D-MAP: Distributed Maximum a Posteriori Probability Estimation of Dynamic Systems

Felicia Y. Jakubiec and Alejandro Ribeiro

***Abstract*—This paper develops a framework for the estimation of a time-varying random signal using a distributed 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 which is assumed to be log-concave satisfies a smoothness condition. We implement the D-MAP algorithm for a linear and a nonlinear system model to show that the performance corroborates with theoretical findings.**

***Index Terms*—Distributed estimation, wireless sensor networks.**

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. Consensus relies on iterative averaging of neighboring values and can be shown to determine linear minimum (LM) mean squared error (MSE) estimators [2]. Consensus algorithms are well studied for linear static estimation problems, e.g., [3]–[5], and have also been adapted for linear dynamic estimation [6]–[8]. Variants of consensus algorithms include the concept of running consensus in which consensus iterations are performed alongside the collection of sensor data [9], [10] and gossip algorithms in which data exchanges happen between pairs of neighbors only [11], [12]. Being based on linear operations, gossip algorithms solve LMMSE estimation and have been extended to dynamic settings as well [13], [14]. A drawback of most consensus methods in [6]–[8], [13], [14] for estimation of time varying signals is the assumption that communications occur in a time scale separate from the timeline of the dynamic system. This is necessary because consensus [2] and gossip [11] are iterative algorithms. Thus, their implementation in a dynamic setting requires assuming that an infinite number of communication steps occur between subsequent states of the dynamic system. An approach that doesn't suffer from this drawback is the application of diffusion algorithms [15] to online Kalman filtering [16] and target tracking problems [17]. In the particular case of target tracking, specific algorithms have also been developed in the context of robot localization[18]–[20].

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 either dual gradient descent [21] or the alternating direction method of multipliers [22]. This latter approach is adapted in [23] to deal with linear Gaussian autoregressive (AR) models giving rise to distributed Kalman filter implementations. Alternative constructions of distributed Kalman filters rely on sensor scheduling and focus on the effects of quantization [24], [25]—see also [26] for a tutorial presentation of Kalman filtering in distributed sensor networks. An important advantage of the price mediation methods in [21], [22] is that they can be used for general, i.e., not necessarily linear, maximum likelihood estimation problems. The generalization to dynamic systems in [23], however, is restricted to linear estimation.

This paper generalizes the price mediation algorithms of [21], [22] 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

Manuscript received April 25, 2012; revised August 23, 2012; accepted September 12, 2012. Date of publication October 03, 2012; date of current version December 25, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefano Marano. Work in this paper is supported by NSF CCF-1017454 and AFOSR MURI FA9550-10-1-0567. Part of the results in this paper were presented at International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, March 25–30, 2012 [1].

The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: life@seas.upenn.edu; aribeiro@seas.upenn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2012.2222398

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. To avoid memory growth we introduce a time window to remove older observations and signals from the estimation problem. The global MAP is then reformulated as a constrained optimization problem in which local estimates are required to coincide with each other. Under the assumption that log-likelihood functions are concave, the 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 linear Gaussian AR process in Section II-B and to a nonlinear variant in which estimates rely on quantized observations in Section II-C.

Convergence properties of D-MAP are studied in Section III. Since we implement gradient descent in the dual domain, our focus is to study the distance between dual iterates and the optimal dual variables. The proximity between D-MAP and centralized MAP estimates is proportional to this distance as we show in Theorem 2. In each step of dual gradient descent the dual iterate is pulled closer to the optimal dual variable. However, the optimal multiplier changes between iterations due to variations in the signal of interest. Once the algorithm reaches steady state, we expect the distance between dual iterates and the time-varying optimal dual variable to settle on some gap. This expectation is formalized in Theorem 1 where 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. The size of the optimality neighborhood is characterized in terms of the condition number of the log-likelihood function, the connectedness of the sensor network, and a parameter bounding the changes in the gradient of the log-likelihood as a function of time. This result implies that the stochastic process of distances between centralized MAP and D-MAP estimates becomes small on average [cf. (i)] and that for almost all realizations the distances become small infinitely often [cf. (ii)]. If the log-likelihood is smooth in time in the sense that the parameter bounding the changes in its gradient vanishes with time, these neighborhoods become arbitrarily small by proper selection of the sampling time. The proof of Theorem 1 is presented in Section III-A.

