Distributed Maximum a Posteriori Probability Estimation for Tracking of Dynamic Systems

Felicia Y. Jakubiec and Alejandro Ribeiro
Department of Electrical and Systems Engineering, University of Pennsylvania

***Abstract*—We present a framework for the estimation of time-varying random signals with wireless sensor networks. Given a continuous time model, sensors collect noisy observations according to the discrete-time equivalent system defined by the sampling period of observations. Estimation is performed locally using a maximum a posteriori probability estimator (MAP) within a time window. To incorporate information from neighboring sensors we introduce Lagrange multipliers to penalize the disagreement between estimates. We show that the distributed (D-)MAP algorithm is able to track dynamical signals with an error characterized in terms of problem constants. This error vanishes with the sampling period if the log-likelihood function satisfies a smoothness condition.**

I. Introduction

We consider the problem of estimating a time varying signal with a distributed sensor network collecting noisy observations of the signal of interest. To track this dynamical system we implement a distributed estimation algorithm in which sensors rely on local observations and communication with neighboring nodes. We meet this goal using maximum a posteriori probability (MAP) estimates and design a mechanism to incorporate global information into local estimates. At each time step t sensors estimate the state of the system at the same time t while coming close to the optimal centralized MAP that would be computed if all 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, relying on iterative averaging of neighboring values. They are well studied for linear static estimation problems, e.g. [2], [3], and have also been adapted for linear dynamic estimation [4]. Variants of consensus algorithms include gossip algorithms in which data exchanges happen between pairs of neighbors only [5]. As iterative algorithms, most consensus methods in [2], [4], [5] for estimation of time varying signals assume that communications occur in a time scale separate from the timeline of the dynamic system. An approach that doesn't suffer from this drawback are diffusion algorithms [6].

An alternative approach to incorporate global information into local estimates is to introduce Lagrange multipliers effectively setting a price on disagreement that sensors try to minimize. This approach can be rendered optimal by introducing Lagrange multiplier updates based in dual gradient descent [7] or the alternating direction method of multipliers [8] – see also [9] for a tutorial review. This paper generalizes the price mediation algorithms of [8]–[10] to dynamic nonlinear MAP estimation problems. Our work also differs from most existing works on dynamic estimation in that we use 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 the process, and as consequence 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 II starts by introducing the dynamical model in continuous time and its equivalent sampled model in discrete time and follows by formulating the global MAP estimation problem. Under the assumption that log-likelihood functions are concave, the equivalent constrained optimization problem is convex allowing us to work in the dual domain. The distributed (D)-MAP algorithm is then obtained by implementing gradient descent in the dual function as discussed in Section II-A. To clarify discussion we particularize D-MAP to the estimation of a nonlinear, quantized variant of Gaussian AR process in which estimates rely on quantized observations in Section II-B. Section III discusses the convergence properties of the D-MAP algorithm, where our focus lies on studying the distance between dual iterates and the optimal dual variables. Specifically, we prove that: (i) The Lagrange multipliers converge in mean to a close neighborhood around the optimal multipliers. (ii) The Lagrange multipliers almost surely visit a near optimality region infinitely often. Finally, numerical experiments for the quantized observations model of Section II-B are given in Section IV.

II. 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 k 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) $\mathrm{P}(\mathbf{x}_a(\tau)|\mathbf{s}_{ak}(\tau))$ is known at each sensor. We further assume that the process of time-varying signal values $\mathbf{s}_a(\tau)$ can be described by a differential equation of the form

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

where $\mathbf{u}_a(\tau)$ denotes a stationary white driving input signal. For any time step h and given current state $\mathbf{s}_a(\tau)$, (1) determines a time-invariant transition pdf which we denote as $\mathrm{P}(\mathbf{s}_a(\tau+h)|\mathbf{s}_a(\tau))$. We assume that this pdf as well as the observation model pdf $\mathrm{P}(\mathbf{x}_a(\tau)|\mathbf{s}_a(\tau))$ are log-concave, i.e. the logarithms $\ln \mathrm{P}(\mathbf{x}_a(\tau)|\mathbf{s}_a(\tau))$ and $\ln \mathrm{P}(\mathbf{s}_a(\tau+h)\,|\mathbf{s}_a(\tau))$ are concave functions of the signal values $\mathbf{s}_a(\tau)$ and $\mathbf{s}_a(\tau+h)$.

