigma_1\left(\sum_{k=1}^{K}\mathbf{X}_k\right) \geq r\right) \leq \begin{cases} d\exp\left(-3\frac{r^2}{16K\zeta^2}\right) & \frac{r}{K} \leq 2\zeta^2 \\ d\exp\left(-3\frac{r}{8}\right) & \frac{r}{K} \geq 2\zeta^2 \end{cases}. \tag{24}$$

Now, note that

$$\sigma_1\left(\sum_{k=1}^{K}\mathbf{X}_k\right) = \sigma_1\left(\sum_{k=1}^{K}\mathbf{Z}_k\mathbf{Z}_k^\top - K\mathbf{\Sigma}\right),$$

so $\sigma_1\left(\sum_{k=1}^{K}\mathbf{X}_k\right) \geq r$ is implied by

$$\left|\frac{1}{K}\sigma_1\left(\sum_{k=1}^{K}\mathbf{Z}_k\mathbf{Z}_k^\top\right) - \sigma_1\left(\mathbf{\Sigma}\right)\right| \geq \frac{r}{K}.$$

Therefore

$$\mathbb{P}\left(\left|\frac{\sigma_1(\mathbf{Q}_K)}{K} - \zeta^2\right| \geq r'\right) \leq \begin{cases} d\exp\left(-\frac{3Kr'^2}{16\zeta^2}\right) & r' \leq 2\zeta^2 \\ d\exp\left(-\frac{3Kr'}{8}\right) & r' \geq 2\zeta^2 \end{cases}. \tag{25}$$

Integrating (25) yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_K)}{K}\right] &= \int_0^\infty \mathbb{P}\left(\frac{\sigma_1(\mathbf{Q}_K)}{K} \geq x\right) dx \\
&\leq 3\zeta^2 + \int_{3\zeta^2}^\infty \mathbb{P}\left(\frac{\sigma_1(\mathbf{Q}_t)}{K} - \zeta^2 \geq x - \zeta^2\right) dx \\
&\leq 3\zeta^2 + \int_{2\zeta^2}^\infty \mathbb{P}\left(\frac{\sigma_1(\mathbf{Q}_t)}{K} - \zeta^2 \geq r'\right) dr' \\
&\leq 3\zeta^2 + \int_{2\zeta^2}^\infty d \exp\left(-\frac{3}{8} K r'\right) dr' \\
&= 3\zeta^2 + \frac{8}{3} \cdot \frac{d}{K} \exp\left(-\frac{3}{4}\zeta^2 K\right)
\end{aligned}$$

For $K > \frac{4}{3\zeta^2}\log d$,

$$\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_K)}{K}\right] \leq 3\zeta^2 + \frac{8}{3} \cdot \frac{3}{4} \cdot \frac{\zeta^2}{\log d} \leq 5\zeta^2.$$

■

B. Decomposing the expected suboptimality gap

The proof in part follows [3]. It is easy to verify that because $\mathbf{P}$ is doubly stochastic the average of the iterates across the nodes at time t, the average of the iterates across the nodes in (21) satisfies the following update rule:

$$\bar{\mathbf{w}}(t+1) = \bar{\mathbf{w}}(t) - \eta_t \sum_{i=1}^m \frac{\mathbf{g}_i(t)}{m}. \tag{26}$$

We emphasize that in Algorithm 1 we do not perform a final averaging across nodes at the end as in (21). Rather, we analyze the average at a single node across its iterates (sometimes called Polyak averaging). Analyzing (21) provides us with a way to understand how the objective $J(\mathbf{w}_i(t))$ evaluated at any node i's iterate $\mathbf{w}_i(t)$ compares to the minimum value $J(\mathbf{w}^*)$. The details can be found in Section A-G.

To simplify notation, we treat all expectations as conditioned on the sample S. Then (26),

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\mathbf{w}}(t+1) - \mathbf{w}^*\right\|^2 \middle| \mathcal{F}_t\right] \\
&\quad = \mathbb{E}\left[\left\|\bar{\mathbf{w}}(t) - \mathbf{w}^*\right\|^2 | \mathcal{F}_t\right] \\
&\qquad + \eta_t^2 \mathbb{E}\left[\left\|\sum_{i=1}^m \frac{\mathbf{g}_i(t)}{m}\right\|^2 \middle| \mathcal{F}_t\right] \\
&\qquad - 2\eta_t (\bar{\mathbf{w}}(t) - \mathbf{w}^*)^\top \sum_{i=1}^m \frac{\mathbb{E}\left[\mathbf{g}_i(t)|\mathcal{F}_t\right]}{m} \\
&\quad = \mathbb{E}\left[\left\|\bar{\mathbf{w}}(t) - \mathbf{w}^*\right\|^2 | \mathcal{F}_t\right] \\
&\qquad + \eta_t^2 \mathbb{E}\left[\left\|\sum_{i=1}^m \frac{\mathbf{g}_i(t)}{m}\right\|^2 \middle| \mathcal{F}_t\right] \\
&\qquad - 2\eta_t \sum_{i=1}^m (\bar{\mathbf{w}}(t) - \mathbf{w}^*)^\top \frac{\mathbb{E}\left[\mathbf{g}_i(t)|\mathcal{F}_t\right]}{m}.
\end{aligned} \tag{27}$$

Note that $\nabla J_i(\mathbf{w}_i(t)) = \mathbb{E}\left[\mathbf{g}_i(t)|\mathcal{F}_t\right]$, so for the last term, for each i we have

$$\begin{aligned}
&\nabla J_i(\mathbf{w}_i(t))^\top (\bar{\mathbf{w}}(t) - \mathbf{w}^*) \\
&\quad = \nabla J_i(\mathbf{w}_i(t))^\top \left(\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right) \\
&\qquad + \nabla J_i(\mathbf{w}_i(t))^\top \left(\mathbf{w}_i(t) - \mathbf{w}^*\right) \\
&\quad \geq - \left\|\nabla J_i(\mathbf{w}_i(t))\right\| \left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \\
&\qquad + \nabla J_i(\mathbf{w}_i(t))^\top \left(\mathbf{w}_i(t) - \mathbf{w}^*\right) \\
&\quad \geq - \left\|\nabla J_i(\mathbf{w}_i(t))\right\| \left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \\
&\qquad + J_i(\mathbf{w}_i(t)) - J_i(\mathbf{w}^*) + \frac{\mu}{2}\left\|\mathbf{w}_i(t) - \mathbf{w}^*\right\|^2 \\
&\quad = - \left\|\nabla J_i(\mathbf{w}_i(t))\right\| \left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \\
&\qquad + J_i(\mathbf{w}_i(t)) - J_i(\bar{\mathbf{w}}(t)) \\
&\qquad + \frac{\mu}{2}\left\|\mathbf{w}_i(t) - \mathbf{w}^*\right\|^2 + J_i(\bar{\mathbf{w}}(t)) - J_i(\mathbf{w}^*) \\
&\quad \geq - \left\|\nabla J_i(\mathbf{w}_i(t))\right\| \left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \\
&\qquad + \nabla J_i(\bar{\mathbf{w}}(t))^\top \left(\mathbf{w}_i(t) - \bar{\mathbf{w}}(t)\right) \\
&\qquad + \frac{\mu}{2}\left\|\mathbf{w}_i(t) - \mathbf{w}^*\right\|^2 + J_i(\bar{\mathbf{w}}(t)) - J_i(\mathbf{w}^*) \\
&\quad \geq - \left(\left\|\nabla J_i(\mathbf{w}_i(t))\right\| + \left\|\nabla J_i(\bar{\mathbf{w}}(t))\right\|\right) \left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \\
&\qquad + \frac{\mu}{2}\left\|\mathbf{w}_i(t) - \mathbf{w}^*\right\|^2 + J_i(\bar{\mathbf{w}}(t)) - J_i(\mathbf{w}^*),
\end{aligned} \tag{28}$$

where the second and third lines comes from applying the Cauchy-Shwartz inequality and strong convexity, the fifth line comes from the definition of subgradient, and the last line is another application of the Cauchy-Shwartz inequality.

