# Scalable Distributed Optimization: Dual Averaging and Spectral Gap-Constrained Convergence

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 9, 2024

**Abstract**

This paper develops a dual averaging subgradient algorithm for distributed convex optimization over networked systems. The proposed algorithm is designed for scalability and robustness in large-scale environments such as sensor networks, multi-agent systems, and distributed machine learning. We provide a convergence analysis that links the algorithm's performance to spectral properties of the underlying communication graph, allowing for a more detailed understanding of how network topology affects the convergence rate. Furthermore, the proposed method handles both deterministic and stochastic communication protocols, ensuring flexibility for varying network conditions. Simulation results validate the theoretical findings, showing optimal performance under various network structures.

**Keywords:** Distributed optimization, Dual averaging, Spectral gap, Network scaling, Convergence analysis

## 1 Introduction

Distributed optimization has emerged as a key methodology for solving large-scale problems where data or computation is inherently distributed across multiple agents. Applications abound in areas such as sensor networks, distributed machine learning, and multi-agent coordination, where the underlying optimization problems are often convex in nature. The primary challenge in such environments is to design algorithms that are both communication-efficient and capable of leveraging the network's topology to achieve fast convergence.

The literature on distributed optimization includes various methods like primal and dual decomposition, consensus averaging, and subgradient methods. However, much of the existing work either assumes idealized network conditions or lacks rigorous analysis regarding the impact of network structure on convergence rates. This paper addresses these limitations by proposing a novel dual averaging subgradient method tailored for distributed settings, focusing on how network scaling and topology influence the convergence behavior.

**Backgrounds:** The focus of this paper is the development and analysis of distributed algorithms for solving convex optimization problems that are defined over networks. Such network-structured optimization problems arise in a variety of application 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 [BT89, LWHS02, LOT03, RN04, XBK07]. Common to these problems is the necessity for completely decentralized computation that is locally light—so as to avoid overburdening small sensors or flooding busy networks—and robust to periodic link or node

failures. As a second example, data sets that are too large to be processed quickly by any single processor present related challenges. A canonical example that arises in statistical machine learning is the problem of minimizing a loss function averaged over a large dataset (e.g., optimization in support vector machines [CV95]). With terabytes of data, it is desirable to assign smaller subsets of the data to different processors, and the processors must communicate to find parameters that minimize 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. Primal and dual decomposition methods that lend themselves naturally to a distributed paradigm have been known for at least fifty years, and their behavior is well understood (e.g., [DW60, Ber99]). The 1980s saw significant interest in distributed detection, consensus, and minimization. The seminal work of Tsitsiklis and colleagues [Tsi84, TBA86, BT89] analyzed algorithms for minimization of a smooth function $f$ known to several agents while distributing processing of components of the parameter vector $x \in \mathbb{R}^n$. An important special case of network optimization—with much faster convergence rates than those known for general distributed optimization—is consensus averaging, where each processor in the network must agree on a single (vector-valued) variable. This is recovered from our objective (1) by setting $f_i(x) = \|x - \theta_i\|_2^2$. A number of researchers have obtained sharp convergence results for distributed consensus algorithms by studying network topology and using spectral properties of random walks or path averaging arguments on the underlying graph structure (e.g., see [BGPS06, BDTV10, DSW08] and references therein). Allowing stochastic gradients also lets us tackle distributed averaging with noise [XBK07]. Mosk-Aoyama et al. [MARS10] 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 using an algorithm similar to dual decomposition. More recently, a few researchers have shifted focus to problems in which each processor locally has its own convex (potentially non-differentiable) objective function [NO09, LO09]. Whereas these initial papers treated the case of unconstrained optimization, more recent work by Ram et al. [RNV10] analyzes a projected subgradient algorithm for distributed minimization of non-smooth functions with constraints.

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 [NO09, LO09, RNV10], 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. Our analysis splits the convergence rate of the algorithm into two terms: an optimization term and a network deviation term. We obtain the optimization penalty using techniques based on the optimization literature, specifically building on results due to Nesterov [Nes09]. This splitting approach can also be adapted to naturally handle issues such as constrained optimization, stochastic communication, and stochastic optimization due to elegant properties of the dual averaging algorithm. On the other hand, the network scaling terms are obtained using techniques from analysis of Markov chains coupled with the distributed communication protocol. We show that the network deviation terms we derive are sharp for our algorithm; in the special case of the consensus problem, these terms are known to be near-optimal [BGPS06].

By comparison to previous work, our convergence results and proofs are different, and our

characterization of the network scaling terms are often much stronger. In particular, the convergence rates given by the papers [NO09, LO09] grow exponentially in the number of nodes $n$ in the network. Nedić et al. [NOOT09] and Ram et al. [RNV10] provide a much tighter analysis that yields convergence rates that scale polynomially in the network size, but are independent of the network topology (apart from requiring strong connectedness and degree independent of $n$). Specifically, Corollary 5.5 in the paper [RNV10] guarantees that their projected subgradient algorithm—under the assumptions that the number of time steps is known a priori and the stepsize is chosen optimally—obtains an $\epsilon$-optimal solution to the optimization problem 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. [JRJ09] analyze a low communication peer-to-peer protocol that attains rates dependent on network structure. However, in their algorithm only one node has a current parameter value, while all nodes in our algorithm maintain good estimates of the optimum at all time. This is important in online, streaming, and control problems where nodes are expected to act or answer queries in real time. In additional comparison to previous work, our analysis gives network scaling terms that are often substantially sharper. Our development yields an algorithm with convergence rate that scales inversely in the spectral gap of the network. By exploiting known results on spectral gaps for graphs with $n$ nodes, we show that (disregarding logarithmic factors) our algorithm obtains an $\epsilon$-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. Moreover, simulation results show an excellent agreement with these theoretical predictions.

Our analysis covers several settings for distributed minimization. We begin by studying fixed communication protocols, which are of interest in a variety of areas such as cluster computing or sensor networks with a fixed hardware-dependent protocol. We also investigate randomized communication protocols as well as randomized network failures, which are often essential to handle gracefully in wireless sensor networks and large clusters with potential node failures. Randomized communication also 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 an analysis of our algorithm with stochastic gradient information, which is not difficult when combined with our initial theorems. We describe an extension for (structured) regularized objectives that often arise in machine learning problems in Appendix D.

**Contributions:** Our work makes two primary contributions. First, we introduce a distributed dual averaging algorithm that generalizes traditional subgradient methods by incorporating network-based averaging mechanisms. This leads to a more efficient handling of large-scale, decentralized systems. Second, we provide a detailed analysis of how the algorithm's convergence rate is affected by network topology, utilizing spectral gap analysis to quantify this relationship.

**Organization:** The remainder of this paper is organized as follows. Section 2 formalizes the distributed optimization problem and introduces the dual averaging algorithm. Section 3 presents the main convergence results, including an analysis of how the spectral gap of the network impacts the scaling of convergence rates. Section 4 compares our approach to other distributed optimization algorithms. In Section 5, we state and prove the basic convergence results, which are further extended in Section 6 to cover cases depending on the spectral gap. Sections 7 and 8 discuss extensions of the

algorithm to stochastic settings, including noisy communication and stochastic gradients. Finally, Section 9 offers simulation results to validate our theoretical findings and demonstrate the practical effectiveness of our approach.

**Notation:** We collect some notation used throughout the paper. We use $\mathbb{1}$ to denote the all-ones vector. We also use standard asymptotic notation for sequences. If $a_n$ and $b_n$ are positive sequences, then $a_n = \mathcal{O}(b_n)$ means that $\limsup_n a_n/b_n < \infty$, whereas $a_n = \Omega(b_n)$ means that $\liminf_n a_n/b_n > 0$. On the other hand, $a_n = o(b_n)$ means that $\lim_n a_n/b_n = 0$ and $a_n = \omega(b_n)$ means that $\lim_n a_n/b_n = \infty$. Finally, we write $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$.

# 2 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.

## 2.1 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 minimize the sum

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

subject to the constraint that $x \in \mathbb{R}^d$ belongs to some closed convex set $\mathcal{X}$—that is, solve the problem

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

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, and as motivation for the analysis to follow, let us consider a few here. A first example is a sensor network, in which each agent represents a sensor mote, equipped with a radio transmitter for communication, some basic sensing devices, and some local memory and computational power. In environmental applications of sensor networks, each mote $i$ might take a measurement $y_i$ of the temperature, and the global objective could be to compute the median of the measurements $\{y_1, y_2, \ldots, y_n\}$. This median computation problem can be formulated as minimizing the scalar objective function $\frac{1}{n} \sum_{i=1}^{n} f_i(x)$, where $f_i(x) = |x - y_i|$. Similar formulations apply to the problem of computing other statistics such as means, variances, quantiles and other $M$-estimators.

A second motivating example is the machine learning problem first described in Section 1. 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 the $i$th processor, and assuming that each

subset is of equal size (or that the $f_i$ are normalized suitably), the average $f$ is the empirical loss over the entire dataset. Here we use cluster computing as our computational model, where each processor is a node in the cluster, and the graph $G$ contains edges between those processors that are directly connected with small network latencies. A special case of our optimization problem within this computational model is the distributed perceptron, recently considered by McDonald et al. [MHM10].

### 2.2 Standard dual averaging

Our algorithm is based on a projected dual averaging algorithm [Nes09], designed for minimization of a (potentially nonsmooth) convex function $f$ subject to the constraint $x \in \mathcal{X}$. We begin by describing the standard version of this 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* $\psi : \mathbb{R}^d \to \mathbb{R}$ that is assumed to be 1-strongly convex with respect to some norm $\|\cdot\|$—more precisely, the proximal function satisfies

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

In addition, we assume that $\psi(x) \geq 0$ for all $x \in \mathcal{X}$ and that $\psi(0) = 0$; these are standard assumptions that can be made without loss of generality. Examples of such proximal function 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 $\ell_2$-norm for $x \in \mathbb{R}^d$.
- the entropic function $\psi(x) = \sum_{i=1}^d x_i \log x_i - x_i$, which is strongly convex with respect to the $\ell_1$-norm for $x$ in the probability simplex, $\{x \mid x \succeq 0, \langle x, \mathbb{1}\rangle = 1\}$.

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

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

There are many cost functions $f_i$ that satisfy this type of Lipschitz condition. For instance, it holds for any convex function on a compact domain $\mathcal{X}$, or for any polyhedral function on an arbitrary domain [HUL96a]. Note that the Lipschitz condition (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

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

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

$$\Pi_{\mathcal{X}}^{\psi}(z, \alpha) := \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ \langle z, x\rangle + \frac{1}{\alpha}\psi(x) \right\} \tag{4}$$

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 $\psi$ 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 and lazy projection algorithms developed in the context of online optimization [KV05], though in this form the algorithm seems to be originally due to Nesterov [Nes09]. In Section 5, we show that a simple analysis of the convergence of the above procedure allows us to relate it to the distributed algorithm we describe.

### 2.3 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 $i^{th}$ 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

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

Using this notation, given the non-increasing sequence $\{\alpha(t)\}_{t=0}^{\infty}$ of positive stepsizes, each node $i \in V = \{1, 2, \ldots, n\}$ performs the updates

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

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

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 own neighborhood $N(i)$, and then computes the next local iterate $x_i(t+1)$ by a projection defined by the proximal function $\psi$ 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*

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

Note that this quantity is locally defined at node $i$ and so can be computed in a distributed manner. From the definition of updates, it is clear that each element of the sequence $\{z_i(t)\}_{t=0}^{\infty}$ is essentially a weighted average of the gradients seen so far, which is a natural extension of dual averaging. At the same time, as we shall see, the averaging of the dual parameters in the sequence $\{z_i(t)\}_{t=0}^{\infty}$ allows us to neatly sidestep the complexity arising from non-linearity of projections. We will thus be able to generalize the algorithm from equations (5a) and (5b) to the case where $P$ is random and varies with time as well as when the vectors $g_i(t)$ are noisy versions of subgradients, satisfying only $\mathbb{E}[g_i(t)] \in \partial f_i(x_i(t))$.

# 3 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 at length in the sections that follow.

## 3.1 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 $\overline{z}(t) := \frac{1}{n}\sum_{i=1}^{n} z_i(t)$, and we recall the definition (6) of the local average $\widehat{x}_i(T)$.

