EXACT NONPARAMETRIC DECENTRALIZED ONLINE OPTIMIZATION

Hrusikesha Pradhan[*]*, Amrit Singh Bedi*[*]*, Alec Koppel*[†] *and Ketan Rajawat*[*]

[*]Department of Electrical Engineering, IIT Kanpur
[†] Computational and Information Sciences Directorate, U.S. Army Research Laboratory

ABSTRACT

We consider online learning over decentralized networks, where nodes observe unique, possibly correlated, observation stream. We focus on the case where agents learn a regression *function* that belongs to a reproducing kernel Hilbert space (RKHS). In this setting, a decentralized network aims to learn nonlinear statistical models that are optimal in terms of a global stochastic convex functional that aggregates data across the network, with only access to a local data stream. We incentivize coordination while respecting network heterogeneity through the introduction of nonlinear proximity constraints. To solve it, we propose applying a functional variant of stochastic primal-dual (Arrow-Hurwicz) method which yields a decentralized algorithm. To handle the fact that the RKHS parameterization has complexity comparable to the iteration index, we project the primal iterates onto Hilbert subspaces that are greedily constructed from the observation sequence of each node. The resulting proximal stochastic variant of Arrow-Hurwicz is shown to converge in expectation, both in terms of primal sub-optimality and constraint violation to a neighborhood that depends on a given constant step-size selection. Experiments on a correlated random field estimation problem validate our theoretical results.

***Index Terms*—** non-parametric function learning, online learning, multi-agent systems, saddle point method, random field estimation

1. INTRODUCTION

In decentralized optimization, agents of an interconnected network $\mathcal{G} = (V, \mathcal{E})$ seek to cooperate to minimize a cost that is global to the network while only communicating with their neighbors. This framework has gained much attention in distributed signal processing and networked control both for the case that agents objectives are static [1] and depend on data streams [2], i.e., the online setting. Here we focus on a generalization of the online case where agents are not necessarily of the same type, since they may represent heterogeneous devices [3, 4] as in cellular devices [5], autonomous robots [6], or social media accounts [7]. In this case, constraining all agents' decisions to be equal to one another can cause degradation in local estimation performance, whereas ignoring the behavior of neighbors would fail to exploit the relevant information overlap. This work seeks to strike a balance between these issues for the specific case that each agent's decisions are defined not by a standard parameter vector but instead a nonlinear regression function that belongs to a reproducing kernel Hilbert space (RKHS).

Setting aside the constraints, to solve stochastic programs, assuming no closed form exists, requires iterative tools. The simplest, gradient descent, requires evaluating an expectation which depends on infinitely many data realizations. This issue may be overcome through stochastic gradient descent (SGD) [8]. SGD is widely used in large-scale learning problems for this reason [9], but its limiting properties are intrinsically tied to the parameterization of the statistical model (decision variable) one chooses. For vector-valued parameterizations, i.e., linear statistical models, the convergence of SGD is well-understood [10] as a consequence of convexity.

However, the optimization problems induced by nonlinear statistical models, which have much richer descriptive capability, are more challenging. Dictionary learning [11] and deep networks [12] toss aside convexity in the interest of descriptive richness, which has led to a flurry of interest in nonconvex stochastic optimization [13]. Generally, overcoming non-convexity requires adding noise that degrades the quality of a parameter estimate [14]. Alternatively, one may preserve convexity while obtaining nonlinearity through the "kernel trick," a quirk of RKHS [15]. This fact motivates our focus on RKHS. Owing to the Representer Theorems [16], we may transform the function variable to an inner product of weights and kernel evaluations at samples. Unfortunately, the complexity of this representation is proportionate with the sample size N, which for online settings $N \to \infty$. To mitigate this bottleneck, we apply hard-thresholding projections that greedily project functions onto sparse subspaces extracted from the history of data observations, as in [17], which is shown to nearly preserve global convergence.

