PROXIMITY WITHOUT CONSENSUS IN ONLINE MULTI-AGENT OPTIMIZATION

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

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

ABSTRACT

We consider stochastic optimization problems in multi-agent settings, where a network of agents aims to learn decision variables which are optimal in terms of a global objective, while giving preference to locally and sequentially observed information. To do so, we formulate a problem where each agent minimizes a global objective while enforcing network proximity constraints, which includes consensus optimization as a special case. We propose a stochastic variant of the saddle point algorithm proposed by Arrow and Hurwicz to solve it, which yields a decentralized algorithm that is shown to asymptotically converge to a primal-dual optimal pair of the problem in expectation when a diminishing algorithm step-size is chosen. Moreover, the algorithm converges linearly to a neighborhood when a constant step-size is chosen. We apply this method to the problem of sequentially estimating a correlated random field in a sensor network, which corroborates these performance guarantees.

1. INTRODUCTION

We consider online multi-agent optimization problems, where a group of interconnected 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. We consider 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 possibly distinct local signals.

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. To do so, various information mixing strategies among the nodes have been proposed in which agents combine local gradient steps with a weighted average of their neighbors variables [1–3], dual reformulations where each agent ascends in the dual domain [4, 5], and primal-dual methods which combine primal descent with dual ascent [6–11]. Stochastic approximation methods have successfully generalized these methods to the online setting [2, 12–14].

In distributed optimization problems, agent agreement may not always be the primary goal. In large-scale settings where one aims to leverage parallel processing architectures to alleviate computational bottlenecks, agreement constraints are suitable. In contrast, if there are different priors on information received at distinct subsets of agents, then requiring the network to reach a common decision may to degrade local predictive accuracy. Moreover, there are tradeoffs in complexity and communications, and it may be that only a subset of nodes requires a solution. In this paper, we seek to solve problems in which each agent aims to minimize a global cost $\sum_i f_i$ subject to a network proximity constraint, which 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 propose a stochastic saddle point method [6, 7] to solve online multi-agent optimization problems with network proximity constraints. Moreover, we establish that this algorithm converges in expectation to a primal-dual optimal pair of this problem when a diminishing step-size is used, and to a neighborhood of the saddle point of the Lagrangian when a constant step-size is used (Section 4). All proofs are given in [15]. Numerical analysis on a spatially correlated sequential estimation problem in a sensor network demonstrates the proposed method's practical utility (Section 5).

Thanks to NSF CCF-1017454, NSF CCF-0952867, & ONR N00014-12-1-0997, ASEE SMART.

2. PROBLEM FORMULATION

Begin by considering 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 convex loss function $f_i : \mathcal{X}_i \times \Theta_i \to \mathbb{R}$ that is parameterized by a decision variable $\mathbf{x}_i \in \mathcal{X}_i \subset \mathbb{R}^p$ and a random variable $\boldsymbol{\theta}_i \in \Theta_i$ with a proper distribution. Throughout, we assume $\mathcal{X}_i$ to be compact and convex and the functions $f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)$ to be m-strongly convex in $\mathbf{x}_i$ for almost all given $\boldsymbol{\theta}_i$. 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,

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

In the online settings considered here the functions $f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)$ are termed instantaneous because they are observed at particular points in time; see Section 3. Moreover, $F_i(\mathbf{x}_i)$ is said to be an average function.

When we consider the network as a whole we can define the stacked vector $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, the product set $\mathcal{X} = \mathcal{X}_1 \times, \ldots \times \mathcal{X}_N$, 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

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

That (1) and (2) describe the same problem is true because 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 problems work on the hypothesis that all agents observe the same parameter and modify (2) by introducing consensus constraints of the form

