DECENTRALIZED ONLINE OPTIMIZATION WITH HETEROGENEOUS DATA SOURCES

Alec Koppel[†], *Brian M. Sadler*[⋆], *and Alejandro Ribeiro*[†]

[†]Department of Electrical and Systems Engineering, University of Pennsylvania
[⋆]U.S. Army Research Laboratory, Adelpi, MD 20783, USA

ABSTRACT

We consider stochastic optimization problems in decentralized settings, where a network of agents aims to learn decision variables which are optimal in terms of a global objective which depends on possibly heterogeneous streaming observations received at each node. Consensus optimization techniques implicitly operate on the hypothesis that each node aims to learn a common parameter vector, which is inappropriate for this context. Motivated by this observation, we formulate a problem where each agent minimizes a global objective while enforcing network proximity constraints that may encode correlation structures among the observations at each node. To solve this problem, we propose a decentralized stochastic saddle point algorithm inspired by Arrow and Hurwicz. We establish that under a constant step-size regime the time-average suboptimality and constraint violation are contained in a neighborhood whose radius vanishes with the iteration index. Further, the time-average primal vectors converge to the optimal objective while satisfying the network proximity constraints. We apply this method to an online source localization problem and show it outperforms consensus-based schemes.

1. INTRODUCTION

We consider online multi-agent optimization problems, where a group of agents aim to minimize a global objective $f = \sum_i f_i$ which may be written as a sum of local objectives f_i available at different nodes i of a network $\mathcal{G} = (V, \mathcal{E})$. The problem is online because information upon which the local objectives depend is sequentially and locally received by each agent. Many collaborative pattern recognition and estimation tasks correspond to cases where the data distribution at each node is distinct, but correlated with other nodes. Consensus optimization implicitly operates on the hypothesis that the observations across the network are independently and identically distributed, and hence will fail when data heterogeneity is present. Thus, we focus on the setting where agents aim to keep their decision variables *close* to one another but *not coincide* in order to minimize this global objective while giving preference to distinct local signals.

Consensus optimization has a rich history in networked control systems, [3–5] wireless systems [6, 7], sensor networks [8, 9], and others. Prior approaches to this problem require each agent to keep a local copy of the global decision variable, and approximately enforce an agreement constraint between the local copies at each iteration. Information mixing strategies for this setting include weighted averaging [10–12], dual reformulations [13, 14], and primal-dual methods [15–19]. To handle optimization problems with streaming data, stochastic methods have been developed [20, 21].

We propose a stochastic variant of the saddle point method [15, 16] (Section 3) to solve online multi-agent optimization problems with network proximity constraints formulated in Section 2. This method allows agents the leeway to select actions which are good with respect to a global cost while not ignoring the structure of locally observed information. We establish that under a constant step-size regime the time-average suboptimality and constraint violation are contained in a neighborhood whose radius vanishes with the iteration index (Section 4). We also apply this method to a source localization problem (Section 5).

2. PROBLEM FORMULATION

We consider agents i of a symmetric and connected network $\mathcal{G} = (V, \mathcal{E})$ with $|V| = N$ nodes and $|\mathcal{E}| = M$ edges and denote as $n_i := \{j : (i,j) \in \mathcal{E}\}$ the neighborhood of agent i. Each of the agents is associated with a (non-strongly) convex loss function $f_i : \mathcal{X} \times \Theta_i \to \mathbb{R}$ that is parameterized by a decision variable $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ and a random variable $\boldsymbol{\theta}_i \in \Theta_i$ with a proper distribution. Throughout, we assume $\mathcal{X}$ is a compact convex subset of $\mathbb{R}^p$ associated with the p-dimensional parameter vector of agent i. The functions $f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)$ for different $\boldsymbol{\theta}_i$ are interpreted as observations of a stochastic model with a possible goal for agent i being the computation of the optimal local estimate,

$$\mathbf{x}_i^{\mathrm{L}} := \underset{\mathbf{x}_i \in \mathcal{X}}{\operatorname{argmin}}\, F_i(\mathbf{x}_i) := \underset{\mathbf{x}_i \in \mathcal{X}}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)] \,. \tag{1}$$

Here the functions $f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)$ are termed instantaneous because they are observed at particular points in time associated with realizations of the random variable $\boldsymbol{\theta}_i$; see Section 3. $F_i(\mathbf{x}_i) := \mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)]$ denotes the local average function at node i.

