Abstract—We consider whether distributed subgradient methods can achieve a linear speedup over a centralized subgradient method. While it might be hoped that distributed network of $n$ nodes that can compute $n$ times more subgradients in parallel compared to a single node might, as a result, be $n$ times faster, existing bounds for distributed optimization methods are often consistent with a slowdown rather than speedup compared to a single node.

We show that a distributed subgradient method has this “linear speedup” property when using a class of square-summable-but-not-summable step-sizes which include $1/t^β$ when $β \in (1/2, 1)$; for such step-sizes, we show that after a transient period whose size depends on the spectral gap of the network, the method achieves a performance guarantee that does not depend on the network or the number of nodes. We also show that the same method can fail to have this “asymptotic network independence” property under the optimally decaying step-size $1/\sqrt{t}$ and, as a consequence, can fail to provide a linear speedup compared to a single node with $1/\sqrt{t}$ step-size.

I. INTRODUCTION

We consider the standard setting of distributed convex optimization: $f_1(x), \ldots, f_n(x)$ are convex functions from $\mathbb{R}^d$ to $\mathbb{R}$, with node $i$ of the network the only node which can compute subgradients of the function $f_i(x)$. The goal is to compute a minimizer

$$x^* \in \arg \min_{x \in \Omega} \left( F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right), \quad (1)$$

where $\Omega$ is a closed convex set. The underlying method must be decentralized, relying only on local subgradient computations and peer-to-peer message exchanges. In particular, we will consider the “standard model” of distributed optimization where at each step, node $i$ computes a subgradient of its local function, possibly performs a projection step onto the set $\Omega$, and broadcasts a message to its neighbors.

This problem was first analyzed in [6], where a distributed subgradient method was proposed for the unconstrained case when $\Omega = \mathbb{R}^n$. The case with the constraint $\Omega$ was first analyzed in [7]. Both papers proposed methods inspired by the “average consensus” literature, where nodes mix subgradient steps on their local functions with consensus steps which move them in the direction of their neighbors.

Distributed optimization methods have attracted considerable attention since the publication of [6] for several reasons. First, many problems in multi-agent control involve nodes acting to maximize a global objective from local information, and Eq. (1) is thought to be among the simplest problems of this type. Second, empirical loss minimization in machine learning reduces exactly to Eq. (1) (see the discussion in Section I of [10]) and it is hoped that solving such problems in a distributed setup might result in speed-ups.

Over the past decade, thousands of papers have been written on different variations of this problem, and it would be impossible to survey all this related work; instead, we refer the reader to the recent survey [5]. Instead, we launch into a discussion of the main motivating concern of this paper, namely how the performance of distributed optimization methods compares to their centralized counterparts. We begin by discussing the available guarantees for the centralized subgradient method, so that we can contrast those guarantees to the available distributed bounds in our survey of previous work, which will follow.

A. The subgradient method

The (centralized) projected subgradient method run on the function $F(x)$ takes the form

$$y(t + 1) = P_\Omega \left[ y(t) - \alpha(t) g_F(t) \right],$$

where $g_F(t)$ is a subgradient of the function $F(\cdot)$ at $y(t)$, and $P_\Omega$ is the projection onto $\Omega$.

The standard reference for an analysis of this method is the set of lecture notes [1]. It is usually assumed that $\|g_F(t)\|_2 \leq L$ for all $t$, i.e., all subgradients are bounded; and $\Omega$ is assumed to have diameter at most $D$. The function $F(x)$ may have more than one minimizer over $\Omega$; we select one minimizer arbitrarily and call it $x^*$.

The step-size $\alpha(t)$ needs to be properly chosen. There are two choices that are typically analyzed in this setting. One is to set $\alpha(t) = 1/\sqrt{t}$, which turns out to be the optimal decay rate. The other is to choose $\alpha(t)$ to be “square summable but not summable” as in the following assumption.

Assumption 1. The sequence $\alpha(t)$ satisfies

$$\sum_{t=1}^{+\infty} \alpha^2(t) < \infty \quad \text{and} \quad \sum_{t=1}^{+\infty} \alpha(t) = +\infty$$

We now briefly summarize the standard analysis of the method from [1], which the reader can consult for details. The analysis is based on the following recurrence relation, to the

\begin{align*}
    P_\Omega \left[ y(t) - \alpha(t) g_F(t) \right] &\leq y(t) - \alpha(t) g_F(t) + \alpha(t) \max_{z \in \Omega} \left\{ g_F(z) - g_F(t) \right\} \\
    &\leq y(t) - \alpha(t) g_F(t) + \alpha(t) \max_{z \in \Omega} \left\{ L(t) - L(t-1) \right\} \\
    &\leq y(t) - \alpha(t) g_F(t) + \alpha(t) L(T) \\
    &= y(t) - \alpha(t) g_F(t) + \alpha(t) L(t).
\end{align*}
effect that, up to second order terms, the method gets closer to the set of minimizers at every step:

$$||y(k+1) - x^*||_2^2 \leq ||y(k) - x^*||_2^2 - 2\alpha(k)(F(y(k)) - F^*) + L^2\alpha^2(k),$$

It is standard to re-arrange this into a telescoping sum as

$$2\alpha(k)(F(y(k)) - F^*) \leq ||y(k) - x^*||_2^2 - ||y(k + 1) - x^*||_2^2 + L^2\alpha^2(k), \quad (2)$$

and sum it up over \(k = 1, \ldots, t\). Indeed, defining

$$y_\alpha(t) := \sum_{k=1}^{t} \frac{\alpha(k) y(k)}{\sum_{k=1}^{t} \alpha(k)}$$

summing up Eq. (2) and appealing to the convexity of \(F(x)\) we can obtain that

$$F(y_\alpha(t)) - F^* \leq \frac{D^2 + L^2 \sum_{k=1}^{t} \alpha^2(k)}{2 \sum_{k=1}^{t} \alpha(k)}, \quad (3)$$

where \(||y(0) - x^*||_2^2 \leq D^2\) (as \(\Omega\) was assumed to have diameter \(D\)). Finally, by Assumption 1 the right-hand side goes to zero, and so we obtain that the subgradient method works. We remark again that the details of this argument can be found in any source on the subject, in particular in [1].

