SIAM J. Optim.
Vol. 19, No. 2, pp. 911–940

DECENTRALIZED RESOURCE ALLOCATION IN DYNAMIC NETWORKS OF AGENTS[*]

HARIHARAN LAKSHMANAN[†] AND DANIELA PUCCI DE FARIAS[‡]

Abstract. We consider the problem of n agents that share m common resources. The objective is to derive an optimal allocation that maximizes a global objective expressed as a separable concave objective function. We propose a decentralized, asynchronous gradient-descent method that is suitable for implementation in the case where the communication between agents is described in terms of a dynamic network. This communication model accommodates situations such as mobile agents and communication failures. The method is shown to converge provided that the objective function has Lipschitz-continuous gradients. We further consider a randomized version of the same algorithm for the case where the objective function is nondifferentiable but has bounded subgradients. We show that both algorithms converge to near-optimal solutions and derive convergence rates in terms of the magnitude of the gradient of the objective function. We show how to accommodate nonnegativity constraints on the resources using the results derived. Experimental results with the problems of varying dimensions suggest that the algorithms are competitive with centralized approaches and scale well with problem size.

Key words. decentralized algorithms, resource allocation, nondifferentiable optimization

AMS subject classifications. 90C25, 91B32

DOI. 10.1137/060662228

1. Introduction.

We consider the problem of n agents that share m common resources. Agent i has utility function f_i. The optimal allocation of resources for maximizing the average of the utilities among agents is given by the following optimization problem:

$$\max_{\lambda_i \in \Re^m, i=1,\dots,n} \; f(\lambda) = \frac{1}{n}\sum_{i=1}^{n} f_i(\lambda_i)$$

$$\text{such that (s.t.)} \sum_{i=1}^{n} \lambda_i = B, \qquad (1)$$

where $B \in \Re^m$ corresponds to the total amount of resources.

We propose decentralized, asynchronous algorithms for a solution of (1). The first method applies in the case where $f_i, i = 1, \dots, n$ are concave and differentiable with Lipschitz-continuous gradients. The second method applies in the case where $f_i, i = 1, \dots, n$ are concave but not necessarily differentiable. We establish asymptotic convergence and convergence rates of both algorithms under mild conditions for communications among agents.

We assume that agents communicate through a network of dynamic topology in order to solve (1). At each iteration t, communication is represented by an undirected graph $G(t)$, where nodes correspond to agents and edges correspond to communication

[*] Received by the editors May 21, 2006; accepted for publication (in revised form) January 7, 2008; published electronically September 11, 2008. This research was partly supported by a Draper Laboratory grant and by NSF CAREER Award 0447766. A preliminary version of this work appears in the Proceedings of the American Control Conference 2006 [7].

http://www.siam.org/journals/siopt/19-2/66222.html

[†] Corresponding author. Department of Civil and Environmental Engineering, MIT, Cambridge, MA 02139 (lhari@mit.edu).

[‡] Department of Mechanical Engineering, MIT, Cambridge, MA 02139 (pucci@mit.edu).

links. We assume that communication is symmetric, so that if agent i communicates with agent j, then agent j also communicates with agent i. We also assume that the union of the communication graphs is connected over any sufficiently large, bounded period of time. This ensures that resource allocation to every agent is periodically influenced either directly or indirectly by the resource allocation to every other agent. This model of communication accommodates several practical scenarios, arising for instance, if agents are mobile and have limited communication range, or if communication links are subject to failure.

The decentralized algorithm for solving (1) in the case of differentiable utility functions has a simple gradient-descent structure. Starting with an initial feasible resource allocation, agents trade resources with their neighbors at each iteration in proportion to the difference in gradient for the respective utility functions. The algorithm has a natural interpretation. The local gradient computed by each agent can be thought of as the price the agent is willing to pay for additional resources. At each iteration, agents trade resources with their neighbors in proportion to the prices each is willing to pay for the resources.

It can be shown that a large class of separable convex optimization problems with linear constraints can be transformed to equivalent resource allocation problems. However, the functions f_i in the transformed resource allocation problem are usually not differentiable. Motivated by this setting, we consider the case where f_i is no longer differentiable but has bounded subgradients. It is shown in this case that a randomized version of a decentralized subgradient-descent algorithm converges with probability one to a near-optimal solution.

The subgradient-descent algorithm for the case of nondifferentiable utility functions can be interpreted as a stochastic approximation version of the gradient-descent method for differentiable functions applied to a smoothed version of the problem. The particular form of smoothing developed in this paper is motivated by several considerations. Adequate smoothing schemes must lead to a close approximation to the original function. Furthermore, as we build on the results for differentiable problems with a Lipschitz-continuous gradient, the gradient of the resulting smooth function must satisfy the same assumption with an adequate Lipschitz constant. Finally, another consideration in this paper is the computational effort involved in computing the gradient for the smoothed function. With this in mind, we propose a smooth approximation of the form $\hat{f}_i = \mathrm{E}[f_i(\lambda_i + Z_i)]$, where Z_i are vectors of zero-mean normal random variables. We show that, with an appropriate choice for the variance of Z_i, $\hat{f}_i$ is within ϵ of f_i, and its gradient is Lipschitz-continuous, with a Lipschitz constant on the order of $O(\sqrt{\log m}/\epsilon)$ so that it scales gracefully on the dimension m of variable λ_i. In addition, this form of smoothing lends itself to an application of a stochastic approximation scheme for gradient descent which, at each iteration, only requires an evaluation of a subgradient of f_i at a single point λ_i.

A comprehensive treatment of algorithms for various classes of resource allocation problems can be found in [13]. The algorithms introduced and analyzed in [13] are centralized in the sense that a central agent is assumed to have complete information about the problem and computes the optimal solution. In [1] and [5], decentralized resource allocation problems in the context of economics are investigated. The main difference in the approaches of [1, 5] as compared to the one presented here is the presence of a central agent who coordinates the computations performed by individual agents. A setting that is closer to ours is presented in [11], which introduces a completely decentralized algorithm for a resource allocation problem with twice differentiable separable convex objective functions. The algorithm assumes a symmetric and fixed communication graph for the agents at all iterations and performs a gradient projection at each iteration onto a subspace related to the communication graph. The same setting is considered in [8], which proposes a decentralized, weighted gradient algorithm for resource allocation problems with objective functions that are twice differentiable with bounded second derivatives. Dynamic communication graphs are considered in [6], which proposes an application-specific decentralized gradient algorithm for the problem of file allocation in distributed computer systems. Asynchronous gradient-descent methods are also considered in [14] for problems of unconstrained optimization with differentiable objective.

Most of the references regarding resource allocation problems in the literature, including the ones mentioned above, contain nonnegativity constraints on the resources (i.e., they require $\lambda_i \geq 0 \ \forall i$), whereas in our formulation resources may be negative. In Section 4, we show how the results in this paper can be applied to problems with nonnegativity constraints. The main motivation for problem (1) is that this formulation arises naturally in problems where a single, generic optimization problem must be solved in a decentralized way; this is the case, for instance, in problems of sequential decision making in teams of mobile agents as considered in [7].

A distributed algorithm for nondifferentiable optimization is presented in [9]. It is shown that a projected subgradient algorithm applied by each agent converges to the optimal solution. An important difference between the work presented in [9] and the work presented here is that the first requires that the long-run frequency of updates performed by each agent to be the same. Smoothing schemes for nondifferentiable optimization can also be found in the literature. [10] proposes a smoothing scheme for functions f_i described as the maximum of differentiable functions. The smoothed function is within ϵ of f_i and has Lipschitz constant on the order of $O(1/\epsilon)$, independent of the dimensions of the problem. A caveat of this approach is that computing the gradient of the smoothed function may require multiple evaluations of the subgradients of the original function. The particular form of smoothing considered here can also be found in the literature (see, e.g., [12]); however, we are unaware of results concerning the Lipschitz constant of the resulting smoothed function, which we develop in this paper.

The paper is organized as follows. In section 2, we describe the structure of communication among agents. In section 3, we introduce and we analyze the decentralized gradient-descent algorithm for problem (1) with differentiable objective functions that have Lipschitz-continuous gradients and its randomized version for problem (1) with nondifferentiable objective functions. In section 4 we describe a method to accommodate the nonnegativity constraints on the variables based on the results developed in section 3. In section 5, we present the results of numerical experiments, which illustrate the practical performance of the developed algorithms. In section 6, we conclude the paper. All proofs can be found in the appendix.

2. Communication between agents.

In this section we describe the communication structure between agents. At iteration t, each agent i communicates with a set of agents denoted by $N_i(t)$. We assume that communication is symmetric; i.e., whenever agent i communicates with agent j, agent j also communicates with agent i. The communication between agents at time t can be represented by an undirected graph $G(t) = (N, E(t))$, where $N = \{1, \ldots, n\}$ represents the set of agents and the edge $(i, j) \in E(t)$ if and only if agent i communicates with agent j at time t. Let $E_{k,l} = \cup_{t=k}^{t=l-1} E(t)$. For a decentralized scheme to converge, the update of the variable associated with any agent must be periodically influenced by information from every other agent. This is ensured by the following assumption.

Assumption 2.1. *There exists a strictly increasing sequence $\{T_z\}$ of natural numbers, with $T_1 = 1$ such that $G = (N, E_{T_z,T_{z+1}})$ is connected for all z and $(T_{z+1} - T_z) \leq \kappa$, where κ is some natural number.*

3. Decentralized resource allocation.

We assume that (1) has an optimal solution. Let $\lambda \in \Re^{nm} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_i \in \Re^m$ for $i = 1, \ldots, n$.

Assumption 3.1. *There exists an optimal solution $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*)$ to* (1).

For the rest of the paper, we let $\|\cdot\|$ denote the Euclidean norm.

3.1. The differentiable case.

We now develop a decentralized algorithm for the case where f_i is concave and differentiable with a Lipschitz-continuous gradient.

Assumption 3.2. *There exists a constant $L > 0$ such that $\|\nabla f_i(\lambda_i) - \nabla f_i(\bar{\lambda}_i)\| \leq L\|\lambda_i - \bar{\lambda}_i\|$, $\forall \lambda_i, \bar{\lambda}_i \in \Re^m$.*

Recall that $f(\lambda) = \frac{1}{n}\sum_{i=1}^n f_i(\lambda_i)$. Hence

$$\begin{aligned}\|\nabla f(\lambda) - \nabla f(\bar{\lambda})\| &= \frac{1}{n}\sqrt{\sum_{i=1}^n \|\nabla f_i(\lambda_i) - \nabla f_i(\bar{\lambda}_i)\|^2} \\ &\leq \frac{1}{n}\sqrt{\sum_{i=1}^n L^2\|\lambda_i - \bar{\lambda}_i\|^2} \\ &= \frac{L}{n}\|\lambda - \bar{\lambda}\|.\end{aligned}$$

The second equality follows from the fact that $\|\lambda - \bar{\lambda}\| = \sqrt{\sum_{i=1}^n \|\lambda_i - \bar{\lambda}_i\|^2}$. Hence $\frac{L}{n}$ is a Lipschitz constant for the function f. The decentralized algorithm that we develop is based on the following lemma, which characterizes an optimal solution to (1) when functions f_i are all differentiable.

Lemma 3.1. *A feasible solution λ^* of* (1) *is an optimal solution if and only if $\nabla f_i(\lambda_i^*) = \nabla f_j(\lambda_j^*)$ for all i, j.*

Let λ_i^t be the value of the variable associated with agent i at iteration t. We consider the following gradient-descent update rule for each agent i:

$$\lambda_i^{t+1} = \lambda_i^t + \gamma \sum_{j \in N_i(t)} \frac{1}{n}(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)). \tag{2}$$

Here γ is a common constant step size that all of the agents use for updates. It should be noted that, to perform updates at iteration t, agent i uses only the gradient information corresponding to its neighbors $N_i(t)$ for iteration t. Furthermore, each intermediate allocation λ^t generated by the algorithm is a feasible solution of (1).

Lemma 3.2. *Suppose λ^1 is a feasible solution for* (1)*. Then λ^t, where λ_i^t is defined by* (2)*, is a feasible solution to* (1) *for all t.*

In order to analyze the convergence properties of the proposed algorithm, it is convenient to define $\tilde{v}(\lambda)$ for any allocation λ as follows:

$$\tilde{v}_i(\lambda) = \sum_{j \in N} \frac{1}{n}(\nabla f_i(\lambda_i) - \nabla f_j(\lambda_j)).$$

Note that $\tilde{v}(\lambda^t)$ is the direction of update when the communication graph $E(t)$ is complete. It can be verified that $\tilde{v}(\lambda^t)$ is also a scaled version of the projection of $\nabla f(\lambda^t)$ onto the subspace $\sum_{i=1}^n \lambda_i^t = B$, hence it represents the centralized update direction at time t. From Lemma 3.1 it can be seen that a feasible solution λ is optimal if and only if $\|\tilde{v}(\lambda)\| = 0$. We now derive a theorem establishing the convergence of the algorithm based on (2). Under mild conditions on the set of optimal solutions, convergence to optimality is guaranteed. We also derive an upper bound on the rate at which the sequence $\{\|\tilde{v}(\lambda^{T_z})\|\}$ converges to zero. Recall that T_z is a sequence of strictly increasing natural numbers such that the union of the communication graphs between iterations T_z and T_{z+1} is connected. In what follows, let $\tilde{v}^t = \tilde{v}(\lambda^t)$.