**Theorem 1** (Basic convergence result)**.** *Let the sequences $\{x_i(t)\}_{t=0}^{\infty}$ and $\{z_i(t)\}_{t=0}^{\infty}$ be generated by the updates* (5a) *and* (5b) *with step size sequence $\{\alpha(t)\}_{t=0}^{\infty}$. Then for any $x^* \in \mathcal{X}$ and for each node $i \in V$, we have*

$$f(\widehat{x}_i(T)) - f(x^*) \leq \frac{1}{T\alpha(T)}\psi(x^*) + \frac{L^2}{2T}\sum_{t=1}^{T}\alpha(t-1) + \frac{2L}{nT}\sum_{t=1}^{T}\sum_{j=1}^{n}\alpha(t)\left\|\overline{z}(t) - z_j(t)\right\|_* + \frac{L}{T}\sum_{t=1}^{T}\alpha(t)\left\|\overline{z}(t) - z_i(t)\right\|_* \tag{7}$$

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 first two terms in the upper bound (7) are optimization error terms that are common to subgradient algorithms. The third and fourth terms 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, roughly, Theorem 1 ensures that as long the bound on the deviation $\left\|\overline{z}(t) - z_i(t)\right\|_*$ is tight enough, for appropriately chosen $\alpha(t)$ (say $\alpha(t) \approx 1/\sqrt{t}$), the error of $\widehat{x}_i(T)$ is small uniformly across all nodes $i \in V$, and asymptotically approaches 0. See Theorem 2 in the next section for a precise statement of rates.

It is worthwhile comparing the optimization error term from the bound (7) to known results. Subgradient descent on the average function $f = \frac{1}{n}\sum_{i=1}^{n} f_i$ has identical convergence rate, as does the randomized version of incremental subgradient descent [NB01]. However, the distributed nature of the algorithm gives a computational advantage over full gradient descent—the gradient computation requires $\mathcal{O}(1)$ computation per computer rather than $\mathcal{O}(n)$ on a single computer. To highlight the benefits compared to incremental subgradient descent, consider the common problem in machine learning and statistics of minimizing a loss on a large dataset. A randomized incremental gradient descent method must access random subsets of data at every iteration, leading to randomized disk seeks with high latency, which the distributed algorithm avoids. In addition, we expect (and empirically see that this is indeed the case) our method to produce more stable iterates, as we observe the gradients of all $n$ functions at every round, albeit with a network communication lag.

[1] The fact that the term $\left\|\overline{z}(t) - z_i(t)\right\|_*$ appears an extra time is no significant difficulty, as we will bound the difference $\overline{z}(t) - z_i(t)$ uniformly for all $i$ when giving concrete convergence results.

### 3.2 Convergence rates and network topology

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

**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 $\alpha(t) = \frac{R\sqrt{1-\sigma_2(P)}}{4L\sqrt{t}}$, we have*

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

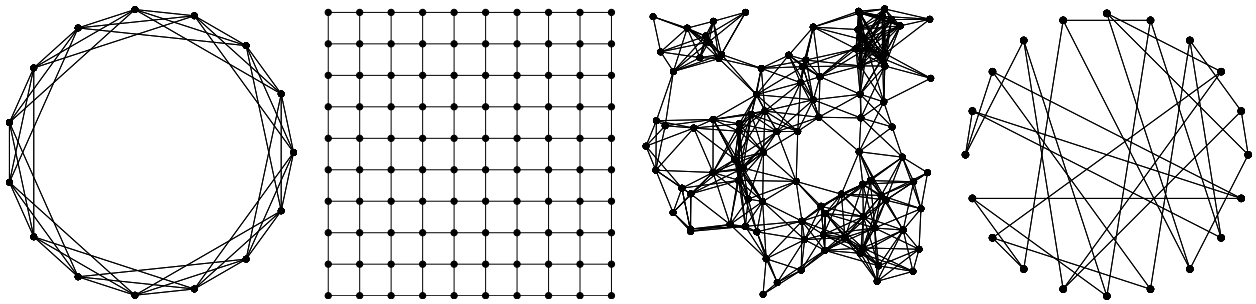
To the best of our knowledge, this theorem is the first to establish a tight connection between the convergence rate of distributed subgradient methods to the spectral properties of the underlying network. In particular, the inverse dependence on the spectral gap $1 - \sigma_2(P)$ is quite natural, since it is well-known to determine the rates of mixing in random walks on graphs [LPW08], 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 different graph topologies that are 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 rather 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. Both the cycle and grid topologies 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., [GK00, Pen03]). There are natural generalizations to dimensions $d > 2$ as well as to cases in which the spatial positions are drawn from a non-uniform distribution.

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 [Chu98]. 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; for instance, a randomly chosen bipartite graph satisfies this property [Alo86], as do random degree regular graphs [FKS89]. In addition, there are several deterministic constructions of expanders that are degree regular (see Section 6.3 of Chung [Chu98] for further details). The deterministic constructions are of interest because they can be used to design a network, while the random constructions are of interest since they are often much simpler.

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



(a) (b) (c) (d)

**Figure 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.

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 [Chu98]. 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 $D = \mathrm{diag}\{\delta_1, \ldots, \delta_n\}$. 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 given by

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

Note that the graph Laplacian $\mathcal{L} = \mathcal{L}(G)$ is always symmetric, positive semidefinite, and satisfies $\mathcal{L} D^{1/2} \mathbb{1} = 0$. Therefore, when the graph is degree-regular ($\delta_i = \delta$ for all $i \in V$), 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 is valid for our theory. For non-degree regular graphs, we need to make a minor modification in order to obtain a doubly stochastic matrix.

Letting $\delta_{\max} = \max_{i \in V} \delta_i$ denote the maximum degree, we define the modified matrix

$$P_n(G) \; := I - \frac{1}{\delta_{\max} + 1}\big(D - A\big) \; = \; I - \frac{1}{\delta_{\max} + 1} D^{1/2} \mathcal{L} D^{1/2} \tag{8}$$

This matrix is symmetric by construction, and moreover, $\sum_{j=1}^n (D_{ij} - A_{ij}) = D_{ii} - \sum_{j=1}^n A_{ij} = 0$ for all $i \in V$, so it is also doubly stochastic. Note that if the graph is $\delta$-regular, then $P_n(G)$ is the standard choice above. Modulo a small technical detail about the ratios of $\delta_{\max}$ to $\delta_i$ and the eigenvalue order of $P$ (see Sec. 6.2), plugging $P_n(G)$ from (8) above into Theorem 2 immediately relates the convergence of distributed dual averaging to the spectral properties of the graph Laplacian, in particular, we have:

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left(\frac{RL}{\sqrt{T}} \frac{\log(Tn)}{\sqrt{\lambda_{n-1}(\mathcal{L}(G))}}\right) \tag{9}$$

The following result summarizes our conclusions for the choice of stochastic matrix in (8) via (9) in application to different network topologies.

**Corollary 1.** *Under the conditions of Theorem 2, we have the following convergence rates:*

*(1) For k-connected paths and cycles,*

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

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

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

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

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

*with high-probability.*

*(4) 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 topology. Instead of stating convergence rates, in order to understand scaling issues as a function of network size and topology, it can be useful to re-state these results in terms of the number of iterations $T_G(\epsilon; n)$ required to achieve error $\epsilon$ for network type $G$ with $n$ nodes. As some special cases, Corollary 1 implies the following scalings:

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

In general, Theorem 2 implies that at most

$$T_G(\epsilon; n) = \mathcal{O}\Big(\frac{1}{\epsilon^2}\cdot\frac{1}{1-\sigma_2(P_n(G))}\Big) \tag{10}$$

iterations are required to achieve an $\epsilon$-accurate solution when using the matrix $P_n(G)$ previously defined in (8).

It is interesting to ask whether the upper bound (10) 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 $\epsilon$-accuracy [NY83], so that 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. For the quadratic proximal function $\psi(x) = \frac{1}{2}\|x\|_2^2$, the following result establishes a lower bound on the number of iterations in terms of graph topology and network structure:

**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 6.3, 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, Proposition 1 implies that our predicted network scaling is sharp. Indeed, in Section 9, we show that the theoretical scalings from Corollary 1—namely, quadratic, linear, and constant in network size $n$—are well-matched in simulations of our algorithm.

### 3.3 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 it $P(t)$ 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 [BGPS06], in which one edge in the network is randomly chosen to communicate at each iteration, allows for 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 7. 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. This yields the scaling

$$f(\widehat{x}_i(T)) - f(x^*) \leq c\,\frac{RL}{\sqrt{T}} \cdot \frac{\log(Tn)}{\sqrt{1-\lambda_2(\mathbb{E}[P(t)^\top P(t)])}} \tag{11}$$

for a universal constant $c$. We can also boost the probability with which this result holds to $1-1/T^k$ for any $k > 1$—without modifying the algorithm—at the cost of incurring a slightly higher constant penalty in the error bound.

The setting of stochastic communication for distributed optimization was previously considered by Lobel and Ozdaglar [LO09]. They established convergence by assuming lower bounds on the entries of $P$ whenever two nodes communicated. As a consequence, 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, and consequently polynomial scaling for any network, as well as faster rates dependent on network structure.

### 3.4 Results for stochastic gradient algorithms

Finally, 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 $\sigma$-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

$$\mathbb{E}\left[\widehat{g}_i(t) \mid \mathcal{F}_{t-1}\right] \in \partial f_i(x_i(t)) \quad \text{and} \quad \mathbb{E}\left[\|\widehat{g}_i(t)\|_*^2 \mid \mathcal{F}_{t-1}\right] \leq L^2 \tag{12}$$

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 result in the case of stochastic gradients. We give the proof and further discussion in Section 8, noting that because we adapt the dual averaging algorithm, the analysis follows quite cleanly from the earlier analysis 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* (12)*. For each $i \in V$, we have*

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

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

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

*If we further assume that the gradient estimates $\widehat{g}_i(t)$ are uncorrelated given $\mathcal{F}_{t-1}$, then with probability at least $1-\delta$,*

$$f(\widehat{x}_i(T)) - f(x^*) \leq \frac{1}{T\alpha(T)}\psi(x^*) + \frac{8L^2}{T}\sum_{t=1}^{T}\alpha(t-1) + \frac{3L^2}{T}\frac{\log(T\sqrt{n})}{1-\sigma_2(P)}\sum_{t=1}^{T}\alpha(t) + \frac{3LR\log\frac{1}{\delta}}{T} + 4LR\sqrt{\frac{\log\frac{1}{\delta}}{nT}}$$

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 (11), but with $\lambda_2(\mathbb{E}[P(t)^\top P(t)])$ replaced by $\sigma_2(P)$.

[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 $\gamma$ is the lower bound on non-zero entries of $P$, so it is at least $4^{n-1}$.

# 4 Related Work

Having stated and discussed our main results in the previous section, we can now more explicitly compare the results in this paper to those in previous work. Our aim here is to give a clear understanding of how our algorithm and results relate to and, in many cases, improve upon prior results. Specifically, with the results of Theorem 2 and Corollary 1 in hand, we can more directly compare our results to other work.

As discussed in the introduction, other researchers have designed algorithms for solving the problem (1). Most previous work [LO09, NO09, NOOT09, RNV10] studies convergence of a (projected) gradient method in which each node $i$ in the network maintains $x_i(t) \in \mathcal{X}$, and at time $t$ performs the update

$$x_i(t+1) = \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \Big\| \sum_{j \in N(i)} P_{ji} x_j(t) - \alpha g_i(t) \Big\|_2^2 \right\} \tag{13}$$

for $g_i(t) \in \partial f_i(x_i(t))$. With the update (13), Corollary 5.5 in the paper [RNV10] shows that

$$f(\widehat{x}_i(T)) - f(x^*) = \mathcal{O}\left( \frac{\alpha n^3 R^2}{T} + \alpha L^2 \right)$$

(we use our notation and assumptions from Theorem 2). The above bound is minimized by setting the stepsize $\alpha \propto \frac{L}{n^{3/2} R \sqrt{T}}$, giving convergence rate $\mathcal{O}(n^{3/2} RL/\sqrt{T})$. It is clear that this convergence rate is substantially slower than all the rates in Corollary 1.

The distributed dual averaging algorithm (5a)–(5b) is quite different from the update (13). The use of the proximal function $\psi$ 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 (e.g. [NY83, 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 (13). Precisely because of this technical issue, earlier works [NO09, LO09] studied unconstrained optimization, and the averaging in $z_i(t)$ seems essential to the faster rates our approach achieves as well as the ease with which we can extend our results to stochastic settings.

