Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright, *Senior Member, IEEE*

Abstract—The goal of decentralized optimization over a network is to optimize a global objective formed by a sum of local (possibly nonsmooth) convex functions using only local computation and communication. It arises in various application domains, including distributed tracking and localization, multi-agent co-ordination, estimation in sensor networks, and large-scale machine learning. We develop and analyze distributed algorithms based on dual subgradient averaging, and we provide sharp bounds on their convergence rates as a function of the network size and topology. Our analysis allows us to clearly separate the convergence of the optimization algorithm itself and the effects of communication dependent on the network structure. We show that the number of iterations required by our algorithm scales inversely in the spectral gap of the network and confirm this prediction's sharpness both by theoretical lower bounds and simulations for various networks. Our approach includes the cases of deterministic optimization and communication as well as problems with stochastic optimization and/or communication.

I. INTRODUCTION

THE focus of this paper is the development and analysis of distributed algorithms for solving convex optimization problems that are defined over networks. Network-structured optimization problems arise in a variety of domains within the information sciences and engineering. For instance, problems such as multi-agent coordination, distributed tracking and localization, estimation problems in sensor networks and packet routing are all naturally cast as distributed convex minimization [1], [2], [3], [4]. Common to these problems is the necessity for completely decentralized computation that is locally light—to avoid overburdening small sensors or flooding busy networks—and robust to periodic link or node failures. As a second example, data sets too large to be processed quickly by any single processor present related challenges. A canonical instance from statistical machine learning is the problem of minimizing a loss function averaged over a large dataset (e.g. support vector machines [5]). With terabytes of data, it is desirable to assign subsets of the data to different processors, and the processors must communicate to find parameters minimizing the loss over the entire dataset. However, the communication should be efficient enough that network latencies do not offset computational gains.

Distributed computation has a long history in optimization, and the 1980s saw significant interest in distributed detection, consensus, and minimization. The seminal work of Tsitsiklis and colleagues [6], [7], [1] analyzed algorithms for minimization of a smooth function f known to several agents while distributing processing of components of the parameter vector $x \in \mathbb{R}^n$. More recently, a few researchers have shifted focus to problems in which each processor locally has its own convex (potentially non-differentiable) objective function [8], [9], [10], [11].

Our paper makes two main contributions. The first contribution is to provide a new simple subgradient algorithm for distributed constrained optimization of a convex function; we refer to it as a *dual averaging subgradient method*, since it is based on maintaining and forming weighted averages of subgradients throughout the network. This approach is essentially different from previously developed methods [8], [9], [10], and these differences facilitate our analysis of network scaling issues, meaning how convergence rates depend on network size and topology. Indeed, the second main contribution of this paper is a careful analysis that demonstrates a close link between convergence of the algorithm and the underlying spectral properties of the network.

By comparison to previous work, our convergence results and proofs are different, and our characterization of network scaling terms is often much stronger. As detailed comparison with past work requires presentation of several of our results, we give only a brief overview of related work here, deferring detailed discussion to Sec. IV. The sharpest results given previously for distributed projected gradient descent are discussed in the papers [12], [10] who show that if the number of time steps is known a priori and the stepsize is chosen optimally, an ϵ-optimal solution to the optimization problem can be reached in $\mathcal{O}(n^3/\epsilon^2)$ time. Since this bound is essentially independent of network topology, it does not capture the intuition that distributed algorithms should converge much faster on "well-connected" networks—expander graphs being a prime example—than on poorly connected networks (e.g., chains, trees or single cycles). Johansson et al. [11] analyze a low communication peer-to-peer protocol that attains rates dependent on network structure; however, in their algorithm only one agent has a current parameter value, while all agents in our algorithm maintain good estimates of the optimum at all times. This is important in online,

JCD, AA, and MJW are with the Department of Electrical Engineering and Computer Sciences at UC Berkeley, Berkeley, CA 94720. Additionally MJW is affiliated with the Department of Statistics, UC Berkeley, Berkeley, CA 94720. Email: {jduchi,alekh,wainwrig}@eecs.berkeley.edu.

streaming, and control problems where agents are expected to act in real time. In additional comparison to previous work, our development yields network scaling terms that are often substantially sharper, specifically, our convergence rate scales inversely in the spectral gap of the network. This allows us to build on known results to show that our algorithm obtains an ϵ-optimal solution in $\mathcal{O}(n^2/\epsilon^2)$ iterations for a single cycle or path, $\mathcal{O}(n/\epsilon^2)$ iterations for a two-dimensional grid, and $\mathcal{O}(1/\epsilon^2)$ iterations for a bounded degree expander graph. We also show that the network deviation terms we derive are tight for our algorithm. Moreover, results on a simulated system identification task using robust linear regression show excellent agreement with our theoretical predictions.

Our analysis covers several settings for distributed minimization. We begin by studying fixed communication protocols, which are of interest in areas such as cluster computing or sensor networks with a fixed hardware-dependent protocol. We also investigate randomized communication protocols and randomized network failures, which are essential to handle gracefully in wireless sensor networks and large clusters with potential node failures. Randomized communication provides an interesting tradeoff between communication savings and convergence rates. In this setting, we obtain much sharper results than previous work by studying the spectral properties of the expected transition matrix of a random walk on the underlying graph. We also present a relatively straightforward extension of our analysis for problems with stochastic gradient information.

The remainder of this paper is organized as follows. Section II is devoted to a formal statement of the problem and description of the dual averaging algorithm, whereas Section III states the main results and consequences of our paper, which we complement in Section IV with a comparison to previous work. In Section V, we state and prove basic convergence results on the dual averaging algorithm, which we then exploit in Section VI to derive concrete results that depend on the spectral gap of the network. Sections VII and VIII treat extensions with noise, in particular algorithms with random communication and stochastic gradients respectively. In Section IX, we present the results of simulations that confirm the sharpness of our analysis.

II. PROBLEM SET-UP AND ALGORITHM

In this section, we provide a formal statement of the distributed minimization problem and a description of the distributed dual averaging algorithm.

A. Distributed minimization

We consider an optimization problem based on functions that are distributed over a network. More specifically, let $G = (V, E)$ be an undirected graph over the vertex set $V = \{1, 2, \ldots, n\}$ with edge set $E \subset V \times V$. Associated with each $i \in V$ is convex function $f_i : \mathbb{R}^d \to \mathbb{R}$, and our overarching goal is to solve the optimization problem

$$\min_x \frac{1}{n} \sum_{i=1}^n f_i(x) \qquad \text{subject to } x \in \mathcal{X}, \qquad (1)$$

where $\mathcal{X}$ is a closed convex set. Each function f_i is convex and hence sub-differentiable, but need not be smooth. We assume without loss of generality that $0 \in \mathcal{X}$, since we can simply translate $\mathcal{X}$. Each node $i \in V$ is associated with a separate agent, and each agent i maintains its own parameter vector $x_i \in \mathbb{R}^d$. The graph G imposes communication constraints on the agents: in particular, agent i has local access to only the objective function f_i and can communicate directly only with its immediate neighbors $j \in N(i) := \{j \in V \mid (i, j) \in E\}$.

Problems of this nature arise in a variety of application domains. A concrete motivating example is the machine learning problem first described in Section I. In this case, the set $\mathcal{X}$ is the parameter space of the statistician or learner. Each function f_i is the empirical loss over the subset of data assigned to processor i, and assuming that each subset is of equal size (or that the f_i are normalized suitably), the function f is the average loss over the entire dataset. Here we use cluster computing as our computational model, where each processor is a node in the cluster, and G contains edges between processors that are directly connected with small latencies.

B. Standard dual averaging

Our algorithm is based on Nesterov's recent dual averaging algorithm [13], [14], designed for minimization of (potentially nonsmooth) convex functions f subject to the constraint $x \in \mathcal{X}$. We begin by describing the standard version of the algorithm and then discuss the extensions for the distributed setting of interest in this paper. The dual averaging scheme is based on a *proximal function* $\psi : \mathcal{X} \to \mathbb{R}$ assumed to be 1-strongly convex with respect to some norm $\|\cdot\|$, that is,

$$\psi(y) \geq \psi(x) + \langle \nabla \psi(x), y - x \rangle + \frac{1}{2} \|x - y\|^2 \quad \text{for } x, y \in \mathcal{X}.$$

In addition, we assume that $\psi \geq 0$ over $\mathcal{X}$ and that $\psi(0) = 0$; these are standard assumptions made without loss of generality. Examples of such proximal and norm pairs include:

- the quadratic $\psi(x) = \frac{1}{2} \|x\|_2^2$, which is the canonical proximal function. Clearly $\frac{1}{2} \|0\|_2^2 = 0$, and $\frac{1}{2} \|x\|_2^2$ is strongly convex with respect to the ℓ_2-norm for $x \in \mathbb{R}^d$.
- the entropic function $\psi(x) = \sum_{j=1}^d x_i \log x_i - x_i$, which is strongly convex with respect to the ℓ_1-norm for x in the probability simplex, $\{x \mid x \succeq 0, \sum_{i=1}^n x_i = 1\}$.

We assume that each function f_i is *L-Lipschitz* with respect to the same norm $\|\cdot\|$—that is,

$$|f_i(x) - f_i(y)| \leq L \|x - y\| \quad \text{for } x, y \in \mathcal{X}. \qquad (2)$$

Many cost functions f_i satisfy this type of Lipschitz condition. For instance, condition (2) holds for any convex function on a compact domain or for a polyhedral function on an arbitrary domain [15]. The bound (2) implies that for any $x \in \mathcal{X}$ and any subgradient $g_i \in \partial f_i(x)$, we have $\|g_i\|_* \leq L$, where $\|\cdot\|_*$ denotes the *dual norm* to $\|\cdot\|$, defined by $\|v\|_* := \sup_{\|u\|=1} \langle v, u \rangle$.

The dual averaging algorithm generates a sequence of iterates $\{x(t), z(t)\}_{t=0}^\infty$ contained within $\mathcal{X} \times \mathbb{R}^d$ according to the following steps. At time step t of the algorithm, it receives a subgradient $g(t) \in \partial f(x(t))$, and then performs the updates

$$z(t+1) = z(t) + g(t) \ \text{ and } \ x(t+1) = \Pi_{\mathcal{X}}^\psi(z(t+1), \alpha(t)), \ (3)$$

where $\{\alpha(t)\}_{t=0}^{\infty}$ is a non-increasing sequence of positive stepsizes, and

$$\Pi_{\mathcal{X}}^{\psi}(z, \alpha) := \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ \langle z, x \rangle + \frac{1}{\alpha} \psi(x) \right\} \qquad (4)$$

is a type of projection. The intuition underlying this algorithm is as follows: given the current iterate $(x(t), z(t))$, the next iterate $x(t+1)$ to chosen to minimize an averaged first-order approximation to the function f, while the proximal function ψ and stepsize $\alpha(t) > 0$ enforce that the iterates $\{x(t)\}_{t=0}^{\infty}$ do not oscillate wildly. The algorithm is similar to the "follow the perturbed leader" algorithms developed in the for online optimization [16], though this specific algorithm seems to be due to Nesterov [13]. In Section V, we show that a simple analysis of the convergence of the above procedure allows us to relate it to the distributed algorithm we describe.

C. Distributed dual averaging

We now consider an appropriate and novel extension of dual averaging to the distributed setting. At each iteration $t = 1, 2, 3, \ldots$, the algorithm maintains n pairs of vectors $(x_i(t), z_i(t)) \in \mathcal{X} \times \mathbb{R}^d$, with the ith pair associated with node $i \in V$. At iteration t, each node $i \in V$ computes an element $g_i(t) \in \partial f_i(x_i(t))$ in the subdifferential of the local function f_i and receives information about the parameters $\{z_j(t), j \in N(i)\}$ associated with nodes j in its neighborhood $N(i)$. Its update of the current estimated solution $x_i(t)$ is based on a convex combination of these parameters. To model this weighting process, let $P \in R^{n \times n}$ be a matrix of non-negative weights that respects the structure of the graph G, meaning that for $i \neq j$, $P_{ij} > 0$ only if $(i, j) \in E$. We assume that P is a doubly stochastic matrix, so that

$$\sum_{j=1}^{n} P_{ij} = \sum_{j \in N(i)} P_{ij} = 1 \text{ for all } i \in V, \text{ and}$$

$$\sum_{i=1}^{n} P_{ij} = \sum_{i \in N(j)} P_{ij} = 1 \text{ for all } j \in V.$$

Using this notation, given the non-increasing sequence $\alpha(t)$ of positive stepsizes, each node $i \in V$ performs the updates

$$z_i(t+1) = \sum_{j \in N(i)} p_{ij} z_j(t) + g_i(t), \quad \text{and} \qquad (5a)$$

$$x_i(t+1) = \Pi_{\mathcal{X}}^{\psi}(z_i(t+1), \alpha(t)), \qquad (5b)$$

where the projection $\Pi_{\mathcal{X}}^{\psi}$ was defined previously (4). In words, node i computes the new dual parameter $z_i(t+1)$ from a weighted average of its own subgradient $g_i(t)$ and the parameters $\{z_j(t), j \in N(i)\}$ in its neighborhood $N(i)$, and then computes the next local iterate $x_i(t+1)$ by a projection defined by the proximal function ψ and stepsize $\alpha(t) > 0$.