Work in this paper is supported by NSF CCF-1017454 and AFOSR MURI FA9550-10-1-0567. Email: {life, aribeiro}@seas.upenn.edu. An extended version of this paper appeared in the Transactions on Signal Processing [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 $\mathbf{x}_k^n = \mathbf{x}_{ak}(nT_s) = [x_{k1}^n, x_{k2}^n, \ldots, x_{kJ}^n]^T$ obtained at the same sampling instances and define the vector $\mathbf{x}^n = [\mathbf{x}_1^{nT}, \ldots, \mathbf{x}_K^{n\,T}]^T$ stacking the observation samples of all nodes for time n. We use $\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right) = \mathrm{P}\left(\mathbf{x}_{ak}(nT_s)|\mathbf{s}_a(nT_s)\right)$ and $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right) = \mathrm{P}\left(\mathbf{s}_a((n+1)T_s)|\mathbf{s}_a(nT_s)\right)$ to denote the kth sensor observation pdf and the state transition pdf, respectively.

In estimation of time-varying processes the goal is to compute estimates $\mathbf{s}^0, \ldots, \mathbf{s}^t$ 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 $\mathbf{s}^{t-W+1}, \ldots, \mathbf{s}^t$ during the window length using the observations $\mathbf{x}^{t-W+1}, \ldots, \mathbf{x}^t$ collected during the same window. 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 $\mathbf{s}(t) := [\mathbf{s}^{t-W+1^T} \ldots \mathbf{s}^{t^T}]^T$ and $\mathbf{x}_k(t) := [\mathbf{x}_k^{t-W+1^T} \ldots \mathbf{x}_k^{t\,T}]^T$ the vector containing all signals and observations during the window for given sensor k respectively, and further define $\mathbf{x}(t) := [\mathbf{x}_1^T(t), \ldots, \mathbf{x}_K^T(t)]^T$ grouping observations for all sensors and all times during the window. Define the vector $\mathbf{s}_{\text{MAP}}(t) = [\mathbf{s}_{\text{MAP}}^{t-W+1}(t)^T \ldots \mathbf{s}_{\text{MAP}}^t(t)^T]^T$ of all MAP estimates between times $t-W+1$ and t as

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

where Bayes' 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

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

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

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

Notice that the estimator $\mathbf{s}_{\text{MAP}}(t)$ 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

$$\begin{aligned}\mathbf{s}_{\text{MAP}}(t) &= \underset{\mathbf{s}}{\operatorname{argmax}}\, f_{(\text{MAP})}(\mathbf{s}, t) \\ &= \underset{\mathbf{s}}{\operatorname{argmax}} \sum_{n=t-W+1}^{t} \left(\sum_{k=1}^{K} \left(\ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)\right) + \ln \mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)\right),\end{aligned} \quad (6)$$

where we defined the function $f_{(\text{MAP})}(\mathbf{s}, t)$ to denote the centralized log-likelihood function at time t whose maximization yields MAP estimates $\mathbf{s}_{\text{MAP}}(t)$. 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 $f_{(\text{MAP})}(\mathbf{s}, t)$ 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 $f_{(\text{MAP})}(\mathbf{s}, t)$ also permits devising a distributed implementation as we discuss in the next section.

