DISTRIBUTED MAXIMUM A POSTERIORI PROBABILITY ESTIMATION OF DYNAMIC SYSTEMS WITH WIRELESS SENSOR NETWORKS

Felicia Y. Jakubiec and Alejandro Ribeiro

Department of Electrical and Systems Engineering, University of Pennsylvania

ABSTRACT

This paper develops a framework for the estimation of a time-varying random signal using a wireless sensor network. Given a continuous time model, sensors collect noisy observations and produce local estimates according to the discrete-time equivalent system defined by the sampling period of observations. Estimation is performed using a maximum a posteriori probability estimator (MAP) within a given window of interest. To mediate the incorporation of information from other sensors we introduce Lagrange multipliers to penalize the disagreement between neighboring estimates. We show that the resulting distributed (D-)MAP algorithm is able to track dynamical signals with a small error. This error is characterized in terms of problem constants and vanishes with the sampling time as long as the log-likelihood function satisfies a smoothness condition.

Index Terms— Wireless sensor networks. Distributed estimation.

1. INTRODUCTION

We consider the problem of estimating a time-varying signal with a distributed sensor network which collects noisy observations of the signal of interest. Our goal is to implement a distributed and adaptive estimation algorithm to track this dynamical system relying on local observations and communication with neighboring nodes. To meet this goal we utilize maximum a posteriori probability (MAP) estimates and design a mechanism to incorporate global information into local estimates. Ultimately, we want sensors to compute estimates at time t which estimate the state of the system at the same time t and are close to the optimal centralized MAP, the estimate computed if all the observations were available at a central location.

The first idea proposed to mediate the incorporation of global information within local estimates is the consensus algorithm in which sensors update their estimates through iterative averaging of neighboring values. Consensus algorithms are well studied for static estimation problems, e.g. [1], and have also been adapted for dynamic estimation [2]. An alternative approach to mediate the incorporation of global information is through the introduction of Lagrange multipliers, effectively setting a price on disagreement which sensors try to minimize; a feat which can be accomplished in a distributed manner using dual subgradient descent techniques [3]. This approach has been generalized to nonlinear estimation and linear Gaussian autoregressive (AR) models [4]. The generalization to linear Gaussian AR models gives rise to distributed Kalman filter implementations; [4]; see also [5] for a tutorial review.

Aside from the distributed Kalman filters in [4], work on distributed estimation for time-varying parameters assumes that communications occur in a time scale separate from the timeline of the dynamic system. This assumption is necessary because the algorithms in [1–3] are iterative. Thus, their implementation in a dynamic setting requires the assumption that an infinite number of communication steps occur between subsequent states of the dynamic system. In this paper we generalize the price mediation algorithms of [3, 4] to nonlinear dynamic estimation problems while using a common time scale for communications and the evolution of the process. When using a single time scale, each iteration of the price update algorithm brings the sensors closer to agreement on the MAP estimate, while at the same time the process, and thus the MAP estimate, drifts to a new value. The technical contribution of this paper is to characterize this tradeoff by showing that local estimates approach the centralized MAP estimator with a small error which we characterize in terms of problem-specific constants.

Section 2 starts by introducing the dynamical model in continuous time and its equivalent model in discrete time, and formulating the global MAP estimation problem. We then proceed to the development of distributed dual gradient descent by separating the global MAP estimation problem into local maximization problems. To clarify the discussion we particularize the algorithm to the estimation of a linear Gaussian AR process (Section 2.2). Due to randomness of the dynamical system being tracked, convergence of the distributed algorithm to the global optimal solution can only be claimed in a probabilistic sense (Section 3). Specifically, we claim that the proposed distributed MAP estimator: (i) converges in mean to a value close to the centralized MAP; and (ii) visits a small neighborhood of the optimal set of solutions infinitely often. The paper closes with numerical results for a scalar linear system (Section 4).

