

# An Accelerated Second-Order Method for Distributed Stochastic Optimization

Artem Agafonov, Pavel Dvurechensky, Gesualdo Scutari, Alexander Gasnikov,  
Dmitry Kamzolov, Aleksandr Lukashevich, and Amir Daneshmand

**Abstract**—We consider distributed stochastic optimization problems that are solved with master/workers computation architecture. Statistical arguments allow to exploit statistical similarity and approximate this problem by a finite-sum problem, for which we propose an inexact accelerated cubic-regularized Newton’s method that achieves lower communication complexity bound for this setting and improves upon existing upper bound. We further exploit this algorithm to obtain convergence rate bounds for the original stochastic optimization problem and compare our bounds with the existing bounds in several regimes when the goal is to minimize the number of communication rounds and increase the parallelization by increasing the number of workers.

**Index Terms**—stochastic optimization, statistical similarity, distributed optimization, stochastic optimization, statistical similarity, distributed optimizations

## I. INTRODUCTION

Distributed optimization lies on the interface between control and optimization with the goal being to find a minimum of some global objective by a network of agents, each of which has access to a local part of the objective and can interact with only neighbouring agents. Many algorithms for this setting under different assumptions were proposed as early as in 1970s [1], [2], [3]. Moreover this setting has many applications including robotics, resource allocation, power system control, control of drone or satellite networks, distributed statistical inference and multiagent reinforcement learning [4], [5], [6], [7], [8]. An important recent application is training large-scale machine learning models, which in the language of optimization requires to solve distributed stochastic optimization problems.

This paper focuses on distributed stochastic optimization problems using master/workers architectures. These computational architectures are common, e.g., in the context of federated learning [9], [10], where for privacy-preserving purposes the dataset is split across multiple workers and computations are coordinated by the master node. To be more precise, we consider the following general stochastic

optimization problem:

$$\min_{x \in \mathbb{R}^d} \mathbf{F}(x) := \mathbb{E}_\xi f(x, \xi), \quad (1)$$

where  $\xi$  is a random variable, e.g. random data,  $f$  is convex and sufficiently smooth, which implies that  $\mathbf{F}$  is convex. We assume that we have access to  $m$  workers,  $T$  rounds of communications (all to all or to the master node), and a total fixed budget of  $N$  realizations of  $\xi$ . Under this assumption the main question is how small we can make the error  $\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*)$  by different algorithms returning a random point  $x^T$ . Here  $x^*$  denotes a solution to (1).

To solve (1) on master/workers architectures, two main approaches are used [11], [12], [13], namely Stochastic Approximation (SA) and Sample Average Approximation (SAA), a.k.a. Monte-Carlo. The division between SA and SAA approaches is made for simplicity and there are algorithms (see [14] and Appendix A) which are based on the SAA idea, but, in fact, use a small number of stochastic gradient for each realization of  $\xi$ , which makes them quite close to SA algorithms. Moreover, in most cases, we are given a dataset and it is our choice whether we see each example once as in the SA approach or multiple times as in the SAA approach. In this paper, we make an attempt to look at SA and SAA approaches from a unified perspective of the actual goal being to solve the stochastic optimization problem (1) with some fixed budget  $N$  of realizations of random variable  $\xi$ .

1) *Stochastic Approximation*: In the SA approach a typical situation is that the total budget of  $N$  realizations of  $\xi$  is distributed between  $T$  communication rounds and  $m$  workers. This leads to so-called intermittent communications with  $n = N/(mT)$  local stochastic gradient steps by each worker in between communication rounds, meaning that between two consecutive communication rounds each worker has access to  $n$  iid stochastic gradients  $\nabla f(x, \xi)$ . The authors of [15] recently obtained for the setting of smooth optimization the following lower bound:<sup>\*†</sup>

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \gtrsim \frac{1}{(N/m)^2} + \frac{1}{\sqrt{N}} + \min \left\{ \frac{1}{T^2}, \frac{1}{\sqrt{N/m}} \right\}. \quad (2)$$

<sup>\*</sup>If  $f$  is a quadratic the last term can be eliminated [16] and such bound is tight.

<sup>†</sup>Here and below we use  $\gtrsim$  and  $\simeq$  for simplicity to highlight the dependence on our main parameters  $m, N, T$  and omit numerical multiplicative constants, logarithmic factors, and other parameters characterizing the problem, e.g. Lipschitz constants of the objective, its gradient, and Hessian, as well as estimates for the norm of the solution.

A.A. and D.K. are with the Moscow Institute of Physics and Technology, Russia (agafonov.ad@phystech.edu, dkamzolov@yandex.ru). P.D. is with the Weierstrass Institute for Applied Analysis and Stochastics, Germany (pavel.dvurechensky@wias-berlin.de). G.S. and A.D. are with the School of Industrial Engineering, Purdue University, USA (gscutari@purdue.edu, adaneshm@purdue.edu). A.G. is with Moscow Institute of Physics and Technology, Russia, Institute for Information Transmission Problems RAS, Russia, Higher school of economics, Russia (gasnikov@yandex.ru). A.L. is with the Skolkovo Institute of Science and Technology, Russia, Russia (aleksandr.lukashevich@skoltech.ru).

They show also that this bound is tight by showing that a special combination of Minibatched Accelerated SGD and Single-Machine Accelerated SGD achieve the bound (2). Under more restrictive assumptions on the smoothness of  $f$  they obtain a lower bound which gives some room for improvement in the last term of (2):

$$\min \left\{ \frac{1}{T^2}, \frac{1}{\sqrt{N/m}}, \frac{1}{(N/m)^{1/4} T^{7/4}} \right\} \quad (3)$$

if Hessian is Lipschitz,

$$\min \left\{ \frac{1}{T^2}, \frac{1}{\sqrt{N/m}}, \frac{1}{TN/m} \right\} \quad (4)$$

if  $f$  is self-concordant.

$$\min \left\{ \frac{1}{T^2}, \frac{1}{\sqrt{N/m}}, \frac{1}{(N/m)^{1/2} T^{3/2}} \right\} \quad (5)$$

if  $f$  is quasi-self-concordant.

The authors of [17], under an additional assumption of stronger local smoothness of  $f$  around  $x^*$ , propose an algorithm with only one round of communication ( $T = 1$ ) with the following guarantee

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \frac{1}{N/m} + \frac{1}{\sqrt{N}}, \quad (6)$$

which, as we will see below, is similar to the SAA-based approach.