A. *Distributed maximum a posteriori probability estimators*

Formulated as in (6), the MAP estimator cannot be implemented in a distributed manner because the MAP estimate $\mathbf{s}_{\text{MAP}}(t)$ is a variable global to the network. In order to propose a distributed algorithm, we rely on dual reformulations that are standard in convex optimization. Start by introducing local estimates $\mathbf{s}_k^n(t)$ for all sensors k and times $n \in [t-W+1, t]$ within the current window and reformulate (6) as the time-varying constrained optimization problem

$$\begin{aligned}\mathbf{s}^*(t) = \underset{\mathbf{s}}{\operatorname{argmax}} &\sum_{n=t-W+1}^{t} \sum_{k=1}^{K} \ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right) \\ s.t. \quad \mathbf{s}_k^n = \mathbf{s}_l^n, \quad &\text{for all } l \in n_k, \\ &\text{for all } n = t-W+1, \ldots, t\end{aligned} \quad (7)$$

where we introduced the vector $\mathbf{s}^*(t)$ stacking local estimates $\mathbf{s}_k^{n*}(t)$ for all sensors and times. For a connected network the constraints $\mathbf{s}_k^n = \mathbf{s}_l^n$ reduce the feasible space of (7) to configurations that 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. Then the centralized problem (6) and the constrained optimization problem (7) are equivalent, i.e. $\mathbf{s}_k^*(t) = \mathbf{s}_{\text{MAP}}(t)$ if the latter exists.

If we denote the edge incidence matrix of the directed network as $\mathbf{C}_k$, we can define a replicated version as $\mathbf{C}$ where each $1, -1$ and 0 in the matrix are replaced by the identity matrix $\mathbf{I}, -\mathbf{I}$ and the zero matrix $\mathbf{0}$ of size J respectively. Then the equality constraints in (7) can be written in the more compact notation $\mathbf{Cs} = \mathbf{0}$. Further defining local objectives $f_k(\mathbf{s}_k, t) = \sum_{n=t-W+1}^{t} \ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + (1/K) \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right)$ and global D-MAP objectives $f_0(\mathbf{s}, t) = \sum_k f_k(\mathbf{s}_k, t)$ we can rewrite (7) as

$$\begin{aligned}\mathbf{s}^*(t) = &\underset{\mathbf{s}}{\operatorname{argmax}}\; f_0(\mathbf{s}, t) = \sum_{k=1}^{K} f_k(\mathbf{s}_k, t), \\ &\text{s.t.} \qquad \mathbf{Cs} = \mathbf{0}.\end{aligned} \quad (8)$$

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

$$\begin{aligned}\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}, t) = \sum_{n=t-W+1}^{t} \sum_{k=1}^{K} \Bigg[&\ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right) \\ &+ \sum_{l \in n_k} \boldsymbol{\lambda}_{kl}^{n\,T}(\mathbf{s}_k^n - \mathbf{s}_l^n) \Bigg]\end{aligned} \quad (9)$$

where $\boldsymbol{\lambda}$ stacks the Lagrange multipliers for all links k, l and times n. Observe that the Lagrangian in (9) 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

$$\boldsymbol{\lambda}(t+1) = \boldsymbol{\lambda}(t) - \epsilon \nabla g(\boldsymbol{\lambda}(t), t), \tag{10}$$

with some given stepsize ϵ. Notice that in (10), $\boldsymbol{\lambda}(t+1)$ 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 $t+1$. 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 $\boldsymbol{\lambda}(t)$. The gradient component associated with link k, l and time n is then given by the corresponding constraint slack

$$\Big[\nabla g(\boldsymbol{\lambda}(t), t)\Big]_{kl}^{n} = \mathbf{s}_k^n(t) - \mathbf{s}_l^n(t). \tag{11}$$

Further notice that because of the symmetry of the network, the last sum in (9) 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 $\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}, t) = \sum_{k=1}^{K} \mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}, t)$ with

$$\mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}, t) = \sum_{n=t-W+1}^{t} \bigg[\ln \mathrm{P}\left(\mathbf{x}_k^n | \mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n | \mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} \mathbf{s}_k^{n^T}(t)(\boldsymbol{\lambda}_{kl}^n - \boldsymbol{\lambda}_{lk}^n) \bigg]. \tag{12}$$

Since separate maximization of the local Lagrangians in (12) 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 (11) 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

$$\mathbf{s}_k(t) = \operatorname*{argmax}_{\mathbf{s}_k} \sum_{n=t-W+1}^{t} \bigg[\ln \mathrm{P}\left(\mathbf{x}_k^n | \mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n | \mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} \mathbf{s}_k^{n^T}(\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)) \bigg]. \tag{13}$$

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

$$\boldsymbol{\lambda}_{kl}^n(t+1) = \boldsymbol{\lambda}_{kl}^n(t) - \epsilon\big(\mathbf{s}_k^n(t) - \mathbf{s}_l^n(t)\big) \tag{14}$$

To implement the primal iteration, sensor k needs access to the local multipliers $\boldsymbol{\lambda}_k^n(t)$ and the 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.