In other related work, Johansson et al. [JRJ09] 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

$$x(t+1) = \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \big\| x(t) - \alpha g_{i(t)}(t) \big\|_2^2 \right\} \tag{14}$$

Johansson et al. show that with optimal setting of $\alpha$ and symmetric transition matrix $P$, MIGD has convergence rate $\mathcal{O}(RL \max_i \sqrt{\frac{n\Gamma_{ii}}{T}})$, where $\Gamma$ 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 $i$th eigenvalue of $P$. The eigenvalues of $\Gamma$ are thus 1 and $1/(1 - \lambda_i(P))$ for $i > 1$, and so we have

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

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 (where $d \geq 2$) 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 $T(\epsilon; n) = \mathcal{O}(n/\epsilon^2)$, essentially a factor of $n$ worse than our results.

# 5 Basic convergence analysis for distributed dual averaging

In this section, we prove convergence of the distributed algorithm based on the updates (5a) and (5b). We begin in Section 5.1 by defining some auxiliary quantities and establishing lemmas useful in the proof, and we prove Theorem 1 in Section 5.2.

## 5.1 Setting up the analysis

Using techniques related to those used in past work [NO09], we establish convergence via two auxiliary sequences, given by

$$\overline{z}(t) \; := \frac{1}{n} \sum_{i=1}^{n} z_i(t) \quad \text{and} \quad y(t) \; := \Pi_{\mathcal{X}}^{\psi}(-\overline{z}(t), \alpha) \tag{15}$$

We begin by showing that the average sum of gradients $\overline{z}(t)$ evolves in a very simple way. In particular, we have

$$\overline{z}(t+1) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \big( p_{ij} z_j(t) + g_i(t) \big)$$

Consider the right-hand side above, let $Z(t) = [z_1(t) \; \cdots \; z_n(t)]$ be the matrix of vectors $z_i$, and denote $P = [p_1 \; \cdots \; p_n]$. Since the matrix $P$ is doubly stochastic, we have

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} z_j(t) \; = \; \frac{1}{n} Z(t) P \mathbb{1} \; = \; \frac{1}{n} Z(t) \mathbb{1} \; = \; \overline{z}(t)$$

which yields the evolution

$$\overline{z}(t+1) = \overline{z}(t) + \frac{1}{n} \sum_{j=1}^{n} g_j(t) \tag{16}$$

Consequently, the (negative of the) averaged dual sequence $\{\overline{z}(t)\}_{t=0}^{\infty}$ evolves almost like standard subgradient descent on the function $f(x) = \sum_{i=1}^{n} f_i(x)/n$, the only difference being $g_i(t)$ 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 (16) of the averaged dual sequence allows us to avoid difficulties with the non-linearity of projection that have been challenging in earlier work.

Before proceeding with the proof of Theorem 1, we state two useful results regarding the convergence of the standard dual averaging algorithm, though we defer their proofs to Appendix A. We begin by giving a convergence guarantee for the single-objective form of the dual averaging algorithm. Let $\{g(t)\}_{t=1}^{\infty} \subset \mathbb{R}^d$ be an arbitrary sequence of vectors, and consider the sequence $\{x(t)\}_{t=1}^{\infty}$ defined by

$$x(t+1) \; := \; \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ \sum_{s=1}^{t} \langle g(s), x \rangle + \frac{1}{\alpha(t)} \psi(x) \right\} = \Pi_{\mathcal{X}}^{\psi} \Big( \sum_{s=1}^{t} g(s), \alpha(t) \Big) \tag{17}$$

**Lemma 1.** *For any non-increasing sequence $\{\alpha(t)\}_{t=0}^{\infty}$ of positive stepsizes, and for any $x^* \in \mathcal{X}$, we have*

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

Next we state a lemma that allows us to restrict our analysis to the easier to analyze centralized sequence $\{y(t)\}_{t=0}^{\infty}$ from (15):

**Lemma 2.** *Consider the sequences $\{x_i(t)\}_{t=1}^{\infty}$, $\{z_i(t)\}_{t=0}^{\infty}$, and $\{y(t)\}_{t=0}^{\infty}$ defined according to equations* (5a), (5b), *and* (15). *Recall that each $f_i$ is $L$-Lipschitz. For each $i \in V$, we have*

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

*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*

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

Equipped with these tools, we now turn the proof of Theorem 1.

### 5.2 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

$$\begin{aligned}
\sum_{t=1}^{T} f(y(t)) - f(x^*) &= \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} f_i(x_i(t)) - f(x^*) + \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \left[ f_i(y(t)) - f_i(x_i(t)) \right] \\
&\leq \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} f_i(x_i(t)) - f(x^*) + \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{L}{n} \left\| y(t) - x_i(t) \right\| \qquad (18)
\end{aligned}$$

where the inequality follows by the $L$-Lipschitz condition on $f_i$.

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

$$\frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} f_i(x_i(t)) - f_i(x^*) \leq \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \langle g_i(t), x_i(t) - x^* \rangle \qquad (19)$$

Breaking up the right hand side of (19) into two pieces, we obtain

$$\sum_{i=1}^{n} \langle g_i(t), x_i(t) - x^* \rangle = \sum_{i=1}^{n} \langle g_i(t), y(t) - x^* \rangle + \sum_{i=1}^{n} \langle g_i(t), x_i(t) - y(t) \rangle \qquad (20)$$

By definition of the updates for $\overline{z}(t)$ and $y(t)$, we have

$$y(t) = \arg\min_{x \in \mathcal{X}} \left\{ \frac{1}{n} \sum_{s=1}^{t-1} \sum_{i=1}^{n} \langle g_i(s), x \rangle + \frac{1}{\alpha(t)} \psi(x) \right\}$$

Thus, we see that the first term in the decomposition (20) can be written in the same way as the bound in Lemma 1, and as a consequence, we have the bound

$$\frac{1}{n}\sum_{t=1}^{T}\left\langle\sum_{i=1}^{n}g_i(t),y(t)-x^*\right\rangle \leq \frac{1}{2}\sum_{t=1}^{T}\alpha(t-1)\left\|\frac{1}{n}\sum_{i=1}^{n}g_i(t)\right\|_*^2+\frac{1}{\alpha(T)}\psi(x^*)$$
$$\leq \frac{L^2}{2}\sum_{t=1}^{T}\alpha(t-1)+\frac{1}{\alpha(T)}\psi(x^*) \tag{21}$$

It remains to control the final two terms in the bounds (18) and (20). Since $\|g_i(t)\|_* \leq L$ by assumption, we have

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{L}{n}\|y(t)-x_i(t)\|+\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle g_i(t),x_i(t)-y(t)\rangle$$
$$\leq \frac{2L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\|y(t)-x_i(t)\|$$
$$= \frac{2L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\Pi_{\mathcal{X}}^{\psi}(-\overline{z}(t),\alpha(t))-\Pi_{\mathcal{X}}^{\psi}(-z_i(t),\alpha(t))\right\|$$

By the $\alpha$-Lipschitz continuity of the projection operator $\Pi_{\mathcal{X}}^{\psi}(\cdot,\alpha)$ (see Appendix A.3), we have

$$\frac{2L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\Pi_{\mathcal{X}}^{\psi}(\overline{z}(t),\alpha(t))-\Pi_{\mathcal{X}}^{\psi}(z_i(t),\alpha)\right\| \leq \frac{2L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\alpha(t)\|\overline{z}(t)-z_i(t)\|_*$$

Combining this bound with (18) and (21) yields the running sum bound

$$\sum_{t=1}^{T}\left[f(y(t))-f(x^*)\right] \leq \frac{1}{\alpha(T)}\psi(x^*)+\frac{L^2}{2}\sum_{t=1}^{T}\alpha(t-1)+\frac{2L}{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\alpha(t)\|\overline{z}(t)-z_j(t)\|_* \tag{22}$$

Applying Lemma 2 to (22) gives that $\sum_{t=1}^{T}[f(x_i(t))-f(x^*)]$ is upper bounded by

$$\frac{1}{\alpha(T)}\psi(x^*)+\frac{L^2}{2}\sum_{t=1}^{T}\alpha(t-1)+\frac{2L}{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\alpha(t)\|\overline{z}(t)-z_j(t)\|_*+L\sum_{t=1}^{T}\alpha(t)\|\overline{z}(t)-z_i(t)\|_*$$

Dividing both sides by $T$ and using convexity of $f$ yields the bound (7).

# 6 Convergence rates, spectral gap, and network topology

In this section, we will give concrete convergence rates for the distributed dual averaging algorithm based on the mixing time of a random walk according to the doubly stochastic matrix $P$. The understanding of the dependence of our convergence rates in terms of the underlying 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 6.1 with the proof of Theorem 2. In Section 6.2, we prove the graph-specific

convergence rates stated in Corollary 1, whereas Section 6.3 contains a proof of the lower bound stated in Proposition 1.

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 $B$, 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. We make frequent use of the following standard inequality: for any positive integer $t = 1, 2, \ldots$ and any $x \in \Delta_n$,

$$\left\|P^t x - \mathbb{1}/n\right\|_{\mathrm{TV}} = \frac{1}{2}\left\|P^t x - \mathbb{1}/n\right\|_1 \leq \frac{1}{2}\sqrt{n}\left\|P^t x - \mathbb{1}/n\right\|_2 \leq \frac{1}{2}\sigma_2(P)^t\sqrt{n} \tag{23}$$

For a brief review of the relevant standard Perron-Frobenius and matrix theory, we refer the reader to Appendix B.

### 6.1 Proof of Theorem 2

We focus on controlling the network error term in the bound (7), namely the quantity

$$\frac{L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\alpha(t)\left\|\overline{z}(t) - z_i(t)\right\|_*$$

Define the matrix $\Phi(t,s) = P^{t-s+1}$ (in the sequel we allow the stochastic matrix $P$ to change as a function of time). Let $[\Phi(t,s)]_{ji}$ be the $j$th entry of the $i$th column of $\Phi(t,s)$. Then via a bit of algebra, we can write

$$z_i(t+1) = \sum_{j=1}^{n}[\Phi(t,s)]_{ji}z_j(s) + \sum_{r=s+1}^{t}\left(\sum_{j=1}^{n}[\Phi(t,r)]_{ji}g_j(r-1)\right) + g_i(t) \tag{24}$$

Clearly the above reduces to the standard update (5a) when $s = t$. Since $\overline{z}(t)$ evolves simply as in (16), we assume that $z_i(0) = \overline{z}(0)$ to avoid notational clutter—we can simply start with $z_i(0) = 0$—and use (24) to see

$$\overline{z}(t) - z_i(t) = \sum_{s=1}^{t-1}\sum_{j=1}^{n}(1/n - [\Phi(t-1,s)]_{ji})g_j(s-1) + \left(\frac{1}{n}\sum_{j=1}^{n}(g_j(t-1) - g_i(t-1))\right) \tag{25}$$

We use the fact that $\left\|g_i(t)\right\|_* \leq L$ for all $i$ and $t$ and (25) to see that

$$\begin{aligned}
\left\|\overline{z}(t) - z_i(t)\right\|_* &= \left\|\sum_{s=1}^{t-1}\sum_{j=1}^{n}(1/n - [\Phi(t-1,s)]_{ji})g_j(s-1) + \left(\frac{1}{n}\sum_{j=1}^{n}g_j(t-1) - g_i(t-1)\right)\right\|_* \\
&\leq \sum_{s=1}^{t-1}\sum_{j=1}^{n}\left\|g_j(s-1)\right\|_*\left|(1/n) - [\Phi(t-1,s)]_{ji}\right| + \frac{1}{n}\sum_{i=1}^{n}\left\|g_j(t-1) - g_i(t-1)\right\|_* \\
&\leq \sum_{s=1}^{t-1} L\left\|[\Phi(t-1,s)]_i - \mathbb{1}/n\right\|_1 + 2L.
\end{aligned} \tag{26}$$

Now we break the sum in (26) 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 (23), $\|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \le \sqrt{n}\sigma_2(P)^{t-s+1}$. Hence, if

$$t - s \ge \frac{\log \epsilon^{-1}}{\log \sigma_2(P)^{-1}} - 1 \quad \text{we immediately have} \quad \|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \le \sqrt{n}\epsilon$$

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

$$\|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \le \frac{1}{T} \tag{27}$$

For larger $s$, we simply have $\|[\Phi(t,s)]_j - \mathbb{1}/n\|_1 \le 2$. 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 in (26) and use (27) to see that since $t - 1 - (t - \widehat{t}) = \widehat{t}$ and there are at most $T$ steps in the summation,

$$\begin{aligned}\|\overline{z}(t) - z_i(t)\|_* &\le L \sum_{s=t-\widehat{t}}^{t-1} \|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1 + L \sum_{s=1}^{t-1-\widehat{t}} \|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1 + 2L \\ &\le 2L\frac{\log(T\sqrt{n})}{\log \sigma_2(P)^{-1}} + 3L \le 2L\frac{\log(T\sqrt{n})}{1-\sigma_2(P)} + 3L \end{aligned} \tag{28}$$

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

Combining (28) with the running sum bound in (22) of the proof of the basic theorem, Theorem 1, we immediately see that for $x^* \in \mathcal{X}$,

$$\sum_{t=1}^{T} f(y(t)) - f(x^*) \le \frac{1}{\alpha(T)}\psi(x^*) + \frac{L^2}{2}\sum_{t=1}^{T}\alpha(t-1) + 6L^2\sum_{t=1}^{T}\alpha(t) + 4L^2\frac{\log(T\sqrt{n})}{1-\sigma_2(P)}\sum_{t=1}^{T}\alpha(t) \tag{29}$$

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

### 6.2 Proof of Corollary 1

The corollary is based on bounding the spectral gap of the matrix $P_n(G)$ from equation (8).

