UCLA CAM Report 13-61

On the Convergence of Decentralized Gradient Descent

Kun Yuan* Qing Ling* Wotao Yin†

Abstract

Consider the consensus problem of minimizing $f(x) = \sum_{i=1}^n f_i(x)$ where each f_i is only known to one individual agent i belonging to a connected network of n agents. All the agents shall collaboratively solve this problem and obtain the solution via data exchanges only between neighboring agents. Such algorithms avoid the need of a fusion center, offer better network load balance, and improve data privacy. We study the decentralized gradient descent method in which each agent i updates its variable $x_{(i)}$, which is a local approximate to the unknown variable x, by taking the average of its neighbors' followed by making a local negative gradient step $-\alpha\nabla f_i(x_{(i)})$. The iteration is

$$x_{(i)}(k+1) \leftarrow \sum_j w_{ij} x_{(j)}(k) - \alpha\nabla f_i(x_{(i)}(k)), \quad \text{for each agent } i,$$

where the coefficients w_{ij} form a symmetric doubly stochastic matrix $W = [w_{ij}] \in \mathbb{R}^{n\times n}$. As agent i does not communicate to non-neighbors, $w_{ij} \neq 0$ only if $i = j$ or j is a neighbor of i. We analyze the convergence of this iteration and derive its rate, assuming that each f_i is proper closed convex and lower bounded, ∇f_i is Lipschitz continuous with constant L_{f_i}, and stepsize α is fixed. Provided that $\alpha < O(1/L_h)$ where $L_h = \max_i\{L_{f_i}\}$, the objective error at the averaged solution, $f(\frac{1}{n}\sum_i x_{(i)}(k)) - f^*$ where f^* is the optimal objective value, reduces at a speed of $O(1/k)$ until it reaches $O(\alpha)$. If f_i are (restricted) strongly convex, then both $\frac{1}{n}\sum_i x_{(i)}(k)$ and each $x_{(i)}(k)$ converge to the global minimizer x^* at a linear rate until reaching an $O(\alpha)$-neighborhood of x^*. We also develop an iteration for decentralized basis pursuit and establish its linear convergence to an $O(\alpha)$-neighborhood of the true sparse signal. This analysis reveals how convergence depends on the stepsize, function convexity, and network spectrum.

1 Introduction

Consider that n agents form a connected network and they collaboratively solve a consensus optimization problem

$$\underset{x}{\text{minimize}} \quad f(x) = \sum_{i=1}^n f_i(x), \tag{1}$$

*K. Yuan and Q. Ling are with Department of Automation, University of Science and Technology of China, Hefei, Anhui 230026, China. kunyuan@mail.ustc.edu.cn and qingling@mail.ustc.edu.cn

†W. Yin is with Department of Mathematics, University of California, Los Angeles, CA 90095, USA. wotaoyin@math.ucla.edu

where $x \in \mathbb{R}^p$ is the common optimization variable and each f_i is only available to agent i. Some pairs of agents with direct communication links can exchange data. Let $\mathcal{X}^*$ denotes the set of solutions to (1), which is assumed to be non-empty, and let f^* denote the optimal objective value.

The traditional (centralized) gradient descent iteration is

$$x(k+1) = x(k) - \alpha \nabla f(x(k)), \tag{2}$$

where α is the stepsize, either fixed or varying with k. To apply iteration (2) to problem (1) under the decentralized situation, one has different choices:

- let a fusion center (which can be one of the agents) carry out iteration (2);
- let all agents carry out the same iteration (2).

In either way, since f_i (and thus ∇f_i) is only known to agent i, in order to obtain $\nabla f(x(k)) = \sum_{i=1}^n \nabla f_i(x(k))$, every agent i must have $x(k)$, compute $\nabla f_i(x(k))$, and then send $\nabla f_i(x(k))$ out. This approach requires synchronizing $x(k)$ and scattering/collecting $\nabla f_i(x(k))$, $i = 1, \ldots, n$, over the entire network. This incurs significant communication traffic, especially if the network is large or sparse, or both. A viable alternative is a decentralized approach, whose communication is confined to between neighbors. Although there is no guarantee that decentralized algorithms use less communication (as they tend to take more iterations), they have advantages in terms of network load balance and better tolerance to the failure of individual agents. In addition, each agent can keep its implementation of f_i private, so to some extent, its data is protected[1].

Decentralized gradient descent [19] does not rely on a fusion center or network-wide communication. It carries out an approximate version of (2) following the strategies below:

- let each agent i hold an approximate *copy* $x_{(i)} \in \mathbb{R}^p$ of $x \in \mathbb{R}^p$ ($x_{(i)} \neq x_{(j)}$ is allowed if $i \neq j$);
- let each agent i update its $x_{(i)}$ to its neighborhood (weighted) average;
- let each agent i compute $-\nabla f_i(x_{(i)})$ and apply it locally to decrease $f_i(x_{(i)})$.

At each iteration k, each agent i performs the following steps

1. computes $\nabla f_i(x_{(i)}(k))$;
2. computes the neighborhood average $x_{(i)}(k+1/2) = \sum_j w_{ij} x_{(j)}(k)$, where $w_{ij} \neq 0$ if and only if j is a neighbor of i or $i = j$;
3. applies $x_{(i)}(k+1) = x_{(i)}(k+1/2) - \alpha \nabla f_i(x_{(i)}(k))$.

Steps 1 and 2 can be carried out in parallel, and their results are used in Step 3. Putting the three steps together, we arrive at our main iteration

$$\boxed{x_{(i)}(k+1) = \sum_j w_{ij} x_{(j)}(k) - \alpha \nabla f_i(x_{(i)}(k)), \quad i = 1, 2, \ldots, n.} \tag{3}$$

[1]Neighbors of i may know some samples of f_i and/or ∇f_i through data exchanges and thus obtain an interpolation of f_i.

If f_i is not differentiable and ∇f_i is replaced by a member of the subdifferential ∂f_i, the resulting iteration is known as the decentralized *subgradient* iteration [19]. Other decentralization methods are reviewed below.

The coefficients w_{ij} form a symmetric, doubly stochastic matrix $W = [w_{ij}]$, which we call the blending matrix. In a multi-agent network, we can restrict communication to between agents with direct links. If agents i and j have a direct link or $i = j$, $w_{ij} > 0$; otherwise, $w_{ij} = 0$. The eigenvalues of W are real and can be sorted in a nonincreasing order $1 = \lambda_1(W) \geq \lambda_2(W) \geq \cdots \geq \lambda_n(W) \geq -1$. Let the second largest magnitude of eigenvalues of W be denoted as

$$\beta = \max\{|\lambda_2(W)|, |\lambda_n(W)|\}. \tag{4}$$

For the design of matrix W and to minimize β in particular, the reader is referred to [3].

Basic questions regarding the decentralized (sub)gradient iteration include: (i) When does $x_{(i)}(k)$ converge? (ii) Does it converge to $x^* \in \mathcal{X}^*$? (iii) When x^* is not the limit, does consensus (i.e., $x_{(i)}(k) = x_{(j)}(k)$, $\forall i \neq j$) hold in the limit? (iv) How do the properties of f_i and the underlying network affect the convergence?

1.1 Background

The study on decentralized optimization can be traced back to the seminal work in the 1980s [28, 29]. Compared to centralized optimization in which a fusion center collects data and takes over computation, decentralized optimization enjoys the advantages of scalability to network sizes, robustness to dynamic topologies, and privacy preservation in data-sensitive applications [6, 16, 21, 30]. These properties are suited for applications where data are collected by and stored in distributed agents, communication to a fusion center is expensive or impossible, and/or agents tend to keep their raw data private; such applications arise in wireless sensor networks [15, 22, 25, 34], multivehicle and multirobot networks [4, 24, 35], smart grids [9, 12], cognitive radio networks [1, 2], etc. The recent research interest in big data processing also motivates the introduction of decentralized optimization to machine learning [7, 26]. Further, the decentralized static optimization problem (1) can be extended to its online or dynamic counterparts if the objective function is an online regret [27, 30] or a dynamic cost [5, 11, 14].

We take spectrum sensing in a cognitive radio network as an example to demonstrate the application of decentralized optimization. Spectrum sensing aims at detecting unused spectrum bands, known as spectrum holes, such that the cognitive radios can opportunistically use those bands. Let x be a vector in which each element corresponds to the magnitude of the corresponding channel. Cognitive radio i takes time-domain measurement of x with $b_i = F^{-1}G_i x + e_i$, where G_i is cognitive radio i's channel fading matrix, F^{-1} is the inverse Fourier transform matrix, and e_i is the measurement noise. To estimate x, a set of geologically nearby cognitive radios collaboratively solve the consensus optimization problem (1). The local objective function of cognitive radio i can be a least squares $f_i(x) = (1/2)\|b_i - F^{-1}G_i x\|^2$ or a regularized least squares $f_i(x) = (1/2)\|b_i - F^{-1}G_i x\|^2 + \phi(x)$, where the regularization term $\phi(x)$ comes from the prior knowledge of x. Decentralized optimization is a good fit since it takes advantage of the fast and energy-

efficient communication between neighboring cognitive radios and, when cognitive radios join and leave the network, no reconfiguration is needed.