Averaging over all the nodes, using convexity of $\|\cdot\|^2$, the definition of $J(\cdot)$, and Jensen's inequality yields the following inequality:

$$\begin{aligned}
&-2\eta_t \sum_{i=1}^m (\bar{\mathbf{w}}(t) - \mathbf{w}^*)^\top \frac{\mathbb{E}[\mathbf{g}_i(t)|\mathcal{F}_t]}{m} \\
&\quad \leq 2\eta_t \sum_{i=1}^m \frac{\left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \left(\left\|\nabla J_i(\mathbf{w}_i(t))\right\| + \left\|\nabla J_i(\bar{\mathbf{w}}(t))\right\|\right)}{m} \\
&\qquad - 2\eta_t \left(\sum_{i=1}^m \frac{J_i(\bar{\mathbf{w}}(t)) - J_i(\mathbf{w}^*)}{m}\right) \\
&\qquad - \mu\eta_t \sum_{i=1}^m \frac{\left\|\mathbf{w}_i(t) - \mathbf{w}^*\right\|^2}{m} \\
&\quad \leq 2\eta_t \sum_{i=1}^m \frac{\left\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\right\| \left(\left\|\nabla J_i(\mathbf{w}_i(t))\right\| + \left\|\nabla J_i(\bar{\mathbf{w}}(t))\right\|\right)}{m} \\
&\qquad - 2\eta_t \left(J(\bar{\mathbf{w}}(t)) - J(\mathbf{w}^*)\right) - \mu\eta_t \left\|\bar{\mathbf{w}}(t) - \mathbf{w}^*\right\|^2
\end{aligned} \tag{29}$$

Substituting inequality (29) in recursion (27),

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{w}}(t+1)-\mathbf{w}^*\|^2\,|\mathcal{F}_t\right] \\
&\le \mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\,|\mathcal{F}_t\right] \\
&\quad + \eta_t^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\frac{\mathbf{g}_i(t)}{m}\right\|^2 \,\Big|\, \mathcal{F}_t\right] \\
&\quad + 2\eta_t\sum_{i=1}^{m}\frac{\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\|\left(\|\nabla J_i(\mathbf{w}_i(t))\|+\|\nabla J_i(\bar{\mathbf{w}}(t))\|\right)}{m} \\
&\quad - 2\eta_t\left(J(\bar{\mathbf{w}}(t))-J(\mathbf{w}^*)\right) - \mu\eta_t\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2 .
\end{aligned} \tag{30}$$

Taking expectations with respect to the entire history $\mathcal{F}_t$,

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{w}}(t+1)-\mathbf{w}^*\|^2\right] \\
&\le \mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\frac{\mathbf{g}_i(t)}{m}\right\|^2\right] \\
&\quad + 2\eta_t\cdot \\
&\quad \sum_{i=1}^{m}\frac{\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\|\left(\|\nabla J_i(\mathbf{w}_i(t))\|+\|\nabla J_i(\bar{\mathbf{w}}(t))\|\right)\right]}{m} \\
&\quad - 2\eta_t\left(\mathbb{E}\left[J(\bar{\mathbf{w}}(t))-J(\mathbf{w}^*)\right]\right) - \mu\eta_t\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\right] \\
&\le -2\eta_t\left(\mathbb{E}\left[J(\bar{\mathbf{w}}(t))-J(\mathbf{w}^*)\right]\right) \\
&\quad + (1-\mu\eta_t)\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\left\|\sum_{i=1}^{m}\frac{\mathbf{g}_i(t)}{m}\right\|^2\right] \\
&\quad + \frac{2\eta_t}{m}\sum_{i=1}^{m}\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\|^2\right]} \\
&\qquad\qquad \cdot\sqrt{\mathbb{E}\left[\left(\|\nabla J_i(\mathbf{w}_i(t))\|+\|\nabla J_i(\bar{\mathbf{w}}(t))\|\right)^2\right]}
\end{aligned} \tag{31}$$

This lets us bound the expected suboptimality gap $\mathbb{E}\left[J(\bar{\mathbf{w}}(t))-J(\mathbf{w}^*)\right]$ via three terms:

$$\begin{aligned}
\text{T1} &= \frac{(\eta_t^{-1}-\mu)}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\right] \\
&\qquad - \frac{\eta_t^{-1}}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t+1)-\mathbf{w}^*\|^2\right]
\end{aligned} \tag{32}$$

$$\text{T2} = \frac{\eta_t}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{m}\frac{\mathbf{g}_i(t)}{m}\right\|^2\right] \tag{33}$$

$$\begin{aligned}
\text{T3} &= \frac{1}{m}\sum_{i=1}^{m}\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\|^2\right]} \\
&\qquad \cdot\sqrt{\mathbb{E}\left[\left(\|\nabla J_i(\mathbf{w}_i(t))\|+\|\nabla J_i(\bar{\mathbf{w}}(t))\|\right)^2\right]},
\end{aligned} \tag{34}$$

where

$$\mathbb{E}\left[J(\bar{\mathbf{w}}(t))-J(\mathbf{w}^*)\right] \le \text{T1}+\text{T2}+\text{T3}. \tag{35}$$

The remainder of the proof is to bound these three terms separately.

C. Network Error Bound

We need to prove an intermediate bound first to handle term T3.

Lemma 8: Fix a Markov matrix $\mathbf{P}$ and consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex we have the following inequality for the expected squared error between the iterate $\mathbf{w}_i(t)$ at node i at time t and the average $\bar{\mathbf{w}}(t)$ defined in Algorithm 1:

$$\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\|^2\right]} \le \frac{2L}{\mu}\cdot\frac{\sqrt{m}}{b}\cdot\frac{\log(2bet^2)}{t}, \tag{36}$$

where $b=(1/2)\log(1/\lambda_2(\mathbf{P}))$.

Proof: We follow a similar analysis as others [3, Prop. 3] [5, IV.A] [22]. Let $M=\frac{1}{m}\mathbf{1}\mathbf{1}^\top$. Let $\mathbf{W}(t)$ be the $m\times d$ matrix whose i-th row is $\mathbf{w}_i(t)$ and $\mathbf{G}(t)$ be the $m\times d$ matrix whose i-th row is $\mathbf{g}_i(t)$. Then the iteration can be compactly written as

$$\mathbf{W}(t+1)=\mathbf{P}(t)\mathbf{W}(t)-\eta_t\mathbf{G}(t)$$

and the network average matrix $\bar{\mathbf{W}}(t)=M\mathbf{W}(t)$. Then we can write the difference using the fact that $\mathbf{P}(t)=\mathbf{P}$ for all t:

$$\begin{aligned}
&\bar{\mathbf{W}}(t+1)-\mathbf{W}(t+1)= \\
&\quad = (M-I)\left(\mathbf{P}\mathbf{W}(t)-\eta_t\mathbf{G}(t)\right) \\
&\quad = (M-\mathbf{P})\mathbf{W}(t)-\eta_t(M-I)\mathbf{G}(t) \\
&\quad = (M-\mathbf{P})\left(\mathbf{P}\mathbf{W}(t-1)-\eta_{t-1}\mathbf{G}(t-1)\right) \\
&\qquad - \eta_t(M-I)\mathbf{G}(t) \\
&\quad = \left(M-\mathbf{P}^2\right)\mathbf{W}(t-1)-\sum_{s=t-1}^{t}\eta_s\left(M-\mathbf{P}^{t-s}\right)\mathbf{G}(s).
\end{aligned}$$