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

For a connected network this constraint makes all variables $\mathbf{x}_i$ equal – hence the definition as a consensus problem. This is overly restrictive, however. 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 of the form $h_i(\mathbf{x}_i, \mathbf{x}_j)$ and a tolerance γ_{ij}. These are used to couple the decisions of agent i to those of its neighbors $j \in n_i$ through the definition of

the optimal estimates as the solution of the constrained optimization problem

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

The consensus constraints in (3) are a particular example of a proximity function $h_i(\mathbf{x}_i,\mathbf{x}_j)$ but so is the norm constraint $\|\mathbf{x}_i - \mathbf{x}_j\| \leq \gamma_{ij}$. 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 (relevant) information of neighboring nodes without the detrimental effect of trying to incorporate the information of far away nodes that is only weakly correlated with the parameter that i tries to estimate. An important observation here is that the workhorse distributed gradient descent [1–3] and dual decomposition methods [4, 5] can't be used to solve (4) because they require the constraints $h_i(\mathbf{x}_i,\mathbf{x}_j)$ to be linear.

The goal of this paper is to solve (4) in distributed online settings where nodes don't know the distribution of the random variable $\boldsymbol{\theta}_i$ but observe local instantaneous functions $f_i(\mathbf{x}_i,\boldsymbol{\theta}_i)$ sequentially. Before developing this algorithm, we discuss a representative example to clarify ideas.

Example 1 (LMMSE Estimation of a Random Field) Consider a sequential estimation problem in a sensor network, where observations $\boldsymbol{\theta}_{i,t} \in \mathbb{R}^q$ of a Gauss-Markov Random Field are collected by agent i at time t. Observations at node i are noisy linear transformations $\boldsymbol{\theta}_{i,t} = \mathbf{H}_i\mathbf{x}_i + \mathbf{w}_{i,t}$ of a signal $\mathbf{x}_i \in \mathbb{R}^p$ contaminated with Gaussian noise $\mathbf{w}_{i,t} \sim \mathcal{N}(0,\sigma^2\mathbf{I})$ independently distributed across nodes and time. The random field couples the variables $\mathbf{x}_i$ of different nodes. We capture this correlation with a covariance matrix $\mathbf{R}_x$ whose elements are assumed to decay with the distance between sensors i and j. If the communication graph between sensors is also assumed to be given by proximity, this means that estimates of neighboring nodes are more strongly correlated than estimates of nonadjacent agents.

Ignoring neighboring observations, the minimum mean square error local estimation problem at node i can then be written in the form of (1) with $f_i(\mathbf{x}_i,\boldsymbol{\theta}_i) = \|\mathbf{H}_i\mathbf{x}_i - \boldsymbol{\theta}_i\|^2$. The quality of these estimates can be improved using the correlated information of adjacent nodes but would be hurt by trying to make estimates uniformly equal across the network. The mathematical formulation

$$\mathbf{x}^* := \underset{\mathbf{x}\in\mathcal{X}^N}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_i}\left[\|\mathbf{H}_{i,t}\mathbf{x}_i - \boldsymbol{\theta}_{i,t}\|^2\right] \quad \text{s.t.} \quad (1/2)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \gamma_{ij}, \quad \text{for all } j \in n_i. \tag{5}$$

captures this specification as it makes the estimate $\mathbf{x}_i^*$ of node i close to the estimates $\mathbf{x}_j^*$ of neighboring nodes $j \in n_i$ but not so close to the estimates $\mathbf{x}_k^*$ of nonadjacent nodes $k \notin n_i$.

3. ALGORITHM DEVELOPMENT

Our goal is to solve (4) in a decentralized online manner. One way to achieve this would be to solve it by enforcing the constraints exactly, but doing so would require global coordination. Instead, we consider the Lagrangian relaxation of (4), stated as

$$\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i,\boldsymbol{\theta}_i)] + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\in n_i} \boldsymbol{\lambda}_{ij}\left(h(\mathbf{x}_i,\mathbf{x}_j) - \gamma_{ij}\right) \tag{6}$$