THEOREM 3.1. *Suppose that Assumptions* 3.1 *and* 3.2 *hold. With a step size of* $\gamma = \frac{1}{2L}$,

1. *the sequence* $\{f(\lambda^t)\}$ *is monotonically nondecreasing;*
2. *the sequence* $\{\|\tilde{v}^{T_z}\|\}$ *converges to* 0*;*
3. $\min_{z=1,\dots,p}(\|\tilde{v}^{T_z}\|^2) \leq \frac{3Ln^4\kappa(f(\lambda^*)-f(\lambda^1))}{4p}$ $\forall p$*;*
4. *if the set of optima is bounded,* $\{f(\lambda^t)\}$ *converges to* $f(\lambda^*)$*.*

3.2. The nondifferentiable case. In this section we consider concave objective functions that are not required to be differentiable at all points. To motivate our interest in such functions we consider the following optimization problem:

$$\max_{x_i, i=1,\dots,n} \frac{1}{n}\sum_{i=1}^{n} g_i(x_i)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \mathbf{A_i} x_i \leq B, \tag{3}$$

where $x_i \in \Re^q$, $\mathbf{A_i} \in \Re^{m\times q}, i = 1,\dots,n$, $B \in \Re^m$, and $g_i(x_i)$ is a concave function, $i = 1,\dots,n$. Define $f_i(\lambda_i)$ as the optimal value for the following optimization problem:

$$f_i(\lambda_i) = \max_{x_i\in\Re^q} g_i(x_i)$$

$$\text{s.t.} \quad \mathbf{A_i} x_i \leq \lambda_i. \tag{4}$$

With this definition of f_i we see that problem (3) is equivalent to problem (1). Note that, if there are linear constraints that involve only the variables $x_{ij}, j = 1,\dots,q$ for some i, then these constraints could be included directly in the problem defining f_i. Suppose that $f_i(\lambda_i)$ is well defined and is finite for all λ_i. It can then be shown that $f_i(\lambda_i)$ is a concave function. Thus we can potentially apply the decentralized algorithm developed in the previous section for finding an optimal solution to (3). However, $f_i(\lambda_i)$ is typically nondifferentiable even when $g_i(x_i)$ is. Hence Theorem 3.1 does not immediately apply to (3), as it relies on the assumption that the objective function is differentiable with a Lipschitz-continuous gradient. This motivates us to consider cases where f_i, $i = 1,\dots,n$ are not necessarily differentiable at all points.

In this section, we relax Assumption 3.2 and consider the case where f_i, $i = 1,\dots,n$ are nondifferentiable. We introduce a smooth approximation for f_i that is amenable to optimization via stochastic approximations and propose a randomized version of (2) to solve the smoothed problem. We show that the new scheme converges to a near-optimal solution of the original problem in a tractable number of iterations.

We assume that f_i, $i = 1,\dots,n$ are concave and differentiable outside a set of measure zero. Denote by $\partial f_i(\lambda_i)$ the set of the subgradients of f_i at λ_i. Let $\nabla f_i(\lambda_i)$ be an element chosen arbitrarily from $\partial f_i(\lambda_i)$ for each λ_i. Let $\|\cdot\|_1$ denote the l_1 norm and recall that $\|\cdot\|$ denotes the Euclidean norm. We make the following assumption.

ASSUMPTION 3.3. *For all i and λ_i,* $\sup_{i,\lambda_i}\{\|v\|_1 : v \in \partial f_i(\lambda_i)\} \leq L < \infty$.

Note that $\sup_{i,\lambda_i}\{\|v\| : v \in \partial f_i(\lambda_i)\} \leq L < \infty$, since $\|v\| \leq \|v\|_1$ for all v. We now consider approximating f_i by a suitable differentiable function. In particular, let

$$\hat{f}_i(\lambda_i) = \mathrm{E}[f_i(\lambda_i + Z_i)],$$

where each $Z_i = (Z_{ij})_{j=1,\ldots,m}$ is a vector of m independently and identically distributed (i.i.d.) normal random variables [4], with a zero mean and variance equal to

$$\sigma = \frac{\sqrt{2}\epsilon}{\sqrt{\pi \log(m+1)}},$$

where ϵ is a parameter related to the accuracy of the approximation as will be clear from the following lemma. The following lemma shows that $\hat{f}_i$ is a concave and differentiable approximation to f_i and that its gradient $\nabla \hat{f}_i$ can be expressed in terms of ∇f_i.

Lemma 3.3. *Let f_i and $\hat{f}_i$ be as given above. Then the following hold:*

1. *$\hat{f}_i$ is concave and differentiable, with gradient $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i)]$;*
2. *$f_i(\lambda_i) \geq \hat{f}_i(\lambda_i) \geq f_i(\lambda_i) - 2.8\epsilon L$;*
3. *$\|\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_i(\bar{\lambda}_i)\| \leq \frac{\sqrt{\log(m+1)}L}{\epsilon}\|\lambda_i - \bar{\lambda}_i\|$.*

Bearing in mind the previous lemma, we consider the problem of maximizing

$$\begin{aligned} \max_{\lambda} \hat{f}(\lambda) &= \sum_{i=1}^{n} \frac{1}{n} \hat{f}_i(\lambda_i) \\ \text{s.t.} \sum_{i=1}^{n} \lambda_i &= B. \end{aligned} \tag{5}$$

Since $\hat{f}_i$ is differentiable with a Lipschitz-continuous gradient, Theorem 3.1 ensures that the update rule (2) leads to convergence. However, note that computing the gradient of $\hat{f}_i$ requires evaluating the expected value $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i)]$, which is, in general, computationally expensive. Due to the special form of the smoothing scheme and, in particular, the fact that $\nabla \hat{f}_i$ is expressed as the expected value of the subgradient of f_i, we consider instead of (2) a stochastic approximation version of the update. In particular, we let

$$\lambda_i^{t+1} = \lambda_i^t + \gamma_t \sum_{j \in N_i(t)} \frac{1}{n}(\nabla f_i(\lambda_i^t + Z_i^t) - \nabla f_j(\lambda_j^t + Z_j^t)), \tag{6}$$

where $Z_i^t,\ t = 1, 2, \ldots$ is a sequence of i.i.d. vectors with the same distribution as Z_i.

For each λ, let $\tilde{v}(\lambda)$ be given by

$$\tilde{v}_i(\lambda) = \sum_{j \in N} \frac{1}{n}(\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_j(\lambda_j)).$$

Let $\tilde{v}^t = \tilde{v}(\lambda^t)$, and note that $\tilde{v}^t$ corresponds to the expected direction of update when the communication graph is complete. From Lemma 3.1, it is clear that a feasible solution λ is optimal for (5) if and only if $\|\tilde{v}(\lambda)\| = 0$. Furthermore, from Lemma 3.3, if λ is optimal for (5), then it is also near-optimal for (1). The following theorem establishes that, if all agents apply (6), then $\|\tilde{v}^t\|$ converges to zero.

We make the following assumption on the step sizes γ_t.

ASSUMPTION 3.4. *The step sizes γ_t satisfy $\gamma_t = \frac{\epsilon}{(2L\sqrt{\log(m+1)})}\beta_t$, where $0 \leq \beta_{t+1} \leq \beta_t \leq 1 \forall t$, $\sum_t \beta_t = \infty$, and $\sum_t \beta_t^2 < \infty$.*

THEOREM 3.2. *Suppose that Assumptions* 3.3 *and* 3.4 *hold. Then with probability* 1*:*

1. *the sequence $\{\|\tilde{v}^{T_z}\|\}$ converges to* 0*;*
2. $\min_{z=1,\dots,p} \mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \frac{\frac{n^4\kappa L\sqrt{\log(m+1)}}{\epsilon}\left[3(f(\lambda^*)-f(\lambda^1)+2.8\epsilon L)+\sum_{t=1}^{t=\kappa p}\frac{4L\beta_t^2\epsilon}{\sqrt{\log(m+1)}}\right]}{4\sum_{z=2}^{p+1}\beta_{\kappa z}}$ $\forall p$*;*
3. *if the set of the optima of* (1) *is bounded, then* $\lim_{t\to\infty} f(\lambda^t) \geq f(\lambda^*) - 2.8\epsilon L$.

It is worth noting some aspects of Theorem 3.2. Unlike in the differentiable case, we cannot guarantee a monotonic increase in the objective function values. Hence the rate of convergence of the sequence $\{\mathrm{E}[\|\tilde{v}^{T_z}\|]\}$ to zero does not have as far-reaching implications as its counterpart in Theorem 3.1. Nevertheless, Theorem 3.2 ensures convergence to a near-optimal solution with probability one. Another substantial difference is on the assumption on step sizes and the corresponding effect on convergence rates. It is easy to see that convergence is ensured if $\beta_t = \frac{1}{t^q}$ for $0.5 < q \leq 1$. The resulting theoretical rate of convergence is clearly dependent on q; when $0.5 < q < 1$, $\frac{1}{x^q}$ is a decreasing function for $x \geq 1$. Hence, for $k \geq 1, \frac{1}{k^q} \geq \int_k^{k+1} \frac{1}{x^q}dx$, and so $\sum_{k=1}^c \frac{1}{k^q} \geq \int_1^{c+1} \frac{1}{x^q}dx = \frac{(c+1)^{1-q}-1}{1-q}$. Thus the number of iterations needed for $\mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \epsilon$ is polynomial in the problem parameters. Similarly, when $q = 1$, $\frac{1}{x^q}$ is just $\frac{1}{x}$ and is a decreasing function as well for $x \geq 1$. Hence, for $k \geq 1$, $\frac{1}{k+1} \leq \int_k^{k+1} \frac{1}{x}dx$, and so $\sum_{k=1}^c \frac{1}{k} \leq 1 + \int_1^c \frac{1}{x}dx = \log(c) + 1$, and so the number of iterations needed for $\mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \epsilon$ is exponential in the problem parameters. As is often observed in stochastic approximation methods, the impact of the choice of step sizes on the speed of the convergence of the algorithm is also verified in the numerical experiments.

4. Decentralized resource allocation with nonnegativity constraints.

In this section, we use the results developed for (1) to solve the following resource allocation problem with nonnegativity constraints:

$$\begin{aligned} \max_{\lambda_i\in\Re^m, i=1,\dots,n} \quad & f(\lambda) = \frac{1}{n}\sum_{i=1}^n f_i(\lambda_i) \\ \text{s.t.} \quad & \sum_{i=1}^n \lambda_i = B, \\ & \lambda_i \geq 0, i = 1,\dots,n. \end{aligned} \tag{7}$$

We assume that f_i is concave and differentiable outside a set of measure zero. Also let Assumption 3.3 hold for f.

We now define $g_i(\lambda_i)$ as follows:

$$g_i(\lambda_i) = f_i(\lambda_i) + \sum_{j=1}^m L_g \min(\lambda_{ij}, 0),$$

where $L_g > 2L$. The following lemma shows that the function $g(\lambda)$ satisfies Assumption 3.3 and is necessary for applying the stochastic approximation version of the gradient-descent algorithm developed in 3.2.

LEMMA 4.1. *Under assumption* 3.3 *for* f,

1. *for all* i, $g_i(\lambda_i)$ *is concave and differentiable outside a set of measure zero;*
2. *for all* i *and* λ_i, $\sup_{i,\lambda_i}\{\|v\|_1 : v \in \partial g_i(\lambda_i)\} \leq L_m < \infty$, *where* $L_m = L + mL_g$.

It can be noted from the definition of g_i that if $\lambda_i \geq 0$, then $g_i(\lambda_i) = f_i(\lambda_i)$. The term $L_g \min(\lambda_{ij}, 0)$ in the above definition can be thought of as a penalty for negative λ_{ij}. This term ensures that solving (1) with g has a nonnegative optimal solution and is equivalent to solving (7) with f.

LEMMA 4.2. *The set of optimal solutions for* (1) *with* g *as the objective function is the same as the set of optimal solutions to* (7) *with* f *as the objective function.*

Since the set of the feasible solutions of (7) is bounded and closed and since f is assumed to be continuous, there exists an optimal solution to (7). Thus any algorithm that finds an optimal solution to (1) with g as the objective function also yields an optimal solution to (7) with f as the objective function.

Lemma 4.1 ensures that we can apply the stochastic approximation version of the gradient-descent algorithm for (1) with g as the objective function. Hence an optimal solution for (7) with f as the objective function can be found by applying the stochastic approximation version of the gradient-descent algorithm developed in 3.2 for (1) with g as the objective function. It should be pointed out that the Lipschitz constant of the smoothed problem, and consequently the convergence rate, is now of the order $O(\frac{m\sqrt{m}}{\epsilon})$ as compared to $O(\frac{\sqrt{m}}{\epsilon})$ for the results of section 3.2.

5. Numerical experiments.

In this section, we present the results of numerical experiments, which illustrate the performance of the algorithms presented in the previous sections. We compare the proposed algorithms to centralized algorithms that use direction $\tilde{v}(\lambda)$ as the direction of update. Recall that $\tilde{v}(\lambda)$ is the direction of update if the current resource allocation is λ and the communication graph is complete. Recall also that, when f_i is differentiable, $\tilde{v}(\lambda)$ is the projection of ∇f onto the subspace $\sum_{i=1}^{n} \lambda_i^t = B$. Thus the centralized algorithm reduces to the classic gradient-descent method of nonlinear optimization in this case. We define $p^t = (\frac{f^t - f^0}{f^* - f^0}) \times 100$, where f^t is the objective function value after t iterations and f^* is the objective function value of the optimal solution, and we investigate how p^t converges to 100 in the centralized and decentralized algorithms.

5.1. Problem with differentiable objective function.

We first consider a problem studied in [8], which is an instance of (1), with

$$f_i(x_i) = -\left(\frac{1}{2}a_i(x_i - c_i)^2 + \log(1 + e^{b_i(x_i - d_i)})\right), i = 1, \ldots, n.$$

The second derivative $f_i^{''}$ is given by

$$f_i^{''}(x_i) = -\left(a_i + b_i^2 \frac{e^{b_i(x_i - d_i)}}{(1 + e^{b_i(x_i - d_i)})^2}\right), \qquad i = 1, \ldots, n.$$

It can be verified that $f_i^{''}(x_i)$ has a lower bound $-(a_i + \frac{1}{4}b_i^2)$, $i = 1, \ldots, n$. It can be shown that, if a one-dimensional function is differentiable and its gradient is bounded by some constant, then the function is Lipschitz-continuous with the same constant. Since f_i is twice differentiable and $f_i^{''}$ is bounded, it follows that $f_i^{'}$ is Lipschitz-continuous, with constant $(a_i + \frac{1}{4}b_i^2)$, if we assume that $a_i \geq 0$. It follows that $f^{'}$ is Lipschitz-continuous, with constant $\frac{L}{n}$, where $L = \max_i(a_i + \frac{1}{4}b_i^2)$. Thus f satisfies Assumption 3.2.