Continuing the expansion and using the fact that $\mathbf{W}(1)=\mathbf{0}$,

$$\begin{aligned}
&\bar{\mathbf{W}}(t+1)-\mathbf{W}(t+1) \\
&\quad = \left(M-\mathbf{P}^t\right)\mathbf{W}(1)-\sum_{s=1}^{t}\eta_s\left(M-\mathbf{P}^{t-s}\right)\mathbf{G}(s) \\
&\quad = -\sum_{s=1}^{t}\eta_s\left(M-\mathbf{P}^{t-s}\right)\mathbf{G}(s) \\
&\quad = -\sum_{s=1}^{t-1}\eta_s\left(M-\mathbf{P}^{t-s}\right)\mathbf{G}(s)-\eta_t(M-I)\mathbf{G}(t).
\end{aligned} \tag{37}$$

Now looking at the norm of the i-th row of (37) and using the bound on the gradient norm:

$$\begin{aligned}
&\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\| \\
&\quad \le \left\|\sum_{s=1}^{t-1}\eta_s\sum_{j=1}^{m}\left(\frac{1}{m}-(\mathbf{P}^{t-s})_{ij}\right)\mathbf{g}_j(s)\right. \\
&\qquad\qquad \left. +\,\eta_t\left(\sum_{j=1}^{m}\frac{1}{m}\mathbf{g}_j(t)-\mathbf{g}_i(t)\right)\right\|
\end{aligned} \tag{38}$$

$$\le \sum_{s=1}^{t-1}\frac{L}{\mu s}\cdot\left\|\frac{1}{m}-(\mathbf{P}^{t-s})_i\right\|_1+\frac{2L}{\mu t}. \tag{39}$$

We handle the term $\left\|\frac{1}{m} - (\mathbf{P}^{t-s})_i\right\|_1$ using a bound on the mixing rate of Markov chains (c.f. (74) in Tsianos and Rabbat [22]):

$$\sum_{s=1}^{t-1} \frac{L}{\mu s} \cdot \left\|\frac{1}{m} - (\mathbf{P}^{t-s})_i\right\|_1 \le \frac{L\sqrt{m}}{\mu} \sum_{s=1}^{t-1} \left(\sqrt{\lambda_2(\mathbf{P})}\right)^{t-s} \frac{1}{s}. \tag{40}$$

Define $a = \sqrt{\lambda_2(\mathbf{P})} \le 1$ and $b = -\log(a) > 0$. Then we have the following identities:

$$\sum_{\tau=1}^{t} \frac{a^{t-\tau+1}}{\tau} = \sum_{\tau=1}^{t} \frac{a^{\tau}}{t-\tau+1} = \sum_{\tau=1}^{t} \frac{\exp(-b\tau)}{t-\tau+1}. \tag{41}$$

Now using the fact that when $x > -1$ we have $\exp(-x) < 1/(1+x)$ and using the integral upper bound we get

$$\begin{aligned}
&\sum_{\tau=1}^{t} \frac{a^{t-\tau+1}}{\tau} \le \sum_{\tau=1}^{t} \frac{1}{(1+b\tau)(t-\tau+1)} \\
&\le \frac{1}{(1+b)t} + \int_1^t \frac{d\tau}{(1+b\tau)(t-\tau+1)} \\
&= \frac{1}{(1+b)t} + \left[\frac{\log(b\tau+1) - \log(t-\tau+1)}{bt+b+1}\right]_{\tau=1}^{t} \\
&= \frac{1}{(1+b)t} + \frac{\log(bt+1) - \log(b+1) + \log(t)}{bt+b+1} \\
&\le \frac{\log(et(bt+1))}{bt} \\
&\le \frac{\log(2bet^2)}{bt}.
\end{aligned} \tag{42}$$

Using (40) and (42) in (39) we get

$$\begin{aligned}
\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\| &\le \frac{L\sqrt{m}}{\mu} \frac{\log(2bet^2)}{bt} + \frac{2L}{\mu t} \\
&\le \frac{2L\sqrt{m}}{\mu} \frac{\log(2bet^2)}{bt}.
\end{aligned} \tag{43}$$

Therefore we have

$$\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]} \le \frac{2L\sqrt{m}}{\mu} \frac{\log(2bet^2)}{bt}. \tag{44}$$

D. Bounds for expected gradient norms

1) Bounding $\mathbb{E}\left[\|\nabla J_i(\bar{\mathbf{w}}(t))\|^2\right]$*:* Let $\beta_{j,t} \in \partial\ell(\bar{\mathbf{w}}(t)^\top \mathbf{x}_{i,j})$ denote a subgradient for the j-th point at node i and $\boldsymbol{\beta}_t = (\beta_{1,t}, \beta_{2,t}, \ldots, \beta_{n,t})^\top$ be the vector of subgradients at time t. Let $\mathbf{Q}_{S_i}$ be the $n \times n$ Gram matrix of the data set S_i. From the definition of $\|\nabla J_i(\bar{\mathbf{w}}(t))\|$ and using the Lipschitz property of the loss functions, we have the following bound:

$$\begin{aligned}
&\|\nabla J_i(\bar{\mathbf{w}}(t))\|^2 \le \left\| \sum_{j \in S_i} \frac{\beta_{j,t}\mathbf{x}_{i,j}}{n} + \mu \bar{\mathbf{w}}(t) \right\|^2 \\
&\le 2 \left\| \sum_{j \in S_i} \frac{\beta_{j,t}\mathbf{x}_{i,j}}{n} \right\|^2 + 2\mu^2 \|\bar{\mathbf{w}}(t)\|^2 \\
&= \frac{2\sum_{j\in S_i}\sum_{j'\in S_i} \beta_{j,t}\beta_{j',t}\mathbf{x}_{i,j}^\top \mathbf{x}'_{i,j}}{n^2} + 2\mu^2 \|\bar{\mathbf{w}}(t)\|^2 \\
&= \frac{2}{n^2} \boldsymbol{\beta}_t^\top \mathbf{Q}_{S_i} \boldsymbol{\beta}_t + 2\mu^2 \|\bar{\mathbf{w}}(t)\|^2 \\
&\le \frac{2}{n^2} \|\boldsymbol{\beta}_t\|^2 \sigma_1(\mathbf{Q}_{S_i}) + 2\mu^2 \|\bar{\mathbf{w}}(t)\|^2 \\
&\le 2L^2 \frac{\sigma_1(\mathbf{Q}_{S_i})}{n} + 2\mu^2 \|\bar{\mathbf{w}}(t)\|^2.
\end{aligned} \tag{45}$$

We rewrite the update (26) in terms of $\{\mathbf{x}_{i,t}\}$, the points sampled at the nodes at time t:

$$\bar{\mathbf{w}}(t+1) = \bar{\mathbf{w}}(t)(1 - \mu\eta_t) - \eta_t \sum_{i=1}^{m} \frac{\partial\ell(\mathbf{w}_i(t)^\top \mathbf{x}_{i,t})\mathbf{x}_{i,t}}{m}. \tag{46}$$

Now from equation (46), after unrolling the recursion as in Shalev-Shwarz et al. [4] we see

$$\bar{\mathbf{w}}(t) = \frac{1}{\mu(t-1)} \sum_{\tau=1}^{t-1} \frac{\sum_{i=1}^{m} \partial\ell(\mathbf{w}_i(\tau)^\top \mathbf{x}_{i,\tau})\mathbf{x}_{i,\tau}}{m}. \tag{47}$$

Let $\gamma_\tau^i \in \partial\ell(\mathbf{w}_i(\tau)^\top \mathbf{x}_{i,\tau})$ the subgradient set for the ith node computed at time τ, then we have

$$\|\bar{\mathbf{w}}(t)\| \le \frac{1}{\mu(t-1)} \cdot \frac{1}{m} \sum_{i=1}^{m} \left\| \sum_{\tau=1}^{t-1} \gamma_\tau^i \mathbf{x}_{i,\tau} \right\|. \tag{48}$$