**Lemma 3.** *The matrix $P_n(G)$ satisfies the bound*

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

**Proof** By a theorem of Ostrowski on congruent matrices (cf. Theorem 4.5.9, [HJ85]), we have

$$\lambda_k(D^{1/2}\mathcal{L}D^{1/2}) \in \left[\min_i \delta_i\lambda_k(\mathcal{L}), \max_i \delta_i\lambda_k(\mathcal{L})\right] \tag{30}$$

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 (8), 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 (30) 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

$$1 - \frac{\delta_{\min}}{\delta_{\max}+1}\lambda_{n-1}(\mathcal{L}) \quad \text{and} \quad \frac{\delta_{\max}}{\delta_{\max}+1}\lambda_1(\mathcal{L}) - 1.$$

□

Much of spectral graph theory is devoted to bounding $\lambda_{n-1}(\mathcal{L})$ sufficiently far away from zero, and Lemma 3 allows us to conveniently leverage such results for bounding the convergence rate of our algorithm.

Note that computing the upper bound in Lemma 3 requires controlling both $\lambda_{n-1}(\mathcal{L})$ and $\lambda_1(\mathcal{L})$. In order to circumvent this complication, we use the well-known idea of a "lazy" random walk [Chu98, LPW08], in which we replace $P$ by $\frac{1}{2}(I+P)$. The resulting symmetric matrix has the same eigenstructure as $P$, and moreover, we have

$$\sigma_2\Big(\frac{1}{2}(I+P)\Big) = \lambda_2\Big(\frac{1}{2}(I+P)\Big) = \lambda_2\Big(I - \frac{1}{2(\delta_{\max}+1)}D^{1/2}\mathcal{L}D^{1/2}\Big) \leq 1 - \frac{\delta_{\min}}{2(\delta_{\max}+1)}\lambda_{n-1}(\mathcal{L}) \quad (31)$$

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 the $n$ 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 Chapter 3 of Gray [Gra06]) imply that it has $m$th eigenvalue

$$\lambda_m(\mathcal{L}) = 1 - \frac{1}{2k}\sum_{j=1}^{k}\exp\left(-2\pi ijm/n\right) - \frac{1}{2k}\sum_{j=1}^{k}\exp\left(-2\pi i(n-j)m/n\right) = 1 - \frac{1}{k}\sum_{j=1}^{k}\cos\left(\frac{2\pi jm}{n}\right)$$

For $m = n-1$ and $k = o(n)$, the last equation can be massaged into [BGPS06, Section VI.A]

$$\lambda_{n-1}(\mathcal{L}) = 1 - \cos\left(\frac{2\pi k}{n}\right) + \Theta\left(\frac{k^4}{n^4}\right)$$

By performing a Taylor expansion of $\cos(\cdot)$, we see that $\lambda_{n-1}(\mathcal{L}) = \Theta\left(\frac{k^2}{n^2}\right)$ for $k = o(n)$.

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 (see Lemma 5 in Appendix C), 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 3 to see that for regular $k$-connected paths or cycles with $k \leq \sqrt{n}$,

$$\sigma_2(P) = 1 - \Theta\left(\frac{k^2}{n^2}\right) \quad (32)$$

Substituting the bound (32) into Theorem 2 yields the claim of Corollary 1(a).

**Regular grids:** Now consider the case of a $\sqrt{n}$-by-$\sqrt{n}$ grid, focusing specifically 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 [Chu98, Section 2.6] to analyze the eigen-structure of the Laplacian. In particular, the toroidal $\sqrt{n}$-by-$\sqrt{n}$ $k$-connected grid is simply the Cartesian product of two regular $k$-connected cycles 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 [Chu98, Theorem 2.13]. Thus, based on the preceding discussion of $k$-connected cycles, we conclude that if $k = o(\sqrt{n})$, then we have $\lambda_{n-1}(\mathcal{L}) = \Theta(k^2/n)$. For a non-toroidal $\sqrt{n}$-by-$\sqrt{n}$ grid (in which the network is *not* "wrapped" on its boundaries, as in Figure 1(b)), we use the previous discussion of regular $k$-connected paths, since the grid is the Cartesian product of two $k$-connected paths of $\sqrt{n}$ nodes. We immediately see that $\lambda_{n-1}(\mathcal{L}) = \Theta(k^2/n)$. In both cases, for $\sqrt{n}$-by-$\sqrt{n}$ $k$-connected grids, we use Lemma 3 and (31) to see that for $k \leq n^{1/4}$,

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

The result in Corollary 1(b) immediately follows.

**Random geometric graphs:** Using the proof of Lemma 10 from Boyd et al. [BGPS06], we see that for any $\epsilon > 0$, if $r = \sqrt{\log^{1+\epsilon} n/(n\pi)}$, then with probability at least $1 - 2/n^{c-1}$,

$$\min_i \delta_i \geq \log^{1+\epsilon} n - \sqrt{2}c \log n \quad \text{and} \quad \max_i \delta_i \leq \log^{1+\epsilon} n + \sqrt{2}c \log n \tag{34}$$

Thus, letting $\mathcal{L}$ be the graph Laplacian of a random geometric graph, if we can bound $\lambda_{n-1}(\mathcal{L})$, (34) coupled with Lemma 3 will control the convergence rate of our algorithm.

Recent work of von Luxburg et al. [vLRH10] gives concentration results on the second-smallest eigenvalue of a geometric graph. In particular, their Theorem 3 says that there are universal constants $c_1, \ldots, c_5 > 0$ such that with probability at least $1 - c_1 n \exp(-c_2 n r^2) - c_3 \exp(-c_4 n r^2)/r^2$, $\lambda_{n-1}(\mathcal{L}) \geq c_5 r^2$. Parsing this a bit, we see 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 (34), we see that for $r = (\log^{1+\epsilon} n/n)^{1/2}$,

$$\frac{\min_i \delta_i}{\max_i \delta_i} = \Theta(1) \quad \text{and} \quad \lambda_{n-1}(\mathcal{L}) = \Omega\left(\frac{\log^{1+\epsilon} n}{n}\right)$$

with high probability. Combining the above equation with Lemma 3 and (31), we have

$$\sigma_2(P) = 1 - \Omega\left(\frac{\log^{1+\epsilon} n}{n}\right) \tag{35}$$

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 [Chu98, Chapter 6] removes any penalty due to network communication (up to logarithmic factors), and hence yields Corollary 1(d).

### 6.3 Proof of Proposition 1

We now give a proof of Proposition 1, which shows that the dependence of our convergence rates on the spectral gap is tight. The 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 communication matrix $P$.

Recall that $\mathbb{1} \in \mathbb{R}^n$ is the eigenvector of $P$ corresponding to its largest eigenvalue (equal to 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 6.2, 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, we have $g_i(t) = c + w_i$ for all $t = 1, 2, \ldots$. Defining the vector $g = (c\mathbb{1} + w) \in \mathbb{R}^n$, we have the evolution

$$\begin{aligned} z(t+1) = Pz(t) + g = P^2 z(t-1) + Pg + g = \cdots = \sum_{\tau=0}^t P^\tau g \\ = \sum_{\tau=0}^{t-1} P^\tau (w + c\mathbb{1}) = \sum_{\tau=0}^{t-1} P^\tau w + ct\mathbb{1} = \sum_{\tau=0}^{t-1} \sigma_2(P)^\tau w + ct\mathbb{1} \end{aligned} \tag{36}$$

since $P\mathbb{1} = \mathbb{1}$.

In order to establish a lower bound, it suffices to show that at least one node is far from the optimum after $t$ steps, and we focus on node 1. Since $w_1 = -1$, the evolution (36) guarantees that

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

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

$$x_i(t+1) = \arg\min_{x \in \mathcal{X}} \left\{ z_i(t+1)x + \frac{1}{2\alpha(t)}x^2 \right\} = \arg\min_{x \in \mathcal{X}} \left\{ \big(x + \alpha(t) z_i(t+1)\big)^2 \right\}$$

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 little calculation shows that we require $t = \Omega((1 - \sigma_2(P))^{-1})$ in order to drive $z_1(t)$ below zero.

# 7 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 7.1 with basic convergence results in this setting, and then prove Theorem 3. Section 7.2 contains a more detailed treatment of the case of gossip algorithms, and Section 7.3 contains the setting of edge failures.

## 7.1 Basic convergence analysis

Recall that Theorem 1 involves the sum $\frac{2L}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\alpha(t)\left\|\overline{z}(t)-z_i(t)\right\|_*$. In Section 6, we showed how to control this sum when communication between agents occurs on a static underlying network structure via a doubly-stochastic matrix $P$. We now relax the assumption that $P$ is fixed and instead let $P(t)$ vary over time.

### 7.1.1 Markov chain mixing for stochastic communication

We use $P(t) = [p_1(t) \ \cdots \ p_n(t)]$ to denote the doubly stochastic symmetric matrix at iteration $t$. The update employed by the algorithm, modulo changes in $P$, is given by the usual updates (5a) and (5b)—namely,

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

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 $\overline{z}(t+1) = \overline{z}(t) - \frac{1}{n}\sum_{i=1}^{n} g_i(t)$ from equation (16), and moreover, (25) holds essentially unchanged:

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

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

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

which we establish by recalling and modifying a few known results [BGPS06].

Let $\Delta_n$ denote the $n$-dimensional probability simplex and the vector $u(0) \in \Delta_n$ be arbitrary. Consider the random sequence $\{u(t)\}_{t=0}^{\infty}$ generated by the recursion $u(t+1) = P(t)u(t)$. 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 $v(t+1)$, we have

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

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

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

$$\leq \epsilon^{-2} \frac{\big\|v(0)\big\|_2^2 \lambda_2 \big(\mathbb{E}P(t)^2\big)^t}{\big\|u(0)\big\|_2^2}$$

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

### 7.1.2 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 $\widehat{t}$ such that for $t - s \geq \widehat{t}$, with exceedingly high probability, $\Phi(t,s)$ is close to the uniform matrix $\mathbb{1}\mathbb{1}^T/n$. We then break the summation from 1 to $T$ into two separate terms, separated by the cut-off point $\widehat{t}$. Throughout this derivation, we let $\lambda_2 = \lambda_2(\mathbb{E}[P(t)^2])$, where we have suppressed the dependence of $\lambda_2$ on graph structure $G$ to ease notation.