To incentivize coordination, multi-agent optimization constraints agents' decisions to be close to or equal to each other. Methods for solving the resulting constrained problems can be classified into primal-only via penalty method [18, 19], dual methods [20, 21] that reformulate the consensus constraint in the dual domain, and primal-dual approaches [22, 23] which alternate primal/dual descent/ascent

steps on the Lagrangian. Approximations of dual methods, i.e., ADMM, have also been used [24]. When we go beyond linear equality constraints due to the requirements of heterogeneous networks, only the approximate primal methods and exact primal-dual approaches are viable, due to the fact that dual methods and ADMM require solving a nonlinear argmin in the inner-loop per iteration for this case, which has prohibitive complexity. Hence, in this work, motivated by the fact that we seek *exact* solutions to the constrained problem, we adopt a primal-dual approach.

In this work, we consider nonlinear proximity constraints [1] which incentivize nearby agents to make decision that are close but not coincide, so as to gracefully tradeoff estimation performance on local and global data. This problem has been solved for parametric settings in [1, 25], and in the functional RKHS setting for consensus [26]. Our contribution here is to demonstrate that RKHS proximal primal-dual method [27] applies to multi-agent optimization (Sec 2), and thus yields a new principled learning methodology for collaborative learning systems (Sec 3). We demonstrate that the resulting algorithm yields both convergence in expectation in terms of primal sub-optimality and constraint violation when used with constant step-sizes (Sec 4). We validate this approach for estimating a spatially correlated random field (Sec 5).

2. DECENTRALIZED FUNCTIONAL STOCHASTIC PROGRAMMING

We consider an expected risk minimization problem of learning a function f over a network of agents. This problem arise in various applications such as robotics [28], sensor networks [29], and communication systems [30]. Let us define a symmetric, connected, and directed network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = V$ nodes, $|\mathcal{E}| = M$ edges, and $n_i := \{j : (i,j) \in \mathcal{E}\}$ as the neighborhood of agent i. At time instant t, each agent $i \in \mathcal{V}$ observes a local data realization $(\mathbf{x}_{i,t}, y_{i,t})$ from random pair $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ and aims to learn local f_i. The quality of this local estimator is quantified by an associated convex loss function $\ell_i : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The overall global objective is written as sum of local objectives as

$$f^* := \operatorname*{argmin}_{f \in \mathcal{H}} \sum_{i \in \mathcal{V}} \left(\mathbb{E}_{\mathbf{x}_i, y_i} \Big[\ell_i(f(\mathbf{x}_i), y_i) \Big] + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}} \right) \quad (1)$$

where $\mathcal{H}$ denotes RKHS, and the expectation is over the data samples $(\mathbf{x}_i, y_i)$. In (1), global loss is defined by averaging over the local loss of all the agents $i \in \mathcal{V}$.

Specifically, an RKHS is a Hilbert space $\mathcal{H}$ is equipped with a reproducing kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (see [31,32]) that satisfies

$$(i)\ \ \langle f, \kappa(\mathbf{x}_i, .) \rangle_{\mathcal{H}} = f(\mathbf{x}_i), \quad (ii)\ \ \mathcal{H} = \overline{\operatorname{span}\{\kappa(\mathbf{x}_i, .)\}} \quad (2)$$

where $\langle ., . \rangle_{\mathcal{H}}$ denotes the Hilbert inner product for $\mathcal{H}$. Moreover, it is also assumed that the kernel is positive semidefinite, i.e. $\kappa(\mathbf{x}_i, \mathbf{x}_i') \geq 0$ for all $\mathbf{x}_i, \mathbf{x}_i' \in \mathcal{X}$. Function estimation problems of this type can be reduced to a parametric form via the Representer Theorem [33, 34].

