Convergence Rate Analysis of Distributed Gradient Methods for Smooth Optimization

Dušan Jakovetić, João Xavier, and José M. F. Moura

***Abstract*— We derive the convergence rate of a distributed gradient algorithm for smooth optimization in networked systems. We assume that each agent in the network has a convex cost function $f_i(x)$, known only to agent i, and the goal is to minimize the sum $\sum_{i=1}^N f_i(x)$ of all agents' costs; such problem formulation arises in various networked applications, like, e.g., distributed inference or source localization in sensor networks. With the distributed gradient algorithm under study, each agent, at each iteration k, performs a weighted average of its own and its neighbors' solution estimates, and performs a step in the direction of the negative of its local function's gradient. We establish a novel result that the distributed gradient algorithm has the convergence rate $O(1/k^{2/3})$, in terms of the cost function optimality gap, under the assumption of convex f_i's with Lipschitz continuous and bounded gradients.**

***Keywords*— Convergence rate analysis, Consensus, Distributed optimization, Gradient methods.**

I. INTRODUCTION

We consider distributed optimization in networked systems where N agents are situated in a generic, connected network; each agent i (out of N agents) has a convex cost function $f_i : \mathbb{R}^d \to \mathbb{R}$, known only to agent i, and the goal is to minimize the sum $\sum_{i=1}^N f_i(x)$ of the individual agents' costs. The latter problem encompasses many applications in networked systems, like distributed estimation and source localization in sensor networks [1], or distributed learning, e.g., of a linear classifier [2].

The literature proposes distributed (sub)gradient methods to solve the described or related problems, see [3], and more recent works [4], [2]. These methods are attractive as they are fully distributed and have simple, computationally cheap iterations k. The literature mostly analyzes the convergence rates of these methods for a wide class of nondifferentiable convex f_i's that: 1) have bounded subgradients, for unconstrained problems; or 2) are Lipschitz over the constraint set, for constrained problems. Specifically, reference [2] shows that the distributed algorithm therein achieves the convergence rate (in terms of the optimality gap at the cost function) $O\left(\frac{\log k}{\sqrt{k}}\right)$ for the second class of the f_i's. In this paper, we analyze distributed gradient algorithms under a more restrictive class $\mathcal{F}$ of differentiable f_i's with Lipschitz continuous and bounded gradients. Such a class is still very interesting as it subsumes many important costs, like the logistic loss in classification, e.g., [5], or the Huber loss in robust estimation, e.g., [1]. It is natural to expect that distributed gradient algorithms achieve a faster guaranteed convergence rate under this "smaller" class of functions, as this situation is typical in conventional, centralized optimization. For example, the centralized (sub)gradient method has the rate $O\left(\frac{1}{\sqrt{k}}\right)$ on the class of nondifferentiable convex costs, while the rate increases to $O\left(\frac{1}{k}\right)$ when the cost function is convex, differentiable, with Lipschitz continuous derivative. To date, a corresponding result for smooth optimization has not been established for distributed gradient methods.

Main contribution. In this paper, we show that the distributed (sub)gradient algorithm in [4] achieves the rate $O\left(\frac{1}{k^{2/3}}\right)$ on the class $\mathcal{F}$ of convex f_i's with Lipschitz continuous and bounded gradients. We consider a family of the algorithm step-sizes $\alpha_k = c/(k+1)^\tau$, $c > 0$, parameterized by $\tau \in [0,1]$, and we show that the choice $\tau = 1/3$ assures the best guaranteed rate – $O\left(\frac{1}{k^{2/3}}\right)$. Interestingly, the obtained rate $O\left(\frac{1}{k^{2/3}}\right)$ for the distributed gradient algorithm differs (is worse) from the corresponding rate of the centralized gradient algorithm $O\left(\frac{1}{k}\right)$. The upper bound $O\left(\frac{1}{k^{2/3}}\right)$ that we obtain here for the algorithm in [4] essentially cannot be improved. Namely, reference [6] demonstrated that a worst case optimality gap under the class $\mathcal{F}$ is no better than $\Omega\left(\frac{1}{k^{2/3}}\right)$. (See the paragraph with heading Notation for the meaning of symbols O, Ω.)