1.2 Related methods

Besides the distributed subgradient method [19], the distributed stochastic subgradient projection algorithm [23] is able to handle constraints; the fast distributed gradient methods [10] adopts Nesterov's acceleration; the distributed online gradient descent algorithm[2] [27] has inner loops for fine search; and, the dual averaging subgradient method [7] carries a projection operation after averaging and descending. Unsurprisingly, going from the traditional centralized computation to the decentralized one incurs more assumptions, weaker convergence rates, and slower convergence. All of the above algorithms work under the assumption of bounded (sub)gradients (and [7] further requires f_i to be Lipschitz continuous). Unbounded gradients can potentially diverge the algorithms. When using a fixed stepsize, the above algorithms (and iteration (3) in particular) do not converge to x^* but its neighborhood, whose size is monotonic in the stepsize. This motivates the use of certain diminishing stepsizes in [7, 10, 27] to guarantee convergence to x^*. The rates of convergence are generally weaker than their counterparts in centralized computation. With diminishing stepsizes, [10] shows an outer loop complexity of $O(1/k^2)$ with Nesterov acceleration where the inner loop performs a substantial search job, without which the rate reduces to $O(\log(k)/k)$.

1.3 Contribution and notation

This paper studies the convergence of iteration (3) under the following assumption.

Assumption 1. *a) For $i = 1, \dots, n$, f_i is proper closed convex, differentiable, and has a minimizer, and ∇f_i is Lipschitz continuous with finite constant L_{f_i}.*

b) The network has a synchronized clock in the sense that (3) is applied to all agents at the same time intervals, the network is connected, and the blending matrix W is symmetric doubly stochastic with $\beta < 1$ (see (4) for the definition.)

We do not assume bounded ∇f_i but provide a stepsize condition that gives bounded ∇f_i:

$$\alpha < O(1/L_h), \tag{5}$$

where $L_h = \max\{L_{f_1}, \dots, L_{f_n}\}$.

Assumption 1 and condition (5) suffice for "near" convergence at a rate $O(1/k)$. Specifically, the objective error evaluated at the mean solution, $f(\frac{1}{n}\sum_{i=1}^n x_{(i)}(k)) - f^*$, reduces at $O(1/k)$ until reaching $O(\frac{\alpha}{1-\beta})$.

For "near" linear convergence, we further assume that f is strongly convex with modulus $\mu_f > 0$, namely,

$$\langle \nabla f(x_a) - \nabla f(x_b), x_a - x_b \rangle \geq \mu_f \|x_a - x_b\|^2, \quad \forall x_a, x_b \in \mathrm{dom} f,$$

[2]Here we consider its decentralized batch version.

or is restricted strongly convex [13] with modulus $\nu_f > 0$, namely,

$$\langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle \geq \nu_f \|x - x^*\|^2, \quad \forall x \in \mathrm{dom} f,\ x^* = \mathrm{Proj}_{\mathcal{X}^*}(x), \tag{6}$$

where $\mathrm{Proj}_{\mathcal{X}^*}(x)$ is the projection of x onto the solution set $\mathcal{X}^*$. Note that $\nabla f(x^*) = 0$, and such functions find applications in sparse optimization and statistical regression; see [33] for some examples. In both cases, we show that the mean-solution error $\|\frac{1}{n}\sum_i x_{(i)}(k) - x^*\|$ and the individual-solution error $\|x_{(i)}(k) - x^*\|$ both reduce geometrically until reaching $O(\frac{\alpha}{1-\beta})$. Note that $\mathcal{X}^*$ is a singleton if f is strongly convex but not necessarily so if f is restricted strongly convex.

When a fixed α is used, $x_{(1)}(k), \ldots, x_{(n)}(k)$ do not generally equal one another either at each k or as $k \to \infty$. One can of course call an additional average consensus algorithm after iteration (3) stops.

Some of our results can be extended to the case of $\alpha \to 0$ whereas the main convergence analysis will be significantly different. Therefore, we leave $\alpha \to 0$ to future work.

As an application of these results, we derive in Section 3 a novel algorithm for the basis pursuit problem with decentralized data. It converges linearly until reaching an $O(\frac{\alpha}{1-\beta})$-neighborhood to the sparse solution.

In Section 4 we present numerical results on decentralized least-squares and decentralized basis pursuit problems to verify our convergence analysis.

Throughout the rest of this paper, we let

$$[x_{(i)}] := \begin{bmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(n)} \end{bmatrix} \in \mathbb{R}^{np} \quad \text{and} \quad h(k) := \begin{bmatrix} \nabla f_1(x_{(1)}(k)) \\ \nabla f_2(x_{(2)}(k)) \\ \vdots \\ \nabla f_n(x_{(n)}(k)) \end{bmatrix} \in \mathbb{R}^{np}.$$

2 Convergence analysis

2.1 Bounded gradients

Previous methods and analysis [7, 10, 19, 23, 27] assume bound gradients or subgradients of f_i. The assumption indeed plays a key role in the convergence analysis. For decentralized gradient descent iteration (3), it gives *bounded* deviation from mean $\|x_{(i)}(k) - \frac{1}{n}\sum_j x_{(j)}(k)\|$. It is necessary in the convergence analysis of subgradient methods, whether they are centralized or decentralized. But as we show below, the boundedness of ∇f_i is not guaranteed but is a consequence of bounded stepsize α, with dependence on the spectral properties of W. We derive a tight bound on α for $\nabla f_i(x_{(i)}(k))$ to be bounded.

Example. Consider $x \in \mathbb{R}$ and a network formed by 3 connected agents (every pair of agents are directly linked). Consider the following consensus optimization problem

$$\underset{x}{\text{minimize}}\, f(x) = \sum_{i=1,2,3} f_i(x), \quad \text{where } f_i(x) = \frac{L_h}{2}(x-1)^2,$$

where L_h is a positive constant. This is a trivial average consensus problem with $\nabla f_i(x_{(i)}) = L_h(x_{(i)} - 1)$ and $x^* = 1$. Take any $\tau \in (0, 1/3)$ and let the blending matrix be

$$W = \begin{bmatrix} 1-2\tau & \tau & \tau \\ \tau & \tau & 1-2\tau \\ \tau & 1-2\tau & \tau \end{bmatrix},$$

which is symmetric doubly stochastic. We have $\lambda_3(W) = 3\tau - 1 \in (-1, 0)$. Start from $(x_1, x_2, x_3) = (1, 0, 2)$. Simple calculations yield

- if $\alpha < (1 + \lambda_3(W))/L_h$, then $x_{(i)}(k)$ converges to x^*, $i = 1, 2, 3$. (The consensus among $x_{(i)}(k)$ as $k \to \infty$ is due to design.)
- if $\alpha > (1 + \lambda_3(W))/L_h$, then $x_{(i)}(k)$ diverges, $i = 1, 2, 3$.
- if $\alpha = (1 + \lambda_3(W))/L_h$, then $(x_1(k), x_2(k), x_3(k))$ equals $(1, 2, 0)$ for odd k and $(1, 0, 2)$ for even k.

Clearly, if $x_{(i)}$ converges, then $\nabla f_i(x_{(i)})$ converges and thus stays bounded. In the above example $\alpha = (1 + \lambda_3(W))/L_h$ is the critical stepsize.

As each $\nabla f_i(x_{(i)})$ is Lipschitz continuous with constant L_{f_i}, $h(k)$ is Lipschitz continuous with constant

$$L_h = \max_i \{L_{f_i}\}.$$

We formally show that $\alpha < (1 + \lambda_n(W))/L_h$ ensures bounded $h(k)$.

Theorem 1. *Under Assumption 1, if the stepsize*

$$\alpha \leq (1 + \lambda_n(W))/L_h, \tag{7}$$

starting from $x_{(i)}(0) = 0$, $i = 1, 2, \ldots, n$, we have

$$\|h(k)\| \leq D := \sqrt{2L_h \sum_i (f_i(0) - f_i^*)} \tag{8}$$

for all $k = 1, 2, \ldots$.