In decentralized optimization, there is no global co-ordination among all the nodes and only local communication is possible with neighbors n_i. Hence, each node try to estimate the global function f^* through the local estimation f_i. In literature [26], consensus constraint is imposed to solve the problem in (1) in a decentralized manner. But it is advocated in the recent works [1, 25] that if the agents are receiving data from disparate distributions, then forcing consensus may degrade the performance since it do not take care of the diversity of data streams. Therefore, enforcing consensus constraints results in sub-optimal solutions. This motivate us to introduce convex local proximity constraints of the form $h_{ij}(f_i, f_j) \leq \gamma_{ij}$ among the neighbors in the network, where γ_{ij} is a tolerance parameter. This modifies the (1) into

$$\begin{aligned} f^\star = &\operatorname*{argmin}_{\{f_i\} \in \mathcal{H}} \sum_{i \in \mathcal{V}} \left(\mathbb{E}_{\mathbf{x}_i, y_i} \Big[\ell_i(f_i(\mathbf{x}_i), y_i) \Big] + \frac{\lambda}{2} \|f_i\|^2_{\mathcal{H}} \right) \\ &\text{s.t. } \mathbb{E}_{\mathbf{x}_i} \Big[h_{ij}(f_i(\mathbf{x}_i), f_j(\mathbf{x}_i)) \Big] \leq \gamma_{ij}, \text{ for all } j \in n_i. \end{aligned} \quad (3)$$

We seek to solve (3) in a decentralized manner with unknown distribution of the random pair $(\mathbf{x}_i, y_i)$ but their independent samples $(\mathbf{x}_{i,n}, y_{i,n})$ are observed sequentially at each agent. For brevity, let us denote $\mathcal{H}^V$ as a product RKHS of function $f(\cdot) = [f_1(\cdot); \cdots; f_V(\cdot)]$ which is stacked version of V functions. The function f evaluated at any $\mathbf{x}$ results in a vector $f(\mathbf{x}) = [f_1(\mathbf{x}_1); \cdots; f_V(\mathbf{x}_V)] \in \mathbb{R}^V$. Additionally, let $\mathbf{x} = [\mathbf{x}_1; \cdots; \mathbf{x}_V] \in \mathcal{X}^V \subset \mathbb{R}^{Vp}$ and $\mathbf{y} = [y_1; \cdots y_V] \in \mathbb{R}^V$. Next, we shift our focus to develop an online decentralized algorithm to solve the problem in (3).

3. ALGORITHM DEVELOPMENT

To solve (3), the stochastic augmented Lagrangian function at time instant t evaluated at realization $(\mathbf{x}_{i,t}, y_{i,t})$

$$\begin{aligned} \hat{\mathcal{L}}_t(f, \boldsymbol{\mu}) := &\sum_{i \in \mathcal{V}} \Bigg[\ell_i(f_i(\mathbf{x}_{i,t}), y_{i,t}) + \frac{\lambda}{2} \|f_i\|^2_{\mathcal{H}} \\ &+ \sum_{j \in n_i} \left\{ \Big[\mu_{ij} (h_{ij}(f_i(\mathbf{x}_{i,t}), f_j(\mathbf{x}_{i,t})) - \gamma_{ij}) \Big] - \frac{\delta \eta}{2} \mu_{ij}^2 \right\} \Bigg] \end{aligned} \quad (4)$$

where $\boldsymbol{\mu}$ is a lagrange multiplier , whose entries μ_{ij} are defined for each $(i,j) \in \mathcal{E}$ from (3). Before continuing, we first need to extend the existing Representer theorem result from unconstrained to multi-agent unconstrained nonparametric optimization problem which is summarized in Corollary 1.

Corollary 1 *Let $\mathcal{H}$ be a RKHS equipped with a kernel function κ and $\mathcal{S}$ be the observations $\{\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\}_{i \in \mathcal{V}}$ stacked across the network for times $t = 1, \ldots, T$. Consider the*

sample average approximation of (3)*, and its associated Lagrangian relaxation. The each ith component of the solution to the resulting saddle-point problem can be expressed as*

$$f_i^* = \sum_{t=1}^{T} w_{i,t}\kappa(\mathbf{x}_{i,t}, .) \tag{5}$$

where $w_{i,t}$ are real-valued coefficients.