In the sequel, we show convergence of the local sequence $\{x_i(t)\}_{t=1}^{\infty}$ to the optimum of (1) via the *running local average*

$$\widehat{x}_i(T) = \frac{1}{T} \sum_{t=1}^{T} x_i(t). \qquad (6)$$

Note that this quantity is locally defined at node i and can be computed in a distributed manner.

III. MAIN RESULTS AND CONSEQUENCES

We will now state the main results of this paper and illustrate some of their consequences. We give the proofs and a deeper investigation of related corollaries in the sections that follow.

A. Convergence of distributed dual averaging

We start with a result on the convergence of the distributed dual averaging algorithm that provides a decomposition of the error into an optimization term and the cost associated with network communication. In order to state this theorem, we define the averaged dual variable $\bar{z}(t) := \frac{1}{n} \sum_{i=1}^{n} z_i(t)$, and we recall the definition (6) of the local average $\widehat{x}_i(T)$.

Theorem 1 (Basic convergence). *Let the sequences $\{x_i(t)\}_{t=0}^{\infty}$ and $\{z_i(t)\}_{t=0}^{\infty}$ be generated by the updates (5a)–(5b) with step size sequence $\{\alpha(t)\}_{t=0}^{\infty}$, where ψ is strongly convex with respect to the norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$. For any $x^* \in \mathcal{X}$ and for each node $i \in V$, we have*

$$f(\widehat{x}_i(T)) - f(x^*) \leq \text{OPT} + \text{NET}, \qquad (7)$$

where

$$\text{OPT} = \frac{1}{T\alpha(T)} \psi(x^*) + \frac{L^2}{2T} \sum_{t=1}^{T} \alpha(t-1) \quad and \qquad (8)$$

$$\text{NET} = \frac{L}{T} \sum_{t=1}^{T} \alpha(t) \left[\frac{2}{n} \sum_{j=1}^{n} \|\bar{z}(t) - z_j(t)\|_* + \|\bar{z}(t) - z_i(t)\|_* \right]$$

Theorem 1 guarantees that after T steps of the algorithm, every node $i \in V$ has access to a locally defined quantity $\widehat{x}_i(T)$ such that the difference $f(\widehat{x}_i(T)) - f(x^*)$ is upper bounded by a sum of four terms. The two terms in the OPT portion of the upper bound (7) are optimization error terms common to subgradient algorithms. The third and fourth (NET) are penalties incurred due to having different estimates at different nodes in the network, and they measure the deviation of each node's estimate of the average gradient from the true average gradient.[1] Thus, Theorem 1 ensures that as long the bound on the deviation $\|\bar{z}(t) - z_i(t)\|_*$ is tight enough, for appropriately chosen $\alpha(t)$ (say $\alpha(t) \propto 1/\sqrt{t}$), the error of $\widehat{x}_i(T)$ is small uniformly across all nodes, and asymptotically approaches zero. See Theorem 2 for a precise statement of rates.

B. Convergence rates and network topology

We now turn to investigation of the effects of network topology on convergence rates. In this section,[2] we assume that the network topology is static and that communication occurs via a fixed doubly stochastic weight matrix P at every round. Since P is doubly stochastic, it has largest singular value $\sigma_1(P) = 1$ (see [17, Chapter 8]). The following result shows that the convergence rate of the distributed dual averaging algorithm is controlled by the *spectral gap* $\gamma(P) := 1 - \sigma_2(P)$ of the matrix P, where $\sigma_2(P)$ is the second largest singular value of P.

---

[1]The fact that the term $\|\bar{z}(t) - z_i(t)\|_*$ appears an extra time is insignificant, as we will bound the difference $\bar{z}(t) - z_i(t)$ uniformly for all i.

[2]In later sections, we weaken these conditions.

Theorem 2 (Rates based on spectral gap)**.** *Under the conditions and notation of Theorem 1, suppose moreover that* $\psi(x^*) \leq R^2$. *With step size choice* $\alpha(t) = \frac{R\sqrt{1-\sigma_2(P)}}{4L\sqrt{t}}$,

$$f(\widehat{x}_i(T)) - f(x^*) \leq 8\frac{RL}{\sqrt{T}} \; \frac{\log(T\sqrt{n})}{\sqrt{1-\sigma_2(P)}} \qquad \text{for all } i \in V.$$

This theorem establishes a tight connection between the convergence rate of distributed subgradient methods and the spectral properties of the underlying network. The inverse dependence on the spectral gap $1 - \sigma_2(P)$ is quite natural, since it is well-known to determine the rates of mixing in random walks on graphs [18], and the propagation of information in our algorithm is integrally tied to the random walk on the underlying graph with transition probabilities specified by P.

Using Theorem 2, one can derive explicit convergence rates for several classes of interesting networks, and Figure 1 illustrates four graph topologies of interest. As a first example, the k-connected cycle in panel (a) is formed by placing n nodes on a circle and connecting each node to its k neighbors on the right and left. For small k, the cycle graph is poorly connected, and our analysis will show that this leads to slower convergence rates than other graphs with better connectivity. The grid graph in two dimensions is obtained by connecting nodes to their k nearest neighbors in axis-aligned directions. For instance, panel (b) shows an example of a degree 4 grid graph in two-dimensions. The cycle and grid are possible models for clustered computing as well as sensor networks.

In panel (c), we show a random geometric graph, constructed by placing nodes uniformly at random in $[0,1]^2$ and connecting any two nodes separated by a distance less than some radius $r > 0$. These graphs are used to model the connectivity patterns of devices, such as wireless sensor motes, that can communicate with all nodes in some fixed radius ball, and have been studied extensively (e.g., [19], [20]). There are natural generalizations to dimensions $d > 2$ as well as to cases in which the spatial positions are drawn non-uniformly.

Finally, panel (d) shows an instance of a bounded degree expander, which belongs to a special class of sparse graphs that have very good mixing properties [21]. Expanders are an attractive option for the network topology in distributed computation since they are known to have large spectral gaps. For many random graph models, a typical sample is an expander with high probability; examples include random bipartite [22] and random degree-regular graphs [23]. In addition, there are several deterministic constructions of degree regular expanders (see Section 6.3 of Chung [21]). The deterministic constructions are of interest because they can be used to design a network, while the random constructions are often much simpler.

In order to state explicit convergence rates, we need to specify a particular choice of the matrix P that respects the graph structure. Although many such choices are possible, here we focus on the graph Laplacian [21]. First, we let $A \in \mathbb{R}^{n \times n}$ be the symmetric adjacency matrix of the undirected graph G, satisfying $A_{ij} = 1$ when $(i,j) \in E$ and $A_{ij} = 0$ otherwise. For each node $i \in V$, we let $\delta_i = |N(i)| = \sum_{j=1}^{n} A_{ij}$ denote the degree of node i, and we define the diagonal

matrix $D = \text{diag}\{\delta_1, \ldots, \delta_n\}$. We assume that the graph is connected, so that $\delta_i \geq 1$ for all i, and hence D is invertible. With this notation, the *normalized graph Laplacian* is

$$\mathcal{L}(G) = I - D^{-1/2}AD^{-1/2}.$$

The graph Laplacian $\mathcal{L} = \mathcal{L}(G)$ is symmetric, positive semidefinite, and satisfies $\mathcal{L}D^{1/2}\mathbb{1} = 0$, where $\mathbb{1}$ is the all ones vector. When the graph is degree-regular ($\delta_i = \delta$ for $i \in V$), the standard random walk with self loops on G given by the matrix $P := I - \frac{\delta}{\delta+1}\mathcal{L}$ is doubly stochastic and valid for our theory. For non-regular graphs, we make a minor modification in order to obtain a doubly stochastic matrix: let $\delta_{\max} = \max_{i \in V} \delta_i$ denote G's maximum degree and define

$$P_n(G) = I - \frac{1}{\delta_{\max}+1}(D-A) = I - \frac{1}{\delta_{\max}+1}D^{1/2}\mathcal{L}D^{1/2}. \tag{9}$$

This matrix is symmetric by construction and it is also doubly stochastic. Note that if the graph is δ-regular, then $P_n(G)$ is the standard choice above. Plugging $P_n(G)$ into Theorem 2 immediately relates the convergence of distributed dual averaging to the spectral properties of the graph Laplacian; in particular, we have

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left(\frac{RL}{\sqrt{T}} \frac{\log(Tn)}{\sqrt{\lambda_{n-1}(\mathcal{L}(G))}}\right), \tag{10}$$

where $\lambda_{n-1}(\mathcal{L}(G))$ is the second smallest eigenvalue of $\mathcal{L}(G)$. The next result summarizes our conclusions for the choice of stochastic matrix (9) via (10) for different network topologies.

Corollary 1. *Under the conditions of Theorem 2, we have the following convergence rates:*
(a) For k-connected paths and cycles,

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left(\frac{RL}{\sqrt{T}} \frac{n\log(Tn)}{k}\right).$$

(b) For k-connected $\sqrt{n} \times \sqrt{n}$ grids,

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left(\frac{RL}{\sqrt{T}} \frac{\sqrt{n}\log(Tn)}{k}\right).$$

(c) For random geometric graphs with connectivity radius $r = \Omega(\sqrt{\log^{1+\epsilon} n/n})$ for any $\epsilon > 0$, with high-probability

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left(\frac{RL}{\sqrt{T}}\sqrt{\frac{n}{\log n}}\log(Tn)\right).$$

(d) For expanders with bounded ratio of minimum to maximum node degree,

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left(\frac{RL}{\sqrt{T}}\log(Tn)\right).$$

Note that up to logarithmic factors, the optimization term in the convergence rate is always of the order $RL/\sqrt{T}$, while the remaining terms vary depending on the network. In order to understand scaling issues as a function of network size and topology, it can be useful to re-state convergence rates in terms of the number of iterations $T_G(\epsilon; n)$ required to achieve error ϵ for a network type G with n nodes. In particular, Corollary 1 implies the following scalings (to logarithmic factors):
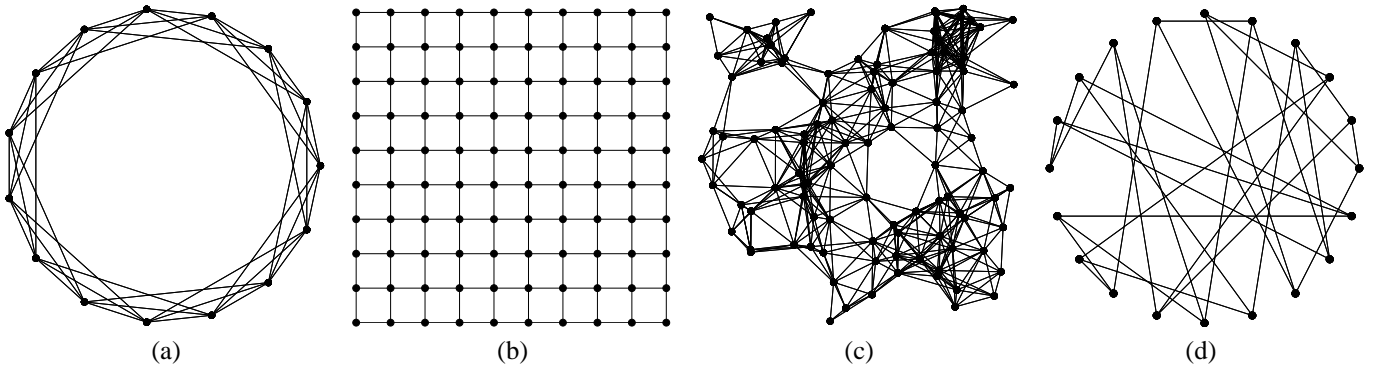
Fig. 1. Illustration of some graph classes of interest in distributed protocols. (a) A 3-connected cycle. (b) Two-dimensional grid with 4-connectivity, and non-toroidal boundary conditions. (c) A random geometric graph. (d) A random 3-regular expander graph.

- for the single cycle graph, $T_{\text{cycle}}(\epsilon; n) = \mathcal{O}(n^2/\epsilon^2)$,
- for the two-dimensional grid, $T_{\text{grid}}(\epsilon; n) = \mathcal{O}(n/\epsilon^2)$,
- for a bounded degree expander, $T_{\text{exp}}(\epsilon; n) = \mathcal{O}(1/\epsilon^2)$.