Using the probabilistic bound (39), note that

$$t - s \geq \frac{3 \log \epsilon^{-1}}{\log \lambda_2^{-1}} - 1 \quad \text{implies} \quad \mathbb{P}\big[\big\|\Phi(t,s)e_i - \mathbb{1}/n\big\|_2 \geq \epsilon\big] \leq \epsilon$$

Consequently, if we make the choice

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

then we are guaranteed that if $t - s \geq \widehat{t} - 1$, then

$$\mathbb{P}\big[\big\|\Phi(t,s)e_i - \mathbb{1}/n\big\|_2 \geq 1/(T^2 n)\big] \leq (T^2 n)^2 \lambda_2^{\frac{3\log(T^2 n)}{-\log \lambda_2}} \;=\; (T^2 n)^2 \big(e^{\log \lambda_2}\big)^{\frac{\log(T^6 n^3)}{-\log \lambda_2}} \;=\; \frac{1}{T^2 n} \qquad (40)$$

Recalling the bound (26), we have

$$\begin{aligned}
\|\overline{z}(t) - z_i(t)\|_* &\leq L \sum_{s=1}^{t-1} \|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1 + 2L \\
&= L \sum_{s=t-\widehat{t}}^{t-1} \|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1 + L \sum_{s=1}^{t-1-\widehat{t}} \|\Phi(t-1,s)e_i - \mathbb{1}/n\|_1 + 2L \\
&\leq 2L \frac{3\log(T^2 n)}{1-\lambda_2} + \underbrace{L\sqrt{n} \sum_{s=1}^{t-1-\widehat{t}} \big\|\Phi(t-1,s)e_i - \mathbb{1}/n\big\|_2}_{S} + 2L \qquad (41)
\end{aligned}$$

It remains to bound the sum $S$. For any fixed pair $s' < s$, since the matrices $P(t)$ are doubly stochastic, we have

$$\begin{aligned}
\big\|\Phi(t-1,s')e_i - \mathbb{1}/n\big\|_2 &= \big\|\Phi(s-1,s')\Phi(t-1,s)e_i - \mathbb{1}/n\big\|_2 \\
&\leq \big|\!\big|\!\big|\Phi(s-1,s')\big|\!\big|\!\big|_2 \, \big\|\Phi(t-1,s)e_i - \mathbb{1}/n\big\|_2 \\
&\leq \big\|\Phi(t-1,s)e_i - \mathbb{1}/n\big\|_2
\end{aligned}$$

where the final inequality uses the bound $\|\Phi(s-1,s')\|_2 \leq 1$. From the bound (40), we have the bound $\left\|\Phi(t-1,t-\widehat{t}-1)e_i - \mathbb{1}/n\right\|_2 \leq \frac{1}{T^2 n}$ with probability at least $1-1/(T^2 n)$. Since $s$ ranges between 1 and $t-\widehat{t}$ in the summation $S$, we conclude that

$$S \leq L\sqrt{n}\, T\, \frac{1}{T^2 n} = \frac{L\sqrt{n}}{T^2 n}$$

and hence assuming that $n \geq 3$,

$$\|\overline{z}(t) - z_i(t)\|_* \leq L\frac{6\log(T^2 n)}{1-\lambda_2} + L\sqrt{n}\frac{1}{Tn} + 2L$$

with probability at least $1-1/(T^2 n)$. Applying the union bound over all iterations $t = 1,\ldots,T$ and nodes $i = 1,\ldots,n$, we obtain

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

Recalling the master bound from Theorem 1 completes the proof.
In the remainder of this section, we give some applications of the stochastic framework outlined above, showing a few sampling schemes and giving bounds on their convergence rates.

## 7.2 Gossip-like protocols

Gossip algorithms are procedures for achieving consensus in a network robustly and quickly by randomly selecting one edge $(i,j)$ in the network for communication at each iteration [BGPS06]. Once nodes $i$ and $j$ are selected, their values are averaged. Gossip algorithms drastically reduce communication in the network, yet they still enjoy fast convergence and are robust to changes in topology.

### 7.2.1 Partially asynchronous gossip protocols

In a partially asynchronous iterative method, agents synchronize their iterations [BT89]. This is the model of standard gossip protocols, 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 $P(t) = I - \frac{1}{2}(e_i - e_j)(e_i - e_j)^T$. It is clear that $P(t)^T P(t) = P(t)$, 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 as in Section 6.2. At round $t$, edge $(i,j)$ (with $A_{ij} = 1$) is chosen with probability $1/\langle \mathbb{1}, A\mathbb{1}\rangle$. Thus,

$$\begin{aligned}\mathbb{E}P(t) &= \frac{1}{\langle \mathbb{1}, A\mathbb{1}\rangle}\sum_{(i,j):A_{ij}=1} I - \frac{1}{2}(e_i - e_j)(e_i - e_j)^T = I - \frac{1}{\langle \mathbb{1}, A\mathbb{1}\rangle}(D - A) \\ &= I - \frac{1}{\langle \mathbb{1}, A\mathbb{1}\rangle}D^{1/2}(I - D^{-1/2}AD^{-1/2})D^{1/2} = I - \frac{1}{\langle \mathbb{1}, A\mathbb{1}\rangle}D^{1/2}\mathcal{L}D^{1/2} \qquad (42)\end{aligned}$$

since $\sum_{(i,j):A_{ij}=1}(e_i - e_j)(e_i - e_j)^T = 2(D - A)$. Using an identical argument as that for Lemma 3, 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. In a clustered computing environment with some centralized control, it is possible to select more than one edge per round so long as no two edges share vertices (for example, by selecting a random maximal matching) and still have $P(t)^T P(t) = P(t)$. For a $\delta$-regular graph, choosing a random maximal matching achieves a spectral gap within constant factors of the spectral gap of the underlying graph but uses only $\Theta(1/\delta)$ as much communication.

#### 7.2.2 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). We assume that no two agents have clocks tick at the same time. This communication corresponds to a gossip protocol with stochastic subgradients, and its convergence can be described simply by combining (42) with Theorem 4. This type of algorithm is well-suited to completely decentralized environments, such as sensor networks.

### 7.3 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, since computation of $\mathbb{E}P(t)^2$ is in general non-trivial, we work with the model of lazy random walks described in Section 6.2. In the lazy random walk model, the communication matrix at each round is $\frac{1}{2}I + \frac{1}{2}P(t)$, which is symmetric PSD since $\sigma_1(P(t)) \leq 1$. Further, for any symmetric PSD stochastic matrix $P$, $P^2 \preceq P$. With that in mind, we see that $\mathbb{E}(\frac{1}{2}I + \frac{1}{2}P(t))^2 \preceq \frac{1}{2}I + \frac{1}{2}\mathbb{E}P(t)$, and applying Weyl's Theorem for the eigenvalues of a Hermitian matrix [HJ85, Theorem 4.3.1],

$$\lambda_2\left(\mathbb{E}\left(\frac{1}{2}I + \frac{1}{2}P(t)\right)^2\right) \leq \lambda_2\left(\frac{1}{2}I + \frac{1}{2}\mathbb{E}P(t)\right) = \frac{1}{2} + \frac{1}{2}\lambda_2(\mathbb{E}P(t)) \tag{43}$$

Thus any bound on $\lambda_2(\mathbb{E}P(t))$ 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 (8). With these definitions, $\mathbb{E}A(t) = \frac{2\delta_{\max}+1}{(\delta_{\max}+1)^2}A$, $\mathbb{E}D(t) = \frac{2\delta_{\max}+1}{(\delta_{\max}+1)^2}D$, and $A - D =$

$(P - I)(\delta_{\max} + 1)$. We have

$$\mathbb{E}P(t) = I - (\delta_{\max} + 1)^{-1}(\mathbb{E}D(t) - \mathbb{E}A(t)) = \left(\frac{\delta_{\max}}{\delta_{\max} + 1}\right)^2 I + \frac{2\delta_{\max} + 1}{(\delta_{\max} + 1)^2} P$$

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 decreases (and hence convergence rate may slow) by a factor proportional to the maximum degree in the graph. This is not surprising, since 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 $\rho$ 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 easily have

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

and $\lambda_2(\mathbb{E}P(t)) = \rho + (1 - \rho)\lambda_2(P)$. Applying (11), we see that we lose at most a factor of $\sqrt{1 - \rho}$ in the convergence rate.

# 8 Stochastic Gradient Optimization

In this section, we show that 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, 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 smooths 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 in the analysis to address passing noisy gradients through nonlinear projections [RNV10].

## 8.1 Proof of Theorem 4

We begin by using convexity and the Lipschitz continuity of the $f_i$ (see equations (18) and (19)), thereby obtaining that the running sum $S(T) = \sum_{t=1}^{T} f(y(t)) - f(x^*)$ is upper bounded as

$$\begin{aligned} S(T) &\leq \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \langle g_i(t), x_i(t) - x^* \rangle + \sum_{t=1}^{T} \sum_{i=1}^{n} L \left\| y(t) - x_i(t) \right\| \\ &= \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \langle \widehat{g}_i(t), x_i(t) - x^* \rangle + \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} L \left\| y(t) - x_i(t) \right\| + \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^* \rangle \end{aligned} \tag{44}$$

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 nothing in

Lemma 1 assumes that $\widehat{g}_i(t)$ is related to $f_i(x_i(t))$. So we upper bound the first term in (44) with

$$\frac{1}{\alpha(T)}\psi(x^*) + \frac{1}{2}\sum_{t=1}^{T}\alpha(t-1)\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{g}_i(t)\right\|_*^2 + \sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\langle \widehat{g}_i(t), x_i(t) - y(t)\rangle \tag{45}$$

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

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

Further, $x_i(t) \in \mathcal{F}_{t-1}$ and $y(t) \in \mathcal{F}_{t-1}$ by assumption for $j \in [n]$ and $s \le t-1$, so

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

Recalling that $\|x_i(t) - y(t)\| \le \alpha(t)\|\overline{z}(t) - z_i(t)\|_*$, we proceed by putting expectations around the norm terms in (26) and (28) to see that

$$\frac{1}{\alpha(t)}\mathbb{E}\|y(t) - x_i(t)\| \le \mathbb{E}\|\overline{z}(t) - z_i(t)\|_* \le \sum_{s=1}^{t-1} L\|[\Phi(t-1,s)]_i - \mathbb{1}/n\|_1 + 2L \le L\frac{\log(T\sqrt{n})}{1-\sigma_2(P)} + 3L$$

Coupled with the above arguments, we can bound the expectation of (44) by

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} f(y(t)) - f(x^*)\right] \le \frac{1}{\alpha(T)}\psi(x^*) + \frac{L^2}{2}\sum_{t=1}^{T}\alpha(t-1) + \left(2L^2\frac{\log(T\sqrt{n})}{1-\sigma_2(P)} + 6L^2\right)\sum_{t=1}^{T}\alpha(t) \\ + \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\left[\langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^*\rangle\right]\end{aligned} \tag{46}$$

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

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

which completes the proof of the first statement of the theorem.

To show that the statement holds with high-probability when $\mathcal{X}$ is compact and $\|\widehat{g}_i(t)\|_* \le L$, it is sufficient to establish that the sequence $\langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^*\rangle$ is a bounded martingale, and then apply Azuma's inequality [Azu67]. (Here we are exploiting the fact that 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

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

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

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^*\rangle \ge \epsilon\right] \le \exp\left(-\frac{\epsilon^2}{16Tn^2L^2R^2}\right)$$

Dividing by $T$ and setting the probability above equal to $\delta$, we obtain that with probability at least $1-\delta$,

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

The second statement of the theorem is now obtained by appealing to Lemma 2. 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.

Proving the last statement of the theorem—the concentration result with uncorrelated noise at each node—requires a martingale extension of Bernstein's inequality [Fre75]. Indeed, one form of Freedman's inequality states that if $X_1, \ldots, X_T$ is a martingale difference sequence, $|X_i| \leq B$ uniformly, and $V \geq \sum_{t=1}^{T} \mathrm{Var}(X_t \mid \mathcal{F}_{t-1})$, then for any $v, \epsilon > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t \geq \epsilon \;\text{ and }\; V \leq v\right) \leq \exp\left(-\frac{\epsilon^2}{2v + 2B\epsilon/3}\right)$$

To extend the above bound to our setting, we recall that $\frac{1}{n}\sum_{i=1}^{n}\langle g_i(t)-\widehat{g}_i(t), x_i(t)-x^*\rangle$ is Martingale difference sequence uniformly bounded by $2LR$. Further, since the expectation is zero, we have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\langle g_i(t)-\widehat{g}_i(t), x_i(t)-x^*\rangle \mid \mathcal{F}_{t-1}\right)\\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}^{n}\langle g_i(t)-\widehat{g}_i(t), x_i(t)-x^*\rangle\langle g_j(t)-\widehat{g}_j(t), x_j(t)-x^*\rangle \mid \mathcal{F}_{t-1}\right]\\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\langle g_i(t)-\widehat{g}_i(t), x_i(t)-x^*\rangle^2 \mid \mathcal{F}_{t-1}\right] \qquad (48)\\
&\leq \frac{1}{n^2}\sum_{i=1}^{n} 4L^2R^2 = \frac{4L^2R^2}{n}.
\end{aligned}$$

The decorrelation equality in (48) follows by our assumption that $\widehat{g}_i(t)$ and $\widehat{g}_j(t)$ are uncorrelated given $\mathcal{F}_{t-1}$, and that $x_i(t)$, $g_i(t)$, and $x^* \in \mathcal{F}_{t-1}$. Substituting $4TL^2R^2/n$ as an upper bound for the variance in Freedman's inequality, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle g_i(t)-\widehat{g}_i(t), x_i(t)-x^*\rangle \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{8TL^2R^2/n + 8LR\epsilon/3}\right)$$

To find a $\delta$ so that $\exp(\cdot)$ term is less than or equal to $\delta$, we solve

$$\delta \geq \exp\left(-\frac{\epsilon^2}{8L^2R^2/n + 8LR\epsilon/3}\right) \quad \text{or} \quad \epsilon^2 - \epsilon\frac{8LR\log\frac{1}{\delta}}{3} - \frac{8TL^2R^2\log\frac{1}{\delta}}{n} \geq 0 \qquad (49)$$

Solving the above quadratic in $\epsilon$, we have equality with zero for

$$\epsilon = \frac{(8/3)LR\log\frac{1}{\delta} \pm \sqrt{(8/3)^2L^2R^2\log^2\frac{1}{\delta} + (32/n)TL^2R^2\log\frac{1}{\delta}}}{2}$$

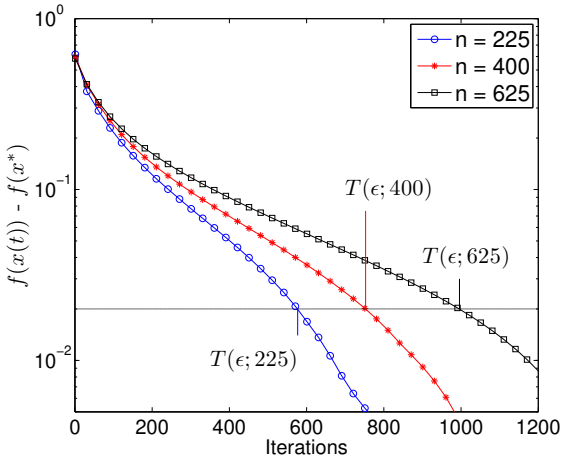



**Figure 2.** Plot of the function error versus the number of iterations for a grid graph. Each curve corresponds to a grid with a different number of nodes ($n \in \{225, 400, 600\}$. As expected, larger graphs require more iterations to reach a pre-specified tolerance $\epsilon > 0$, as defined by the iteration number $T(\epsilon; n)$. The network scaling problem is to determine how $T(\epsilon; n)$ scales as a function of $n$.