When we consider the network as a whole we can define the stacked vector $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, which is an element of the product set $\mathbb{R}^{Np}$, and the aggregate function $F(\mathbf{x}) := \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)]$. It then follows that the set of problems in (2) is equivalent to the aggregate problem

$$\mathbf{x}^{\mathrm{L}} = \underset{\mathbf{x} \in \mathcal{X}^N}{\operatorname{argmin}}\, F(\mathbf{x}) := \underset{\mathbf{x} \in \mathcal{X}^N}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)] \,. \tag{2}$$

Further define the stacked instantaneous function as $f(\mathbf{x}, \boldsymbol{\theta}) = \sum_i f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)$. (1) and (2) describe the same problem because

This work in this paper is supported by NSF CCF-1017454, NSF CCF-0952867, ONR N00014-12-1-0997, ARL MAST CTA, and ASEE SMART. Part of the results in this paper appeared in [1]. All proofs are given in [2].

there is no coupling between the variables $\mathbf{x}_i$ at different agents. In many situations, however, the parameters $\mathbf{x}_i^{\mathrm{L}}$ that different agents want to estimate are related. It then makes sense to couple decisions of different agents as a means of letting agents exploit each others' observations. Consensus optimization works on the hypothesis that all agents seek common parameters $\mathbf{x}_i$ for all $i \in V$, whereby we modify (2) via consensus constraints

$$\mathbf{x}_i = \mathbf{x}_j, \text{ for all } j \in n_i \,. \tag{3}$$

For a connected network this constraint makes all variables $\mathbf{x}_i$ equal – hence the definition as a consensus problem. This hypothesis only makes sense in cases where agents observe information drawn from a common distribution, which may be overly restrictive. In general, parameters of nearby nodes are expected to be close but are not necessarily all equal, as is the situation in, e.g., the estimation of a smooth field that is albeit not uniform. To model this situation we introduce a convex local proximity function with real-valued range of the form $h_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and a tolerance $\gamma_{ij} \geq 0$. These are used to couple the decisions of agent i to its neighbors $j \in n_i$ through the definition of the optimal estimates via the constrained problem

$$\mathbf{x}^* := \underset{\mathbf{x}\in\mathcal{X}^N}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)] \text{ s.t. } h_{ij}(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma_{ij}, \tag{4}$$

where the constraint is taken for all $j \in n_i$. In the formulation in (4) we assume that the proximity function $h_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ that couples node i to node j is equivalent to the proximity function $h_{ji}(\mathbf{x}_j, \mathbf{x}_i)$ that couples node j to node i. I.e., we assume that for all $\mathbf{x}_i$ and $\mathbf{x}_j$ we have $h_{ij}(\mathbf{x}_i, \mathbf{x}_j) = h_{ji}(\mathbf{x}_j, \mathbf{x}_i)$. This implies that the constraints $h_{ij}(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma_{ij}$ and $h_{ji}(\mathbf{x}_j, \mathbf{x}_i) \leq \gamma_{ji}$ are redundant. We also define the stacked constraint $h : \mathcal{X}^N \to \mathbb{R}^M$. We keep them separate for algorithm symmetry (Section 3).

The consensus constraints in (3) are a particular example of a proximity function $h_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ but so is the norm constraint $\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \gamma_{ij}$. This latter choice makes the estimates $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$ of neighboring nodes close to each other but not equal. Implicitly, this allows i to incorporate the information of neighboring nodes without the detrimental effect of trying to incorporate the information of far away nodes that may only weakly correlated with the parameter that i tries to estimate.

3. ALGORITHM DEVELOPMENT

Recall that a decentralized algorithm is one in which node i has access to local functions $f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)$ and local constraints $h_{ij}(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma_{ij}$ and exchanges information with neighbors $j \in n_i$ only. Recall also that the algorithm is further said to be online if the distribution of θ_i is unknown and agent i has access to independent observations $\boldsymbol{\theta}_{i,t}$ that are acquired sequentially. Our goal is to develop an online decentralized algorithm to solve (4). To achieve this we consider the approximate Lagrangian relaxation of (4) which we state as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{N} \Bigg[\mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)] \tag{5}$$
$$+ \frac{1}{2} \sum_{j\in n_i} \left(\lambda_{ij} \left(h_{ij}(\mathbf{x}_i, \mathbf{x}_j) - \gamma_{ij} \right) - \frac{\delta\epsilon_t}{2} \lambda_{ij}^2 \right) \Bigg],$$

where $\lambda_{ij} \in \mathbb{R}^+$ is a nonnegative Lagrange multiplier associated with the proximity constraint between node i and node j. Observe that (5) *does not* define the Lagrangian of the optimization problem (4), but instead defines an *augmented Lagrangian* due to the presence of the last term on the right-hand side. This last term $-(\delta\epsilon_t/2)\lambda_{ij}^2$, with scalar parameters δ and ϵ_t, is a regularizer on the dual variable, whose utility arises in controlling the accumulation of constraint violation.