Numerical experiments are presented in Section IV. Section IV-A implements D-MAP for the linear model of Sections II-B and IV-B for the quantized observations model of Section II-C. In both cases D-MAP results in smaller MSE than estimates that rely on local observations only. The advantage is most noticeable when comparing the worst MSE across different sensors in a given realization. Section V closes the paper with concluding remarks.

II. Problem Formulation

Consider a symmetric sensor network $\mathcal{G}$ with K sensors and let n_k denote the set of neighbors of sensor k composing an edge set $\mathcal{E}$ with $|\mathcal{E}|$ edges. 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) $\mathrm{P}\left(\mathbf{x}_a(\tau)|\mathbf{s}_{ak}(\tau)\right)$ is known at each sensor. We further assume that the process of time-varying signal values $\mathbf{s}_a(\tau)$ can be described by 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 $\mathrm{P}\left(\mathbf{s}_a(\tau+h)|\mathbf{s}_a(\tau)\right)$. 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 $\ln \mathrm{P}\left(\mathbf{s}_a(\tau+h)\,|\mathbf{s}_a(\tau)\right)$ are concave functions of the signal values $\mathbf{s}_a(\tau)$ and $\mathbf{s}_a(\tau+h)$.

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 $\mathbf{x}^n = [{\mathbf{x}_1^n}^T, \ldots, {\mathbf{x}_K^n}^T]^T$ 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. Observe that this probabilistic description of the discrete time model is obtained from the continuous time model introduced above. 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. To clarify 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 $\mathbf{x}_k(t) := [{\mathbf{x}_k^{t-W+1}}^T \ldots {\mathbf{x}_k^t}^T]^T$ 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 $\mathbf{x}(t) := [\mathbf{x}_1^T(t), \ldots, \mathbf{x}_K^T(t)]^T$ grouping observations for all sensors and all times during the window. 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. With this notational convention

define the vector $\mathbf{s}_{\mathrm{MAP}}(t) = [\mathbf{s}_{\mathrm{MAP}}^{t-W+1}(t)^T \ldots \mathbf{s}_{\mathrm{MAP}}^{t}(t)^T]^T$ of all MAP estimates between times $t-W+1$ and t as

$$\begin{aligned}\mathbf{s}_{\mathrm{MAP}}(t) &= \arg\max_{\mathbf{s}} \mathrm{P}\left(\mathbf{s}|\mathbf{x}(t)\right) \\ &= \arg\max_{\mathbf{s}} \mathrm{P}\left(\mathbf{x}(t)|\mathbf{s}\right)\mathrm{P}\left(\mathbf{s}\right), \end{aligned} \tag{2}$$

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

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

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

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

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

$$\mathbf{s}_{\mathrm{MAP}}(t) = \arg\max_{\mathbf{s}} \prod_{n=t-W+1}^{t}\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)\prod_{k=1}^{K}\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right). \tag{5}$$

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

$$\begin{aligned}\mathbf{s}_{\mathrm{MAP}}(t) &= \arg\max_{\mathbf{s}} f_{(\mathrm{MAP})}(\mathbf{s},t) \\ &= \arg\max_{\mathbf{s}} \sum_{n=t-W+1}^{t}\Bigg(\sum_{k=1}^{K}\left(\ln\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)\right) \\ &\qquad + \ln\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)\Bigg), \end{aligned} \tag{6}$$

where we defined the function $f_{(\mathrm{MAP})}(\mathbf{s},t)$ to denote the centralized log-likelihood function at time t whose maximization yields MAP estimates $\mathbf{s}_{\mathrm{MAP}}(t)$. Throughout the text we omit summation borders to simplify notation. We use $\sum_n(\cdot)$ to stand for $\sum_{n=t-W+1}^{t}(\cdot)$, $\sum_k(\cdot)$ for $\sum_{k=1}^{K}(\cdot)$ and $\sum_{n,k}(\cdot)$ for the joint sum $\sum_{n=t-W+1}^{t}\sum_{k=1}^{K}(\cdot)$.