Let us in turn bound for each node i the term $\left\|\sum_{\tau=1}^{t-1} \gamma_\tau^i \mathbf{x}_{i,\tau}\right\|$. Let $\gamma_\tau^i \in \partial\ell(\mathbf{w}_i(\tau)^\top \mathbf{x}_{i,\tau})$ denote a subgradient for the point sampled at time τ at node i and $\boldsymbol{\gamma}^i = (\gamma_1^i, \gamma_2^i, \ldots, \gamma_{t-1}^i)^\top$ be the vector of subgradients up to time $t-1$. We have

$$\begin{aligned}
\left\| \sum_{\tau=1}^{t-1} \gamma_\tau^i \mathbf{x}_{i,\tau} \right\|^2 &= \sum_{\tau,\tau'} \gamma_\tau^i \gamma_{\tau'}^i \mathbf{x}_{i,\tau}^\top \mathbf{x}_{i,\tau'} \\
&= (\boldsymbol{\gamma}^i)^\top \mathbf{Q}_{i,t-1} \boldsymbol{\gamma}^i \\
&\le \left\|\boldsymbol{\gamma}^i\right\|^2 \sigma_1(\mathbf{Q}_{i,t-1}) \\
&\le (t-1)L^2 \sigma_1(\mathbf{Q}_{i,t-1}),
\end{aligned} \tag{49}$$

where $\mathbf{Q}_{i,t-1}$ is the $(t-1) \times (t-1)$ Gram submatrix corresponding to the points sampled at the i-th node until time $t-1$.

Further bounding (48):

$$\begin{aligned}
\|\bar{\mathbf{w}}(t)\|^2 &\le \left(\frac{1}{\mu(t-1)} \frac{\sum_{i=1}^{m} \sqrt{(t-1)L^2 \sigma_1(\mathbf{Q}_{i,t-1})}}{m} \right)^2 \\
&\le \frac{L^2}{\mu^2} \left(\frac{1}{m} \sum_{i=1}^{m} \sqrt{\frac{\sigma_1(\mathbf{Q}_{i,t-1})}{t-1}} \right)^2.
\end{aligned}$$

Since as stated before everything is conditioned on the sample split we take expectations w.r.t the history and the random split and using the Cauchy-Schwarz inequality again, and the fact that the points are sampled i.i.d. from the same distribution,

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{w}}(t)\|^2\right] \\
&\leq \frac{L^2}{\mu^2}\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}\left[\frac{\sqrt{\sigma_1(\mathbf{Q}_{i,t-1})\sigma_1(\mathbf{Q}_{j,t-1})}}{t-1}\right] \\
&\leq \frac{L^2}{\mu^2}\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sqrt{\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_{i,t-1})}{t-1}\right]\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_{j,t-1})}{t-1}\right]} \\
&= \frac{L^2}{\mu^2}\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_{i,t-1})}{t-1}\right]. \qquad (50)
\end{aligned}$$

The last line follows from the expectation over the sampling model: the data at node i and node j have the same expected covariance since they are sampled uniformly at random from the total data.

Taking the expectation in (45) and substituting (50) we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla J_i(\bar{\mathbf{w}}(t))\|^2\right] \leq 2L^2\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_{S_i})}{n}\right] \\
+ 2L^2\mathbb{E}\left[\frac{\sigma_1(\mathbf{Q}_{i,t-1})}{t-1}\right]. \qquad (51)
\end{aligned}$$

Since S_i is a uniform random draw from S and by assuming both t and n to be greater than $4/(3\rho^2)\log(d)$, applying Lemma 7 gives us

$$\mathbb{E}\left[\|\nabla J_i(\bar{\mathbf{w}}(t))\|^2\right] \leq 20L^2\rho^2. \qquad (52)$$

2) Bounding $\mathbb{E}\left[\|\nabla J_i(\mathbf{w}_i(t))\|^2\right]$*:* We have just as in the previous subsection

$$\|\nabla J_i(\mathbf{w}_i(t))\|^2 \leq 2L^2\frac{\sigma_1(\mathbf{Q}_{S_i})}{n} + 2\mu^2\|\mathbf{w}_i(t)\|^2.$$

Using the triangle inequality, the fact that $(a_1+a_2)^2 \leq 2a_1^2 + 2a_2^2$, the bounds (43) and (50), and Lemma 7:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_i(t)\|^2\right] &\leq 2\mathbb{E}\left[\|\mathbf{w}_i(t)-\bar{\mathbf{w}}(t)\|^2\right] + 2\mathbb{E}\left[\|\bar{\mathbf{w}}(t)\|^2\right] \\
&\leq \frac{8L^2m}{\mu^2}\frac{\log^2(2bet^2)}{b^2(t-1)^2} + \frac{5L^2\rho^2}{\mu^2}. \qquad (53)
\end{aligned}$$

Since the second term does not scale with t, from (53) we can infer that for the second term to dominate the first we require

$$\frac{t}{\log(t)} > \sqrt{\frac{8}{5}}\frac{\sqrt{m}}{\rho b}.$$

This gives us

$$\mathbb{E}\left[\|\mathbf{w}_i(t)\|^2\right] \leq \frac{10L^2\rho^2}{\mu^2}, \qquad (54)$$

and therefore

$$\mathbb{E}\left[\|\nabla J_i(\mathbf{w}_i(t))\|^2\right] \leq 30L^2\rho^2. \qquad (55)$$

E. Bound for $T2$

Because the gradients are bounded,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=1}^{m}\frac{\mathbf{g}_i(t)}{m}\right\|^2\right] = \mathbb{E}\left[\sum_{i,j}\frac{\mathbf{g}_i(t)^\top\mathbf{g}_i(t)}{m^2}\right] \\
&= \sum_{i=1}^{m}\frac{\mathbb{E}\left[\|\mathbf{g}_i(t)\|^2\right]}{m^2} + \sum_{i\neq j}\frac{\mathbb{E}\left[\mathbf{g}_i(t)^\top\mathbf{g}_j(t)\right]}{m^2} \\
&\leq \frac{L^2}{m} + \sum_{i\neq j}\frac{\mathbb{E}\left[\mathbf{g}_i(t)^\top\mathbf{g}_j(t)\right]}{m^2} \\
&= \frac{L^2}{m} + \frac{\sum_{i\neq j}\mathbb{E}_{\mathcal{F}_{t-1}}\left[\mathbb{E}\left[\mathbf{g}_i(t)^\top\mathbf{g}_j(t)|\mathcal{F}_{t-1}\right]\right]}{m^2}.
\end{aligned}$$

Now using the fact that the gradients $\mathbf{g}_i(t)$ are unbiased estimates of $\nabla J_i(\mathbf{w}_t)$ and that $\mathbf{g}_i(t)$ and $\mathbf{g}_j(t)$ are independent given past history and inequality (55) for node i and j we get

$$\begin{aligned}
&\frac{\sum_{i\neq j}\mathbb{E}_{\mathcal{F}_{t-1}}\left[\mathbb{E}\left[\mathbf{g}_i(t)^\top\mathbf{g}_j(t)|\mathcal{F}_{t-1}\right]\right]}{m^2} \\
&= \sum_{i\neq j}\frac{\mathbb{E}_{\mathcal{F}_{t-1}}\left[\nabla J_i(\mathbf{w}_i(t))^\top\nabla J_j(\mathbf{w}_j(t))\right]}{m^2} \\
&\leq \sum_{i\neq j}\frac{\sqrt{\mathbb{E}_{\mathcal{F}_{t-1}}\left[\|\nabla J_i(\mathbf{w}_i(t))\|^2\right]}\sqrt{\mathbb{E}_{\mathcal{F}_{t-1}}\left[\|\nabla J_j(\mathbf{w}_j(t))\|^2\right]}}{m^2} \\
&= \frac{(m-1)}{m}\cdot 30L^2\rho^2 \\
&\leq 30L^2\rho^2. \qquad (56)
\end{aligned}$$

Therefore to bound the term T2 in (35) we can use

$$\mathbb{E}\left[\left\|\sum_{i=1}^{m}\frac{\mathbf{g}_i(t)}{m}\right\|^2\right] \leq \frac{L^2}{m} + 30L^2\rho^2. \qquad (57)$$