The proof of Corollary 1 is similar to the proof of [27, Theorem 1] and presented in [35]. The result in (5) allows us to write f in terms of kernel evaluations at training data points and hence shifting the search over an infinite space into one over a set of weights $\mathbf{w}_i \in \mathbb{R}^T$ which are finite. We propose to apply stochastic saddle point algorithm developed in [1] to nonparametric problem in (3) and use Corollary 1 to derive the decentralized updates of the algorithm summarized in Proposition 1.

Proposition 1 *Alternating primal/dual stochastic descent/ascent steps applied to problem* (3) *yields in the following updates*

$$f_{i,t+1} = f_{i,t}(1-\eta\lambda) - \eta\Bigg[\ell_i'(f_{i,t}(\mathbf{x}_{i,t}), y_{i,t}) + \sum_{j\in n_i} \mu_{ij} h_{ij}'(f_{i,t}(\mathbf{x}_{i,t}), f_{j,t}(\mathbf{x}_{i,t}))\Bigg]\kappa(\mathbf{x}_{i,t}, \cdot) \tag{6}$$

$$\mu_{ij,t+1} = \Big[\mu_{ij,t}(1-\delta\eta^2) + \eta\Big(h_{ij}(f_{i,t}(\mathbf{x}_{i,t}), f_{j,t}(\mathbf{x}_{i,t})) - \gamma_{ij}\Big)\Big]_+ \tag{7}$$

The detailed derivation of the updates in Proposition 1 are provided in [35]. The detailed pseudocode of how this implementation yields a decentralized method for collaborative network learning is presented in Algo.1. Additionally, we need the step size $\eta < 1/\lambda$ for regularization parameter $\lambda > 0$ in (1) and the sequence of $(f_t, \boldsymbol{\mu}_t)$ is initialized by $f_0 = 0 \in \mathcal{H}^V$ and $\boldsymbol{\mu} = 0 \in \mathbb{R}_+^M$. With this initialization, Representer theorem allows us to write function $f_{i,t}$ in terms of linear expansion of kernels evaluated at feature vectors $\mathbf{x}_{i,t}$ observed so far as

$$f_{i,t}(\mathbf{x}) = \sum_{n=1}^{t-1} w_{i,n}\kappa(\mathbf{x}_{i,n}, \mathbf{x}) = \mathbf{w}_{i,t}^T \boldsymbol{\kappa}_{\mathbf{X}_{i,t}}(\mathbf{x}) \ . \tag{8}$$

The notation $\mathbf{X}_{i,t} = [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,t-1}] \in \mathbb{R}^{p\times(t-1)}$, $\boldsymbol{\kappa}_{\mathbf{X}_{i,t}}(\cdot) = [\kappa(\mathbf{x}_{i,1}, \cdot), \ \ldots \ , \kappa(\mathbf{x}_{i,t-1}, \cdot)]^T$, and $\mathbf{w}_{i,t} = [w_{i,1}, \ \ldots \ w_{i,t-1}] \in \mathbb{R}^{t-1}$ is introduced on the right-hand side of (8) for compact representation. Using the kernel expansion in (8), the functional update in (6) boils down to the following V parallel parametric updates on both kernel dictionaries $\mathbf{X}_i$ and $\mathbf{w}_i$:

$$\mathbf{X}_{i,t+1} = [\mathbf{X}_{i,t}, \ \ \mathbf{x}_{i,t}] \ , \tag{9}$$

$$[\mathbf{w}_{i,t+1}]_u = \begin{cases} (1-\eta\lambda)[\mathbf{w}_{i,t}]_u, & 0 \le u \le t-1 \\ -\eta\Big(\ell_i'(f_{i,t}(\mathbf{x}_{i,t}), y_{i,t}) + \sum_{j\in n_i} \mu_{ij} h_{ij}'(f_{i,t}(\mathbf{x}_{i,t}), f_{j,t}(\mathbf{x}_{i,t}))\Big), & u = t \end{cases}$$