Proof. Our proof is based on the auxiliary function

$$\xi_\alpha([x_{(i)}]) := -\frac{1}{2} \sum_{ij} w_{ij} x_{(i)}^T x_{(j)} + \sum_i \left(\frac{1}{2} \|x_{(i)}\|^2 + \alpha f_i(x_{(i)}) \right). \tag{9}$$

Function ξ_α is convex since all f_i are convex and the sum of the rest terms $\frac{1}{2} \left(\sum_i \|x_{(i)}\|^2 - \sum_{ij} w_{ij} x_{(i)}^T x_{(j)} \right)$ is also convex (and uniformly nonnegative) due to $\lambda_1(W) = 1$. In addition, $\nabla \xi_\alpha$ is Lipschitz continuous with constant $L_{\xi_\alpha} \leq (1 - \lambda_n(W)) + \alpha L_h$. An important observation is that iteration (3) can be written as

$$x_{(i)}(k+1) = \sum_j w_{ij} x_j(k) - \alpha \nabla f_i(x_{(i)}(k)) = x_{(i)}(k) - \nabla_i \xi_\alpha([x_{(i)}(k)]).$$

Since $\beta < 1$, we have $\lambda_n(W) > -1$ and $(L_{\xi_\alpha}/2 - 1) \le 0$. Hence,

$$\begin{aligned}\xi_\alpha([x_{(i)}(k+1)]) &\le \xi_\alpha([x_{(i)}(k)]) + \nabla\xi_\alpha([x_{(i)}(k)])^T([x_{(i)}(k+1) - x_{(i)}(k)]) + \frac{L_{\xi_\alpha}}{2}\|[x_{(i)}(k+1) - x_{(i)}(k)]\|^2 \\ &= \xi_\alpha([x_{(i)}(k)]) + (L_{\xi_\alpha}/2 - 1)\|\nabla_i\xi_\alpha([x_{(i)}(k)])\|^2 \\ &\le \xi_\alpha([x_{(i)}(k)]),\end{aligned}$$

which explains how iteration (3) reduces the underlying potential function $\xi_\alpha([x_{(i)}])$.

Recall that $\frac{1}{2}\left(\sum_i \|x_{(i)}\|^2 - \sum_{ij} w_{ij} x_{(i)}^T x_{(j)}\right)$ is nonnegative, so we have the uniform bound

$$\sum_i f_i(x_{(i)}(k)) \le \alpha^{-1}\xi_\alpha([x_{(i)}(k)]) \le \cdots \le \alpha^{-1}\xi_\alpha([x_{(i)}(0)]) = \alpha^{-1}\xi_\alpha(0) = \sum_i f_i(0). \tag{10}$$

On the other hand, for any differentiable convex function g with minimizer x^* and Lipschitz constant L_g, we have $g(x_a) \ge g(x_b) + \nabla g^T(x_b)(x_a - x_b) + \frac{1}{2L_g}\|\nabla g(x_a) - \nabla g(x_b)\|^2$ and $\nabla g(x^*) = 0$. Then, $\|\nabla g(x)\|^2 \le 2L_g(g(x) - g^*)$ where $g^* := g(x^*)$. Applying this inequality and (10), we obtain

$$\|h(k)\|^2 = \sum_i \|\nabla f_i(x_{(i)}(k))\|^2 \le \sum_i 2L_{f_i} \cdot (f_i(x_{(i)}(k)) - f_i^*) \le 2L_h\left(\sum_i f_i(0) - \sum_i f_i^*\right),$$

where $f_i^* := f_i(x^*)$. This completes the proof. □

In the above theorem, we choose $x_{(i)}(0) = 0$ for convenience. Otherwise, a different bound for $\|h(k)\|$ can still be obtained.

Dependence on stepsize. In (3), the negative gradient step $-\alpha\nabla f_i(x_{(i)})$ does not diminish at $x_{(i)} = x^*$. Even if we let $x_{(i)} = x^*$ for all i, $x_{(i)}$ will immediately change once (3) is applied. Therefore, the term $-\alpha\nabla f_i(x_{(i)})$ prevents the consensus of $x_{(i)}$. Even worse, because both terms in the right-hand side of (3) change $x_{(i)}$, they can possibly add up to an uncontrollable amount and cause $x_{(i)}(k)$ to diverge. The local averaging term is itself stable, so the only choice is to limit the size of $-\alpha\nabla f_i(x_{(i)})$ by bounding α.

Network spectrum. One can design W so that $\lambda_n(W) > 0$ and thus simply bound (7) to

$$\alpha \le 1/L_h,$$

which no longer requires the knowledge of the spectral property of the underlying network. Given any blending matrix $\tilde{W}$ satisfying $1 = \lambda_1(\tilde{W}) > \lambda_2(\tilde{W}) \ge \cdots \ge \lambda_n(\tilde{W}) > -1$ (cf. [3]), the new blending matrix $W = (\tilde{W} + I)/2$ satisfies $1 = \lambda_1(W) > \lambda_2(W) \ge \cdots \ge \lambda_n(W) > 0$. The same argument applies to the results throughout the paper.

2.2 Bounded deviation from mean

Let

$$\bar{x}(k) := \frac{1}{n}\sum_{i=1}^n x_{(i)}(k)$$

be the *mean* of $x_1(k), \ldots, x_n(k)$. We will later analyze the error in terms of $\bar{x}(k)$ and then $x_{(i)}(k)$. To enable that analysis, we shall show that the deviation from mean $\|x_{(i)}(k) - \bar{x}(k)\|$ is bounded uniformly over

i and k. Hence, any bound of $\|\bar{x}(k) - x^*\|$ will give a bound of $\|x_{(i)}(k) - x^*\|$. Intuitively, if the deviation from mean is unbounded, then there is no approximate consensus among $x_1(k), \ldots, x_n(k)$. Without this approximate consensus, descending individual $f_i(x_{(i)}(k))$ does not contribute to the descent of $f(\bar{x}(k))$ and thus convergence is out of the question. Hence, it is a key step to bound the deviation $\|x_{(i)}(k) - \bar{x}(k)\|$.

Theorem 2. *If $\|h(k)\| \leq D$ for all k and $\beta < 1$, then the total deviation from mean is bounded, namely,*

$$\|x_{(i)}(k) - \bar{x}(k)\| \leq \frac{\alpha D}{1-\beta}, \quad \forall k, \forall i.$$

Proof. Recall the definition of $[x_{(i)}]$ and $h(k)$, from equation (3) we have

$$[x_{(i)}(k+1)] = (W \otimes I)[x_{(i)}(k)] - \alpha h(k),$$

where $\otimes$ means the Kronecker product. From it, we obtain

$$[x_{(i)}(k)] = -\alpha \sum_{s=0}^{k-1} (W^{k-1-s} \otimes I) h(s). \tag{11}$$

Besides, letting $[\bar{x}(k)] = [\bar{x}(k); \cdots; \bar{x}(k)] \in \mathbb{R}^{np}$, it follows that

$$[\bar{x}(k)] = \frac{1}{n}((1_n 1_n^T) \otimes I))[x_{(i)}(k)].$$

As a result,

$$\begin{aligned}
\|x_{(i)}(k) - \bar{x}(k)\| &\leq \|[x_{(i)}(k)] - [\bar{x}(k)]\| \\
&= \|[x_{(i)}(k)] - \frac{1}{n}((1_n 1_n^T) \otimes I))[x_{(i)}(k)]\| \\
&= \| - \alpha \sum_{s=0}^{k-1} (W^{k-1-s} \otimes I) h(s) + \alpha \sum_{s=0}^{k-1} \frac{1}{n}((1_n 1_n^T W^{k-1-s}) \otimes I) h(s)\| \\
&= \| - \alpha \sum_{s=0}^{k-1} (W^{k-1-s} \otimes I) h(s) + \alpha \sum_{s=0}^{k-1} \frac{1}{n}((1_n 1_n^T) \otimes I) h(s)\| \qquad (12) \\
&= \alpha \| \sum_{s=0}^{k-1} ((W^{k-1-s} - \frac{1}{n} 1_n 1_n^T) \otimes I) h(s)\| \\
&\leq \alpha \sum_{s=0}^{k-1} \|W^{k-1-s} - \frac{1}{n} 1_n 1_n^T\| \|h(s)\| \\
&= \alpha \sum_{s=0}^{k-1} \beta^{k-1-s} \|h(s)\|,
\end{aligned}$$

where (12) holds since W is doubly stochastic. Since $\|h(k)\| \leq D$ and $\beta < 1$, finally we have

$$\|x_{(i)}(k) - \bar{x}(k)\| \leq \alpha \sum_{s=0}^{k-1} \beta^{k-1-s} \|h(s)\| \leq \alpha \sum_{s=0}^{k-1} \beta^{k-1-s} D \leq \frac{\alpha D}{1-\beta},$$

which completes the proof. □

The proof of Theorem 2 utilizes the spectral property of the blending matrix W. The upper bound of the deviation from mean is proportional to the stepsize α and monotonically increasing with respect to the second largest eigenvalue modulus β. Similar results can be found in the analysis of decentralized first-order algorithms, e.g., the distributed stochastic subgradient projection algorithm (Lemma 4.1 in [23]) and the dual averaging subgradient method (Theorem 2 in [7]).

