2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

OR-SAGA: Over-relaxed stochastic average gradient mapping algorithms for finite sum minimization

Ion Necoara and Andrei Patrascu

Abstract— **In this paper we derive a family of stochastic average gradient methods for solving unconstrained convex problems with the objective function expressed as a finite sum, that uses a proximal operator oracle for each function instead of a gradient oracle. Our work is building on the recently introduced SAGA type method based on average gradient mapping. First we show that SAGA algorithm based on a proximal operator oracle can be interpreted as a stochastic variant of the alternating direction method of multipliers (ADMM). Using this ADMM interpretation, we derive a family of primal SAGA type schemes based on over-relaxation of the average gradient mapping. We analyze the convergence behavior of the proposed algorithms for the case when each objective function component is strongly convex and with Lipschitz continuous gradients. We prove that on this class of problems we achieve linear convergence rate. Numerical evidence supports the effectiveness of our over-relaxations in applications.**

I. Introduction

The randomness in most of the practical optimization applications led the stochastic optimization field to become an essential tool for many applied mathematics areas, such as machine learning and statistics [10], control and signal processing [12], [13], sensor networks [3], etc. In particular, most of the learning problems are formulated as stochastic optimization where the objective functions are expressed in terms of finite sums of the form:

$$f^* = \min_{x \in \mathbb{R}^n} f(x) + h(x), \text{ with } f = \frac{1}{N}\sum_{i=1}^{N} f_i, \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are smooth convex functions and h is convex, not necessarily smooth. This is an important optimization problem arising for example in:

1. *Stochastic linear model predictive control*. In this application we deal with stochastic linear systems:

$$z_t = A_t z_{t-1} + B_t u_{t-1} \quad t \geq 0, \tag{2}$$

where z_t/u_t denotes the current state/input. We assume that the system matrices A_t and B_t depend on the realization of a finite support random variable $\xi_t : \Omega_t \to \Sigma_t$ with known distribution, whose outcome is known only after the input u_{t-1} is fed to the system. In the usual non-stochastic settings, the optimality of an input signal $\mathbf{u} = (u_0 \cdots u_{T-1}) \in \mathcal{U}$ of length T (prediction horizon), where through the convex set $\mathcal{U}$ one imposes input constraints to the system, is measured in terms of a cost function:

$$f(\mathbf{u}) = \mathbb{E}\left[\sum_{t=0}^{T-1} \ell(z_t, u_t) + \ell_T(z_T)\right].$$

However, in the stochastic settings, *scenario reduction* techniques [5], [19] are usually employed in order to generate sequential finite approximations of the distribution. The support finiteness of the variable ξ_t implies that all the combinations of outcomes can be arranged in a tree structure $\mathcal{N}$ of height T, in other words a *scenario tree*. Since the states and inputs are dependent of the random variables, all the possible inputs and states can also be arranged on the scenario tree. We denote $\mathcal{N}_t \subseteq \mathcal{N}$ the subset of nodes at level t and identify a node (scenario) $\nu \in \mathcal{N}_t$ with the t-tuple $(\xi_1, \cdots, \xi_t)$. To the t-tuple node $\nu = (\xi_1, \cdots, \xi_t)$ we associate the variables $u_\nu = u(\xi_1, \cdots, \xi_t)$ and $z_\nu = z(\xi_1, \cdots, \xi_t)$, and denote its ancestor by $a(\nu)$. Clearly, there are no inputs associated to the leaf nodes, as well as no state variables are defined at the root. For simplicity, we define $\mathcal{N}^z = \mathcal{N} \backslash \mathcal{N}_0$ and $\mathcal{N}^u = \mathcal{N} \backslash \mathcal{N}_T$ as the sets of node indices corresponding to state and input variables. The data are denoted by A_ν and B_ν for all $\nu \in \mathcal{N}^z$. With respect to this notation, the stochastic model predictive control problem with uniform distribution is synthesized as follows:

$$\begin{aligned} \min_{u} \; & \sum_{t=0}^{T} \frac{1}{|\mathcal{N}|} \sum_{\nu \in \mathcal{N}^u} \ell_\nu(u_\nu, z_\nu) + \frac{1}{|\mathcal{N}|} \sum_{\nu \in \mathcal{N}_T} \ell_\nu(z_\nu) \\ & \text{s.t. } z_\nu = A_\nu z_{a(\nu)} + B_\nu u_{a(\nu)}, u_\nu \in \mathcal{U} \; \forall \nu \in \mathcal{N}^z. \end{aligned}$$

For a simpler formulation we use $f_\nu(u) = \sum\limits_{t=1}^{T} \ell_t(u_{\nu_t}, z_{\nu_t}) + \ell_T(z_{\nu_T})$. Therefore, the resulting optimization problem is written as a finite sum:

$$\min_{\mathbf{u} = \{u_\mu\}_{\mu \in \mathcal{N}^u}} \sum_{\nu \in \mathcal{N}} f_\nu(\mathbf{u}) + \sum_{\mu \in \mathcal{N}^u} I_{\mathcal{U}}(u_\mu)$$

where we use the notation $I_{\mathcal{U}}$ for the indicator function of the convex set $\mathcal{U}$. Clearly, this problem is of the form (1).