Algorithm 1 DFSSPA: Decentralized Functional Stochastic Saddle Point Algorithm

Require: $\{\mathbf{x}_t, \mathbf{y}_t, \epsilon_t\}_{t=0,1,2,\ldots}$, η and δ
  initialize $f_{i,0}(\cdot) = 0, \mathbf{D}_{i,0} = [], \mathbf{w}_0 = []$, i.e. initial dictionary, coefficients are empty for each $i \in \mathcal{V}$
  for $t = 0, 1, 2, \ldots$ **do**
    loop in parallel for agent $i \in \mathcal{V}$
      Observe local training example realization $(\mathbf{x}_{i,t}, y_{i,t})$
      Send $\mathbf{x}_{i,t}$ to the neighboring nodes, $j \in n_i$ and receive $f_{j,t}(\mathbf{x}_{i,t})$
      Receive $\mathbf{x}_{j,t}$ from the neighbouring nodes, $j \in n_i$ and send $f_{i,t}(\mathbf{x}_{j,t})$
      Compute unconstrained stochastic grad. step [cf. (6)]

$$\tilde{f}_{i,t+1}(\cdot) = f_{i,t}(1-\eta\lambda) - \eta\Bigg[\ell_i'(f_{i,t}(\mathbf{x}_{i,t}), y_{i,t}) + \sum_{j\in n_i} \mu_{ij} h_{ij}'(f_{i,t}(\mathbf{x}_{i,t}), f_{j,t}(\mathbf{x}_{i,t}))\Bigg]\kappa(\mathbf{x}_{i,t}, \cdot)$$

      Update dual variables for $j \in n_i$

$$\mu_{ij,t+1} = \Big[\mu_{ij,t}(1-\delta\eta^2) + \eta\Big(h_{ij}(f_{i,t}(\mathbf{x}_{i,t}), f_{j,t}(\mathbf{x}_{i,t})) - \gamma_{ij}\Big)\Big]_+$$

      Update params: $\tilde{\mathbf{D}}_{i,t+1} = [\mathbf{D}_{i,t}, \ \ \mathbf{x}_{i,t}]$, $\tilde{\mathbf{w}}_{i,t+1}$ [cf. (9)]
      Greedily compress function using matching pursuit

$$(f_{i,t+1}, \mathbf{D}_{i,t+1}, \mathbf{w}_{i,t+1}) = \textbf{KOMP}(\tilde{f}_{i,t+1}, \tilde{\mathbf{D}}_{i,t+1}, \tilde{\mathbf{w}}_{i,t+1}, \epsilon_t)$$