We propose applying a stochastic saddle point algorithm to (5) which operates by alternating primal and dual stochastic gradient descent and ascent steps respectively. Define the stacked dual variable as $\boldsymbol{\lambda} := [\lambda_1; \cdots; \lambda_M] \in \mathbb{R}^M$. Moreover, denote the network aggregate random variable as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1; \cdots; \boldsymbol{\theta}_N]$. Particularized to the Lagrangian stated in (5), the stochastic saddle point method takes the form

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{X}^N}\Big[\mathbf{x}_t - \epsilon_t \nabla_{\mathbf{x}} \hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)\Big] \,, \tag{6}$$

$$\boldsymbol{\lambda}_{t+1} = \Big[\boldsymbol{\lambda}_t + \epsilon_t \nabla_{\boldsymbol{\lambda}} \hat{\mathcal{L}}_t(\mathbf{x}_{t+1}, \boldsymbol{\lambda}_t)\Big]_+ \,, \tag{7}$$

where $\nabla_{\mathbf{x}} \hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)$ and $\nabla_{\boldsymbol{\lambda}} \hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)$, are the primal and dual stochastic gradients of the augmented Lagrangian with respect to $\mathbf{x}$ and $\boldsymbol{\lambda}$, respectively. These stochastic subgradients are approximations of the gradients of (5) evaluated at the current realization of the random variable $\boldsymbol{\theta}$. The notation $\mathcal{P}_{\mathcal{X}^N}(\mathbf{x})$ denotes component-wise orthogonal projection of the individual primal variables $\mathbf{x}_i$ onto the given convex compact set $\mathcal{X}$, and $[\cdot]_+$ denotes the projection onto the M-dimensional nonnegative orthant $\mathbb{R}_+^M$. As an abuse of notation, we also use $[\cdot]_+$ to denote scalar positive projection where appropriate.

The method stated in (6) - (7) can be implemented with decentralized computations across the network, as we state next.

Proposition 1 *Let $\mathbf{x}_{i,t}$ be the ith component of the primal iterate $\mathbf{x}_t$ and $\lambda_{ij,t}$ the i,jth component the dual iterate $\boldsymbol{\lambda}_t$. The primal variable update is equivalent to N parallel local variable updates*

$$\mathbf{x}_{i,t+1} = \mathcal{P}_{\mathcal{X}}\Big[\mathbf{x}_{i,t} - \epsilon_t \Big(\nabla_{\mathbf{x}_i} f_i(\mathbf{x}_{i,t}; \boldsymbol{\theta}_{i,t}) \tag{8}$$
$$+ \sum_{j\in n_i} (\lambda_{ij,t} + \lambda_{ji,t}) \nabla_{\mathbf{x}_i} h_{ij}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) \Big) \Big] \,.$$

Likewise, the dual updates in (7) *are equivalent to M updates*

$$\lambda_{ij,t+1} = \Big[(1 - \epsilon_t\delta)\lambda_{ij,t} + \epsilon_t \left(h_{ij}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) - \gamma_{ij} \right) \Big]_+ . \tag{9}$$

With primal variables $\mathbf{x}_{i,t}$ and Lagrange multipliers $\lambda_{ij,t}$ maintained and updated by node i, Proposition 1 implies that the saddle point method in (6)-(7) can be translated into a decentralized protocol in which: (i) The primal and dual variables variables of distinct agents are decoupled from one another. (ii) The updates require exchanges of information among neighbors.

4. CONVERGENCE ANALYSIS

We turn to establishing that the objective error sequence and constraint violation incurred by the saddle point algorithm defined by (6)-(7) when used with a constant step-size are contained within a neighborhood whose average radius vanishes with the iteration index. To establish these results, we note some facts of the problem setting, and introduce a few assumptions.