We propose applying a stochastic saddle point algorithm to (6) which operates by alternating primal and dual stochastic gradient descent and ascent steps respectively [6]. To do so, consider the stochastic approximation of the Lagrangian evaluated at observed realizations $\boldsymbol{\theta}_{i,t}$ of the random variables $\boldsymbol{\theta}_i$, which we define as

$$\hat{\mathcal{L}}_t(\mathbf{x},\boldsymbol{\lambda}) = \sum_{i=1}^{N} f_i(\mathbf{x}_i,\boldsymbol{\theta}_{i,t}) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\in n_i} \boldsymbol{\lambda}_{ij}\left(h(\mathbf{x}_i,\mathbf{x}_j) - \gamma_{ij}\right) \tag{7}$$

We allow for the case that the proximity constants $\gamma_{ij} = \gamma_{ij,t}$ are time-varying as well, and selected according to some distributional inference on $\boldsymbol{\theta}_i$.

Define the stacked primal and dual variables respectively as $\mathbf{x} := [\mathbf{x}_1;\cdots;\mathbf{x}_N] \in \mathbb{R}^{Np}$ and $\boldsymbol{\lambda} := [\boldsymbol{\lambda}_1;\cdots;\boldsymbol{\lambda}_M] \in \mathbb{R}^{Mp}$. Moreover, denote the network aggregate random variable as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1;\cdots;\boldsymbol{\theta}_N]$. Particularized to the stochastic Lagrangian stated in (7), the stochastic saddle point method takes the form

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{X}}\left[\mathbf{x}_t - \epsilon_t \nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t,\boldsymbol{\lambda}_t)\right], \tag{8}$$

$$\boldsymbol{\lambda}_{t+1} = \mathcal{P}_{\Lambda}\left[\boldsymbol{\lambda}_t + \epsilon_t \nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}(\mathbf{x}_{t+1},\boldsymbol{\lambda}_t)\right], \tag{9}$$

where $\nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t,\boldsymbol{\lambda}_t)$ and $\nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}(\mathbf{x}_t,\boldsymbol{\lambda}_t)$, are the primal and dual stochastic gradients of the Lagrangian with respect to $\mathbf{x}$ and $\boldsymbol{\lambda}$, respectively. These stochastic subgradients are approximations of the gradients of (6) evaluated at the current realization of the random variable $\boldsymbol{\theta}$. The notation $\mathcal{P}_{\Lambda}(\boldsymbol{\lambda})$ denotes projection of dual variables on a given convex compact set Λ, which we assume be written as a Cartesian product of sets Λ_{ij} so that the projection of $\boldsymbol{\lambda}$ into Λ is equivalent to the separate projection of the components $\boldsymbol{\lambda}_{ij}$ into the sets Λ_{ij}. The notation $\mathcal{P}_{\mathcal{X}}(\mathbf{x})$ denotes projection onto the set of feasible primal variables $\mathcal{X} = \mathcal{X}_1\times,\ldots\times\mathcal{X}_N$.

The method stated in (8) - (9) yields an effective tool for decentralized computation across the network, as we state in the following proposition.

Proposition 1 *The gradient computations in* (8)-(9) *may be separated along the local primal variables* $\mathbf{x}_{i,t}$ *associated with node* i*, yielding* N *parallel updates*

$$\mathbf{x}_{i,t+1} = \mathcal{P}_{\mathcal{X}_i}\Big[\mathbf{x}_{i,t} - \epsilon_t\Big(\nabla_{\mathbf{x}_i} f_i(\mathbf{x}_{i,t};\boldsymbol{\theta}_{i,t}) + \sum_{j\in n_i} \boldsymbol{\lambda}_{ij,t}^T \nabla_{\mathbf{x}_i} h_i(\mathbf{x}_{i,t},\mathbf{x}_{j,t})\Big)\Big], \tag{10}$$

Moreover, the dual gradients in the update of $\boldsymbol{\lambda}_{ij,t}$ *in* (9) *may be separated into* M *parallel updates associated with edge* (i,j)

$$\boldsymbol{\lambda}_{ij,t+1} = \mathcal{P}_{\Lambda}\left[\boldsymbol{\lambda}_{ij,t} + \epsilon_t\left(h_i(\mathbf{x}_{i,t+1},\mathbf{x}_{j,t+1}) - \gamma_{ij}\right)\right]. \tag{11}$$

which allows for distributed computation across the network.