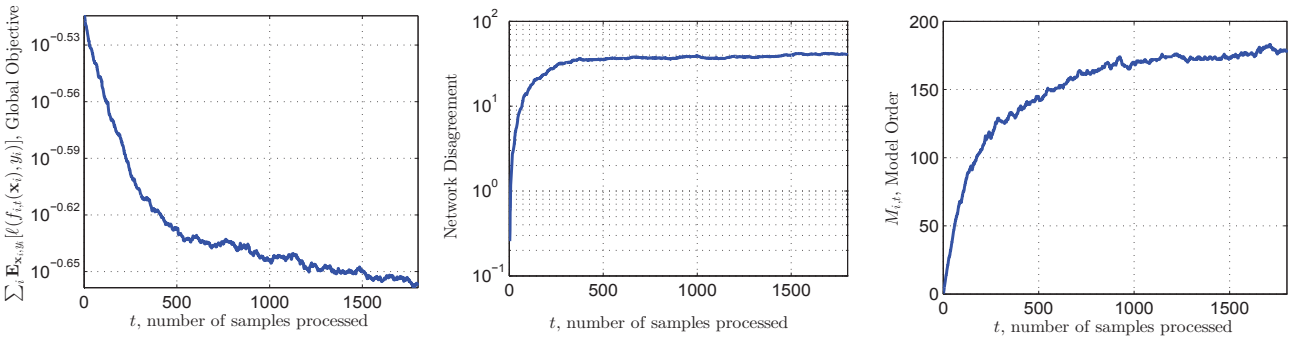
    end loop
  end for

From (9) it can be observed that each time one more column gets added to the columns in $\mathbf{X}_{i,t}$, thereby increasing the dimension everytime leading to intractable memory growth. We define the number of data points $M_{i,t}$, i.e., the number of columns of $\mathbf{X}_{i,t}$ at time t as the *model order*. For the stochastic functional gradient update in (6), the model order $M_{i,t} = t-1$ grows unbounded with iteration index t. This is a conventional challenge in bridging the gap between nonparametric statistics and optimization tools. Next by taking the motivation from [17], feed each agent's function into a variant of (kernel orthogonal matching pursuit) KOMP [36] that explicitly enforces the projection to be contained within a finite Hilbert norm ball, which may be interpreted as a proximal projection [17].

4. CONVERGENCE RESULTS

In this section, we establish that the proposed algorithm converges in expectation both in terms of objective suboptimality and constraint violation to a fixed-radius neighborhood when used with a constant step-size. Before doing so, we state some technical conditions required.

[AS1] The feature space $\mathcal{X} \subset \mathbb{R}^p$ and target domain $\mathcal{Y} \subset \mathbb{R}$



(a) Global loss function w.r.t Samples (b) Network Disagreement v.s. Samples (c) Model Order $\mathbf{M}_{i,t}$ v.s. Samples

Fig. 1: Convergence in terms of primal sub-optimality, constraint violation, and model complexity, for estimating a spatially correlated random field with a nonlinear observation model with parsimony constant $P = 0.8$, Gaussian kernel with 0.04, $\lambda = \delta = 10^{-5}$, and $\eta = 0.05$.

are compact, and the kernel map may be bounded as

$$\sup_{\mathbf{x}\in\mathcal{X}} \sqrt{\kappa(\mathbf{x},\mathbf{x})} = X < \infty \tag{10}$$

[AS2] The local losses $\ell_i(f_i(\mathbf{x}), y)$ are convex and differentiable with respect to the first (scalar) argument $f_i(\mathbf{x})$ on $\mathbb{R}$ for all $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. Moreover, the instantaneous losses $\ell_i : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ are C_i-Lipschitz continuous for all $z \in \mathbb{R}$ for a fixed $y \in \mathcal{Y}$

$$|\ell_i(z, y) - \ell_i(z', y)| \leq C_i|z - z'| \tag{11}$$

with $C := \max_i C_i$ as the largest modulus of continuity.

[AS3] The constraint functions h_{ij} for all $(i, j) \in \mathcal{E}$ are all uniformly L_h-Lipschitz continuous in its first (scalar) argument; i.e., for any z, $z' \in \mathbb{R}$, there exist constant L_h, such that

$$|h_{ij}(z, y) - h_{ij}(z', y)| \leq L_h|z - z'| \tag{12}$$

[AS4] The output $f_{i,t+1}$ of the KOMP update has hilbert norm bounded by $R_{\mathcal{B}} \leq \infty$, and the optimal f_i^* lies in the ball $\mathcal{B}$ with radius $R_{\mathcal{B}}$.

With these stated, we are ready to state the main result of this work.

Theorem 1 *Assume that the assumptions **AS1-AS4** hold. Denote $(f_t, \boldsymbol{\mu}_t)$ as the primal-dual sequence from Algorithm 1 with constant step-size $\eta = 1/\sqrt{T}$ and approximation budget $\epsilon_t = \epsilon = P\eta^2$, where the scalar $P > 0$ is termed as the parsimony constant. Further define $S(f_t) = \sum_{i\in\mathcal{V}} \mathbb{E}\big[\ell_i(f_{i,t}(\mathbf{x}_{i,t}), y_{i,t})\big] + \frac{\lambda}{2}\|f_{i,t}\|^2_{\mathcal{H}}$ as the objective in* (1) *with f^* defined in* (3)*. The objective sub-optimality $\mathbb{E}[S(f_t) - S(f^*)]$ grows sub-linearly with iteration index T*

$$\sum_{t=1}^{T} \mathbb{E}[S(f_t) - S(f^*)] \leq \mathcal{O}(\sqrt{T}). \tag{13}$$