2) *Sample Average Approximation*: The alternative SAA approach [18], [19] is based on sampling in advance  $N$  realizations of random function  $f(x, \xi)$  and approximating the expectation in (1) by a regularized finite-sum

$$\min_{x \in \mathbb{R}^d} F(x) = \frac{1}{N} \sum_{k=1}^N f(x, \xi^k) + \frac{\mu}{2} \|x - x_0\|^2, \quad (7)$$

where  $\|\cdot\|$  is the Euclidean norm. If the regularization parameter  $\mu \simeq 1/\sqrt{N}$ , solving the problem (7) with sufficient accuracy, we obtain the solution of (1) (see Sec. II for details). This motivates developing fast algorithms for problem (7) with the ultimate goal being to obtain solution to the original stochastic optimization problem (1).

For the SAA approach we assume that we have in total  $N$  realizations  $\{\xi^k\}_{k=1}^N$  of the random variable  $\xi$ ,  $T$  communication rounds, and  $m$  workers. Each worker can perform  $n$  local steps between two communication rounds with each step using one gradient  $f(x, \xi^k)$  for a particular realization  $\xi^k$ . The difference with the SA approach is that it is possible to use the same realization  $\xi^k$  in different local steps and also that, in general,  $n \neq N/(mT)$ .

Although the SAA approach allows using each observation multiple times, there are Variance Reduced (VR) methods, which applied to the problem (7), typically use only a logarithmic number of gradients for each observation  $\xi^t$ .

The following convergence rate for the original problem (1) was obtained in [9] by using SVRG algorithm

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( - \min \left\{ \frac{nT}{\sqrt{N}}, \frac{mnT}{N} \right\} \right) + \frac{1}{\sqrt{N}}. \quad (8)$$

Note, that parameters  $n$  and  $T$  appear in the inequality (8) only as a combination  $nT$ . Therefore, the minimum number of communication rounds  $T$  is achieved when  $T = 1$ . Further, from the bound (8) we see that when the number of workers  $m$  increases, there is a limit for possible improvement in the bound. Indeed, if  $m \gtrsim \sqrt{N}$  the minimum in the exponent in (8) is achieved on the first term, and no improvement in the bound is guaranteed. Moreover, the similar limit  $m \simeq \sqrt{N}$  can be obtained via SA-based method from [17] with the guarantee (6).

Accelerated (non distributed) Variance Reduced schemes [20] can not improve the above bound (8). Non-accelerated distributed Variance Reduced method from [21] applied to the special type of the problem (7)

$$\min_{x \in \mathbb{R}^d} F(x) = \frac{1}{m} \sum_{i=1}^m \frac{1}{K} \sum_{j=1}^K f(x, \xi^{i,j}) + \frac{\mu}{2} \|x - x_0\|^2, \quad (9)$$

where  $K = N/m$ , gives only the following bound for an approximate solution to problem (1) (see Appendix A for details):

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( - \min \left\{ \frac{T}{\sqrt{N}}, \frac{mnT}{N} \right\} \right) + \frac{1}{\sqrt{N}}. \quad (10)$$

The RHS of the above inequality consists of two terms. The first one (optimization error) corresponds to the inexact solution of the approximation (9) and the second one (statistical error) comes from statistical reasoning of how well (9) approximates (1). Ideally, optimization error and statistical error should be of the same order. Indeed, if, due to a small budget of communications, the optimization error dominates, the sample size  $N$  should have been chosen smaller. On the other hand, there is no much sense in optimizing below the statistical error. Thus, to minimize  $T$ , we are interested in the regime when optimization and statistical errors are of the same order. This is achieved when  $T \simeq \max\{N^{1/2}, N/(mn)\}$  (recall that  $\simeq$  hides also logarithmic factors). At the same time, we would like to maximize the number of workers to scale up the computations. If  $m$  is too large, i.e.  $m \gtrsim N^{1/2}$ , the minimum in the exponent is achieved at the first term and there is no improvement in convergence rate with the increase of  $m$ . Therefore, the best possible choice is  $m \simeq \sqrt{N}$  (where we set  $n = 1$  to maximize  $m$ ) and  $T \simeq \sqrt{N}$ . In the rest of the paper, we follow the same scheme to estimate a sufficient number of communication rounds and the possible numbers of workers.

Optimal (accelerated) distributed Variance Reduced method from [14] gives

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( - \min \left\{ \frac{T}{N^{1/4}}, \frac{mnT}{N}, \frac{\sqrt{mn}T}{N^{3/4}} \right\} \right) + \frac{1}{\sqrt{N}}. \quad (11)$$

From (11) following the same reasoning as before we derive that to minimize  $T$  and  $m$ , we should choose  $T \simeq N^{1/4}$  and  $m \simeq N$ . Interestingly, the same result is achieved by the standard accelerated gradient method [22] applied to the finite-sum problem with  $N$  terms in the sum.

3) *Exploiting statistical similarity*: Recent advances in distributed optimization for solving problem (7) are achieved by distributing  $N$  realizations of  $f(x, \xi)$  between  $m$  workers each having  $n = N/m$  realizations. Then problem (7) takes the form

$$\min_{x \in \mathbb{R}^d} F(x) = \frac{1}{m} \sum_{k=1}^m f_k(x) + \frac{\mu}{2} \|x - x_0\|^2, \quad (12)$$

where  $\mu \simeq 1/\sqrt{N} = 1/\sqrt{mn}$ ,  $f_k(x) = \frac{1}{n} \sum_{j=1}^n f(x, \xi^{k,j})$ . Using probabilistic arguments statistical similarity is shown between  $f_k$  and the whole sum. More formally,  $\|\nabla^2 f_k(x) - \frac{1}{m} \sum_{i=1}^m \nabla^2 f_i(x)\| \leq \beta$ , where  $\beta \simeq 1/\sqrt{n}$ . This idea have been recently extensively exploited for optimization problems (mainly) over master/workers architectures, under the name of statistical preconditioning [23], [24], [25], [26], [27], [28], [29], [30]. These papers focus on solving the finite-sum problem (7) and most of them do not achieve the lower communication complexity bound for this setting obtained in [31]:<sup>‡</sup>

$$\Omega \left( \sqrt{1 + \frac{\beta}{\mu} \ln \left( \frac{\Delta F(x^0)}{\varepsilon'} \right)} \right) \quad (13)$$

to solve problem (7) with accuracy  $\varepsilon'$ , where  $\Delta F(x^0) := F(x^0) - F(x_F^*)$  and  $x_F^*$  is the solution to this problem. The authors of [32] propose a distributed implementation of the damped Newton Method called DISCO in the master/workers architecture for minimizing  $M$ -self-concordant functions and achieve the complexity bound

$$O \left( \left( M^2 \Delta F(x^0) + \ln \frac{1}{\varepsilon'} \right) \sqrt{1 + \frac{\beta}{\mu}} \right). \quad (14)$$