B. Quantized observations

To illustrate the D-MAP algorithm in (13) and (14) consider its application to a quantized model based on the following Gaussian AR model. Consider the case where the state $\mathbf{s}_a(\tau)$ and signal $\mathbf{x}_{ak}(t)$ evolve and are related according to

$$\dot{\mathbf{s}}_a(\tau) = \mathbf{A}_a \, \mathbf{s}_a(\tau) + \mathbf{u}_a(\tau) \tag{15}$$

$$\mathbf{x}_{ak}(\tau) = \mathbf{H}_{ak} \, \mathbf{s}_a(\tau) + \mathbf{n}_{ak}(\tau) \tag{16}$$

where $\mathbf{u}_a(\tau)$ is the driving noise drawn from a zero-mean Wiener process with covariance matrix $\mathbf{Q}_a$, and $\mathbf{n}_a(\tau)$ 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 (15) between times $n T_s$ and $(n+1)T_s$ with initial condition $\mathbf{s}^n$ to get [1]

$$\mathbf{s}^{n+1} = \mathbf{A} \, \mathbf{s}^n + \mathbf{u}^n \tag{17}$$

$$\mathbf{x}_k^n = \mathbf{H}_k \, \mathbf{s}^n + \mathbf{n}_k^n \tag{18}$$

where the discrete model parameters depend on the sampling time T_s and can be explicitly computed. The matrices in (17)-(18) 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 $\mathbf{Q} = \mathbb{E}\left[\mathbf{u}^{n^T}\mathbf{u}^n\right] = (\mathbf{Q}_a/2)\mathbf{A}_a^{-1}(\exp(2\mathbf{A}_a T_s) - \mathbf{I})$ and $\mathbf{R}_k = \mathbb{E}\left[\mathbf{n}_k^{n^T}\mathbf{n}_k^n\right] = \mathbf{R}_{ak}/T_s$. Consider the quantization of the linear model in (17)-(18) where sensors are attached to a single-level quantizer that produces binary observations $\mathbf{y}_k^n = [y_{k1}^n, \ldots, y_{kJ}^n]$ with elements $y_{kj}^n \in \{0, 1\}$. To model the quantization process we introduce the threshold level $\theta_{0,kj}$ used to quantize the jth component x_{kj}^n of the vector observation $\mathbf{x}_k^n$ in (18). The binary variable y_{kj}^n indicates whether the analog observation x_{kj}^n exceeds the threshold $\theta_{0,kj}$, i.e. $y_{kj}^n = \mathbb{I}\{x_{kj}^n \geq \theta_{0,kj}\}$. For simplicity of exposition assume the observation noise is uncorrelated so that the covariance matrix $\mathbf{R}_k$ takes on the diagonal form $\mathbf{R}_k = \operatorname{diag}(r_{k1}, r_{k2} \ldots, r_{kJ})$. Then the log-likelihood $\ln \mathrm{P}\big(\mathbf{y}_k^n | \mathbf{s}^n\big)$ can be computed as

$$\ln \mathrm{P}\left(\mathbf{y}_k^n | \mathbf{s}^n\right) = \sum_{j=1}^{J} \Big(y_{kj}^n \ln \mathrm{P}\left(y_{kj}^n = 1 | \mathbf{s}^n\right) + (1 - y_{kj}^n) \ln\left(1 - \mathrm{P}\left(y_{kj}^n = 1 | \mathbf{s}^n\right)\right) \Big). \tag{19}$$