F. Bound for $T3$

Applying (44), (52), and (55) to T3 in (35), as well as Lemma 8 and the fact that $(a_1+a_2)^2 \leq 2a_1^2 + 2a_2^2$ we obtain the following bound:

$$\begin{aligned}
T3 &\leq \frac{1}{m}\sum_{i=1}^{m}\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}_i(t)\|^2\right]} \\
&\qquad\cdot\sqrt{\mathbb{E}\left[(\|\nabla J_i(\mathbf{w}_i(t))\| + \|\nabla J_i(\bar{\mathbf{w}}(t))\|)^2\right]} \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\frac{2L\sqrt{m}}{\mu}\frac{\log(2bet^2)}{bt}\cdot 10L\rho \\
&\leq \frac{20L^2}{\mu}\cdot\frac{\sqrt{m}}{b}\cdot\frac{\log(T)}{t}\cdot\rho. \qquad (58)
\end{aligned}$$

G. Combining the Bounds

Finally combining (57) and (58) in (35) and applying the step size assumption $\eta_t = 1/(\mu t)$:

$$\begin{aligned}
&\mathbb{E}\left[J(\bar{\mathbf{w}}(t)) - J(\mathbf{w}^*)\right] \\
&\le \frac{(\eta_t^{-1}-\mu)}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\right] \\
&\quad - \frac{\eta_t^{-1}}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t+1)-\mathbf{w}^*\|^2\right] \\
&\quad + \left(\frac{30L^2\rho^2}{\mu}+\frac{L^2}{\mu m}\right)\cdot\frac{1}{t} \\
&\quad + \frac{20L^2}{\mu}\cdot\frac{\sqrt{m}}{b}\cdot\frac{\log(2bet^2)}{t}\cdot\rho \\
&\le \frac{\mu(t-1)}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t)-\mathbf{w}^*\|^2\right] \\
&\quad - \frac{\mu t}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(t+1)-\mathbf{w}^*\|^2\right] + K_0\cdot\frac{L^2}{\mu t},
\end{aligned} \tag{59}$$

where $K_0 = \left(30\rho^2 + 1/m + \left(60\cdot\sqrt{m\rho^2}\cdot\log(T)\right)/b\right)$, using $t \le T$ and assuming $T > 2be$.

Let us now define two new sequences, the average of the average of iterates over nodes from $t = 1$ to T and the average for any node $i \in [m]$

$$\hat{\mathbf{w}}(T) = \frac{1}{T}\sum_{t=1}^{T}\bar{\mathbf{w}}(t) \tag{60}$$

$$\hat{\mathbf{w}}_i(T) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_i(t). \tag{61}$$

Then summing (59) from $t = 1$ to T, using the convexity of J and collapsing the telescoping sum in the first two terms of (59),

$$\begin{aligned}
&\mathbb{E}\left[J(\hat{\mathbf{w}}(T)) - J(\mathbf{w}^*)\right] \\
&\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[J(\bar{\mathbf{w}}(t)) - J(\mathbf{w}^*)\right] \\
&\le -\frac{\mu T}{2}\mathbb{E}\left[\|\bar{\mathbf{w}}(T+1)-\mathbf{w}^*\|^2\right] + K_0\cdot\frac{L^2}{\mu}\cdot\frac{\sum_{t=1}^{T}1/t}{T} \\
&\le K_0\cdot\frac{L^2}{\mu}\cdot\frac{\log(T)}{T}.
\end{aligned} \tag{62}$$

Now using the definition of subgradient, Cauchy-Schwarz, and Jensen's inequality we have

$$\begin{aligned}
&J(\hat{\mathbf{w}}_i(T)) - J(\mathbf{w}^*) \\
&\le J(\hat{\mathbf{w}}(T)) - J(\mathbf{w}^*) + \nabla J(\hat{\mathbf{w}}_i(T))^\top(\hat{\mathbf{w}}_i(t) - \hat{\mathbf{w}}(T)) \\
&\le J(\hat{\mathbf{w}}(T)) - J(\mathbf{w}^*) + \|\nabla J(\hat{\mathbf{w}}_i(T)\|\,\|\hat{\mathbf{w}}_i(t) - \hat{\mathbf{w}}(T)\| \\
&\le J(\hat{\mathbf{w}}(T)) - J(\mathbf{w}^*) \\
&\qquad + \|\nabla J(\hat{\mathbf{w}}_i(T))\|\cdot\sum_{t=1}^{T}\frac{\|\mathbf{w}_i(t)-\bar{\mathbf{w}}(t)\|}{T}.
\end{aligned} \tag{63}$$

To proceed we must bound $\mathbb{E}\left[\|\nabla J(\hat{\mathbf{w}}_i(T))\|^2\right]$ in a similar way as the bound (52). First, let $\alpha_i = \partial\ell(\hat{\mathbf{w}}_i(T)^\top\mathbf{x}_i)$ denote the subgradient for the i-th loss function of $J(\cdot)$ in (1), evaluated at $\hat{\mathbf{w}}_i(T)$, and $\boldsymbol{\alpha}_T = (\alpha_1, \alpha_2, \ldots, \alpha_N)^\top$ be the vector of subgradients. As before,

$$\begin{aligned}
\|\nabla J(\hat{\mathbf{w}}_i(T))\|^2 &= \left\|\frac{1}{N}\sum_{i=1}^{N}\alpha_i\mathbf{x}_i + \mu\hat{\mathbf{w}}_i(T)\right\|^2 \\
&\le \frac{2}{N^2}\boldsymbol{\alpha}^\top\mathbf{Q}\boldsymbol{\alpha} + 2\mu^2\|\hat{\mathbf{w}}_i(T)\|^2 \\
&\le 10L^2\rho^2 + 2\mu^2\|\hat{\mathbf{w}}_i(T)\|^2 \\
&\le 10L^2\rho^2 + 2\mu^2\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{w}_i(t)\|^2.
\end{aligned}$$

Taking expectations of both sides and using (54) as before:

$$\mathbb{E}\left[\|\nabla J(\hat{\mathbf{w}}_i(T))\|^2\right] \le 30L^2\rho^2.$$

Taking expectations of both sides of (63) and using the Cauchy-Schwarz inequality, (62), the preceding gradient bound, Lemma 8 and the definition of K_0 we get

$$\begin{aligned}
&\mathbb{E}\left[J(\hat{\mathbf{w}}_i(T)) - J(\mathbf{w}^*)\right] \\
&\le K_0\cdot\frac{L^2}{\mu}\cdot\frac{\log(T)}{T} + \frac{2\sqrt{30}L^2}{\mu}\cdot\frac{\sqrt{m}}{b}\cdot\rho\cdot\frac{\log(T)}{T}\cdot\sum_{t=1}^{T}\frac{1}{t} \\
&\le \left(K_0 + \frac{2\sqrt{30}\cdot\sqrt{m\rho^2}\cdot\log T}{b}\right)\cdot\frac{\log T}{T} \\
&\le \left(30\rho^2 + \frac{1}{m} + \frac{70\sqrt{m\rho^2}\cdot\log T}{b}\right)\cdot\frac{L^2}{\mu}\cdot\frac{\log T}{T}.
\end{aligned} \tag{64}$$

Recalling that $b = \log(1/\lambda_2(\mathbf{P})) \ge 1 - \lambda_2(P)$, assuming $T > 2be$ and subsuming the first term in the third and taking expectations with respect to the sample split the above bound can be written as

$$\begin{aligned}
\mathbb{E}\left[J(\hat{\mathbf{w}}_i(T)) - J(\mathbf{w}^*)\right] \le &\left(\frac{1}{m} + \frac{100\sqrt{m\rho^2}\cdot\log T}{1-\lambda_2(P)}\right) \\
&\cdot\frac{L^2}{\mu}\cdot\frac{\log T}{T}.
\end{aligned} \tag{65}$$