2. PROBLEM FORMULATION

Consider a connected symmetric sensor network $\mathcal{G}$ composed of K sensors and let n_k denote the set of neighbors of sensor k. The network is deployed to estimate a $J \times 1$ continuous time-varying vector signal $\mathbf{s}_a(\tau) = [s_{a1}(\tau),\, s_{a2}(\tau),\, ,\ldots,\, s_{aJ}(\tau)]^T$. Each sensor collects a $J \times 1$ vector observation which we denote as $\mathbf{x}_{ak}(\tau) = [x_{ak1}(\tau),\, x_{ak2}(\tau),\, ,\ldots,\, x_{akJ}(\tau)]^T$. We assume that observations $\mathbf{x}_{ak}(\tau)$ collected at different sensors are conditionally independent given the signal $\mathbf{s}_a(\tau)$ and that the conditional probability density function (pdf) $\mathrm{P}\left(\mathbf{x}_a(\tau)|\mathbf{s}_a(\tau)\right)$ is known at each sensor. We further assume that the process of time-varying signal values $\mathbf{s}_a(\tau)$ can be described by the differential equation

$$\dot{\mathbf{s}}_a(\tau) = f_{a\mathbf{s}}(\mathbf{s}_a(\tau), \mathbf{u}_a(\tau)), \tag{1}$$

where $\mathbf{u}_a(\tau)$ denotes white driving input noise. For any time h we can compute the transition pdf $\mathrm{P}\left(\mathbf{s}_a(\tau+h)|\mathbf{s}_a(\tau)\right)$ from (1). We assume that this pdf as well as the observation model pdf $\mathrm{P}\left(\mathbf{x}_a(\tau)|\mathbf{s}_a(\tau)\right)$ are log-concave, i.e. the logarithms $\ln \mathrm{P}\left(\mathbf{x}_a(\tau)|\mathbf{s}_a(\tau)\right)$ and $\ln \mathrm{P}\left(\mathbf{s}_a(\tau+h)\right.$ $\left.|\mathbf{s}_a(\tau)\right)$ are concave functions of the signal values $\mathbf{s}(\tau)$ and $\mathbf{s}(\tau+1)$.

To estimate $\mathbf{s}_a(\tau)$ we consider the equivalent discrete time model $\mathbf{s}^n = \mathbf{s}_a(nT_s)$ obtained by sampling $\mathbf{s}_a(\tau)$ at intervals of length T_s. Likewise, we consider discrete-time observations $\mathbf{x}_k^n = \mathbf{x}_{ak}(nT_s)$ obtained at the same sampling instances and define the vector $\mathbf{x}^n = [\mathbf{x}_1^{n\,T}, \ldots, \mathbf{x}_K^{n\,T}]^T$ stacking the observation samples of all nodes for time n. The probabilistic description of the discrete time model can be obtained from the continuous time model introduced above. We use $\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)$ and $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)$ to denote the kth sensor observation and state transition pdfs, respectively.

In estimation of time-varying processes the goal is to compute estimates $\hat{\mathbf{s}}^0, \ldots, \hat{\mathbf{s}}^t$ of all observed signals given all collected observations $\mathbf{x}^0, \ldots, \mathbf{x}^t$. To avoid excessive memory growth we introduce a time window of length W and focus instead on computing estimates $\hat{\mathbf{s}}^{t-W+1}, \ldots, \hat{\mathbf{s}}^t$ during the window length using all the observations $\mathbf{x}^{t-W+1}, \ldots, \mathbf{x}^t$ collected during the window. To simplify notation let t denote the current time index so that the window of interest includes observations and signals between times $t-W+1$ and t. Denote as $\mathbf{x}_k(t) := [\mathbf{x}_k^{t-W+1}(t)^T \ldots \mathbf{x}_k^t(t)^T]^T$ the vector containing all observations during the window for given sensor k, recall the definition of the vector $\mathbf{x}^n := [\mathbf{x}_1^{n\,T}, \ldots, \mathbf{x}_K^{n\,T}]^T$ grouping observations of all sensors during given time $n \in [t-W+1, t]$, and further define

Supported by NSF CCF-1017454 and AFOSR MURI FA9550-10-1-0567.

$\mathbf{x}(t) := [\mathbf{x}_1^T(t), \ldots, \mathbf{x}_K^T(t)]^T$ grouping observations for all sensors and all times during the window. Unless otherwise noted, in a symbol of the form $\mathbf{x}_k^n(t)$, the argument denotes the current time t, the superscript a time $n \in [t-W+1, t]$ during the window of interest, and the subscript k a given sensor. If a symbol does not have a subscript it is supposed to group homologous variables for all sensors. If it misses a superscript it groups all times between $t-W+1$ and t, and if it misses both it groups all sensors and all window times. According to this notational convention we can define the vector $\hat{\mathbf{s}}_{\text{MAP}}(t) = [\hat{\mathbf{s}}_{\text{MAP}}^{t-W+1}(t)^T \ldots \hat{\mathbf{s}}_{\text{MAP}}^{t}(t)^T]^T$ of all MAP estimates between times $t-W+1$ and t as