Let $\mathbf{h}_k^T$ denote the k-th row of the observation matrix $\mathbf{H}_k$, then $\mathrm{P}\big(y_{kj}^n = 1 | \mathbf{s}^n\big)$ can be computed by noting that the pdf of x_{kj}^n is normal with mean $\mathbf{h}_k^T \mathbf{s}^n$ and variance r_{kj},

$$\mathrm{P}\left(y_{kj}^n = 1 | \mathbf{s}^n\right) = \frac{1}{\sqrt{2\pi r_{kj}}} \int_{\theta_{0,kj}}^{\infty} \exp\left(-\frac{1}{2}(x - \mathbf{h}_k^T \mathbf{s}^n) r_{kj}^{-1} (x - \mathbf{h}_k^T \mathbf{s}^n)\right) dx. \tag{20}$$

The resulting primal iteration in (13) then takes the form

$$\mathbf{s}_k(t) = \operatorname*{argmax}_{\mathbf{s}_k} \sum_{n=t-W+1}^{t} \ln \mathrm{P}\left(\mathbf{y}_k^n | \mathbf{s}^n\right) + \frac{1}{K}(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1})^T \mathbf{Q}^{-1} (\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1}) + \sum_{l \in n_k} \mathbf{s}_k^{n^T}(\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)), \tag{21}$$

with $\ln \mathrm{P}\left(\mathbf{y}_k^n | \mathbf{s}^n\right)$ as given in (19). The dual iteration is given by (14). It is not possible to get a closed form expression for the primal iteration, but the maximand in (21) is a concave function of $\mathbf{s}^n$. The maximum arguments $\mathbf{s}_k(t)$ can then be numerically determined using Newton's method.

III. Convergence Properties

To determine the optimality of (13)-(14), we want to assess how the D-MAP algorithm compares to the centralized MAP. Therefore we want to compare the distance $\|\mathbf{s}_k(t)-\mathbf{s}_{\text{MAP}}(t)\|$ between the primal iterate $\mathbf{s}_k(t)$ computed by sensor k at time t with the corresponding centralized MAP estimator $\mathbf{s}_{\text{MAP}}(t)$. Given the equivalence of (6) and (7) we know that $\mathbf{s}_{\text{MAP}}(t) = \mathbf{s}_k^*(t)$ from where it follows that the distance of interest satisfies

$$\|\mathbf{s}_k(t) - \mathbf{s}_{\text{MAP}}(t)\| = \|\mathbf{s}_k(t) - \mathbf{s}_k^*(t)\| \leq \|\mathbf{s}(t) - \mathbf{s}^*(t)\|. \quad (22)$$

The rightmost term in (22) is the distance between the current primal iterate $\mathbf{s}(t)$ and the optimal primal arguments $\mathbf{s}^*(t)$ of (7). As such it can be related to the distance between the current dual iterate $\boldsymbol{\lambda}(t)$ and the set of optimal dual variables $\boldsymbol{\Lambda}^*(t)$. This section is devoted to the characterization of the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$ between $\boldsymbol{\lambda}(t)$ and a specific sequence of optimal dual variables $\boldsymbol{\lambda}_0^*(t) \in \boldsymbol{\Lambda}^*(t)$. The derivation of these results requires making the following assumptions on the edge incidence matrix $\mathbf{C}_k$, the log-likelihood functions $f_0(\mathbf{s}, t)$ and the initial Lagrange multipliers $\boldsymbol{\lambda}(0)$.

(A1) The sensor network is connected. Equivalently, the edge-incidence matrix $\mathbf{C}_k$ has $K-1$ nonzero singular values $0 < \gamma_2 = \gamma \leq \cdots \leq \gamma_K = \Gamma$.

(A2) The eigenvalues of the Hessians $\nabla^2 f_0(\mathbf{s}, t)$ of the distributed log-likelihood functions $f_0(\mathbf{s}, t)$ are upper bounded by the Lipschitz constant $1/m$ so that for arbitrary vectors $\mathbf{s}$ and $\mathbf{r}$ and all times t we can write

$$f_0(\mathbf{s}, t) \leq f_0(\mathbf{r}, t)+\nabla f_0(\mathbf{r}, t)^T(\mathbf{s}-\mathbf{r})+\frac{1}{2m}\|\mathbf{s}-\mathbf{r}\|^2. \quad (23)$$