■

Appendix B
Proof of Lemma 2

Proof: Let us define the product of the sequence of random matrices $\{\mathbf{P}(\tau) : s \le \tau \le t\}$:

$$\Phi(s:t) = \mathbf{P}(t)\cdots\mathbf{P}(s). \tag{66}$$

Then proceeding as in proof of Lemma 8 and using the step size $\eta_t = 1/(\mu t)$, we get

$$\begin{aligned}
\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\| \le &\left\|\sum_{s=1}^{t-1}\eta_s\sum_{j=1}^{m}\left(\frac{1}{m} - \Phi(s:t)_{ij}\right)\mathbf{g}_j(s)\right. \\
&\left. + \eta_t\left(\sum_{j=1}^{m}\frac{1}{m}\mathbf{g}_j(t) - \mathbf{g}_i(t)\right)\right\|
\end{aligned} \tag{67}$$

$$\le \sum_{s=1}^{t-1}\frac{L}{\mu s}\cdot\left\|\frac{1}{m} - \Phi(s:t)\mathbf{e}_i\right\|_1 + \frac{2L}{\mu t}. \tag{68}$$

Let $\mathbf{e}_i$ be a vector with 0's everywhere except at the the ith position, then

$$\mathbb{E}\left[\left\|\frac{1}{m} - \Phi(s:t)\mathbf{e}_i\right\|_1\right] \le \sqrt{m}\mathbb{E}\left[\left\|\frac{1}{m} - \Phi(s:t)\mathbf{e}_i\right\|_2\right]. \tag{69}$$

Consider the recursion $\mathbf{u}(t+1) = \mathbf{P}(t)\mathbf{u}(t)$ and let $\mathbf{v}(t+1) = \mathbf{P}(t)\mathbf{u}(t) - \frac{1}{m}$ then we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{v}(t+1)^\top \mathbf{v}(t+1)|\mathbf{v}(t)\right] &= \mathbb{E}\left[\mathbf{v}(t)^\top \mathbf{P}^2(t)\mathbf{v}(t)|\mathbf{v}(t)\right] \\ &= \mathbf{v}(t)^\top \mathbb{E}\left[\mathbf{P}^2(t)\right]\mathbf{v}(t) \\ &\le \|\mathbf{v}(t)\|^2 \lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right),\end{aligned} \tag{70}$$

since $\mathbf{v}(t)$ is orthogonal to the largest eigenvector of $\mathbf{P}(t)$.

Taking expectations w.r.t to $\mathbf{v}(t)$ we get

$$\mathbb{E}\left[\|\mathbf{v}(t+1)\|^2\right] \le \mathbb{E}\left[\|\mathbf{v}(t)\|^2\right]\lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right). \tag{71}$$

Recursively expanding (71) we obtain

$$\mathbb{E}\left[\|\mathbf{v}(t+1)\|^2\right] \le \|\mathbf{v}(0)\|^2 \lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right)^{t-s+1}. \tag{72}$$

Consider an initial vector $\mathbf{u}(0) = \mathbf{e}_i$. We see that $\|\mathbf{v}(t+1)\|^2 = \left\|\frac{1}{m} - \Phi(s:t)_i\right\|_2$, this finally gives us

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{m} - \Phi(s:t)_i\right\|_1\right] \\ &\le \sqrt{m}\mathbb{E}\left[\left\|\frac{1}{m} - \Phi(s:t)_i\right\|_2\right] \\ &\le \sqrt{m}\left\|\mathbf{e}_i - \frac{1}{m}\right\|^2 \lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right)^{t-s+1} \\ &\le \sqrt{m}\lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right)^{t-s+1}.\end{aligned} \tag{73}$$

Proceeding like the proof of Lemma 2 where $a = \lambda_2\left(\mathbb{E}\left[\mathbf{P}^2(t)\right]\right)$ and $b = -\log(a)$ we get

$$\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]} \le \frac{2L\sqrt{m}}{\mu}\frac{\log(2bet^2)}{bt}. \tag{74}$$

■

Appendix C
Proof of Theorem 3

The proof follows easily from the proof of Theorem 1.

Proof: Since (35) still holds, we merely apply Lemma 2 in (35) and continue in the same way as the proof of Theorem 1. ■

Appendix D
Proof of Theorem 4

We will first establish the network lemma for scheme (13).

Lemma 9: Fix a Markov matrix $\mathbf{P}$ and consider Algorithm 1 when the objective $J(\mathbf{w})$ is strongly convex and the frequency of communication satisfies

$$1/\nu > \frac{4}{3\rho^2}\log(d) \tag{75}$$

we have the following inequality for the expected squared error between the iterate $\mathbf{w}_i(t)$ at node i at time t and the average $\bar{\mathbf{w}}(t)$ defined in Algorithm 1 for scheme (13)

$$\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]} \le \frac{4L\sqrt{5m\rho^2}}{\mu}\cdot\frac{\log(2bet^2)}{bt} \tag{76}$$

where $b = (1/2)\log(1/\lambda_2(\mathbf{P}))$.

Proof: It is easy to see that we can write the update equation in Algorithm 1 as

$$\mathbf{w}_i(t+1) = \sum_{j=1}^m \tilde{\mathbf{P}}_{ij}(t)\mathbf{w}_j(t) - \eta_t \mathbf{g}_i^{1/\nu}(t) \tag{77}$$

where

$$\tilde{\mathbf{P}}_{ij}(t) = \begin{cases} P_{ij}(t) & \text{when } i \ne j \\ P_{ii}(t) - \frac{1}{mt} & \text{when } i = j \end{cases} \tag{78}$$

and $\mathbf{g}_i(t) = \mathbf{g}_i^{1/\nu}(t) + \mu\mathbf{w}_i(t)$.

We need first a bound on $\left\|\mathbf{g}_j^{1/\nu}(s)\right\|$ using the definition of the minibatch (sub)gradient:

$$\begin{aligned}\left\|\mathbf{g}_i^{1/\nu}(s)\right\|^2 &= \left\|\frac{\sum_{i_{k_s}\in H_s^i}\partial\ell(\mathbf{w}_i(s)^\top\mathbf{x}_{k_{i_s}})\mathbf{x}_{k_{i_s}}}{1/\nu}\right\|^2 \\ &\le L^2\nu\left\|\mathbf{Q}_{1/\nu}\right\|\end{aligned} \tag{79}$$

From (39) and the minibatch (sub)gradient bound

$$\begin{aligned}&\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\| \\ &\le \left\|\sum_{s=1}^{t-1}\eta_s\sum_{j=1}^m\left(\frac{1}{m} - (\tilde{\mathbf{P}}^{t-s})_{ij}\right)\mathbf{g}_j^{1/\nu}(s)\right\| \\ &\quad + \eta_t\left\|\left(\sum_{j=1}^m\frac{\mathbf{1}}{m}\mathbf{g}_j^{1/\nu}(t) - \mathbf{g}_i^{1/\nu}(t)\right)\right\| \\ &\le L\sqrt{\nu\left\|\mathbf{Q}_{1/\nu}\right\|}\sum_{s=1}^{t-1}\frac{\left\|\frac{\mathbf{1}}{m} - (\tilde{\mathbf{P}}^{t-s})_i\right\|_1}{\mu s} + \frac{2L\sqrt{\nu\left\|\mathbf{Q}_{1/\nu}\right\|}}{\mu t} \\ &\le L\sqrt{\nu\left\|\mathbf{Q}_{1/\nu}\right\|} \\ &\qquad \sum_{s=1}^{t-1}\frac{\left\|\frac{\mathbf{1}}{m} - (\mathbf{P}^{t-s})_i\right\|_1 + \left\|(\mathbf{P}^{t-s})_i - (\tilde{\mathbf{P}}^{t-s})_i\right\|_1}{\mu s} \\ &\quad + \frac{2L\sqrt{\nu\left\|\mathbf{Q}_{1/\nu}\right\|}}{\mu t} \\ &\le 2L\sqrt{\nu\left\|\mathbf{Q}_{1/\nu}\right\|}\sum_{s=1}^{t-1}\frac{\left\|\frac{\mathbf{1}}{m} - (\mathbf{P}^{t-s})_i\right\|_1}{\mu s} + \frac{2L\sqrt{\nu\left\|\mathbf{Q}_{1/\nu}\right\|}}{\mu t}\end{aligned}$$