In particular, noting that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, it is sufficient that

$$\epsilon \geq \frac{4}{3} LR \log \frac{1}{\delta} + \frac{\sqrt{6}}{3} LR \log \frac{1}{\delta} + 2\sqrt{2} LR \sqrt{\frac{T}{n} \log \frac{1}{\delta}}$$

for the inequality in (49) to be satisfied. Thus with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \langle g_i(t) - \widehat{g}_i(t), x_i(t) - x^* \rangle < 3LR \log \frac{1}{\delta} + 4LR \sqrt{\frac{T}{n} \log \frac{1}{\delta}}$$

Dividing by $T$ completes the proof of the last statement of Theorem 4.

## 9 Simulations

In this section, we report experimental results on the network scaling behavior of the distributed dual averaging algorithm as a function of the graph structure and number of processors $n$. These results illustrate the excellent agreement of the empirical behavior with our theoretical predictions.

For all experiments reported here, we consider distributed minimization of a sum of hinge loss functions; it is this optimization problem that underlies the widely-used support vector machine method for classification [CV95]. In a classification problem, we are given $n$ pairs of the form $(b_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$, where $b_i \in \mathbb{R}^d$ corresponds to a feature vector and $y_i \in \{-1, +1\}$ is the associated label. The goal is to use these samples to estimate a linear classifier, meaning a function of the form $b \mapsto \operatorname{sign} \langle b, x \rangle$ based on some weight vector $x \in \mathbb{R}^d$. In methods based on support vector machines, the weight vector is chosen by minimizing a sum of hinge loss functions associated with each pair $(b_i, y_i)$. In particular, given the shorthand notation $[c]_+ := \max\{0, c\}$, the hinge loss associated with a linear classifier based on $x$ is given by $f_i(x) = [1 - y_i \langle b_i, x \rangle]_+$. The global objective is a sum of $n$ such terms, namely

$$f(x) := \frac{1}{n} \sum_{i=1}^{n} [1 - y_i \langle b_i, x \rangle]_+ \tag{50}$$

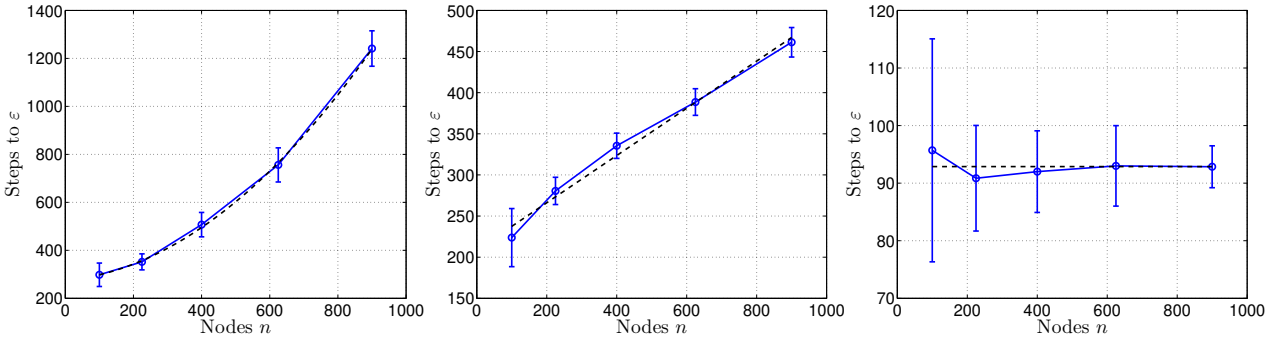



**Figure 3.** Each plot shows the number of iterations required to reach a fixed accuracy $\epsilon$ (vertical axis) versus the network size $n$ (horizontal axis). Each panel shows the same plot for a different graph topology: (a) single cycle; (b) two-dimensional grid; and (c) bounded degree expander. Step sizes were chosen according to the spectral gap, and dotted lines show predictions of Corollary 1.

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 quadratic constraint on the minimization problem (50), and for the simulations considered here, 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 SVM classification problem as follows. For each $i = 1, 2, \ldots, n$, we first draw a random vector $b_i \in \mathbb{R}^d$ from the uniform distribution over the unit sphere. We then randomly generate a random Gaussian vector $w \sim N(0, I_{d\times d})$, and then let $a_i = \text{sign}(\langle w, b_i\rangle)b_i$, randomly flipping the sign of 5% of the $a_i$. Note that these choices yield a function $f$ that is Lipschitz with parameter $L = 1$. Although this is a specific ensemble of problems, we have observed qualitatively similar behavior for other problem ensembles. In order to study the effect of graph size and topology, we perform simulations with three different graph structures, namely cycles, grids, and random 5-regular expanders [FKS89], with the number of nodes $n$ ranging from 100 to 900. In all cases, we use the optimal setting of the step size $\alpha$ specified in Theorem 2 and Corollary 1.

Figure 2 provides plots 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, these plots also show how the convergence time scales as a function of the graph size $n$. In particular, for a given class of optimization problems, define $T_G(\epsilon; n)$ to be the number of iterations required to obtain $\epsilon$-accurate solution for a graph $G$ with $n$ nodes. As shown in Figure 2, for any fixed $\epsilon > 0$, the function $T_G(\epsilon; n)$ shifts to the right as $n$ is increased, and the goal of network scaling analysis is to gain a precise understanding of this shifting.

As discussed following Corollary 1, for cycles, grids, and expanders, we have the following upper bounds on the quantity $T_G(\epsilon; n)$:

$$T_{\text{cycle}}(\epsilon; n) = \mathcal{O}\left(\frac{n^2}{\epsilon^2}\right), \quad T_{\text{grid}}(\epsilon; n) = \mathcal{O}\left(\frac{n}{\epsilon^2}\right), \quad \text{and} \quad T_{\text{expander}}(\epsilon; n) = \mathcal{O}\left(\frac{1}{\epsilon^2}\right) \tag{51}$$

In Figure 3, we compare these 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 blue curve is the average of 20 trials, and the bars show standard errors. For comparison, the dotted black line shows the theoretical prediction (51).

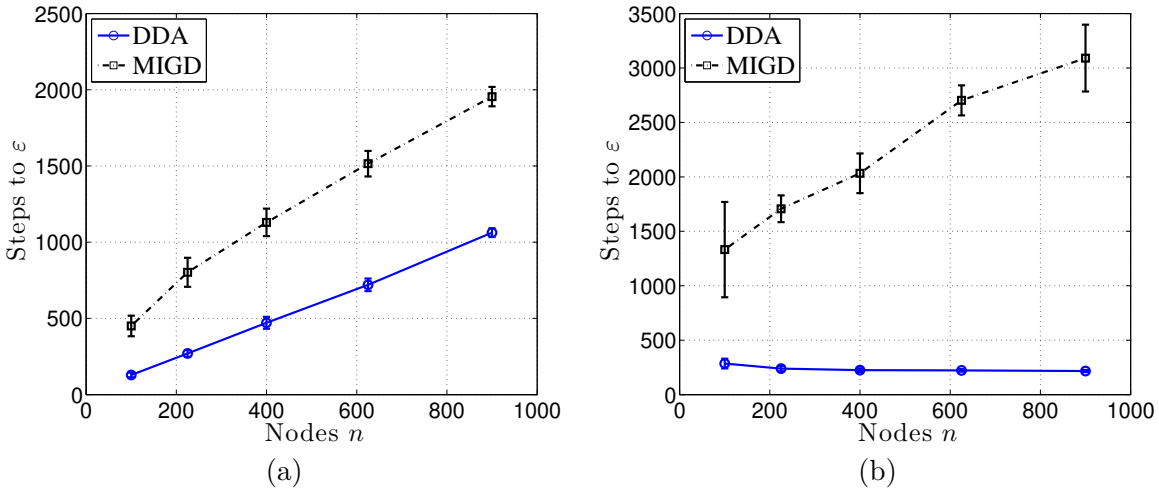



**Figure 4.** Each plot shows the number of iterations required to reach a fixed accuracy $\epsilon$ (vertical axis) versus network size $n$ (horizontal axis) for distributed dual averaging (DDA) and Markov incremental gradient descent (MIGD) [JRJ09]. The panels show the same plot for different graph topologies: (a) two-dimensional grid, (b) bounded degree expander.

Note that the agreement between the empirical behavior and theoretical predictions is excellent 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 having constant network scaling.

Though our focus in this paper is mostly a theoretical one, in our final set of experiments we compare the distributed dual averaging method (DDA) that we present to the Markov incremental gradient descent (MIGD) method [JRJ09] and the distributed projected gradient method [RNV10], which seem to have the sharpest convergence rates currently in the literature. 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 [RNV10] 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 such as the expander in Fig. 4(b)—the dual averaging algorithm gives improved performance.

## 10 Conclusions and Discussion

In this paper, we developed and analyzed a distributed dual averaging algorithm for minimizing the sum of local convex functions over a network. This algorithm is computationally efficient, scalable, and robust to noisy environments, with applications in multi-agent coordination, distributed tracking, localization, sensor networks, and large-scale machine learning. Our analysis provides sharp convergence bounds, showing a direct relationship between the convergence rates and the spectral properties of the underlying network topology, specifically the spectral gap. This allows for a clear distinction between the optimization behavior and the impact of communication constraints.

In addition to handling deterministic communication, our work extends to stochastic communication protocols, such as when communication occurs only along a random subset of the edges

at each round. This flexibility enables tradeoffs between communication overhead and convergence speed. We also analyzed the algorithm's robustness in the presence of noisy gradients, demonstrating its effectiveness in stochastic optimization scenarios. The theoretical predictions of our algorithm were confirmed through both theoretical lower bounds and extensive simulations across different network structures.

Several open questions remain. Future work could explore how similar techniques might be applied to other network-based optimization problems by leveraging different structures for local information aggregation. Additionally, further research could investigate how other classical optimization methods could be adapted for distributed systems to exploit problem structures more efficiently.

# References


[Alo86] N. Alon, *Eigenvalues and expanders*, Combinatorica **6** (1986), 83–96.

[Azu67] K. Azuma, *Weighted sums of certain dependent random variables*, Tohoku Mathematical Journal **68** (1967), 357–367.

[BDTV10] F. Bénézit, A. Dimakis, P. Thiran, and M. Vetterli, *Order-optimal consensus through randomized path averaging*, IEEE Transactions on Information Theory **56** (2010), no. 10, 5150–5167.

[Ber99] D.P. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.

[BGPS06] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE Transactions on Information Theory **52** (2006), no. 6, 2508–2530.

[BT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*, Prentice-Hall, Inc., 1989.

[Chu98] F.R.K. Chung, *Spectral graph theory*, AMS, 1998.

[CV95] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning **20** (1995), no. 3, 273–297.

[DS91] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, The Annals of Probability **1** (1991), no. 1, 36–61.