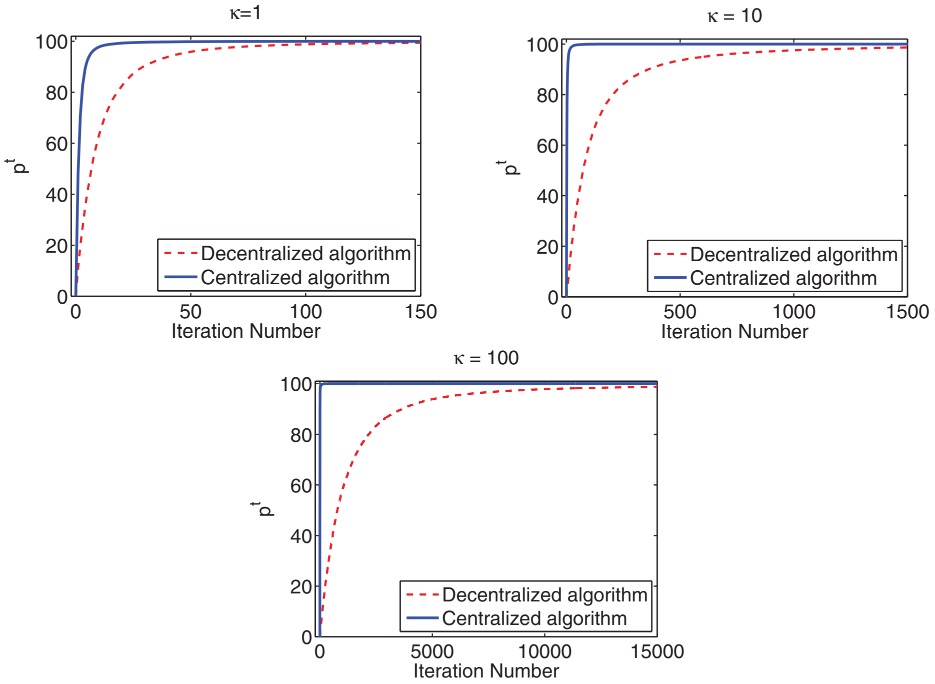

FIG. 1. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various κ.*

We choose problem instances with 20 agents and as in [8]; the coefficients a_i, b_i, c_i, and d_i are generated randomly, with uniform distributions on $[0, 2]$, $[-2, 2]$, $[-10, 10]$, and $[-10, 10]$, respectively. Recall that, for our algorithm to converge, the union of communication graphs should be connected periodically. For a chosen κ, we let the edges $(i, i+1), i = 1, \ldots, n-1$ be a part of the communication graph $E(t)$ for some arbitrarily chosen t such that $m\kappa < t \leq (m+1)\kappa, m = 0, 1, \ldots$. This ensures that $G = (N, E_{m\kappa+1,m(\kappa+1)+1})$ is connected (recall that $E_{k,l} = \cup_{t=k}^{t=l-1} E(t)$). We let every other edge (i, j), with $j \neq i+1$, be a part of at the most one communication graph between iterations $m\kappa + 1$ and $(m+1)\kappa$, with a probability e_p. The parameter e_p controls the density of the graph, $G = (N, E_{m\kappa+1,m(\kappa+1)+1})$. The step size is chosen to be $\frac{1}{2L}$, with L as defined above. Figure 1 shows the convergence behavior of the algorithm for various values of the parameter κ, with $e_p = 0.1$. p^t in the figure represents the average of p^t for 10 randomly chosen problems. It can be seen from the figure that the performance of the decentralized algorithm is comparable to the centralized algorithm for $\kappa = 1$, even though the communication graph is not dense $(e_p = 0.1)$.

Figure 2 shows a comparison of the convergence behavior of the algorithms for problems with a varying number of agents. We fix $\kappa = 1$ in these problems, and $e_p = 0.1$. The other parameters are chosen as described above. We notice from Figure 2 that the scaling of the performance of decentralized algorithms with increasing number of agents is much better than $O(n^4)$ promised by Theorem 3.1.

5.2. Decentralized optimization of linear programming problems. We now consider a decentralized solution of linear programming problems using the randomized version of the decentralized subgradient-descent algorithm developed in this

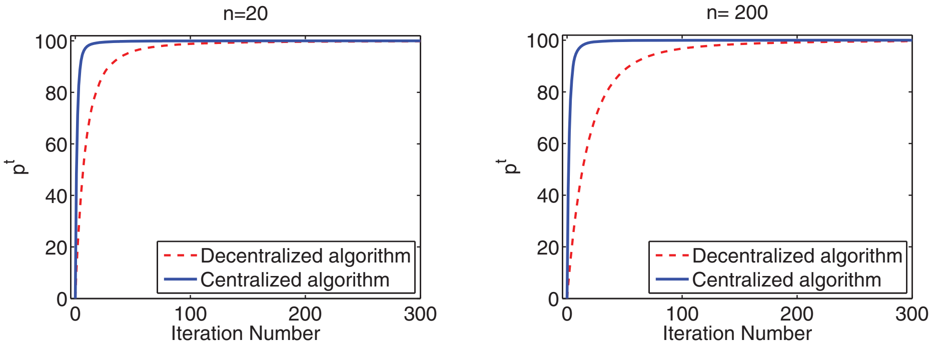

FIG. 2. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various n.*

paper. This problem arises naturally in sequential decision-making problems in teams of mobile agents [7]. Consider the following linear programming problem:

$$\begin{aligned} &\max_{x_i, i=1,\ldots,n} \frac{1}{n}\sum_{i=1}^{n} C_i^T x_i \\ &\text{s.t.} \quad \sum_{i=1}^{n} \mathbf{A_i} x_i \le B, \end{aligned} \tag{8}$$

where $C_i, x_i \in \Re^q$, $\mathbf{A_i} \in \Re^{m\times q}, i = 1,\ldots,n$, and $B \in \Re^m$. It can be seen that (8) belongs to the class of problems identified by (3). Recall that, for a given $\lambda_i \in \Re^m$, $f_i(\lambda_i)$ is the optimal value of the following optimization problem:

$$\begin{aligned} &\max_{x_i\in\Re^q} C_i^T x_i \\ &\text{s.t.} \quad \mathbf{A_i} x_i \le \lambda_i. \end{aligned} \tag{9}$$

Suppose that the dual feasible sets defined by $S_i = \{\nu_i | \mathbf{A_i^T}\nu_i = C_i, \nu_i \ge 0\}$ are nonempty and bounded. It is known from linear programming theory that $f_i(\lambda_i) = \min_{p=1,\ldots,P} \lambda_i^T \nu_{ip}$, where ν_{ip} are the extreme points of the polyhedra defined by S_i. Hence $f_i(\lambda_i)$ is nondifferentiable and concave. Further ν_i' is a subgradient of $f_i(\lambda_i)$ at λ_i if and only if it is an optimal solution to the dual problem [3]. Thus if S_i is bounded, it can be seen that Assumption 3.3 is satisfied, and the convergence analysis of section 3.2 holds.

Let the columns of $\mathbf{A_i}$ be denoted as $\mathbf{a_{ij}}$, $j = 1,\ldots,q$. Also let $C_i = [C_{ij}], j = 1,\ldots,q$. Suppose that the column $\mathbf{a_{ik}} > 0$ and $C_{ik} > 0$ for some k such that $1 \le k \le q$, and suppose S_i is nonempty. The corresponding dual constraint is $\mathbf{a_{ik}}^T\nu_i = C_{ik}$ showing that S_i is bounded. For the experiments we choose $\mathbf{a_{i1}} = \mathbf{1}$, $i = 1,\ldots,n$, where $\mathbf{1}$ is a vector of ones of the appropriate size. We also choose $C_{i1} = 200$, $i = 1,\ldots,n$. The rest of the constraint matrix and the cost vector are chosen arbitrarily while ensuring that S_i is nonempty.

Although the theoretical results require a randomization of the direction of update, it was observed that both of the decentralized and the centralized versions of the algorithm converge without the required randomization. Unlike the decentralized algorithm for the differentiable case, there is flexibility in choosing step sizes. It was observed in the experiments that the practical performance of both the centralized

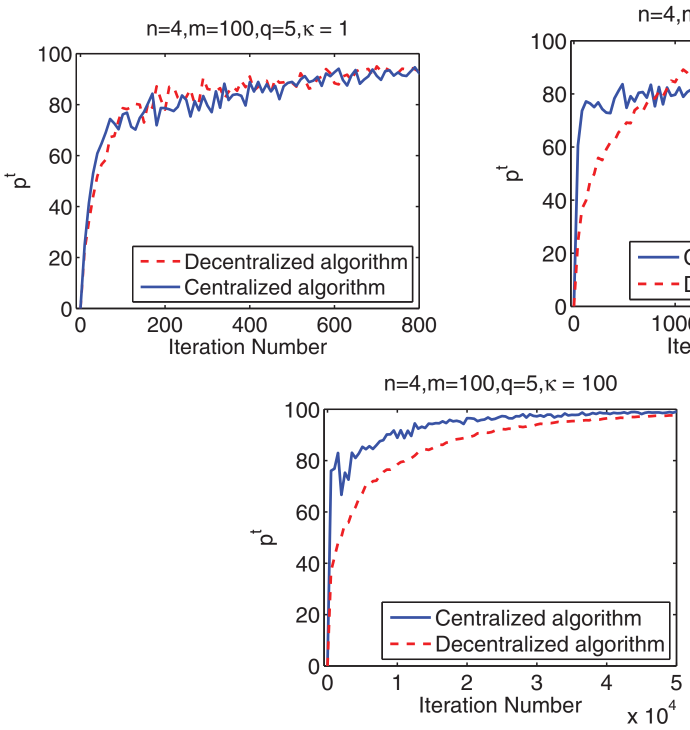

FIG. 3. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various κ.*

algorithm and the decentralized algorithm, with or without the randomization of the direction of update, depends dramatically on the choice of step sizes. We present the results of the experiments where the direction of update was not randomized, as it provides better insight into the convergence behavior of the algorithm. It was observed that convergence was obtained in this case so long as $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$. We choose step sizes of the form $\gamma^t = \frac{\theta(t)}{2L\sqrt{\log m+1}}\beta^t$, where L is the common Lipschitz constant of the functions f_i, $i = 1, \ldots, n$. Since $C_{i1} = 200$ and $\mathbf{a_{i1}} = \mathbf{1}$ for all i, it can be verified from the dual constraint $\mathbf{a_{i1}}^T \nu_i = C_{i1}$ that $L = C_{i1} = 200$. β^t was chosen to be of the form $\frac{1}{1+w(t)t^{0.51}}$. Thus $\theta(t)$ and $w(t)$ control the rate at which γ^t goes to 0. We chose $w(t)$ as a monotonically nondecreasing function bounded above, and $\theta(t)$ as a monotonically nonincreasing function bounded below. This ensures that $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$. For our experiments, we chose $w(0) = 0$ and $w(z\kappa + j) = w(z\kappa)$ for $z = 0, 1, \ldots, j = 1, 2, \ldots, \kappa - 1$, and $w((z+1)\kappa) = \min\{w(z\kappa) + r_w, w_{\max}\}$. For all of the experiments we chose $r_w = 0.0001$ and $w_{\max} = 0.1$. We also chose $\theta(t+1) = \max\{\theta(t) - r_\theta, \theta_{\min}\}$. For these experiments, we chose $\theta(0) = 30$, $\theta_{\min} = 3$, and $r_\theta = 0.1$. We ensured that the union of the communication graphs are connected periodically in the same manner as described in section 5.1. For these experiments, we choose $e_p = 0.5$. Figure 3 presents a comparison of the performance of the decentralized algorithm with the centralized algorithm, for varying κ. In the figures, n represents the number of agents, q represents the number of variables per agent, and m represents the number of constraints.

Figure 4 presents a comparison of the performance of the decentralized algorithm, with the centralized algorithm for varying n. All parameters except $\theta(0)$ were chosen

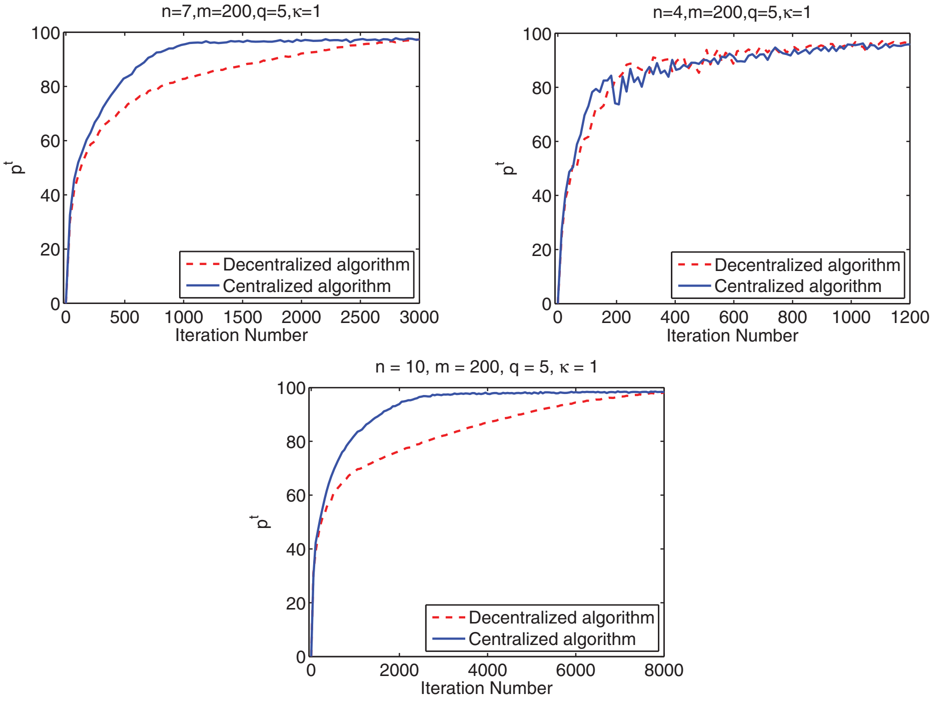

FIG. 4. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various n.*

as described previously. $\theta(0)$ was chosen to be 50 for the experiments of Figure 4. It can be observed that the performance of the decentralized algorithm scales well with increasing n. The numerical experiments suggest that κ has a greater effect on the practical performance of the algorithm than n.