Continuing as in the proof of Lemma 8, taking expectations and using Lemma 7, for $1/\nu > \frac{4}{3\rho^2}\log(d)$ we have

$$\begin{aligned}\sqrt{\mathbb{E}\left[\|\bar{\mathbf{w}}(t) - \mathbf{w}_i(t)\|^2\right]} &\le \frac{4L\sqrt{m\nu\mathbb{E}\left[\left\|\mathbf{Q}^{1/\nu}\right\|\right]}}{\mu}\frac{\log(2bet^2)}{bt} \\ &\le \frac{4L\sqrt{5m\rho^2}}{\mu}\frac{\log(2bet^2)}{bt}\end{aligned} \tag{80}$$

■

For the scheme (13) all the steps until bound (35) from proof of Theorem 3 remain the same. The difference in the rest of the proof arises primarily from the mini batch gradient norm factor in Lemma 9. We have the same decomposition as (35) with T1, T2, and T3 as in (32), (33), and (34). The gradient norm bounds also don't change since the minibatch gradient is also an unbiased gradient of the true gradient $\nabla J(\cdot)$. Thus substituting Lemma 9 in the above and following the same steps as in proof of Theorem 3, replacing T by νT where T is now the total iterations including the communication as well as the minibatch gathering rounds, we get Theorem 4.

A. Proof of Lemma 5

In the proof we will first show that the iterate of Algorithm 1 is asymptotically normal by showing it is close to the iterate of distributed algorithm of [9] and then use the corresponding multivariate normality result of Bianchi et al. [9, Theorem 5]. Finally using smoothness and strong convexity we shall get Lemma 5.

We need to verify that Algorithm 1 satisfies all the assumptions necessary (Assumptions **1**, **4**, **6**, **7**, **8a**, and **8b** in Bianchi et al. [9]) for the result to hold.

- Assumption **1** requires the weight matrix $\mathbf{P}(t)$ to be row stochastic almost surely, identically distributed over time, and that $\mathbb{E}[\mathbf{P}(t)]$ is column stochastic. Our Markov matrix is constant over time and doubly stochastic. Assumption **1b** follows because $\mathbf{P}$ is constant and independent of the stochastic gradients, which are sampled uniformly with replacement.
- Assumption **4** requires square integrability of the gradients as well as a regularity condition. In our setting, this follows since the sampled gradients are bounded almost everywhere.
- Assumption **6** imposes some analytic conditions at the optimum value. These hold since the gradient is assumed to be differentiable and the Hessian matrix at $\mathbf{w}^*$ is positive definite with its smallest eigenvalue is at least $\mu > 0$ (this follows from strong convexity).
- Assumption **7** of Bianchi et al. [9] follows from our existing assumptions.
- Assumptions **8a** and **8b** are standard stochastic approximation assumptions on the step size that are easily satisfied by $\eta_t = \frac{1}{\mu t}$.

It is easy to show that the average over the nodes of the iterates $\tilde{\mathbf{w}}_i(t)$, $\mathbf{w}_i(t)$ for Algorithm 1 and distributed algorithm of [9] are the same and satisfy

$$\begin{aligned}\bar{\tilde{\mathbf{w}}}(t+1) &= \bar{\tilde{\mathbf{w}}}(t) - \eta_t \frac{\sum_{i=1}^m \mathbf{g}_i(t)}{m} \\ \bar{\mathbf{w}}_i(t+1) &= \bar{\mathbf{w}}_i(t+1) - \eta_t \frac{\sum_{i=1}^m \mathbf{g}_i(t)}{m} \qquad (81)\end{aligned}$$

Now note that

$$\mathbf{w}_i(t) - \mathbf{w}^* = \underbrace{\mathbf{w}_i(t) - \bar{\mathbf{w}}_i(t)}_{\textbf{T1=Network Error}} + \underbrace{\bar{\mathbf{w}}_i(t) - \mathbf{w}^*}_{\textbf{T2=Asymptotically Normal}} \qquad (82)$$

From Lemma 8 we know that the network error (**T1**) decays and from update equation (81) we know that the averaged iterates for both the versions are the same . Then the proof of Theorem 5 of Bianchi et al. [9] shows that the term **T2**, under the above assumptions when appropriately normalized converges to a centered Gaussian distribution. Equation (82) then implies

$$\sqrt{\mu t}\left(\mathbf{w}_i(t) - \mathbf{w}^*\right) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{H}\right), \qquad (83)$$

where $\mathbf{H}$ is the solution to the equation

$$\nabla J^2(\mathbf{w}^*)\mathbf{H} + \mathbf{H}\nabla J^2(\mathbf{w}^*)^T = \mathbf{C}. \qquad (84)$$

Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then for $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$ we have $\mathbf{X} = \mathbf{Y}\mathbf{H}^{1/2}$ and thus

$$\|\mathbf{X}\|^2 = \mathbf{Y}^\top \mathbf{H} \mathbf{Y} \qquad (85)$$

Then it is well known that $\|\mathbf{X}\|^2 \sim \chi^2(\mathrm{Tr}(\mathbf{H}))$ and so $\mathbb{E}\left[\|\mathbf{X}\|^2\right] = \mathrm{Tr}(\mathbf{H})$.

Let us now consider the suboptimality at the iterate $\sum_{j=1}^m P_{ij}\mathbf{w}_j(t)$. It is easy to see that for a differentiable and strongly convex function

$$J\left(\sum_{j=1}^m P_{ij}\mathbf{w}_j(t)\right) - J(\mathbf{w}^*) \leq \frac{G}{2}\left\|\sum_{j=1}^m P_{ij}\mathbf{w}_j(t) - \mathbf{w}^*\right\|^2. \qquad (86)$$

Now it is easy to see from (83) that for a node $j \in \mathcal{N}(i)$

$$P_{ij}\sqrt{\mu t}\left(\mathbf{w}_j(t) - \mathbf{w}^*\right) \sim \mathcal{N}\left(\mathbf{0}, (P_{ij})^2\mathbf{H}\right). \qquad (87)$$

This implies that

$$\sum_{j \in \mathcal{N}(i)} P_{ij}\sqrt{\mu t}\left(\mathbf{w}_j(t) - \mathbf{w}^*\right) \sim \mathcal{N}\left(\mathbf{0}, \left(\sum_{j \in \mathcal{N}(i)} (P_{ij})^2\right)\mathbf{H}\right). \qquad (88)$$

Then taking expectation w.r.t to the distribution (88) and using standard properties of norms of multivariate normal variables,

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{j \in \mathcal{N}(i)} P_{ij}\sqrt{\mu t}\left(\mathbf{w}_j(t) - \mathbf{w}^*\right)\right\|^2\right] \\ &= \left(\sum_{j \in \mathcal{N}(i)} (P_{ij})^2\right)\mathrm{Tr}\left(\mathbf{H}\right). \qquad (89)\end{aligned}$$

Then substituting in bound (86) and taking the limit we finally get

$$\begin{aligned}&\limsup_{T \to \infty} T \cdot \mathbb{E}\left[J\left(\sum_{j=1}^m P_{ij}\mathbf{w}_j(T)\right) - J(\mathbf{w}^*)\right] \\ &\leq \sum_{j \in \mathcal{N}(i)} (P_{ij})^2 \cdot \mathrm{Tr}\left(\mathbf{H}\right) \cdot \frac{G}{\mu}. \qquad (90)\end{aligned}$$