Furthermore, the time-aggregation of the expected constraint violation grows sub-linearly with iteration T as

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

Theorem 1 result establishes that the iterates generated by the proposed algorithm f_t converges to the optimal value $f^\star$ as $T \to \infty$ and yields increasing feasible iterates as time progresses [see [35] for proof]. The use of KOMP-based projections also allows us to establish that the limiting model order of each agents function is finite in a manner similar to Corollary 1 in [26]. This result is omitted here in the interest of brevity, but will be included in future work.

5. NUMERICAL RESULTS

This section details an application of estimating a spatially planar correlated Gaussian random field in a given region $\mathcal{G} \subset \mathbb{R}^2$ space. The estimation is performed using the proposed algorithm Algo. 1 and convergence is demonstrated. A planar field is a random function of spatial components u (for x-axis) and z (for y-axis) across a region $\mathcal{G}$. Moreover, random field is parameterized by the correlation matrix $\mathbf{R}_x$, which depends on the location of the sensors. Each element of $[\mathbf{R}_x]_{ij}$ is assumed to have a structure of the form $\Omega(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|l_i - l_j\|}$, where l_i and l_j are the respective locations of sensor i and j in region $\mathcal{G}$ [37]. All the sensors collect observations $(\mathbf{x}_{i,t}, y_{i,t})$ where $y_{i,t}$ is a noisy non-linear transformation of the the original field $\mathbf{x}_{i,t}$. The observation model can be written as $y_{i,t} = h_i\|\mathbf{x}_{i,t}\|^2 + n_{i,t}$, where $n_{i,t} \sim \mathcal{N}(0, \sigma^2)$ is i.i.d with $\sigma^2 = 2$. For the simulation purpose, we considered a network of randomly connected 10 nodes which are spatially distributed in a 10×10 meter square area. The instantaneous observation $\mathbf{x}_t$ across the network is given by $\mathbf{x}_t = \pi + \mathbf{C}^T\mathbf{v}_t$, where $\pi = \{1/V, 2/V, \ldots, 1\}$ is a fixed mean vector of length V, $\mathbf{C}$ is the Cholesky factorization of the correlation matrix $\mathbf{R}_x$, and $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. We select tolerance parameter to be $\gamma_{ij} = \Omega(\mathbf{x}_i, \mathbf{x}_j)$. The observation $\mathbf{x}_{i,t}$ is a scalar $(p = 1)$. After solving the random field estimation problem over a network using the proposed algorithm, the simulation results are presented in Fig. 1. These results show the convergence of the global objective to the optimal, constraint violation going to zero, and finiteness of the model order.

6. REFERENCES

[1] A. Koppel, B. M. Sadler, and A. Ribeiro, "Proximity without consensus in online multiagent optimization," *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 3062–3077, 2017.

[2] A. S. Bedi and K. Rajawat, "Asynchronous incremental stochastic dual descent algorithm for network resource allocation," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2229–2244, 2018.

[3] J. Liu, Q. Chen, and H. D. Sherali, "Algorithm design for femtocell base station placement in commercial building environments," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2951–2955.

[4] A. Ghosh and S. Sarkar, "Pricing for profit in internet of things," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 2211–2215.

[5] R. J. Kozick and B. M. Sadler, "Source localization with distributed sensor arrays and partial spatial coherence," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 601–616, 2004.

[6] M. Schwager, P. Dames, D. Rus, and V. Kumar, "A multi-robot control policy for information gathering in the presence of unknown hazards," in *Robotics Research*. Springer, 2017, pp. 455–472.

[7] S.-W. Seong, J. Seo, M. Nasielski, D. Sengupta, S. Hangal, S. K. Teh, R. Chu, B. Dodson, and M. S. Lam, "Prpl: a decentralized social networking infrastructure," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*. ACM, 2010, p. 8.