A variation on this argument can get rid of the dependence on \(L\) in Eq. (3). This requires the following assumption.

**Assumption 2.** There is a constant \(C_\alpha\) such that for all positive integers \(t\),

$$\sum_{k=1}^{t} \alpha(k) \leq C_\alpha \sum_{k=\lceil t/2 \rceil}^{t} \alpha(k).$$

This assumption can be motivated by observing that it is satisfied by step-sizes that decay polynomially as \(\alpha(t) = 1/t^\beta\).

With this assumption in place, one can set \(t' = \lfloor t/2 \rfloor\) and instead sum Eq. (2) from \(t'\) to \(t\). Defining the running average from time \(t'\) to \(t\) as

$$y'_\alpha(t) := \sum_{k=t'}^{t} \frac{\alpha(k) y(k)}{\sum_{k=t'}^{t} \alpha(k)}$$

this immediately yields the following proposition.

**Proposition I.1.** Suppose Assumptions 1 and 2 on the step-size are satisfied, \(F(x)\) is a convex function whose subgradients are upper bounded by \(L\) in the Euclidean norm, and \(t\) is large enough so that we have the upper bound

$$\sum_{k=\lfloor t/2 \rfloor}^{t} \alpha^2(k) \leq \frac{D^2}{L^2}.$$ 

Then

$$F(y'_\alpha(t)) - F^* \leq \frac{D^2 C_\alpha}{\sum_{k=1}^{t} \alpha(k)}. \quad (4)$$

This result has no dependence on \(L\), but at the expense of multiplying the dependence on \(D\) by the constant \(C_\alpha\). For example, if \(\alpha(t) = 1/t^{3/4}\), it is an exercise to verify that one can take \(C_\alpha = 6\). We note that since the step-size \(\alpha^2(t)\) is square summable, Eq. (4) is guaranteed to hold for large enough \(t\).

The bound of this proposition suggests to take \(\alpha(t)\) decaying as slowly as possible (so that \(\sum_{k=1}^{t} \alpha(k)\) grows as fast as possible) while still keeping \(\alpha(t)\) square summable but not summable. There is no optimal choice, but in general one wants to pick \(\alpha(t) = 1/t^\beta\) where \(\beta\) is close to \(1/2\), but not \(1/2\) since \(\alpha(t) = 1/\sqrt{t}\) is not square summable. The result will be a decay rate of \(F(y'_\alpha(t)) - F^* = O(t^{1/2 - \beta})\).

One can redo the above argument with the rate of decay of \(\alpha(t) = 1/\sqrt{t}\) to obtain an optimal rate of decay. In that case, because this is not a square summable step-size, the dependence on \(L\) cannot be avoided. However, since \(\sum_{k=1}^{t} \alpha(k) = O(1)\), we can simply repeat all the steps above to give the bound

$$F(y_\alpha(t)) - F^* \leq O \left( \frac{D^2 + L^2}{\sqrt{t}} \right), \quad (5)$$

One can also choose \(\alpha(t)\) depending on the constants \(D\) and \(L\) to obtain better scaling with respect to those constants; however, in this paper, we will essentially be restricting our attention to unoptimized step-sizes of the form \(\alpha(t) = 1/t^\beta\).

We next compare these results for the centralized subgradient to available convergence times in the distributed case.

**B. Convergence times of distributed subgradient methods**

A number of distributed subgradient methods have been proposed in the literature, with the simplest being

$$x(t + 1) = W(x(t) - \alpha(t)g(t)), \quad (6)$$

which was analyzed in [6]. Here \(x(t)\) is an \(n \times d\) matrix, with the \(i\)th row of \(x(t)\) being controlled by agent \(i\); we will use \(x_i(t)\) to denote the same \(i\)th row. The matrix \(g(t)\) is also \(n \times d\) and it’s \(i\)th row, which we will denote by \(g_i(t)\), is a subgradient of the function \(f_i(x)\) at \(x = x_i(t)\). The matrix \(W\) is doubly stochastic and needs to satisfy some connectivity and non-aperiodicity conditions; it suffices to assume that \(W\) has positive diagonal and the directed graph corresponding to the positive entries of \(W\) is strongly connected.

It was shown in [5] that, for small enough constant stepsize \(\alpha(t) = \alpha\), this method converges in an error that scales linearly in \(\alpha\). The projected version

$$x(t + 1) = P_{\Omega_i} [Wx(t) - \alpha(t)g(t)], \quad (7)$$

was studied in [7]: here the projection operator \(P_{\Omega_i}\) acts independently on each row of the matrix, \(g(t)\) is composed of subgradients evaluated at \(Wx(t)\). It was shown that, under an appropriately decaying step-size, this scheme results convergence to an optimal solution.

A number of follow-up articles analyzed variations on projected gradient/subgradient methods in the distributed setting. A primal-dual approach was explored in [15] and more recently in [13]. Using the dual subgradient method instead of the ordinary subgradient method was studied in [2]. An analysis that applies to the stochastic case was given in [12]. Methods applying to non-identical constraints and communication delays were studied in [3]. How finite-time convergence may be achieved (in continuous time) was analyzed in [3]. A control inspired approach based on log-barrier functions
was proposed in [14]. In [11] a continuous-time approach was proposed which used the differences $P_\Omega(x_i(t)) - x_i(t)$ along with the subgradients as inputs to drive the system. A convergence time analysis was given in [4], but under the assumption that the function $F(x)$ is strongly convex.