6. Discussion.

In this paper, we proposed a decentralized gradient-descent algorithm for a general class of resource allocation problems. We first considered the case where the objective functions have Lipschitz-continuous gradients. We showed that the proposed algorithm converges and established that the rate at which the gradient projection converges to zero as a function of the number of agents in the network. Motivated by the need to develop decentralized algorithms for general convex optimization problems, we proposed a randomized subgradient-descent algorithm for the resource allocation problem with a possibly nondifferentiable objective function. We established an asymptotic convergence of the algorithm to a near-optimal solution and derived a convergence rate. Numerical experiments, both in the differentiable and in the nondifferentiable settings, suggested that the decentralized algorithms are competitive with the centralized versions of gradient and subgradient descent. The experiments also suggested that the performance of the decentralized randomized subgradient-descent algorithm depends dramatically on the choice of step sizes; how to set them up optimally while taking into account the structure of the communication network is a topic for future research.

An appealing feature of the developed algorithms is that the communication topology of the network of agents is allowed to be dynamic provided that the union of communication graphs of the agents is connected within a bounded time. This makes the algorithm particularly suitable in settings involving mobile agents or communication failures. We finally note that the formulation considered here goes beyond traditional resource allocation problems. In particular, we show in a related paper [7] that this algorithm is particularly suitable for a decentralized solution of a linear programming-based method for approximate dynamic programming for the problems of sequential decision making in the systems of mobile agents.

Appendix A. Proofs.

LEMMA 3.1. *A feasible solution λ^* of* (1) *is an optimal solution if and only if $\nabla f_i(\lambda_i^*) = \nabla f_j(\lambda_j^*)$ for all i, j.*

Proof. First note that we can eliminate one of the variables in (1) to make it unconstrained. For instance, if we let $\lambda_n = B - \sum_{i=1}^{n-1} \lambda_i$, (1) is equivalent to

$$\min_{\lambda} \bar{f}(\lambda) = \frac{1}{n} \sum_{i=1}^{n-1} \left(f_i(\lambda_i) + f_n \left(B - \sum_{i=1}^{n-1} \lambda_i \right) \right).$$

This is an unconstrained convex and differentiable optimization problem; hence a solution λ^* is optimal if and only if $\nabla \bar{f}(\lambda^*) = 0$. Noting that

$$\nabla_{\lambda_i} \bar{f}(\lambda^*) = \frac{1}{n} \left(\nabla f_i(\lambda_i^*) - \nabla f_n \left(B - \sum_{j=1}^{n-1} \lambda_j^* \right) \right),$$

we conclude that λ^* is optimal if and only if

$$\nabla f_i(\lambda_i^*) = \nabla f_j(\lambda_j^*) = \nabla f_n \left(B - \sum_{j=1}^{n-1} \lambda_j^* \right) = \nabla f_n(\lambda_n^*) \ \forall i, j < n. \qquad \square$$

LEMMA 3.2. *Suppose λ^1 is a feasible solution for* (1). *Then λ^t, where λ_i^t is defined by* (2), *is a feasible solution to* (1) *for all t.*

Proof. Suppose λ^t is a feasible solution for (1). Then

$$\begin{aligned} \sum_i \lambda_i^{t+1} &= \sum_i \lambda_i^t - \frac{\gamma}{n} \sum_i \sum_{j \in N_i(t)} (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)) \\ &= B - \frac{\gamma}{n} \sum_{(i,j) \in E(t)} (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t) + \nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)) \\ &= B. \end{aligned}$$

The second equality follows from the assumption that communication is symmetric. Thus λ^{t+1} is a feasible solution for (1), and the lemma follows by induction. $\square$

THEOREM 3.1. *Suppose that Assumptions* 3.1 *and* 3.2 *hold. With a step size of $\gamma = \frac{1}{2L}$,*

1. *the sequence $\{f(\lambda^t)\}$ is monotonically nondecreasing;*
2. *the sequence $\{\|\tilde{v}^{T_z}\|\}$ converges to* 0*;*
3. $\min_{z=1,\dots,p}(\|\tilde{v}^{T_z}\|^2) \leq \frac{3Ln^4\kappa(f(\lambda^*)-f(\lambda^1))}{4p}$ $\forall p$*;*
4. *if the set of optima is bounded, $\{f(\lambda^t)\}$ converges to $f(\lambda^*)$.*

The proof is based on a series of lemmas. Let the direction of update at time t be v^t. It can be seen from (2) that

$$v_i^t = \frac{1}{n} \sum_{j \in N_i(t)} (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)).$$

We first show that v^t is aligned to the direction of the gradient.

LEMMA A.1. $\nabla f(\lambda^t)^T v^t = \frac{1}{n^2}\sum_{(i,j)\in E(t)} \|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2$.

Proof. We have that

$$\begin{aligned}\nabla f(\lambda^t)^T v^t &= \sum_{i\in N}\frac{1}{n}\nabla f_i(\lambda_i^t)^T\left(\frac{1}{n}\sum_{j\in N_i(t)}\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\right)\\ &= \frac{1}{n^2}\sum_{i\in N}\nabla f_i(\lambda_i^t)^T\left(\sum_{j\in N_i(t)}\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\right). \qquad (10)\end{aligned}$$

Since communication is symmetric, for every term of the form $\nabla f_i(\lambda_i^t)^T(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))$ in the above summation, there is a corresponding term of the form $\nabla f_j(\lambda_j^t)^T (\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t))$. Hence,

$$\begin{aligned}\nabla f(\lambda^t)^T v^t &= \frac{1}{n^2}\sum_{(i,j)\in E(t)}\nabla f_i(\lambda_i^t)^T(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\\ &\quad + \nabla f_j(\lambda_j^t)^T(\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t))\\ &= \frac{1}{n^2}\sum_{(i,j)\in E(t)}\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2. \qquad \square\end{aligned}$$

We now prove a lemma that establishes a relationship between $\|v^t\|$ and $\nabla f(\lambda^t)^T v^t$. We can interpret $\gamma\nabla f(\lambda^t)^T v^t$ as the approximate increase in the objective of (1) when using the direction v^t and a sufficiently small step size γ.

LEMMA A.2. $\|v^t\|^2 \le 2n\nabla f(\lambda^t)^T v^t$.

Proof. Using the Cauchy–Schwarz inequality, $(\sum_{i=1}^k c_i)^2 \le k\sum_{i=1}^k c_i^2$,

$$\begin{aligned}\|v_i^t\|^2 &\le |N_i(t)|\sum_{j\in N_i(t)}\frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\\ &\le n\sum_{j\in N_i(t)}\frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2},\\ \|v^t\|^2 &= \sum_i\|v_i^t\|^2\\ &\le n\sum_i\sum_{j\in N_i(t)}\frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\\ &= 2n\sum_{(i,j)\in E(t)}\frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\\ &= 2n\nabla f(\lambda^t)^T v^t.\end{aligned}$$

The last equality comes from Lemma A.1. $\square$

We now prove a lemma that establishes a relationship between $\|\tilde{v}^t\|$ and $\nabla f(\lambda^t)^T\tilde{v}^t$.

LEMMA A.3. $\|\tilde{v}^t\|^2 = n\nabla f(\lambda^t)^T\tilde{v}^t$.

Proof. We first have that

$$\begin{aligned}
\|\tilde{v}_i^t\|^2 &= \frac{\| \sum_{j\in N}(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\|^2}{n^2}\\
&= \sum_{j\in N} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\\
&\quad + 2 \sum_{((j,l)\in N^2, j<l)} \frac{(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))^T(\nabla f_i(\lambda_i^t) - \nabla f_l(\lambda_l^t))}{n^2}\\
&= \sum_{j\in N} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\\
&\quad + \sum_{((j,l)\in N^2, j<l)}\\
&\quad \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2 + \|\nabla f_i(\lambda_i^t) - \nabla f_l(\lambda_l^t)\|^2 - \|\nabla f_j(\lambda_j^t) - \nabla f_l(\lambda_l^t)\|^2}{n^2}\\
&= (n-1)\left(\sum_{j\in N} \frac{\|(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\|^2}{n^2}\right)\\
&\quad - \sum_{((j,l)\in N^2, j<l,(j,l\neq i))} \frac{\|\nabla f_j(\lambda_j^t) - \nabla f_l(\lambda_l^t)\|^2}{n^2},\\
\|\tilde{v}^t\|^2 &= \sum_{i\in N} \|\tilde{v}_i^t\|^2\\
&= \sum_{i\in N}\left((n-1)\left(\sum_{(j\in N)} \frac{\|(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\|^2}{n^2}\right)\right.\\
&\qquad \left. - \sum_{((j,l)\in N^2, j<l, j,l\neq i)} \frac{\|\nabla f_j(\lambda_j^t) - \nabla f_l(\lambda_l^t)\|^2}{n^2}\right).
\end{aligned}$$

We note that $\|\tilde{v}^t\|^2 = \sum_{((i,j)\in N^2, i<j)} c_{ij}((\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2)/n^2)$. To determine c_{ij}, note that the term $\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2$ appears with a coefficient $(n-1)$ in $\|\tilde{v}_i^t\|^2$, $(n-1)$ in $\|\tilde{v}_j^t\|^2$, and with a coefficient -1 in $\|\tilde{v}_k^t\|^2$ for all $(k \in N, k \neq i, j)$. Hence, $c_{ij} = (n-1) + (n-1) - (n-2) = n$. Therefore,

$$\begin{aligned}
\|\tilde{v}^t\|^2 &= \sum_{((i,j)\in N^2, i<j)} c_{ij} \left(\frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\right)\\
&= n \sum_{((i,j)\in N^2, i<j)} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\\
&= n\nabla f(\lambda^t)^T \tilde{v}^t. \qquad \square
\end{aligned}$$

Consider a decentralized direction of update v^t derived from an arbitrary connected graph $G = (N, E(t))$. We now compare the ratio of the approximate increase in the objective of (1) using v^t as the direction of update and for a sufficiently small step size γ to the approximate increase in the objective using $\tilde{v}^t$ as the direction of update for the same step size. This ratio is given by

$$\frac{(\nabla f(\lambda^t)^T v^t)}{(\nabla f(\lambda^t)^T \tilde{v}^t)} = \frac{\sum_{(i,j)\in E(t)}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)}{\sum_{((i,j)\in N^2, i<j)}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)}. \tag{11}$$

The following lemma shows that this ratio is bounded away from 0 by a factor that depends only on the number of agents.

Lemma A.4. *For all connected graphs $G = (N, E)$,*

$$\sum_{(i,j)\in E} \|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2 \geq \frac{8}{n^3} \sum_{((i,j)\in N^2, i<j)} \|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2.$$

For any vector X, let $(X)_k$ denote its kth component. We note that $(\sum_{(i,j)\in E}\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)/(\sum_{((i,j)\in N^2,i<j)}\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)$ is of the form $(\sum_{k=1}^m b_k)/(\sum_{k=1}^m c_k)$, where $b_k = \sum_{(i,j)\in E}((\nabla f_j(\lambda_j^t))_k - (\nabla f_i(\lambda_i^t))_k)^2$ and $c_k = \sum_{((i,j)\in N^2,i<j)}((\nabla f_j(\lambda_j^t))_k - (\nabla f_i(\lambda_i^t))_k)^2$, and we recall that $\lambda_j^t \in \Re^m$ for all $j \in N$. Let $r_k = \frac{b_k}{c_k}$. We show that if $c_k > 0$, then $r_k \geq \frac{8}{n^3}$. We define $r(E) = (\sum_{(i,j)\in E}(p_j - p_i)^2)/(\sum_{((i,j)\in N^2,i<j)}(p_j - p_i)^2)$ for arbitrary values of the scalars p_i, $i = 1, \dots, n$, such that $\sum_{((i,j)\in N^2,i<j)}(p_j - p_i)^2 > 0$. We show that $r(E) \geq \frac{8}{n^3}$, which establishes that when $c_k > 0$, $r_k \geq \frac{8}{n^3}$. This result is based on a series of lemmas. We first establish that, for any fixed value of $p_i, i = 1, \dots, n$, the worst possible value of r is achieved when G corresponds to a chain whose nodes have monotone values of p_i. Then we compute the worst possible value of r with respect to possible values of p_i.

We also assume that $p_i \neq p_j$ for all $i \neq j$, without loss of generality; since $\sum_{((i,j)\in N^2,i<j)}(p_j - p_i)^2 > 0$ by assumption, for any set of values $p_i, i = 1, \dots, n$, we can always perturb the values to make them strictly distinct while making $r(E)$ in the resulting graph arbitrarily close to that in the original problem.