2. *Learning problems*. Another example is the binary classification with some loss function ℓ, e.g. squared Hinge loss $\ell(\tau, s) = \max^2\{0, 1 - s\tau\}$:

$$\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} \ell(a_i^T x + \mu, y_i) + h(x), \tag{3}$$

where $\{(a_i, y_i) : 1 \leq i \leq N\}$ is the data set containing the observations and class labels, μ is a bias term, and h is a

The research leading to these results has received funding from Executive Agency for Higher Education, Research and Innovation Funding (UEFISCDI), Romania: PNIII-P4-PCE-2016, project ScaleFreeNet, no. 39/2017.

I. Necoara and A. Patrascu are with Automatic Control and Systems Engineering Department, University Politehnica Bucharest, Romania.

978-3-9524-2699-9 ©2018 EUCA

given regularizer, e.g. $h(x) = \lambda\|x\|^2$. This problem is also of the form (1). The reader can find many other applications that fit into the framework (1).

The recent success of certain stochastic optimization methods for problem (1) has motivated increasingly great efforts into developments of new algorithms or into analyzing deeper the existing ones. A widely used approach for solving stochastic problems is based on first order methods. These schemes are typically the method of choice in practice for e.g. many machine learning or control applications due to their cheap iteration, simplicity, and superior empirical performance. One of the most popular algorithms for solving stochastic optimization problems is the stochastic gradient descent (SGD) method which independently samples an unbiased estimate of the gradient and then takes a step along this direction with a certain stepsize length [9], [10]. Stochastic first order methods based on proximal operator oracle has been also analyzed recently e.g. in [14], [20]. However, these methods do not converge linearly when the objective function is smooth and strongly convex. Therefore, recently variance reduction schemes have been developed that converge linearly on this class of objective functions. There are several variance reduction techniques that have been proposed in the last few years. For example, SVRG is a widely used variance reduction scheme due to its simple implementation [8]. However, despite the fact that SVRG is simple, several parameters need to be selected. First, we need to choose the number of epochs m_s at each state $s \in [S_{\max}]$. A simple choice is $m_s = N$, the number of data points, or more general $m_s = \mathcal{O}(N)$. Second, we need to choose the learning rate η, which usually depends on the assumptions on the functions f_i and h. We note that the original SVRG in [8] did not incorporate a proximal operator of h, and its extension to $h \neq 0$ has been given in [21]. Moreover, along with the standard SVRG or its prox variants, accelerated versions were also considered in [15] and [1]. Stochastic dual coordinate ascent methods for regularized finite sum minimization have been also proposed e.g. in [18]. Recently, in [17] the authors have proposed a stochastic average gradient method, known as SAG, and later extended to composite settings, called SAGA, [7], which do not require tuning m_s. Moreover, SAGA has been recently extended to work with proximal operator oracle of each function instead of a gradient oracle in [6].

In this paper we derive a family of over-relaxed stochastic average gradient mapping methods for solving finite sum minimization problems, that is the algorithms use a proximal operator oracle for each function in the sum. Our work is extending the recent results for the SAGA method based on exact average gradient mapping [6]. First we show that SAGA algorithm based on a proximal operator oracle can be interpreted as a stochastic dual variant of the alternating direction method of multipliers (ADMM). Using this ADMM interpretation, we derive a family of primal SAGA type schemes based on over-relaxation of the average gradient mapping. We analyze the convergence behavior of the proposed algorithms for the case when each objective function component is strongly convex and with Lipschitz continuous gradients and $h = 0$. We prove that on this class of problems we achieve linear convergence rate. Numerical evidence supports the effectiveness of our over-relaxations in applications.

A. Problem formulation

Although the previous applications are modeled by an optimization problem with composite objective function, we consider in the rest of our paper $h = 0$ as in [6], and we leave for future work the composite case $h \neq 0$. However, notice that for smooth regularizer, such as $h = \lambda\|x\|_2^2$, the function h can be included in the finite sum. Thus, in this paper we analyze the finite sum model:

$$f^\star = \min_{x\in\mathbb{R}^n} f(x) \quad \left(= \frac{1}{N}\sum_{i=1}^{N} f_i \right). \tag{4}$$

We make the following assumptions on f:

Assumption 1.1: Each function f_i is $\sigma > 0$ strongly convex, i.e. the following inequality holds:

$$\langle \nabla f_i(x) - \nabla f_i(y), x - y\rangle \geq \sigma\|x-y\|^2 \qquad \forall x, y \in \mathbb{R}^n.$$

Assumption 1.2: Each function f_i has $L > 0$ Lipschitz gradient, i.e. the following inequality holds:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x-y\| \qquad \forall x, y \in \mathbb{R}^n.$$

For the closed convex function f_i the corresponding proximal operator is given by:

$$\operatorname{prox}_{\rho f_i}(z) = \arg\min_{x\in\mathbb{R}^n} f_i(x) + 1/(2\rho)\|x-z\|^2.$$

We assume that the proximal operator can be computed efficiently for each function f_i in the sum above.