Our interest is in the convergence rate of these methods; in particular, we want to see if the parallelization inherent in having $n$ nodes query subgradients at the same time helps convergence. A useful benchmark is the consider a single node, which knows all the functions $f_i(x), i = 1, \ldots, n$, and can compute the gradient of one of these functions at every time step. We will call the rate obtained in this setup by performing full-batch subgradient descent (i.e., by computing the gradient of $F(x)$ by querying the subgradients of $f_1(x), \ldots, f_n(x)$ in $n$ steps) the single-node rate. The single node rate consists in multiplying all the rates obtained in the previous section by $n$, consistent with $n$ steps to compute a single subgradient of $F(x)$. For example, the bound of Proposition 1 becomes

$$F(y_n(t)) - F^* \leq \frac{nD^2C_\alpha}{\sum_{k=1}^n \alpha(k)}.$$

Ideally, one hopes for a factor $n$ speedup over the single node rate, since the $n$-node network can compute $n$ subgradients in parallel at every step. This corresponds to a convergence time that removes the factor of $n$ from the last equations.

Most of the existing convergence analyses do not attempt to write out all the scalings for the convergence times of distributed optimization methods; many papers write out the scaling with $t$ but do not focus on scaling with the number of nodes. Unfortunately, once those scalings are traced out within the course of the proof, they tend to scale with $(1 - \sigma)^{-1}$, where $\sigma$ is the second-largest singular value associated with the matrix $W$. The quantity $(1 - \sigma)^{-1}$ can scale as much as $O(n^2)$ in the worst-case over all graphs [5], so the underlying scaling is actually worse than the single-node rate.

A concrete examples of this comes from the survey paper [5], where a worse case rate is explicitly written out. The unconstrained case is studied, with step-size $\alpha = 1/\sqrt{T}$ and the algorithm is run for $T$ steps. It is shown in [5] that

$$F(y_n(t)) - F^* \leq O\left(\frac{D^2 + L^2(1-\sigma)^{-1}}{\sqrt{T}}\right)$$

Comparing this with Eq. 5, we see that, in the worst case when $(1 - \sigma)^{-1} \approx \Theta(n^2)$, this is a factor of $n$ slower than the single node rate – in spite of the fact that the network can compute $n$ gradients in parallel. Similar issues affect all the upper bounds in this setting that have been derived in the previous literature, in particular the bounds derived in [2] for dual subgradient, in [13] for the standard setting of distributed optimization where a single message exchange in neighbors is possible per step, and those implicit in [7] for square-summable-but-not-summable step-sizes.

In the paper [9], it was shown how, for a particular way to choose the matrix $W$ in a distributed way, it is possible to replace the $(1 - \sigma)^{-1}$ with an $O(n)$ factor, matching the single-node rate. The idea was to use Nesterov acceleration, which allows one to replace $(1 - \sigma)^{-1}$ with $(1 - \sigma)^{-1/2}$, and argue that for a certain particularly chosen weights the latter quantity is $O(n)$. However, this required slightly stronger assumptions (namely, knowing either the total number of nodes or a reasonably accurate upper bound on it; in general, it requires knowing something about the spectral gap). While this does not offer a speedup over the single-node rate, at least it matches it.

Finally, we mention that our paper is closest to the recent work [8], which is also concerned with the very same question, and gives bounds that seek to isolate the effect of the graph topology.

C. Our contributions

We will analyze a minor variation of Eq. (6) and Eq. (7):

$$x(t + 1) = WP_\Omega [x(t) - \alpha(t)g(t)],$$

This is slightly more natural than Eq. (7), since $g(t)$ here is the subgradient evaluated at $x(t)$, and not at $Wx(t)$ as in Eq. (7). This makes analysis somewhat neater.

Our main result will be to show that a linear speedup is achieved by this iteration on a class of square-summable-but-not-summable stepsizes which include $\alpha(t) = 1/t^\beta$ with $\beta \in (1/2,1)$. This is done by showing that, provided $t$ is large enough, we can give a performance bound that does not depend on $(1 - \sigma)^{-1}$, i.e., is network independent. We will also show that the same assertions fail for the optimally decaying step-size $\alpha(t) = 1/\sqrt{t}$.

We next give a formal statement of our main results. First, let us state our assumptions formally as follows.

Assumption 3. Each function $f_i(x) : \mathbb{R}^d \to \mathbb{R}$ is a convex with all of its subgradients bounded by $L$ in the Euclidean norm. Moreover, the set $\Omega$ is a closed convex set. Each node begins with an identical initial condition $x_i(0) \in \Omega$.

Assumption 4. The matrix $W$ is nonnegative, doubly stochastic, and with positive diagonal. The graph corresponding to the positive entries of $W$ is strongly connected.

Secondly, we will be making an additional assumption on step-size, motivated by the fact that it holds for step-sizes of the form $\alpha(t) = 1/t^\beta$.

Assumption 5. The sequence $\alpha(t)$ is nonincreasing and there is a constant $C_{\alpha'}$ such that

$$\alpha([t/2]) \leq C_{\alpha'} \alpha(t).$$

This assumption essentially bounds how much $\alpha(t)$ can decrease over the period of $t/2, \ldots, t$.

Finally, let us introduce the notation

$$\pi(t) = \frac{1}{n} \sum_{i=1}^n x_i(t),$$

for the average of the iterates at time $t$. We adopt a general convention that, for a vector or a matrix, putting an overline will mean referring to the average of the rows.

Similarly to what was done in the previous subsection, we define

$$x_\alpha(t) = \frac{\sum_{k=\tau}^t \alpha(k)\pi(k)}{\sum_{k=\tau}^t \alpha(k)}$$
Our first main result is the following theorem.

**Theorem I.2** (Asymptotic Network Independence with Square Summable Step-Sizes). Suppose Assumptions 7, 2 and 5 on the step-size, Assumption 3 on the functions, and Assumption 4 on the mixing matrix $W$ all hold.