A consequence of Theorem 2 is that the distance between the following two quantities is also bounded

$$g(k) := \frac{1}{n}\sum_i \nabla f_i(x_{(i)}(k)),$$
$$\bar{g}(k) := \frac{1}{n}\sum_i \nabla f_i(\bar{x}(k)).$$

Corollary 1. *Under Assumption 1, if $\|h(k)\| \le D$ for all k and $\beta < 1$, then*

$$\|\nabla f_i(x_{(i)}(k)) - \nabla f_i(\bar{x}(k))\| \le \frac{\alpha D L_{f_i}}{1-\beta},$$
$$\|g(k) - \bar{g}(k)\| \le \frac{\alpha D L_h}{1-\beta}.$$

Proof. Since Assumption 1 holds,

$$\|\nabla f_i(x_{(i)}(k)) - \nabla f_i(\bar{x}(k))\| \le L_{f_i}\|x_{(i)}(k) - \bar{x}(k)\| \le \frac{\alpha D L_{f_i}}{1-\beta}.$$

The last inequality holds per Theorem 2. On the other hand,

$$\|g(k) - \bar{g}(k)\| = \|\frac{1}{n}\sum_i \big(\nabla f_i(x_{(i)}(k)) - \nabla f_i(\bar{x}(k))\big)\| \le \frac{1}{n}\sum_i L_{f_i}\|x_{(i)}(k) - \bar{x}(k)\| \le \frac{\alpha D L_h}{1-\beta},$$

which completes the proof. □

We are interested in $g(k)$ since $-\alpha g(k)$ updates the average of $x_{(i)}(k)$. To see this, taking the average of (3) over i and noticing $W = [w_{ij}]$ is doubly stochastic give us

$$\bar{x}(k+1) = \frac{1}{n}\sum_{i=1}^n x_{(i)}(k+1) = \frac{1}{n}\sum_i\sum_j w_{ij}x_{(j)} - \frac{\alpha}{n}\sum_i \nabla f_i(x_{(i)}(k)) = \bar{x}(k) - \alpha g(k). \tag{13}$$

On the other hand, since the exact gradient of $\frac{1}{n}\sum_i f_i(\bar{x}(k))$ is $\bar{g}(k)$, iteration (13) can be viewed as an inexact gradient descent iteration (using $g(k)$ instead of $\bar{g}(k)$) for problem

$$\underset{x}{\text{minimize}} \;\; \bar{f}(x) := \frac{1}{n}\sum_i f_i(x). \tag{14}$$

It is easy to see that $\bar{f}$ is Lipschitz continuous with constant

$$L_{\bar{f}} = \frac{1}{n}\sum_{i=1}^n L_{f_i}.$$

If any f_i is strongly convex, then so is $\bar{f}$ with modulus $\mu_{\bar{f}} = \frac{1}{n}\sum_i \mu_{f_i}$. Based on this observation, we next bound $f(\bar{x}(k)) - f^*$ and $\|\bar{x}(k) - x^*\|$.

2.3 Bounded distance to minimum

We consider the convex, restricted strongly convex, and strongly convex cases. In the former two cases, the solution x^* may be non-unique, so we use the set of solutions $\mathcal{X}^*$. Define two errors for our analysis

- objective error $\bar{r}(k) := \bar{f}(\bar{x}(k)) - \bar{f}^* = \frac{1}{n}(f(\bar{x}(k)) - f^*)$ where $\bar{f}^* := \bar{f}(x^*)$, $x^* \in \mathcal{X}^*$;
- solution error $\bar{e}(k) := \bar{x}(k) - x^*(k)$ where $x^*(k) = \mathrm{Proj}_{\mathcal{X}^*}(\bar{x}(k)) \in \mathcal{X}^*$.

Theorem 3. *Under Assumption 1, if $\alpha \le \min\{(1+\lambda_n(W))/L_h, 1/L_{\bar{f}}\} = O(1/L_h)$, then while*

$$\bar{r}(k) > C\sqrt{2} \cdot \frac{\alpha L_h D}{(1-\beta)} = O\left(\frac{\alpha}{1-\beta}\right)$$

(constants C and D are defined in (15) and (8), respectively), the reduction of $\bar{r}(k)$ obeys

$$\bar{r}(k+1) \le \bar{r}(k) - O(\alpha \bar{r}^2(k)),$$

and therefore,

$$\bar{r}(k) \le O\left(\frac{1}{\alpha k}\right),$$

i.e., $\bar{r}(k)$ decreases at a minimal rate of $O(\frac{1}{\alpha k}) = O(1/k)$ until reaching $O(\frac{\alpha}{1-\beta})$.

Proof. First we show that $\|\bar{e}(k)\| \le C$. To this end, recall the definition of $\xi_\alpha([x_{(i)}])$ in (9). Let $\hat{\mathcal{X}}$ denote its set of minimizer(s), which is nonempty since each f_i has a minimizer from Assumption 1. Following the arguments in [20, pp. 69] and with the bound on α, we have $d(k) \le d(k-1) \le \cdots \le d(0)$, where $d(k) := \|[x_{(i)}(k) - \hat{x}_i]\|$ and $[\hat{x}_i] \in \hat{\mathcal{X}}$. Using $\|a_1 + \cdots + a_n\| \le \sqrt{n}\|(a_1; \ldots; a_n)\|$, we have

$$\begin{aligned}
\|\bar{e}(k)\| = \|\bar{x}(k) - x^*(k)\| &= \|\frac{1}{n}\sum_i (x_{(i)}(k) - x^*)\| \le \frac{1}{\sqrt{n}}\|[x_{(i)}(k) - x^*]\| \\
&\le \frac{1}{\sqrt{n}}(\|[x_{(i)}(k) - \hat{x}_i]\| + \|[\hat{x}_i - \hat{x}^*]\|) \\
&\le \frac{1}{\sqrt{n}}(\|[x_{(i)}(0) - \hat{x}_i]\| + \|[\hat{x}_i - \hat{x}^*]\|) =: C
\end{aligned} \tag{15}$$

Next we show the convergence of $\bar{r}(k)$. By assumption, we have $1 - \alpha L_{\bar{f}} \ge 0$, and thus

$$\begin{aligned}
\bar{r}(k+1) &\le \bar{r}(k) + \langle \bar{g}(k), \bar{x}(k+1) - \bar{x}(k)\rangle + \frac{L_{\bar{f}}}{2}\|\bar{x}(k+1) - \bar{x}(k)\|^2 \\
&\overset{(13)}{=} \bar{r}(k) - \alpha\langle \bar{g}(k), g(k)\rangle + \frac{\alpha^2 L_{\bar{f}}}{2}\|g(k)\|^2 \\
&= \bar{r}(k) - \alpha\langle \bar{g}(k), \bar{g}(k)\rangle + \frac{\alpha^2 L_{\bar{f}}}{2}\|\bar{g}(k)\|^2 + 2\alpha\frac{1-\alpha L_{\bar{f}}}{2}\langle \bar{g}(k), \bar{g}(k) - g(k)\rangle + \frac{\alpha^2 L_{\bar{f}}}{2}\|\bar{g}(k) - g(k)\|^2 \\
&\le \bar{r}(k) - \alpha(1 - \frac{\alpha L_{\bar{f}}}{2} - \delta\frac{1-\alpha L_{\bar{f}}}{2})\|\bar{g}(k)\|^2 + \alpha(\frac{\alpha L_{\bar{f}}}{2} + \delta^{-1}\frac{1-\alpha L_{\bar{f}}}{2})\|\bar{g}(k) - g(k)\|^2,
\end{aligned}$$

where the last inequality follows from $\pm 2a^T b \le \delta^{-1}\|a\|^2 + \delta\|b\|^2$ for any $\delta > 0$. Although we can later optimize over $\delta > 0$, for simplicity, we take $\delta = 1$. Since $\alpha \le (1+\lambda_n(W))/L_h$, we can apply Theorem 1 and then Corollary 1 to the last term above, and we obtain

$$\bar{r}(k+1) \le \bar{r}(k) - \frac{\alpha}{2}\|\bar{g}(k)\|^2 + \frac{\alpha^3 D^2 L_h^2}{2(1-\beta)^2}.$$

Since $\|\bar{e}(k)\| \leq C$ as shown above, from $\bar{r}(k) = \bar{f}(\bar{x}(k)) - \bar{f}^* \leq \langle \bar{g}(k), \bar{x}(k) - x^*(k) \rangle = \langle \bar{g}(k), \bar{e}(k) \rangle$, we have that

$$\|\bar{g}(k)\| \geq \|\bar{g}(k)\| \frac{\|\bar{e}(k)\|}{C} \geq \frac{|\langle \bar{g}(k), \bar{e}(k) \rangle|}{C} \geq \frac{\bar{r}(k)}{C},$$

which gives

$$\bar{r}(k+1) \leq \bar{r}(k) - \frac{\alpha}{2C^2} \bar{r}^2(k) + \frac{\alpha^3 D^2 L_h^2}{2(1-\beta)^2}.$$

Hence, while $\frac{\alpha}{2C^2} \bar{r}^2(k) > 2 \cdot \frac{\alpha^3 D^2 L_h^2}{2(1-\beta)^2}$ or equivalently $\bar{r}(k) > C\sqrt{2} \cdot \frac{\alpha L_h D}{(1-\beta)}$, we have $\bar{r}(k+1) \leq \bar{r}(k) - O(\alpha \bar{r}^2(k))$. Dividing both sides by $\bar{r}(k)\bar{r}(k+1)$ gives $\frac{1}{\bar{r}(k)} + O(\frac{\alpha \bar{r}(k)}{\bar{r}(k+1)}) \leq \frac{1}{\bar{r}(k+1)}$. Hence, $\frac{1}{\bar{r}(k)}$ increase at $\Omega(\alpha k)$, or $\bar{r}(k)$ reduces at $O(1/(\alpha k))$, which completes the proof. □

Theorem 3 shows that until reaching $f^* + O(\frac{\alpha}{1-\beta})$, $f(\bar{x}(k))$ reduces at the rate of $O(1/(\alpha k))$. For fixed α, there is tradeoff between convergence rate and optimality. Again, upon the stopping of iteration (3), $\bar{x}(k)$ is not available to any of the agents but can be obtained by invoking an average consensus algorithm.