Unlike [32], we propose to reach (13) by using cubic-regularized Newton's method at the central node. Moreover, most of the above statistical preconditioning papers focus on (7) and do not account for the actual goal of solving the original stochastic optimization problem (1). Under an appropriate choice of the parameter  $\beta \simeq 1/\sqrt{n} = 1/\sqrt{N/m}$ , when combined with statistical reasoning, the best result in the literature for (1) corresponds to the guarantee

$$\mathbb{E} \mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( -\frac{T}{m^{1/(2\kappa(\mathbf{F}))}} \right) + \frac{1}{\sqrt{N}}, \quad (15)$$

where  $\kappa(\mathbf{F}) \in [1, 2]$  and, in general [27],  $\kappa(\mathbf{F}) \simeq 1$ . Moreover, the methods achieving this bound [28], [29], [30] require to solve rather difficult auxiliary problems at each node.

<sup>‡</sup>Here and below we use  $O(\cdot)$ ,  $\Omega(\cdot)$  notation to denote bounds which hold up to constant factors, and  $\tilde{O}(\cdot)$ ,  $\tilde{\Omega}(\cdot)$  notation to denote bounds which hold up to constant and polylogarithmic factors.

In [19] the latter drawback is overcome by applying *non-accelerated* cubic regularized Newton step [33] at each node in order to solve (7), which results in the bound

$$\mathbb{E} \mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( -\min \left\{ \frac{T}{N^{1/4}}, \frac{T}{m^{1/2}} \right\} \right) + \frac{1}{\sqrt{N}} \quad (16)$$

for the problem (1).

#### A. Our contribution

The main contribution of this paper is two-fold. First, we focus on finite-sum problems under statistical similarity and propose a master/workers distributed algorithm for such problems. The main idea is to implement inexact accelerated cubic regularized Newton's method [34], [35], [36] at the master node for functions with  $L$ -Lipschitz Hessian, which allows to obtain communication complexity bound

$$O \left( \sqrt{\frac{\beta}{\mu} \ln \frac{1}{\varepsilon'}} + \left( \frac{L^2 \Delta F(x_0)}{\mu^3} \right)^{1/6} \right) \quad (17)$$

that is better than the bound (14) in [32] since  $M = L/\mu^{3/2}$ , and matches the dependence on  $\beta$  and  $\mu$  in the lower bound (13). Since the size of the message between nodes remains  $\mathcal{O}(d)$  as for first-order methods, our approach allows to reduce communication complexity without additional communication overhead compared to first-order methods.

Second, we apply this method in order to solve the original stochastic optimization problem (1) and obtain an algorithm that converges according to the following bound

$$\mathbb{E} \mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( -\min \left\{ \frac{T}{N^{1/6}}, \frac{T}{m^{1/4}} \right\} \right) + \frac{1}{\sqrt{N}}. \quad (18)$$

Under additional assumption of  $\mu$ -strong convexity of the original problem (1), the proposed algorithm provides the bound

$$\mathbb{E} \mathbf{F}(x^T) - \mathbf{F}(x^*) \lesssim \exp \left( -\min \left\{ T\mu^{1/3}, T\frac{\mu^{1/2}N^{1/4}}{m^{1/4}} \right\} \right) + \frac{1}{\mu N}.$$

In Table I we summarize a comparison with the related works described above. Note, that in our work we are motivated by two goals: minimize the number of communications  $T$  and maximize the number of workers  $m$  to achieve better parallelization possibilities.

## II. FINITE-SUM APPROXIMATION FOR STOCHASTIC OPTIMIZATION PROBLEM

In this section, we state the main assumptions about stochastic optimization problem (1), discuss its finite-sum approximation obtained via the SAA approach, as well as introduce and motivate the concept of statistical similarity for finite-sum optimization problems.

We consider stochastic minimization problem

$$\min_{x \in \mathbb{R}^d} \mathbf{F}(x) := \mathbb{E}_{\xi} f(x, \xi), \quad (19)$$

TABLE I  
COMPARISON BETWEEN DIFFERENT METHODS AND BOUNDS FOR  
PROBLEM (1)

	Bound	$T$	$m$
SA lower bound [15]	(2)	$N^{1/4}$	$N^{3/4}$
SGD [17]	(6)	1	$N^{1/2}$
SVRG [9]	(8)	1	$N^{1/2}$
Non-accelerated VR [21]	(10)	$N^{1/2}$	$N^{1/2}$
Accelerated VR [14]	(11)	$N^{1/4}$	$N$
Cubic Newton [19]	(16)	$N^{1/4}$	$N^{1/2}$
Accelerated Cubic Newton [this paper]	(18)	$N^{1/6}$	$N^{2/3}$

where  $\xi$  is a random variable,  $f(x, \xi)$  is convex function w.r.t.  $x \in \mathbb{R}^d$  for all  $\xi$ . We assume that, for all  $\xi$ ,  $f(x, \xi)$  is  $L_0$ -Lipschitz continuous, i.e.,

$$\|f(x, \xi) - f(y, \xi)\| \leq L_0 \|x - y\| \quad \forall x, y \in \mathbb{R}^d$$

and has  $L$ -Lipschitz Hessian, i.e.

$$\|\nabla^2 f(x, \xi) - \nabla^2 f(y, \xi)\| \leq L \|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

Note that the first of these assumptions implies that  $\mathbf{F}$  is  $L_0$ -Lipschitz continuous. We will also consider the case when  $f(x, \xi)$  is  $\mu$ -strongly convex for all  $\xi$  as functions of  $x$ .

#### A. Finite-sum approximation

To solve problem (19) we apply the SAA approach, i.e. sample  $N$  iid realizations  $\xi^{k,j}$  from an unknown distribution of  $\xi$ , for each worker node  $k = 1, \dots, m$  define local objective  $f_k(x) = \frac{1}{n} \sum_{j=1}^n f(x, \xi^{k,j})$ , where  $n = N/m$ , and approximate the original stochastic optimization problem (19) by the finite-sum problem

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{m} \sum_{k=1}^m f_k(x) + \frac{\mu}{2} \|x - x_0\|^2. \quad (20)$$

We use  $x^*$ ,  $x_F^*$  to denote the solutions of problems (19) and (20) respectively. We assume that  $x^*$  and  $x_F^*$  lie inside the Euclidean ball with center at  $x_0$  and radius  $R$ .