We contrast this paper with related work on distributed gradient methods [6], [2], [4], [5]. Reference [4] studies the same algorithm as we do here, but for nondifferentiable functions; reference [2] considers a different, dual averaging algorithm. Finally, references [6], [5] also analyze a different, accelerated distributed gradient method.

Paper organization. The next paragraph introduces notation. Section II describes the network and optimization models and outlines the algorithm. Section III states our main result on the convergence rate, while Section IV proves the result. Finally, Section V concludes the paper.

Notation. We denote by: $\mathbb{R}^d$ the d-dimensional real coordinate space, $d \geq 1$; A_{ij} the entry in the i-th row and j-th column of a matrix A; a_i the i-th entry of a vector a;

The work of the first and second authors is partially supported by: the Carnegie Mellon—Portugal Program under a grant from the Fundação para a Ciência e Tecnologia (FCT), Portugal; by FCT grants CMU-PT/SIA/0026/2009; and by ISR/IST plurianual funding (POSC program, FEDER). The work of the first and third authors is partially supported by NSF under grants CCF-1011903 and CCF-1018509, and by AFOSR grant FA95501010291. D. Jakovetić holds a fellowship from FCT.

D. Jakovetić is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal, and with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA djakovet@andrew.cmu.edu

J. Xavier is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal jxavier@isr.ist.utl.pt

J. M. F. Moura is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA moura@ece.cmu.edu

$(\cdot)^\top$ the transpose; $\|\cdot\| = \|\cdot\|_2$ the Euclidean (respectively, spectral) norm of its vector (respectively, matrix) argument ($\|\cdot\|$ also denotes the modulus of a scalar); $\nabla \mathcal{J}(x)$ the gradient evaluated at x of a function $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$, $d \geq 1$. Finally, for two positive sequences η_n and χ_n, $\eta_n = O(\chi_n)$ means that $\limsup_{n\to\infty} \frac{\eta_n}{\chi_n} < \infty$; $\eta_n = \Omega(\chi_n)$ means that $\liminf_{n\to\infty} \frac{\eta_n}{\chi_n} > 0$; and $\eta_n = \Theta(\chi_n)$ means that $\eta_n = O(\chi_n)$ and $\eta_n = \Omega(\chi_n)$.

II. Model and Distributed Gradient Algorithm

A. Optimization and network models

Optimization model assumes that N agents solve:

$$\text{minimize} \quad f(x) := \textstyle\sum_{i=1}^N f_i(x). \tag{1}$$

The function $f_i : \mathbb{R}^d \to \mathbb{R}$ is known only by agent i. We impose the following structure on the f_i's.

Assumption 1 (Optimization model) (a) For all i, $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex, and problem (1) is solvable.

(b) For all i, function f_i has Lipschitz continuous first derivative with constant $L \in (0, +\infty)$, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$, $\forall x, y \in \mathbb{R}^d$.

(c) For all i, gradient $\nabla f_i(x)$ is uniformly bounded by constant $G \in [0, \infty)$, i.e., $\|\nabla f_i(x)\| \leq G, \;\; \forall x \in \mathbb{R}^d$.

From Assumption 1 (b), it follows that, e.g., [7]:

$$f_i(z) \leq f_i(x) + \nabla f_i(x)^\top (z - x) + \frac{L\|z - x\|^2}{2}, \forall x, z \in \mathbb{R}^d. \tag{2}$$

Also, under Assumption 1, function $f(x) = \sum_{i=1}^N f_i(x)$ is also convex, and has Lipschitz continuous gradient with constant NL. Denote by $x^\star$ a solution to (3), and by $f^\star = \inf_{x \in \mathbb{R}^d} f(x) = f(x^\star)$ the optimal value. For notational simplicity, from now on we set $d = 1$ so that the optimization variable is scalar, but all our results hold for a generic d.