First, the dual stochastic gradient is independent of $\boldsymbol{\theta}_{i,t}$ [cf. (9)], and hence for all t, $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}_t,\boldsymbol{\lambda}_t)=\nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}_t(\mathbf{x}_t,\boldsymbol{\lambda}_t)$. Also pertinent to analyzing the performance of the stochastic saddle point method is the fact that the primal stochastic gradient of the Lagrangian is an unbiased estimator of the true primal gradient. Let $\mathcal{F}_t$ be a sigma algebra that measures the history of the algorithm up until time t, i.e. a collection that contains at least the variables $\{\mathbf{x}_u,\boldsymbol{\lambda}_u,\boldsymbol{\theta}_u\}_{u=1}^t\subseteq\mathcal{F}_t$. The primal stochastic gradient is an unbiased estimate of the true primal gradient, i.e.

$$\mathbb{E}\left[\nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t,\boldsymbol{\lambda}_t)\,\middle|\,\mathcal{F}_t\right]=\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_t,\boldsymbol{\lambda}_t)\;. \tag{10}$$

Furthermore, the compactness of sets $\mathcal{X}$ allows us to bound the magnitude of the iterates $\mathbf{x}_{i,t}$ by a constant R/N, which implies that the network-wide iterates may be bounded as $\|\mathbf{x}_t\|\leq R$ for all t . To prove convergence of the stochastic saddle point method, some conditions are required of the network, loss functions, and constraints, which we now state.

(A1) (Network connectivity) The network $\mathcal{G}$ is symmetric and connected with diameter D.

(A2) (Smoothness) The stacked instantaneous objective is Lipschitz continuous in expectation with constant L_f, i.e. for distinct primal variables $\mathbf{x},\tilde{\mathbf{x}}\in\mathcal{X}$ and all $\boldsymbol{\theta}$, we have

$$\mathbb{E}\left[\|f(\mathbf{x},\boldsymbol{\theta})-f(\tilde{\mathbf{x}},\boldsymbol{\theta})\|\right]\leq L_f\|\mathbf{x}-\tilde{\mathbf{x}}\|\;, \tag{11}$$

Moreover, the stacked constraint function $h(\mathbf{x})$ is Lipschitz continuous with modulus L_h. That is, for distinct primal variables $\mathbf{x},\tilde{\mathbf{x}}\in\mathcal{X}$, we may write

$$\|h(\mathbf{x})-h(\tilde{\mathbf{x}})\|\leq L_h\|\mathbf{x}-\tilde{\mathbf{x}}\|. \tag{12}$$

Assumption 1 ensures that the network is connected. Assumption 2 states that the stacked objective and constraints are sufficiently smooth, and have bounded gradients.

The following theorem bounds the sub-optimality and constraint violation of the saddle point iterates, when a specific constant algorithm step-size is chosen, as we state next.

Theorem 1 *Denote $(\mathbf{x}_t,\boldsymbol{\lambda}_t)$ as the sequence generated by the saddle point algorithm in* (6)-(7) *and let Assumptions 1 - 2 hold. Suppose the algorithm is run for T iterations with constant step-size $\epsilon_t=1/\sqrt{T}$, then the time aggregation of the objective function error sequence $F(\mathbf{x}_t)-F(\mathbf{x}^*)$, with $\mathbf{x}^*$ defined as in* (4)*, grows sublinearly with final iteration index T as*

$$\sum_{t=1}^{T}[F(\mathbf{x}_t)-F(\mathbf{x}^*)]\leq\mathcal{O}(\sqrt{T}). \tag{13}$$

Moreover, the time-aggregation of the constraint violation of the algorithm grows sublinearly with the final iteration index T as

$$\sum_{(i,j)\in\mathcal{E}}\Big[\sum_{t=1}^{T}\Big(h_{ij}(\mathbf{x}_{i,t},\mathbf{x}_{j,t})-\gamma_{ij}\Big)\Big]_+\leq\mathcal{O}(T^{3/4}). \tag{14}$$

Theorem 1, whose proof is inspired by [22], establishes that the stochastic saddle point method, when run with a fixed algorithm step-size, yields an objective function error sequence whose difference is bounded by a constant strictly less times than T, the final iteration index. Moreover, the time-accumulation of the constraint violation incurred by the algorithm is strictly smaller than T, the final iteration index. Thus, for larger T, the time-average difference between $F(\mathbf{x}_t)$ and $F(\mathbf{x}^*)$ goes to null, as does the average constraint violation. Theorem 1 also allows us to establish convergence of the average iterates to a specific level of accuracy dependent on the total number of iterations T, as we state next.