Proof: See [15], Appendix A. □

Proposition 1 states that the saddle point method requires that individuals only coordinate their decision variables with their neighbors, and hence may be implemented in a distributed manner. Observe that the constraint functions h_i may be selected as any convex function of the decision variables of node i and its neighbors $j \in n_i$. Moreover, the dual variable ascends along the local constraint slack given by the network proximity of node i to its neighbors $j \in n_i$. For the case presented in Example 1 with quadratic constraints $(1/2)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \gamma_{ij}$, the update in (10) take the form

$$\mathbf{x}_{i,t+1} = \mathcal{P}_{\mathcal{X}_i}\Big[\mathbf{x}_{i,t} - \epsilon\Big(2\mathbf{H}_{i,t}^T\left(\mathbf{H}_{i,t}\mathbf{x}_{i,t} - \boldsymbol{\theta}_{i,t}\right) + \sum_{j\in n_i}\boldsymbol{\lambda}_{ij,t}\Big(\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\Big)\Big]. \tag{12}$$

Moreover, the dual update which ascends along the constraint violation, for the quadratic constraint case, may be stated as

$$\boldsymbol{\lambda}_{ij,t+1} = \mathcal{P}_{\Lambda}\left[\boldsymbol{\lambda}_{ij,t} + \epsilon_t \left((1/2)\|\mathbf{x}_{i,t+1} - \mathbf{x}_{j,t+1}\|^2 - \gamma_{ij}\right)\right] . \quad (13)$$

The factorization properties of the Lagrangian [cf. (6)] allow for distributed computation, from which new decentralized estimation schemes may be derived, as with the updates in (12) - (13).

4. CONVERGENCE ANALYSIS

We turn to establishing that the saddle point algorithm in (8)-(9) asymptotically converges to a saddle point of the Lagrangian [cf. (6)], which implies that we solve the problem stated in (4) in a decentralized online manner. The analysis is done and the results are stated in terms of projected gradients. Consider then the feasible primal set $\mathcal{X}$ and define the projected primal stochastic gradient of the Lagrangian onto the set $\mathcal{X}$ as

$$\mathcal{P}_{\mathcal{X}}[\nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t)] = \begin{cases} \nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t) & \text{if } \nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t) \in \mathcal{X}^{\circ} \\ \nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t)^{\|} & \text{if } \nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t) \in \partial\mathcal{X}. \end{cases} \quad (14)$$

where $\nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t)^{\|}$ denotes the component of the vector parallel to the set $\mathcal{X}$, whose interiors and boundaries are denoted as $\mathcal{X}^{\circ}$ and $\partial\mathcal{X}$ respectively. The projected dual stochastic gradient $\mathcal{P}_{\Lambda}[\nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t)]$ is similarly defined.

In order to obtain convergence results, some conditions are required of the network, data distribution, loss functions, and stochastic approximation errors which we state below.

(A1) The Lagrangian has Lipschitz gradients in the primal and dual variables, i.e. the following holds

$$\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) - \nabla_{\mathbf{x}}\mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda})\| \leq L_{\mathbf{x}}\|\mathbf{x} - \tilde{\mathbf{x}}\| . \quad (15)$$

$$\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}, \tilde{\boldsymbol{\lambda}})\| \leq L_{\boldsymbol{\lambda}}\|\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}\| . \quad (16)$$

for distinct primal and dual variables $\mathbf{x} \neq \tilde{\mathbf{x}}$, $\boldsymbol{\lambda} \neq \tilde{\boldsymbol{\lambda}}$.