Communication model. We associate with problem (1) a network $\mathcal{V}$ of N agents, described by the graph $\mathcal{G} = (\mathcal{V}, E)$, where $E \subset \mathcal{V} \times \mathcal{V}$ is the set of links.

Assumption 2 (Network model) The graph $\mathcal{G}$ is connected, undirected, and simple (no self/multiple links.)

We also associate to graph $\mathcal{G}$ a symmetric, stochastic (rows sum to one and all the entries are non-negative), $N \times N$ weight matrix W, with, for $i \neq j$, $W_{ij} > 0$ if and only if, $\{i, j\} \in E$, and $W_{ii} = 1 - \sum_{j \neq i} W_{ij}$. Denote also $\widetilde{W} := W - J$. We require that $\mu := \|\widetilde{W}\| < 1$, which is, for a connected $\mathcal{G}$, true for any W with strictly positive diagonal entries W_{ii}'s, $\forall i$; a popular choice are the Metropolis weights, e.g., [8].

B. Distributed Gradient Algorithm in [4]

We now review the distributed (sub)gradient algorithm in [4]. Each agent i updates over iterations k its estimate $x_i(k)$ of a solution. With the initialization $x_i(0) \in \mathbb{R}$, the update rule at agent i is, for $k = 0, 1, ...$:

$$x_i(k+1) = \sum_{j \in O_i} W_{ij} x_j(k) - \alpha_k \nabla f_i(x_i(k)), \tag{3}$$

where α_k is the step-size, specified further ahead, and O_i is the neighborhood set of agent i (including i). The algorithm operation is as follows. Each agent i, at each iteration k, broadcasts its variable $x_i(k)$ to all its neighbors $j \in O_i - \{i\}$, receives $x_j(k)$ from all its neighbors $j \in O_i - \{i\}$, and 3) updates $x_i(k)$ via (3). To avoid notational clutter in the subsequent analysis, we let $x_i(0) = 0$, for all i, i.e., all agents initialize their estimates to zero.

Matrix form. We now write algorithm (3) in matrix form. Introduce the map $F : \mathbb{R}^N \mapsto \mathbb{R}^N$ as:

$$F(x) = F(x_1, x_2, ..., x_N) = (f_1(x_1), f_2(x_2), ..., f_N(x_N))^\top.$$

Then, (3) in matrix form is:

$$x(k+1) \;\; = \;\; Wx(k) - \alpha_k \nabla F(x(k)), \;\; k = 0, 1, ...(4)$$

Step size. We consider a family of step-sizes: $\alpha_k = \frac{c}{(k+1)^\tau}$,$k = 0, 1, ...$, where $c > 0$ is a constant. The nonnegative parameter $\tau \in [0, 1]$ is a degree of freedom that we later optimize for the best convergence rate.

Local weighted running averages. We derive the convergence rate with respect to the local weighted running averages, defined by: $x_{i,\mathrm{ra}}(k) = \left(\sum_{t=0}^{k-1} \alpha_t\right)^{-1} \sum_{t=0}^{k-1} \alpha_t x_i(t)$, for $k = 1, ...,$ and $x_{i,\mathrm{ra}}(0) = 0$. Note that the quantity $x_{i,\mathrm{ra}}(k)$ is locally available to agent i and is, with $x_{i,\mathrm{ra}}(0) = 0$, efficiently calculated recursively for $k = 1, ...$ by:

$$\begin{aligned} s_i(k+1) &= s_i(k) + \alpha_k, \;, k = 0, 1, ..., \; s_i(0) = 0 \\ x_{i,\mathrm{ra}}(k+1) &= \frac{s_i(k) x_{i,\mathrm{ra}}(k) + \alpha_k x_i(k)}{s_i(k+1)}. \end{aligned}$$

III. Convergence rate: Statement of the result

We now state our main result on the convergence rate of the distributed algorithm (3) under Assumptions 1 and 2.