(A3) The eigenvalues of the Hessians $\nabla^2 f_0(\mathbf{s}, t)$ of the distributed log-likelihood functions $f_0(\mathbf{s}, t)$ are lower bounded by the strong convexity constant $1/M$ so that for arbitrary vectors $\mathbf{s}$ and $\mathbf{r}$ and all times t it holds

$$f_0(\mathbf{s}, t) \geq f_0(\mathbf{r}, t)+\nabla f_0(\mathbf{r}, t)^T(\mathbf{s}-\mathbf{r})+\frac{1}{2M}\|\mathbf{s}-\mathbf{r}\|^2. \quad (24)$$

(A4) The Lagrange multipliers are initialized at some value $\boldsymbol{\lambda}(0) \in \text{Im}(\mathbf{C}^T)$ in the image of the transpose edge incidence matrix $\mathbf{C}^T$.

(A5) Consider the gradients $\nabla f_0(\mathbf{s}^*(t), t)$ and $\nabla f_0(\mathbf{s}^*(t+1), t+1)$ 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 $\mathbf{s}^*(t)$ and $\mathbf{s}^*(t+1)$. The expected value of the norm of this difference given past observations up to time t is bounded by a vanishing constant $\delta(T_s)$. Denoting by $\mathbf{x}(0:t) = \mathbf{x}^0 \ldots \mathbf{x}^t$ the past observations, it holds

$$\lim_{T_s\to 0} \mathbb{E}\left[\left\|\nabla f_0\big(\mathbf{s}^*(t), t\big) - \nabla f_0\big(\mathbf{s}^*(t+1), t+1\big)\right\| \;\middle|\; \mathbf{x}(0:t)\right] \leq \delta(T_s), \quad (25)$$

for some $\delta(T_s)$ function with $\lim_{T_s\to 0}\delta(T_s) = 0$.

Assumption (A1) is typical in distributed algorithms, where γ^2 is the spectral gap of the network graph characterizing the diffusion of information in distributed algorithms. Assumptions (A2) and (A3) are customary in the analysis of descent algorithms except that we require them of the primal objectives $f_0(\mathbf{s}, t)$ while we descend on the dual functions $g(\boldsymbol{\lambda}, t)$. They can be translated into similar statements of the dual Hessian using the extremal singular values γ and Γ. Assumption (A4) is easy to ensure as it suffices to make $\boldsymbol{\lambda}(0) = \mathbf{0}$. Assumption (A5) limits the variability of the log-likelihood function $f_0(\mathbf{s}, t)$. This is a reasonable requirement because descending along the gradient $\nabla g(\boldsymbol{\lambda}(t), t)$ of the dual function $g(\boldsymbol{\lambda}, t)$ corresponding to time t is sensible only if this function is close to the dual function $g(\boldsymbol{\lambda}, t+1)$ corresponding to time $t+1$. Having close dual functions can be satisfied if the primal functions $f_0(\mathbf{s}, t)$ and $f_0(\mathbf{s}, t+1)$ are close.

The main result of this paper concerns the difference between dual iterates $\boldsymbol{\lambda}(t)$ to some optimal dual variables $\boldsymbol{\lambda}_0^*(t) \in \boldsymbol{\Lambda}^*(t)$. Theorem 1 states two asymptotic stochastic bounds on the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$ (see [1] for the proof). The first is a mean bound that holds across ensemble averages, and the second bound holds almost surely for individual realizations.

Theorem 1 *Let $\boldsymbol{\lambda}(t)$ denote the vector with current dual iterates obtained at time t from* (14) *and $\boldsymbol{\lambda}_0^*(t) \in \text{Im}(\mathbf{C}^T)$ denote the unique optimal argument of the dual function $g(\boldsymbol{\lambda}, t)$ that lies in the image of the transposed replicated edge incidence matrix $\mathbf{C}^T$. Assume the step size $\epsilon < 1/\Gamma^2 M$. If assumptions (A1)-(A5) hold, the expected value of the distance between the dual multipliers $\boldsymbol{\lambda}(t)$ and the optimal multipliers $\boldsymbol{\lambda}_0^*(t)$ at time t satisfies*

$$\liminf_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|\right] \leq \frac{1+\epsilon m\gamma^2}{\epsilon m\gamma^3}\delta(T_s). \quad (26)$$