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

[9] L. Bottou, "Online learning and stochastic approximations," *On-line learning in neural networks*, vol. 17, no. 9, p. 142.

[10] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[11] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Trans. Img. Proc.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[12] S. Haykin, "Neural networks: A comprehensive foundation," 1994.

[13] P. Jain, P. Kar *et al.*, "Non-convex optimization for machine learning," *Foundations and Trends® in Machine Learning*, vol. 10, no. 3-4, pp. 142–336, 2017.

[14] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points?online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[15] K. Slavakis, P. Bouboulis, and S. Theodoridis, "Online learning in reproducing kernel hilbert spaces," *Signal Processing Theory and Machine Learning*, pp. 883–987, 2013.

[16] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," *Subseries of Lecture Notes in Computer Science Edited by JG Carbonell and J. Siekmann*, p. 416, 2001.

[17] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online learning with kernels via sparse projections in function space," *arXiv preprint arXiv:1612.04111*, 2016.

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

[19] 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.

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

[21] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed optimization via dual averaging," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, 2013, pp. 1484–1489.

[22] A. Koppel, F. Y. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5149–5164, 2015.

[23] S. Lee and M. M. Zavlanos, "Distributed primal-dual methods for online constrained optimization," in *American Control Conference (ACC), 2016*. IEEE, 2016, pp. 7171–7176.

[24] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization." *IEEE Trans. Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.

[25] A. S. Bedi, A. Koppel, and K. Rajawat, "Beyond consensus and synchrony in online network optimization via saddle point method," *arXiv preprint arXiv:1707.05816*, 2017.

[26] A. Koppel, S. Paternain, C. Richard, and A. Ribeiro, "Decentralized online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3240–3255, June 2018.

[27] A. Koppel, K. Zhang, H. Zhu, and T. M. Baser, "Projected stochastic primal-dual method for constrained online learning with kernels," *IEEE Trans. Signal Process.*, vol. (submitted), Apr 2018, available at http://koppel.bitballoon.com/.

[28] A. S. Bedi, P. Sarma, and K. Rajawat, "Tracking moving agents via inexact online gradient descent algorithm," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 202–217, Feb 2018.

[29] R. J. Kozick and B. M. Sadler, "Source localization with distributed sensor arrays and partial spatial coherence," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 601–616, March 2004.

[30] A. Ribeiro, "Ergodic Stochastic Optimization Algorithms for Wireless Communication and Networking," vol. 58, no. 12, pp. 6369–6386, Dec. 2010.

[31] J.-B. Li, S.-C. Chu, and J.-S. Pan, *Kernel Learning Algorithms for Face Recognition*. Springer, 2014.

[32] K. Müller, T. Adali, K. Fukumizu, J. C. Principe, and S. Theodoridis, "Special issue on advances in kernel-based learning for signal processing [from the guest editors]," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 14–15, 2013. [Online]. Available: http://dx.doi.org/10.1109/MSP.2013.2253031

[33] V. Norkin and M. Keyzer, "On stochastic optimization and statistical learning in reproducing kernel hilbert spaces by support vector machines (svm)," *Informatica*, vol. 20, no. 2, pp. 273–292, 2009.

[34] R. Wheeden, R. Wheeden, and A. Zygmund, *Measure and Integral: An Introduction to Real Analysis*, ser. Chapman & Hall/CRC Pure and Applied Mathematics. Taylor & Francis, 1977. [Online]. Available: https://books.google.com/books?id=YDkDmQ_hdmcC

[35] "Technical report for exact nonparamteric decentralized online optimization," 2018. [Online]. Available: https://bit.ly/2tFY3P0

[36] P. Vincent and Y. Bengio, "Kernel matching pursuit," *Machine Learning*, vol. 48, no. 1, pp. 165–187, 2002.

[37] M. Dong, L. Tong, and B. M. Sadler, "Information retrieval and processing in sensor networks: Deterministic scheduling versus random access," *IEEE Transactions on Signal Processing*, vol. 55, no. 12, pp. 5806–5820, Dec 2007.