Theorem 1 Consider algorithm (3) under Assumptions 1 and 2. Set the step-size to $\alpha_k = \frac{c}{(k+1)^\tau}$, with $c = 1/(2L)$ and $\tau = 1/3$. Then, at every agent i: $f\left(x_{i,\mathrm{ra}}(k)\right) - f^\star = O\left(1/k^{2/3}\right)$, where $x_{i,\mathrm{ra}}(k) = \left(\sum_{t=0}^{k-1} \alpha_t\right)^{-1} \sum_{t=0}^{k-1} \alpha_t x_i(t)$.

Theorem 1 establishes the rate $O\left(\frac{1}{k^{2/3}}\right)$ for arbitrary functions f_i's that obey Assumption 1. Hence, compared with the results in [2], we can see that adding smoothness to the f_i's (as in Assumption 1 (b)) improves the rate from $O\left(\frac{\log k}{k^{1/2}}\right)$ to $O\left(\frac{1}{k^{2/3}}\right)$. We also note that an accelerated distributed gradient method, different from (3), proposed in [6], achieves a faster rate $O\left(\frac{\log k}{k}\right)$ under Assumptions 1–2, [6].

Note that we choose the step-size parameters $c = 1/(2L)$ and $\tau = 1/3$. As can be seen from the proofs, Theorem 1 still holds if we take any $c \in (0, 1/(2L))$. We will also see that $\tau = 1/3$ is optimal, in the sense that it minimizes an upper bound on the optimality gap (See the proof of Theorem 1 and equation (9).)

Finally, we note that the obtained bound $O\left(\frac{1}{k^{2/3}}\right)$ is essentially tight, i.e., algorithm (3) cannot achieve a rate faster than $O\left(\frac{1}{k^{2/3}}\right)$ under Assumptions 1–2. We can show this by creating a certain worst-case lower bound on the optimality gap at $x_i(k)$, which turns out to be $\Omega\left(\frac{1}{k^{2/3}}\right)$. This is considered in [6].

IV. Convergence rate: Proof

Our convergence rate analysis is based on the evolution of the (hypothetical) global average $\overline{x}_i(k) = \frac{1}{N}\sum_{i=1}^N x_i(k)$ of the agents' estimates. Subsection IV-A studies the optimality gap at $\overline{x}(k)$ through an inexact oracle framework that we introduce. Subsection IV-B upper bounds the agents' disagreement – how far are the $x_i(k)$'s from $\overline{x}(k)$. Finally, Subsection IV-C proves Theorem 1 by combining the results from Subsections IV-A and IV-B.

A. Inexact oracle framework

We base our analysis on the framework of inexact first order oracle, and on the analysis of gradient methods under first order oracle. We first need the following definition from [6] and that is a variation of Definition 1 in [9].

Definition 2 (Pointwise inexact first order oracle) Consider a function $\phi : \mathbb{R} \to \mathbb{R}$ that is convex and has Lipschitz continuous gradient with constant L_ϕ. We say that a pair $\left(\widetilde{\phi}_x, \widetilde{g}_x\right) \in \mathbb{R} \times \mathbb{R}$ is a (L_x, δ_x) inexact oracle of ϕ at x if:

$$\begin{aligned} \phi(z) &\geq \widetilde{\phi}_x + \widetilde{g}_x^\top (z-x), \quad \forall z \in \mathbb{R} \qquad (5)\\ \phi(z) &\leq \widetilde{\phi}_x + \widetilde{g}_x^\top (z-x) + \frac{L_x}{2}\|z-x\|^2 + \delta_x, \ \forall z \in \mathbb{R}^d. \end{aligned}$$

Note that the pair $(\phi(x), \nabla\phi(x))$ satisfies Definition 2 with $(L_x = L_\phi, \delta_y = 0)$. Also, if $\left(\widetilde{\phi}_x, \widetilde{g}_x\right)$ is a (L_x, δ_x) inexact oracle at x, then it is also a (L_x', δ_x) local inexact oracle at x, with $L_x' \geq L_x$. We will use the following Lemma on the convergence of the (centralized) gradient method under inexact oracle. The Lemma easily follows from Theorem 2 and the result in equation (33) in [9].