$$\hat{\mathbf{s}}_{\text{MAP}}(t) = \operatorname*{argmax}_{\mathbf{s}} \mathrm{P}\left(\mathbf{s}|\mathbf{x}(t)\right) = \operatorname*{argmax}_{\mathbf{s}} \mathrm{P}\left(\mathbf{x}(t)|\mathbf{s}\right) \mathrm{P}\left(\mathbf{s}\right), \quad (2)$$

where Bayes's rule is used in the second equality. Recalling the conditional independence of the observations at different sensors, the conditional probability in (2) can be rewritten as

$$\mathrm{P}\left(\mathbf{x}(t)|\mathbf{s}\right) = \prod_{n=t-W+1}^{t} \prod_{k=1}^{K} \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right), \quad (3)$$

Similarly, using the Markov property of the continuous model according to which $\mathbf{s}^n$ only depends on $\mathbf{s}^{n-1}$ but not on previous data we can write the prior distribution in (2) as

$$\mathrm{P}\left(\mathbf{s}\right) = \prod_{n=t-W+1}^{t} \Big(\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right) \Big) \quad (4)$$

Substituting (3) and (4) for the corresponding terms in (2) leads to

$$\hat{\mathbf{s}}_{\text{MAP}}(t) = \operatorname*{argmax}_{\mathbf{s}} \prod_{n=t-W+1}^{t} \Bigg(\prod_{k=1}^{K} \Big(\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right) \Big) \mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right) \Bigg). \quad (5)$$

Notice that the estimator $\hat{\mathbf{s}}_{\text{MAP}}(t)$ is obtained through the maximization of a time-varying objective, because observations $\mathbf{x}_k^n$ shift to the left as time t progresses. Since the logarithm is a monotonously increasing function, we can alternatively write the MAP estimate in (5) as

$$\hat{\mathbf{s}}_{\text{MAP}}(t) = \operatorname*{argmax}_{\mathbf{s}} f_0(\mathbf{s}, t) = \operatorname*{argmax}_{\mathbf{s}} \sum_{n=t-W+1}^{t} \Bigg(\sum_{k=1}^{K} \Big(\ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right) \Big) + \ln \mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right) \Bigg), \quad (6)$$

where we introduced $f_0(\mathbf{s}, t)$ to denote the maximization objective at time t. Since we assume that the probability distributions $\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)$ and $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)$ are log-concave, the likelihood function $f_0(\mathbf{s}, t)$ is concave. Thus, the computational complexity of solving (6) is approximately cubic in the window size and the dimension of the signal vector $\mathbf{s}^n$. This means that computation of MAP estimates at a centralized location can be carried at manageable computational complexity even for large window sizes. Concavity of $f_0(\mathbf{s}, t)$ also permits devising a distributed implementation as we discuss in the next section.

2.1. Distributed maximum a posteriori probability estimators

Formulated as in (6), the MAP estimator cannot be implemented in a distributed manner because the MAP estimate $\hat{\mathbf{s}}_{\text{MAP}}(t)$ is a variable global to the network. We therefore introduce local estimates $\hat{\mathbf{s}}_k^n(t)$ for all sensors k and window times $n \in [t-W+1, t]$ within the current window and reformulate (6) as the time-varying constrained optimization problem

$$\hat{\mathbf{s}}(t) = \operatorname*{argmax}_{\mathbf{s}} \sum_{n=t-W+1}^{t} \left(\sum_{k=1}^{K} \ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right) \right) \quad (7)$$
$$s.t. \quad \mathbf{s}_k^n = \mathbf{s}_l^n, \qquad \forall\, l \in n_k, \forall\, n = t-W+1, \ldots, t$$

where we introduced the vector $\hat{\mathbf{s}}(t)$ stacking local estimates $\hat{\mathbf{s}}_k^n(t)$ for all sensors and times. Observe that if we denote the edge incidence matrix of the directed network as $\mathbf{C}$ the equality constraints in (7) can be written as $\mathbf{C}\mathbf{s} = \mathbf{0}$.

For a connected network the constraints $\mathbf{s}_k^n = \mathbf{s}_l^n$ reduce the feasible space of (7) to configurations which have the same values at all sensors, i.e., they require $\mathbf{s}_k^n = \mathbf{s}_l^n$ for all pairs of sensors k, l and times n. This further implies equivalence of (7) and (6) in the sense that if the optimization problem in (6) has a solution $\hat{\mathbf{s}}_{\text{MAP}}(t)$, every element $\hat{\mathbf{s}}_k(t)$ of the estimator in (7) is equal to the MAP estimator, i.e. $\hat{\mathbf{s}}_k(t) = \hat{\mathbf{s}}_{\text{MAP}}(t)$.