In general, Theorem 2 implies that at most

$$T_G(\epsilon; n) = \mathcal{O}\left(\frac{1}{\epsilon^2} \frac{1}{1 - \sigma_2(P_n(G))}\right) \qquad (11)$$

iterations are required to achieve an ϵ-accurate solution when using communication matrix $P_n(G)$. A detailed comparison of these results with the previous work is provided in Section IV.

It is interesting to ask whether the upper bound (11) from our analysis is actually a sharp result, meaning that it cannot be improved (up to constant factors). On one hand, it is known that (even for centralized optimization algorithms), any subgradient method requires at least $\Omega\left(\frac{1}{\epsilon^2}\right)$ iterations to achieve ϵ-accuracy [24], so the $1/\epsilon^2$ term is unavoidable. The next proposition addresses the complementary issue, namely whether the inverse spectral gap term is unavoidable for the dual averaging algorithm, by establishing a lower bound on the number of iterations in terms of graph topology:

Proposition 1. *Consider the dual averaging algorithm (5a) and (5b) with quadratic proximal function and communication matrix $P_n(G)$. For any graph G with n nodes, the number of iterations $T_G(c; n)$ required to achieve a fixed accuracy $c > 0$ is lower bounded as*

$$T_G(c; n) = \Omega\left(\frac{1}{1 - \sigma_2(P_n(G))}\right).$$

The proof of this result, given in Section VI-C, involves constructing a "hard" optimization problem and lower bounding the number of iterations required for our algorithm to solve it. In conjunction with Corollary 1, and the bound (11), Proposition 1 implies that our predicted network scaling is sharp. Indeed, in Section IX, we show that the theoretical scalings from Corollary 1—namely, quadratic, linear, and constant in network size n—are well-matched in simulations.

C. Extensions to stochastic communication links

Our results also extend to the case when the communication matrix P is time-varying and random—that is, the matrix $P(t)$ is potentially different for each t and randomly chosen (but $P(t)$ still obeys the constraints imposed by G). Such stochastic

communication is of interest for a variety of reasons. If there is an underlying dense network topology, we might want to avoid communicating along every edge at each round to decrease communication and network congestion. For instance, the use of a gossip protocol [25], in which one edge in the network is randomly chosen to communicate at each iteration, allows a more refined trade-off between communication cost and number of iterations. Communication in real networks also incurs errors due to congestion or hardware failures, and we can model such errors by a stochastic process.

The following theorem provides a convergence result for the case of time-varying random communication matrices. In particular, it applies to sequences $\{x_i(t)\}_{t=0}^{\infty}$ and $\{z_i(t)\}_{t=0}^{\infty}$ generated by the dual averaging algorithm with updates (5a) and (5b) with step size sequence $\{\alpha(t)\}_{t=0}^{\infty}$, but in which p_{ij} is replaced with $p_{ij}(t)$.

Theorem 3 (Stochastic communication)**.** *Let $\{P(t)\}_{t=0}^{\infty}$ be an i.i.d. sequence of doubly stochastic matrices, and define $\lambda_2(G) := \lambda_2(\mathbb{E}[P(t)^\top P(t)])$. For any $x^* \in \mathcal{X}$ and $i \in V$, with probability at least $1 - 1/T$, we have*

$$f(\widehat{x}_i(T)) - f(x^*) \leq \frac{1}{T\alpha(T)}\psi(x^*) + \frac{L^2}{2T}\sum_{t=1}^{T}\alpha(t-1)$$

$$+ \frac{3L^2}{T}\left(\frac{6\log(T^2 n)}{1 - \lambda_2(G)} + \frac{1}{T\sqrt{n}} + 2\right)\sum_{t=1}^{T}\alpha(t).$$

We provide a proof of the theorem in Section VII. Note that the upper bound from the theorem is valid for any sequence of non-increasing positive stepsizes $\{\alpha(t)\}_{t=0}^{\infty}$. The bound consists of three terms, with the first growing and the last two shrinking as the stepsize choice is reduced. If we assume that $\psi(x^*) \leq R^2$, then we can optimize the tradeoff between these competing terms, and we find that the stepsize sequence $\alpha(t) \propto \frac{R\sqrt{1-\lambda_2}}{L\sqrt{t}}$ approximately minimizes the bound bound in the theorem. For some universal constant c, this yields

$$f(\widehat{x}_i(T)) - f(x^*) \leq c\frac{RL}{\sqrt{T}}\frac{\log(Tn)}{\sqrt{1 - \lambda_2(\mathbb{E}[P(t)^\top P(t)])}}. \qquad (12)$$

Stochastic communication for distributed optimization was previously considered by Lobel and Ozdaglar [9]; however, their bounds grew exponentially in the number of nodes n in

the network.[3] In contrast, the rates given here for stochastic communication are directly comparable to the convergence rates in the previous section for fixed transition matrices. More specifically, we have inverse dependence on the spectral gap of the expected network—consequently achieving polynomial scaling for any network—as well as faster rates dependent on network structure.

D. Results for stochastic gradient algorithms

For our last main result, we show that none of our convergence results rely on the gradients being correct. Specifically, we can straightforwardly extend our results to the case of noisy gradients corrupted with zero-mean bounded-variance noise. This setting is especially relevant in situations such as distributed learning or wireless sensor networks, when data observed is noisy. Let $\mathcal{F}_t$ be the σ-field containing all information up to time t, that is, $g_i(1), \ldots, g_i(t) \in \mathcal{F}_t$ and $x_i(1), \ldots, x_i(t+1) \in \mathcal{F}_t$ for all i. We define a stochastic oracle that provides gradient estimates satisfying

$$\mathbb{E}\left[\widehat{g}_i(t) \mid \mathcal{F}_{t-1}\right] \in \partial f_i(x_i(t)) \text{ and } \mathbb{E}\left[\|\widehat{g}_i(t)\|_*^2 \mid \mathcal{F}_{t-1}\right] \leq L^2. \tag{13}$$

As a special case, this model includes an additive noise oracle that takes an element of the subgradient $\partial f_i(x_i(t))$ and adds to it bounded variance zero-mean noise. Theorem 4 gives our convergence result in the case of stochastic gradients. We give the proof and further discussion in Section VIII, noting that because we adapt dual averaging, the analysis follows quite cleanly from that for the previous three theorems.

Theorem 4 (Stochastic gradient updates). *Let the sequence $\{x_i(t)\}_{t=1}^{\infty}$ be as in Theorem 1, except that at each round of the algorithm agent i receives a vector $\widehat{g}_i(t)$ from an oracle satisfying condition (13). For each $i \in V$ and any $x^* \in \mathcal{X}$,*

$$\mathbb{E}\left[f(\widehat{x}_i(T))\right] - f(x^*) \leq \frac{1}{T\alpha(T)}\psi(x^*) + \frac{8L^2}{T}\sum_{t=1}^{T}\alpha(t-1)$$
$$+ \frac{3L^2}{T}\frac{\log(T\sqrt{n})}{1-\sigma_2(P)}\sum_{t=1}^{T}\alpha(t).$$

If we assume in addition that $\|\widehat{g}_i(t)\|_* \leq L$ and that $\mathcal{X}$ has finite radius $R := \sup_{x\in\mathcal{X}}\|x - x^*\|$, then with probability at least $1 - \delta$,

$$f(\widehat{x}_i(T)) - f(x^*) \leq \frac{1}{T\alpha(T)}\psi(x^*) + \frac{8L^2}{T}\sum_{t=1}^{T}\alpha(t-1)$$
$$+ \frac{3L^2}{T}\frac{\log(T\sqrt{n})}{1-\sigma_2(P)}\sum_{t=1}^{T}\alpha(t) + 8LR\sqrt{\frac{\log\frac{1}{\delta}}{T}}.$$

As with the case of stochastic communication covered by Theorem 3, it should be clear that by choosing the stepsize $\alpha(t) \propto \frac{R\sqrt{1-\sigma_2(P)}}{L\sqrt{t}}$, we have essentially the same optimization error guarantee as the bound (12), but with $\lambda_2(\mathbb{E}[P(t)^2])$ replaced by $\sigma_2(P)$. It is also possible to substantially tighten

[3]More precisely, inspection of the constant C in equation (37) of their paper shows that it is of order $\gamma^{-2(n-1)}$, where γ is the lower bound on non-zero entries of P, so it is at least 4^{n-1}.

the deviation probabilities if we assume that the noise of the subgradient estimates is uncorrelated, which we show in Theorem 4 of the long version of this paper [26]. Specifically, the δ-dependent terms in the second bound of Theorem 4 above are replaced by a term that is $\mathcal{O}(\frac{LR\log\frac{1}{\delta}}{T} + \sqrt{\frac{\log\frac{1}{\delta}}{nT}})$.

IV. RELATED WORK

We now turn to surveying some past work with the aim of giving a clear understanding of how our algorithm and results relate to and, in many cases, improve upon it. We first note that the classical problem of consensus averaging [6], [12], [25] is a special case of the problem (1) when $f_i(x) = \|x - \theta_i\|_2^2$. Allowing stochastic gradients also lets us tackle distributed averaging with noise [4]. Mosk-Aoyama et al. [27] consider a problem related to our setup, minimizing $\sum_{i=1}^{n} f_i(x_i)$ for $x_i \in \mathbb{R}$ subject to linear equality constraints, and they obtain rates of convergence dependent on network-conductance.

As discussed in the introduction, other researchers have designed algorithms for solving the problem (1). Earlier works considering our setup include the papers [9], [8]; however, the convergence rates there grow exponentially in the number of nodes n in the network. Nedić et al. [12] and Ram et al. [10] substantially sharpen these earlier results; specifically, Corollary 5.5 in the paper [10] shows that their projected subgradient algorithm can obtain an ϵ-optimal solution to the optimization problem in $\mathcal{O}(n^3/\epsilon^2)$ time with optimal choice of stepsize. All the above papers study convergence of a gradient method in which each node i maintains $x_i(t) \in \mathcal{X}$, and at time t performs the update

$$x_i(t+1) = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\left\{\frac{1}{2}\Big\|\sum_{j\in N(i)}P_{ji}x_j(t) - \alpha g_i(t)\Big\|_2^2\right\} \tag{14}$$

for $g_i(t) \in \partial f_i(x_i(t))$. The distributed dual averaging algorithm (5a)–(5b) is quite different from the update (14). The use of the proximal function ψ allows us to address problems with non-Euclidean geometry, which is useful, for example, for very high-dimensional problems or where the domain $\mathcal{X}$ is the simplex [24, Chapter 3]. The differences between the algorithms become more pronounced in the analysis. Since we use dual averaging, we can avoid some technical difficulties introduced by the projection step in the update (14); precisely because of this technical issue, earlier works [8], [9] studied unconstrained optimization, and the averaging in $z_i(t)$ seems essential to the faster rates our approach achieves.

In other related work, Johansson et al. [11] establish network-dependent rates for Markov incremental gradient descent (MIGD), which maintains a single vector $x(t)$ at all times. A token $i(t)$ determines an active node at time t, and at time step $t+1$ the token moves to one of its neighbors $j \in N(i(t))$, each with probability $P_{ji(t)}$. Letting $g_{i(t)}(t) \in \partial f_{i(t)}(x(t))$, the update is

$$x(t+1) = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\left\{\frac{1}{2}\|x - \alpha g_{i(t)}(t)\|_2^2\right\}. \tag{15}$$

Johansson et al. show that with optimal setting of α and symmetric transition matrix P, MIGD has convergence rate $\mathcal{O}(\max_i \sqrt{\frac{n\Gamma_{ii}}{T}})$, where Γ is the return time matrix

$\Gamma = (I - P + \mathbb{1}\mathbb{1}^\top/n)^{-1}$. In this case, let $\lambda_i(P) \in [-1, 1]$ denote the ith eigenvalue of P. The eigenvalues of Γ are thus 1 and $1/(1 - \lambda_i(P))$ for $i > 1$, and we thus have

$$n \max_{i=1,\dots,n} \Gamma_{ii} \geq \operatorname{tr}(\Gamma) = 1 + \sum_{i=2}^{n} \frac{1}{1 - \lambda_i(P)}$$
$$> \max \left\{ \frac{1}{1 - \lambda_2(P)}, \frac{1}{1 - \lambda_n(P)} \right\} = \frac{1}{1 - \sigma_2(P)}.$$

Consequently, the bound in Theorem 2 is never weaker, and for certain graphs, our results are substantially tighter, as shown in Corollary 1. For d-dimensional grids ($d \geq 2$) we have $T(\epsilon; n) = \mathcal{O}(n^{2/d}/\epsilon^2)$, whereas MIGD scales as $T(\epsilon; n) = \mathcal{O}(n/\epsilon^2)$. For well-connected graphs, such as expanders and the complete graph, the MIGD algorithm scales as $T(\epsilon; n) = \mathcal{O}(n/\epsilon^2)$, a factor of n worse than our results.