Lemma A.5. *The graph $G = (N, E)$ that minimizes r over all possible sets E, under the constraint that G is a connected graph, is a tree.*

Proof. Take an arbitrary graph (N, E), and suppose that it is not a tree. Then we can convert it into a tree (N, E') by removing some edges from E. It is clear that $r(E') \leq r(E)$, therefore (N, E) cannot be optimal. □

Lemma A.6. *If a certain graph (N, E) contains edges ij and jk such that $p_j < \min(p_i, p_k)$ or $p_j > \max(p_i, p_k)$, then it does not minimize r.*

Proof. Consider the first situation and suppose, without loss of generality, that $p_j < p_i < p_k$. Let $E' = E\backslash\{jk\} \cup \{ik\}$. The difference in the numerator of $r(E)$ and $r(E')$ is equal to $(p_j - p_k)^2 - (p_i - p_k)^2$, which is greater than 0. Therefore (N, E) cannot be optimal. A similar analysis holds when $p_j > \max(p_i, p_k)$. □

Lemma A.7. *If a node j contains more than two neighbors, then it has two neighbors i and k such that $p_j < \min(p_i, p_k)$ or $p_j > \max(p_i, p_k)$.*

Proof. Suppose that i, k, and l are neighbors of j. Then at least two among the three values p_i, p_k, and p_l must be less than or greater than p_j. □

Lemma A.8. *Consider the chain that links nodes $1, \dots, n$ in increasing order of p_i. Then it minimizes r over all possible connected graphs.*

Proof. From the previous lemmas, we conclude that the optimal graph is a tree. Moreover, each node in the optimal tree must have at most two neighbors. We conclude that the optimal graph is a chain. From Lemma A.6, the nodes in the chain are in increasing or decreasing order of p_i, and the lemma follows. □

Proof of Lemma A.4. Without loss of generality, suppose that $p_1 < p_2 < \cdots < p_n$. Let $\Delta_i = p_{i+1} - p_i$. Note that, for all $j > i$, $p_j - p_i = \sum_{k=i}^{j-1} \Delta_k$. In view of the previous lemmas, we have the following for every connected graph (N, E):

$$\begin{aligned}
r(E) &= \frac{\sum_{(i,j)\in E}(p_i - p_j)^2}{\sum_{((i,j)\in N^2, i<j)}(p_i - p_j)^2} \\
&\geq \frac{\sum_{i=1}^{n-1}(p_{i+1} - p_i)^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(p_i - p_j)^2} = \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\sum_{k=i}^{j-1}\Delta_k\right)^2} \\
&\geq \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=i}^{j-1}(j-i)\Delta_k^2} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{i=1}^{n-1}\sum_{k=i}^{n-1}\sum_{j=k+1}^{n}(j-i)\Delta_k^2} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\sum_{i=1}^{k}\sum_{j=k+1}^{n}(j-i)} = \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\sum_{i=1}^{k}\sum_{j=1}^{n-k}(j+k-i)} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\sum_{i=1}^{k}\frac{(n-k)(n-k+1)}{2} + (k-i)(n-k)} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2(\frac{k(n-k)(n-k+1)}{2} + \frac{(n-k)(k)(k-1)}{2})} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\frac{k(n-k)(n)}{2}} \\
&\geq \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\frac{n^3}{8}} \\
&= \frac{8}{n^3}.
\end{aligned}$$

The second inequality follows from the Cauchy–Schwarz inequality.

We note from the definitions that if $c_k = 0$, then $b_k = 0$. Thus for $k = 1, \dots, m$, either $b_k = c_k = 0$, or $r_k \geq \frac{8}{n^3}$. The Lemma is trivially true, if for $k = 1, \dots, m$, $b_k = c_k = 0$. Suppose there exist some $\bar{k} \in (1, \dots, m)$ such that $c_{\bar{k}} > 0$. Let K be the set of integers from 1 to m such that $c_k > 0$ for $k \in K$. K is not empty since it contains $\bar{k}$.

$$\begin{aligned}
\frac{\sum_{(i,j)\in E}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)}{\sum_{((i,j)\in N^2, i<j)}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)} &= \frac{\sum_{k=1}^{m} b_k}{\sum_{k=1}^{m} c_k} \\
&= \frac{\sum_{k\in K} b_k}{\sum_{k\in K} c_k} \\
&\geq \frac{\sum_{k\in K}\frac{8}{n^3}c_k}{\sum_{k\in K} c_k} \\
&= \frac{8}{n^3}. \qquad \square
\end{aligned}$$

Let E_{T_z} be a subset of the edge set $E_{T_z,T_{z+1}}$ such that the graph (N, E_{T_z}) is a tree. By Assumption (2.1), the graph $(N, E_{T_z,T_{z+1}})$ is connected, and so E_{T_z} is well defined. Let the decentralized direction of update derived using $G = (N, E_{T_z})$ be denoted by $\bar{v}^{T_z}$. The following lemma shows that the approximate increase in the objective in period $[T_z, T_{z+1}]$ using the direction of update v^t and a sufficiently small step size γ is comparable to the approximate increase in objective when the direction $\bar{v}^{T_z}$ is used for update at time T_z.

LEMMA A.9. $\nabla f(\lambda^{T_z})^T \bar{v}^{T_z} \leq \frac{3}{2}\kappa \sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t$.

Proof. We have that

$$\begin{aligned} \|\nabla f(\lambda^{t+1}) - \nabla f(\lambda^t)\|^2 &\leq \frac{L^2}{n^2}\|\gamma v^t\|^2 = \frac{1}{4n^2}\|v^t\|^2 \\ &\leq \frac{1}{4n^2} 2n(\nabla f(\lambda^t)^T v^t) = \frac{1}{2n}\nabla f(\lambda^t)^T v^t. \end{aligned} \tag{12}$$

The first inequality is true because of Assumption 3.2. The first equality is true because $\gamma = \frac{1}{2Ln}$. The second inequality follows from Lemma A.2. Let $t^i_{T_z}$ be the earliest time between time periods T_z and $T_{z+1} - 1$ such that there is an edge $(i, j) \in E_{T_z}$ for agent i. It is clear that $T_z \leq t^i_{T_z} \leq T_{z+1} - 1$. Also, by definition, for $l = T_z, T_z + 1, \ldots, (t^i_{T_z} - 1)$, there is no edge $(i, p) \in E(l)$. Thus $\lambda_i^{t^i_{T_z}} = \lambda_i^{T_z}$, and $\nabla f_i(\lambda_i^{t^i_{T_z}}) = \nabla f_i(\lambda_i^{T_z})$. Letting $w_{ij}(t) = \frac{1}{n}(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))$, we have that

$$\begin{aligned} \|w_{ij}(T_z)\| &= \frac{1}{n}\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{T_z})\| \\ &\leq \frac{1}{n}(\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\| + \|\nabla f_j(\lambda_j^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{T_z})\|) \\ &\leq \frac{1}{n}\left(\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\| + \sum_{t=T_z}^{t^i_{T_z}-1} \|\nabla f_j(\lambda_j^{t+1}) - \nabla f_j(\lambda_j^t)\|\right). \end{aligned}$$

From the Cauchy–Schwarz inequality,

$$\begin{aligned} \|w_{ij}(T_z)\|^2 \;&\leq\; \frac{(t^i_{T_z} - T_z + 1)}{n^2}\left(\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2\right. \\ &\qquad\qquad \left. + \sum_{t=T_z}^{t^i_{T_z}-1} \|\nabla f_j(\lambda_j^{t+1}) - \nabla f_j(\lambda_j^t)\|^2\right) \\ &\leq\; \kappa\left(\frac{\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2}{n^2} + \sum_{t=T_z}^{t^i_{T_z}-1} \frac{\|\nabla f_j(\lambda_j^{t+1}) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\right) \\ &\leq\; \kappa\left(\frac{\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2}{n^2} + \frac{1}{2n}\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t\right). \end{aligned}$$

The last inequality comes from (12) and from the fact that $\|\nabla f(\lambda^{t+1}) - \nabla f(\lambda^t)\|^2 = \sum_{i=1}^n \frac{\|\nabla f(\lambda_i^{t+1}) - \nabla f(\lambda_i^t)\|^2}{n^2}$. We finally have that

$$\begin{aligned}\nabla f(\lambda^{T_z})^T \bar{v}^{T_z} &= \sum_{(i,j)\in E_{T_z}} \|w_{ij}(T_z)\|^2 \\ &\le \sum_{(i,j)\in E_{T_z}} \kappa \left(\frac{\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2}{n^2} + \frac{1}{2n} \sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right) \\ &\le \kappa \left(\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t + \frac{n-1}{2n} \sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right) \\ &\le \frac{3}{2}\kappa \left(\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right).\end{aligned}$$

The first equality comes from Lemma A.1, with v^t replaced by $\bar{v}^{T_z}$. The second inequality comes from the fact that E_{T_z} is a subset of $E_{T_z,T_{z+1}}$ and from Lemma A.1. It is clear that Lemma A.1 is valid for all decentralized directions of update v derived using some communication graph G, where $v_i = \sum_{j\in N(i)} \frac{1}{n}(\nabla f_i(\lambda_i) - \nabla f_j(\lambda_j))$ and $N(i)$ is the set of neighbors of i in G. Hence Lemma A.1 is valid for $\bar{v}^{T_z}$. The second inequality holds because of Lemma A.1 and because there are exactly $n-1$ edges in the set E_{T_z}, as $G = (N, E_{T_z})$ is a tree. □

Proof of Theorem 3.1.

Proof of 1 *of Theorem* 3.1. First note that

$$\begin{aligned}f(\lambda^{t+1}) - f(\lambda^t) &\ge \gamma \nabla f(\lambda^t)^T v^t - \frac{L}{2n}\|\gamma v^t\|^2 \\ &\ge \gamma \nabla f(\lambda^t)^T v^t - \frac{\gamma^2 L}{2n} 2n \nabla f(\lambda^t)^T v^t \\ &= \frac{1}{2L}\nabla f(\lambda^t)^T v^t - \frac{1}{4L}\nabla f(\lambda^t)^T v^t = \frac{1}{4L}\nabla f(\lambda^t)^T v^t. \qquad (13)\end{aligned}$$

The first inequality comes from the descent lemma for differentiable functions [2]. The second inequality comes from Lemma A.2. The first equality comes from the fact that $\gamma = \frac{1}{2L}$. Since $\nabla f(\lambda^t)^T v^t$ is nonnegative, the sequence $\{f(\lambda^t)\}$ is monotonic and nondecreasing establishing the first part of the theorem.

Proof of 2 *of Theorem* 3.1. Since (1) is assumed to have an optimal solution, $f(\lambda^t)$ is bounded from above. We conclude from the first claim that $\{f(\lambda^t)\}$ converges, and $\{\nabla f(\lambda^t)^T v^t\}$ must converge to zero.

We now have that

$$\begin{aligned}\|\tilde{v}^{T_z}\|^2 &= n \nabla f(\lambda^{T_z})^T \tilde{v}^{T_z} \\ &\le \frac{n^4}{8} \nabla f(\lambda^{T_z})^T \bar{v}^{T_z} \le \frac{3n^4}{16}\kappa \left(\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right),\end{aligned}$$

where the first equality follows from Lemma A.3, the first inequality follows from Lemma A.4 and Lemma A.1, and the second inequality follows from Lemma A.9. The last inequality and the convergence of $\{\nabla f(\lambda^t)^T v^t\}$ to zero establishes the second part of the theorem.

Proof of 3 *of Theorem* 3.1. Note that

$$\begin{aligned} f(\lambda^{T_{z+1}}) - f(\lambda^{T_z}) &\geq \frac{1}{4L} \sum_{t=T_z}^{t=T_{z+1}-1} \nabla f(\lambda^t)^T v^t \\ &\geq \frac{1}{6L\kappa} \nabla f(\lambda^{T_z})^T \bar{v}^{T_z} \\ &\geq \frac{4}{3Ln^3\kappa} \nabla f(\lambda^{T_z})^T \tilde{v}^{T_z} \\ &= \frac{4}{3Ln^4\kappa} \|\tilde{v}^{T_z}\|^2. \end{aligned}$$

The first inequality comes from (13). The second inequality comes from Lemma A.9. The third inequality comes from Lemma A.4 and Lemma A.1, and the equality comes from Lemma A.3. Thus,

$$\begin{aligned} \sum_{z=1}^{p} f(\lambda^{T_{z+1}}) - f(\lambda^{T_z}) &\geq \frac{4}{3Ln^4\kappa} \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2 \\ f(\lambda^{T_{p+1}}) - f(\lambda^1) &\geq \frac{4}{3Ln^4\kappa} \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2 \\ f(\lambda^*) - f(\lambda^1) &\geq \frac{4}{3Ln^4\kappa} \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2. \end{aligned}$$

The last inequality, together with the fact that $p(\min_{z=1,\ldots,p} \|\tilde{v}^{T_z}\|^2) \leq \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2$, proves the third claim.