Then if $t$ is large enough so that the tail sum satisfies the upper bound

$$
\sum_{k=\lceil t/2 \rceil}^{+\infty} \alpha(t) \leq \frac{D^2(1-\sigma)}{10C_\alpha^2 L^2}
$$

and also

$$
t \geq \Omega \left( \frac{1}{1-\sigma} \log \left( (1-\sigma) t \sigma_{\text{max}} / (C_\alpha \sigma(t)) \right) \right)
$$

we have the network-independent bound

$$
F(x'_\alpha(t)) - F^* \leq \frac{D^2 C^* \alpha}{\sum_{k=0}^t \alpha(k)}
$$

In particular, if $\alpha(t) = 1/t^\beta$ where $\beta$ lies in the range $(1/2, 1)$, then when $t$ additionally satisfies

$$
t^{2\beta-1} \geq \Omega \beta \left( \frac{L^2}{D^2(1-\sigma)} \right)
$$

we have the network-independent bound

$$
F(x'_\alpha(t)) - F^* \leq O \left( \frac{D^2}{1+t^{-\beta}} \right),
$$

where the subscript of $\beta$ denotes that the constants in the $O(\cdot)$ and $\Omega(\cdot)$ notation depend on $\beta$.

At the risk of being repetitive, we note that the performance guaranteed by this theorem is asymptotically network independent, as the only dependence on the spectral gap $1 - \sigma$ is in the transient. The point of the theorem is to contrast Eq. (11) with Proposition (11). The two guarantees are within constant factors of each other, which implies that the distributed optimization method analyzed in Theorem 1.2 gives us a linear time speedup over the single-node rate using the same-step size. Likewise, Eq. (12) gives a network-independent bound (though, again, the size of the transient until it holds depends on the network), and can be thought of as a linear-time speedup over the corresponding single-node rate.

Such linear speedups are significant in that they provide a strong motivation for distributed optimization: one can claim that, over a network with $n$ nodes, the distributed optimization is $n$ times faster than a centralized one, at least provided $t$ is large enough.

Note that the lower bound of Eq. (11) is not fully explicit, as $t$ actually appears on both sides. However, for large enough $t$ the inequality always holds, as otherwise $\alpha(t) \leq C_1 t e^{-C_2 t}$ for all $t$ where $C_1, C_2$ depend on $1 - \sigma, \sigma_{\text{max}}$ and $C_\alpha$; and this would contradict the non-summability of $\alpha(t)$ in Assumption 1.

Unfortunately, this theorem does not apply to $\beta = 1/2$ which, as discussed earlier, is the best node of decay for the subgradient method. In fact, our next theorem will show something quite different occurs when $\beta = 1/2$: we will construct a counterexample where the distributed method has network dependent performance regardless of how large $t$ is.

We next give a formal statement of this result. Our first step is to describe how we will choose the matrix $W$ depending on the underlying graph. Let us adopt the convention that, given an undirected graph $G = (V,E)$ without self-loops, we will define the symmetric stochastic matrix $W_{G,\epsilon}$ as

$$
[W_{G,\epsilon}]_{ij} = \begin{cases} \epsilon & (i,j) \in E \\ 0 & \text{else} \end{cases},
$$

and we set diagonal entries $[W_G]_{ii}$ to whatever values result in a stochastic matrix. Clearly, $\epsilon$ should be smaller than the largest degree in $G$.

We next define $G'_n$ to be a graph on $2n$ nodes obtained as follows: two complete graphs on nodes $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are joined by connecting $u_i$ to $v_i$. We note that because the largest degree in this graph is $n + 1$, any $\epsilon$ used to construct a stochastic $W_{G,\epsilon}$ should be upper bounded by $1/(n + 1)$.

Our second main result shows that when we run Eq. (9) on this graph $G'_n$, then with an appropriate choice of functions we will never obtain a performance independent $\epsilon^{-1}$; and since, as we remarked, $\epsilon^{-1}$ grows as $\Omega(n)$, the performance will always scale with $n$ no matter how long we wait.

**Theorem I.3** (Lack of Asymptotic Network Independence with $1/\sqrt{t}$ Step-Size). Consider the distributed optimization method of Eq. (9) with

- The functions $f_i(x) = \gamma |x|$, when $i \in \{u_1, \ldots, u_n\}$ and $f_i(x) = \frac{1}{2} |x - 1|$, for $i \in \{v_1, \ldots, v_n\}$
- Step-size $\alpha(t) = 1/\sqrt{t}$
- Constraint set $\Omega = [-a,a]$
- Initial conditions $x_i(0) = 0.$

Then, there exists a choice of the constants $\gamma > 1$ and an independent of $n$ or $\epsilon$ such that, on the graph $G'_n$ for any choice of $\epsilon$ sufficiently small, there exists an infinite sequence $g_i(t)$ such that:

- $g_i(t)$ is a subgradient of $f_i(x) = x_i(t)$
- For $i \in \{v_1, \ldots, v_n\}$, $x_i(t) - x^*$ is a nonnegative sequence that does not depend on $i$ and satisfies

$$
x_i - x^* = \Omega \left( \frac{\epsilon^{-1}}{\sqrt{t}} \right), \text{ for all } i \in \{v_1, \ldots, v_n\},
$$

for all large enough $t$.

- $x_i(t) = x^*$ for all $i \in \{u_1, \ldots, u_n\}$ and all $t$.

Observe that for the distributed optimization problem discussed in this theorem, we have that $x^* = 0$. An immediate consequence is that not only does the average $(1/n) \sum_{i=1}^n x_i(t) - x^*$ scale as $O(\epsilon^{-1}/\sqrt{t})$ but so does any convex combination over various $t$'s (in particular, the quantities $x_\alpha(t)$ or $x'_\alpha(t)$ discussed earlier). It then follows
that $F(\cdot) - F^*$ for all of these quantities also scales linearly with $\epsilon^{-1}$. In particular, since $\epsilon \leq 1/(n+1)$, the performance of Eq. [3] with $1/\sqrt{t}$ step-size does not attain a speedup over the corresponding single-node rate. This is to be contrasted with Eq. [11] and Eq. [12] where, provided one waits long enough, attain a linear speedup over the single-node rate.