II. Proximal SAGA is a variant of stochastic ADMM

In this section we show that SAGA algorithm based on a proximal operator oracle [6] can be interpreted as a stochastic dual variant of the alternating direction method of multipliers (ADMM). Indeed, using the Fenchel conjugates of the functions f_i given by:

$$f_i^*(z) = \max_{y\in\mathbb{R}^n} \langle z, y\rangle - f_i(y),$$

we can rewrite equivalently the finite sum model (4) through its dual problem as follows:

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} f_i(x) \\
&= \min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} \max_{u_i\in\mathbb{R}^n} \langle x, u_i\rangle - f_i^*(u_i) \\
&= \min_{x\in\mathbb{R}^n} \max_{u\in\mathbb{R}^{nN}} \langle x, \frac{1}{N}\sum_{i=1}^{N} u_i\rangle - \frac{1}{N}\sum_{i=1}^{N} f_i^*(u_i).
\end{aligned}$$

By observing that the minimization in variable x requires $\sum_{u=1}^N u_i = 0$, the previous min-max (saddle point) problem can be written equivalently as:

$$\min_{u\in\mathbb{R}^{nN}} \frac{1}{N}\sum_{i=1}^N f_i^*(u_i) \;\text{ s.t. }\; \frac{1}{N}\sum_{i=1}^N u_i = 0.$$

It is well-known that an efficient strategy for solving the previous equality constrained problem is through the alternating direction method of multipliers (ADMM) [4]. Therefore, we consider the following augmented Lagrangian problem for some given scalar $\rho > 0$:

$$\max_{x\in\mathbb{R}^n}\min_{u\in\mathbb{R}^{nN}} \sum_{i=1}^N f_i^*(u_i) - \langle x, \sum_{i=1}^N u_i\rangle + \frac{\rho}{2N}\left\|\sum_{i=1}^N u_i\right\|^2.$$

For this augmented Lagrangian formulation we propose a new over-relaxed stochastic ADMM type scheme. The over-relaxation is given by the parameter $\alpha > 0$, so that our algorithm has the following update rules:
1. Choose uniformly random the index i and update:

$$\begin{aligned} u_i^{k+1} &= \arg\min_{u_i} f_i^*(u_i) - \langle x^k, u_i\rangle \\ &\qquad + \frac{\rho}{2}\left\|\alpha\sum_{j=1}^N u_j^k + u_i - u_i^k\right\|^2 \\ u_j^{k+1} &= u_j^k \quad \forall j \neq i. \end{aligned}$$

2. Update also x as:

$$x^{k+1} = x^k - \rho\left(\alpha\sum_{j=1}^N u_j^k + u_i^{k+1} - u_i^k\right).$$

Clearly, these update rules are completely different from the classical ADMM schemes [4] and the reader might ask if such strategy works at all. We will see that indeed such a scheme is convergent. Note that deterministic over-relaxed ADMM schemes have been proposed previously in the literature, see e.g. [16]. Also, notice that if $\alpha = 1$, we recover the stochastic variant of the method analyzed in [2]. Now, in order to write the proposed algorithm in a simpler form and in the primal framework, we recall the definition of the proximal operator of a function g:

$$\text{prox}_{\rho g}(z) = \arg\min_{x\in\mathbb{R}^n} g(x) + 1/(2\rho)\|x - z\|^2.$$

Then, the following basic relations are satisfied by the proximal operator [4]:

$$\text{prox}_g(z) + \text{prox}_{g^*}(z) = z, \quad (\rho g)^*(z) = \rho g^*(z/\rho).$$

Using now the notation $s_i^k = \frac{1}{\rho}x^k + u_i^k - \alpha\sum_{j=1}^N u_j^k$, then the previous updates can be simplified:

$$\begin{aligned} u_i^{k+1} &= \text{prox}_{\rho^{-1}f_i^*}(s_i^k) = s_i^k - \text{prox}_{(\rho^{-1}f_i^*)^*}(s_i^k) \\ &= s_i^k - \rho^{-1}\text{prox}_{\rho f_i}(\rho s_i^k). \end{aligned}$$

Now if we redefine the over-relaxed point $s_i^k \leftarrow \rho s_i^k$, i.e. $s_i^k = x^k + \rho u_i^k - \rho\alpha\sum_{j=1}^N u_j^k$, then we get:

$$u_i^{k+1} = \frac{1}{\rho}\left(s_i^k - \text{prox}_{\rho f_i}(s_i^k)\right),$$

which is the so-called gradient mapping of primal partial function f_i at the point $\text{prox}_{\rho f_i}(s_i^k)$. Moreover, the primal variable x is updates as follows:

$$x^{k+1} = x^k - \rho\left(\alpha\sum_{j=1}^N u_j^k + u_i^{k+1} - u_i^k\right) = \text{prox}_{\rho f_i}(s_i^k).$$

In conclusion, the proposed over-relaxed stochastic ADMM scheme, which we call *Over-Relaxed Stochastic Average Gradient mapping Algorithm* (OR-SAGA) for reasons that will be clear immediately (see Lemma 2.1 below), takes the following compact form:

```
Algorithm OR-SAGA ($x^0, u^0, \rho, \alpha$)
For $k \geq 1$ do:
1. Choose uniformly random an index $i \in [N]$
2. Compute the over-relaxed point:
       $s_i^k = x^k + \rho(u_i^k - \alpha\sum_{j=1}^N u_j^k)$
3. Update the proximal map for $f_i$:
       $u_i^{k+1} = \frac{1}{\rho}\left(s_i^k - \text{prox}_{\rho f_i}(s_i^k)\right)$ & $u_j^{k+1} = u_j^k \ \forall j \neq i$
4. Update the primal variable $x$:
       $x^{k+1} = \text{prox}_{\rho f_i}(s_i^k)$.
```

Note that OR-SAGA is based on the relaxation of the stochastic average gradient mapping, i.e. $(\alpha N)(1/N\sum_{j=1}^N u_j^k)$, where each u_j^k represents the proximal map of f_j at a specific point. Next lemma states that for a certain choice of α the previous OR-SAGA scheme becomes the SAGA algorithm based on the proximal operator oracle analyzed in [6]. The proof follows immediately by taking $\alpha = 1/N$ in our OR-SAGA scheme.