V. BASIC CONVERGENCE ANALYSIS FOR DISTRIBUTED DUAL AVERAGING

In this section, we prove convergence of the distributed algorithm based on the updates (5a)–(5b). We begin in Section V-A by defining auxiliary quantities and establishing lemmas useful in the proof then prove Theorem 1 in Section V-B.

A. Setting up the analysis

Using techniques related to those used in past work [8], we establish convergence via the two auxiliary sequences

$$\bar{z}(t) := \frac{1}{n} \sum_{i=1}^{n} z_i(t) \quad \text{and} \quad y(t) := \Pi_{\mathcal{X}}^{\psi}(\bar{z}(t), \alpha). \quad (16)$$

We begin by showing that the sequence $\bar{z}(t)$ evolves in a very simple way. We have

$$\bar{z}(t+1) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(p_{ij}(z_j(t) - \bar{z}(t)) \right) + \bar{z}(t)$$
$$+ \frac{1}{n} \sum_{j=1}^{n} g_j(t) = \bar{z}(t) + \frac{1}{n} \sum_{j=1}^{n} g_j(t), \quad (17)$$

where the second equality follows from double-stochasticity of P. Consequently, the averaged dual sequence $\{\bar{z}(t)\}_{t=0}^{\infty}$ evolves almost as it would for centralized dual averaging applied to $f(x) = \sum_{i=1}^{n} f_i(x)/n$, the difference being that $g_i(t)$ is a subgradient at $x_i(t)$ (which need not be the same as the subgradient $g_j(t)$ at $x_j(t)$). The simple evolution (17) of the averaged dual sequence alleviates difficulties with the non-linearity of projection that have been previously challenging.

Before proceeding with the proof of Theorem 1, we state a few useful results regarding the convergence of the standard dual averaging algorithm. We begin with a result about Lipschitz continuity of the projection mapping (4), recalling that $\|\cdot\|_*$ is dual norm to $\|\cdot\|$.

Lemma 2. For an arbitrary pair $u, v \in \mathbb{R}^d$, we have $\left\| \Pi_{\mathcal{X}}^{\psi}(u, \alpha) - \Pi_{\mathcal{X}}^{\psi}(v, \alpha) \right\| \leq \alpha \|u - v\|_*$.

This result is standard in convex analysis (e.g. [15, Theorem X.4.2.1], or [13, Lemma 1]). We next state the convergence guarantee for the standard dual averaging algorithm. Let

$\{g(t)\}_{t=1}^{\infty} \subset \mathbb{R}^d$ be an arbitrary sequence of vectors, and consider the sequence $\{x(t)\}_{t=1}^{\infty}$:

$$x(t+1) := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \sum_{s=1}^{t} \langle g(s), x \rangle + \frac{1}{\alpha} \psi(x) \right\}$$
$$= \Pi_{\mathcal{X}}^{\psi} \left(\sum_{s=1}^{t} g(s), \alpha \right). \quad (18)$$

Lemma 3. For a non-increasing sequence $\{\alpha(t)\}_{t=0}^{\infty}$ of positive stepsizes, and for any $x^* \in \mathcal{X}$,

$$\sum_{t=1}^{T} \langle g(t), x(t) - x^* \rangle \leq \frac{1}{2} \sum_{t=1}^{T} \alpha(t-1) \|g(t)\|_*^2 + \frac{1}{\alpha(T)} \psi(x^*).$$

The lemma is a consequence of Theorem 2 and Eq. (3.3) in Nesterov [13]. We include a simple proof in Appendix A of the long version of this paper [26]. Finally, we state a lemma that allows us to restrict our analysis to the centralized sequence $\{y(t)\}_{t=0}^{\infty}$ from (16).

Lemma 4. Consider the sequences $\{x_i(t)\}_{i=1}^{\infty}$, $\{z_i(t)\}_{t=0}^{\infty}$, and $\{y(t)\}_{t=0}^{\infty}$ defined according to the updates (5a), (5b), and (16), where each f_i is L-Lipschitz. For each $i \in V$,

$$\sum_{t=1}^{T} f(x_i(t)) - f(x^*)$$
$$\leq \sum_{t=1}^{T} f(y(t)) - f(x^*) + L \sum_{t=1}^{T} \alpha(t) \|\bar{z}(t) - z_i(t)\|_*.$$

Similarly, with the definitions $\widehat{y}(T) := \frac{1}{T} \sum_{t=1}^{T} y(t)$ and $\widehat{x}_i(T) := \frac{1}{T} \sum_{t=1}^{T} x_i(t)$, we have

$$f(\widehat{x}_i(T)) - f(x^*)$$
$$\leq f(\widehat{y}(T)) - f(x^*) + \frac{L}{T} \sum_{t=1}^{T} \alpha(t) \|\bar{z}(t) - z_i(t)\|_*.$$

Proof: Using the L-Lipschitz continuity of the f_i, we note

$$f(x_i(t)) - f(x^*) = f(y(t)) - f(x^*) + f(x_i(t)) - f(y(t))$$
$$\leq f(y(t)) - f(x^*) + L \|x_i(t) - y(t)\|,$$

and we then use Lemma 2, which gives $\|x_i(t) - y(t)\| \leq \alpha(t) \|\bar{z}(t) - z_i(t)\|_*$. The second statement follows analogously after using the triangle inequality. ∎

B. Proof of Theorem 1

Our proof is based on analyzing the sequence $\{y(t)\}_{t=0}^{\infty}$. Given an arbitrary $x^* \in \mathcal{X}$, we have

$$n \sum_{t=1}^{T} f(y(t)) - f(x^*)$$
$$= \sum_{t=1}^{T} \sum_{i=1}^{n} f_i(x_i(t)) - f(x^*) + \sum_{t=1}^{T} \sum_{i=1}^{n} [f_i(y(t)) - f_i(x_i(t))]$$
$$\leq \sum_{t=1}^{T} \sum_{i=1}^{n} [f_i(x_i(t)) - f(x^*) + L \|y(t) - x_i(t)\|], \quad (19)$$

the inequality following by the L-Lipschitz condition on f_i.

Now let $g_i(t) \in \partial f_i(x_i(t))$ be a subgradient of f_i at $x_i(t)$. Using convexity, we have the bound

$$\sum_{i=1}^{n} f_i(x_i(t)) - f_i(x^*) \leq \sum_{i=1}^{n} \langle g_i(t), x_i(t) - x^* \rangle. \quad (20)$$

Breaking the right hand side of (20) into two pieces, we obtain

$$\sum_{i=1}^{n} \langle g_i(t), x_i(t) - x^* \rangle$$
$$= \sum_{i=1}^{n} \langle g_i(t), y(t) - x^* \rangle + \sum_{i=1}^{n} \langle g_i(t), x_i(t) - y(t) \rangle. \quad (21)$$

By definition of the updates for $\bar{z}(t)$ and $y(t)$, we have $y(t) = \Pi_{\mathcal{X}}^{\psi}(\frac{1}{n} \sum_{s=1}^{t-1} \sum_{i=1}^{n} g_i(s), \alpha)$. Thus, we see that the first term in the decomposition (21) can be written in the same way as the bound in Lemma 3, and as a consequence, we have

$$\frac{1}{n} \sum_{t=1}^{T} \left\langle \sum_{i=1}^{n} g_i(t), y(t) - x^* \right\rangle$$
$$\leq \frac{1}{2} \sum_{t=1}^{T} \alpha(t-1) \left\| \frac{1}{n} \sum_{i=1}^{n} g_i(t) \right\|_*^2 + \frac{1}{\alpha(T)} \psi(x^*)$$
$$\leq \frac{L^2}{2} \sum_{t=1}^{T} \alpha(t-1) + \frac{1}{\alpha(T)} \psi(x^*). \quad (22)$$

It remains to control the final two terms in the bounds (19) and (21). Since $\|g_i(t)\|_* \leq L$,

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \frac{L}{n} \|y(t) - x_i(t)\| + \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \langle g_i(t), x_i(t) - y(t) \rangle$$
$$\leq \frac{2L}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \|y(t) - x_i(t)\|.$$

By definition of $y(t)$ and $x_i(t)$ as projections of $\bar{z}(t)$ and $z_i(t)$, respectively, the α-Lipschitz continuity of the projection operator $\Pi_{\mathcal{X}}^{\psi}(\cdot, \alpha)$ (see Lemma 2) implies

$$\frac{2L}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \|y(t) - x_i(t)\| \leq \frac{2L}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \alpha(t) \|\bar{z}(t) - z_i(t)\|_*.$$

Combining this bound with (19) and (22) yields the running sum bound

$$\sum_{t=1}^{T} [f(y(t)) - f(x^*)] \leq \frac{1}{\alpha(T)} \psi(x^*) + \frac{L^2}{2} \sum_{t=1}^{T} \alpha(t-1)$$
$$+ \frac{2L}{n} \sum_{t=1}^{T} \sum_{j=1}^{n} \alpha(t) \|\bar{z}(t) - z_j(t)\|_*. \quad (23)$$

Applying Lemma 4 to (23) and recalling the definition (8) of OPT gives that $\sum_{t=1}^{T} [f(x_i(t)) - f(x^*)]$ is upper bounded by

$$\text{OPT} + L \sum_{t=1}^{T} \alpha(t) \left[\frac{2}{n} \sum_{j=1}^{n} \|\bar{z}(t) - z_j(t)\|_* + \|\bar{z}(t) - z_i(t)\|_* \right].$$

Dividing both sides by T and using the convexity of f yields our desired result (7).

VI. CONVERGENCE RATES, SPECTRAL GAP, AND NETWORK TOPOLOGY

In this section, we give concrete convergence rates for the distributed dual averaging algorithm based on the mixing time of a random walk according to the matrix P. The understanding of the dependence of our convergence rates in terms of network topology is crucial, because it can provide important cues to the system administrator in a clustered computing environment or for the locations and connectivities of sensors in a sensor network. We begin in Section VI-A with the proof of Theorem 2, which we follow in Sections VI-B and VI-C with proofs of the graph-specific convergence rates stated in Corollary 1 and the lower bound of Proposition 1, respectively.

Throughout this section, we adopt the following notational conventions. For an $n \times n$ matrix B, we call its singular values $\sigma_1(B) \geq \sigma_2(B) \geq \cdots \geq \sigma_n(B) \geq 0$. For a real symmetric matrix B, we use $\lambda_1(B) \geq \lambda_2(B) \geq \ldots \geq \lambda_n(B)$ to denote the n real eigenvalues of B. We let $\Delta_n = \{x \in \mathbb{R}^n \mid x \succeq 0, \sum_{i=1}^{n} x_i = 1\}$ denote the n-dimensional probability simplex. Let $\mathbb{1}$ denote the vector of all ones. For stochastic P, we have the following inequality: for any positive integer $t = 1, 2, \ldots$ and $x \in \Delta_n$,

$$\|P^t x - \mathbb{1}/n\|_1 \leq \sqrt{n} \|P^t x - \mathbb{1}/n\|_2 \leq \sigma_2(P)^t \sqrt{n}. \quad (24)$$

See the book [17] for a review of relevant Perron-Frobenius theory.