We remark that this theorem has some similarities with Eq. (75) of [3], which considers the speed at which a decentralized optimization method can move towards infinity when no minimizer exists under a constant step-size, and finds it can be network-dependent.

II. PROOFS OF THE MAIN RESULTS

In this section, we provide proofs of Theorems [1,2] and [3]. Our first step is to rewrite Eq. (9) in a way that will be easier to analyze. We define

$$s(t) = \frac{x(t) - P_{\Omega} [x(t) - \alpha(t)g(t)]}{\alpha(t)}$$

so that Eq. (9) can be written as

$$x(t + 1) = W [x(t) - \alpha(t) s(t)]$$

(13)

Consistent with our previous notation, we will use $s_i(t)$ to denote the $i$th row of the matrix $s(t)$.

In this formulation, we no longer have to explicitly deal with the projection, which is incorporated into the definition of $s(t)$. As we will see, the quantity $s(t)$, which is typically known as the “gradient mapping” in the case where the functions are smooth, has some properties similar to the properties of the a subgradient. The next lemma is our first statement to this effect, showing that $s_i(t)$ inherits any upper bound on $g_i(t)$.

Lemma II.1. If $\|g_i(t)\| \leq L$ then $\|s_i(t)\| \leq L$.

Proof. We first observe that for all $i, t$ we have that $x_i(t) \in \Omega$. Indeed, $x_i(t)$ is obtained as a convex combination of vectors projected onto $\Omega$ and so itself belongs to $\Omega$ by convexity. We then use this, along with the fact that projection onto convex sets is nonexpansive, to argue that

$$\|s_i(t)\|_2 = \frac{\|x_i(t) - P_{\Omega} [x_i(t) - \alpha(t)g_i(t)]\|_2}{\|\alpha(t)\|_2}$$

$$\leq \frac{\|x_i(t) - P_{\Omega} [x_i(t) - \alpha(t)g_i(t)]\|_2}{\|\alpha(t)\|_2}$$

$$\leq \frac{\|\alpha(t)g_i(t)\|_2}{\|\alpha(t)\|_2}$$

$$\leq \|g_i(t)\|_2 \leq L.$$ 

Next, we note that it is standard that the subgradient $g_i(t)$ of the convex function $f_i(x)$ at $x_i(t)$ satisfies the relation

$$g_i(t)(x_i(t) - x^*)^T \geq f(x_i(t)) - f_i(x^*).$$

(14)

Our next lemma shows that $s_i(t)$ satisfies the same inequality up to a term that scales with the step-size.

Lemma II.2. Under Assumption [3] we have that

$$\alpha(t)s_i(t)(x_i(t) - x^*)^T \geq \alpha(t)f(x_i(t)) - f_i(x^*) - \frac{\alpha^2(t)}{2} L^2.$$ 

Proof. We start from the relation

$$x_i(t) - \alpha(t)g_i(t) = P_{\Omega} [x_i(t) - \alpha(t)g_i(t)],$$

which is just a rearrangement of the definition of $s(t)$. Our next step is to subtract $x^*$ and take the squared Euclidean norm of both sides. On the left-hand side, we have

$$\|x_i(t) - x^*\|^2 \leq 2\alpha(t)s_i(t)(x_i(t) - x^*)^T + \alpha^2(t)\|g_i(t)\|^2_2.$$ 

Putting these two facts together, we obtain the inequality

$$\|x_i(t) - x^*\|^2 \leq -2\alpha(t)g_i(t)(x_i(t) - x^*)^T + \alpha^2(t)\|g_i(t)\|^2_2$$

or

$$2\alpha(t)s_i(t)(x_i(t) - x^*)^T \geq 2\alpha(t)g_i(t)(x_i(t) - x^*)^T - \alpha^2(t)\|g_i(t)\|^2_2.$$ 

Now using Assumption [3] and Eq. (14), we obtain

$$2\alpha(t)s_i(t)(x_i(t) - x^*) \geq 2\alpha(t)(f_i(x_i(t)) - f_i(x^*)) - \alpha^2(t)L^2,$$

which proves the lemma.

The final lemma we will need bounds the distance between each $x_i(t)$ and $\overline{x}(t)$ as $O(\alpha(t))$ (where the constant inside this $O(\cdot)$-notation will depend on the matrix $W$). Such bounds are standard in the distributed optimization literature.

We introduce some new notation which we will find convenient to use. We will use $1$ to denote the all-ones vector in $\mathbb{R}^n$, so that $1^T t$ has the same dimensions as $x(t)$. We adopt the notation $\sigma$ to denote the second-largest singular value of the matrix $W$; under Assumption [4] we have that $\sigma < 1$ while the largest singular value is $1$ corresponding to the all-ones vector (formally, this follows from Lemma 4 of [4]). In the sequel, we will use the inequality

$$\|W(y - \overline{y})\|_2 \leq \sigma^2\|y - \overline{y}\|_2 \leq \sigma^2\|y\|_2.$$ 

(15)

With these preliminaries in place, we have the following lemma.