Lemma 2.1: For the relaxation parameter $\alpha = 1/N$ we recover the proximal SAGA scheme from [6].

III. Convergence analysis of OR-SAGA

In this section we analyze the convergence behavior of the proposed over-relaxed stochastic average gradient mapping algorithm (OR-SAGA) for different values of the over-relaxation parameter α. Moreover, in the convergence analysis we consider smooth and strongly convex functions f, i.e. Assumptions 1.1 and 1.2 are valid in this section. Let us define $v_i^* = x^* + \rho u_i^*$, where x^* is the primal solution and $u^* = (u_1^* \cdots u_N^*)$ is a dual solution. Then, we observe that $x^* = v_i^* - \rho\text{prox}_{\rho^{-1}f_i^*}\left(\frac{1}{\rho}v_i^*\right)$. We will use the following auxiliary results:

Lemma 3.1: Given $\alpha > 0$ the following holds:

$$\begin{aligned}&\mathbb{E}[\langle \rho(u_i^k - \alpha \sum_{j=1}^N u_j^k) - \rho u_i^*, s_i^k - v_i^* \rangle] \\ &= \frac{\rho^2}{N}\sum_{j=1}^N \|u_j^k - u_j^*\|^2 + \rho^2\left(\alpha^2 - \frac{2\alpha}{N}\right)\|\sum_{j=1}^N u_j^k\|^2 \\ &\quad + \rho\left(\frac{1}{N} - \alpha\right)\langle \sum_{j=1}^N u_j^k, x^k - x^* \rangle.\end{aligned}$$

Proof: From the definition of s_i^k and v_i^*, we get:

$$\begin{aligned}&\mathbb{E}[\langle \rho(u_i^k - \alpha \sum_{j=1}^N u_j^k) - \rho u_i^*, s_i^k - v_i^* \rangle] \\ &= \rho\mathbb{E}[\langle u_i^k - \alpha\sum_{j=1}^N u_j^k - u_i^*, x^k - x^* \rangle] \\ &\quad + \rho^2\mathbb{E}[\|u_i^k - \alpha\sum_{j=1}^N u_j^k - u_i^*\|^2]. \qquad (5)\end{aligned}$$

The second term in (5) can be written as:

$$\begin{aligned}&\mathbb{E}[\|u_i^k - \alpha\sum_{j=1}^N u_j^k - u_i^*\|^2] \\ &= \mathbb{E}[\|u_i^k - u_i^*\|^2] + \left(\alpha^2 - \frac{2\alpha}{N}\right)\|\sum_{j=1}^N u_j^k\|^2.\end{aligned}$$

On the other hand, first term in (5) can be written as:

$$\begin{aligned}&\mathbb{E}[\langle u_i^k - \alpha\sum_{j=1}^N u_j^k - u_i^*, x^k - x^* \rangle] \\ &= \left(\frac{1}{N} - \alpha\right)\langle \sum_{j=1}^N u_j^k, x^k - x^* \rangle,\end{aligned}$$

which leads to our statement. ■

The following lemma has been proved in [6], [11].

Lemma 3.2: The following inequalities (firm non-expansiveness) hold under Assumptions 1.1 and 1.2:

(i) $\langle \mathrm{prox}_{\rho^{-1}f_i^*}(x/\rho) - \mathrm{prox}_{\rho^{-1}f_i^*}(y/\rho), x - y \rangle$
$\geq \left(1 + \frac{1}{L\rho}\right)\|\mathrm{prox}_{\rho^{-1}f_i^*}(x/\rho) - \mathrm{prox}_{\rho^{-1}f_i^*}(y/\rho)\|^2$

(ii) $\langle x - \mathrm{prox}_{\rho^{-1}f_i^*}(x/\rho) - y + \mathrm{prox}_{\rho^{-1}f_i^*}(y/\rho), x - y \rangle$
$\geq (1+\rho\sigma)\|x - \mathrm{prox}_{\rho^{-1}f_i^*}(x/\rho) - y + \mathrm{prox}_{\rho^{-1}f_i^*}(y/\rho)\|^2,$

for all $x, y \in \mathbb{R}^n$ and any scalar $\rho > 0$.

Let us define the following constant:

$$\kappa = \frac{1}{1 + \sigma\rho} \in (0,\ 1).$$

The proof of the next lemma follows similar lines as in [6].