Theorem 3 shares similarity with the nearly sublinear convergence of the distributed subgradient method [19] and the dual averaging subgradient method [7]. However, [19] and [7] assume bounded (sub)gradients of f_i. In Theorem 3, we remove this assumption using the fact that a bounded stepsize leads to bounded gradients (cf. Theorem 1).

Next, we bound $\|\bar{e}(k+1)\|$ by assuming restricted or standard strong convexities in a unified framework. To start, we present a lemma.

Lemma 1. *Suppose $\nabla \bar{f}$ is Lipschitz continuous with constant $L_{\bar{f}}$. We have*

$$\langle x - x^*, \nabla \bar{f}(x) - \nabla \bar{f}(x^*) \rangle \geq c_1 \|\nabla \bar{f}(x) - \nabla \bar{f}(x^*)\|^2 + c_2 \|x - x^*\|^2$$

(where $x^* \in \mathcal{X}^*$ and $\nabla \bar{f}(x^*) = 0$) for the following cases:

a) ([20, Theorem 2.1.12]) if $\bar{f}$ is strongly convex with modulus $\mu_{\bar{f}}$, then $c_1 = \frac{1}{\mu_{\bar{f}} + L_{\bar{f}}}$ and $c_2 = \frac{\mu_{\bar{f}} L_{\bar{f}}}{\mu_{\bar{f}} + L_{\bar{f}}}$;

b) ([33, Lemma 2]) if $\bar{f}$ is restricted strongly convex with modulus $\nu_{\bar{f}}$, then $c_1 = \frac{\theta}{L_{\bar{f}}}$ and $c_2 = (1-\theta)\nu_{\bar{f}}$ for any $\theta \in [0, 1]$.

Theorem 4. *Under Assumptions 1, if f is either strongly convex with modulus μ_f or restricted strongly convex with modulus ν_f, and if stepsize $\alpha \leq \min\{(1 + \lambda_n(W))/L_h, c_1\} = O(1/L_h)$ and $\beta < 1$, then we have*

$$\|\bar{e}(k+1)\|^2 \leq c_3^2 \|\bar{e}(k)\|^2 + c_4^2,$$

where

$$c_3^2 = 1 - \alpha c_2 + \alpha\delta - \alpha^2 \delta c_2, \quad c_4^2 = \alpha^3(\alpha + \delta^{-1}) \frac{L_h^2 D^2}{(1-\beta)^2}, \quad D = \sqrt{2L_h \sum_i (f_i(0) - f_i^*)},$$

constants c_1 and c_2 are given in Lemma 1, $\mu_{\bar{f}} = \mu_f/n$ and $\nu_{\bar{f}} = \nu_f/n$, and δ is any positive constant. In particular, if we set $\delta = \frac{c_2}{2(1-\alpha c_2)}$ such that $c_3 = \sqrt{1 - \frac{\alpha c_2}{2}} \in (0, 1)$, then we have

$$\|\bar{e}(k)\| \leq c_3^k \|\bar{e}(0)\| + O(\frac{\alpha}{1-\beta}).$$

Proof. Recalling that $x^*(k+1) = \text{Proj}_{\mathcal{X}^*}(\bar{x}(k+1))$ and $\bar{e}(k+1) = \bar{x}(k+1) - x^*(k+1)$, we have

$$\begin{aligned}
\|\bar{e}(k+1)\|^2 &\le \|\bar{x}(k+1) - x^*(k)\|^2 \\
&= \|\bar{x}(k) - x^*(k) - \alpha g(k)\|^2 \\
&= \|\bar{e}(k) - \alpha\bar{g}(k) + \alpha(\bar{g}(k) - g(k))\|^2 \\
&= \|\bar{e}(k) - \alpha\bar{g}(k)\|^2 + \alpha^2\|\bar{g}(k) - g(k)\|^2 + 2\alpha(\bar{g}(k) - g(k))^T(\bar{e}(k) - \alpha\bar{g}(k)) \\
&\le (1+\alpha\delta)\|\bar{e}(k) - \alpha\bar{g}(k)\|^2 + \alpha(\alpha + \delta^{-1})\|\bar{g}(k) - g(k)\|^2,
\end{aligned}$$

where the last inequality follows again from $\pm 2a^Tb \le \delta^{-1}\|a\|^2 + \delta\|b\|^2$ for any $\delta > 0$. The bound of $\|\bar{g}(k) - g(k)\|^2$ follows from Corollary 1 and Theorem 1, and we shall bound $\|\bar{e}(k) - \alpha\bar{g}(k)\|^2$, which is a standard exercise; we repeat below for completeness. Applying Lemma 1 and noticing $\bar{g}(x) = \nabla\bar{f}(x)$ by definition, we have

$$\begin{aligned}
\|\bar{e}(k) - \alpha\bar{g}(k)\|^2 &= \|\bar{e}(k)\|^2 + \alpha^2\|\bar{g}(k)\|^2 - 2\alpha\bar{e}(k)^T\bar{g}(k) \\
&\le \|\bar{e}(k)\|^2 + \alpha^2\|\bar{g}(k)\|^2 - \alpha c_1\|\bar{g}(k)\|^2 - \alpha c_2\|\bar{e}(k)\|^2 \\
&= (1 - \alpha c_2)\|\bar{e}(k)\|^2 + \alpha(\alpha - c_1)\|\bar{g}(k)\|^2.
\end{aligned}$$

We shall pick $\alpha \le c_1$ so that $\alpha(\alpha - c_1)\|\bar{g}(k)\|^2 \le 0$. Then from the last two inequality arrays, we have

$$\begin{aligned}
\|\bar{e}(k+1)\|^2 &\le (1+\alpha\delta)(1 - \alpha c_2)\|\bar{e}(k)\|^2 + \alpha(\alpha + \delta^{-1})\|\bar{g}(k) - g(k)\|^2 \\
&\le (1 - \alpha c_2 + \alpha\delta - \alpha^2\delta c_2)\|\bar{e}(k)\|^2 + \alpha^3(\alpha + \delta^{-1})\frac{L_h^2 D^2}{(1-\beta)^2}.
\end{aligned}$$

Note that if f is strongly convex, $c_1c_2 = \frac{\mu_{\bar{f}} L_{\bar{f}}}{(\mu_{\bar{f}} + L_{\bar{f}})^2} < 1$; if f is restricted strongly convex, $c_1c_2 = \frac{\theta(1-\theta)\nu_{\bar{f}}}{L_{\bar{f}}} < 1$ because $\theta \in [0,1]$ and $\nu_{\bar{f}} < L_{\bar{f}}$, therefore we have $c_1 < 1/c_2$. When $\alpha < c_1$, $(1+\alpha\delta)(1-\alpha c_2) > 0$.

Next, since

$$\|\bar{e}(k)\|^2 \le c_3^{2k}\|\bar{e}(0)\|^2 + \frac{1 - c_3^{2k}}{1 - c_3^2}c_4^2 \le c_3^{2k}\|\bar{e}(0)\|^2 + \frac{c_4^2}{1 - c_3^2},$$

and thus

$$\|\bar{e}(k)\| \le c_3^k\|\bar{e}(0)\| + \frac{c_4}{\sqrt{1 - c_3^2}}.$$

If we set

$$\delta = \frac{c_2}{2(1 - \alpha c_2)},$$

then we obtain

$$c_3^2 = 1 - \frac{\alpha c_2}{2} < 1,$$

$$\frac{c_4}{\sqrt{1 - c_3^2}} = \frac{\alpha L_h D}{1 - \beta}\sqrt{\frac{\alpha(\alpha + \frac{2(1-\alpha c_2)}{c_2})}{\frac{\alpha c_2}{2}}} = \frac{\alpha L_h D}{1 - \beta}\sqrt{\frac{4}{c_2^2} - \frac{2}{c_2}\alpha} = O(\frac{\alpha}{1 - \beta}),$$

and completes the proof. □

Remark 1. *As a result, if f is strongly convex, then $\bar{x}(k)$ geometrically converges until reaching an $O(\frac{\alpha}{1-\beta})$-neighborhood of the unique solution x^*; on the other hand, if f is restricted strongly convex, then $\bar{x}(k)$ geometrically converges until reaching an $O(\frac{\alpha}{1-\beta})$-neighborhood of the optimal solution set $\mathcal{X}^*$.*

2.4 Local agent convergence

Corollary 2. *Under Assumption 1, if f is either strongly convex or restricted strongly convex, the stepsize $\alpha < \min\{(1+\lambda_n(W))/L_h, c_1\}$ and $\beta < 1$, we have*

$$\|x_{(i)}(k) - x^*(k)\| \leq c_3^k \|x^*(0)\| + \frac{c_4}{\sqrt{1-c_3^2}} + \frac{\alpha D}{1-\beta},$$

where $x^*(0), x^*(k) \in \mathcal{X}^*$ are solutions defined at the beginning of subsection 2.3 and the constants c_3, c_4 and D are the same as given in Theorem 4.