[DSW08] A. G. Dimakis, A. Sarwate, and M. J. Wainwright, *Geographic gossip: Efficient averaging for sensor networks*, IEEE Transactions on Signal Processing **53** (2008), 1205–1216.

[DW60] G. B. Dantzig and P. Wolfe, *Decomposition principle for linear programs*, Operations Research **8** (1960), 101–111.

[FKS89] J. Friedman, J. Kahn, and E. Szemerédi, *On the second eigenvalue of random regular graphs*, Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing, ACM, 1989, pp. 587–598.

[Fre75] D. A. Freedman, *On tail probabilities for martingales*, The Annals of Probability **3** (1975), no. 1, 100–118.

[GK00] P. Gupta and P. Kumar, *The capacity of wireless networks*, IEEE Transactions on Information Theory **46** (2000), no. 2, 388–404.

[Gra06] R. Gray, *Toeplitz and circulant matrices: A review*, Foundations and Trends in Communications and Information Theory **2** (2006), no. 3, 155–239.

[HJ85] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[HUL96a] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I*, Springer, 1996.

[HUL96b] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II*, Springer, 1996.

[JRJ09] B. Johansson, M. Rabi, and M. Johansson, *A randomized incremental subgradient method for distributed optimization in networked systems*, SIAM Journal on Optimization **20** (2009), no. 3, 1157–1170.

[KV05] A. Kalai and S. Vempala, *Efficient algorithms for online decision problems*, Journal of Computer and System Sciences **71** (2005), no. 3, 291–307.

[LO09] I. Lobel and A. Ozdaglar, *Distributed subgradient methods over random networks*, Tech. Report 2800, MIT LIDS, 2009.

[LOT03] V. Lesser, C. Ortiz, and M. Tambe (eds.), *Distributed Sensor Networks: A Multiagent Perspective*, vol. 9, Kluwer Academic Publishers, 2003.

[LPW08] D. Levin, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, 2008.

[LWHS02] D. Li, K. Wong, Y. Hu, and A. Sayeed, *Detection, classification and tracking of targets in distributed sensor networks*, IEEE Signal Processing Magazine, 2002, pp. 17–29.

[MARS10] D. Mosk-Aoyama, T. Roughgarden, and D. Shah, *Fully distributed algorithms for convex optimization problems*, SIAM Journal on Optimization **20** (2010), no. 6, 3260–3279.

[MHM10] R. McDonald, K. Hall, and G. Mann, *Distributed training strategies for the structured perceptron*, North American Chapter of the Association for Computational Linguistics (NAACL), 2010.

[NB01] A. Nedic and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM Journal on Optimization **12** (2001), no. 1, 109–138.

[Nes04] Y. Nesterov, *Introductory lectures on convex optimization*, Kluwer Academic Publishers, 2004.

[Nes09] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Mathematical Programming A **120** (2009), no. 1, 261–283.

[NO09] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Transactions on Automatic Control **54** (2009), 48–61.

[NOOT09] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *On distributed averaging algorithms and quantization effects*, IEEE Transactions on Automatic Control **54** (2009), no. 11, 2506 –2517.

[NY83] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.

[Pen03] M. Penrose, *Random Geometric Graphs*, Oxford University Press, 2003.

[RN04] M. Rabbat and R. Nowak, *Distributed optimization in sensor networks*, The 3rd International Symposium on Information Processing in Sensor Networks, 2004, pp. 20–27.

[RNV10] S. S. Ram, A. Nedić, and V. V. Veeravalli, *Distributed stochastic subgradient projection algorithms for convex optimization*, Journal of Optimization Theory and Applications **147** (2010), no. 3, 516–545.

[TBA86] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Transactions on Automatic Control **31** (1986), 803–812.

[Tsi84] J. Tsitsiklis, *Problems in decentralized decision making and computation*, Ph.D. thesis, Massachusetts Institute of Technology, 1984.

[vLRH10] U. von Luxburg, A. Radl, and M. Hein, *Hitting times, commute distances, and the spectral gap for large random geometric graphs*, 2010.

[XBK07] L. Xiao, S. Boyd, and S. J. Kim, *Distributed average consensus with least-mean-square deviation*, Journal of Parallel and Distributed Computing **67** (2007), no. 1, 33–46.

[Xia10] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, Journal of Machine Learning Research **11** (2010), 2543–2596.

# A The Dual Averaging Algorithm

In this section, we give a simple convergence proof for the basic (non-distributed) dual averaging algorithm (3). In particular, we recall the updates

$$z(t+1) = z(t) + g(t) \quad \text{and} \quad x(t+1) = \arg\min_{x \in \mathcal{X}} \left\{ \langle z(t+1), x \rangle + \frac{1}{\alpha(t)} \psi(x) \right\}$$

Recall our assumptions without any loss of generality that $0 \in \mathcal{X}$, $\psi \geq 0$ and $\psi(0) = 0$. Let $\mathbb{I}(x \in \mathcal{X})$ be the $\{0, \infty\}$-valued indicator function for membership in $\mathcal{X}$, and for each $\alpha > 0$, let $\psi_\alpha^*$ denote the conjugate dual of the convex function $\frac{1}{\alpha}\psi(x) + \mathbb{I}(x \in \mathcal{X})$. By definition, the conjugate takes the form

$$\psi_\alpha^*(z) = \sup_{x \in \mathcal{X}} \left\{ \langle z, x \rangle - \frac{1}{\alpha} \psi(x) \right\} = \sup_x \left\{ \langle z, x \rangle - \frac{1}{\alpha} \psi(x) - \mathbb{I}(x \in \mathcal{X}) \right\} \tag{52}$$

The definition (4) of the projection $\Pi_{\mathcal{X}}^{\psi}$ shows that the supremum (52) is uniquely attained by $\Pi_{\mathcal{X}}^{\psi}(-z, \alpha)$. Moreover, for any fixed $z$, we note that at $x = 0$, $\langle z, x \rangle - \frac{1}{\alpha}\psi(x) - \mathbb{I}(x \in \mathcal{X}) = 0$. Thus, we can restrict the supremum in (52) to the set

$$\left\{ x \;\middle|\; \frac{1}{\alpha}\psi(x) + \mathbb{I}(x \in \mathcal{X}) - \langle z, x \rangle \leq 0 \right\} = \mathcal{X} \cap \left\{ x \;\middle|\; \frac{1}{\alpha}\psi(x) - \langle z, x \rangle \leq 0 \right\},$$

which is compact since $\mathcal{X}$ is closed and $\psi$ is strongly convex. Thus, since the supremum is uniquely attained and $\langle z, x \rangle$ is differentiable in $z$, $\nabla \psi_\alpha^*(-z) = \Pi_{\mathcal{X}}^{\psi}(z, \alpha)$ [HUL96a, Theorem 4.4.2].

This fact has two important consequences. First, since the projection is Lipschitz-continuous (see Lemma 4), we have the bound

$$\|\nabla \psi_\alpha^*(-z) - \nabla \psi_\alpha^*(-z-g)\| = \left\| \Pi_{\mathcal{X}}^{\psi}(z, \alpha) - \Pi_{\mathcal{X}}^{\psi}(z+g, \alpha) \right\| \leq \alpha \|g\|_*$$

Consequently, an integration argument (e.g., [Nes04, Lemma 1.2.3]) yields the upper bound

$$\psi_\alpha^*(-z-g) \leq \psi_\alpha^*(-z) - \langle g, \nabla \psi_\alpha^*(-z) \rangle + \frac{1}{2}\alpha \|g\|_*^2 \tag{53}$$

The second consequence is that we have

$$x(t) = \nabla \psi_{\alpha(t-1)}^*(-z(t)) = \Pi_{\mathcal{X}}(z(t), \alpha(t-1))$$

## A.1 Proof of Lemma 1

To bound the sequence of inner products, we note that for any $x^* \in \mathcal{X}$, we have

$$-\sum_{t=1}^{T} \langle g(t), x^* \rangle \leq \sup_{x \in \mathcal{X}} \left\{ -\sum_{t=1}^{T} \langle g(t), x \rangle - \frac{1}{\alpha(T)} \psi(x) \right\} + \frac{1}{\alpha(T)} \psi(x^*)$$

$$= \psi^*_{\alpha(T)}(-z(T+1)) + \frac{1}{\alpha(T)}\psi(x^*) \tag{54}$$

By definition of the conjugate function $\psi^*_\alpha$, whenever we have $\alpha(t) \leq \alpha(t-1)$, then we are guaranteed that $\psi^*_{\alpha(t)}(z) \leq \psi^*_{\alpha(t-1)}(z)$ for all $z \in \mathbb{R}^d$. Thus, using the upper bound (53) and the relations $x(t) = \nabla\psi^*_{\alpha(t-1)}(-z(t))$ and $z(t+1) = z(t) + g(t)$, we obtain

$$\begin{aligned}
\psi^*_{\alpha(t)}(-z(t+1)) &\leq \psi^*_{\alpha(t-1)}(-z(t+1)) \\
&= \psi^*_{\alpha(t-1)}(-z(t) - g(t)) \\
&\leq \psi^*_{\alpha(t-1)}(-z(t)) - \langle g(t), x(t)\rangle + \frac{1}{2}\alpha(t-1)\left\|g(t)\right\|_*^2
\end{aligned}$$

Rearranging terms yields

$$\langle g(t), x(t)\rangle \leq \psi^*_{\alpha(t-1)}(-z(t)) - \psi^*_{\alpha(t)}(-z(t+1)) + \frac{1}{2}\alpha(t-1)\left\|g(t)\right\|_*^2 \tag{55}$$

Finally, we combine the upper bound on $\langle g(t), x(t)\rangle$ from equation (55) with the earlier bound (54), thereby obtaining that for any $x^* \in \mathcal{X}$, the sum $S(T) = \sum_{t=1}^T \langle g(t), x(t) - x^*\rangle$ is upper bounded as

$$\begin{aligned}
S(T) &\leq \sum_{t=1}^T \langle g(t), x(t)\rangle + \psi^*_{\alpha(T)}(-z(T+1)) + \frac{1}{\alpha(T)}\psi(x^*) \\
&\leq \frac{1}{2}\sum_{t=1}^T \alpha(t-1)\left\|g(t)\right\|_*^2 + \sum_{t=1}^T \left[\psi^*_{\alpha(t-1)}(-z(t)) - \psi^*_{\alpha(t)}(-z(t+1))\right] + \psi^*_{\alpha(T)}(-z(T+1)) + \frac{1}{\alpha(T)}\psi(x^*) \\
&= \frac{1}{2}\sum_{t=1}^T \alpha(t-1)\left\|g(t)\right\|_*^2 + \frac{1}{\alpha(T)}\psi(x^*)
\end{aligned}$$

The last line exploited the facts that $z(1) = 0$ and $\psi^*_\alpha(0) = 0$. This completes the proof of the claim.

## A.2 Proof of Lemma 2

Via the $L$-Lipschitz continuity of the $f_i$, we can write

$$\begin{aligned}
\sum_{t=1}^T f(x_i(t)) - f(x^*) &= \sum_{t=1}^T f(y(t)) - f(x^*) + \sum_{t=1}^T f(x_i(t)) - f(y(t)) \\
&\leq \sum_{t=1}^T f(y(t)) - f(x^*) + \sum_{t=1}^T L\left\|x_i(t) - y(t)\right\|
\end{aligned}$$

For the second bound, we again use the $L$-Lipschitz continuity of the $f_i$ and the triangle inequality,

$$\begin{aligned}
f(\widehat{x}_i(T)) - f(x^*) &= f(\widehat{y}(T)) - f(x^*) + f(\widehat{x}_i(T)) - f(\widehat{y}(T)) \\
&\leq f(\widehat{y}(T)) - f(x^*) + L\left\|\widehat{x}_i(T) - \widehat{y}(T)\right\| \leq f(\widehat{y}(T)) - f(x^*) + \frac{L}{T}\sum_{t=1}^T \left\|x_i(t) - y(t)\right\|
\end{aligned}$$

Lipschitz-continuity of the projection (Lemma 4) shows that $\left\|x_i(t) - y(t)\right\| \leq \alpha(t)\left\|\overline{z}(t) - z_i(t)\right\|_*$ which gives both the desired results.

### A.3 Lipschitz continuity of projections

The following lemma on the Lipschitz-continuity of the projection operator is well-known, but we state and prove it for completeness.