Since the equality constraints are linear and the maximand is concave in the variables $\mathbf{s}_k^n$, the optimization problem in (7) is convex. Thus, we can equivalently work with the Lagrangian dual problem of (7). To do so, associate the Lagrange multiplier $\boldsymbol{\lambda}_{kl}^n$ with the constraint $\mathbf{s}_k^n = \mathbf{s}_l^n$ for the optimization problem at time t and define the Lagrangian as

$$\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}, t) = \sum_{n=t-W+1}^{t} \sum_{k=1}^{K} \Bigg[\ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} {\boldsymbol{\lambda}_{kl}^n}^T (\mathbf{s}_k^n - \mathbf{s}_l^n) \Bigg] \quad (8)$$

where $\boldsymbol{\lambda}$ stacks the Lagrange multipliers for all links k, l and times n. Observe that the Lagrangian in (8) is time-varying because it depends on the observations $\mathbf{x}(t)$ collected during the current window. The dual function, which is also time-varying, is defined as the maximum of the Lagrangian with respect to primal variables, $g(\boldsymbol{\lambda}, t) = \operatorname{argmax}_{\mathbf{s}} \mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}, t)$, and the dual problem is defined as the minimization of the dual function. We denote as $\boldsymbol{\Lambda}^*(t)$ the set of optimal dual variables for the dual function $g(\boldsymbol{\lambda}, t)$.

Because the dual function is convex, we can use a gradient descent algorithm to update multipliers $\boldsymbol{\lambda}$ so that they approach the optimal multiplier set $\boldsymbol{\Lambda}^*(t)$. Since we want to handle communications in the same timeline as the samples of the signal we consider dual iterates $\boldsymbol{\lambda}(t)$ which we want to update according to the gradient descent algorithm

$$\boldsymbol{\lambda}(t+1) = \boldsymbol{\lambda}(t) - \epsilon \nabla g(\boldsymbol{\lambda}(t), t), \quad (9)$$

with some given stepsize ϵ. Notice that in (9), $\boldsymbol{\lambda}(t+1)$ is updated according to the gradient of the dual function $g(\boldsymbol{\lambda}(t), t)$ at time t, but we are interested in its optimality with respect to the dual function $g(\boldsymbol{\lambda}, t+1)$ at time $t+1$. We analyze this mismatch in Section 3.

To compute the gradient of the dual function consider the Lagrangian primal maximizers $\mathbf{s}(t) := \operatorname{argmax} \mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}(t), t)$ for given dual iterate $\boldsymbol{\lambda}(t)$. As is well-known, e.g., [6], the gradient component associated with link k, l and time n is then given by the corresponding constraint slack

$$\Big[\nabla g(\boldsymbol{\lambda}(t), t) \Big]_{kl}^n = \mathbf{s}_k^n(t) - \mathbf{s}_l^n(t). \quad (10)$$

Further notice that because of the symmetry of the network, the last sum in (8) can be rearranged so that it is expressed as a sum of primal variables $\mathbf{s}_k^n$ instead of as a sum of dual variables $\boldsymbol{\lambda}_{kl}^n$. If we do so, the Lagrangian can be separated into a sum of local Lagrangians, i.e., we can write $\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}, t) = \sum_k \mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}, t)$ with

$$\mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}, t) = \sum_{n=t-W+1}^{t} \Bigg[\ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} {\mathbf{s}_k^n}^T(t) (\boldsymbol{\lambda}_{kl}^n - \boldsymbol{\lambda}_{lk}^n) \Bigg]. \quad (11)$$

Since separate maximization of the local Lagrangians in (11) results in the maximization of their sum, it follows that the Lagrangian maximizers $\mathbf{s}_k^n(t)$ necessary to compute the dual gradient components in (10) can be determined in a distributed manner. This permits the definition of a distributed MAP (D-MAP) algorithm which we formulate as an iterative application of the following items.