B. Proof of Theorem 6

The the covariance of the gradient noise under the sampling with replacement model is

$$\begin{aligned}\mathbf{C} &= \mathbb{E}\left[\mathbf{g}_i(t)\mathbf{g}_i(t)^\top\right] - \nabla J(\mathbf{w}_i(t))\nabla J(\mathbf{w}_i(t))^\top \\ &= \frac{\sum_{i=1}^N \beta_{i,t}\mathbf{x}_i\mathbf{x}_i^T}{N} + \frac{\mu}{N}\sum_{i=1}^N \beta_{i,t}\left(\mathbf{x}_i\mathbf{w}_i(t)^\top + \mathbf{w}_i(t)\mathbf{x}_i^\top\right) \\ &\quad + \mu^2\mathbf{w}_i(t)\mathbf{w}_i(t)^\top - \nabla J(\mathbf{w}_i(t))\nabla J(\mathbf{w}_i(t))^\top.\end{aligned} \tag{91}$$

Thus we can bound the spectral norm of $\mathbf{C}$ as

$$\begin{aligned}\sigma_1(\mathbf{C}) &\leq L^2\rho^2 + 2\mu L\mathbb{E}\left[\|\mathbf{w}_i(t)\|\right] + \mu^2\mathbb{E}\left[\|\mathbf{w}_i(t)\|^2\right] \\ &\quad + \mathbb{E}\left[\|\nabla J(\mathbf{w}_i(t))\|^2\right].\end{aligned} \tag{92}$$

Now from bound (54) since $T \to \infty$ we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_i(t)\|^2\right] &\leq \frac{10L^2\rho^2}{\mu^2} \\ \mathbb{E}\left[\|\nabla J_i(\mathbf{w}_i(t))\|^2\right] &\leq 30L^2\rho^2.\end{aligned}$$

Putting everything together we get

$$\sigma_1(\mathbf{C}) \leq 50\rho L^2. \tag{93}$$

Next note that $H = C\left(\nabla^2 J(\mathbf{w}^*)\right)^{-1}/2$. From the completeness and uniform weight assumptions on the graph, we have

$$\sum_{j\in\mathcal{N}(i)} (P_{ij})^2 = \frac{1}{m}. \tag{94}$$

Thus substituting in Lemma 5, using (93) gives us

$$\begin{aligned}&\limsup_{t\to\infty} t\cdot\mathbb{E}\left[J\left(\sum_{j=1}^m P_{ij}\mathbf{w}_j(t)\right) - J(\mathbf{w}^*)\right] \\ &\leq \frac{1}{m}\cdot\frac{\mathrm{Tr}\left(\left(\mathbf{C}\nabla^2 J(\mathbf{w}^*)\right)^{-1}\right)}{2}\cdot\frac{G}{\mu} \\ &\leq \frac{25\rho L^2}{m}\cdot\mathrm{Tr}\left(\nabla^2 J(\mathbf{w}^*)^{-1}\right)\cdot\frac{G}{\mu}.\end{aligned}$$

References

[1] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," ArXiV, extended version of ICML paper arXiv:1109.5647 [cs.LG], 2012. [Online]. Available: http://arxiv.org/abs/1109.5647

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, UK: Cambridge, 2014.

[3] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, January 2009. [Online]. Available: http://dx.doi.org/10.1109/TAC.2008.2009515

[4] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM," *Mathematical Programming, Series B*, vol. 127, no. 1, pp. 3–30, October 2011. [Online]. Available: http://dx.doi.org/10.1007/s10107-010-0420-4

[5] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, March 2011. [Online]. Available: http://dx.doi.org/10.1109/TAC.2011.2161027

[6] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, December 2011. [Online]. Available: http://dx.doi.org/10.1007/s10957-010-9737-7

[7] Y. Zhang, J. Duchi, and M. Wainwright, "Communication-efficient algorithms for statistical optimization," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1511–1519. [Online]. Available: http://books.nips.cc/papers/files/nips25/NIPS2012_0716.pdf

[8] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *Proceedings of the 31st International Conference on Machine Learning*, ser. JMLR Workshop and Conference Proceedings, E. P. Xing and T. Jebara, Eds., vol. 32, 2014, pp. 1000–1008. [Online]. Available: http://jmlr.org/proceedings/papers/v32/shamir14.html

[9] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7405–7418, 2013. [Online]. Available: http://dx.doi.org/10.1109/TIT.2013.2275131

[10] J. Liu, S. J. Wright, C. Re, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," in *Proceedings of the 31st International Conference on Machine Learning*, ser. JMLR Workshop and Conference Proceedings, L. Getoor and T. Scheffer, Eds., vol. 32, 2014. [Online]. Available: http://jmlr.org/proceedings/papers/v32/liud14.pdf

[11] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, "Parallel coordinate descent for l_1-regularized loss minimization," in *Proceedings of the 28th International Conference on Machine Learning*, ser. JMLR Workshop and Conference Proceedings, L. Getoor and T. Scheffer, Eds., vol. 28, 2011. [Online]. Available: http://www.select.cs.cmu.edu/publications/paperdir/icml2011-bradley-kyrola-bickson-guestrin.pdf

[12] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 873–881. [Online]. Available: http://books.nips.cc/papers/files/nips24/NIPS2011_0574.pdf

[13] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro, "Mini-batch primal and dual methods for SVMs," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, ser. JMLR Workshop and Conference Proceedings, S. Dasgupta and D. McAllester, Eds., vol. 28, 2013, pp. 1022–1030. [Online]. Available: http://jmlr.org/proceedings/papers/v28/takac13.html

[14] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan, "Better mini-batch algorithms via accelerated gradient methods," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 1647–1655. [Online]. Available: http://papers.nips.cc/paper/4432-better-mini-batch-algorithms-via-accelerated-gradient-methods

[15] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches," *Journal of Machine Learning Research*, vol. 13, pp. 165–202, January 2012. [Online]. Available: http://jmlr.org/papers/v13/dekel12a.html

[16] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic convex optimization," in *Proceedings Conference on Learning Theory*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2009.

[17] L. Yu, N. Wang, and X. Meng, "Real-time forest fire detection with wireless sensor networks," in *Wireless Communications, Networking and Mobile Computing Proceedings*, vol. 2, no. 11, 2005, pp. 1214 – 1217.

[18] A. Mokhtari and A. Ribeiro, "DSA: decentralized double stochastic averaging gradient algorithm," *Journal of Machine Learning Research*, vol. 17, no. 61, pp. 1–35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-292.html

[19] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[20] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," HAL, Tech. Rep. hal-00860051, January 2015. [Online]. Available: https://hal.inria.fr/hal-00860051v2

[21] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proceedings International Conference on Machine Learning*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1511–1519.

[22] K. I. Tsianos and M. G. Rabbat, "Distributed strongly convex optimization," ArXiV, Tech. Rep. arXiv:1207.3031 [cs.DC], July 2012. [Online]. Available: http://arxiv.org/abs/1207.3031

[23] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, “Communication/computation tradeoffs in consensus-based distributed optimization,” in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1952–1960.

[24] J. A. Blackard and D. J. Dean, “Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables,” *Computers and Electronics in Agriculture*, vol. 24, no. 3, pp. 131–151, December 1999. [Online]. Available: http://dx.doi.org/10.1016/S0168-1699(99)00046-0

[25] M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[26] M.-R. Amini, N. Usunier, and C. Goutte, “Learning from multiple partially observed views - an application to multilingual text categorization,” in *Advances in Neural Information Processing Systems 22*, 2009.

[27] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. [Online]. Available: http://dx.doi.org/10.1145/1961189.1961199

[28] S. Boyd, L. Xiao, and P. Diaconis, “Fastest mixing markov chain on a graph,” *SIAM Review*, vol. 46, no. 4, pp. 667–689, 2004. [Online]. Available: http://dx.doi.org/10.1137/S0036144503423264

[29] A. S. Bijral and N. Srebro, “On doubly stochastic graph optimization,” in *NIPS Workshop on Analyzing Networks and Learning with Graphs*, 2009.

[30] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, August 2012. [Online]. Available: http://dx.doi.org/10.1007/s10208-011-9099-z