Proof of 4 *of Theorem* 3.1. If the set of optima of (1) is bounded, $\{\lambda : \|\tilde{v}(\lambda)\| \leq C\}$ is a bounded set for some $C > 0$. We conclude that λ^{T_z} has a converging subsequence $\lambda^{T_{z_k}}$. Let $\bar{\lambda}$ be the limit of $\lambda^{T_{z_k}}$. Since $\|\tilde{v}(\cdot)\|$ is a continuous function and $\|\tilde{v}(\lambda^{T_{z_k}})\|$ converges to zero, we conclude that $\tilde{v}(\bar{\lambda}) = 0$ and $\bar{\lambda}$ is optimal. Since f is continuous, we conclude that $\{f(\lambda^{T_{z_k}})\}$ converges to $f(\bar{\lambda}) = f(\lambda^*)$. Since $\{f(\lambda^t)\}$ converges, we conclude that it must converge to $f(\lambda^*)$. □

LEMMA 3.3. *Let f_i and $\hat{f}_i$ be as defined in section* 3.2. *Then the following hold:*

1. *$\hat{f}_i$ is concave and differentiable, with gradient $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i)]$;*
2. *$f_i(\lambda_i) \geq \hat{f}_i(\lambda_i) \geq f_i(\lambda_i) - 2.8\epsilon L$;*
3. *$\|\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_i(\bar{\lambda}_i)\| \leq \frac{\sqrt{\log(m+1)}L}{\epsilon} \|\lambda_i - \bar{\lambda}_i\|$.*

Proof of 1 *of Lemma* 3.3. For all $a \in [0, 1]$, we have that

$$\begin{aligned} \hat{f}_i(a\lambda_i + (1-a)\bar{\lambda}_i) &= \mathrm{E}[f_i(a\lambda_i + (1-a)\bar{\lambda}_i + Z_i)] \\ &\geq \mathrm{E}[af_i(\lambda_i + Z_i) + (1-a)f_i(\bar{\lambda}_i + Z_i)] = a\hat{f}_i(\lambda_i) + (1-a)\hat{f}_i(\bar{\lambda}_i). \end{aligned}$$

It follows that $\hat{f}_i$ is concave. Since f_i is nondifferentiable only on a set of measure zero, we have that

$$\begin{aligned} (\nabla f_i(\lambda_i + Z_i))_j &= \lim_{\delta \uparrow 0} \frac{f_i(\lambda_i + Z_i + \delta e_j) - f_i(\lambda_i + Z_i)}{\delta} \\ &= \lim_{\delta \downarrow 0} \frac{f_i(\lambda_i + Z_i + \delta e_j) - f_i(\lambda_i + Z_i)}{\delta}, \end{aligned}$$

with probability 1, where e_j is the vector with all entries equal to zero except for the jth entry, which is equal to one. Hence

$$\begin{aligned}\lim_{\delta\uparrow 0}\frac{\mathrm{E}[f_i(\lambda_i+Z_i+\delta e_j)]-\mathrm{E}[f_i(\lambda_i+Z_i)]}{\delta} &= \mathrm{E}\left[\lim_{\delta\uparrow 0}\frac{f_i(\lambda_i+Z_i+\delta e_j)-f_i(\lambda_i+Z_i)}{\delta}\right]\\ &= \mathrm{E}[(\nabla f_i(\lambda_i+Z_i))_j]\\ &= \mathrm{E}\left[\lim_{\delta\downarrow 0}\frac{f_i(\lambda_i+Z_i+\delta e_j)-f_i(\lambda_i+Z_i)}{\delta}\right]\\ &= \lim_{\delta\downarrow 0}\frac{\mathrm{E}[f_i(\lambda_i+Z_i+\delta e_j)]-\mathrm{E}[f_i(\lambda_i+Z_i)]}{\delta}.\end{aligned}$$

Note that $|\frac{f_i(\lambda_i+Z_i+\delta e_j)-f_i(\lambda_i+Z_i)}{\delta}| \le L$. Hence the exchanges between limit and expectation are valid by the bounded convergence theorem. It follows that $\hat{f}_i$ is differentiable, and its gradient is given by

$$\nabla\hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i+Z_i)].$$

Proof of 2 *of Lemma* 3.3. First, we have that

$$\begin{aligned}\hat{f}_i(\lambda_i) &= \mathrm{E}[f_i(\lambda_i+Z_i)]\\ &\le f_i(\lambda_i+EZ_i) = f_i(\lambda_i),\end{aligned}$$

where the inequality follows from the concavity of f_i and Jensen's inequality [4].

For the lower bound on $\hat{f}_i$, we have that

$$\begin{aligned}\hat{f}_i(\lambda_i) &= \mathrm{E}[f_i(\lambda_i+Z_i)]\\ &= \mathrm{E}[f_i(\lambda_i-Z_i)]\\ &\ge \mathrm{E}[f_i(\lambda_i)-Z_i^T\nabla f_i(\lambda_i-Z_i)]\\ &\ge f_i(\lambda_i)-\mathrm{E}[\max_j|Z_{ij}|]L, \qquad (14)\end{aligned}$$

where $|Z_{ij}|$ is the modulus function. The first inequality follows from the concavity of f and the fact that $\nabla f_i(\lambda_i-Z_i)$ is a subgradient of f at λ_i-Z_i. The second inequality follows from the fact that $\|\nabla f_i(\lambda_i-Z_i)\|_1 \le L$.

We now show that $\mathrm{E}[\max_j|Z_{ij}|] \le 2.8\epsilon$. Note that this inequality and (14) prove the claim.

We first place a bound on $P(|Z_{ij}|>c)$, for $c>0$. We have that

$$\begin{aligned}P(|Z_{ij}|>c) &= \int_c^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz + \int_{-\infty}^{-c}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz = 2\int_c^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz\\ &= 2e^{-\frac{c^2}{2\sigma^2}}\int_0^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2+2zc}{2\sigma^2}}dz\\ &= 2e^{-\frac{c^2}{2\sigma^2}}\left(\int_0^c\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2+2zc}{2\sigma^2}}dz+\int_c^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2+2zc}{2\sigma^2}}dz\right)\\ &\le 2e^{-\frac{c^2}{2\sigma^2}}\left(\int_0^c\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{zc}{\sigma^2}}dz+\int_c^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz\right)\\ &= 2e^{-\frac{c^2}{2\sigma^2}}\left(\frac{\sigma}{\sqrt{2\pi}c}\left(1-e^{-\frac{c^2}{\sigma^2}}\right)+\frac{1}{2}P(|Z_{ij}|>c)\right).\end{aligned}$$

Hence

$$P(|Z_{ij}| > c) \leq 2 \left(\frac{\frac{e^{-\frac{c^2}{2\sigma^2}}\sigma}{\sqrt{2\pi}c}\left(1 - e^{-\frac{c^2}{\sigma^2}}\right)}{1 - e^{-\frac{c^2}{2\sigma^2}}} \right)$$
$$= \frac{2e^{-\frac{c^2}{2\sigma^2}}\sigma}{\sqrt{2\pi}c}\left(1 + e^{-\frac{c^2}{2\sigma^2}}\right) \leq \frac{2e^{-\frac{c^2}{2\sigma^2}}\sqrt{2}\sigma}{\sqrt{\pi}c}.$$

It follows that, for all $c \geq \frac{2\epsilon}{\sqrt{\pi}}$,

$$P(\max_j |Z_{ij}| > c) \leq 2m\frac{e^{-\frac{c^2}{2\sigma^2}}\sqrt{2}\sigma}{\sqrt{\pi}c} \leq 2(m+1)\frac{e^{-\frac{c^2\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}}$$
$$= \frac{2e^{-\frac{\left(c^2 - \frac{4\epsilon^2}{\pi}\right)\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}} \leq \frac{2e^{-\frac{\left(c - \frac{2\epsilon}{\sqrt{\pi}}\right)^2\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}}.$$

The first inequality follows from the union bound [4]. The second inequality follows from $c \geq \frac{2\epsilon}{\sqrt{\pi}}$. The last inequality follows from $(c - \frac{2\epsilon}{\sqrt{\pi}})^2 \leq c^2 - \frac{4\epsilon^2}{\pi}$ for all $c > \frac{2\epsilon}{\sqrt{\pi}}$.

Finally,

$$\begin{aligned}
&\mathrm{E}[\max_j |Z_{ij}|] \\
&\quad= \int_0^\infty P(\max_j |Z_{ij}| > z)dz \\
&\quad= \int_0^{\frac{2\epsilon}{\sqrt{\pi}}} P(\max_j |Z_{ij}| > z)dz + \int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty P(\max_j |Z_{ij}| > z)dz \\
&\quad\leq \frac{2\epsilon}{\sqrt{\pi}} + \int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty \frac{2e^{-\frac{\left(z-\frac{2\epsilon}{\sqrt{\pi}}\right)^2\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}}dz \\
&\quad= \frac{2\epsilon}{\sqrt{\pi}} + \frac{4\epsilon}{\sqrt{\pi}\log(m+1)}\int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty \frac{\sqrt{\log(m+1)}}{2\epsilon}e^{-\frac{\left(z-\frac{2\epsilon}{\sqrt{\pi}}\right)^2\pi\log(m+1)}{4\epsilon^2}}dz \\
&\quad= \frac{2\epsilon}{\sqrt{\pi}} + \frac{4\epsilon}{\sqrt{\pi}\log(m+1)}\int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\pi\log(m+1)}}{\sqrt{2}\epsilon}e^{-\frac{\left(\frac{\sqrt{\pi\log(m+1)}\left(z-\frac{2\epsilon}{\sqrt{\pi}}\right)}{\sqrt{2}\epsilon}\right)^2}{2}}dz \\
&\quad= \frac{2\epsilon}{\sqrt{\pi}} + \frac{2\epsilon}{\sqrt{\pi}\log(m+1)} \\
&\quad\leq 2.8\epsilon.
\end{aligned}$$

The last equality comes from the identity $\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{t^2}{2}}dt = \frac{1}{2}$.

Proof of 3 *of Lemma* 3.3. Denote by $p(\cdot)$ the probability density function for Z_i; i.e., the joint probability density function for $Z_{i1}, \ldots, Z_{im}$. Then we have that

$$\nabla\hat{f}_i(\lambda_i) = \int_{\Re^m} p(z)\nabla f_i(\lambda_i + z)dz$$
$$= \int_{\Re^m} p(z - \lambda_i)\nabla f_i(z)dz.$$

It follows that

$$\begin{aligned}\|\nabla\hat{f}_i(\lambda_i)-\nabla\hat{f}_i(\bar{\lambda}_i)\| &= \left\|\int_{\Re^m}(p(z-\lambda_i)-p(z-\bar{\lambda}_i))\nabla f_i(z)dz\right\| \\ &\leq \int_{\Re^m}|p(z-\lambda_i)-p(z-\bar{\lambda}_i)|\|\nabla f_i(z)\|dz \\ &\leq L\int_{\Re^m}|p(z-\lambda_i)-p(z-\bar{\lambda}_i)|dz. \qquad (15)\end{aligned}$$

Since $p(\cdot)$ is the joint distribution of m i.i.d. zero-mean Gaussian random variables, $p(z)$ is strictly decreasing on $\|z\|$. Hence

$$\begin{aligned}&\int_{\Re^m}|p(z-\lambda_i)-p(z-\bar{\lambda}_i)|dz \\ &\quad= \int_{\{z\in\Re^m:\|z-\lambda_i\|<\|z-\bar{\lambda}_i\|\}}\big(p(z-\lambda_i)-p(z-\bar{\lambda}_i)\big)\,dz \\ &\qquad+\int_{\{z\in\Re^m:\|z-\lambda_i\|>\|z-\bar{\lambda}_i\|\}}\big(p(z-\bar{\lambda}_i)-p(z-\lambda_i)\big)\,dz \\ &\quad= 2\int_{\{z\in\Re^m:\|z-\lambda_i\|<\|z-\bar{\lambda}_i\|\}}\big(p(z-\lambda_i)-p(z-\bar{\lambda}_i)\big)\,dz \\ &\quad= 2\int_{\{z\in\Re^m:\|z\|<\|z-(\bar{\lambda}_i-\lambda_i)\|\}}p(z)dz-2\int_{\{z\in\Re^m:\|z\|>\|z-(\lambda_i-\bar{\lambda}_i)\|\}}p(z)dz \\ &\quad= 2P(\|Z_i\|<\|Z_i-(\bar{\lambda}_i-\lambda_i)\|)-2P(\|Z_i\|>\|Z_i-(\lambda_i-\bar{\lambda}_i)\|) \\ &\quad= 2P(2Z_i^T(\bar{\lambda}_i-\lambda_i)<\|\bar{\lambda}_i-\lambda_i\|^2)-2P(2Z_i^T(\bar{\lambda}_i-\lambda_i)<-\|\bar{\lambda}_i-\lambda_i\|^2) \\ &\quad= 2P(-0.5\|\bar{\lambda}_i-\lambda_i\|<V<0.5\|\bar{\lambda}_i-\lambda_i\|), \qquad (16)\end{aligned}$$

where

$$V=\frac{Z_i^T(\bar{\lambda}_i-\lambda_i)}{\|\bar{\lambda}_i-\lambda_i\|}.$$

It is easy to verify that V is normal with a zero mean and variance equal to $\sigma=\frac{\sqrt{2}\epsilon}{\sqrt{\pi\log(m+1)}}$. It follows that

$$\begin{aligned}P(-0.5\|\bar{\lambda}_i-\lambda_i\|<V<0.5\|\bar{\lambda}_i-\lambda_i\|) &\leq \frac{1}{\sqrt{2\pi}\sigma}\|\bar{\lambda}_i-\lambda_i\| \\ &= \frac{\sqrt{\log(m+1)}}{2\epsilon}\|\bar{\lambda}_i-\lambda_i\|. \qquad (17)\end{aligned}$$

The claim follows from (15), (16), and (17). □

THEOREM 3.2. *Suppose that Assumptions* 3.3 *and* 3.4 *hold. Then with probability* 1*:*

1. *the sequence* $\{\|\tilde{v}^{T_z}\|\}$ *converges to* 0*;*
2. $\min_{z=1,\dots,p}\mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \frac{\frac{n^4\kappa L\sqrt{\log(m+1)}}{\epsilon}\left[3(f(\lambda^*)-f(\lambda^1)+2.8\epsilon L)+\sum_{t=1}^{t=\kappa p}\frac{4L\beta_t^2\epsilon}{\sqrt{\log(m+1)}}\right]}{4\sum_{z=2}^{p+1}\beta_{\kappa z}}$ $\forall p$*;*
3. *if the set of the optima of* (1) *is bounded, then* $\lim_{t\to\infty}f(\lambda^t)\geq f(\lambda^*)-2.8\epsilon L$*.*

The proof has the same structure as the proof of Theorem 3.1. Let the expected direction of update at time t be v^t:

$$v_i^t = \frac{1}{n} \sum_{j \in N_i(t)} (\nabla \hat{f}_i(\lambda_i^t) - \nabla \hat{f}_j(\lambda_j^t)).$$

Let δ^t be the random variable denoting the difference between the actual and the expected directions of update:

$$\delta_i^t = \sum_{j \in N_i(t)} \frac{1}{n} (\nabla f_i(\lambda_i^t + Z_i^t) - \nabla f_j(\lambda_j^t + Z_j^t)) - v_i^t.$$

Let $\mathcal{F}_t$ be the sigma-algebra [4] generated by $Z_i^\tau,\ i = 1, \dots, n, \tau = 1, \dots, t$. We have the following result about δ_i^t.