Primal iteration. Given dual iterate $\boldsymbol{\lambda}(t)$ at time t, determine primal Lagrangian maximizers as

$$\mathbf{s}_k(t) = \underset{\mathbf{s}_k}{\operatorname{argmax}} \sum_{n=t-W+1}^{t} \Big[\ln \mathrm{P}\left(\mathbf{x}_k^n | \mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n | \mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} \mathbf{s}_k^{nT} (\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)) \Big]. \quad (12)$$

Dual iteration. Given primal iterates $\mathbf{s}(t)$ update dual iterates as

$$\boldsymbol{\lambda}_{kl}^n(t+1) = \boldsymbol{\lambda}_{kl}^n(t) - \epsilon\big(\mathbf{s}_k^n(t) - \mathbf{s}_l^n(t)\big) \quad (13)$$

To implement the primal iteration, sensor k needs access to local multipliers $\boldsymbol{\lambda}_k^n(t)$ and multipliers $\boldsymbol{\lambda}_l^n(t)$ for neighboring sensors $l \in n_k$. Likewise, to implement the dual iteration, only local $\mathbf{s}_k^n(t)$ and neighboring $\mathbf{s}_l^n(t)$ primal variables are needed.

2.2. Linear Gaussian autoregressive model

To illustrate the D-MAP algorithm in (12) and (13) consider its application to a linear Gaussian autoregressive model. In this case the state $\mathbf{s}_a(\tau)$ and signal $\mathbf{x}_{ak}(t)$ evolve and are related according to

$$\dot{\mathbf{s}}_a(t) = \mathbf{A}_a\, \mathbf{s}_a(t) + \mathbf{u}_a(t) \quad (14)$$

$$\mathbf{x}_{ak}(t) = \mathbf{H}_{ak}\, \mathbf{s}_a(t) + \mathbf{n}_{ak}(t) \quad (15)$$

where $\mathbf{u}_a(t)$ is the driving noise drawn from a zero-mean Wiener process with covariance matrix $\mathbf{Q}_a$, and $\mathbf{n}_a(t)$ represents Gaussian observation noise drawn from a zero-mean Wiener process with covariance matrix $\mathbf{R}_a$. An equivalent discrete-time model can be obtained by solving the differential equation (14) between times $n\,T_s$ and $(n+1)T_s$ with initial condition $\mathbf{s}^n$ to get [6]

$$\mathbf{s}^{n+1} = \mathbf{A}\, \mathbf{s}^n + \mathbf{u}^n \quad (16)$$

$$\mathbf{x}_k^n = \mathbf{H}_k\, \mathbf{s}^n + \mathbf{n}_k^n \quad (17)$$

where the discrete model parameters depend on the sampling time T_s and can be explicitly computed. The matrices in (16)-(17) are given by $\mathbf{A} = \exp(\mathbf{A}_a\, T_s)$, $\mathbf{H}_k = \mathbf{H}_{ak}$, while the driving and observation noises are white Gaussian with covariance matrices $\mathbf{Q} = \mathbb{E}\left[\mathbf{u}^{n^T}\mathbf{u}^n\right] = (\mathbf{Q}_a/2)\mathbf{A}_a^{-1}(\exp(2\mathbf{A}_a\, T_s) - I)$ and $\mathbf{R}_k = \mathbb{E}\left[\mathbf{n}_k^{n^T}\mathbf{n}_k^n\right] = \mathbf{R}_{ak}/T_s$.

Because the addition of Gaussian random variables yields a Gaussian random variable, the discrete prior conditional probabilities are Gaussian for all times n and any sensor k. Consequently, the MAP estimator in (6) can be reduced to the maximization of the quadratic form

$$\hat{\mathbf{s}}_{\mathrm{MAP}}(t) = \underset{\mathbf{s}}{\operatorname{argmax}} \sum_{n=t-W+1}^{t} \Big(\sum_{k=1}^{K} (\mathbf{x}_k^n - \mathbf{H}_k\, \mathbf{s}^n)^T \mathbf{R}_k^{-1} (\mathbf{x}_k^n - \mathbf{H}_k\, \mathbf{s}^n) + (\mathbf{s}^n - \mathbf{A}\, \mathbf{s}^{n-1})^T \mathbf{Q}^{-1} (\mathbf{s}^n - \mathbf{A}\, \mathbf{s}^{n-1}) \Big). \quad (18)$$

Substituting this specific function into the primal iteration in (12), it follows that for linear Gaussian autoregressive models the primal iteration of the D-MAP algorithm becomes

$$\mathbf{s}_k(t) = \underset{\mathbf{s}_k}{\operatorname{argmax}} \sum_{n=t-W+1}^{t} \Big((\mathbf{x}_k^n - \mathbf{H}_k \mathbf{s}^n)^T \mathbf{R}_k^{-1} (\mathbf{x}_k^n - \mathbf{H}_k \mathbf{s}^n) + \frac{1}{K} (\mathbf{s}^n - \mathbf{A}\mathbf{s}^{n-1})^T \mathbf{Q}^{-1} (\mathbf{s}^n - \mathbf{A}\mathbf{s}^{n-1}) + \sum_{l \in n_k} \mathbf{s}_k^{nT} (\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)) \Big). \quad (19)$$

The dual iteration is identical to (13).