Lemma II.3. Suppose Assumptions [1,2,3] on the step-size, Assumption [4] on the functions, and Assumption [7] on the mixing matrix $W$ hold all. When

$$t \geq \Omega \left(\frac{1}{1 - \sigma} \log \left((1 - \sigma)\alpha_{\max}/(C_{\alpha}(\alpha(t))\right)\right)$$

we have that

$$\|x(t) - 1\sigma\|_W \leq \frac{2C_{\alpha}(\alpha(t))L\sqrt{n}}{1 - \sigma}.$$ 

Proof. Recall that, by assumption, the initial conditions are identical; let us denote them all by \( x_1 \). Thus starting from Eq. (13) we have that

\[ x(t) = W^t x_1 - \alpha(1) W^t s(1) - \cdots - \alpha(t-1) W s(t-1) \]

we can use the fact that multiplication by a doubly stochastic matrix doesn’t affect the mean of a vector to obtain that

\[ \pi(t) = x_1 - \alpha(1) \pi(1) - \cdots - \alpha(t-1) \pi(t-1) \]

so that using Eq. (15),

\[ ||x(t) - 1\pi(t)||_F \leq \sum_{k=1}^{t-1} \alpha(k) a^{t-k} ||s(k)||_F. \]

Let us break this sum at \( t' = t - 1 \) \([-t/2]\) and bound each of the two pieces separately. The first piece, over the range \( t = 1, \ldots, t' \) is bounded simply using the fact that all subgradients are upper bounded by \( L \) in the Euclidean norm and consequently \( ||s(t-k)||_F \leq L \sqrt{n} \); whereas the second piece, over the last \( t/2 \) steps, is upper bounded as a geometric sum. The result is

\[ ||x(t) - 1\pi(t)||_F \leq \frac{t}{2} \alpha_{\text{max}} L \sqrt{n} \sigma^{[t/2]} + L \sqrt{n} C_{\alpha}(t) \frac{1}{1 - \sigma}. \]

We next use that \( x^{1/(1-\sigma)} \leq e^{-1} \) when \( x \in [0, 1] \) as well as Assumption 5 to obtain that

\[ ||x(t) - 1\pi(t)||_F \leq t \alpha_{\text{max}} L \sqrt{n} e^{-t(1-\sigma)/2} + L \sqrt{n} C'_{\alpha}(t) \frac{1}{1 - \sigma}. \]

When \( t \geq \Omega((1-\sigma)^{-1} \log \left( (1-\sigma) (\alpha_{\text{max}} / (C'_{\alpha}(t))) \right) \), the first term is upper bounded by the second and the lemma is proved.

With these pieces in place, we are now ready to give a proof of our main result.

Proof of Theorem 1.2. Starting from Eq. (13) we obtain

\[ \pi(t+1) - x^* = \pi(t) - \alpha(t) \pi(t) - x^* \]

so that

\[ ||\pi(t+1) - x^*||_2 = ||\pi(t) - x^*||_2 + \alpha^2(t)||\pi(t)||^2 - 2\alpha(t) \pi(t) (\pi(t) - x^*)^T ||\pi(t)||_2^2 \]

\[ = ||\pi(t) - x^*||_2 + \alpha^2(t)||\pi(t)||^2 - 2\alpha(t) \frac{1}{n} \sum_{i=1}^n s_i (\pi(t) - x^*)^T \]

\[ = ||\pi(t) - x^*||_2 + \alpha^2(t)||\pi(t)||^2 - 2\alpha(t) \frac{1}{n} \sum_{i=1}^n s_i (x_i(t) - x^*)^T \]

\[ + 2\alpha(t) \frac{1}{n} \sum_{i=1}^n s_i (x_i(t) - \pi(t))^T \]

\[ \leq ||\pi(t) - x^*||_2 + \alpha^2(t)L^2 \]

\[ - 2\alpha(t) \frac{1}{n} \sum_{i=1}^n f_i(x_i(t)) - f_i(x^*) \]

\[ + L^2 \alpha^2(t) + 2\alpha(t) \frac{1}{n} \sum_{i=1}^n L||x_i(t) - \pi(t)||_2, \]

where, in the above sequence of inequalities, we used Lemma 11.1 to bound the norm of \( ||\pi(t)||_2 \) and Lemma 11.2 to bound \( s_i(t) (x_i(t) - x^*) \). Now using the fact that each \( f_i(\cdot) \) is \( L \)-Lipschitz, which follows from Assumption 5, we have

\[ ||\pi(t+1) - x^*||_2 \leq ||\pi(t) - x^*||_2 + 2\alpha^2(t)L^2 \]

\[ - 2\alpha(t) \frac{1}{n} \sum_{i=1}^n f_i(\pi(t)) - f_i(x^*) \]

\[ + 4\alpha(t)L \frac{1}{n} \sum_{i=1}^n ||x_i(t) - \pi(t)||_2, \]

We next bound the very last term in the sequence of inequalities above. Our starting point is the observation that

\[ \sum_{i=1}^n ||x_i(t) - \pi(t)||_2 \leq \sqrt{n} ||x(t) - \pi(t)||_F, \]

which follows by an application of Cauchy-Schwarz. We then use Lemma 11.3 to bound the right-hand side above. This yields that, for \( t \) large enough to satisfy the assumptions of that lemma,

\[ ||\pi(t+1) - x^*||_2 \leq ||\pi(t) - x^*||_2 + 2\alpha^2(t)L^2 \]

\[ - 2\alpha(t) \frac{1}{n} \sum_{i=1}^n f_i(\pi(t)) - f_i(x^*) \]

\[ + 4\alpha(t)L 2C_{\alpha}(t)L \frac{1}{1 - \sigma}, \]

implying that

\[ 2\alpha(t) \frac{1}{n} \sum_{i=1}^n f_i(\pi(t)) - f_i(x^*) \leq ||\pi(t) - x^*||_2^2 - ||\pi(t+1) - x^*||_2^2 \]

\[ + 2\alpha^2(t)L^2 + 8\alpha^2(t)C'_{\alpha}(t)L \frac{1}{1 - \sigma}, \]

As before, let \( t' = \lfloor t/2 \rfloor \). We sum the last inequality up over \( k = t', \ldots, t \) to obtain

\[ \sum_{k=t'}^t \alpha(k) ||F(x(k)) - F^*||^2 \leq ||\pi(t') - x^*||_2^2 + \frac{10C'_{\alpha}(t)L^2}{1 - \sigma} \sum_{k=t'}^t \alpha^2(k), \]
where we used that $C'_a \geq 1$ (because $a(t)$ is nonincreasing) and that $\sigma < 1$ to combine the terms involving $a^2(t)$.