Corollary 1 *Let $\bar{\mathbf{x}}_T=(1/T)\sum_{t=1}^{T}\mathbf{x}_t$ be the vector formed by averaging the primal iterates $\mathbf{x}_t$ over times $t=1,\ldots,T$.. Under Assumptions 1 - 2, with constant algorithm step-size $\epsilon_t=1/\sqrt{T}$, the objective function evaluated at $\bar{\mathbf{x}}_T$ satisfies*

$$F(\bar{\mathbf{x}}_T)-F(\mathbf{x}^*)\leq\mathcal{O}(1/\sqrt{T}) \tag{15}$$

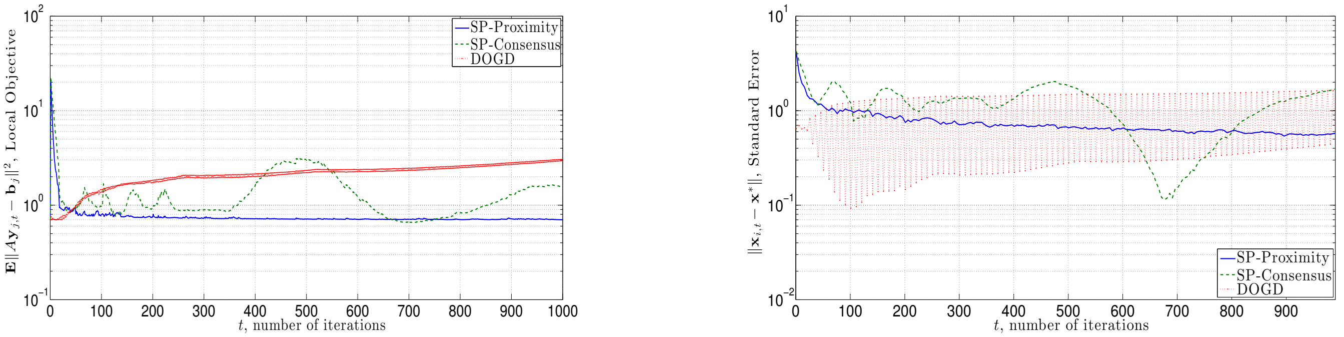
Further, the constraint violation of the average vector $\bar{\mathbf{x}}_T$ satisfies

$$\sum_{(i,j)\in\mathcal{E}}\Big[h_{ij}(\bar{\mathbf{x}}_{i,T},\bar{\mathbf{x}}_{j,T})-\gamma_{ij}\Big]_+=\mathcal{O}(T^{-\frac{1}{4}}). \tag{16}$$

Corollary 1 shows that the average saddle point primal iterates $\bar{\mathbf{x}}_t$ converge to within a margin $\mathcal{O}(1/\sqrt{T})$ in terms of objective function evaluation to the optimal objective $F(\mathbf{x}^*)$, where T is the number of iterations. Moreover, the primal average vector also yields the bound on the network proximity constraint violation as $\mathcal{O}(T^{-1/4})$. To summarize, when a constant step-size is used, with increasing T we have the following relations: the instantaneous iterates converge on average, whereas the average iterates converge to the optimal objective. This pattern also holds in terms of the algorithm's feasibility – the time average constraint violation approaches null with increasing T, whereas the constraint violation of the average vector approaches null.

5. SOURCE LOCALIZATION

We now consider the use of the stochastic saddle point method given in (6) - (7) to solve an online source localization problem. In particular, we consider an array of N sensors, where $\mathbf{l}_i\in\mathbb{R}^p$ denotes the position of the sensor i in some deployed environment $\mathcal{A}\subset\mathbb{R}^p$. Each node, assuming it is aware of its location $\mathbf{l}_i$, seeks to locate a source signal $\mathbf{x}\in\mathbb{R}^p$ through its access to noisy range observations of the form $r_{i,t}=\|\mathbf{x}-\mathbf{l}_i\|+\varepsilon_{i,t}$ where $\varepsilon_t=[\varepsilon_{1,t};\cdots;\varepsilon_{N,t}]$ is some unknown noise vector. Range-based source localization has been studied in a variety of fields, from wireless communications to geophysics [24, 25]. We consider the squared range-based least squares (SR-LS) problem $\mathbf{x}^*:=\operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^p}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{r}_i}\left(\|\mathbf{l}_i-\mathbf{x}\|^2-r_i^2\right)^2$. This problem is non-convex, but may be approximated as a quadratic program (see [26], Section II-B) by expanding the square in the first term in the previously stated objective and modifying the argument inside the expectation $(\alpha-2\mathbf{l}_i^T\mathbf{x}+\|\mathbf{l}_i\|^2-r_i^2)^2$ with the constraint $\|\mathbf{x}\|=\alpha$. We relax this constraint and consider the change of variables $\mathbf{A}_i=[-2\mathbf{l}_i^T;1]$, vector $\mathbf{b}\in\mathbb{R}^N$ with ith entry as