(A2) The gradients of the Lagrangian are bounded as

$$\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})\| \leq G_{\mathbf{x}} , \quad \|\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})\| \leq G_{\boldsymbol{\lambda}} . \quad (17)$$

(A3) There exists a constant A so that for all times t the norm of the difference between descent (respectively, ascent) along stochastic gradients followed by set projections and descent (ascent) along projected stochastic gradients are almost surely bounded by $A\epsilon_t^2$, i.e.

$$\left\|(\mathbf{x}_{t+1} - \mathbf{x}_t) - \epsilon_t\mathcal{P}_{\mathcal{X}}[\nabla_{\mathbf{x}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t)]\right\| \leq A\epsilon_t^2 \quad (18)$$

$$\left\|(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - \epsilon_t\mathcal{P}_{\Lambda}[\nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}(\mathbf{x}_t, \boldsymbol{\lambda}_t)]\right\| \leq A\epsilon_t^2 \quad (19)$$

(A4) The projected primal and dual stochastic gradients of the Lagrangian are unbiased estimators of the projected gradients of the Lagrangian, i.e.

$$\mathbb{E}\left[\mathcal{P}_{\mathcal{X}}[\nabla_{\mathbf{x}}\hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\Lambda}_t)]\right] = \mathcal{P}_{\mathcal{X}}[\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\Lambda}_t)] , \quad (20)$$

$$\mathbb{E}\left[\mathcal{P}_{\Lambda}[\nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\Lambda}_t)]\right] = \mathcal{P}_{\Lambda}[\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\Lambda}_t)] . \quad (21)$$

The second moments of the stochastic gradients conditional on $\mathcal{F}_t$, a sigma algebra that measures the history of the system up until time t, are bounded by S_x^2 for all times t, which allows us to write

$$\max\{\mathbb{E}[\|\nabla_{\mathbf{x}}\hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)\|^2 \,|\, \mathcal{F}_t], \mathbb{E}[\|\nabla_{\boldsymbol{\lambda}}\hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)\|^2 \,|\, \mathcal{F}_t]\} \leq S_x^2. \quad (22)$$

Assumption 1 requires the gradients to be uniformly well behaved and Assumption 2 requires them to be bounded. Since the primal and dual domains $\mathcal{X}$ and Λ are compact, both of these assumptions are valid in most practical situations. The condition in Assumption 3 is a technical condition which permits analyzing the algorithm in terms of projected stochastic gradients rather than an analysis that uses regular (non-projected) gradients followed by set projections. This technical assumption is not restrictive either. It just means that the difference between applying a projected gradient and applying a regular gradient followed by projection scales with the stepsize. The bias and variance conditions in (22) of Assumption 4 is standard in stochastic optimization literature and valid in all but pathological cases.

With these conditions in place, we are ready to study the convergence properties of the algorithm stated in (8) - (9). We first consider the case where the algorithm step-size ϵ_t is diminishing, i.e. $\epsilon_t = O(1/t)$, which implies

$$\text{(i) } \sum_{t=0}^{\infty} \epsilon_t = \infty \text{ and (ii) } \sum_{t=0}^{\infty} \epsilon_t^2 < \infty \quad (23)$$

When the algorithm step-size satisfies (23), the saddle point iterates converge in expectation to a primal-dual optimal pair of (4).

Theorem 1 *Denote $(\mathbf{x}_t, \boldsymbol{\lambda}_t)$ as the sequence generated by the saddle point algorithm in* (8)-(9)*. Suppose Assumptions 1 - 4 hold and the step-size $\epsilon_t = O(1/t)$, then $(\mathbf{x}_t, \boldsymbol{\lambda}_t)$ asymptotically converges in expectation to a KKT point of the problem as*

$$\lim_{t\to\infty} \mathbb{E}[\|\mathcal{P}_{\mathcal{X}}[\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t)]\|] = 0 , \quad (24)$$

$$\lim_{t\to\infty} \mathbb{E}[\|\mathcal{P}_{\Lambda}[\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t)]\|] = 0 . \quad (25)$$

Proof: See [15], Section IV. □

Theorem 1 guarantees that the saddle point algorithm stated in (8) - (9) asymptotically converges in expectation to a primal-dual optimal pair of the problem stated in (4) when a diminishing step-size $\epsilon_t = O(1/t)$ is used. As a consequence, individuals in the network successfully learn global information while satisfying the network proximity constraint on average. Instead, if we select a constant step-size $\epsilon_t = \epsilon$ then the algorithm in (8) and (9) converges asymptotically to a neighborhood of the optimal. Moreover, with a suitably small constant step-size, the convergence rate is linear, as we state next.