Lemma 3 Consider the gradient algorithm for a solvable problem of unconstrained minimization of a convex function $\phi : \mathbb{R} \mapsto \mathbb{R}$: $y(k+1) = y(k) - \beta_k h_k$, $k = 0, 1, ...$, where $y(0) \in \mathbb{R}$, $\beta_k = 1/R_k$, and (ϕ_k, h_k) is a (R_k, δ_k) inexact first order oracle of ϕ at $y(k)$. Then, for a solution $y^\star$:

$$\phi(y_{\mathrm{ra}}(k)) - \phi(y^\star) \leq \frac{\frac{1}{2}\|y(0) - y^\star\|^2 + \sum_{t=0}^{k-1}\beta_t \delta_t}{\sum_{t=0}^{k-1}\beta_t},$$

where $y_{\mathrm{ra}}(k) = \frac{1}{\sum_{t=0}^{k-1}\beta_t}\sum_{t=0}^{k-1}\beta_t y(t)$.

Casting algorithm (3) **in the inexact oracle framework**. Consider algorithm (3) and denote by $\overline{x}(k) := \frac{1}{N}\sum_{i=1}^N x_i(k)$ and $\overline{x}_{\mathrm{ra}}(k) := \frac{1}{N}\sum_{i=1}^N x_{i,\mathrm{ra}}(k)$. Note that the latter quantities are not available to any agent, but only serve for convergence analysis. Then, multiplying (4) from the left by $(1/N)\mathbf{1}^\top$, we get that $\overline{x}(k)$ evolves as:

$$\overline{x}(k+1) = \overline{x}(k) - \frac{\alpha_k}{N}\sum_{i=1}^N \nabla f_i(x_i(k)), \ k = 0, 1, ..., \quad (6)$$

with $\overline{x}(0) = 0$. Denote by $\widetilde{x}_i(k) = x_i(k) - \overline{x}(k)$, and $\widetilde{x}(k) = (\widetilde{x}_1(k), ..., \widetilde{x}_N(k))^\top$. We refer to $\widetilde{x}(k)$ as the disagreement vector, as it says how much the estimates at different agents mutually differ. We now show that (6) is a gradient method under inexact oracle.

Lemma 4 Let Assumption 1 hold, set $\alpha_k = 1/(2L(k+1)^\tau)$, $\tau \in [0, 1]$, and consider $\widetilde{f}_k := \sum_{i=1}^N \left\{ f_i(x_i(k)) + \nabla f_i(x_i(k))^\top (\overline{x}(k) - x_i(k)) \right\}$, $\widetilde{g}_k := \sum_{i=1}^N \nabla f_i(x_i(k))$. Then, $(\widetilde{f}_k, \widetilde{g}_k)$ is a (L_k, δ_k) inexact oracle of $f = \sum_{i=1}^N f_i$ at point $\overline{x}(k)$ with constants $L_k = \frac{N}{\alpha_k}$ and $\delta_k = L\|\widetilde{x}(k)\|^2$.

For a proof, see the proof of Lemma 3 in [6]. Revisiting algorithm (6), we see that it is the gradient method to minimize f under inexact oracle with the amount of "inexactness" specified by Lemma 4. Hence, we can apply Lemma 3 to (6) to derive the optimality gap at $\overline{x}_{\mathrm{ra}}(k)$. In order to "close" the analysis, we need to find δ_k, i.e., we need to upper bound the disagreement quantity $\|\overline{x}(k)\|$.

B. Disagreement estimate

We have the following Lemma that upper bounds the disagreement estimate.

Lemma 5 Consider algorithm (3) under Assumptions 1 and 2 and the step-size $\alpha_k = \frac{1}{(2L)(k+1)^\tau}$, $\tau \in [0, 1]$, $k = 0, 1, ...$ Then, for $k = 1, 2, ...$:

$$\|\widetilde{x}(k)\| \leq \frac{\sqrt{N}GC}{2L\,k^\tau},$$

where $C = C(W) \in [0, \infty)$ depends only on matrix W.