(a) Local objective vs. iteration t

(b) Standard error $\|\mathbf{x}_{i,t}-\mathbf{x}^*\|$ vs. iteration t

Fig. 1: Comparison of proximity and consensus algorithms on the source localization problem [cf.(17)] for an $N = 64$ node grid network deployed as an 8×8 square in a 1000×1000 meter region for the case that the noise perturbing observations by node i is zero-mean Gaussian, with a variance proportional its distance to the source as $\sigma^2 = 2\|\mathbf{l}_i - \mathbf{x}^*\|$, where $\mathbf{l}_i$ is the location of node i. We run the proximity-constrained saddle point method [cf (19), the saddle point method with consensus constraints (3), and Distributed Online Gradient Descent (DOGD) [23], a weighted averaging gradient consensus scheme. The proximity-constrained saddle point method performs best in terms of objective convergence and standard error.

$\mathbf{b}_i = r_i^2 - \|\mathbf{l}_i\|^2$, and $\mathbf{y} = [\mathbf{x}; \alpha] \in \mathbb{R}^{p+1}$, which allows the problem to be approximated as

$$\mathbf{y}^* := \underset{\mathbf{y}\in\mathbb{R}^{p+1}}{\text{argmin}} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{b}_i}\left(\|\mathbf{A}_i\mathbf{y} - \mathbf{b}_i\|^2\right) . \tag{17}$$

We solve (17) in decentralized settings, in which case each sensor keeps a local copy $\mathbf{y}_i$ of the global source estimate $\mathbf{y}$ based on information that is available with local information only and via message exchange with neighboring sensors.

Frequently in application settings, measurement quality is better for sensors nearer to the source. Thus, we consider a setup where the noise variance perturbing the range measurements depends on the sensor to source distance. Further, each sensor i weights the importance of neighboring sensors $j \in n_i$ by aiming to keep its estimate $\mathbf{x}_i$ within an ℓ_2 ball centered at its neighbors estimate $\mathbf{x}_j$, whose radius is given by the pairwise minimum estimated distance to the source $\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \min\{\|\mathbf{x}_i - \mathbf{l}_i\|^2, \|\mathbf{x}_j - \mathbf{l}_j\|^2\}$ for all $j \in n_i$ which, while nonconvex, may be convexified by rearranging the terms and replacing the resulting max by the log-sum-exp function which yields

$$\begin{aligned} &\min_{\mathbf{y}\in\mathbb{R}^{N(p+1)}} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{b}_i}\left(\|\mathbf{A}_i\mathbf{y}_i - \mathbf{b}_i\|^2\right) , && (18)\\ &\text{s.t.} \qquad (1/2)\Big(\|\mathbf{y}_i - \mathbf{y}_j\|^2 + \log\Big(e^{\|\mathbf{y}_i - \mathbf{l}_i\|^2} + e^{\|\mathbf{y}_j - \mathbf{l}_j\|^2}\Big)\Big) \leq 0, \end{aligned}$$

where the constraint for node i is with respect to all of its neighbors $j \in n_i$. The problem in (18) is of the form (4). Define $g(\mathbf{y}_i, \mathbf{y}_j)$ as the constraint function the left-hand side of the inequality in (18). Then the local primal update (6) is given as

$$\begin{aligned} \mathbf{y}_{i,t+1} = \mathbf{y}_{i,t} - \epsilon_t\Big(&2\mathbf{A}_{i,t}^T(\mathbf{A}_{i,t}\mathbf{y}_{i,t} - \mathbf{b}_{i,t}) && (19)\\ &+ \sum_{j\in n_i} \boldsymbol{\lambda}_{ij,t}\Big(\frac{e^{\|\mathbf{y}_{i,t}-\mathbf{l}_i\|^2}(\mathbf{y}_{i,t} - \mathbf{l}_i)}{e^{\|\mathbf{y}_{i,t}-\mathbf{l}_i\|^2} + e^{\|\mathbf{y}_{j,t}-\mathbf{l}_j\|^2}} + (\mathbf{y}_{i,t} - \mathbf{y}_{j,t})\Big), \end{aligned}$$