Proof. From Theorems 2 and 4 we have

$$\begin{aligned}
&\|x_{(i)}(k) - x^*(k)\| \\
\leq &\|\bar{x}(k) - x^*(k)\| + \|x_{(i)}(k) - \bar{x}(k)\| \\
\leq &c_3^k \|x^*(0)\| + \frac{c_4}{\sqrt{1-c_3^2}} + \frac{\alpha D}{1-\beta},
\end{aligned}$$

which completes the proof. □

Remark 2. *Similar to Theorem 4 and Remark 1, if we set $\delta = \frac{c_2}{2(1-\alpha c_2)}$, and if f is strongly convex, then $x_i(k)$ geometrically converges to an $O(\frac{\alpha}{1-\beta})$-neighborhood of the unique solution x^*; if f is restricted strongly convex, then $x_i(k)$ geometrically converges to an $O(\frac{\alpha}{1-\beta})$-neighborhood of the optimal solution set $\mathcal{X}^*$.*

3 Decentralized basis pursuit

3.1 Problem statement

We derive an algorithm for solving a decentralized basis pursuit problem to illustrate the application of iteration (3).

Consider a multi-agent network of n agents who collaboratively find a sparse representation y of a given signal $b \in \mathbb{R}^p$ that is known to all the agents. Each agent i has a part $A_i \in \mathbb{R}^{p \times q_i}$ of the entire dictionary $A \in \mathbb{R}^{p \times q}$, where $q = \sum_{i=1}^n q_i$, and shall recover the corresponding $y_i \in \mathbb{R}^{q_i}$. Let

$$y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^q, \quad A := \begin{bmatrix} | & & | \\ A_1 & \dots & A_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{p \times q}.$$

The problem is

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & \|y\|_1, \\ \text{subject to} \quad & \sum_{i=1}^{n} A_i y_i = b, \end{aligned} \tag{16}$$

where $\sum_{i=1}^{n} A_i y_i = Ay$.

Problem (16) finds applications in, for example, collaborative spectrum sensing [1], sparse event detection [17], and seismic modeling [18]. Take seismic modeling [18] as an example. There are n distributed sources of random ambient vibrations whose signals are measured by p receivers. The measurement is $b = \sum_i A_i y_i$, where A_i is known by source i and the sparse vector $y = [y_1; \ldots; y_n]$ is to be determined. The n sources are geographically far apart and A_i are much larger in size than y_i and b. A decentralized algorithm avoids the need of long-distance communication of large amounts of data. In general, (16) arises where the observation b is a linear combination of distributed sparse sources y_i that shall be recovered.

Developing efficient decentralized algorithms to solve (16) is nontrivial since the objective function is neither differentiable nor strongly convex, and the constraint couples all the agents. Paper [18] proposes a decentralized ADMM algorithm in which every agent needs to solve an optimization subproblem at each iteration, which generally requires much more computing power that computing a gradient. In this paper, we turn to an equivalent and tractable reformulation by appending a strongly convex term and solving its Lagrange dual problem by decentralized gradient descent. Consider the augmented form of (16) motivated by [13]

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & \|y\|_1 + \frac{1}{2\gamma}\|y\|^2, \\ \text{subject to} \quad & Ay = b, \end{aligned} \tag{17}$$

where the regularization parameter $\gamma > 0$ is chosen so that (17) returns a solution to (16). Indeed, provided that $Ay = b$ is consistent, there always exists $\gamma_{\min} > 0$ such that the solution to (17) is also a solution to (16) for any $\gamma \geq \gamma_{\min}$ [8, 31]. Setting $\gamma = 10\|y^o\|_\infty$ or larger, where y^o is the true signal, is shown to work well with recovery guarantees in [13] given that A satisfies certain properties that are commonly assumed in compressive sensing. The Lagrange dual of (17), casted as a minimization (instead of maximization) problem, is

$$\underset{x}{\text{minimize}} \;\; f(x) := \frac{\gamma}{2}||A^T x - \text{Proj}_{[-1,1]}(A^T x)||^2 - b^T x, \tag{18}$$

where $x \in \mathbb{R}^p$ is the dual variable and $\text{Proj}_{[-1,1]}$ denotes element-wise projection onto interval $[-1, 1]$.

We turn (18) into the form of (1):

$$\underset{x}{\text{minimize}} \;\; f(x) = \sum_{i=1}^{n} f_i(x), \;\; \text{where } f_i(x) := \frac{\gamma}{2}||A_i^T x - \text{Proj}_{[-1,1]}(A_i^T x)||^2 - \frac{1}{n} b^T x. \tag{19}$$

Function f_i is defined with A_i and b, where matrix A_i is the private information of agent i. The local objective functions f_i are differentiable with gradients given as

$$\nabla f_i(x) = \gamma A_i \text{Shrink}(A_i^T x) - \frac{b}{n}, \tag{20}$$

where $\text{Shrink}(z)$ is the shrinkage operator defined as $\max(|z| - 1, 0)\text{sign}(z)$ component-wise.

Applying iteration (3) to problem (19) starting with $x_{(i)}(0) = 0$, we obtain the iteration

$$\boxed{x_{(i)}(k+1) = \sum_j w_{ij} x_{(j)}(k) - \alpha \left(A_i y_i(k) - \frac{b}{n} \right), \quad \text{where } y_i(k) = \gamma \text{Shrink}(A_i^T x_{(i)}(k)).} \tag{21}$$

Note that the primal solution $y_i(k)$ is iteratively updated, as a middle step for the update of $x_{(i)}(k+1)$.

If the basis pursuit problem is noise-polluted, the equality constraint $Ax = b$ in (17) can be replaced by an inequality constraint $\|Ax - b\| \leq \sigma$, where σ is an estimate of the noise magnitude, or a penalty term $\frac{\mu}{2}\|Ax - b\|^2$ can be introduced to the objective. A dual approach similar to the above can be applied to these noise-polluted case to design either centralized [13] or decentralized basis pursuit algorithms.

3.2 Dual and primal convergence

Now we show that the local objective functions f_i satisfy Assumption 1.

$$\begin{aligned} \|\nabla f_i(x_a) - \nabla f_i(x_b)\| =& \|\gamma A_i \text{Shrink}(A_i^T x_a) - \gamma A_i \text{Shrink}(A_i^T x_b)\| \\ \leq& \gamma \|A_i\| \|\text{Shrink}(A_i^T x_a) - \text{Shrink}(A_i^T x_b)\|. \end{aligned} \tag{22}$$

Again $\|\text{Shrink}(A_i^T x_a) - \text{Shrink}(A_i^T x_b)\| \leq \|A_i\|\|x_a - x_b\|$ due to the nonexpansiveness of the shrinkage operator. Hence we have

$$\|\nabla f_i(x_a) - \nabla f_i(x_b)\| \leq \gamma \|A_i\|^2 \|x_a - x_b\|, \tag{23}$$

which implies that the Lipschitz constant is $L_{f_i} = \gamma \|A_i\|^2$.

Given that $Ay = b$ is consistent, [13] proves that $f(x)$ is restricted strongly convex. Define y^* as the unique solution to (17), $\text{supp}(y^*)$ as the support of y^*, and y_i^* as the ith part of y^*. There exists a positive constant ν_f such that for any point x and its projection onto the optimal solution set of (18) $\text{Proj}_{\mathcal{X}^*}(x)$ it holds

$$[x - \text{Proj}_{\mathcal{X}^*}(x)]^T [\nabla f(x) - \nabla f(\text{Proj}_{\mathcal{X}^*}(x))] \geq \nu_f \|x - \text{Proj}_{\mathcal{X}^*}(x)\|^2, \tag{24}$$

where

$$\nu_f = \lambda_A \min_{i \in \text{supp}(y^*)} \frac{\gamma |y_i^*|}{|y_i^*| + 2\gamma}, \tag{25}$$

and λ_A is the smallest positive eigenvalue of $U^T U$ with U being any nonzero submatrix of A with p rows.