A. Proof of Theorem 2

We focus on controlling the network error terms in the bound (7), $\sum_{i=1}^{n} \alpha(t) \|\bar{z}(t) - z_i(t)\|_*$. Define the matrix $\Phi(t, s) = P^{t-s+1}$ and $\overline{\Phi}(t, s) = \frac{\mathbb{1}\mathbb{1}^T}{n} - \Phi(t, s)$. Let $[\Phi(t, s)]_{ji}$ be the jth entry of the ith column of $\Phi(t, s)$. Then

$$z_i(t+1) = \sum_{j=1}^{n} [\Phi(t, s)]_{ji} z_j(s)$$
$$+ \sum_{r=s+1}^{t} \left(\sum_{j=1}^{n} [\Phi(t, r)]_{ji} g_j(r-1) \right) + g_i(t). \quad (25)$$

The above clearly reduces to the standard update (5a) when $s = t$. Since $\bar{z}(t)$ evolves simply as in (17), we assume that $z_i(0) = 0$ to avoid notational clutter and use (25) to see

$$\bar{z}(t) - z_i(t) = \sum_{s=1}^{t-1} \sum_{j=1}^{n} (1/n - [\Phi(t-1, s)]_{ji}) g_j(s-1)$$
$$+ \left(\frac{1}{n} \sum_{j=1}^{n} (g_j(t-1) - g_i(t-1)) \right). \quad (26)$$

We have $\|g_i(t)\|_* \leq L$ for all i and t, so the equality (26) and definition of $\overline{\Phi}$ imply

$$\|\bar{z}(t) - z_i(t)\|_* \leq \left\| \sum_{s=1}^{t-1} \sum_{j=1}^{n} [\overline{\Phi}(t-1, s)]_{ji} g_j(s-1) \right\|_*$$
$$+ \left\| \frac{1}{n} \sum_{j=1}^{n} g_j(t-1) - g_i(t-1) \right\|_*.$$

This in turn is further bounded by

$$\sum_{s=1}^{t-1}\sum_{j=1}^{n}\|g_j(s-1)\|_* \,|[\overline{\Phi}(t-1,s)]_{ji}|$$

$$+\frac{1}{n}\sum_{i=1}^{n}\|g_j(t-1)-g_i(t-1)\|_*$$

$$\leq \sum_{s=1}^{t-1} L \,\|[\Phi(t-1,s)]_i - \mathbb{1}/n\|_1 + 2L. \tag{27}$$

Now we break the sum in (27) into two terms separated by a cutoff point $\widehat{t}$. The first term consists of "throwaway" terms, that is, timesteps s for which the Markov chain with transition matrix P has not mixed, while the second consists of steps s for which $\|[\Phi(t-1,s)]_i - \mathbb{1}/n\|_1$ is small. Note that the indexing on $\Phi(t-1,s) = P^{t-s+1}$ implies that for small s, $\Phi(t-1,s)$ is close to uniform. From (24), $\|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \leq \sqrt{n}\sigma_2(P)^{t-s+1}$. Hence, if

$$t-s \geq \frac{\log \epsilon^{-1}}{\log \sigma_2(P)^{-1}} - 1, \quad \|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \leq \sqrt{n}\epsilon.$$

By setting $\epsilon^{-1} = T\sqrt{n}$, for $t - s + 1 \geq \frac{\log(T\sqrt{n})}{\log \sigma_2(P)^{-1}}$, we have

$$\|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \leq \frac{1}{T}. \tag{28}$$

For larger s, we simply have $\|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \leq 2$. The above suggests that we split the sum at $\widehat{t} = \frac{\log T\sqrt{n}}{\log \sigma_2(P)^{-1}}$. We break apart the sum (27) and use (28) to see that, since $t-1-(t-\widehat{t}) = \widehat{t}$ and there are at most T steps in the summation,

$$\|\bar{z}(t) - z_i(t)\|_* \leq L\sum_{s=t-\widehat{t}}^{t-1}\|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1$$

$$+ L\sum_{s=1}^{t-1-\widehat{t}}\|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1 + 2L$$

$$\leq 2L\frac{\log(T\sqrt{n})}{\log \sigma_2(P)^{-1}} + 3L \leq 2L\frac{\log(T\sqrt{n})}{1-\sigma_2(P)} + 3L. \tag{29}$$

The last inequality follows from the concavity of $\log(\cdot)$, since $\log \sigma_2(P)^{-1} \geq 1 - \sigma_2(P)$.

Combining (29) with the running sum bound (23) in the proof of Theorem 1, we immediately see that for $x^* \in \mathcal{X}$,

$$\sum_{t=1}^{T} f(y(t)) - f(x^*) \leq \frac{1}{\alpha(T)}\psi(x^*) + \frac{L^2}{2}\sum_{t=1}^{T}\alpha(t-1)$$

$$+ 6L^2\sum_{t=1}^{T}\alpha(t) + 4L^2\frac{\log(T\sqrt{n})}{1-\sigma_2(P)}\sum_{t=1}^{T}\alpha(t). \tag{30}$$

Appealing to Lemma 4 allows us to obtain the same result on the sequence $x_i(t)$ with slightly worse constants. Note that $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T} - 1$. Thus, using the assumption that $\psi(x^*) \leq R^2$, using convexity to bound $f(\widehat{y}(T)) \leq \frac{1}{T}\sum_{t=1}^{T} f(y(t))$ (and similarly for $\widehat{x}_i(T)$), and setting $\alpha(t)$ as in the statement of the theorem completes the proof.

B. Proof of Corollary 1

The corollary is based on bounding the spectral gap of the matrix $P_n(G)$ from Eq. (9). We begin with a technical lemma.

Lemma 5. *Let* $\overline{\delta} = \delta_{\max}$. *The matrix P satisfies*

$$\sigma_2(P_n(G)) \leq \max\left\{1 - \frac{\min_i \delta_i}{\overline{\delta}+1}\lambda_{n-1}(\mathcal{L}), \frac{\overline{\delta}}{\overline{\delta}+1}\lambda_1(\mathcal{L}) - 1\right\}.$$

Proof: By a theorem of Ostrowski on congruent matrices (cf. Theorem 4.5.9, [17]), we have

$$\lambda_k(D^{1/2}\mathcal{L}D^{1/2}) \in \left[\min_i \delta_i \lambda_k(\mathcal{L}), \max_i \delta_i \lambda_k(\mathcal{L})\right]. \tag{31}$$

Since $\mathcal{L}D^{1/2}\mathbb{1} = 0$, we have $\lambda_n(\mathcal{L}) = 0$, and so it suffices to focus on $\lambda_1(D^{1/2}\mathcal{L}D^{1/2})$ and $\lambda_{n-1}(D^{1/2}\mathcal{L}D^{1/2})$. From the definition (9), the eigenvalues of P are of the form $1 - (\delta_{\max} + 1)^{-1}\lambda_k(D^{1/2}\mathcal{L}D^{1/2})$. The bound (31) coupled with the fact that all the eigenvalues of $\mathcal{L}$ are non-negative implies that $\sigma_2(P) = \max_{k<n}\left\{\left|1 - (\delta_{\max} + 1)^{-1}\lambda_k(D^{1/2}\mathcal{L}D^{1/2})\right|\right\}$ is upper bounded by the larger of $1 - \frac{\delta_{\min}}{\delta_{\max}+1}\lambda_{n-1}(\mathcal{L})$ and $\frac{\delta_{\max}}{\delta_{\max}+1}\lambda_1(\mathcal{L}) - 1$. ∎

Much of spectral graph theory is devoted to bounding $\lambda_{n-1}(\mathcal{L})$ sufficiently far away from zero, and Lemma 5 allows us to leverage such results. Computing the upper bound in Lemma 5 requires controlling both $\lambda_{n-1}(\mathcal{L})$ and $\lambda_1(\mathcal{L})$. To circumvent this complication, we use the well-known idea of a lazy random walk [18], in which we replace P by $\frac{1}{2}(I+P)$. The resulting symmetric matrix has the same eigenstructure as P. Further, $\frac{1}{2}(I+P)$ is positive semidefinite so

$$\sigma_2\left(\frac{1}{2}(I+P)\right) = \lambda_2\left(\frac{1}{2}(I+P)\right),$$

and hence,

$$\sigma_2\left(\frac{1}{2}(I+P)\right) = \lambda_2\left(I - \frac{1}{2(\delta_{\max}+1)}D^{1/2}\mathcal{L}D^{1/2}\right)$$

$$\leq 1 - \frac{\delta_{\min}}{2(\delta_{\max}+1)}\lambda_{n-1}(\mathcal{L}). \tag{32}$$

Consequently, it is sufficient to bound only $\lambda_{n-1}(\mathcal{L})$, which is more convenient from a technical standpoint. The convergence rate implied by the lazy random walk through Theorem 2 is no worse than twice that of the original walk, which is insignificant for the analysis in this paper.

We are now equipped to address each of the graph classes covered by Corollary 1.

Cycles and paths: Recall the regular k-connected cycle from Figure 1(a), constructed by placing n nodes on a circle and connecting every node to k neighbors on the right and left. For this graph, the Laplacian $\mathcal{L}$ is a circulant matrix with diagonal entries 1 and off-diagonal non-zero entries $-1/2k$. Known results on circulant matrices (see [28, Chapter 3] or [25, Section VI.A]) imply that $\lambda_{n-1}(\mathcal{L}) = 1 - \cos\left(\frac{2\pi k}{n}\right) + \Theta\left(\frac{k^4}{n^4}\right)$. A Taylor expansion of $\cos(\cdot)$ gives that $\lambda_{n-1}(\mathcal{L}) = \Theta\left(\frac{k^2}{n^2}\right)$.

Now consider the regular k-connected path, a path in which each node is connected to the k neighbors on its right and left. By computing Cheeger constants [26, Lemma 6], we see that if $k \leq \sqrt{n}$, then $\lambda_{n-1}(\mathcal{L}) = \Theta(k^2/n^2)$. Note also that for the k-connected path on n nodes, $\min_i \delta_i = k$ and $\delta_{\max} = 2k$. Thus,

we can combine the previous two paragraphs with Lemma 5 to see that for k-connected paths or cycles with $k \leq \sqrt{n}$,

$$\sigma_2(P) = 1 - \Theta\left(\frac{k^2}{n^2}\right). \tag{33}$$

Substituting bound (33) into Theorem 2 yields Corollary 1(a).

Regular grids: Now consider the case of a $\sqrt{n}$-by-$\sqrt{n}$ grid, focusing in particular on regular k-connected grids, in which any node is joined to every node that is fewer than k horizontal or vertical edges away in an axis-aligned direction. In this case, we use results on Cartesian products of graphs [21, Section 2.6] to analyze the eigen-structure of the Laplacian. In particular, the $\sqrt{n}$-by-$\sqrt{n}$ k-connected grid is simply the Cartesian product of two regular k-connected paths of $\sqrt{n}$ nodes. The second smallest eigenvalue of a Cartesian product of graphs is half the minimum of second-smallest eigenvalues of the original graphs [21, Theorem 2.13]. Thus, based on the preceding discussion of k-connected paths, we conclude that if $k \leq n^{1/4}$, then we have $\lambda_{n-1}(\mathcal{L}) = \Theta(k^2/n)$, and we use Lemma 5 and (32) to see that

$$\sigma_2(P) = 1 - \Theta\left(\frac{k^2}{n}\right). \tag{34}$$

The result in Corollary 1(b) immediately follows.

Random geometric graphs: Using the proof of Lemma 10 from Boyd et al. [25], we see that for any ϵ and $c > 0$, if $r = \sqrt{\log^{1+\epsilon} n/(n\pi)}$, then with probability at least $1 - 2/n^{c-1}$,

$$\log^{1+\epsilon} n - \sqrt{2}c \log n \leq \delta_i \leq \log^{1+\epsilon} n + \sqrt{2}c \log n \tag{35}$$

for all i. Thus, letting $\mathcal{L}$ be the graph Laplacian of a random geometric graph, if we can bound $\lambda_{n-1}(\mathcal{L})$, (35) coupled with Lemma 5 will control the convergence rate of our algorithm.

Recent work of von Luxburg et al. [29] gives concentration results on the second-smallest eigenvalue of a geometric graph. In particular, their Theorem 3 says that if $r = \omega(\sqrt{\log n/n})$, then with exceedingly high probability, $\lambda_{n-1}(\mathcal{L}) = \Omega(r) = \omega(\log n/n)$. Using (35), we see that for $r = (\log^{1+\epsilon} n/n)^{1/2}$, the ratio $\frac{\min_i \delta_i}{\max_i \delta_i} = \Theta(1)$ and $\lambda_{n-1}(\mathcal{L}) = \Omega\left(\frac{\log^{1+\epsilon} n}{n}\right)$ with high probability. Combining the above equation with Lemma 5 and (32), we have

$$\sigma_2(P) = 1 - \Omega\left(\frac{\log^{1+\epsilon} n}{n}\right).$$

Thus we have obtained the result of Corollary 1(c). Our bounds show that a grid and a random geometric graph exhibit the same convergence rate up to logarithmic factors.

Expanders: The constant spectral gap in expanders [21, Chapter 6] removes any penalty due to network communication (to logarithmic factors), yielding Corollary 1(d).

C. Proof of Proposition 1

Our proof is based on construction of a set of objective functions f_i that force convergence to be slow by using the second eigenvector of the matrix P. Olshevsky and Tsitsiklis [30] independently use similar techniques to prove a lower bound for distributed consensus.

Recall that $\mathbb{1} \in \mathbb{R}^n$ is the eigenvector of P corresponding to its largest eigenvalue, 1. Let $v \in \mathbb{R}^n$ be the eigenvector of P corresponding to its second singular value, $\sigma_2(P)$. By using the lazy random walk defined in Section VI-B, we may assume without loss of generality that $\lambda_2(P) = \sigma_2(P)$. Let $w = \frac{v}{\|v\|_\infty}$ be a normalized version of the second eigenvector of P, and note that $\sum_{i=1}^n w_i = 0$. Without loss of generality, we assume that there is an index i for which $w_i = -1$ (otherwise we can flip signs in what follows); moreover, by re-indexing as needed, we may assume that $w_1 = -1$. We set $\mathcal{X} = [-1, 1] \subset \mathbb{R}$ and define the univariate functions $f_i(x) := (c + w_i)x$, so that the global problem is to minimize $\frac{1}{n} \sum_{i=1}^n f_i(x) = \frac{1}{n} \sum_{i=1}^n (c + w_i)x = cx$ for some constant $c > 0$ to be chosen. Note that each f_i is $c + 1$-Lipschitz. By construction, we see immediately that $x^* = -1$ is optimal for the global problem.