Dividing both sides by $2 \sum_{k=t'}^t a(t)$ and using convexity of $F(x)$, we obtain

$$F(\pi(t')) - F^* \leq \frac{\|x(t') - x^\star\|^2}{2} + \frac{10C_a' L^2}{1 - \sigma} \sum_{k=t'}^t a^2(k),$$

where $\sigma$ denotes the constant on the right-hand side of (3).

This step-size satisfies all the assumptions we have made. We then have that

$$\sum_{k=t'}^t \frac{1}{k^\beta} \geq \Omega\left(\left(1 - \frac{1}{2 - \beta + 1}\right)^{-\beta + 1}\right),$$

where the subscript of $\beta$ denotes the constant on the right-hand side of (3).

Therefore when

$$t^{2\beta - 1} \geq \Omega\left(\left(\frac{L^2}{D^2(1 - \sigma)}\right)^\beta\right)$$

we have that

$$F(\pi(t')) - F^* \leq O(\beta)$$

This proves Eq. (12) and the proof is now complete.

Finally, all that remains is to give a proof of Theorem 1.3 showing that the above results for $\beta \in (1/2, 1)$ become false once we set $\beta = 1/2$. The proof below will construct an explicit example where the dependence on spectral gap never disappears, no matter how large $t$ is.

The argument will use the following technical lemma, whose proof we postpone.

**Lemma II.4.** Consider the update rule determined by $y(1) = 0$ and

$$y(t + 1) = (1 - \epsilon) y(t) - \frac{(1/2)(1 - \epsilon)\text{sign}(y(t) - 1) + \epsilon\Delta(t)}{\sqrt{t}},$$

where

$$\Delta(t) = \frac{\epsilon\sqrt{t}y(t) - (\epsilon/2)\text{sign}(y(t) - 1)}{1 - \epsilon}.\tag{16}$$

Then for small enough $\epsilon$, we have that $y(t) \in [0, O(1)]$ and

$$y(t) = \Theta\left(\frac{\epsilon^{-1}}{\sqrt{t}}\right).\tag{17}$$

**Proof of Theorem II.3** We argue that

$$\begin{cases} x_i(t) = 0 & \text{if } i \in \{u_1, \ldots, u_n\} \\ x_i(t) = y(t) & \text{if } i \in \{v_1, \ldots, v_n\} \end{cases}$$

is a valid trajectory of Eq. (6). Here $y(t)$ is from Lemma II.4 and by “valid” we mean that there exists a sequence of subgradients $g_i(t)$, with $g_i(t)$ being a valid subgradient of $f_i(x)$ at $x_i(t)$, resulting in the values in Eq. (18) for all $t$.

Our proof is by induction. At time $t = 1$, we just have $x_1(t) = 0$, so there is nothing to prove. Suppose Eq. (18) is a valid trajectory over times $1, \ldots, t$, and let us consider time $t + 1$. Observe that $x(t) - \alpha(t)g(t) \in [-\gamma, O(\gamma)]$ since $x_i(t) \in [0, O(1)]$ by Lemma II.4 and subgradients are in $[-1/2, 1/2]$ for nodes $v_1, \ldots, v_n$ and between $[-\gamma, +\gamma]$ for the rest (recall that $\gamma > 1$ so that all constants are $O(\gamma)$). Thus choosing $a = \Theta(\gamma)$ appropriately, the projection step can be omitted from Eq. (9).

We next focus on nodes $i \in \{u_1, \ldots, u_n\}$. For these nodes, $x_{u_i}(t + 1) = x_{u_i}(t) + \epsilon(x_{u_i}(t) - x_{u_i}(t)) = (1 - \epsilon)g_{u_i}(t) + \epsilon g_{v_i}(t)$

or

$$\sqrt{t}x_{u_i}(t + 1) = \epsilon\sqrt{t}y(t) - (1 - \epsilon)g_{u_i}(t) - \epsilon g_{v_i}(t),$$

where, in the last two equations, we used that $x_{u_i}(t) = x_{u_i}(t)$ for all $a, b$ by the inductive hypothesis, and $x_v(t) = y(t)$. Therefore, to have $x_{u_i}(t + 1) = 0$ as specified by Eq. (18), we need to have

$$g_{u_i}(t) = (1 - \epsilon)^{-1}\left(\epsilon\sqrt{t}y(t) - \epsilon g_{v_i}(t)\right).\tag{19}$$

Is this a valid choice of subgradient at $x_{u_i}(t) = 0$? Observe that, for small enough $\epsilon$, the right-hand side is $O(1)$ by Lemma II.4 regardless of how we (later) choose $g_{v_i}(t) \in [-1/2, 1/2]$, so as long as we ultimately choose $\gamma$ bigger than this, this is indeed valid.

We now turn to the second line of Eq. (18): we need to show that this will also be valid with an appropriate choice of $g_{i,j}(t)$. By induction, we have that for $i \in \{v_1, \ldots, v_n\}$

$$x_{v_i}(t + 1) = x_{v_i}(t) + \epsilon(0 - x_{v_i}(t)) - \frac{(1 - \epsilon)g_{v_i}(t) + \epsilon g_{u_i}(t)}{\sqrt{t}},$$

where we used that $x_{u_i}(t) = 0$ by the inductive hypothesis, and $x_{v_i}(t) = x_{v_i}(t)$ for all $i, j$ also by the inductive hypothesis. We want to show that there is a choice of $g_{v_i}(t)$ that turns the left-hand side into $y(t + 1)$. But the choice $g_{v_i}(t) = (1/2)\text{sign}(y(t) - 1)$ is valid and turns the left-hand side of the above equation into Eq. (16), so it certainly results in $x_{v_i}(t + 1) = y(t + 1)$.

To summarize, we have shown how to choose valid subgradients at each step so that Eq. (9) turns into the recursion relation satisfied by Eq. (18). The proof is now complete.