For the objective function in (18), $L_h = \max\{\gamma\|A_i\|^2 : i = 1, 2, \cdots, n\}$, $L_{\bar{f}} = \frac{\gamma}{n}\sum_{i=1}^n \|A_i\|^2$ and $\nu_{\bar{f}} = \nu_f/n$. Based on Theorem 4, we have the following convergence result of iteration (21); specifically, any local dual solution $x_{(i)}(k)$ linearly converges to a neighborhood of the solution set of (18) and the primal solution $y(k) = [y_1(k); \cdots; y_n(k)]$ linearly converges to a neighborhood of the unique solution of (17).

Theorem 5. *Consider $x_{(i)}(k)$ generated by iteration (21) and $\bar{x}(k) := \frac{1}{n}\sum_{i=1}^n x_{(i)}(k)$. The unique solution of (17) is y^* and the projection of $\bar{x}(k)$ onto the optimal solution set of (18) is $\bar{x}^*(k) = Proj_{\mathcal{X}^*}(\bar{x}(k))$. If the stepsize $\alpha < \min\{(1+\lambda_n(W))/L_h, c_1\}$, we have*

$$\|x_{(i)}(k) - \bar{x}^*(k)\| \le c_3^k\|\bar{x}^*(0)\| + \left(\frac{c_4}{\sqrt{1-c_3^2}} + \frac{\alpha D}{1-\beta}\right),$$

where the constants c_3 and c_4 are the same as given in Theorem 4. In particular, if we set $\delta = \frac{c_2}{2(1-\alpha c_2)}$ such that $c_3 = \sqrt{1-\frac{\alpha c_2}{2}} \in (0,1)$, then $\frac{c_4}{\sqrt{1-c_3^2}} + \frac{\alpha D}{1-\beta} = O(\frac{\alpha}{1-\beta})$. On the other hand, the primal solution satisfies

$$\|y(k) - y^*\| \le n\gamma \max_i \left(\|A_i\|\|x_{(i)}(k) - \bar{x}^*(k)\|\right). \tag{26}$$

Proof. The results on dual convergence is a corollary of Corollary 2. Hence we focus on primal convergence (26).

Given any dual solution $\bar{x}(k)$, the primal solution of (17) is $y^* = \gamma\text{Shrink}(A^T\bar{x}^*(k))$. Recall that $y(k) = [y_1(k); \cdots; y_n(k)]$ and $y_i(k) = \gamma\text{Shrink}(A_i^T x_{(i)}(k))$

$$\begin{aligned}\|y(k) - y^*\| =& \|(\gamma\text{Shrink}(A_1^T x_1(k)); \cdots; \gamma\text{Shrink}(A_n^T x_n(k))) - \gamma\text{Shrink}(A^T\bar{x}^*(k))\| \\ \le& \gamma\sum_{i=1}^n \|\text{Shrink}(A_i^T x_{(i)}(k)) - \text{Shrink}(A_i^T\bar{x}^*(k))\|.\end{aligned} \tag{27}$$

Due to contraction of the shrinkage operator we have the bound $\|\text{Shrink}(A_i^T x_{(i)}(k)) - \text{Shrink}(A_i^T\bar{x}^*(k))\| \le \|A_i\|\|x_{(i)}(k) - \bar{x}^*(k)\| \le \max_i \left(\|A_i\|\|x_{(i)}(k) - \bar{x}^*(k)\|\right)$. Combining this inequality with (27) we get (26) that completes the proof. □

4 Numerical experiments

In this section, we report our numerical results of iteration (3) on a decentralized least-squares problem and iteration (21) on a decentralized basis pursuit problem.

We generate a network consisting of n agents with $\frac{n(n-1)}{2}\eta$ edges that are uniformly randomly chosen, where $n = 100$ and $\eta = 0.3$ are chosen for all the tests. We ensure a connected network.

4.1 Decentralized gradient descent for least squares

We apply iteration (3) to the least-squares problem

$$\underset{x\in\mathbb{R}^3}{\text{minimize}} \quad \frac{1}{2}\|b - Ax\|^2 = \sum_{i=1}^n \frac{1}{2}\|b_i - A_i x\|^2. \tag{28}$$

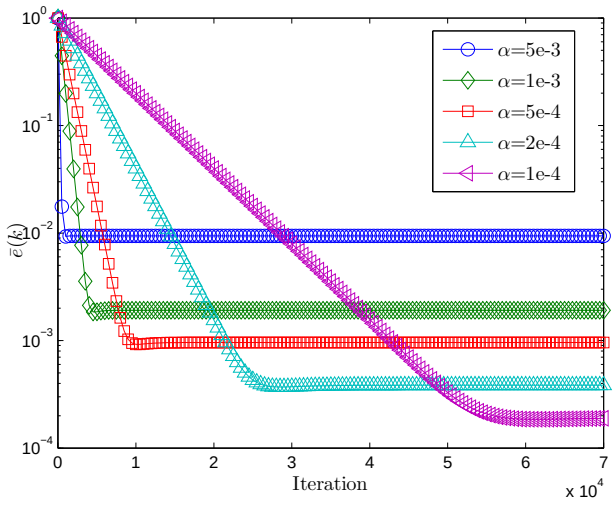

Figure 1: Comparison of the decentralized gradient descent algorithm with different fixed stepsizes.

The true signal $x^* \in \mathbb{R}^3$ whose entries are sampled i.i.d. from the Gaussian distribution $\mathcal{N}(0,1)$. $A_i \in \mathbb{R}^{3\times 3}$ is the linear sampling matrix of agent i whose elements are sampled i.i.d. from $\mathcal{N}(0,1)$, and $b_i = A_i x^* \in \mathbb{R}^3$ is the measurement vector of agent i.

For problem (28), let $f_i(x) = \frac{1}{2}\|b_i - A_i x\|^2$. For any $x_a,\ x_b \in \mathbb{R}^3$, $\|\nabla f_i(x_a) - \nabla f_i(x_b)\| = \|A_i^T A_i(x_a - x_b)\| \leq \|A_i^T A_i\|\|x_a - x_b\|$, so hence $\nabla f_i(x)$ is Lipschitz continuous. In addition, $\frac{1}{2}\|b - Ax\|_2^2$ is strongly convex since A has full column rank, with probability 1.

Fig. 1 depicts the convergence of the error $\bar{e}(k)$ corresponding to five different stepsizes. It shows that $\bar{e}(k)$ reduces linearly until reaching an $O(\alpha)$-neighborhood, which agrees with Theorem 4. Not surprisingly, a smaller α causes the algorithm to converge more slowly.

The final accuracy is clearly proportional to α. For example, the limit accuracy for $\alpha =$ 5e-3 is 10 times that for $\alpha =$ 5e-4. That for $\alpha =$ 1e-3 is 10 times that for $\alpha =$ 1e-4.

Fig. 2 compares our theoretical stepsize bound that ensures convergence (cf. Theorem 1) to what α needs to be in practice. The theoretical bound in this experimental network is $\min\{\frac{1+\lambda_n(W)}{L_h}, c_1\} = 0.1038$. In Fig. 2, we choose $\alpha = 0.1038$ and then a slightly larger $\alpha = 0.12$. We observe convergence with $\alpha = 0.1038$ but clear divergence with $\alpha = 0.12$. This shows that our bound on α is quite close to the actual requirement.

4.2 Decentralized gradient descent for basis pursuit

In this subsection we test iteration (21) on the decentralized basis pursuit problem (16).

Let $y \in \mathbb{R}^{100}$ be the unknown signal whose entries are sampled i.i.d. from $\mathcal{N}(0,1)$. The entries of the measurement matrix $A \in \mathbb{R}^{50\times 100}$ are also sampled i.i.d. from $\mathcal{N}(0,1)$. Each agent holds one column of A. $b = Ay \in \mathbb{R}^{50}$ is the measurement vector. We use the same network as the last test, and each agent i holds

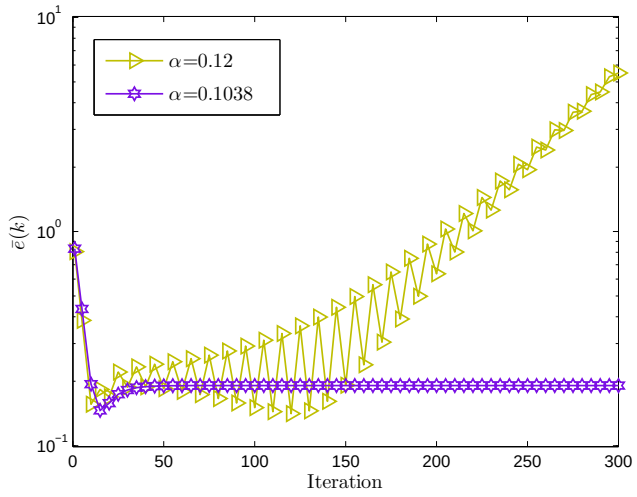

Figure 2: Comparison of the decentralized gradient descent algorithm with stepsizes $\alpha = 0.1038$ and $\alpha = 0.12$.

the ith column of A.