Now consider the evolution of the $\{z(t)\}_{t=0}^\infty \subset \mathbb{R}^n$ as generated by the update (5a). By construction $g_i(t) = c + w_i$ for all t. Defining the vector $g = (c\mathbb{1} + w) \in \mathbb{R}^n$, we have

$$z(t+1) = Pz(t) + g = P^2 z(t-1) + Pg + g = \sum_{\tau=0}^t P^\tau g$$

$$= \sum_{\tau=0}^{t-1} P^\tau(w + c\mathbb{1}) = \sum_{\tau=0}^{t-1} P^\tau w + ct\mathbb{1}$$

$$= \sum_{\tau=0}^{t-1} \sigma_2(P)^\tau w + ct\mathbb{1} \tag{36}$$

since $P\mathbb{1} = \mathbb{1}$. In order to establish a lower bound, it suffices to show at least one node is far from the optimum after t steps, and we focus on node 1. Since $w_1 = -1$, we have by (36)

$$z_1(t+1) = -\sum_{\tau=0}^{t-1} \sigma_2(P)^\tau + ct = -\frac{1 - \sigma_2(P)^{t-1}}{1 - \sigma_2(P)} + ct. \tag{37}$$

Recalling that $\psi(x) = \frac{1}{2}x^2$ for this scalar setting, we have

$$x_i(t+1) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ z_i(t+1)x + \frac{1}{2\alpha(t)}x^2 \right\}$$

$$= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ (x + \alpha(t)z_i(t+1))^2 \right\}.$$

Hence $x_1(t)$ is the projection of $-\alpha(t)z_1(t+1)$ onto $[-1, 1]$, and unless $z_1(t) > 0$ we have $f(x_1(t)) - f(-1) \geq c > 0$. If t is overly small, the relation (37) will guarantee that $z_1(t) \leq 0$, so that $x_1(t)$ is far from the optimum. If we choose $c \leq 1/3$, then a simple calculation shows that we require $t = \Omega((1 - \sigma_2(P))^{-1})$ to drive $z_1(t)$ above zero.

VII. CONVERGENCE RATES FOR STOCHASTIC COMMUNICATION

In this section, we develop theory appropriate for stochastic and time-varying communication, which we model by a sequence $\{P(t)\}_{t=0}^\infty$ of random matrices. We begin in Section VII-A with basic convergence results and then prove Theorem 3. Section VII-B contains analysis of gossip algorithms, and we analyze random edge failures in Section VII-C.

A. Basic convergence analysis

Recall that Theorem 1 involves the sum $\frac{2L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\alpha(t)\left\|\bar{z}(t)-z_i(t)\right\|_*$. In Section VI, we showed how to control this sum when communication between agents occurs on a static underlying network structure via a fixed doubly-stochastic matrix P. We now relax this assumption and instead let $P(t)$ vary over time.

1) Markov chain mixing for stochastic communication: We use $P(t) = [p_1(t) \cdots p_n(t)]$ to denote the doubly stochastic matrix at iteration t. The update employed by the algorithm, modulo changes in P, is given by the updates (5a) and (5b):

$$z_i(t+1) = \sum_{j=1}^{n} p_{ij}(t)z_j(t) + g_i(t), \quad x_i(t+1) = \Pi_{\mathcal{X}}^{\psi}(z_i(t+1), \alpha).$$

In this case, our analysis makes use of the modified definition $\Phi(t,s) = P(s)P(s+1)\cdots P(t)$. However, we still have the evolution of $\bar{z}(t+1) = \bar{z}(t) + \frac{1}{n}\sum_{i=1}^{n} g_i(t)$ from equation (17), and moreover, (26) holds essentially unchanged:

$$\bar{z}(t) - z_i(t) = \sum_{s=1}^{t-1}\sum_{j=1}^{n}(1/n - [\Phi(t-1,s)]_{ji})g_j(s-1)$$
$$+ \frac{1}{n}\sum_{j=1}^{n}(g_j(t-1) - g_i(t-1)). \quad (38)$$

To show convergence for the random communication model, we must control the convergence of $\Phi(t,s)$ to the uniform distribution. We first claim that

$$\mathbb{P}\left[\left\|\Phi(t,s)e_i - \mathbb{1}/n\right\|_2 \geq \epsilon\right] \leq \epsilon^{-2}\lambda_2\left(\mathbb{E}[P(t)^\top P(t)]\right)^{t-s+1}, \quad (39)$$

which we establish by modifying a few known results [25].

Let Δ_n denote the n-dimensional probability simplex and $u(0) \in \Delta_n$ be arbitrary. Consider the random sequence $\{u(t)\}_{t=0}^{\infty}$ generated by $u(t+1) = P(t)u(t)$. Let $v(t) := u(t) - \mathbb{1}/n$ correspond to the portion of $u(t)$ orthogonal to the all 1s vector. Calculating the second moment of $v(t+1)$,

$$\mathbb{E}\left[\langle v(t+1), v(t+1)\rangle \mid v(t)\right] = \mathbb{E}\left[v(t)^\top P(t)^\top P(t)v(t) \mid v(t)\right]$$
$$= v(t)^\top \mathbb{E}[P(t)^\top P(t)]v(t)$$
$$\leq \left\|v(t)\right\|_2^2 \lambda_2\left(\mathbb{E}P(t)^\top P(t)\right)$$

since $\langle v(t), \mathbb{1}\rangle = 0$, $v(t)$ is orthogonal to the first eigenvector of $P(t)$, and $P(t)^\top P(t)$ is symmetric and doubly stochastic. Applying Chebyshev's inequality yields

$$\mathbb{P}\left[\frac{\|u(t) - \mathbb{1}/n\|_2}{\|u(0)\|_2} \geq \epsilon\right] \leq \frac{\mathbb{E}\|v(t)\|^2}{\|u(0)\|_2^2 \epsilon^2}$$
$$\leq \epsilon^{-2}\frac{\|v(0)\|_2^2 \lambda_2\left(\mathbb{E}P(t)^\top P(t)\right)^t}{\|u(0)\|_2^2}.$$

Replacing $u(0)$ with e_i and noting that $\|e_i - \mathbb{1}/n\|_2 \leq 1$ yields the claim (39).

2) Proof of Theorem 3: Using the claim (39), we now prove the main theorem of this section, following an argument similar to the proof of Theorem 2. We begin by choosing a (non-random) time index $\hat{t}$ such that for $t - s \geq \hat{t}$, with high probability, $\Phi(t,s)$ is close to the uniform matrix $\mathbb{1}\mathbb{1}^\top/n$. We

then break the summation from 1 to T into two separate terms, separated by the cut-off point $\hat{t}$. Throughout this derivation, we let $\lambda_2 = \lambda_2(\mathbb{E}[P(t)^\top P(t)])$ to ease notation. Using the probabilistic bound (39), note that $t - s \geq \frac{3\log\epsilon^{-1}}{\log\lambda_2^{-1}} - 1$ implies $\mathbb{P}\left[\left\|\Phi(t,s)e_i - \mathbb{1}/n\right\|_2 \geq \epsilon\right] \leq \epsilon$. Consequently, the choice

$$\hat{t} := \frac{3\log(T^2n)}{\log\lambda_2^{-1}} = \frac{6\log T + 3\log n}{\log\lambda_2^{-1}} \leq \frac{6\log T + 3\log n}{1 - \lambda_2},$$

guarantees that if $t - s \geq \hat{t} - 1$, then

$$\mathbb{P}\left[\left\|\Phi(t,s)e_i - \mathbb{1}/n\right\|_2 \geq \frac{1}{T^2n}\right] \leq (T^2n)^2\lambda_2^{\frac{3\log(T^2n)}{-\log\lambda_2}} = \frac{1}{T^2n}. \quad (40)$$

Recalling the definition $\overline{\Phi}(t,s) = \frac{\mathbb{1}\mathbb{1}^\top}{n} - \Phi(t,s)$ and the bound (27), we have

$$\left\|\bar{z}(t) - z_i(t)\right\|_* \leq L\underbrace{\sum_{s=1}^{t-1}\left\|\Phi(t-1,s)e_i - \mathbb{1}/n\right\|_1}_{\mathcal{T}} + 2L$$

Breaking $\mathcal{T}$ into the sum up to $\hat{t}$ and from $\hat{t}$ to $t-1$ gives

$$\mathcal{T} = \sum_{s=t-\hat{t}}^{t-1}\left\|\overline{\Phi}(t-1,s)e_i\right\|_1 + \sum_{s=1}^{t-1-\hat{t}}\left\|\overline{\Phi}(t-1,s)e_i\right\|_1$$
$$\leq 2\frac{3\log(T^2n)}{1 - \lambda_2} + \sqrt{n}\sum_{s=1}^{t-1-\hat{t}}\left\|\Phi(t-1,s)e_i - \mathbb{1}/n\right\|_2,$$

and hence

$$\left\|\bar{z}(t) - z_i(t)\right\|_* \leq 2L\frac{3\log(T^2n)}{1 - \lambda_2}$$
$$+ L\sqrt{n}\underbrace{\sum_{s=1}^{t-1-\hat{t}}\left\|\Phi(t-1,s)e_i - \mathbb{1}/n\right\|_2}_{\mathcal{S}} + 2L. \quad (41)$$

Now for any fixed pair $s' < s$, since the matrices $P(t)$ are doubly stochastic, we have

$$\left\|\Phi(t-1,s')e_i - \mathbb{1}/n\right\|_2$$
$$= \left\|\Phi(s-1,s')\Phi(t-1,s)e_i - \mathbb{1}/n\right\|_2$$
$$\leq \left\|\Phi(s-1,s')\right\|_2\left\|\Phi(t-1,s)e_i - \mathbb{1}/n\right\|_2$$
$$\leq \left\|\Phi(t-1,s)e_i - \mathbb{1}/n\right\|_2,$$

where the final inequality uses the bound $\|\Phi(s-1,s')\|_2 \leq 1$. From the bound (40), we have the bound $\left\|\Phi(t-1,t-\hat{t}-1)e_i - \mathbb{1}/n\right\|_2 \leq \frac{1}{T^2n}$ with probability at least $1 - 1/(T^2n)$. Since s ranges between 1 and $t - \hat{t}$ in the summation $\mathcal{S}$, we conclude that $\mathcal{S} \leq L\sqrt{n}T\frac{1}{T^2n} = \frac{L\sqrt{n}}{Tn}$. Hence, assuming that $n \geq 3$, we have $\left\|\bar{z}(t) - z_i(t)\right\|_* \leq L\frac{6\log(T^2n)}{1-\lambda_2} + L\frac{1}{T\sqrt{n}} + 2L$ with probability at least $1 - 1/(T^2n)$. Applying the union bound over all iterations $t = 1, \ldots, T$ and nodes $i = 1, \ldots, n$,

$$\mathbb{P}\left[\max_{t,i}\left\|\bar{z}(t) - z_i(t)\right\|_* > \frac{6L\log(T^2n)}{1 - \lambda_2} + \frac{L}{T\sqrt{n}} + 2L\right] \leq \frac{1}{T}.$$

Recalling the master result in Theorem 1 completes the proof.

B. Gossip-like protocols

Gossip algorithms are procedures for achieving consensus in a network robustly by randomly selecting one edge (i,j) in the network for communication at each iteration; upon selection, nodes i and j average their values [25]. Gossip algorithms drastically reduce communication in the network but enjoy fast convergence and are robust to changes in topology.

1) Partially asynchronous gossip protocols: In a partially asynchronous iterative method, agents synchronize their iterations [1], which is the model of standard gossip, where computation proceeds in rounds, and in each round communication occurs on one random edge. In our framework, this corresponds to using the random transition matrix $P(t) = I - \frac{1}{2}(e_i - e_j)(e_i - e_j)^\top$. It is clear that $P(t)^\top P(t) = P(t)$, since $P(t)$ is a projection matrix.