**Lemma 4.** *For an arbitrary pair $u, v \in \mathbb{R}^d$, we have*

$$\left\| \Pi_{\mathcal{X}}^{\psi}(u, \alpha) - \Pi_{\mathcal{X}}^{\psi}(v, \alpha) \right\| \leq \alpha \left\| u - v \right\|_*$$

**Proof** Lemma 4 is essentially an immediate consequence of the relationship between strong-convexity and Lipschitz continuity of the gradient for conjugate functions [HUL96b, Theorem X.4.2.1], but we give a short proof for completeness. For an arbitrary pair $u, v \in \mathbb{R}^d$, denote $w = \Pi_{\mathcal{X}}^{\psi}(u, \alpha)$ and $x = \Pi_{\mathcal{X}}^{\psi}(v, \alpha)$. By the first-order optimality conditions for convex minimization, for any $y \in \mathcal{X}$, we have

$$\left\langle u + \frac{1}{\alpha} \nabla \psi(w), y - w \right\rangle \geq 0 \quad \text{and} \quad \left\langle v + \frac{1}{\alpha} \nabla \psi(x), y - x \right\rangle \geq 0$$

Setting $y = x$ and $y = w$ in these two inequalities (respectively) yields

$$\langle \alpha u + \nabla \psi(w), x - w \rangle \geq 0 \quad \text{and} \quad \langle \alpha v + \nabla \psi(x), w - x \rangle \geq 0$$

Adding the above two inequalities, we obtain the bound

$$\langle \nabla \psi(w) - \nabla \psi(x), w - x \rangle \leq \alpha \langle u - v, x - w \rangle \leq \alpha \left\| v_1 - v_2 \right\|_* \left\| w - x \right\| \tag{56}$$

On the other hand the strong convexity of $\psi$ implies that $\psi(w) \geq \psi(x) + \langle \psi(x), w - x \rangle + \frac{1}{2} \left\| x - w \right\|^2$, with an analogous bound with the roles of $x$ and $w$ exchanged. Some algebra then leads to

$$\langle \nabla \psi(w) - \nabla \psi(x), w - x \rangle \geq \left\| w - x \right\|^2$$

which, when combined with (56), gives the desired result. □

## B Background on stochastic matrices

In this section, we briefly review some well-known properties of stochastic matrices; we refer the reader to Chapter 8 of Horn and Johnson [HJ85] for additional detail. For an $n \times n$ matrix $A$, we let its singular values be $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A)$, and for a real symmetric $A$, we define the eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$. Let $\mathbb{1}$ be the all ones vector. In our setting, $P = [p_1 \ \cdots \ p_n] \in \mathbb{R}^{n \times n}$ is a doubly stochastic matrix, so that $P\mathbb{1} = \mathbb{1}$ and $\mathbb{1}^T P = \mathbb{1}^T$. We have $\sigma_1(P) = 1$, $\lambda_1(P^T P) = 1$, and $1 - \sigma_2(P)$ is the spectral gap, which is known to determine the mixing properties of the Markov chain induced by $P$ [LPW08].

In order to establish the connection between mixing and spectral gap, define the uniform matrix $F := \frac{1}{n} \mathbb{1}\mathbb{1}^T$. Observe that $F$ is idempotent ($F^2 = F$), and moreover it satisfies $PF = FP = F$. By construction, the eigenspectrum of $P - F$ is equal to that of $P$ except that the largest eigenvalue 1 is removed. Similarly, the eigenspectrum of $(P - F)^T (P - F) = P^T P - F^T P - P^T F + F^T F = P^T P - F^T F$ is identical to that of $P^T P$ but with $\lambda_1(P^T P) = 1$ removed. Given these properties, a

simple calculation yields that for any integer $t = 1, 2, \ldots$, we have $(P-F)^t = P^t - F$. Consequently, for any $x \in \mathbb{R}^n$, we have

$$\left\|P^t x - Fx\right\|_2 = \left\|(P-F)^t x\right\|_2 \leq \sigma_1(P-F)\left\|(P-F)^{t-1}x\right\|_2 \leq \quad \cdots \quad \leq \sigma_2(P)^t \left\|x\right\|_2$$

If we take $x = e_i$, denoting a canonical basis vector for $i = 1, \ldots, n$, then we see that $\left\|P^t - L\right\|_\infty \leq \sigma_2(P)^t$. Taking $x \in \Delta_n$, the $n$-dimensional simplex, gives

$$\left\|P^t x - \mathbb{1}/n\right\|_{\mathrm{TV}} = \frac{1}{2}\left\|P^t x - \mathbb{1}/n\right\|_1 \leq \frac{1}{2}\sqrt{n}\left\|P^t x - \mathbb{1}/n\right\|_2 \leq \frac{1}{2}\sigma_2(P)^t\sqrt{n}$$

which establishes the bound (23). (The $\sqrt{n}$ factor in the bound is standard in the Markov chain literature, e.g., [DS91, Proposition 3].)

# C Eigenvalues of paths

Let $G$ be a graph and $S$ be a subset of the nodes in the graph. Let $E(S, S^c)$ denote the set of edges crossing between $S$ and $S^c$, and let the volume of $S$ be the sum of the degrees of the nodes in $S$, that is, $\mathrm{vol}(S) = \sum_{i \in S} \delta_i$. The Cheeger constant of a graph $G$ is defined as

$$h_G \ := \min_{S \subset V} \frac{\mathrm{card}(E(S, S^c))}{\min\{\mathrm{vol}(S), \mathrm{vol}(S^c)\}} \tag{57}$$

If $\mathcal{L}$ is the Laplacian of $G$, then $2h_G \geq \lambda_{n-1}(\mathcal{L}) > \frac{1}{2}h_G^2$ (e.g., see Lemma 2.1 and Theorem 2.2 in Chung [Chu98]).

**Lemma 5.** *Let $G$ be a $k$-connected path with $n$ nodes and $k \leq \sqrt{n}$. Then its normalized graph Laplacian $\mathcal{L}$ satisfies $\lambda_{n-1}(\mathcal{L}) = \Theta(k^2/n^2)$.*

**Proof** We invoke Theorem 4.13 in Chung [Chu98] to conclude that $\lambda_{n-1}(\mathcal{L}) = \mathcal{O}(k^2/n^2)$, since $G$ is a subgraph of the $k$-connected cycle. It thus suffices to prove that the Cheeger constant is lower bounded as $h_G = \Omega(k/n)$.

Let $S$ be the set of nodes achieving the minimum in the definition (57). To make the rest of the proof easier, assume that the degree of each node is $2k$. (We may do so without loss of generality, since it only has the effect of increasing $\mathrm{vol}(S)$ and $\mathrm{vol}(S^c)$ in the Cheeger constant calculation, and so any Cheeger constant calculated under this assumption lower bounds the true Cheeger constant.)

First, note that one of the nodes in $S$ must be against the end of the path—if not, shifting the nodes in $S$ in one direction (taking into account that we must pick the direction in which more nodes are brought near the end of the path) can only decrease $\mathrm{card}(E(S, S^c))$. Now we show that all of the nodes in $S$ must be directly adjacent to one another. Suppose the nodes are not adjacent. Since $k \leq \sqrt{n}$, there must be a pair of nodes in $S$ with a distance of at least $k$. Let $i \in S^c$ be between those two nodes, and let $S_\ell$ denote the nodes to the left of $i$ and $S_r$ the nodes to the right. Collapsing all the nodes in $S_r$ to the rightmost end of the path and all the nodes in $S_\ell$ to the leftmost end can only decrease $\mathrm{card}(E(S, S^c))$. If $|S| \geq k$, then at least one of the sets $S_r$ and $S_\ell$ shares $k(k-1)/4$ edges with $S^c$. Otherwise, if $|S| < k$, then $\mathrm{card}(E(S, S^c)) \geq k$ and $\mathrm{vol}(S) \leq k^2$, so $\mathrm{card}(E(S, S^c))/\mathrm{vol}(S) \geq 1/k$. Under the assumption $k^2 \leq n$, we have $1/k \leq k/n$, from which the result follows. □

# D Composite Objectives

In this section, we show how to generalize the dual averaging algorithm to incorporate composite objectives, specifically those of the form $f+\varphi$ for known $\varphi$. Though it is possible to perform similar derivations to those in Lemma 1, for brevity we refer to recent work of Xiao [Xia10]. Nonetheless, the algorithm is conceptually very similar to the dual averaging algorithm (updates (5a) and (5b)), and equally as simple to write. We assume that $\varphi$ is closed convex and non-negative, and $\mathcal{X}$ is closed. We define the composite projection operator $\Pi^t_{\mathcal{X}}$ as

$$\Pi^t_{\mathcal{X}}(z) = \operatorname*{arg\,min}_{x\in\mathcal{X}} \left\{ \langle z, x\rangle + t\varphi(x) + \frac{1}{\alpha(t)}\psi(x) \right\} \tag{58}$$

The mapping $\Pi^t_{\mathcal{X}}$ is $\alpha(t)$-Lipschitz with respect to $\|\cdot\|$ and $\|\cdot\|_*$, that is,

$$\left\|\Pi^t_{\mathcal{X}}(z_1) - \Pi^t_{\mathcal{X}}(z_2)\right\| \le \alpha(t)\left\|z_1 - z_2\right\|_* \tag{59}$$

As in Lemma 4, (59) is a consequence of the fact that the conjugate dual of a $1/\alpha(t)$-strongly convex function has $\alpha(t)$-Lipschitz continuous gradient with respect to the associated dual norm, and the gradient of the conjugate of $t\varphi(x) + \frac{1}{\alpha(t)}\psi(x)$ is simply $\Pi^t_{\mathcal{X}}(z)$ [HUL96b, Theorem X.4.2.1].

The distributed algorithm based on the update (58) is essentially identical to the dual averaging algorithm discussed in the main body of the paper. Each agent $i$ maintains the gradient vector

$$z_i(t+1) = \sum_{j=1}^n p_{ij}(t) z_j(t) - g_i(t) \quad \text{where} \quad \mathbb{E}g_i(t) \in \partial f_i(x_i(t)) \tag{60}$$

The update to $x_i(t+1)$ is then

$$x_i(t+1) = \Pi^t_{\mathcal{X}}(-z_i(t+1)) \tag{61}$$

As in (16), we have $\overline{z}(t+1) = \overline{z}(t) - \frac{1}{n}\sum_{i=1}^n g_j(t)$. The following proposition, a simplification of [Xia10, Section B.2], allows us to give a convergence guarantee for the algorithm described by (60) and (61).

**Proposition 2.** *Let $\alpha(t)$ be a decreasing sequence and $g(t) \in \mathbb{R}^d$ be an arbitrary sequence of vectors. If $x(t+1) = \Pi^t_{\mathcal{X}}(\sum_{\tau=1}^t g(t))$, then for any $x^* \in \mathcal{X}$,*

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

The above proposition, combined with the techniques used to derive Theorem 1, allow us to easily prove convergence of distributed composite-objective dual averaging. As earlier, let $y(t) = \Pi^t_{\mathcal{X}}(-\overline{z}(t))$, and assume that the $f_i$ are $L$-Lipschitz with respect to $\|\cdot\|$. Then as in (18), (19), and (20), for any $x^* \in \mathcal{X}$, we immediately have

$$\begin{aligned}
&\sum_{t=1}^T \left[f(y(t)) + \varphi(y(t)) - f(x^*) - \varphi(x^*)\right] \\
&\le \sum_{t=1}^T \frac{1}{n}\left\langle \sum_{i=1}^n g_i(t), y(t) - x^* \right\rangle + \sum_{t=1}^T \varphi(y(t)) - \varphi(x^*) + \sum_{t=1}^T\sum_{i=1}^n \frac{2L}{n}\left\|y(t) - x_i(t)\right\|
\end{aligned}$$

By definition of $y(t)$, we see that Proposition 2 bounds the above by

$$\frac{1}{\alpha(T)}\psi(x^*) + \frac{1}{2}\sum_{t=1}^{T}\alpha(t-1)L^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\frac{2L}{n}\left\|y(t) - x_i(t)\right\|$$

Finally, we use the fact that the mapping $\Pi_{\mathcal{X}}^t$ is $\alpha(t)$-Lipschitz to see that for the distributed composite-objective projection algorithm of (60) and (61),

$$\sum_{t=1}^{T} f(y(t))+\varphi(y(t))-f(x^*)-\varphi(x^*) \leq \frac{1}{\alpha(T)}\psi(x^*)+\frac{1}{2}\sum_{t=1}^{T}\alpha(t-1)L^2+\frac{2L}{n}\sum_{t=1}^{T}\alpha(t)\sum_{i=1}^{n}\left\|\overline{z}(t) - z_i(t)\right\|_* \tag{62}$$

Any of the techniques in the prequel can be used to bound (62).