Furthermore, for almost all realizations of the observation process $\mathbf{x}(t)$ it holds

$$\liminf_{t\to\infty} \|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\| \leq \frac{1+\epsilon m\gamma^2}{\epsilon m\gamma^3}\delta(T_s) \quad a.s. \quad (27)$$

The first result in (26) states that the mean across different realizations of the process $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$ 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 (25) vanishes with decreasing sampling time. It is therefore possible to approximate $\boldsymbol{\lambda}_0^*$ 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 $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$.

Coming back to the original goal, the distance between D-MAP estimates $\mathbf{s}(t)$ and MAP estimates $\mathbf{s}^*(t)$ can be bounded by the dual suboptimality distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$. Combining this result with the bounds in Theorem 1 characterizes the steady state behavior of D-MAP. D-MAP estimates are close to MAP estimates on average [cf. (26)] and for almost all realizations of the dynamic system of interest D-MAP estimates are close to MAP estimates infinitely often [cf. (27)]. The bound depends on the condition number m/M of the primal objective, the spectral radius γ of the edge incidence matrix, and the objective smoothness parameter $\delta(T_s)$.

IV. Simulation Results

We implement the D-MAP algorithm in (13)-(14) for the quantized observations model of Section II-B. We compare

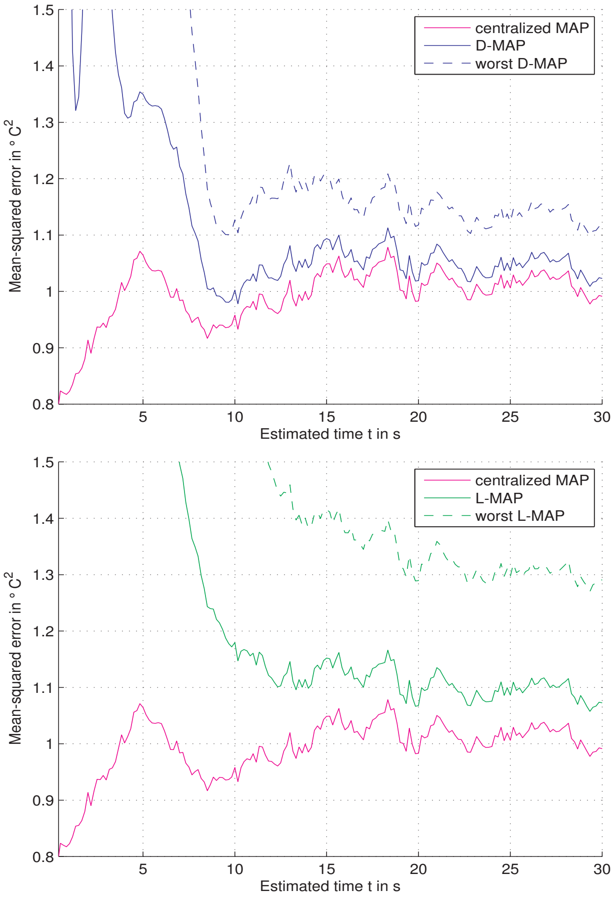

Fig. 1. Empirical mean squared error (MSE) for centralized MAP, D-MAP, and L-MAP and worst empirical MSE for times $\tau \in [0, 30\,\mathrm{s}]$, averaged over 10^3 simulation runs for a binary quantized model. The empirical MSE of D-MAP estimates is closer than that of L-MAP estimates to the empirical MSE of centralized MAP estimates.