3. CONVERGENCE PROPERTIES

To determine the optimality of the algorithm described in (12) and (13), feasibility and optimality of the solution need to be assessed. Therefore we want to compare the primal iterate $\mathbf{s}_k^n(t)$ computed by sensor k at time t for the signal value at time n with the corresponding centralized MAP estimator $\hat{\mathbf{s}}_{MAP}^n(t)$. Given the equivalence of (6) and (7), $\hat{\mathbf{s}}_{\mathrm{MAP}}^n(t) = \hat{\mathbf{s}}_k^n(t)$ which means that the distance of interest is given by

$$\|\mathbf{s}_k^n(t) - \hat{\mathbf{s}}_{\mathrm{MAP}}^n(t)\| = \|\mathbf{s}_k^n(t) - \hat{\mathbf{s}}_k^n(t)\|. \quad (20)$$

The norm in the right hand side is the distance between the current primal iterate and the optimal primal value of (7). As such it can be easily related to the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\|$ between the current dual iterate $\boldsymbol{\lambda}(t)$ and the set of optimal dual variables $\boldsymbol{\Lambda}^*(t)$. Characterizing this latter distance is the purpose of this section.

To derive this result we make the following assumptions on the log likelihood $f_0(\mathbf{s}, t)$ and dual $g(\boldsymbol{\lambda}, t)$ functions.

(A1) The dual functions are strongly convex for all t with strong convexity parameter m,

$$g(\boldsymbol{\mu}, \boldsymbol{t}) \geq g(\boldsymbol{\lambda}, t) + \nabla g(\boldsymbol{\lambda}, t)^T (\boldsymbol{\mu} - \boldsymbol{\lambda}) + \frac{m}{2} \|\boldsymbol{\mu} - \boldsymbol{\lambda}\|^2 \quad (21)$$

(A2) The gradients of the dual function $g_{\mathbf{x}}(\boldsymbol{\lambda})$ are Lipschitz continuous with Lipschitz constant M,

$$\|\nabla g(\boldsymbol{\mu}, t) - \nabla g(\boldsymbol{\lambda}, t)\| \leq M \|\boldsymbol{\mu} - \boldsymbol{\lambda}\| \quad (22)$$

(A3) Consider the gradients of the log likelihood functions $f_0(\mathbf{s}, t)$ and $f_0(\mathbf{s}, t+1)$ at subsequent times t and $t+1$ evaluated at corresponding optimal points $\hat{\mathbf{s}}_{\mathrm{MAP}}(t)$ and $\hat{\mathbf{s}}_{\mathrm{MAP}}(t+1)$. The expected value of the norm of this difference is bounded by a constant $\delta(T_s)$,

$$\mathbb{E}\left[\left\| \frac{\partial f_0}{\partial \mathbf{s}} \big(\hat{\mathbf{s}}_{\mathrm{MAP}}(t)\big) - \frac{\partial f_{0,t_2}}{\partial \mathbf{s}} \big(\hat{\mathbf{s}}_{\mathrm{MAP}}(t+1)\big) \right\| \right] \leq \delta(T_s) \quad (23)$$

Assumption (A2) is mild. If the dual function is differentiable, then it implies that the Hessian is upper bounded by M. If the objective function is strongly convex, then the Lipschitz constant M exists. If assumption (A2) is fulfilled, then due to specific characteristics of the optimization problem in (7), (A1) is fulfilled for the $\boldsymbol{\lambda}$ considered. Assumption (A3) is also a reasonable requirement. For the linear Gaussian autoregressive model of Section 2.2, the constant $\delta(T_s)$ is proportional to $\sqrt{T_s}$, [6].