Let A be the adjacency matrix of the graph G and D be the diagonal matrix of its degrees. At round t, edge (i,j) (with $A_{ij} = 1$) is chosen with probability $1/\langle \mathbb{1}, A\mathbb{1} \rangle$. Thus,

$$\mathbb{E}P(t) = \frac{1}{\langle \mathbb{1}, A\mathbb{1} \rangle} \sum_{(i,j):A_{ij}=1} I - \frac{1}{2}(e_i - e_j)(e_i - e_j)^\top$$

$$= I - \frac{1}{\langle \mathbb{1}, A\mathbb{1} \rangle}(D - A) = I - \frac{1}{\langle \mathbb{1}, A\mathbb{1} \rangle}D^{1/2}\mathcal{L}D^{1/2} \quad (42)$$

since $\sum_{(i,j):A_{ij}=1}(e_i - e_j)(e_i - e_j)^\top = 2(D - A)$. Using an identical argument as that for Lemma 5, we see that (42) implies that $\lambda_2(\mathbb{E}P(t)) \leq 1 - \frac{\min_i \delta_i}{\langle \mathbb{1}, A\mathbb{1} \rangle}\lambda_{n-1}(\mathcal{L})$. Note that $\langle \mathbb{1}, A\mathbb{1} \rangle = \langle \mathbb{1}, D\mathbb{1} \rangle$, so that for approximately regular graphs, $\langle \mathbb{1}, A\mathbb{1} \rangle \approx n\delta_{\max}$, and $\min_i \delta_i / \langle \mathbb{1}, A\mathbb{1} \rangle \approx 1/n$. Thus, at the expense of a factor of roughly $1/n$ in convergence rate, we can reduce the number of messages sent per round from the number of edges in the graph, $\Theta(n\delta_{\max})$, to one.

2) Totally asynchronous gossip protocol: Now we relax the assumption that agents have synchronized clocks, so the iterations of the algorithm are no longer synchronized. Suppose that each agent has a random clock ticking at real-valued times, and at each clock tick, the agent randomly chooses one of its neighbors to communicate with. Further assume that each agent computes an iterative approximation to $g_i \in \partial f_i(x_i(t))$, and that the approximation is always unbiased (an example of this is when f_i is the sum of several functions, and agent i simply computes the subgradient of each function sequentially). This communication corresponds to a gossip protocol with stochastic subgradients, and its convergence can be described by combining (42) with Theorem 4. This type of algorithm is well-suited to completely decentralized environments, such as sensor networks.

C. Random edge inclusion and failure

The two communication "protocols" we analyze now make selection of each edge at each iteration of the algorithm independent. We begin with random edge inclusions and follow by giving convergence guarantees for random edge failures. For both protocols, computation of $\mathbb{E}P(t)^\top P(t)$ is in general non-trivial, so we work with the model of lazy random walks in Section VI-B. We observe that for any PSD

stochastic matrix P, $P^2 \preceq P$, so [17, Theorem 4.3.1]

$$\lambda_2\left(\mathbb{E}\left(\frac{1}{2}I + \frac{1}{2}P(t)\right)^\top\left(\frac{1}{2}I + \frac{1}{2}P(t)\right)\right)$$

$$\leq \sigma_2\left(\frac{1}{2}I + \frac{1}{2}\mathbb{E}P(t)\right) \leq \frac{1}{2} + \frac{1}{2}\sigma_2(\mathbb{E}P(t)). \quad (43)$$

Thus any bound on $\sigma_2(\mathbb{E}P(t))$ provides an upper bound on the convergence rate of the distributed dual averaging algorithm with random communication, as in Theorem 3.

Consider the communication protocol in which with probability $1 - \delta_i/(\delta_{\max} + 1)$, node i does not communicate, and otherwise the node picks a random neighbor. If a node i picks a neighbor j, then j also communicates back with i to ensure double stochasticity of the transition matrix. We let $A(t)$ be the random adjacency matrix at time t. When there is an edge (i,j) in the underlying graph, the probability that node i picks edge (i,j) is $1/(\delta_{\max} + 1)$, and thus $\mathbb{E}A(t)_{ij} = \frac{2\delta_{\max}+1}{(\delta_{\max}+1)^2}$. The random communication matrix is $P(t) = I - (\delta_{\max} + 1)^{-1}(D(t) - A(t))$. Let A and D be the adjacency matrix and degree matrix of the underlying (non-stochastic) graph and P be communication matrix defined in (9). With these definitions, it is easily shown that

$$\mathbb{E}P(t) = I - (\delta_{\max} + 1)^{-1}(\mathbb{E}D(t) - \mathbb{E}A(t))$$

$$= \left(\frac{\delta_{\max}}{\delta_{\max} + 1}\right)^2 I + \frac{2\delta_{\max} + 1}{(\delta_{\max} + 1)^2}P,$$

and hence $1 - \lambda_2(\mathbb{E}P(t)) = \frac{2\delta_{\max}+1}{(\delta_{\max}+1)^2}(1 - \lambda_2(P))$. Using (43), we see that the spectral gap—and hence our convergence guarantee—decreases by a factor proportional to the maximum degree in the graph. The amount of communication performed decreases by the same factor.

A related model we can analyze is that of a network in which at every time step of the algorithm, an edge fails with probability ρ independently of the other edges. We assume we are using the model of communication in the prequel, so $P(t) = I - (\delta_{\max} + 1)^{-1}D(t) + (\delta_{\max} + 1)^{-1}A(t)$. Let A, D, and P be as before and $\mathcal{L}$ be the Laplacian of the underlying graph; we have

$$\mathbb{E}P(t) = I - \frac{1-\rho}{\delta_{\max} + 1}D - \frac{1-\rho}{\delta_{\max} + 1}A$$

$$= I - \frac{1-\rho}{\delta_{\max} + 1}D^{1/2}\mathcal{L}D^{1/2} = \rho I + (1 - \rho)P.$$

Applying the convergence guarantee (12), we see that we lose at most a factor of $\sqrt{1 - \rho}$ in the convergence rate.

VIII. STOCHASTIC GRADIENT OPTIMIZATION

The algorithm we have presented naturally generalizes to the case in which the agents do not receive true subgradient information but only an unbiased estimate of a subgradient of f_i. That is, we now assume that during round t agent i receives a vector $\widehat{g}_i(t)$ with $\mathbb{E}\widehat{g}_i(t) = g_i(t) \in \partial f_i(x_i(t))$. The proof is made significantly easier by the dual averaging algorithm, which by virtue of the simplicity of its dual update smoothes the propagation of errors from noisy estimates of individual subgradients throughout the network. This was a difficulty in prior work, where significant care was needed to pass noisy gradients through nonlinear projections [10].

A. Proof of Theorem 4

We begin by using convexity and the Lipschitz continuity of the f_i (equations (19) and (20)); hence the running sum $\sum_{t=1}^{T} f(y(t)) - f(x^*)$ is bounded by

$$\sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \langle g_i(t), x_i(t) - x^* \rangle + \sum_{t=1}^{T} \sum_{i=1}^{n} L \| y(t) - x_i(t) \|$$

$$= \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \Big[\langle \widehat{g}_i(t), x_i(t) - x^* \rangle + L \| y(t) - x_i(t) \|$$

$$+ \langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^* \rangle \Big]. \quad (44)$$

We bound the first two terms of (44) using the same derivation as that for Theorem 1. In particular, $\sum_{i=1}^{n} \langle \widehat{g}_i(t), x_i(t) - x^* \rangle = \sum_{i=1}^{n} \langle \widehat{g}_i(t), y(t) - x^* \rangle + \sum_{i=1}^{n} \langle \widehat{g}_i(t), x_i(t) - y(t) \rangle$, and Lemma 3 applies to arbitrary $\widehat{g}_i(t)$. So we bound the first term in (44) with

$$\frac{1}{\alpha(T)} \psi(x^*) + \frac{1}{2} \sum_{t=1}^{T} \alpha(t-1) \left\| \frac{1}{n} \sum_{i=1}^{n} \widehat{g}_i(t) \right\|_*^2$$

$$+ \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \langle \widehat{g}_i(t), x_i(t) - y(t) \rangle. \quad (45)$$

Hölder's inequality implies that $\mathbb{E}[\| \widehat{g}_i(t) \|_* \| \widehat{g}_j(s) \|_*] \leq L^2$ and $\mathbb{E}\| \widehat{g}_i(t) \|_* \leq L$ for any i, j, s, t. We use the two inequalities to bound (45). We have

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \widehat{g}_i(t) \right\|_*^2 \leq \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{E} \left[\| \widehat{g}_i(t) \|_* \| \widehat{g}_j(t) \|_* \right] \leq L^2.$$

Further, $x_i(t) \in \mathcal{F}_{t-1}$ and $y(t) \in \mathcal{F}_{t-1}$ by assumption, so

$$\mathbb{E} \langle \widehat{g}_i(t), x_i(t) - y(t) \rangle \leq \mathbb{E} \| \widehat{g}_i(t) \|_* \| x_i(t) - y(t) \|$$

$$= \mathbb{E} \left(\mathbb{E}[\| \widehat{g}_i(t) \|_* \mid \mathcal{F}_{t-1}] \| x_i(t) - y(t) \| \right)$$

$$\leq L \mathbb{E} \| x_i(t) - y(t) \|.$$

Recalling that $\| x_i(t) - y(t) \| \leq \alpha(t) \| \bar{z}(t) - z_i(t) \|_*$, we proceed by putting expectations around the norm terms in (27) and (29) to see that

$$\frac{1}{\alpha(t)} \mathbb{E} \| y(t) - x_i(t) \| \leq \mathbb{E} \| \bar{z}(t) - z_i(t) \|_*$$

$$\leq \sum_{s=1}^{t-1} L \| [\Phi(t-1, s)]_i - \mathbb{1}/n \|_1 + 2L$$

$$\leq L \frac{\log(T\sqrt{n})}{1 - \sigma_2(P)} + 3L.$$

Let us define $e_i(t) = g_i(t) - \widehat{g}_i(t)$. Then coupled with the above arguments, we can bound the expectation of (44) by

$$\mathbb{E} \left[\sum_{t=1}^{T} f(y(t)) - f(x^*) \right] \leq \frac{1}{\alpha(T)} \psi(x^*) + \frac{L^2}{2} \sum_{t=1}^{T} \alpha(t-1)$$

$$+ \left(2L^2 \frac{\log(T\sqrt{n})}{1 - \sigma_2(P)} + 6L^2 \right) \sum_{t=1}^{T} \alpha(t)$$

$$+ \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E} \left[\langle e_i(t), x_i(t) - x^* \rangle \right]. \quad (46)$$

Taking the expectation for the final term in the bound (46), we recall that $x_i(t) \in \mathcal{F}_{t-1}$, so

$$\mathbb{E} \left[\langle e_i(t), x_i(t) - x^* \rangle \right]$$

$$= \mathbb{E} \left[\langle \mathbb{E}(g_i(t) - \widehat{g}_i(t) \mid \mathcal{F}_{t-1}), x_i(t) - x^* \rangle \right] = 0, \quad (47)$$

which proves the first statement of the theorem.

To show that the statement holds with high probability when $\mathcal{X}$ is compact and $\| \widehat{g}_i(t) \|_* \leq L$, it is sufficient to establish that the sequence $\langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^* \rangle$ is a bounded difference martingale and apply Azuma's inequality [31]. (Under compactness and bounded norm conditions, our previous bounds on terms in the decomposition (45) now hold for the analogous terms in the decomposition (46) without taking expectations.)

By assumption on the compactness of $\mathcal{X}$ and the Lipschitz assumptions on f_i, we have

$$\langle e_i(t), x_i(t) - x^* \rangle \leq \| g_i(t) - \widehat{g}_i(t) \|_* \| x_i(t) - x^* \| \leq 2LR.$$

Recalling (47), we conclude that the last sum in the decomposition (46) is a bounded difference martingale, and Azuma's inequality implies that

$$\mathbb{P} \left[\sum_{t=1}^{T} \sum_{i=1}^{n} \langle e_i(t), x_i(t) - x^* \rangle \geq \epsilon \right] \leq \exp \left(-\frac{\epsilon^2}{16Tn^2L^2R^2} \right).$$

Dividing by T and setting the probability above to δ, we obtain

$$\frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^* \rangle \leq 4LR \sqrt{\frac{\log \frac{1}{\delta}}{T}},$$

with probability at least $1 - \delta$.

The second statement of the theorem is now obtained by appealing to Lemma 4. By convexity, we have $f(\widehat{x}_i(T)) \leq \frac{1}{T} \sum_{t=1}^{T} f(x_i(t))$, thereby completing the proof.

IX. SIMULATIONS

We report experimental results on the convergence behavior of the distributed dual averaging algorithm as a function of the graph structure and number of nodes n as well as giving comparison of distributed dual averaging to the methods in the papers [11], [10] in this section. These results illustrate the excellent agreement of the empirical behavior with our theoretical predictions and improved algorithmic performance.