LEMMA A.10. *For all* t, $\mathrm{E}[\delta^t | \mathcal{F}_{t-1}] = 0$ *and* $\mathrm{E}[\|\delta^t\|^2 | \mathcal{F}_{t-1}] < 8nL^2$, *with probability* 1.

Proof. $\mathrm{E}[\delta_i^t | \mathcal{F}_{t-1}] = 0$ follows from $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i^t)]$ for all i. Moreover,

$$\begin{aligned}
&\mathrm{E}[\|\delta_i^t\|^2 | \mathcal{F}_{t-1}] \\
&= \mathrm{E}\left[\left\| \sum_{j \in N_i(t)} \frac{\nabla f_i(\lambda_i^t + Z_i^t) - \nabla \hat{f}_i(\lambda_i^t) - \nabla f_j(\lambda_j^t + Z_j^t) + \nabla \hat{f}_j(\lambda_j^t)}{n} \right\|^2 | \mathcal{F}_{t-1} \right] \\
&= \frac{\mathrm{E}\left[\left\| N_i(t) \left(\nabla f_i(\lambda_i^t + Z_i^t) - \nabla \hat{f}_i(\lambda_i^t) \right) - \sum_{j \in N_i(t)} \left(\nabla f_j(\lambda_j^t + Z_j^t) - \nabla \hat{f}_j(\lambda_j^t) \right) \right\|^2 | \mathcal{F}_{t-1} \right]}{n^2} \\
&\leq \frac{N_i(t)^2 \mathrm{E}[\|\nabla f_i(\lambda_i^t + Z_i^t) - \nabla \hat{f}_i(\lambda_i^t)\|^2 | \mathcal{F}_{t-1}] + \sum_{j \in N_i(t)} \mathrm{E}[\|\nabla f_j(\lambda_j^t + Z_j^t) - \nabla \hat{f}_j(\lambda_j^t)\|^2 | \mathcal{F}_{t-1}]}{n^2} \\
&< 8L^2.
\end{aligned}$$

The last inequality follows from $N_i(t) < n$ and

$$\|\nabla f_j(\lambda_j^t + Z_j^t) - \nabla \hat{f}_j(\lambda_j^t)\| \leq \|\nabla f_j(\lambda_j^t + Z_j^t)\| + \|\nabla \hat{f}_j(\lambda_j^t)\| \leq 2L.$$

Finally,

$$\mathrm{E}[\|\delta^t\|^2 | \mathcal{F}_{t-1}] = \sum_i \mathrm{E}[\|\delta_i^t\|^2 | \mathcal{F}_{t-1}] < 8nL^2. \qquad \square$$

The following results follow immediately from Lemmas A.1–A.4 applied with $\hat{f}_i$ replacing f_i for all i:

$$\nabla \hat{f}(\lambda^t)^T v^t = \frac{1}{n^2} \sum_{(i,j) \in E(t)} \|\nabla \hat{f}_i(\lambda_i^t) - \nabla \hat{f}_j(\lambda_j^t)\|^2 \tag{18}$$

$$\|v^t\|^2 \leq 2n \nabla \hat{f}(\lambda^t)^T v^t \tag{19}$$

$$\|\tilde{v}^t\|^2 = n \nabla \hat{f}(\lambda^t)^T \tilde{v}^t \tag{20}$$

$$\begin{aligned}
&\sum_{(i,j) \in E} \|\nabla \hat{f}_j(\lambda_j^t) - \nabla \hat{f}_i(\lambda_i^t)\|^2 \geq \frac{8}{n^3} \sum_{((i,j) \in N^2, i<j)} \|\nabla \hat{f}_j(\lambda_j^t) - \nabla \hat{f}_i(\lambda_i^t)\|^2 \\
&\forall E : (N, E) \text{ is connected.}
\end{aligned} \tag{21}$$

Let E_{T_z} be a subset of the edge set $E_{T_z,T_{z+1}}$ such that the graph (N, E_{T_z}) is a tree. By Assumption 2.1, the graph $(N, E_{T_z,T_{z+1}})$ is connected and so E_{T_z} is well defined. As before, let the decentralized direction of update derived using $G = (N, E_{T_z})$ be denoted by $\bar{v}^{T_z}$. The following result is the counterpart of Lemma A.9.

LEMMA A.11. *Let* $L_\epsilon = \frac{\sqrt{\log(m+1)}L}{\epsilon}$.
$\nabla\hat{f}(\lambda^{T_z})^T\bar{v}^{T_z} \le \ \kappa\sum_{t=T_z}^{t=T_{z+1}-1}[(1+2L_\epsilon^2\gamma_t^2)\mathrm{E}[\nabla\hat{f}(\lambda^t)^T v^t|\mathcal{F}_{T_z-1}] + 8L^2L_\epsilon^2\gamma_t^2]$.

Proof. We have that

$$
\begin{aligned}
\mathrm{E}[\|\nabla\hat{f}(\lambda^{t+1}) - \nabla\hat{f}(\lambda^t)\|^2|\mathcal{F}_{t-1}] &\le \frac{L_\epsilon^2}{n^2}\mathrm{E}[\|\gamma_t(v^t+\delta^t)\|^2|\mathcal{F}_{t-1}] \\
&= \frac{L_\epsilon^2\gamma_t^2}{n^2}(\|v^t\|^2 + \mathrm{E}[\|\delta^t\|^2|\mathcal{F}_{t-1}]) \\
&\le \frac{L_\epsilon^2\gamma_t^2}{n^2}(2n\nabla\hat{f}(\lambda^t)^T v^t + 8nL^2). \qquad (22)
\end{aligned}
$$

It follows from Lemma 3.3 that L_ϵ is a Lipschitz constant for the functions $\hat{f}_i$, $i = 1,\ldots,n$. Hence $\frac{L_\epsilon}{n}$ is a Lipschitz constant for $\hat{f}$, and the first inequality follows from this. The second inequality follows from (19) and Lemma A.10.

Let $t^i_{T_z}$ be the earliest time between the time periods T_z and $T_{z+1}-1$ such that there is an edge $(i,j) \in E_{T_z}$ for agent i. It is clear that $T_z \le t^i_{T_z} \le T_{z+1}-1$. Also, by definition, for $l = T_z, T_z+1, \ldots, (t^i_{T_z}-1)$, there is no edge $(i,p) \in E(l)$. Thus $\lambda_i^{t^i_{T_z}} = \lambda_i^{T_z}$, and $\nabla\hat{f}_i(\lambda_i^{t^i_{T_z}}) = \nabla\hat{f}_i(\lambda_i^{T_z})$. Let $w_{ij}(t) = \frac{1}{n}(\nabla\hat{f}_i(\lambda_i^t) - \nabla\hat{f}_j(\lambda_j^t))$. Then

$$
\begin{aligned}
\|w_{ij}(T_z)\| &= \frac{1}{n}\|\nabla\hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla\hat{f}_j(\lambda_j^{T_z})\| \\
&\le \frac{1}{n}(\|\mathrm{E}[\nabla\hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla\hat{f}_j(\lambda_j^{t^i_{T_z}})|\mathcal{F}_{T_z-1}]\| \\
&\quad + \|\mathrm{E}[\nabla\hat{f}_j(\lambda_j^{t^i_{T_z}}) - \nabla\hat{f}_j(\lambda_j^{T_z})|\mathcal{F}_{T_z-1}]\|) \\
&\le \frac{1}{n}\left(\|\mathrm{E}[\nabla\hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla\hat{f}_j(\lambda_j^{t^i_{T_z}})|\mathcal{F}_{T_z-1}]\|\right. \\
&\quad \left. + \sum_{t=T_z}^{t=t^i_{T_z}-1}\|\mathrm{E}[\nabla\hat{f}_j(\lambda_j^{t+1}) - \nabla\hat{f}_j(\lambda_j^t)|\mathcal{F}_{T_z-1}]\|\right).
\end{aligned}
$$

From the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\|w_{ij}(T_z)\|^2 \ &\le \ \frac{(t^i_{T_z} - T_z + 1)}{n^2}\left(\|\mathrm{E}[\nabla\hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla\hat{f}_j(\lambda_j^{t^i_{T_z}})|\mathcal{F}_{T_z-1}]\|^2\right. \\
&\quad \left. + \sum_{t=T_z}^{t=t^i_{T_z}-1}\|\mathrm{E}[\nabla\hat{f}_j(\lambda_j^{t+1}) - \nabla\hat{f}_j(\lambda_j^t)|\mathcal{F}_{T_z-1}]\|^2\right) \\
&\le \ \frac{\kappa}{n^2}\left(\mathrm{E}[\|\nabla\hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla\hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2|\mathcal{F}_{T_z-1}]\right. \\
&\quad \left. + \sum_{t=T_z}^{t=t^i_{T_z}-1}\mathrm{E}[\|\nabla\hat{f}_j(\lambda_j^{t+1}) - \nabla\hat{f}_j(\lambda_j^t)\|^2)|\mathcal{F}_{T_z-1}]\right)
\end{aligned}
$$

$$\leq \kappa \left(\frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z - 1}]}{n^2} \right.$$
$$\left. + \frac{L_\epsilon^2}{n^2} \sum_{t=T_z}^{t=T_{z+1}-1} \gamma_t^2 (\mathrm{E}[2n \nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8nL^2) \right).$$

The last inequality follows from the fact that $\|\nabla \hat{f}(\lambda^{t+1}) - \nabla \hat{f}(\lambda^t)\|^2 = \frac{1}{n^2}\sum_{k=1}^n \|\nabla \hat{f}_k (\lambda_k^{t+1}) - \nabla \hat{f}_k(\lambda_k^t)\|^2 \geq \frac{1}{n^2}\|\nabla \hat{f}_j(\lambda_j^{t+1}) - \nabla \hat{f}_j(\lambda_j^t)\|^2$ and from (22). We finally have that

$$\begin{aligned} \nabla \hat{f}(\lambda^{T_z})^T \bar{v}^{T_z} &= \sum_{(i,j)\in E_{T_z}} \|w_{ij}(T_z)\|^2 \\ &\leq \sum_{(i,j)\in E_{T_z}} \kappa \left(\frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z - 1}]}{n^2} \right. \\ &\qquad \left. + \frac{L_\epsilon^2}{n^2} \sum_{t=T_z}^{t=T_{z+1}-1} \gamma_t^2 (\mathrm{E}[2n \nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8nL^2) \right) \\ &\leq \kappa \sum_{t=T_z}^{t=T_{z+1}-1} \left[(1 + 2L_\epsilon^2 \gamma_t^2) \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8L^2 L_\epsilon^2 \gamma_t^2 \right]. \end{aligned}$$

In the last inequality, we have used the fact that

$$\begin{aligned} &\sum_{(i,j)\in E_{T_z}} \frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z - 1}]}{n^2} \\ &= \sum_{t=T_z}^{t=T_{z+1}-1} \sum_{\left((i,j)\in E_{T_z}, t^i_{T_z} = t\right)} \frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z - 1}]}{n^2} \\ &\leq \sum_{t=T_z}^{t=T_{z+1}-1} \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}]. \qquad \square \end{aligned}$$

Proof of Theorem 3.2.

Proof of 1 *of Theorem* 3.2. We first have that

$$\begin{aligned} \mathrm{E}[\hat{f}(\lambda^{t+1})|\mathcal{F}_{t-1}] &\geq \hat{f}(\lambda^t) + \gamma_t \nabla \hat{f}(\lambda^t)^T v^t - \frac{L_\epsilon}{2n}\mathrm{E}[\|\gamma_t(v^t + \delta^t)\|^2 | \mathcal{F}_{t-1}] \\ &= \hat{f}(\lambda^t) + \gamma_t \nabla \hat{f}(\lambda^t)^T v^t - \frac{L_\epsilon}{2n}\|\gamma_t v^t\|^2 - \frac{L_\epsilon}{2n}\mathrm{E}[\|\gamma_t \delta^t\|^2 | \mathcal{F}_{t-1}] \\ &\geq \hat{f}(\lambda^t) + (\gamma_t - L_\epsilon \gamma_t^2) \nabla \hat{f}(\lambda^t)^T v^t - 4L_\epsilon L^2 \gamma_t^2. \end{aligned} \tag{23}$$

The first inequality comes from the descent lemma for differentiable functions [2]. The equality follows from $\mathrm{E}[\delta|\mathcal{F}_{t-1}] = 0$ from Lemma A.10. The second inequality follows from Lemma A.10 and (19).

Note that $\nabla \hat{f}(\lambda^t)^T v^t \geq 0$. This and Assumption 3.4 imply that the second term in (23) is also greater than or equal to zero. Moreover,

$$\sum_t 4L_\epsilon L^2 \gamma_t^2 < \infty.$$

Since $\hat{f}$ is bounded from above, we conclude by the supermartingale convergence theorem [4] that $\hat{f}(\lambda^t)$ converges with probability 1. Moreover, $\sum_t (\gamma_t - L_\epsilon \gamma_t^2) \nabla \hat{f}(\lambda^t)^T v^t < \infty$ with probability 1 and since $\sum_t \gamma_t = \infty$, we conclude that $\nabla \hat{f}(\lambda^t)^T v^t$ converges to zero with probability 1. Note that

$$\mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{t-1}] = \nabla \hat{f}(\lambda^t)^T v^t$$

with probability 1, and we conclude that $\mathrm{E}[\nabla \hat{f}(\lambda^t)^T v_t | \mathcal{F}_{t-1}]$ also converges to zero with probability 1.