The main result of this paper concerns the optimality of $\boldsymbol{\lambda}(t)$. Since the optimal multipliers $\boldsymbol{\lambda}^*(t)$ are elements of a set $\boldsymbol{\Lambda}^*(t)$, we use the Hausdorff distance from $\boldsymbol{\lambda}(t)$ to the set $\boldsymbol{\lambda}^*(t)$, defined as

$$\|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\| = \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}^*(t)} \|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}\|, \quad (24)$$

as a measure of optimality of $\boldsymbol{\lambda}(t)$. Our main result concerns this distance and is stated in the following Theorem – see [6] for the proof.

Theorem 1 *Let $\boldsymbol{\lambda}(t)$ denote the vector with current dual iterates obtained at time t from (13) and $\boldsymbol{\Lambda}^*(t)$ the set of optimal multipliers with elements $\boldsymbol{\lambda}^*(t)$. Assume the step size $\epsilon < 1/M$. If assumptions (A1) to (A3) hold, then the expected value of the distance between the dual multipliers $\boldsymbol{\lambda}(t)$ and the set of optimal multipliers $\boldsymbol{\Lambda}^*(t)$ at time t satisfies*

$$\lim_{t \to \infty} \mathbb{E}\left[\|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\|\right] \leq \gamma \frac{1 - \epsilon m}{\epsilon m} \delta(T_s). \quad (25)$$

Furthermore, for almost all realization of the signal process $\mathbf{s}(t)$ it holds

$$\liminf_{t \to \infty} \|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\| \leq \gamma \frac{1 - \epsilon m}{\epsilon m} \delta(T_s) \quad a.s. \quad (26)$$

The constant $\gamma := \mu_{max}(\mathbf{C}^+) > 0$ is the largest eigenvalue of the Moore-Penrose pseudoinverse of the network's edge incidence matrix $\mathbf{C}$.

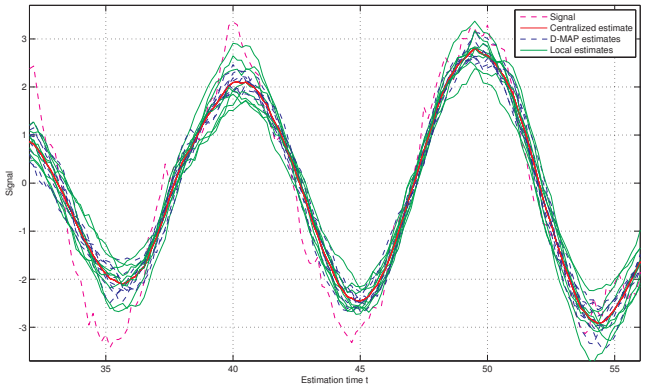

Fig. 1. D-MAP and local estimates compared to centralized estimates and signal for the time $t \in [32, 56]$. The D-MAP is significantly closer to the centralized MAP than the local estimates.

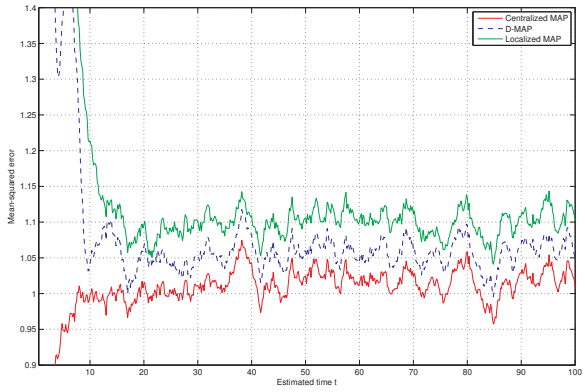

Fig. 2. Difference of mean squared error between D-MAP and centralized MAP compared to difference of mean squared error between local MAP and centralized MAP for the time $t \in [0, 100]$. The D-MAP gives significant improvement over the local MAP.

While Theorem 1 claims convergence of dual variables, the desired convergence result for estimates $\hat{\mathbf{s}}(t)$ follows as a simple corollary. The result is conceptually analogous of Theorem 1 – see [6]. Due to the randomness of the multipliers, it is not possible to bound $\|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\|$ directly. Instead, Theorem 1 shows that there is a tendency for $\|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\|$ to become smaller with increasing time. Specifically, the first result in (25) states that the mean across different realizations of the process becomes small. The second result states that all processes eventually reach the same small value, although they may deviate from this value with some probability. Further notice that for smooth log-likelihood functions having continuous gradients the gradient difference norm in (23) vanishes with decreasing sampling time. It is therefore possible to approximate $\boldsymbol{\Lambda}^*$ arbitrarily by reducing the sampling time. We can then interpret Theorem 1 as a means for selecting T_s to achieve a prescribed error tolerance in the difference $\|\boldsymbol{\lambda}(t) - \boldsymbol{\Lambda}^*(t)\|$. A final interesting observation is the dependence on $\mu_{\max}(\mathbf{C}^+)$ which characterizes networks for which D-MAP performs poorly. The eigenvalue $\mu_{\max}(\mathbf{C}^+)$ is large for networks that are sparsely connected and small for densely connected networks.