Lemma 3.3: Let Assumptions 1.1 and 1.2 hold and let $\{x^k\}_{k\geq 0}$ be the sequence generated by OR-SAGA scheme with parameters $\alpha, \rho > 0$. Then, the following holds:

$$\begin{aligned}&\mathbb{E}[\|x^{k+1} - x^*\|^2] \\ &\leq \frac{\kappa}{\alpha N}\|x^k - x^*\|^2 + \frac{\kappa\rho^2}{N}\sum_{j=1}^N \|u_j^k - u_j^*\|^2 \\ &\quad + \kappa\left(1 - \frac{1}{\alpha N}\right)\|x^k - x^* - \rho\alpha\sum_{j=1}^N u_j^k\|^2 \\ &\quad - \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2].\end{aligned}$$

Proof: From the definition of x^{k+1} and Lemma 3.2(ii) we have the following relations:

$$\begin{aligned}&\mathbb{E}[\|x^{k+1} - x^*\|^2] \\ &= \mathbb{E}[\|s_i^k - \rho\mathrm{prox}_{\rho^{-1}f_i^*}\left(\frac{1}{\rho}s_i^k\right) - v_i^* + \rho\mathrm{prox}_{\rho^{-1}f_i^*}\left(\frac{1}{\rho}v_i^*\right)\|^2] \\ &\leq \mathbb{E}[\frac{\rho}{1+\sigma\rho}\frac{1}{\rho}\langle s_i^k - \rho\mathrm{prox}_{\rho^{-1}f_i^*}\left(\frac{1}{\rho}s_i^k\right) - v_i^* \\ &\qquad + \rho\mathrm{prox}_{\rho^{-1}f_i^*}\left(\frac{1}{\rho}v_i^*\right), s_i^k - v_i^* \rangle] \\ &= \kappa\mathbb{E}[\langle x^{k+1} - x^*, s_i^k - v_i^* \rangle] \\ &= \kappa\mathbb{E}[\langle x^{k+1} - x^k + x^k - x^*, s_i^k - v_i^* \rangle] \\ &= \kappa\mathbb{E}[\langle x^{k+1} - x^k, s_i^k - v_i^* \rangle] + \kappa\mathbb{E}[\langle x^k - x^*, s_i^k - v_i^* \rangle] \\ &= \kappa\mathbb{E}[\langle x^{k+1} - x^k, s_i^k - v_i^* \rangle] + \kappa\|x^k - x^*\|^2 \\ &\quad + \rho\kappa\left(\frac{1}{N} - \alpha\right)\langle \sum_{j=1}^N u_j^k, x^k - x^* \rangle, \qquad (6)\end{aligned}$$

where in the last equality we used $E[s_i^k - v_i^*] = x^k - x^* + \left(\frac{1}{N} - \alpha\right)\sum_j u_j^k$. The first term in (6) can be refined as:

$$\begin{aligned}&\mathbb{E}[\langle x^{k+1} - x^k, s_i^k - v_i^* \rangle] \\ &= \mathbb{E}[\langle x^{k+1} - \rho u_i^* + \rho u_i^* - x^k, s_i^k - v_i^* \rangle] \\ &= \mathbb{E}[\langle x^k - \rho\left(\alpha\sum_{j=1}^N u_j^k - u_i^k\right) - \rho u_i^{k+1} - \rho u_i^* \\ &\qquad + \rho u_i^* - x^k, s_i^k - v_i^* \rangle] \\ &= \mathbb{E}[\langle \rho(u_i^k - \alpha\sum_{j=1}^N u_j^k) - \rho u_i^*, s_i^k - v_i^* \rangle] \\ &\qquad + \rho\mathbb{E}[\langle u_i^* - u_i^{k+1}, s_i^k - v_i^* \rangle] \\ &\leq \frac{\rho^2}{N}\sum_{j=1}^N \|u_j^k - u_j^*\|^2 + \rho^2\left(\alpha^2 - \frac{2\alpha}{N}\right)\|\sum_{j=1}^N u_j^k\|^2 \\ &\qquad + \rho\left(\frac{1}{N} - \alpha\right)\langle \sum_{j=1}^N u_j^k, x^k - x^* \rangle \\ &\qquad - \rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2],\end{aligned}$$

where in the last inequality, for the first term we used Lemma 3.1 and for the second term we used Lemma 3.2(i), since $u_i^{k+1} = \mathrm{prox}_{\rho^{-1}f_i^*}(s_i^k/\rho)$ and $u_i^* = \mathrm{prox}_{\rho^{-1}f_i^*}(v_i^*/\rho)$. By replacing this bound into (6) we obtain:

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1} - x^*\|^2] \\
&\leq \kappa\|x^k - x^*\|^2 + \kappa\frac{\rho^2}{N}\sum_j\|u_j^k - u_j^*\|^2 \\
&+ \kappa\rho^2\left(\alpha^2 - \frac{\alpha}{N}\right)\|\frac{1}{\rho\alpha}(x^k - x^*) - \sum_j u_j^k\|^2 \\
&- \kappa\left(1 - \frac{1}{\alpha N}\right)\|x^k - x^*\|^2 \\
&- \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2] \\
&= \frac{\kappa}{\alpha N}\|x^k - x^*\|^2 + \kappa\frac{\rho^2}{N}\sum_j\|u_j^k - u_j^*\|^2 \\
&+ \kappa\left(1 - \frac{1}{\alpha N}\right)\|x^k - x^* - \rho\alpha\sum_j u_j^k\|^2 \\
&- \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2],
\end{aligned}
$$

which confirms our result. ■

Based on the previous lemma we can now prove linear convergence rates for the iterates of OR-SAGA scheme. For a given scalar $\beta > 0$, let us define:

$$
T^k = \frac{\beta}{N}\sum_{j=1}^N\|u_j^k - u_j^*\|^2 + \|x^k - x^*\|^2.
$$

Theorem 3.4: Let Assumptions 1.1 and 1.2 hold, and let $\{x^k\}_{k\geq 0}$ be the sequence generated by OR-SAGA scheme for given parameters $\alpha, \rho > 0$. Then, the following linear convergence rates hold:

(i) If $\frac{\kappa}{N} < \alpha \leq \frac{1}{N}$, $\zeta = \sqrt{1/N + 1/(N\alpha) - 1}$ and $\frac{1}{\sigma L} \leq \beta \leq \frac{1}{4\sigma^2\zeta}\left(\frac{1}{\alpha N} - 1 + \sqrt{\left(\frac{1}{\alpha N} - 1\right)^2 + \frac{4\sigma\zeta}{L}}\right)^2$, then taking $\rho(\beta) = \sqrt{\frac{1}{4}\left(\frac{1}{L} - \frac{\sigma\beta}{N}\right)^2 + \frac{\beta}{N}} - \frac{1}{2}\left(\frac{1}{L} - \frac{\sigma\beta}{N}\right)$ we have:

$$
\mathbb{E}[T^{k+1}] \leq \frac{\kappa}{\alpha N}T^k, \quad \text{where } \frac{\kappa}{\alpha N} < 1.
$$

(ii) If $\frac{1}{N} \leq \alpha \leq \sqrt{\frac{N-2}{2N^2}}$, $\frac{1+c}{\sigma L(c/N+1)} \leq \beta \leq \left[\frac{\alpha(N-c)-1}{\zeta\sigma\alpha N(1+c)} + \frac{1}{\sigma}\sqrt{\frac{[\alpha(N-c)-1]^2}{\zeta^2\alpha^2N^2(1+c)^2} + \frac{\sigma}{L}}\right]^2$ and $\rho(\beta) = \sqrt{\frac{1}{4}\left(\frac{\beta\sigma}{N} - \frac{1}{L}\right)^2 + \frac{\beta}{N}} - \frac{1}{2}\left(\frac{1}{L} - \frac{\beta\sigma}{N}\right)$, where $\zeta = \sqrt{\frac{1+1/N-1/(\alpha N)}{1+c}} + \sqrt{1/N}$ and $c = 2N^2\alpha\left(\alpha - \frac{1}{N}\right)$, we have:

$$
\mathbb{E}[T^{k+1}] \leq \kappa\left(2 - \frac{1}{\alpha N}\right)T^k, \text{ where } \kappa\left(2 - \frac{1}{\alpha N}\right) < 1.
$$

Proof: (i) If $\frac{\kappa}{N} < \alpha \leq \frac{1}{N}$, then from Lemma 3.3 we have the recurrence:

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1} - x^*\|^2] \\
&\leq \frac{\kappa}{\alpha N}\|x^k - x^*\|^2 + \kappa\frac{\rho^2}{N}\sum_{j=1}^N\|u_j^k - u_j^*\|^2 \\
&- \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2].
\end{aligned}
$$

In order to obtain a recurrence in T^k, we add on both sides the term $\frac{\beta}{N}\mathbb{E}[\sum_{j=1}^N\|u_j^{k+1} - u_j^*\|^2] = \left(1 - \frac{1}{N}\right)\frac{\beta}{N}\sum_{j=1}^N\|u_j^k - u_j^*\|^2 + \frac{\beta}{N}\mathbb{E}[\|u_j^{k+1} - u_j^*\|^2]$ and derive that:

$$
\begin{aligned}
&\mathbb{E}[T^{k+1}] \\
&\leq \frac{\kappa}{\alpha N}T^k + \left(\beta - \frac{\beta}{N} - \frac{\kappa\beta}{\alpha N} + \kappa\rho^2\right)\frac{1}{N}\sum_j\|u_j^k - u_j^*\|^2 \\
&+ \left(\frac{\beta}{N} - \kappa\rho^2 - \frac{\kappa\rho}{L}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2].
\end{aligned}
$$

Then, imposing the coefficients of the last two terms to be nonpositive, we get that for any β satisfying $\frac{1}{\sigma L} \leq \beta \leq \frac{1}{4\sigma^2\zeta}\left(\frac{1}{\alpha N} - 1 + \sqrt{\left(\frac{1}{\alpha N} - 1\right)^2 + \frac{4\sigma\zeta}{L}}\right)^2$, where $\zeta = \sqrt{1/N + 1/(N\alpha) - 1}$, and any ρ satisfying $\frac{1}{2}\left(\frac{\beta\sigma}{N} - \frac{1}{L}\right) + \sqrt{\frac{1}{4}\left(\frac{1}{L} - \frac{\sigma\beta}{N}\right)^2 + \frac{\beta}{N}} \leq \rho \leq \sqrt{\frac{\sigma^2\beta^2}{4}\left(1 - \frac{1}{N}\right)^2 + \beta\left(\frac{1}{\alpha N} - 1 + \frac{1}{N}\right)} - \frac{\sigma\beta}{2}\left(1 - \frac{1}{N}\right)$, we obtain the desired recurrence $\mathbb{E}[T^{k+1}] \leq \kappa/(\alpha N)T^k$.