The dual update [cf. (7)] at edge (i, j) takes the form

$$\boldsymbol{\lambda}_{ij,t+1} = \Big[(1 - \delta\epsilon_t)\boldsymbol{\lambda}_{ij,t} + \epsilon_t g(\mathbf{y}_{i,t}, \mathbf{y}_{j,t})\Big]_+ . \tag{20}$$

We analyze the performance of the saddle point updates (19) - (20) to solve localization problems in a decentralized manner, such that nodes more strongly weight the importance of sensors in closer proximity to the source in the sense of $g(\mathbf{y}_i, \mathbf{y}_j)$. We consider the local objective $\mathbb{E}_{\mathbf{b}_i}\|\mathbf{A}_i\mathbf{y}_i - \mathbf{b}_i\|^2$, as well as the standard error $\|\mathbf{x}_{i,t} - \mathbf{x}^*\|$ to the source signal $\mathbf{x}^*$, where we recover $\mathbf{x}_{i,t}$ from $\mathbf{y}_{i,t}$ by taking its first p elements. We further consider the magnitude of the constraint violation for this problem, which when considering the proximity constrained problem in (18), is given by $\sum_{j\in n_i}(1/2)g(\mathbf{y}_{i,t}, \mathbf{y}_{j,t}) = \sum_{j\in n_i}(1/2)(\|\mathbf{y}_{i,t} - \mathbf{y}_{j,t}\|^2 + \log(e^{\|\mathbf{y}_{i,t}-\mathbf{l}_{i,t}\|^2} + e^{\|\mathbf{y}_{j,t}-\mathbf{l}_j\|^2}))$, and when implementing consensus methods, is given by $\sum_{j\in n_i} h(\mathbf{y}_{i,t}, \mathbf{y}_{j,t}) = \sum_{j\in n_i}\|\mathbf{y}_{i,t} - \mathbf{y}_{j,t}\|$ for a randomly chosen sensor in the network.

We compare the saddle point method on a proximity constrained problem [cf. (19) - (20)] as compared with consensus methods. We compare the saddle point method (with proximity as well as consensus constraints) to distributed online gradient descent (DOGD) [23], a weighted averaging scheme for implementing consensus constraints.We consider problem instances when the number $N = 64$ of sensors is fixed, and are spatially deployed in a grid formation as a 8×8 square in a planar ($p = 2$) region of size 1000×1000. Observation noise at node i is zero-mean Gaussian, with a variance proportional its distance to the source as $\sigma^2 = 2\|\mathbf{l}_i - \mathbf{x}^*\|$, where $\mathbf{l}_i$ is the location of node i, and the true source signal $\mathbf{x}^*$ is located at the average location of the sensors. For the saddle point methods, we find a hybrid step-size strategy to be most effective, and hence set $\epsilon_t = \min(\epsilon, \epsilon t_0/t)$ with $t_0 = 100$ and $\epsilon = 10^{-1.5}$. For DOGD, we find best performance to correspond to using a constant outer step-size $\epsilon = 10^{-1.5}$, along with a halving scheme step-size in the inner averaging loop [23].

We plot the results of this problem instance in Figure 1 for an arbitrarily chosen sensor $i \in V$. The saddle point method with proximity constraints method yields the best performance in terms of objective convergence (Figure 1a) and standard error (Figure 1b) whereas SP-Consensus and DOGD respectively experience numerical oscillations and divergent behavior after a burn-in period of $t = 100$.

6. REFERENCES

[1] A. Koppel, B. M. Sadler, and A. Ribeiro, “Proximity without consensus in online multi-agent optimization,” in *Proc. Int. Conf. Accoustics Speech Signal Process.*, Shanghai China, March 20-25 2016.

[2] A. Koppel, B. M. Sadler, and A. Ribeiro, “Proximity without consensus in online multi-agent optimization,” *Submitted to IEEE Trans. Signal Process, ArXiv preprint 1606.05578*, June. 2016.

[3] Y. Cao, W. Yu, W. Ren, and G. Chen, “An overview of recent progress in the study of distributed multi-agent coordination,” *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 1, pp. 427–438, 2013.