From Lemma 5, we can see that, as long as $\tau > 0$, the agents reach agreement asymptotically, i.e., $\|\widetilde{x}(k)\| \to 0$ as $k \to \infty$. We now prove Lemma 5.

Proof: [Proof of Lemma 5] We first derive the recursive equation for $\widetilde{x}(k)$. Recall $\widetilde{W} = W - J$, and note that $\widetilde{x}(k) = (I - J)x(k)$. Now, multiplying (4) from the left by $(I - J)$, and using $(I - J)W = \widetilde{W}(I - J)$, obtain:

$$\widetilde{x}(k+1) = \widetilde{W}\widetilde{x}(k) + \frac{1}{(k+1)^\tau}u(k), \qquad (7)$$

for $k = 0, 1, ...$, and $\widetilde{x}(0) = 0$, where $u(k) := -\frac{1}{2L}(I - J)\nabla F(x(k))$. Note that, by Assumption 1 (c), and by the sub-multiplicative property of norms, $\|u(k)\| \leq \frac{1}{2L}\|I - J\|\|\nabla F(x(k))\| \leq \frac{\sqrt{N}G}{2L}$. Next, by unwinding the recursion (7), we have, for $k \geq 1$: $\widetilde{x}(k) = \sum_{t=0}^{k-1}\widetilde{W}^t \frac{u(k-1-t)}{(k-t)^\tau}$. Now, using the sub-multiplicative and sub-additive properties of norms, and recalling $\mu := \|\widetilde{W}\|$:

$$\|\widetilde{x}(k)\| \leq \frac{\sqrt{N}G}{2L}\sum_{t=0}^{k-1}\mu^t \frac{1}{(k-t)^\tau}. \qquad (8)$$

We next upper bound the sum $\sum_{t=0}^{k-1}\mu^t \frac{1}{(k-t)^\tau} = \frac{1}{k^\tau} + \frac{\mu}{(k-1)^\tau} + ... + \frac{\mu^{k-1}}{1^\tau}$ by splitting it into two parts – one with t running from $t = 0$ to $t = \lfloor k/2 \rfloor$, and the other with t running from $t = \lfloor k/2 \rfloor + 1$ to $t = k$. (Here $\lfloor a \rfloor$ is the integer part of a.) For the first sum: $\sum_{t=0}^{\lceil k/2 \rceil} \frac{\mu^t}{(k-t)^\tau} \leq \frac{(1+\mu+\mu^2+...+\mu^{\lfloor k/2 \rfloor})}{(k/2)^\tau} \leq \frac{2}{1-\mu} \frac{1}{k^\tau}$. For the second sum: $\sum_{t=\lceil k/2 \rceil+1}^{k} \frac{\mu^t}{(k-t)^\tau} \leq \mu^{k/2}(\frac{1}{1^\tau} + ... + \frac{1}{(k-\lceil k/2 \rceil+1)^\tau}) \leq \mu^{k/2} k^{1-\tau} \leq \left(\sup_{k \geq 1} k \mu^{k/2}\right) \frac{1}{k^\tau}$. Note that, as $\mu \in [0,1)$, $B(\mu) := \sup_{k \geq 1} k \mu^{k/2}$ is finite. Combining the bounds for the first and the second sums, we obtain $\sum_{t=0}^{k-1} \mu^t \frac{1}{(k-t)^\tau} \leq \left(\frac{2}{1-\mu} + B(\mu)\right) \frac{1}{k^\tau} =: C(W) \frac{1}{k^\tau}$, which, combined with (8), gives the desired result. ■

C. Proof of Theorem 1

We now have all the tools needed to prove Theorem 1.

Proof: [Proof of Theorem 1] We first derive the optimality gap at the global running average $\overline{x}_{\mathrm{ra}}(k)$ by applying Lemma 3. Then, we derive the optimality gap at each agent's local running average $x_{i,\mathrm{ra}}(k)$.