Theorem 2 *Denote $(\mathbf{x}_t, \boldsymbol{\lambda}_t)$ as the sequence generated by the saddle point algorithm in* (8) *and* (9) *with constant step-size $\epsilon_t = \epsilon < 1/(2m)$. Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{X}^* \times \boldsymbol{\Lambda}^*$ be a primal-dual optimal pair of the problem stated in* (4)*. If Assumptions 1 - 4 hold, the Lagrangian $\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t)$ converges to a neighborhood of the saddle point $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ as*

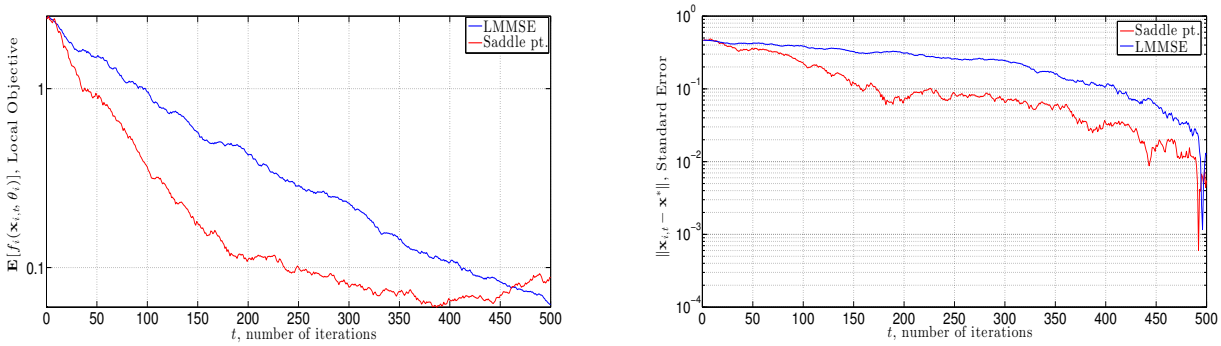
$$\liminf_{t\to\infty} |\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t) - \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)| \leq \frac{\epsilon S_x^2 L_{\boldsymbol{\lambda}} + 2G_{\boldsymbol{\lambda}}^2}{4m} . \quad (26)$$

Moreover, the absolute error sequence of the Lagrangian $\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t) - \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ converges linearly to a neighborhood

$$\mathbb{E}[|\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t) - \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)|] \leq (1 - 2m\epsilon)^t |\mathcal{L}(\mathbf{x}_0, \boldsymbol{\lambda}_0) - \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)| + \frac{\epsilon S_x^2 L_{\mathbf{x}} + 2G_{\boldsymbol{\lambda}}^2}{4m} . \quad (27)$$

Proof: See [15], Section IV. □

Theorem 2 guarantees that the saddle point method as stated in (8) and (9) converge linearly to a neighborhood of the Lagrangian evaluated at a primal-dual optimal pair. The constant $(\epsilon L_{\mathbf{x}} S_x^2 + 2G_{\boldsymbol{\lambda}}^2)/4m$ is dominated by the constant $G_{\boldsymbol{\lambda}}$ which represents the worst-case constraint slack. If we replace $\mathbf{x}_t$ by its time average $\bar{\mathbf{x}}_t = (1/t)\sum_{u=1}^{t} \mathbf{x}_t$, this quantity asymptotically behaves as $O(\epsilon^2)$ in deterministic settings, as is established in Proposition 5.1 (a) in [7].