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 $f_{(\mathrm{MAP})}(\mathbf{s},t)$ 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 $f_{(\mathrm{MAP})}(\mathbf{s},t)$ 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 $\mathbf{s}_{\mathrm{MAP}}(t)$ 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

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

where we introduced the vector $\mathbf{s}^*(t)$ stacking local estimates $\mathbf{s}_k^{n*}(t)$ for all sensors and times. Formulating an equivalent constrained problem with additional variables is common practice in distributed optimization. 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, if the centralized problem in (6) has a solution $\mathbf{s}_{\mathrm{MAP}}(t)$, the solution to the constrained optimization problem (7) is the same. In other words, (7) and (6) are equivalent because every element $\mathbf{s}_k^*(t)$ of the estimator in (7) is equal to the MAP estimator, i.e., $\mathbf{s}_k^*(t) = \mathbf{s}_{\mathrm{MAP}}(t)$.

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 $\mathbf{C}\mathbf{s} = \mathbf{0}$. Further defining local objectives $f_k(\mathbf{s}_k,t) = \sum_n \ln\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + (1/K)\ln\mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right)$ and global D-MAP objectives $f_0(\mathbf{s},t) = \sum_k f_k(\mathbf{s}_k,t)$ we can rewrite (7) as

$$\begin{aligned}\mathbf{s}^*(t) = \arg\max_{\mathbf{s}} \quad & f_0(\mathbf{s},t) = \sum_k f_k(\mathbf{s}_k,t), \\ \text{s.t.} \quad & \mathbf{C}\mathbf{s} = \mathbf{0}. \end{aligned} \tag{8}$$

Since the equality constraints are linear and the maximand is concave in the variables $\mathbf{s}_k^n$, the optimization problem in (7)—and its equivalent form in (8)—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

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

where $\boldsymbol{\lambda}$ stacks the Lagrange multipliers for all links k, l and times n. Recall that $\sum_{n,k}(\cdot)$ stands for the sum $\sum_{n=t-W+1}^{t}\sum_{k=1}^{K}(\cdot)$ as already convened.

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, i.e.,

$$g(\boldsymbol{\lambda},t) = \arg\max_{\mathbf{s}} \mathcal{L}(\mathbf{s},\boldsymbol{\lambda},t), \tag{10}$$

and the dual problem is defined as the minimization of the dual function $D(t) = \min_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}, t)$. More important than the minimum value of the dual function is the optimal dual argument that achieves this minimum. The structure of the primal problem in (7) is such that this minimizing argument is not unique. Define then the *set* of optimal Lagrange multipliers as

$$\boldsymbol{\Lambda}^*(t) = \arg\min_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}, t), \tag{11}$$

and denote as $\boldsymbol{\lambda}^*(t) \in \boldsymbol{\Lambda}^*(t)$ the elements of this set. The set $\boldsymbol{\Lambda}^*(t)$ is a subspace whose dimension is determined by the rank of the replicated edge incidence matrix $\mathbf{C}$; see (36)–(38).

Because the dual function is convex, we can use gradient descent 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

$$\boldsymbol{\lambda}(t+1) = \boldsymbol{\lambda}(t) - \epsilon \nabla g(\boldsymbol{\lambda}(t), t) \tag{12}$$

with some given stepsize ϵ. Notice that in (12), $\boldsymbol{\lambda}(t+1)$ 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 III.