Corollary 1.2 from [37] claims that under the assumption of  $L_0$ -Lipschitz continuity of  $f(x, \xi)$  w.r.t  $x$  and with  $\mu = \frac{L_0 \log N}{RN}$  the following bound holds with probability at least  $1 - \delta$ :

$$\mathbf{F}(x_F^*) - \mathbf{F}(x^*) \leq O\left(\frac{L_0 R}{\sqrt{N}} \log(N/\delta)\right).$$

Using  $L_0$ -Lipschitz continuity of  $\mathbf{F}(x)$ , we obtain that

$$\mathbf{F}(x) - \mathbf{F}(x_F^*) \leq L_0 \|x - x_F^*\|, \quad \forall x.$$

Combining the above two inequalities and plugging  $x = x^T$ , where  $x^T$  is an output of some optimization algorithm after  $T$  communication rounds, we obtain

$$\mathbf{F}(x^T) - \mathbf{F}(x^*) \leq L_0 \|x^T - x_F^*\| + O\left(\frac{L_0 R}{\sqrt{N}} \log(N/\delta)\right). \quad (21)$$

Thus, if we find a good approximation  $x^T$  to the solution  $x_F^*$  of the finite-sum problem (20), we automatically obtain an approximate solution to problem (19).

If we additionally assume that  $f(x, \xi)$  is  $\mu$ -strongly convex for all  $\xi$ , there is no need in additional regularization and the original stochastic problem (19) can be approximated by

$$\min_x \frac{1}{m} \sum_{k=1}^m f_k(x).$$

In that case the bound (21) can be improved [18] to:

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \leq L_0 \|x^T - x_F^*\| + O\left(\frac{L_0^2}{\mu N}\right). \quad (22)$$

#### B. Statistical similarity

Since problem (20) originates from problem (19), we can state and utilize one more important property of the objective function in (20), namely, statistical similarity. Under assumption that the observations  $\xi^{k,j}$  are iid, the following bound holds [38] for all  $k = 1, \dots, m$  with probability at least  $1 - \delta$ :

$$\sup_w \left\| \frac{1}{m} \sum_{j=1}^m \nabla^2 f_j(w) - \nabla^2 f_k(w) \right\| \leq \tilde{O}\left(\sqrt{\frac{32L^2 d}{n}}\right). \quad (23)$$

In the next section we utilize statistical similarity to propose an efficient distributed algorithm for problem (20).

### III. ACHIEVING THE LOWER BOUND FOR FINITE-SUM OPTIMIZATION UNDER STATISTICAL SIMILARITY

Motivated by the connection between the finite sum problem (20) and the original stochastic optimization problem (19) stated in the previous section, we propose in this section a distributed minimization algorithm with master/workers architecture for general finite-sum problems, in particular, problem (20). We also show that this algorithm achieves the lower communication complexity bound in [31] specialized for our setting of master/workers architecture. Moreover, our algorithm achieves communication complexity bound that is better than the one in [32].

To that end we consider a network with  $m$  agents and the following general finite-sum optimization problem

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{m} \sum_{k=1}^m f_k(x), \quad (24)$$

where each worker node  $k$  has access only to its local part  $f_k$  of the objective. Note that this problem statement covers problem (20) as a special case since the regularizer can be equally distributed among the agents. We assume that  $F$  is  $\mu$ -strongly convex and has  $L$ -Lipschitz Hessian. Motivated by subsection II-B, we make in this section the following assumption that each local objective  $f_k$  is a good approximation to the global objective  $F$ .

*Assumption 1:* (statistical similarity) Each local function  $f_k$  is  $\beta$  related to the global objective  $F$ :

$$\|\nabla^2 F(x) - \nabla^2 f_k(x)\| \leq \beta, \quad (25)$$

for all  $x \in \mathbb{R}^d$  and some  $\beta \geq 0$ . In particular, for problem (20) we have that  $\beta = \tilde{O}(\sqrt{d/n})$  with high probability.

To solve problem (24) we propose Restarted Distributed Accelerated Cubic Regularized Newton's Method by extending the methods of [34], [35], [36]. First, we describe Distributed Accelerated Cubic Regularized Newton's Method for minimizing convex functions (Algorithm 1). Then we apply restart technique to obtain linearly convergent Algorithm 2 for minimizing  $\mu$ -strongly convex function in (24).

We choose one of the agents (w.l.o.g. the agent with number 1) to be the central node (server), and all the others are  $m - 1$  to be workers (machines) that are assumed to be connected with the central node. Each one of these nodes stores the part  $f_k$  of the global objective, computes gradients, and passes them to the central node. Then, the server forms the gradient of the global objective  $\nabla F(x)$ , computes the Hessian of its local loss  $f_1(x)$ , constructs the following model of the global objective  $F$ :

$$\tilde{F}_M(x, z) = F(z) + \langle \nabla F(z), x - z \rangle \quad (26)$$

$$+ \frac{1}{2} \langle (\nabla^2 f_1(z) + 3\beta I)(x - z), x - z \rangle + \frac{M}{6} \|x - z\|^3, \quad (27)$$

updates variable  $x$  by minimizing this model and broadcasts it to the workers.

Algorithm 1 is a master/workers generalization of the Accelerated Cubic Newton method under inexact second-order information [36]. We use local objective  $f_1(x)$  on the server as an approximation to the global objective  $F$ . Note, that we need to compute  $\nabla^2 f_1(x)$  only once per iteration, at the point  $w_t$ , but at the same time we need two communication rounds per one iteration. Importantly, each communication round requires sending only vectors. Also, our approach allows to vary sample size on the central node. This can lead to the better performance, since it improves the constant  $\beta$  obtained from statistical similarity.

The next Theorem gives convergence rate of Algorithm 1.

*Theorem 1:* Let Assumption 1 hold and  $F$  be convex function with  $L$ -Lipschitz Hessian and defined in (24). Then after  $t$  iterations of Algorithm 1 we have

$$F(x_t) - F(x_F^*) \leq \frac{98L\|x_0 - x_F^*\|^3}{t^3} + \frac{48\beta\|x_0 - x_F^*\|^2}{t^2} \quad (29)$$

where  $x_F^*$  is a solution to (24).

Note that  $T$  iterations correspond to  $2T$  communication rounds since each iteration requires two communication rounds.