(ii) Now let $\alpha \geq \frac{1}{N}$. From Lemma 3.3 we have:

$$
\begin{aligned}
\mathbb{E}[\|x^{k+1} - x^*\|^2] &\leq \frac{\kappa}{\alpha N}\|x^k - x^*\|^2 + \kappa\frac{\rho^2}{N}\sum_j\|u_j^k - u_j^*\|^2 \\
&+ \kappa\left(1 - \frac{1}{\alpha N}\right)\|x^k - x^* - \rho\alpha\sum_j u_j^k\|^2 \\
&- \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2].
\end{aligned}
$$

By using relation $\|x^k - x^* - \rho\alpha\sum_j u_j^k\|^2 \leq 2\|x^k - x^*\|^2 + 2\rho^2\alpha^2\|\sum_j u_j^k\|^2$, then we further have:

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1} - x^*\|^2] \\
&\leq \kappa\left(2 - \frac{1}{\alpha N}\right)\|x^k - x^*\|^2 + \kappa\frac{\rho^2}{N}\sum_j\|u_j^k - u_j^*\|^2 \\
&+ 2\kappa\rho^2N^2\left(\alpha^2 - \frac{\alpha}{N}\right)\|\frac{1}{N}\sum_j u_j^k\|^2 \\
&- \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2] \\
&\leq \kappa\left(2 - \frac{1}{\alpha N}\right)\|x^k - x^*\|^2 - \kappa\rho^2\left(1 + \frac{1}{L\rho}\right)\mathbb{E}[\|u_i^{k+1} - u_i^*\|^2] \\
&+ \left[\kappa\frac{\rho^2}{N} + 2\kappa\rho^2N\alpha\left(\alpha - \frac{1}{N}\right)\right]\sum_j\|u_j^k - u_j^*\|^2,
\end{aligned}
$$

where in the last inequality we used $\|\frac{1}{N}\sum_j u_j^k\|^2 = \|\frac{1}{N}\sum_j(u_j^k - u_j^*)\|^2 \leq \frac{1}{N}\sum_j\|u_j^k - u_j^*\|^2$. We argue that

$\kappa\left(2-\frac{1}{\alpha N}\right) \in (0,1)$ as follows: first, $\kappa\left(2-\frac{1}{\alpha N}\right) > 0$ is equivalent to $2-\frac{1}{\alpha N} > 0$, that holds since $\alpha \geq \frac{1}{N} > \frac{1}{2N}$; second, $\kappa\left(2-\frac{1}{\alpha N}\right) < 1$ is equivalent to $1-\frac{1}{\alpha N} \leq \rho\sigma$, which is ensured by $\rho < 1/\sigma$ (note that our choices for β and ρ guarantees $\rho < 1/\sigma$). To obtain a recurrence in T^k we add on both sides $\frac{\beta}{N}\mathbb{E}[\sum_{j=1}^{N}\|u_j^{k+1}-u_j^*\|^2] = \left(1-\frac{1}{N}\right)\frac{\beta}{N}\sum_{j=1}^{N}\|u_j^k-u_j^*\|^2 + \frac{\beta}{N}\mathbb{E}[\|u_j^{k+1}-u_j^*\|^2]$ and get:

$$\begin{aligned}\mathbb{E}[T^{k+1}] &\leq \kappa\left(2-\frac{1}{\alpha N}\right)T^k \\ &+\left(\frac{\beta}{N}-\kappa\rho^2-\frac{\kappa\rho}{L}\right)\mathbb{E}[\|u_i^{k+1}-u_i^*\|^2] \\ &+\left(\beta-\frac{\beta}{N}-\kappa\beta(2-\frac{1}{\alpha N})+\kappa\rho^2+\kappa\rho^2 c\right)\frac{1}{N}\sum_j\|u_j^k-u_j^*\|^2,\end{aligned}$$

where $c = 2N^2\alpha\left(\alpha-\frac{1}{N}\right)$. By assuming that $\frac{1}{N} \leq \alpha \leq \sqrt{\frac{N-2}{2N^2}}$, the bounds $\frac{1+c}{\sigma L(c/N+1)} \leq \beta \leq \left[\frac{\alpha(N-c)-1}{\zeta\sigma\alpha N(1+c)} + \frac{1}{2\sigma}\sqrt{\frac{[2\alpha(N-c)-1]^2}{\zeta^2\alpha^2N^2(1+c)^2}+\frac{4\sigma}{L}}\right]^2$ and $\rho(\beta) = \sqrt{\frac{1}{4}\left(\frac{\beta\sigma}{N}-\frac{1}{L}\right)^2+\frac{\beta}{N}} - \frac{1}{2}\left(\frac{1}{L}-\frac{\beta\sigma}{N}\right)$, where $\zeta = \sqrt{\frac{1+1/N-1/(\alpha N)}{1+c}} + \sqrt{1/N}$, we ensure that the last two terms in the previous inequality are non-positive. ■