4. SIMULATION RESULTS

We simulate tracking of a 2 sinusoid signals $\mathbf{s}(t)$ with a sensor network consisting of $K = 8$ sensors, where the discretized signals from the linear Gaussian autoregressive model are rotated by $\pi/30$ to achieve the sinusoidal shape. With an initial value of $\mathbf{s}(0) = [2, 2]^T$, the system parameters are $\mathbf{A}_a = -0.01$, $\mathbf{q}_a = \text{diag}(\boldsymbol{\sigma}_q)$ and $\mathbf{r}_a = \text{diag}(\boldsymbol{\sigma}_r)$ where the entries of $\boldsymbol{\sigma}_q$ are all 0.5, the entries of $\boldsymbol{\sigma}_r$ are 1, and $\mathbf{H}_a$ is a vector of 8 uniformly distributed values between 0.5 and 1.5. At time $t = 0$, the Lagrange multipliers $\boldsymbol{\lambda}^0_{kl}(0)$ were initialized as 1 for all $l \in n_k$, and at each new time step t, $\boldsymbol{\lambda}^t$ is initialized as $\boldsymbol{\lambda}^t(t) = \boldsymbol{\lambda}^{t-1}(t)$. With these parameters, for a randomly drawn graph, the D-MAP algorithm was run for 1000 realizations of $\mathbf{s}$ and $\mathbf{x}$, according to (19) and (13), using a sampling time of $T_s = 0.16s$ and a window size fixed at $W = 3$ sampling points. The total estimation length is set to $100s$, making the most recent iteration time $n = 600$. To compare the performance of the D-MAP algorithm with the MAP estimates which would have been computed at the sensors if only local information was given, i.e. using only the local MAP, Fig. 1 shows an example of signals where the D-MAP and localized estimates are compared at times $t \in [32, 56]$. For reference, the centralized estimates as well as the source signal are given as well. More generally, Fig. 2 compares the MSE of the D-MAP and the local MAP with the centralized MAP for all estimated times $t = 1...100$, averaged over all 1000 simulation runs with randomly generated signals. With a steady-state MSEs of around 1.02 for the centralized MAP, the D-MAP shows an average MSE of 1.05 and thusly presents a significant improvement over the local MAP with an MSE of 1.1. Note also that it takes the local MAP 18 seconds to reach its steady-state performance whereas 9 seconds are enough for the D-MAP.

5. CONCLUSION

This paper proposes an algorithm for the estimation of time-varying signals with a sensor network collecting noisy observations, which is of a distributed and adaptive nature while at the same time incorporating global information from neighboring nodes. We discuss the optimality of Lagrange multipliers, from which the optimality of primal iterates follows as a corollary. When certain smoothness and continuity assumptions on the primal and dual functions are fulfilled, we claim that (i) the Lagrange multipliers converge in mean to the optimal multipliers, (ii) the Lagrange multipliers visit near optimality infinitely often almost surely, where the proximity to optimality depends on the sampling time. Numerical results corroborate theoretical findings, as the D-MAP improves the estimate for the current time in comparison to the local MAP.

6. REFERENCES

[1] S. Kar and J. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, 2010.

[2] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[3] M. Rabbat, R. Nowak, and J. Bucklew, "Generalized consensus computation in networked systems with erasure links," in *IEEE Workshop on Signal Processing Advances in Wireless Communications Proc.*, June 2005, pp. 1088–1092.

[4] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc WSNs with noisy links - Part I: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[5] A. Ribeiro, I. Schizas, S. Roumeliotis, and G. Giannakis, "Kalman filtering in wireless sensor networks," *IEEE Control Systems Magazine*, vol. 30, no. 2, pp. 66–86, 2010.

[6] F. Jakubiec and A. Ribeiro, "Distributed maximum a posteriory probability estimation of dynamic systems with wireless sensor networks," *IEEE Trans. Signal Process.*, 2011, in preparation. [Online]. Available: https://alliance.seas.upenn.edu/~aribeiro/wiki/