*Proof:* To prove the theorem we would like to apply Theorem 11 of [36], which analyzes an accelerated cubic-regularized Newton's method with inexact Hessian. Thus, first we show that in our algorithm the central node indeed runs accelerated cubic-regularized Newton's method with inexact Hessian (Algorithm 2 of [36]). Their algorithm uses an approximation  $H_t$  for the Hessian that satisfies (in their notation  $\mu_u$  instead of  $\lambda$ )

$$\frac{\lambda}{2} I \preceq H_t - \nabla^2 F(w) \preceq \lambda I. \quad (30)$$

---

### Algorithm 1 Accelerated Cubic Newton

---

**Input:**  $x_0 \in \mathbb{R}^d$ ,  $\alpha_0 = 1$ ,  $A_0 = 1$ ,  $L$ ,  $\beta$ .

**Step 0:**

Master node computes  $\nabla F(x_0)$  by collecting  $\nabla f_k(x_0)$  from workers.

$$x_1 = \operatorname{argmin}_{x \in \mathbb{R}^d} \tilde{F}_{4L}(x, x_0),$$

$$y_1 = \operatorname{argmin}_{x \in \mathbb{R}^d} \{\psi_1(x) := F(x_1) + 8\beta\|x - x_0\|^2 + 16L\|x - x_0\|^3\}.$$

Set  $w_0 = x_0$  and  $t = 1$ .

**Step 1:**

Master node computes  $\nabla F(w_t)$  by collecting  $\nabla f_k(w_t)$  from workers.

Set

$$w_t = \left(1 - \frac{3}{t+3}\right) x_t + \frac{3}{t+3} y_t$$

$$x_{t+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \tilde{F}_{4L}(x, w_t).$$

**Step 2:**

Master node computes  $\nabla F(x_{t+1})$  by collecting  $\nabla f_k(x_{t+1})$  from workers.

Define  $A_t = A_{t-1}(1 - 3/(t+3))$ ,

$$\psi_{t+1}(x) := \psi_t(x) + 4\beta\|x - x_0\|^2 \quad (28a)$$

$$+ \frac{3}{A_t(t+3)} (F(x_{t+1}) + \langle \nabla F(x_{t+1}), x - x_{t+1} \rangle). \quad (28b)$$

$$y_{t+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \{\psi_{t+1}(x)\}.$$

**Step 3:** Set  $t \leftarrow t + 1$  and go to step 1.

---

Our assumption of  $\beta$ -similarity (25) allows to choose  $H_t = \nabla^2 f_1(w) + 3\beta I$  to satisfy (30) with  $\lambda = 4\beta$ . Indeed, the equation (25) implies

$$\beta \geq \frac{\|(\nabla^2 F(x) - \nabla^2 f_1(x))x\| \|x\|}{\|x\|^2} \geq \frac{\langle x, (\nabla^2 f_1(x) - \nabla^2 F(x))x \rangle}{\|x\|^2}.$$

Therefore,

$$-\beta I \preceq \nabla^2 f_1(x) - \nabla^2 F(x) \preceq \beta I.$$

Adding  $3\beta I$  to the inequality above we obtain (30) with  $\lambda = 4\beta$ . Thus, we see that the central node using its local Hessian is equivalent to using inexact Hessian of the global objective  $F$ . Since the central node calculates the gradient of the global objective  $F$  by communicating with the nodes, the central node indeed implements cubic steps with inexact Hessian.

Our choice of the algorithm parameters corresponds to the following choice of the parameters in Algorithm 2 of [36] stated in their Theorem 11:

$$\alpha_t = \frac{3}{t+3}, \mu_u = 4\beta, \bar{\mu}_t = 8\beta(t+2), \gamma = L, \beta = 96L,$$

$$\eta = M = 4L.$$

---

**Algorithm 2** Restarted Accelerated Cubic Newton

---

**Input:**  $z_0 \in \mathbb{R}^n$ , strong convexity parameter  $\mu > 0$ , Lipschitz constant for Hessian  $L$ , and  $R_0 > 0$  such that  $\|z_0 - x_F^*\| \leq R_0$ . **For**  $s = 1, 2, \dots$ :

- 1) Set  $x_0 = z_{s-1}$  and  $R_s = \frac{R_0}{2^s}$ .
- 2) Run Algorithm 1 for  $T_s$  iterations, where

$$t_s = 2 \max \left\{ \left( \frac{196LR_{s-1}}{\mu} \right)^{\frac{1}{3}}, 2 \left( \frac{24\beta}{\mu} \right)^{\frac{1}{2}} \right\}. \quad (31)$$

- 3) Set  $z_s = x_{t_s}$ .
- 

Applying Theorem 11 of [36] we obtain the statement of our theorem.  $\blacksquare$

Our next step is to restart Algorithm 1 in order to exploit strong convexity of the objective and obtain linear convergence rate. In each step of Algorithm 2 we run Distributed Accelerated Cubic Regularized Method for the number of iterations, defined in (31). Then, we use its output as the initial point for the next run of Algorithm 1 with reset parameters and so on.

*Theorem 2:* Let the assumptions of Theorem 1 hold and additionally  $F$  be  $\mu$ -strongly convex. Let  $R_0 > 0$  be such that  $\|z_0 - x^*\| \leq R_0$  and  $\{z_s\}_{s \geq 0}$  be generated by Algorithm 2. Then for any  $s \geq 0$  we have

$$\|z_s - x_F^*\| \leq R_0 2^{-s}, \quad (32)$$

$$F(z_s) - F(x_F^*) \leq \mu R_0^2 \cdot 2^{-2s-1} \quad (33)$$

Moreover, the total communication and oracle complexities are

$$O \left( \sqrt{\frac{\beta}{\mu}} \log \frac{F(x_0) - F(x_F^*)}{\varepsilon} + \left( \frac{LR_0}{\mu} \right)^{\frac{1}{3}} \right). \quad (34)$$

*Proof:* We prove by induction that  $\|z_s - x_F^*\| \leq 2^{-s} \|x_0 - x_F^*\| \leq R_s = 2^{-s} R_0$ . For  $s = 0$  this obviously holds. By strong convexity and inequality (29), we obtain that

$$\begin{aligned} \|z_{s+1} - x_F^*\|^2 &\leq \frac{2}{\mu} (F(x_{t_s+1}) - F(x_F^*)) \\ &\leq \frac{2}{\mu} \left( \frac{98L \|z_s - x_F^*\|^3}{t_{s+1}^3} + \frac{48\beta \|z_s - x_F^*\|^2}{t_{s+1}^2} \right) \\ &\leq \frac{2}{\mu} \left( \frac{98LR_s^3}{\left( 2 \left( \frac{196LR_s}{\mu} \right)^{\frac{1}{3}} \right)^3} + \frac{48\beta R_s^2}{\left( 4 \left( \frac{24\beta}{\mu} \right)^{\frac{1}{2}} \right)^2} \right) \leq \frac{R_s^2}{4} = R_{s+1}^2. \end{aligned}$$

Thus, by induction, we obtain that (32), (33) hold.