From previous theorem it follows that for $\alpha \in (\kappa/N, 1/N]$ to obtain an ϵ-solution, the OR-SAGA scheme requires the following number of iterations:

$$\mathcal{O}\left(\frac{1}{1-\kappa/(\alpha N)}\log\left(\frac{1}{\epsilon}\right)\right).$$

On the other hand, for $\alpha \in [1/N, \sqrt{(N-2)/(2N^2)}]$, in order to obtain an ϵ-solution, the OR-SAGA requires the following number of iterations:

$$\mathcal{O}\left(\frac{1}{1-\kappa\left(2-\frac{1}{\alpha N}\right)}\log\left(\frac{1}{\epsilon}\right)\right).$$

Notice that in practice we observe that the over-relaxation $\alpha \geq 1/N$ usually leads to a better behavior of algorithm OR-SAGA (see also the numerical tests in the next section). However, according to Theorem 3.4, the optimal rate is achieved for $\alpha = 1/N$, when $1/(\alpha N) = 2 - 1/(\alpha N)$, and in this case we need $\mathcal{O}\left(\left(\sqrt{\frac{NL}{\sigma}}+N\right)\log\left(\frac{1}{\epsilon}\right)\right)$ iterations to obtain an ϵ-solution.

IV. Numerical experiments

In this section we perform some preliminary numerical tests to evaluate the practical performance of OR-SAGA. We consider smooth and strongly convex functions of the form:

$$f_i(x) = \max{}^2(0, a_i^T x - b_i) + \sigma/2\|x\|^2,$$

where (a_i, b_i) are randomly drawn from normal distribution with zero mean and unit variance, for all $i = 1:N$. Also, the initial iterate x^0 is drawn similarly. We test algorithm OR-SAGA for dimensions $n = 50$ and $N = \{32, 64, 128, 512\}$ for 10 randomly generated problems. We choose the values $\rho = 2$ and for the over-relaxation parameter we consider three choices $\alpha = \{\frac{1}{10N}, \frac{1}{N}, \frac{10}{N}\}$. The average total number of iterations in order to achieve 10^{-2} accuracy are displayed in the table below. From this table we observe that the over-relaxation $\alpha > 1/N$ can bring benefits to the behavior of the OR-SAGA algorithm, allowing to perform less number of iterations than for the choice $\alpha = 1/N$ considered in [6].

$\alpha\backslash N$	32	64	128	512
$1/10N$	5771	4696	5664	8489
$1/N$	2599	4351	4574	9823
$10/N$	2210	3981	4112	8565

References

[1] Z. Allen-Zhu, *Katyusha: The first direct acceleration of stochastic gradient methods*, Journal of Machine Learning Research, in press, 2017.

[2] D. P. Bertsekas. *Incremental aggregated proximal and augmented lagrangian algorithms*, arXiv:1509.09257, 2015.

[3] D. Blatt and A.O. Hero, *Energy based sensor network source localization via projection onto convex sets*, IEEE Transactions on Signal Processing, 54(9): 3614–3619, 2006.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3(1): 1–122, 2011.

[5] G. Calafiore and L. Fagiano, *Stochastic model predictive control of lpv systems via scenario optimization*, Automatica, 49(6), 2013.

[6] A. Defazio, *A simple practical accelerated method for finite sums*, Advances in Neural Information Processing Systems (NIPS), 2016.

[7] A. Defazio, F. Bach and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, Advances in Neural Information Processing Systems (NIPS), 2014.

[8] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in Neural Information Processing Systems (NIPS), 2013.

[9] G. Lan, *An optimal method for stochastic composite optimization*, Mathematical Programing, 133(1): 365–397, 2012.

[10] E. Moulines and F.R. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Advances in Neural Information Processing Systems, 2011.

[11] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, 2004.

[12] I. Necoara, V. Nedelcu and I. Dumitrache, *Parallel and distributed optimization methods for estimation and control in networks*, Journal of Process Control, 21(5), 2011.

[13] I. Necoara, D. Clipici, P. Patrinos and A. Bemporad, *MPC for power systems dispatch based on stochastic optimization*, IFAC World Congress, 2014.

[14] A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*, Journal of Machine Learning Research, 2018.

[15] A. Nitanda, *Accelerated Stochastic Gradient Descent for Minimizing Finite Sums*, Artificial Intel. & Statistics, 2016.

[16] R. Nishihara, L. Lessard, B. Recht, A. Packard and M. Jordan, *A General Analysis of the Convergence of ADMM*, ICML, 2015.

[17] N. Le Roux, M. Schmidt, and F. Bach, *A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets*, Advances in Neural Information Processing Systems, 2012.

[18] S. Shalev-Schwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, Journal of Machine Learning Research, 14: 567–599, 2013.

[19] A. Themelis, S. Villa, P. Patrinos and A. Bemporad, *Stochastic gradient methods for stochastic model predictive control*, European Control Conference, 2016.

[20] P. Toulis, D. Tran and E. Airoldi, *Towards stability and optimality in stochastic gradient descent*, International Conference on Artificial Intelligence and Statistics, 1290–1298, 2016.

[21] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM Journal on Optimization, vol. 24, no. 2, pp. 2057–2075, 2014.