(a) Local objective vs. iteration t

(b) Standard error vs. iteration t

Fig. 1: Saddle point algorithm applied to the problem of estimating a correlated random field. Nodes are deployed uniformly in a square region of size 200×200 squared meters in a grid formation, and node estimators are correlated according to the distance-based model $\rho(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|l_i - l_j\|}$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the decisions of nodes i and j, and l_j are their respective locations. Individual sensors learn global information while remaining close to nodes whose information they deem important. Exploiting the correlation structure of the field yields a reduction in the local estimation error and distance to the optimal LMMSE estimate.

5. NUMERICAL ANALYSIS

Consider the task of estimating a spatially correlated random field in a specified region $\mathcal{A}$ by making use of a sensor network. Interconnected sensors collect observations $\boldsymbol{\theta}_{i,t}$ which are noisy linear transformations of the signal $\mathbf{x}$ they would like to estimate, which are related through the observation model $\boldsymbol{\theta}_{it} = \mathbf{H}_i\mathbf{x} + \mathbf{w}_{i,t}$ with Gaussian noise $\mathbf{w}_{i,t} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_q)$ i.i.d across time and node with $\sigma^2 = 2$, as in Example 1. The random field is parameterized by the correlation matrix $\mathbf{R}_x$, which is assumed to follow a spatial correlation structure of the form $\rho(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|l_i - l_j\|}$, where l_i and l_j are the respective locations of sensor i and sensor j in the deployed region, see, e.g., [16]. Observe that now each node has a unique signal-to-noise ratio based upon its location and that information received at more distant nodes are less important; however, their contribution to the aggregate objective $F(\mathbf{x})$ still incentivizes global coordination.

To solve this problem, we deploy $N = 50$ sensors in a grid formulation, where neighboring nodes have a constant apart from one another in a 1000×1000 meter square region. We make use of the saddle point algorithm [cf. (10) - (11)], whose updates for the random field estimation problem is given by the explicit expressions in (12) and (13), respectively. We select $\gamma_{ij} = \rho(\mathbf{x}_i, \mathbf{x}_j)$. Besides the local and global losses which on average asymptotically converge to their constrained minima in the diminishing step-size regime (Theorem 1) and to a neighborhood of the optima in the constant step-size case (Theorem 2), we also study the standard error to the LMSE estimator $\mathbf{x}^*$, i.e. $\|\mathbf{x}_{i,t} - \mathbf{x}^*\|$.

To compute $\mathbf{x}^*$ for a single time slot, stack observations $\boldsymbol{\theta} = [\boldsymbol{\theta}_1; \cdots; \boldsymbol{\theta}_N]$ and observation models $\mathbf{H} = [\mathbf{H}_1; \cdots; \mathbf{H}_N]$. Then the least mean squared error (LMSE) estimator for a single time slot of this problem is $\mathbf{x}^* = (\mathbf{H}\mathbf{R}_x\mathbf{H}^T + \frac{1}{\sigma^2}\mathbf{I})^{-1}\frac{1}{\sigma^2}\mathbf{I}\boldsymbol{\theta}$. To compute the benchmark LMSE $\mathbf{x}^*$ for a given run, we stack signals $\boldsymbol{\theta}_{i,t}$ for all nodes i and times t at a centralized location into one large linear system and substitute the sample variance $\hat{\sigma}^2$ in the prior computation.

We consider problem instances where observations and signal estimates are scalar ($p = q = 1$), the scalar $\mathbf{H} = 1$, and the a priori signal $\mathbf{x} = \mathbf{1}$ is set a vector of ones, and run the algorithm for for $T = 500$ iterations with a hybrid step-size strategy which is constant for the first t_0 iterations and then attenuates, i.e. $\epsilon_1 = \min(\epsilon, \epsilon t_0/t)$ with $t_0 = 100$ and $\epsilon = 10^{-2}$. The noise level is set to $\sigma^2 = 10$. We compare the performance of the algorithm with that of a simple LMMSE estimator strategy which does not take advantage of the correlation structure of the sensor network.