Next we provide the corresponding complexity bounds. From the above induction bounds, we obtain that after  $S$  restarts the total number of iterations of Algorithm 1, each requiring one call of the second-order oracle, two calls of the first-order oracle and two communication rounds, is no

greater than

$$\begin{aligned} \sum_{s=1}^S t_s &\leq 2 \left( \frac{196LR_0}{\mu} \right)^{\frac{1}{3}} \sum_{s=1}^S 2^{\frac{1-s}{3}} + 4S \left( \frac{24\beta}{\mu} \right)^{\frac{1}{2}} \leq \\ &8 \left( \frac{392LR_0}{\mu} \right)^{\frac{1}{3}} + 4 \sqrt{\frac{24\beta}{\mu}} \log_4 \left[ \frac{F(x_0) - F(x_F^*)}{\varepsilon} \right]. \end{aligned}$$

Therefore, the total communication and oracle complexities are given by (34).  $\blacksquare$

Let us now translate the bound (32) to the language of the number of iterations of Algorithm 1 and the number of communication rounds. Let  $T$  be an even number of communications, which means that we made  $t = T/2$  iterations of Algorithm 1. Let  $\tau_1 = 2 \left( \frac{196LR_0}{\mu} \right)^{\frac{1}{3}}$  and  $\tau_2 = 4 \left( \frac{24\beta}{\mu} \right)^{\frac{1}{2}}$ . Then  $t_s$  in Algorithm 2 satisfies  $t_s \leq \max\{\tau_1, \tau_2\}$  and after  $T$  communication rounds, this algorithm makes  $s \geq \lfloor \frac{T}{2 \max\{\tau_1, \tau_2\}} \rfloor$  restarts and generates a point  $x^T = z_s$  such that

$$\|x^T - x_F^*\| \leq R_0 2^{-\lfloor \frac{T}{2 \max\{\tau_1, \tau_2\}} \rfloor} \leq R_0 2^{-\frac{T}{2 \max\{\tau_1, \tau_2\}}}. \quad (35)$$

At this point it is convenient to compare the complexity bound (34) with the bounds in the literature. Firstly, in terms of the dependence on  $\beta$  and  $\mu$  our algorithm achieves the lower bound (13) obtained in [31]. Secondly, we compare our bound with the bound (14) of the DISCO algorithm [32], which unlike other works [23], [24], [25], [26], [27], [28], [29], [30] also achieves the lower bound in terms of the dependence on  $\beta$  and  $\mu$ . Since the dependence on these parameters in (14) and in our bound (34) are the same, we compare the other parts of the complexity bound.

Let us denote  $\Delta F(x_0) = F(x_0) - F(x_F^*)$ . Using the fact, that a  $\mu$ -strongly-convex function with  $L$ -Lipschitz Hessian is self-concordant with constant  $M = \frac{L}{(2\mu^{3/2})}$ , we can rewrite our bound as

$$\begin{aligned} O \left( \left( \frac{LR_0}{\mu} \right)^{1/3} \right) &\leq O \left( \left( \frac{L}{\mu} \left( \frac{\Delta F(x_0)}{\mu} \right)^{1/2} \right)^{1/3} \right) \\ &= O \left( \left( \frac{L}{\mu^{3/2}} (\Delta F(x_0))^{1/2} \right)^{1/3} \right) = O \left( (M^2 \Delta F(x_0))^{1/6} \right). \end{aligned}$$

The corresponding part of the bound (14) for the DISCO algorithm is much worse:

$$O \left( M^2 \Delta F(x_0) \sqrt{\frac{\beta}{\mu}} \right).$$

#### IV. APPLICATION TO STOCHASTIC OPTIMIZATION PROBLEM

In this section we return back to the stochastic optimization problem (19). As it was described in subsection II-A, if the regularization parameter  $\mu$  in (20) is chosen as  $\mu = \tilde{O} \left( \sqrt{\frac{L_0^2}{mnR^2}} \right)$  (Note that in this case  $N = mn$ ), then an approximate solution to problem (20) is also an approximate solution to the stochastic optimization problem

(19). We apply the algorithm from the previous section to solve problem (20) which satisfies assumptions of Theorem 2. Combining the bound of this theorem with the bound (21), we obtain that the point  $x^T$  generated by Algorithm 2 after  $T$  rounds of communications satisfies

$$\begin{aligned} F(x^T) - F(x^*) &\leq \tilde{O} \left( L_0 R_0 2^{-\frac{T}{2 \max\{\tau_1, \tau_2\}}} + \frac{L_0 R}{\sqrt{N}} \right) \\ &= \tilde{O} \left( L_0 R_0 2^{-\min\{T(\frac{\mu}{L R_0})^{\frac{1}{3}}, T(\frac{\mu}{\beta})^{\frac{1}{2}}\}} + \sqrt{\frac{L_0^2 R^2}{N}} \right). \end{aligned}$$

Further, substituting the value of  $\mu$ , the value of  $\beta$  from (23) (see also Assumption 1), and omitting all the constants except  $N$ ,  $m$ ,  $T$ ,  $n = N/m$ , we obtain

$$F(x^T) - F(x^*) \leq \exp \left( -\min \left\{ \frac{C_1 T}{N^{1/6}}, \frac{C_2 T}{m^{1/4}} \right\} \right) + \frac{C_3}{\sqrt{N}} \quad (36)$$

where constants  $C_1, C_2, C_3$  depend on the parameters  $L_0, L, R_0$  and logarithms of other parameters.

From the bound (36) we can obtain the dependence of the number of communication rounds  $T$  and number of workers  $m$  on the total number of observations  $N$ . We consider the case of full parallelization, i.e. when we use as many workers as possible and perform as less as possible communication rounds. The RHS of (36) consists of two terms, and only the first one, that comes from the solution of the finite-sum approximation (20), depends on  $T$ . So we would like to choose the number of communications such that both terms have the same order. Otherwise, the first term will either be larger than the second one, which means, that we have performed not enough communication rounds, or less, which implies that we have made too many communication rounds, and that does not improve the convergence. Therefore, we get  $T \simeq \max\{N^{1/6}, m^{1/4}\}$ . Recall, that we also would like to maximize  $m$ . If we choose  $m \gtrsim N^{2/3}$ , we will have  $T \simeq m^{1/4}$ . Hence, the number of communication rounds will increase with the number of workers. Therefore, the best possible choice is  $m \simeq N^{2/3}$  and  $T \simeq m^{1/4}$ .