[4] C. Lopes and A. Sayed, “Diffusion least-mean squares over adaptive networks: Formulation and performance analysis,” *Signal Processing, IEEE Transactions on*, vol. 56, pp. 3122–3136, July 2008.

[5] A. Jadbabaie, J. Lin, *et al.*, “Coordination of groups of mobile autonomous agents using nearest neighbor rules,” *Automatic Control, IEEE Transactions on*, vol. 48, no. 6, pp. 988–1001, 2003.

[6] A. Ribeiro, “Optimal resource allocation in wireless communication and networking,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–19, 2012.

[7] A. Ribeiro, “Ergodic stochastic optimization algorithms for wireless communication and networking,” *Signal Processing, IEEE Transactions on*, vol. 58, pp. 6369–6386, Dec 2010.

[8] M. Rabbat and R. Nowak, “Decentralized source localization and tracking [wireless sensor networks],” in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP ’04). IEEE International Conference on*, vol. 3, pp. iii–921–4 vol.3, May 2004.

[9] I. Schizas, A. Ribeiro, and G. Giannakis, “Consensus in ad hoc wsns with noisy links - part i: distributed estimation of deterministic signals,” *IEEE Trans. Signal Process.*, vol. 56, pp. 350–364, Jan. 2008.

[10] D. Jakovetic, J. M. F. Xavier, and J. M. F. Moura, “Fast distributed gradient methods,” *CoRR*, vol. abs/1112.2972, Apr. 2011.

[11] S. Ram, A. Nedic, and V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” *J Optimiz. Theory App.*, vol. 147, pp. 516–545, Sept. 2010.

[12] K. Yuan, Q. Ling, and W. Yin, “On the convergence of decentralized gradient descent,” *ArXiv e-prints 1310.7063*, Oct. 2013.

[13] M. Rabbat, R. Nowak, and J. Bucklew, “Generalized consensus computation in networked systems with erasure links,” in *IEEE 6th Workshop Signal Process. Adv. in Wireless Commun Process.*, pp. 1088–1092, Jun. 5-8 2005.

[14] F. Jakubiec and A. Ribeiro, “D-map: Distributed maximum a posteriori probability estimation of dynamic systems,” *IEEE Trans. Signal Process.*, vol. 61, pp. 450–466, Feb. 2013.

[15] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*, vol. II of *Stanford Mathematical Studies in the Social Sciences*. Stanford University Press, Stanford, Dec. 1958.

[16] A. Nedic and A. Ozdaglar, “Subgradient methods for saddle-point problems,” *J Optimiz. Theory App.*, vol. 142, pp. 205–228, Aug. 2009.

[17] A. Koppel, F. Jakubiec, and A. Ribeiro, “A saddle point algorithm for networked online convex optimization,” *IEEE Trans. Signal Process.*, p. 15, Oct 2015.

[18] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, “D4l: Decentralized dynamic discriminative dictionary learning,” *IEEE Trans. Signal Process.*, vol. (submitted), July 2015. Available at http://www.seas.upenn.edu/ aribeiro/wiki.

[19] Q. Ling and A. Ribeiro, “Decentralized dynamic optimization through the alternating direction method of multipliers,” *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1185–1197, 2014.

[20] H. Robbins and S. Monro, “A stochastic approximation method,” *Ann. Math. Statist.*, vol. 22, pp. 400–407, 09 1951.

[21] J. Tsitsiklis, D. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, 1986.

[22] M. Mahdavi, R. Jin, and T. Yang, “Trading regret for efficiency: online convex optimization with long term constraints,” *Journal of Machine Learning Research*, vol. 13, no. Sep, pp. 2503–2528, 2012.

[23] K. I. Tsianos and M. G. Rabbat, “Distributed strongly convex optimization,” *CoRR*, vol. abs/1207.3031, July 2012.

[24] R. Kozick and B. Sadler, “Accuracy of source localization based on squared-range least squares (sr-ls) criterion,” in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2009 3rd IEEE International Workshop on*, pp. 37–40, Dec 2009.

[25] P. Roux, M. Corciulo, M. Campillo, and D. Dubuq, “Source localization analysis using seismic noise data acquired in exploration geophysics,” *AGU Fall Meeting Abstracts*, p. C2249, Dec. 2011.

[26] A. Beck, P. Stoica, and J. Li, “Exact and approximate solutions of source localization problems,” *Signal Processing, IEEE Transactions on*, vol. 56, pp. 1770–1778, May 2008.