**Proof Sketch of Lemma II.4** That $y(t) \in [0, O(1)]$ for small enough $\epsilon$ follows by observing that, for small enough $\epsilon$, (i) $y(t)$ decreases whenever it is above (ii) $y(t)$ cannot decrease below zero (iii) if $y(t) \in [0, 1]$, then it can increase by at most $O(1/\sqrt{t})$.

Next, since $|\text{sign}(y(t) - 1)| \leq 1$, we have that for small enough $\epsilon$,

$$y(t + 1) \leq \left(1 - \epsilon - \frac{\epsilon^2}{1 - \epsilon}\right)y(t) + \frac{1}{2\sqrt{t}}.$$
Defining \( z(t) \) via
\[
z(t + 1) = (1 - \epsilon)z(t) + \frac{1}{2\sqrt{t}}.
\]
we have that \( y(t) \leq z(t) \). To prove that \( y(t) = O(e^{-1}/\sqrt{t}) \), we simply need to establish the same estimate for \( z(t) \) rather than \( y(t) \). To that end, we multiply both sides of Eq. (20) by \( \sqrt{t+1} \) and use concavity of square root to obtain
\[
\sqrt{t+1}z(t + 1) \leq (1 - \epsilon)\left(\sqrt{t} + \frac{1}{2\sqrt{t}}\right)z(t) + \frac{1}{2}\sqrt{\frac{t+1}{t}}.
\]
Since it is immediate from Eq. (20) that \( z(t) = O(\sqrt{t}) \), the last equation gives implies that
\[
\sqrt{t+1}z(t + 1) \leq (1 - \epsilon)\sqrt{t}z(t) + O(1),
\]
which gives that \( \sqrt{t}z(t) = O(e^{-1}) \). This proves that \( y(t) = O(e^{-1}/\sqrt{t}) \).

Since we now know that \( y(t) \to 0 \), it follows that, for all large enough \( t \), \( \text{sign}(y(t) - 1) = -1 \) and therefore, for such large enough \( t \),
\[
y(t + 1) = \left(1 - \epsilon - \frac{\epsilon^2}{1 - \epsilon}\right)y(t) + \frac{(1 - \epsilon) - \epsilon^2/(1 - \epsilon)}{2\sqrt{t}},
\]
which, for small enough \( \epsilon \), implies that
\[
y(t + 1) \geq (1 - 2\epsilon)y(t) + \frac{1}{4\sqrt{t}},
\]
immediately implying that \( y(t) = \Omega(e^{-1}/\sqrt{t}) \). This completes the proof. \( \square \)

III. A NUMERICAL EXAMPLE

We next briefly give a numerical illustration of what network independence looks like. Specifically, we simulate the example constructed in the proof of Theorem 1.3. We choose the values \( \gamma = 2 \) and \( \alpha = 5 \) and step-size of \( 1/t^\beta \). The optimal step-size choice of \( \beta = 1/2 \) is shown in the first figure and the choice of \( \beta = 3/4 \) is shown in the second figure. The y-axis of both figures shows \( t^{1-\beta}(F(\pi(t)) - F^*) \). Each simulation shows three different values of \( n \).

Comparing the figures illustrates our main result: we can see the qualitative difference in behavior between \( \beta = 1/2 \) and \( \beta = 3/4 \). Indeed, the first figure results in \( t^{1-\beta}(F(\pi(t)) - F^*) \) approaching some number that clearly grows with \( n \). Thus the effect of \( n \) is never forgotten.

On the other hand, in the second figure, after a peak whose height/length may depend on \( n \), every curve drops below 1, i.e., we have that \( t^{1-\beta}(F(\pi(t)) - F^*) \leq 1 \). In other words, in this case, after a transient, the performance satisfies an \( O(t^{-1-\beta}) \) decay bound that does not depend on \( n \). Of course, the size of the transient until this happens could depend on the spectral gap, and on \( n \) when the spectral gap depends on \( n \).

REFERENCES

[1] S. Boyd, L. Xiao, and A. Mutapcic. Subgradient methods. lecture notes of EE3320, Stanford University, Autumn Quarter, 2004:2004–2005, 2003.
[2] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. IEEE Transactions on Automatic control, 57(3):592–606, 2011.
[3] P. Lin, W. Ren, and J. A. Farrell. Distributed continuous-time optimization: nonuniform gradient gains, finite-time convergence, and convex constraint set. IEEE Transactions on Automatic Control, 62(5):2239–2253, 2016.
[4] S. Liu, Z. Qiu, and L. Xie. Convergence rate analysis of distributed optimization with projected subgradient algorithm. Automatica, 83:162–169, 2017.
[5] A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. Proceedings of the IEEE, 106(5):953–976, 2018.
[6] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multiagent optimization. IEEE Transactions on Automatic Control, 54(1):48–61, 2009.
[7] A. Nedić, A. Ozdaglar, and P. A. Parrilo. Constrained consensus and optimization in multi-agent networks. IEEE Transactions on Automatic Control, 55(4):922–938, 2010.
[8] G. Neglia, C. Xu, D. Towsley, and G. Calbi. Decentralized gradient methods: does topology matter? In The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020), 2020.
[9] A. Olshevsky. Linear time average consensus and distributed optimization on fixed graphs. *SIAM Journal on Control and Optimization*, 55(6):3990–4014, 2017.

[10] S. Pu, A. Olshevsky, and I. C. Paschalidis. Asymptotic network independence in distributed optimization for machine learning. *IEEE Signal Processing Magazine, to appear*, 2020.

[11] Z. Qiu, S. Liu, and L. Xie. Distributed constrained optimal consensus of multi-agent systems. *Automatica*, 68:209–215, 2016.

[12] S. S. Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

[13] K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pages 2740–2749, 2018.

[14] J. Wang and N. Elia. A control perspective for centralized and distributed convex optimization. In *2011 50th IEEE conference on decision and control and European control conference*, pages 3800–3805. IEEE, 2011.

[15] M. Zhu and S. Martínez. On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control*, 57(1):151–164, 2011.