Communication requirements in terms of  $T$  and  $m$  for different approaches to solve (1) are presented in Table I. One can see that our result is better than the lower bound for stochastic optimization (2) in both  $T$  and  $m$ . Compared to other state of the art approaches our method outperforms them either in number of communications or number of workers.

In the case when the original stochastic problem (19) is  $\mu$ -strongly convex, we no longer need to add regularization to have convergence. From (22) and (35) we have

$$\begin{aligned} \mathbb{E}F(x^T) - F(x^*) &\leq \\ \exp \left( -\min \left\{ C_1 T \mu^{1/3}, C_2 T \mu^{1/2} n^{1/4} \right\} \right) &+ \left( \frac{C_3}{\mu N} \right), \end{aligned}$$

where constants  $C_1, C_2, C_3$  depend on the parameters  $L_0, L, R_0$  and logarithms of other parameters.

## REFERENCES

- [1] V. Borkar and P. P. Varaiya, "Asymptotic agreement in distributed estimation," *IEEE Transactions on Automatic Control*, vol. 27, no. 3, pp. 650–655, 1982.
- [2] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 42–50, 1984.
- [3] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.
- [4] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *Journal of Optimization Theory and Applications*, vol. 129, no. 3, pp. 469–488, 2006.
- [5] M. Rabbat and R. Nowak, "Decentralized source localization and tracking wireless sensor networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2004, pp. 921–924.
- [6] S. S. Ram, V. V. Veeravalli, and A. Nedic, "Distributed non-autonomous power control through distributed convex optimization," in *IEEE INFOCOM 2009*. IEEE, 2009, pp. 3001–3005.
- [7] T. Kraska, A. Talwalkar, J. C. Duchi, R. Griffith, M. J. Franklin, and M. I. Jordan, "Mlbase: A distributed machine-learning system." in *CIDR*, vol. 1, 2013, pp. 2–1.
- [8] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.
- [9] B. E. Woodworth, J. Wang, A. Smith, B. McMahan, and N. Srebro, "Graph oracle models, lower bounds, and gaps for parallel stochastic optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 8505–8515.
- [10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.
- [11] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009. [Online]. Available: <https://doi.org/10.1137/070704277>
- [12] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*. Society for Industrial and Applied Mathematics, 2009. [Online]. Available: <http://epubs.siam.org/doi/abs/10.1137/1.9780898718751>
- [13] D. Dvinskikh, "Stochastic approximation versus sample average approximation for population wasserstein barycenters," *Optimization methods and Software*, 2020.
- [14] H. Li, Z. Lin, and Y. Fang, "Optimal accelerated variance reduced extra and digging for strongly convex and smooth decentralized optimization," *arXiv preprint arXiv:2009.04373*, 2020.
- [15] B. Woodworth, B. Bullins, O. Shamir, and N. Srebro, "The min-max complexity of distributed stochastic convex optimization with intermittent communication," 2021.
- [16] B. Woodworth, K. K. Patel, and N. Srebro, "Minibatch vs local sgd for heterogeneous distributed learning," *arXiv preprint arXiv:2006.04735*, 2020.
- [17] A. Godichon-Baggioni and S. Saadane, "On the rates of convergence of parallelized averaged stochastic gradient algorithms," *Statistics*, vol. 54, no. 3, pp. 618–635, 2020.
- [18] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic convex optimization," in *COLT*, 2009.
- [19] A. Daneshmand, G. Scutari, P. Dvurechensky, and A. Gasnikov, "Newton method over networks is fast up to the statistical precision," 2021.
- [20] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.
- [21] E. Gorbunov, F. Hanzely, and P. Richtárik, "Local sgd: Unified theory and new efficient methods," *arXiv preprint arXiv:2011.02828*, 2020.
- [22] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.
- [23] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Beijing, China: PMLR, 22–24 Jun 2014, pp. 1000–1008. [Online]. Available: <http://proceedings.mlr.press/v32/shamir14.html>

- [24] S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós, and A. Smola, “Aide: Fast and communication efficient distributed optimization,” *arXiv preprint arXiv:1608.06879*, 2016.
- [25] X.-T. Yuan and P. Li, “On convergence of distributed approximate newton methods: Globalization, sharper bounds and beyond,” *Journal of Machine Learning Research*, vol. 21, no. 206, pp. 1–51, 2020. [Online]. Available: <http://jmlr.org/papers/v21/19-764.html>
- [26] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach, and L. Massoulié, “Statistically preconditioned accelerated gradient method for distributed optimization,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 4203–4227. [Online]. Available: <http://proceedings.mlr.press/v119/hendrikx20a.html>
- [27] R.-A. Dragomir, A. Taylor, A. d’Aspremont, and J. Bolte, “Optimal complexity and certification of bregman first-order methods,” *arXiv preprint arXiv:1911.08510*, 2019.
- [28] Y. Sun, A. Daneshmand, and G. Scutari, “Convergence rate of distributed optimization algorithms based on gradient tracking,” *arXiv preprint arXiv:1905.02637*, 2019.
- [29] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach, and L. Massoulié, “Statistically preconditioned accelerated gradient method for distributed optimization,” *arXiv preprint arXiv:2002.10726*, 2020.
- [30] P. Dvurechensky, D. Kamzolov, A. Lukashovich, S. Lee, E. Ordentlich, C. A. Uribe, and A. Gashnikov, “Hyperfast second-order local solvers for efficient statistically preconditioned distributed optimization,” 2021.
- [31] Y. Arjevani and O. Shamir, “Communication complexity of distributed convex learning and optimization,” in *Advances in neural information processing systems*, 2015, pp. 1756–1764.
- [32] Y. Zhang and L. Xiao, *Communication-Efficient Distributed Optimization of Self-concordant Empirical Loss*. Cham: Springer International Publishing, 2018, pp. 289–341. [Online]. Available: [https://doi.org/10.1007/978-3-319-97478-1\\_11](https://doi.org/10.1007/978-3-319-97478-1_11)
- [33] Y. Nesterov and B. Polyak, “Cubic regularization of newton method and its global performance,” *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006. [Online]. Available: <http://dx.doi.org/10.1007/s10107-006-0706-8>
- [34] Y. Nesterov, “Accelerating the cubic regularization of newton’s method on convex problems,” *Mathematical Programming*, vol. 112, no. 1, pp. 159–181, Mar 2008. [Online]. Available: <https://doi.org/10.1007/s10107-006-0089-x>
- [35] M. Baes, “Estimate sequence methods: extensions and approximations,” Tech. Rep., 2009. [Online]. Available: [http://www.optimization-online.org/DB\\_FILE/2009/08/2372.pdf](http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf)
- [36] S. Ghadimi, H. Liu, and T. Zhang, “Second-order methods with cubic regularization under inexact information,” *arXiv:1710.05782*, 2017.
- [37] V. Feldman and J. Vondrak, “High probability generalization bounds for uniformly stable algorithms with nearly optimal rate,” in *Conference on Learning Theory*. PMLR, 2019, pp. 1270–1279.
- [38] G. Zhang and R. Heusdens, “Distributed optimization using the primal-dual method of multipliers,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 173–187, 2018.
- [39] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” in *International Conference on Machine Learning*. PMLR, 2016, pp. 1225–1234.
- [40] Y. Lei and Y. Ying, “Fine-grained analysis of stability and generalization for stochastic gradient descent,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 5809–5819.