Optimality gap at $\overline{x}_{\mathrm{ra}}(k)$. Using Lemma 5, for $k \geq 1$:

$$\delta_k = L \|\widetilde{x}(k)\|^2 \leq \frac{N G^2 C^2}{4 L k^{2\tau}} \leq \frac{N G^2 C^2}{L (k+1)^{2\tau}}.$$

Also, recall that $L_k = \frac{N}{\alpha_k} = 2NL(k+1)^\tau$. Plugging the values of δ_k and L_k into Lemma 3, after some algebra, for $\tau \in [0,1)$ ($\tau = 1$ gives $f(\overline{x}_{\mathrm{ra}}(k)) - f^\star = O(1/\log k)$ and is hence not optimal):

$$\begin{aligned} & f\left(\overline{x}_{\mathrm{ra}}(k)\right) - f^\star \\ \leq \; & \frac{N L \|x^\star\|^2}{\sum_{t=0}^{k-1}(t+1)^{-\tau}} + \frac{N G^2 C^2 \sum_{t=0}^{k-1}(t+1)^{-3\tau}}{L \sum_{t=0}^{k-1}(t+1)^{-\tau}}, \end{aligned}$$

which, using: $k^{1-\tau} - 1 \leq \sum_{t=0}^{k-1} \frac{1}{(t+1)^\tau} \leq k^{1-\tau}$:

$$f\left(\overline{x}_{\mathrm{ra}}(k)\right) - f^\star \leq \frac{N L \|x^\star\|^2}{k^{1-\tau} - 1} + \frac{N G^2 C^2 \, k^{-2\tau}}{L \left(1 - k^{\tau - 1}\right)}, \; k \geq 2. \quad (9)$$

From now on, we make the choice $\tau = \frac{1}{3}$. We justify this choice. First, for $\tau = 1/3$, both summands on the right hand side in (9) are $\Theta\left(\frac{1}{k^{2/3}}\right)$, which means that

$$f\left(\overline{x}_{\mathrm{ra}}(k)\right) - f^\star = O\left(\frac{1}{k^{2/3}}\right). \quad (10)$$

Further, if we choose $\tau = 1/3 - \delta$, $\delta \in (0, 1/3]$, then, the second summand on the right hand side of (9) is $\Theta\left(\frac{1}{k^{2/3 - 2\delta}}\right)$; on the other hand, if we choose $\tau = 1/3 + \delta$, $\delta \in (0, 2/3)$, then the first summand on the right hand side of (9) is $\Theta\left(\frac{1}{k^{2/3-\delta}}\right)$. In summary, the rate as good as $O\left(\frac{1}{k^{2/3}}\right)$ is guaranteed for $\tau = 1/3$.

Optimality gap at $x_{i,\mathrm{ra}}(k)$. For arbitrary agent i:

$$\begin{aligned} f(x_{i,\mathrm{ra}}(k)) - f^\star &= \left(f(x_{i,\mathrm{ra}}(k)) - f(\overline{x}_{\mathrm{ra}}(k))\right) \quad (11) \\ &+ \left(f(\overline{x}_{\mathrm{ra}}(k)) - f^\star\right). \end{aligned}$$

We proceed with upper bounding $f(x_{i,\mathrm{ra}}(k)) - f(\overline{x}_{\mathrm{ra}}(k))$. By property (2):

$$\begin{aligned} & f(x_{i,\mathrm{ra}}(k)) - f(\overline{x}_{\mathrm{ra}}(k)) \leq \|\nabla f(\overline{x}_{\mathrm{ra}}(k))\| \quad (12) \\ \times \; & \|x_{i,\mathrm{ra}}(k) - \overline{x}_{\mathrm{ra}}(k)\| + \frac{N L \|x_{i,\mathrm{ra}}(k) - \overline{x}_{\mathrm{ra}}(k)\|^2}{2}. \end{aligned}$$