We now have that

$$\begin{aligned}
\|\tilde{v}^{T_z}\|^2 &= n \nabla \hat{f}(\lambda^{T_z})^T \tilde{v}^{T_z} \\
&\le \frac{n^4}{8} \nabla \hat{f}(\lambda^{T_z})^T \bar{v}^{T_z} \\
&\le \frac{n^4}{8} \kappa \sum_{t=T_z}^{t=T_{z+1}-1} \left[(1 + 2L_\epsilon^2 \gamma_t^2) \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z - 1}] + 8L^2 L_\epsilon^2 \gamma_t^2 \right] \\
&\le \frac{n^4}{8} \kappa \sum_{t=T_z}^{t=T_{z+1}-1} \left[1.5 \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z - 1}] + 2L^2 \beta_t^2 \right]. \qquad (24)
\end{aligned}$$

The equality follows from (20). The first inequality follows from (21) and (18). The second inequality follows from Lemma A.11. The third inequality follows from Assumption 3.4 on the step sizes γ_t. We conclude that $\|\tilde{v}^{T_z}\|$ converges to zero with probability 1.

Proof of 2 *of Theorem* 3.2. From (23), we have that

$$\begin{aligned}
\nabla \hat{f}(\lambda^t)^T v^t &\le \frac{\mathrm{E}[\hat{f}(\lambda^{t+1}) | \mathcal{F}_{t-1}] + 4L_\epsilon L^2 \gamma_t^2 - \hat{f}(\lambda^t)}{\gamma_t (1 - L_\epsilon \gamma_t)} \\
&\le \frac{2(\mathrm{E}[\hat{f}(\lambda^{t+1}) | \mathcal{F}_{t-1}] - \hat{f}(\lambda^t)) + \left(\frac{2L^2 \beta_t^2}{L_\epsilon} \right)}{\gamma_t}. \qquad (25)
\end{aligned}$$

In the second inequality we have used $\gamma_t \le \frac{1}{2L_\epsilon}$ from Assumption 3.4.

Combining (24) and (25), we have that

$$\begin{aligned}
\mathrm{E}[\|\tilde{v}^{T_z}\|^2] &\le \frac{n^4}{8} \kappa \sum_{t=T_z}^{t=T_{z+1}-1} \left(\frac{3\mathrm{E}[\hat{f}(\lambda^{t+1}) - \hat{f}(\lambda^t)] + \left(\frac{3L^2 \beta_t^2}{L_\epsilon} \right)}{\gamma_t} + 2L^2 \beta_t^2 \right) \\
&\le \frac{n^4}{8\gamma_{\kappa z}} \kappa \sum_{t=T_z}^{t=T_{z+1}-1} \left[3\mathrm{E}[\hat{f}(\lambda^{t+1}) - \hat{f}(\lambda^t)] + \frac{3L^2 \beta_t^2}{L_\epsilon} + \frac{L^2 \beta_t^2}{L_\epsilon} \right].
\end{aligned}$$

The last inequality follows from Assumption 3.4 on the step sizes. It follows that

$$\begin{aligned}
\sum_{z=1}^{p} \gamma_{\kappa z} \mathrm{E}[\|\tilde{v}^{T_z}\|^2] &\le \frac{n^4}{8} \kappa \sum_{t=1}^{t=T_{p+1}-1} \left[3\mathrm{E}[\hat{f}(\lambda^{t+1}) - \hat{f}(\lambda^t)] + \frac{4L^2 \beta_t^2}{L_\epsilon} \right] \\
&\le \frac{n^4}{8} \kappa \left[3(\hat{f}(\hat{\lambda}) - \hat{f}(\lambda^1)) + \sum_{t=1}^{t=\kappa p} \frac{4L^2 \beta_t^2}{L_\epsilon} \right],
\end{aligned}$$

where $\hat{\lambda}$ denotes an optimal solution of (5). From Lemma 3.3, we have that $\hat{f}(\lambda^1) \geq f(\lambda^1) - 2.8\epsilon L$. We also have that $\hat{f}(\hat{\lambda}) \leq f(\hat{\lambda}) \leq f(\lambda^*)$. It follows that

$$\min_{z=1,\ldots,p} \mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \frac{\frac{n^4 \kappa L \sqrt{\log(m+1)}}{\epsilon}\left[3(f(\lambda^*) - f(\lambda^1) + 2.8\epsilon L) + \sum_{t=1}^{t=\kappa p} \frac{4L\beta_t^2 \epsilon}{\sqrt{\log(m+1)}}\right]}{4\sum_{z=2}^{p+1} \beta_{\kappa z}} \; \forall p.$$

Proof of 3 *of Theorem* 3.2. Since $f \geq \hat{f} \geq f - 2.8\epsilon L$, if (1) has a bounded set of optima, so does (5). Recall from the proof of the first claim that $\hat{f}(\lambda^t)$ converges with probability 1. Using the same argument as in the proof of the fourth claim of Theorem 3.1, we conclude that $\hat{f}(\lambda^t)$ converges to $\hat{f}(\hat{\lambda})$ with probability 1. We conclude that

$$\begin{aligned}\limsup_{t\to\infty} f(\lambda^t) &\geq \hat{f}(\hat{\lambda}) \\ &\geq \hat{f}(\lambda^*) \\ &\geq f(\lambda^*) - 2.8\epsilon L.\end{aligned}$$

The first inequality follows from $f(\lambda^t) \geq \hat{f}(\lambda^t)$ for all t from Lemma 3.3. The second inequality follows from the optimality of $\hat{\lambda}$. The third inequality follows from Lemma 3.3. □

LEMMA 4.1. *Under assumption* 3.3 *for* f,

1. *for all* i, $g_i(\lambda_i)$ *is concave and differentiable outside a set of measure zero;*
2. *for all* i *and* λ_i, $\sup_{i,\lambda_i}\{\|v\|_1 : v \in \partial g_i(\lambda_i)\} \leq L_m < \infty$, *where* $L_m = L + mL_g$.

Proof of 1 *of Lemma* 4.1. Let $h_j(\lambda_i) = L_g \min(\lambda_{ij}, 0)$. It is clear that h_j is a piecewise linear function. Recall that

$$g_i(\lambda_i) = f_i(\lambda_i) + \sum_{j=1}^{m} h_j(\lambda_i).$$

The concavity of g_i follows from the concavity of f and the functions h_j, $j = 1, \ldots, m$. The points of the nondifferentiability of f_i form a set of measure zero. The other points of nondifferentiability of g_i are points λ_i, where $\lambda_{ij} = 0$ for some j. These points form a set of measure zero. Thus g_i is differentiable outside a set of measure zero.

Proof of 2 *of Lemma* 4.1. Let $\mathbf{e}_j$ be the vector whose jth component is 1 and other components are 0. It is clear that, for λ_i with $\lambda_{ij} \neq 0$, h_j is differentiable and $\nabla h_j(\lambda_i) = L_g \mathbf{e}_j$ if $\lambda_{ij} < 0$ and $\nabla h_j(\lambda_i) = \mathbf{0}$ if $\lambda_{ij} > 0$, where $\mathbf{0}$ is the m-dimensional zero vector. For λ_i with $\lambda_{ij} = 0$, $\partial h_j(\lambda_i)$ consists of vectors of the form $\bar{L}\mathbf{e}_j$, where $0 \leq \bar{L} \leq L_g$. Thus, for all j, $\sup_{\lambda_i}\{\|v\|_1 : v \in \partial h_j(\lambda_i)\} = L_g$. It is known from the theory of convex functions that if $u = \sum_{j=1}^{k} u_j$ where u_j, $j = 1, \ldots, k$ are convex functions, then $\partial u(x) = \sum_{j=1}^{k} \partial u_j(x)$. Thus, if $\sup_x\{\|v\|_1 : v \in \partial u_j(x)\} \leq L_j$, then $\sup_x\{\|v\|_1 : v \in \partial u(x)\} \leq \sum_{j=1}^{k} L_j$. By assumption, $\sup_{\lambda_i}\{\|v\|_1 : v \in \partial f_i(\lambda_i)\} \leq L$. Hence $\sup_{\lambda_i}\{\|v\|_1 : v \in \partial g_i(\lambda_i)\} \leq L + \sum_{i=1}^{m} L_g = L + mL_g$.

LEMMA 4.2. *The set of optimal solutions for* (1) *with* g *as the objective function is the same as the set of optimal solutions to* (7) *with* f *as the objective function.*

Proof. Without loss of generality assume that $B > 0$. Consider some optimal solution λ^* for (7) with f as the objective function. Suppose there exists some feasible solution $\hat{\lambda}$ to (1), with $\hat{\lambda}_{ij} < 0$ for some i, j. We show that $g(\hat{\lambda}) < g(\lambda^*)$. This implies that solving (7) with g as the objective function is equivalent to solving (1) with g as the objective function. Since $g(\lambda) = f(\lambda)$ when $\lambda \geq 0$, solving (7) with g is equivalent to solving (7) with f. Thus the set of optimal solutions for (1) with g and for (7) with f are the same proving the lemma.

Consider the following problem,

$$\begin{aligned} &\max_{\lambda_p \in \Re^m, p=1,\ldots,n} g(\lambda) = \frac{1}{n}\sum_{p=1}^{n} g_p(\lambda_p) \\ &\text{s.t.} \qquad \sum_{p=1}^{n} \lambda_p = B, \\ &\qquad\qquad \lambda_p \geq -|\hat{\lambda}_p|, p = 1, \ldots, n. \end{aligned} \tag{26}$$

It can be seen that λ^* and $\hat{\lambda}$ are feasible solutions to (26). We now show that $\hat{\lambda}$ cannot be an optimal solution to (26). Since $B > 0$, there exists some k such that $\hat{\lambda}_{kj} > 0$. Define $\bar{\lambda}$ so that it differs from $\hat{\lambda}$ only in the ij and kj components as follows:

$$\bar{\lambda}_{ij} = \hat{\lambda}_{ij} + \delta,$$

$$\bar{\lambda}_{kj} = \hat{\lambda}_{kj} - \delta.$$

We choose a $\delta > 0$ such that $\bar{\lambda}_{kj} > 0$ and $\bar{\lambda}_{ij} < 0$. It is clear that $\bar{\lambda}$ is a feasible solution to (26). We now have

$$\begin{aligned} g_i(\bar{\lambda}_i) &= f_i(\bar{\lambda}_i) + \sum_{l=1}^{m} h_l(\bar{\lambda}_i) \\ &\geq g_i(\hat{\lambda}_i) + \delta\left(L_g + \left(\nabla f_i(\bar{\lambda}_i)\right)_j\right). \end{aligned}$$

The inequality comes from the concavity of g_i and from the definition of $\bar{\lambda}$. Similarly

$$\begin{aligned} g_k(\bar{\lambda}_k) &= f_k(\bar{\lambda}_k) + \sum_{l=1}^{m} h_l(\bar{\lambda}_k) \\ &\geq g_k(\hat{\lambda}_k) - \delta\left(\nabla f_k(\bar{\lambda}_k)\right)_j. \end{aligned}$$

Hence

$$g_i(\bar{\lambda}_i) + g_k(\bar{\lambda}_k) \geq g_i(\hat{\lambda}_i) + g_k(\hat{\lambda}_k) + \delta\left(L_g + \left(\nabla f_i(\bar{\lambda}_i)\right)_j - \left(\nabla f_k(\bar{\lambda}_k)\right)_j\right).$$

Since $L_g > 2L$, we can conclude from the above that $g(\bar{\lambda}) > g(\hat{\lambda})$, and hence $\hat{\lambda}$ cannot be an optimal solution to (26).

It can be seen from the definition of (26) that its feasible set is bounded. Since g is continuous and since the feasible set of (26) is bounded and closed, it has at least one optimal solution by the extreme value theorem. The above argument, presented to establish the nonoptimality of $\hat{\lambda}$ for (26), holds for any feasible solution for (26) with at least one nonnegative component. Hence all optimal solutions of (26) are nonnegative. Since $g(\lambda) = f(\lambda)$ when $\lambda \geq 0$, solving (26) with g is equivalent to solving (7) with f, and hence λ^* is an optimal solution for (26). Hence $g(\hat{\lambda}) < g(\bar{\lambda}) \leq g(\lambda^*)$. □

Acknowledgments. The second author thanks Dimitri Bertsekas, Rob Freund, and John Tsitsiklis for their helpful comments and suggestions.

REFERENCES

[1] K. Arrow and F. Hahn, *General Competitive Analysis*, Holden Day, San Francisco, 1971.

[2] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[3] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Belmont, MA, 1997.

[4] R. Durrett, *Probability: Theory and Examples*, Duxbury Press, Belmont, CA, 1995.

[5] G. M. Heal, *Planning without prices*, Rev. Econom. Stud., 63 (1969), pp. 343–362.

[6] J. Kurose and R. Simha, *A microeconomic approach to optimal resource allocation computer systems*, IEEE Trans. Comput., 38 (1989).

[7] H. Lakshmanan and D. P. De Farias, *Decentralized approximate dynamic programming for dynamic networks of agents*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006.

[8] L. Xiao and S. Boyd, *Optimal scaling of a gradient method for distributed resource allocation*, J. Optim. Theory Appl., 129 (2006), pp. 469–488.

[9] A. Nedic and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.

[10] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[11] L. Servi, Y. C. Ho, and R. Suri, *A class of center-free resource allocation algorithms*, Large Scale Systems, 1 (1980), pp. 51–62.

[12] T.-S. Tang and M. A. Styblinsky, *Yield optimization for nondifferentiable density functions using convolution techniques*, IEEE Trans. Comput. Aided Design, 7 (1988), pp. 1053–1067.

[13] T. Ibaraki and N. Katoh, *Resource Allocation Problems: Algorithmic Approaches*, MIT Press, Cambridge, MA, 1988.

[14] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, MIT, Cambridge, MA, 1984.