## APPENDIX

### A. SOTA approaches for distributed stochastic optimization

In the recent paper [15] a novel lower bound for the SA approach was obtained:

$$\mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) \geq \frac{LR^2}{(N/m)^2} + \frac{\sigma R}{\sqrt{N}} + \min \left\{ \frac{LR^2}{T^2}, \frac{\sigma R}{\sqrt{N/m}} \right\}, \quad (37)$$

where  $N = nmT$ , since in the SA approach we see each observation once. That bound is matched by a combination

of two versions of Accelerated Gradient Descent [15]. Recall that the convergence rate of batched accelerated SGD is

$$O\left(\frac{LR^2}{t^2} + \frac{\sigma R}{\sqrt{tr}}\right),$$

where  $t$  is the number of iterations and  $r$  is batch size. To obtain (2) one can consider two cases of the distributed setting:

- single-machine  $m = 1$ ,  $t = \frac{N}{m}$ ,  $r = 1$ :

$$O\left(\frac{LR^2}{(N/m)^2} + \frac{\sigma R}{\sqrt{N/m}}\right);$$

- full batch  $r = \frac{N}{T}$ ,  $t = T$ :

$$O\left(\frac{LR^2}{T^2} + \frac{\sigma R}{\sqrt{N}}\right).$$

The lower bound (37) is matched up to a logarithmic factor with the combination of these two regimes.

There is also a lower bound for functions with Lipschitz Hessian, obtained by [15]

$$\frac{1}{(N/m)^2} + \frac{1}{\sqrt{N}} + \min \left\{ \frac{1}{T^2}, \frac{1}{\sqrt{N/m}}, \frac{1}{(N/m)^{1/4} T^{7/4}} \right\}$$

But it is not known whether it is accurate or not, since there is no method on which it reached.

In the paper [17] authors get a better convergence rate considering non-accelerated parallelized SGD with specific step size and only one communication at the end, assuming stronger local smoothness of objective near the solution

$$O\left(\frac{1}{(N/m)} + \frac{1}{\sqrt{N}}\right).$$

Considering the SAA approach, one should note that optimization methods for original stochastic problem can also be applied. The most common example is stochastic gradient descent. Papers [39], [40] show that SGD, used to minimize the empirical risk of the model, also reduces the generalization error, if the number of iterations is not very large (linear in sample size  $N$  [39]).

For the SAA approach Variance Reduction schemes can also be used. In the paper [9] authors propose VR scheme that converges as follows

$$\exp\left(-\min\left\{\frac{C_1 mnT}{N}, \frac{C_2 L_0 R}{L_1} \frac{nT}{\sqrt{N}}\right\}\right) + \frac{C_3 L_0 R}{\sqrt{N}}, \quad (38)$$

where  $C_1, C_2, C_3$  depends on logarithms of parameters  $N, m, T$ .

Accelerated VR algorithm from [14] can be applied to the problem of the form (9). In this case we have  $N = nK$  total observations of stochastic gradient. For the convergence of that variance reduced method parameters  $n, m, T$  must be selected in such a way that after  $s$  iterations the following two conditions are satisfied

- $\sqrt{\frac{L}{\mu}} s \leq C_1 T$  (stochastic updates);



- $\left(K + \sqrt{K\frac{L}{\mu}}\right) s \leq C_2 nT$  (full gradient computation)

to solve (9), where  $C_1, C_2$  may depend on  $K, n, m, T$  only logarithmically. Therefore, we get

$$s \leq \min \left\{ \frac{C_1 T}{\sqrt{L/\mu}}, \frac{C_2 nT}{K + \sqrt{KL/\mu}} \right\}.$$

Using that we have fixed size number of observations  $N = mK$  and  $\mu = \tilde{O}\left(\frac{L_0^2}{NR^2}\right)$ , we obtain

$$s \leq \min \left\{ C_1 \sqrt{\frac{L_0 R}{L}} \frac{T}{N^{1/4}}, C_2 \frac{mnT}{N}, C_2 \sqrt{\frac{L_0 R}{L}} \frac{\sqrt{mnT}}{N^{3/4}} \right\}.$$

From (21), using the fact that convergence of this algorithm is linear, we have

$$\begin{aligned} \mathbb{E}\mathbf{F}(x^T) - \mathbf{F}(x^*) &\leq \left(\frac{1}{2}\right)^s + \tilde{O}\left(\frac{L_0 R}{\sqrt{N}}\right) \leq \\ \exp\left(-\min\left\{\sqrt{C_1 \frac{L_0 R}{L}} \frac{T}{N^{1/4}}, C_2 \frac{mnT}{N}, C_2 \sqrt{\frac{L_0 R}{L}} \frac{\sqrt{mnT}}{N^{3/4}}\right\}\right) & \\ + \tilde{O}\left(\frac{L_0 R}{\sqrt{N}}\right). & \end{aligned} \quad (39)$$

To compare these methods with the proposed one (Algorithms 1-2) and with the lower bound (37) we derive dependence of parameters  $T$  and  $m$  on  $N$ , as we did before. Results are listed in the Table I.

In the case of  $\mu$ -strong convexity of  $f(x, \xi)$  w.r.t.  $x$  for all  $\xi$  offline bounds change since we do not need to regularize finite-sum approximation (see Section II for details). Therefore, convergence rate of Variance Reduction scheme (38) from [9] changes to

$$\exp\left(-\min\left\{C_1 \frac{nT}{L/\mu}, C_2 \frac{mnT}{N}\right\}\right) + O\left(\frac{L_0^2}{\mu N}\right)$$

And convergence rate (39) of accelerated VR method [14] in the strongly convex case is

$$\exp\left(-\min\left\{\frac{C_1 T}{\sqrt{L/\mu}}, \frac{C_2 T}{N/m}, \frac{C_2 T}{\sqrt{(NL)/(m\mu)}}\right\}\right) + O\left(\frac{L_0^2}{\mu N}\right).$$