We upper bound $\|\nabla f(\overline{x}_{\mathrm{ra}}(k))\|$. By the property of convex functions with Lipschitz continuous derivative of constant L_ϕ and a minimizer $y^\star$, [7]: $\|\nabla \phi(y)\|^2 \leq 2 L_\phi (\phi(y) - \phi(y^\star))$, for all $y \in \mathbb{R}$, and using the result that we already showed $f\left(\overline{x}_{\mathrm{ra}}(k)\right) - f^\star \leq \frac{C_0}{k^{2/3}}$, for some constant $C_0 \in (0, \infty)$:

$$\|\nabla f(\overline{x}_{\mathrm{ra}}(k))\| \leq \frac{\sqrt{2NLC_0}}{k^{1/3}} = O\left(\frac{1}{k^{1/3}}\right). \quad (13)$$

It remains to upper bound $\|x_{i,\mathrm{ra}}(k) - \overline{x}_{\mathrm{ra}}(k)\|$. By the convexity of $\| \cdot \|$, Lemma 5, and $\alpha_t = 1/(2L(t+1)^{1/3})$:

$$\begin{aligned} \|x_{i,\mathrm{ra}}(k) - \overline{x}_{\mathrm{ra}}(k)\| &= \left\| \sum_{t=0}^{k-1} \frac{\alpha_t}{\sum_{t=0}^{k-1} \alpha_t} (x_i(t) - \overline{x}(t)) \right\| \\ \leq & \sum_{t=0}^{k-1} \frac{\alpha_t}{\sum_{t=0}^{k-1} \alpha_t} \|x_i(t) - \overline{x}(t))\| \\ \leq & \frac{\sqrt{N} G C}{2L} \frac{\sum_{t=0}^{k-1}(t+1)^{-2/3}}{\sum_{t=0}^{k-1}(t+1)^{-1/3}} \leq \frac{\sqrt{N} G C}{2L} \frac{k^{1/3}}{k^{2/3} - 1} \\ = & O\left(\frac{1}{k^{1/3}}\right). \quad (14) \end{aligned}$$

Finally, combining (12)–(14), we obtain that $f(x_{i,\mathrm{ra}}(k)) - f(\overline{x}_{\mathrm{ra}}(k)) = O\left(\frac{1}{k^{2/3}}\right)$. The latter, combined with (10) and (11), completes the proof of the Theorem. ■

V. Conclusion

We considered distributed optimization in networks where N agents minimize the sum $\sum_{i=1}^N f_i(x)$ of their local costs. We analyzed the convergence rate of the distributed (sub)gradient algorithm proposed in [4], under the class $\mathcal{F}$ of convex f_i's with Lipschitz continuous and bounded gradients. We established that the algorithm achieves the convergence rate $O(1/k^{2/3})$ under the class $\mathcal{F}$, and that the obtained convergence rate bound $O(1/k^{2/3})$ is tight.

References

[1] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *IPSN 2004, 3rd International Symposium on Information Processing in Sensor Networks*, Berkeley, California, USA, April 2004, pp. 20 – 27.

[2] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, March 2012.

[3] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Contr.*, vol. 31, no. 9, pp. 803–812, Sep. 1986.

[4] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, January 2009.

[5] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Distributed Nesterov-like gradient algorithms," in *to appear in proc. CDC'12, 51st IEEE Conference on Decision and Control*, Maui, Hawaii, December 2012.

[6] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Fast distributed gradient methods," November 2011, available at: http://arxiv.org/pdf/1112.2972.pdf.

[7] L. Vandenberghe, "Optimization methods for large-scale systems," 2010, Lecture Notes, available at: http://www.ee.ucla.edu/ vandenbe/ ee236c.html.

[8] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *IPSN '05, Information Processing in Sensor Networks*, Los Angeles, CA, April 2005, pp. 63–70.

[9] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *submitted to Mathematical Programming*, 2011, available at: http://www.optimization-online.org/DB_FILE/2010/12/2865.pdf.