In Figure 1, we plot the results of this numerical estimation experiment. Figure 1a shows the local objective $\mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_{i,t}, \boldsymbol{\theta}_i)]$ of an arbitrarily chosen node $i \in V$ versus iteration t. We observe the numerical behavior of the local objective is similar to the global objective, and is thus omitted. We see that when nodes incorporate the correlation structure of the random field into their estimation strategy via the quadratic proximity constraint with γ_{ij} chosen according to the correlation of node i and its neighbors $j \in n_i$, the estimation performance improves. In particular, to achieve the benchmark $\mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)] \leq .1$, we require $t = 247$ versus $t = 411$ iterations respectively for the case that the correlation structure is exploited via the saddle point algorithm as compared with a simple LMMSE scheme. We observe that for small t the gain is substantial, but for large t the performance is comparable to the LMMSE strategy.

This improved estimation performance is corroborated in the plot of the standard error $\|\mathbf{x}_{i,t} - \mathbf{x}^*\|$ to the optimal estimator as compared with iteration t in Figure 1b. We see that to achieve the benchmark $\|\mathbf{x}_{i,t} - \mathbf{x}^*\| \leq .1$, the saddle point algorithm requires $t = 157$ iterations as compared with $t = 414$ for the LMMSE estimator, more than twice as many. We have observed that the benefit of using the saddle point method as compared with simple LMMSE is more substantial in problem instances where the signal to noise ratio is low, and the region $\mathcal{A}$ is larger.

6. CONCLUSION

We formulated online multi-agent optimization problems with network proximity constraints as a generalization of online consensus optimization, and consider the saddle point method of Arrow and Hurwicz to solve it. We establish this algorithm converges in expectation both in the diminishing and constant algorithm step-size regimes in Theorems 1 - 2, respectively. The flexibility afforded by the saddle point algorithm allows individual nodes in the network to give preference to locally observed information and consider more general agreement constraints which may take advantage of correlation structures in their decision variables.

As an application, we considered a random field estimation problem where the estimators of individual sensors follow a spatial correlation pattern. The saddle point method for this task provides a framework for nodes to incorporate priors on the importance of their neighbors' decisions for their local estimates via proximity constraints based on their correlations. In doing so, we observe empirical performance gains over a simple LMMSE estimator.

7. REFERENCES

[1] D. Jakovetic, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *Automatic Control, IEEE Transactions on*, vol. 59, no. 5, pp. 1131–1146, 2014.

[2] S. Ram, A. Nedic, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J Optimiz. Theory App.*, vol. 147, no. 3, pp. 516–545, Sep. 2010.

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

[4] 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.*, Jun. 5-8 2005, pp. 1088–1092.

[5] F. Jakubiec and A. Ribeiro, "D-map: Distributed maximum a posteriori probability estimation of dynamic systems," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 450–466, Feb. 2013.

[6] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and nonlinear programming*, ser. Stanford Mathematical Studies in the Social Sciences. Stanford University Press, Stanford, Dec. 1958, vol. II.

[7] A. Nedic and A. Ozdaglar, "Subgradient methods for saddle-point problems," *J Optimiz. Theory App.*, vol. 142, no. 1, pp. 205–228, Aug. 2009.

[8] A. Koppel, F. Y. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization." in *Proc. Int. Conf. Acoust. Speech Signal Process.*, May 4-9 2014, pp. 8292–8296.

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

[10] A. Koppel, F. Y. Jakubiec, and A. Ribeiro, "Regret bounds of a distributed saddle point algorithm." in *Proc. Int. Conf. Acoust. Speech Signal Process.*, April 19-24 2015.

[11] 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.

[12] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951.

[13] 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.

[14] 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.

[15] A. Koppel, B. M. Sadler, and A. Ribeiro, "Deviating from consensus in online multi-agent optimization," *in preparation*, 2015.

[16] M. Dong, L. Tong, and B. M. Sadler., "Information retrieval and processing in sensor networks: deterministic scheduling vs. random access," in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, June 2004, pp. 79–.