For all experiments reported here, we consider distributed minimization of a sum of ℓ_1-regression loss functions; these are robust versions of standard linear regression and useful in system identification [32]. In this problem, we are given n pairs of the form $(b_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ and which to estimate the vector that so that $\langle b_i, x \rangle \approx y_i$. That is, we minimize

$$f(x) \quad := \quad \frac{1}{n} \sum_{i=1}^{n} |y_i - \langle b_i, x \rangle| = \frac{1}{n} \| y - Bx \|_1. \quad (48)$$

Setting $L = \max_i \| b_i \|_2$, we note that f is L-Lipschitz and non-smooth at any point with $\langle b_i, x \rangle = y_i$. It is common to impose some type of norm constraint on the solution of (48), so we set $\mathcal{X} = \{ x \in \mathbb{R}^d \mid \| x \|_2 \leq 5 \}$. For a given graph size n, we form a random instance of a regression problem with n data points. In order to study the effect of graph
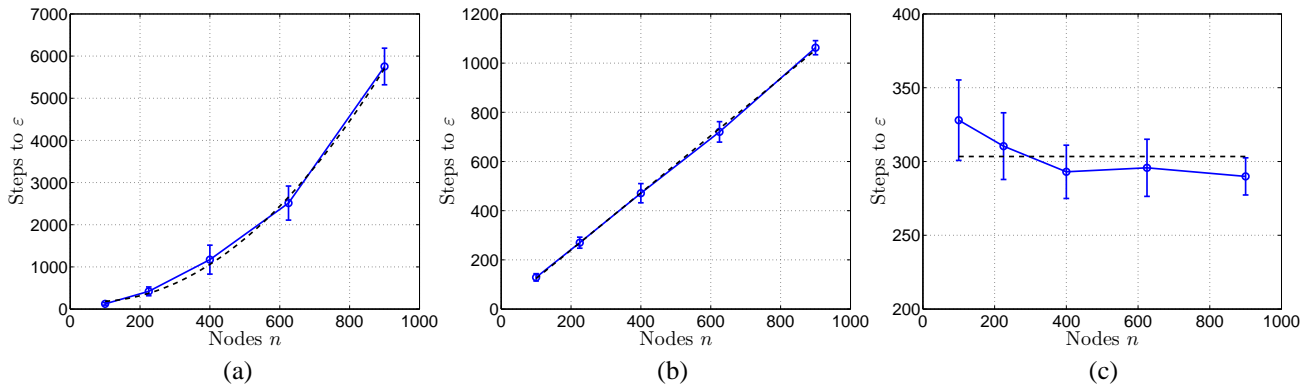
Fig. 3. Each plot shows the number of iterations required to reach a fixed accuracy ϵ (vertical axis) versus the network size n (horizontal axis). Each panel shows the same plot for a different graph topology: (a) single cycle; (b) two-dimensional grid; and (c) bounded degree expander. Step sizes were chosen according to the spectral gap. Dotted lines show predictions of Corollary 1
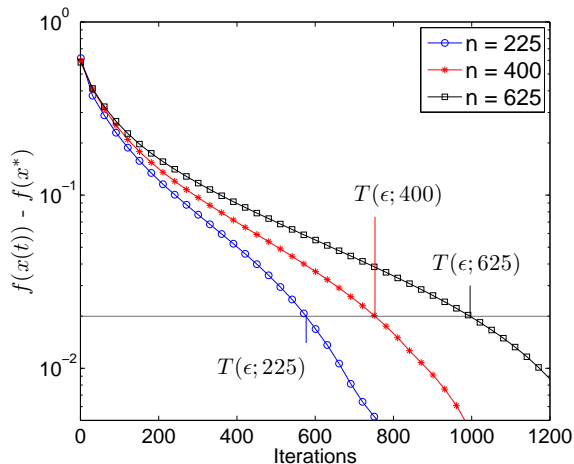
Fig. 2. Plot of the function error versus the number of iterations for a grid graph (see text).

size and topology, we perform simulations with three different graph structures, namely cycles, grids, and random 5-regular expanders [23]. In all cases, we use the setting of the step size α specified in Theorem 2 and Corollary 1.

Figure 2 provides a plot of the function error $\max_i[f(\widehat{x}_i(T) - f(x^*)]$ versus the number of iterations for grid graphs with a varying number of nodes $n \in \{225, 400, 625\}$. In addition to demonstrating convergence, the plots also shows how the convergence time scales as a function of the graph size n. For any fixed $\epsilon > 0$, the function $T_G(\epsilon; n)$ defined in equation (11) shifts to the right as n is increased, and our analysis aims to gain a precise understanding of this shifting.

In Figure 3, we compare our theoretical predictions with the actual behavior of dual subgradient averaging. Each panel shows the function $T_G(\epsilon; n)$ versus the graph size n for the fixed value $\epsilon = 0.1$; the three different panels correspond to different graph types: cycles (a), grids (b) and expanders (c). In each panel, each point on the heavy blue curve is the average of 20 trials, and the bars show standard errors. For comparison, the dotted black line shows the theoretical prediction of Corollary 1. Note the excellent agreement between the empirical behavior and theoretical predictions in all cases. In particular, panel (a) exhibits the quadratic scaling predicted for the cycle, panel (b) exhibits the the linear scaling expected for the grid,

and panel (c) shows that expander graphs have the desirable property of constant network scaling.

Our final set of experiments compares the distributed dual averaging method (DDA) that we present to the Markov incremental gradient descent (MIGD) method [11] and the distributed projected gradient method [10]. In Figure 4, we plot the quantity $T_G(\epsilon; n)$ versus graph size n for DDA and MIGD on grid and expander graphs. We use the optimal stepsize $\alpha(t)$ suggested by the analyses for each method. (We do not plot results for the distributed projected gradient method [10] because the optimal choice of stepsize according to the analysis therein results in such slow convergence that it does not fit on the plots.) Fig. 4 makes it clear that— especially on graphs with good connectivity properties—the dual averaging algorithm gives improved performance.

X. Conclusions and Discussion

In this paper, we proposed and analyzed a distributed dual averaging algorithm for minimizing the sum of local convex functions over a network. It is computationally efficient, and we provided a sharp analysis of its convergence behavior as a function of the properties of the optimization functions and the underlying network topology. Our analysis demonstrates a close connection between convergence rates and mixing times of random walks on the underlying graph; such a connection is natural given the local and graph-constrained nature of our updates. In addition to analysis of deterministic updates, our results also include stochastic communication protocols, for instance when communication occurs only along a random subset of the edges at each round. Such extensions allow for the design of protocols that trade off amount of communication and convergence rate. We also demonstrate that our algorithm is robust to noise by providing an analysis for the case of stochastic optimization with noisy gradients. We confirmed the sharpness of our theoretical predictions by implementation and simulation of our algorithm.

There are several interesting open questions that remain to be explored. For instance, it would be interesting to analyze the convergence properties of other kinds of network-based optimization problems, by combining local information in different structures. It would also be of interest to study what other optimization procedures from the standard setting can be

(a)

(b)
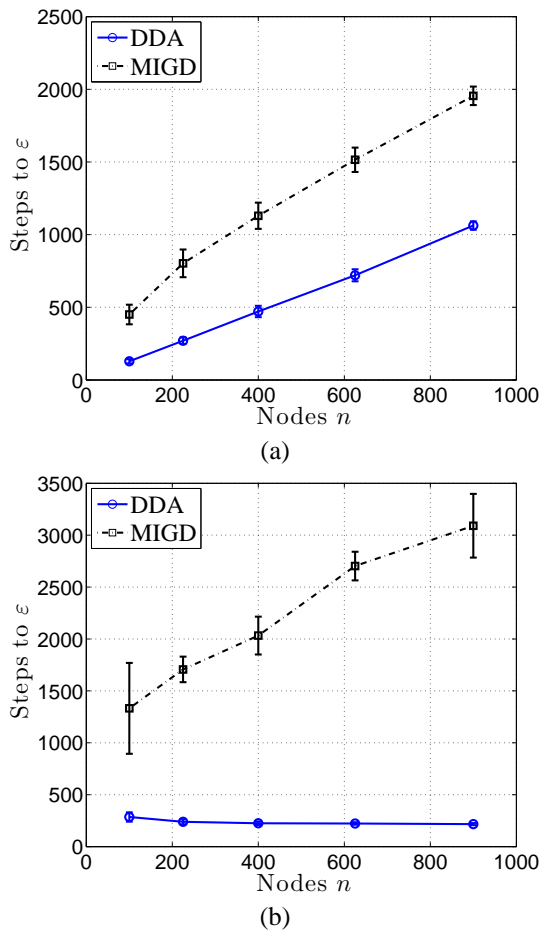
Fig. 4. Number of iterations for distributed dual averaging (DDA) and Markov incremental gradient descent (MIGD) [11] to reach fixed accuracy ϵ versus network size n for (a) two-dimensional grids, (b) bounded degree expanders.

converted into efficient distributed algorithms to better exploit problem structure when possible.

REFERENCES

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., 1989.

[2] V. Lesser, C. Ortiz, and M. Tambe, Eds., *Distributed Sensor Networks: A Multiagent Perspective*. Kluwer Academic Publishers, 2003, vol. 9.

[3] D. Li, K. Wong, Y. Hu, and A. Sayeed, "Detection, classification and tracking of targets in distributed sensor networks," in *IEEE Signal Processing Magazine*, 2002, pp. 17–29.

[4] L. Xiao, S. Boyd, and S. J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, September 1995.

[6] J. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Institute of Technology, 1984.

[7] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, pp. 803–812, 1986.

[8] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multiagent optimization," *IEEE Transactions on Automatic Control*, vol. 54, pp. 48–61, 2009.

[9] I. Lobel and A. Ozdaglar, "Distributed subgradient methods over random networks," MIT LIDS, Tech. Rep. 2800, 2009.

[10] S. S. Ram, A. Nedić, 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, 2010.

[11] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1157–1170, 2009.

[12] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[13] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical Programming A*, vol. 120, no. 1, pp. 261–283, 2009.

[14] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *Journal of Machine Learning Research*, vol. 11, pp. 2543–2596, 2010.

[15] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I & II*. Springer, 1996.

[16] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 291–307, 2005.

[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[18] D. Levin, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[19] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[20] M. Penrose, *Random Geometric Graphs*. Oxford University Press, 2003.

[21] F. Chung, *Spectral Graph Theory*. AMS, 1998.

[22] N. Alon, "Eigenvalues and expanders," *Combinatorica*, vol. 6, pp. 83–96, 1986.

[23] J. Friedman, J. Kahn, and E. Szemerédi, "On the second eigenvalue of random regular graphs," in *Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing*. ACM, 1989, pp. 587–598.

[24] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley, 1983.

[25] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[26] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," 2010. [Online]. Available: http://arxiv.org/abs/1005.2012

[27] D. Mosk-Aoyama, T. Roughgarden, and D. Shah, "Fully distributed algorithms for convex optimization problems," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3260–3279, 2010.

[28] R. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[29] U. von Luxburg, A. Radl, and M. Hein, "Hitting times, commute distances, and the spectral gap for large random geometric graphs," 2010. [Online]. Available: http://arxiv.org/abs/1003.1266

[30] A. Olshevsky and J. N. Tsitsiklis, "A lower bound on distributed averaging," in *49th IEEE Conference on Decision and Control*, 2010.

[31] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 68, pp. 357–367, 1967.

[32] B. T. Polyak and J. Tsypkin, "Robust identification," *Automatica*, vol. 16, pp. 53–63, 1980.

John C. Duchi John C. Duchi received the Bachelor's and Master's degrees in Computer Science from Stanford Univeristy, Stanford, CA in 2007. Since 2008, he has been pursuing a Ph.D. in Computer Science at the University of California, Berkeley. He received the National Defense Science and Engineering Graduate Fellowship in 2009.

Alekh Agarwal Alekh Agarwal received the B. Tech degree in Computer Science and Engineering from Indian Institute of Technology Bombay, Mumbai, India in 2007, and an M. A. in Statistics from University of California Berkeley in 2009 where he is currently pursuing a PhD in Computer Science. He received the Microsoft Research Fellowship in 2009.

Martin J. Wainwright Martin J. Wainwright is currently a professor at University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences. He received a Bachelor's degree in Mathematics from University of Waterloo, Canada, and Ph.D. degree in Electrical Engineering and Computer Science (EECS) from Massachusetts Institute of Technology (MIT). His research interests include coding and information theory, machine learning, mathematical statistics, and statistical signal processing. He has been awarded an Alfred P. Sloan Foundation Fellowship, an NSF CAREER Award, the George M. Sprowls Prize for his dissertation research (EECS department, MIT), a Natural Sciences and Engineering Research Council of Canada 1967 Fellowship, an IEEE Signal Processing Society Best Paper Award in 2008, and several outstanding conference paper awards.