performance of D-MAP estimates $\mathbf{s}_k(t)$ to the centralized MAP estimator $\mathbf{s}_{\mathrm{MAP}}(t)$ in (6), which would be computed if all observations were available at a common location, and local (L-) MAP estimates $\hat{\mathbf{s}}_k(t)$ computed using local observations only. Consider a WSN with $K = 8$ sensors and edges between any two sensors k and l present with probability $1/2$. Sensors collect quantized binary observations as dictated by the model in Section II-B. The signal $\mathbf{s}(\tau) = s(\tau)$ is a scalar temperature reading and the parameters of the linear model serving as basis to the quantized model correspond to the state transition matrix $\mathbf{A}_a = a_a - 0.01/\mathrm{s}$, signal noise variance $\mathbf{Q}_a = q_a = 0.5°C^2/\mathrm{s}^2$, observation noise covariances $\mathbf{R}_{ak} = r_{ak} = 1°C^2$ for every sensor k, and observation matrices $\mathbf{H}_{ak} = h_{ak} = 1$ for all sensors k. We set the sampling time to $T_s = 0.166\,\mathrm{s}$ and the initial temperature to $s(0) = 20°C$. Quantization thresholds are set to $\theta_{0,k1} = \theta_{0,k} = 20°C$ for all sensors k. The system is simulated for 180 observation slots corresponding to a total elapsed time of 30 s. The estimation window is again set to $W = 3$. For D-MAP the Lagrange multipliers are initialized to $\boldsymbol{\lambda}_{kl}(0) = 0$ for all links (k, l). D-MAP estimates are computed according to (21). The stepsize for D-MAP for each edge (k, l) and signal j set to 0.1 times the inverse of its respective diagonal entry in the dual Hessian of the corresponding Gaussian linear model.

Fig. 1 shows simulation results for the described setup. Fig. 1 compares the MSEs over 10^3 simulation runs. The empirical MSE of the centralized MAP is shown along with the average empirical MSEs of D-MAP (left) and L-MAP (right) as well as the worst empirical MSE for D-MAP and L-MAP for times $\tau \in [0, 30\,\mathrm{s}]$. At steady state the MSE of the centralized MAP is $1.00°C^2$. The steady state MSE of D-MAP is about $1.05°C^2$ whereas it is $1.1°C^2$ for the L-MAP. A better transient behavior of the D-MAP can also be observed by the time it takes to reach the steady state MSE which is 9 s for the D-MAP and 12 s for the L-MAP. This improvement in performance is stronger for the worst empirical MSE. While the worst empirical MSE for D-MAP attains a steady-state value of $1.15°C^2$ after time $\tau = 10\,\mathrm{s}$, the L-MAP takes $\tau = 20\,\mathrm{s}$ to approach a worst steady-state MSE of $1.3°C^2$.

V. 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.

References

[1] F. Jakubiec and A. Ribeiro. D-map: Distributed maximum a posteriori probability estimation of dynamic systems. *IEEE Trans. Signal Process.*, 2012 (to appear).

[2] S. Kar and J.M.F. Moura. Distributed consensus algorithms in sensor networks: Quantized data and random link failures. *IEEE Transactions on Signal Processing*, 58(3):1383–1400, 2010.

[3] T.C. Aysal, M.J. Coates, and M.G. Rabbat. Distributed average consensus with dithered quantization. *IEEE Transactions on Signal Processing*, 56(10):4905–4918, 2008.

[4] R. Olfati-Saber, J.A. Fax, and R.M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

[6] C.G. Lopes and A.H. Sayed. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7):3122–3136, 2008.

[7] M.G. Rabbat, R.D. Nowak, and J.A. Bucklew. Generalized consensus computation in networked systems with erasure links. In *IEEE Workshop on Signal Processing Advances in Wireless Communications Proc.*, pages 1088–1092, June 2005.

[8] I.D. Schizas, A. Ribeiro, and G.B. Giannakis. Consensus in ad hoc WSNs with noisy links - Part I: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, 2008.

[9] A. Ribeiro, I.D. Schizas, S.I. Roumeliotis, and G.B. Giannakis. Kalman filtering in wireless sensor networks. *Control Systems Magazine, IEEE*, 30(2):66–86, 2010.

[10] A. F. Sha, A. Ribeiro, and G. Giannakis. Bandwidth-constrained map estimation for wireless sensor networks. In *Proc. Asilomar Conf. Signals Systems Computers*, pages 215–219. Pacific Grove CA, Oct. 28 - Nov. 1 2005.