Fig. 3 depicts the convergence of $\bar{x}(k)$, the mean of the dual variables at iteration k. As stated in Theorem 5, $\bar{x}(k)$ converges linearly to an $O(\alpha)$-neighborhood of the solution set $\mathcal{X}^*$. The limiting errors $\bar{e}(k)$ corresponding to the four values of α are proportional to α. As the stepsize is chosen smaller, the algorithm converges more accurately to $\mathcal{X}^*$. Fig. 4 shows the linear convergence of the primal variable $y(k)$. It is interesting that $y(k)$ corresponding to three different values of α appear to reach the same level of accuracy, which might be related to the error forgetting property of a first-order ℓ_1 algorithm [32] and deserves further investigation.

5 Conclusion

Consensus optimization problems in multi-agent networks arise in applications such as mobile computing, self-driving cars' coordination, cognitive radios, as well as collaborative data mining. Compared to the traditional centralized approach, a decentralized approach offers more balanced communication load and better privacy protection. In this paper, our effort is to provide a mathematical understanding to the decentralized gradient descent method with a fixed stepsize. We give a tight condition for guaranteed convergence, as well as an example to illustrate the fail of convergence when the condition is violated. We provide the analysis of convergence and rates of convergence for problems with different properties and establish the relations between network topology, stepsize, and convergence speed, which shed some light on network design. The numerical observation reasonably matches the theoretical results.

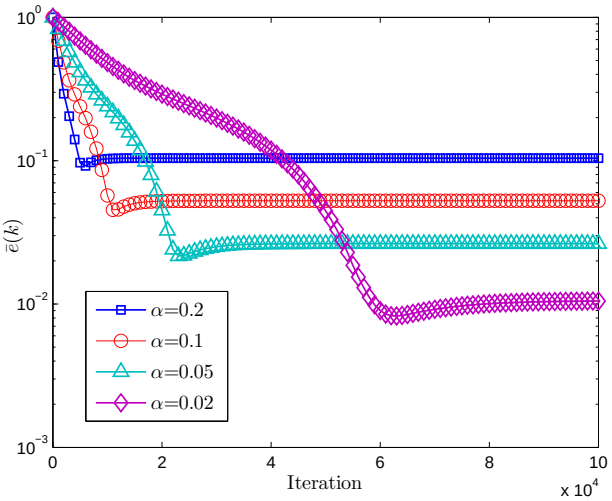

Figure 3: Convergence of the mean value of the dual variable $\bar{x}(k)$ in the decentralized gradient descent algorithm.

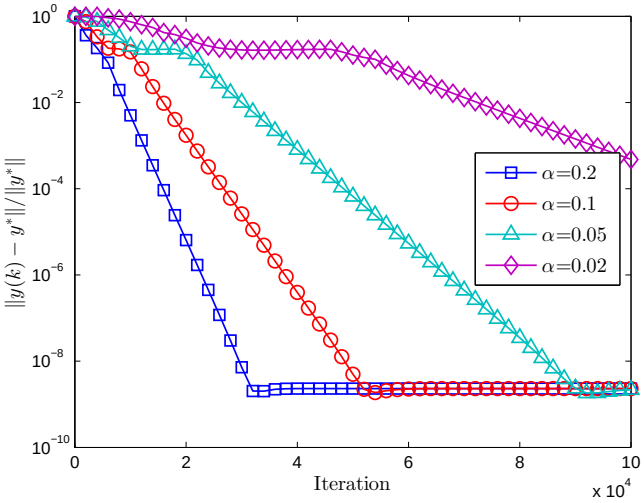

Figure 4: Convergence of the primal variable $y(k)$ in the decentralized gradient descent algorithm. y^* is the optimal solution of problem (17).

Acknowledgements

Q. Ling is supported by NSFC grant 61004137. W. Yin is supported by ARL and ARO grant W911NF-09-1-0383 and NSF grants DMS-0748839 and DMS-1317602. In addition, the authors thank Yangyang Xu for helpful comments.

References

[1] J. A. Bazerque and G. B. Giannakis, *Distributed spectrum sensing for cognitive radio networks by exploiting sparsity*, IEEE Transactions on Signal Processing, 58 (2010), pp. 1847–1862.

[2] J. A. Bazerque, G. Mateos, and G. B. Giannakis, *Group-lasso on splines for spectrum cartography*, IEEE Transactions on Signal Processing, 59 (2011), pp. 4648–4663.

[3] S. Boyd, P. Diaconis, and L. Xiao, *Fastest mixing markov chain on a graph*, SIAM review, 46 (2004), pp. 667–689.

[4] Y. Cao, W. Yu, W. Ren, and G. Chen, *An overview of recent progress in the study of distributed multi-agent coordination*, IEEE Transactions on Industrial Informatics, 9 (2013), pp. 427–438.

[5] R. L. Cavalcante and S. Stanczak, *A distributed subgradient method for dynamic convex optimization problems under noisy information exchange*, IEEE Jounal of Selected Topics in Signal Processing, 7 (2013), pp. 243–256.

[6] J. Chen and A. H. Sayed, *Diffusion adaptation strategies for distributed optimization and learning over networks*, IEEE Transactions on Signal Processing, 60 (2012), pp. 4289–4305.

[7] 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 (2012), pp. 592–606.

[8] M. P. Friedlander and P. Tseng, *Exact regularization of convex programs*, SIAM Journal on Optimization, 18 (2007), pp. 1326–1350.

[9] G. B. Giannakis, V. Kekatos, N. Gatsis, S.-J. Kim, H. Zhu, and B. F. Wollenberg, *Monitoring and optimization for power grids: A signal processing perspective*, arXiv preprint arXiv:1302.0885, (2013).

[10] D. Jakovetic, J. Xavier, and J. M. Moura, *Fast distributed gradient methods*, arXiv preprint arXiv:1112.2972, (2013).

[11] F. Jakubiec and A. Ribeiro, *D-map: Distributed maximum a posteriori probability estimation of dynamic systems*, IEEE Transactions on Signal Processing, 61 (2013), pp. 450–466.

[12] V. Kekatos and G. B. Giannakis, *Distributed robust power system state estimation*, IEEE Transactions on Power Systems, (To appear).

[13] M.-J. Lai and W. Yin, *Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1059–1091.

[14] Q. Ling and A. Ribeiro, *Decentralized dynamic optimization through the alternating direction method of multipliers*, in 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC), IEEE, 2013, pp. 170–174.

[15] Q. Ling and Z. Tian, *Decentralized sparse signal recovery for compressive sleeping wireless sensor networks*, IEEE Transactions on Signal Processing, 58 (2010), pp. 3816–3827.

[16] Q. Ling, Z. Wen, and W. Yin, *Decentralized jointly sparse optimization by reweighted ℓ_q minimization*, IEEE Transactions on Signal Processing, 61 (2013), pp. 1165–1170.

[17] J. Meng, H. Li, and Z. Han, *Sparse event detection in wireless sensor networks using compressive sensing*, in Information Sciences and Systems, 2009. CISS 2009. 43rd Annual Conference on, IEEE, 2009, pp. 181–185.

[18] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Puschel, *Distributed basis pursuit*, IEEE Transactions on Signal Processing, 60 (2012), pp. 1942–1956.

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

[20] Y. Nesterov, *Gradient methods for minimizing composite objective function*, 2007.

[21] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proceedings of the IEEE, 95 (2007), pp. 215–233.

[22] J. B. Predd, S. Kulkarni, and H. V. Poor, *Distributed learning in wireless sensor networks*, IEEE Signal Processing Magazine, 23 (2006), pp. 56–69.

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

[24] W. Ren, R. W. Beard, and E. M. Atkins, *Information consensus in multivehicle cooperative control*, IEEE Control Systems, 27 (2007), pp. 71–82.

[25] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, *Consensus in ad hoc wsns with noisy links – part i: Distributed estimation of deterministic signals*, IEEE Transactions on Signal Processing, 56 (2008), pp. 350–364.

[26] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, *Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning*, in 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2012, pp. 1543–1550.

[27] K. I. Tsianos and M. G. Rabbat, *Distributed strongly convex optimization*, in 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2012, pp. 593–600.

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

[29] J. N. Tsitsiklis, *Problems in decentralized decision making and computation.*, tech. report, DTIC Document, 1984.

[30] F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi, *Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties*, IEEE Transactions on Knowledge and Data Engineering, (To appear).

[31] W. Yin, *Analysis and generalizations of the linearized bregman method*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 856–877.

[32] W. Yin and S. Osher, *Error forgetting of bregman iteration*, Journal of Scientific Computing, 54 (2013), pp. 684–695.

[33] H. Zhang and W. Yin, *Gradient methods for convex minimization: better rates under weaker conditions*, arXiv preprint arXiv:1303.4645, (2013).

[34] F. Zhao, J. Shin, and J. Reich, *Information-driven dynamic sensor collaboration*, IEEE Signal Processing Magazine, 19 (2002), pp. 61–72.

[35] K. Zhou and S. I. Roumeliotis, *Multirobot active target tracking with combinations of relative observations*, IEEE Transactions on Robotics, 27 (2011), pp. 678–695.