To compute the gradient of the dual function consider the Lagrangian primal maximizers $\mathbf{s}(t) = \mathbf{s}(\boldsymbol{\lambda}(t))$ that maximize the Lagrangian $\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}(t), t)$ for given dual iterate $\boldsymbol{\lambda}(t)$

$$\mathbf{s}(t) = \mathbf{s}(\boldsymbol{\lambda}(t)) := \arg\max \mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}(t), t). \tag{13}$$

The gradient of the dual function $\nabla g(\boldsymbol{\lambda}(t), t)$ is then given by the constraint slack associated with these Lagrangian maximizers

$$\nabla g(\boldsymbol{\lambda}(t), t) = \mathbf{C}\mathbf{s}(t). \tag{14}$$

According to the definition of the edge incidence matrix $\mathbf{C}$ the gradient component associated with link k, l and time n is given by the constraint slack corresponding to this link

$$\Big[\nabla g(\boldsymbol{\lambda}(t), t)\Big]_{kl}^{n} = \mathbf{s}_k^n(t) - \mathbf{s}_l^n(t). \tag{15}$$

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 $\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}, t) = \sum_k \mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}, t)$ with

$$\mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}, t) = \sum_n \Bigg[\ln \mathrm{P}\left(\mathbf{x}_k^n | \mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n | \mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} \mathbf{s}_k^{nT}(t)(\boldsymbol{\lambda}_{kl}^n - \boldsymbol{\lambda}_{lk}^n) \Bigg]. \tag{16}$$

Since separate maximization of the local Lagrangians in (16) 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 (15) can be determined in a distributed manner. This permits the definition of the D-MAP algorithm which we formulate as iterative application of the following steps.

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

$$\mathbf{s}_k(t) = \arg\max_{\mathbf{s}_k} \sum_n \Bigg[\ln \mathrm{P}\left(\mathbf{x}_k^n | \mathbf{s}_k^n\right) + \frac{1}{K} \ln \mathrm{P}\left(\mathbf{s}_k^n | \mathbf{s}_k^{n-1}\right) + \sum_{l \in n_k} \mathbf{s}_k^{nT}(\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)) \Bigg]. \tag{17}$$

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

$$\boldsymbol{\lambda}_{kl}^n(t+1) = \boldsymbol{\lambda}_{kl}^n(t) - \epsilon\big(\mathbf{s}_k^n(t) - \mathbf{s}_l^n(t)\big). \tag{18}$$

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.

Remark 1: In (17) the signal estimate $\mathbf{s}_k^{t-W}(t-1)$ computed at time $t-1$ for the signal value at time $t-W$ is not used. Nevertheless, $\mathbf{s}_k^{t-W}(t-1)$ contains information about the signal values $\mathbf{s}_k^{t-W+1}, \ldots, \mathbf{s}_k^t$ in the current window of interest. To exploit this information the term $\left[\mathbf{s}_k^{t-W+1} - \mathbf{A}\mathbf{s}_k^{t-W}(t-1)\right]^T \mathbf{Q}^{-1}\left[\mathbf{s}_k^{t-W+1} - \mathbf{A}\mathbf{s}_k^{t-W}(t-1)\right]$ can be added to the local Lagrangian in (17). This is equivalent to assuming that $\mathbf{s}^{t-W+1}$ given $\mathbf{s}^{t-W}(t-1)$ is normal with covariance $\mathbf{Q}$ and mean $\mathbf{A}\mathbf{s}_k^{t-W}(t-1)$. This term is not considered here to keep the analysis in Section III tractable but it should be added to practical implementations. We consider the effect of this term in the numerical simulations performed in Section IV.

B. Linear Gaussian Autoregressive Model

To illustrate the D-MAP algorithm in (17) and (18) consider its application to a linear time invariant Gaussian AR model. In this case the evolution of the state $\mathbf{s}_a(\tau)$ follows a linear differential equation and the observation $\mathbf{x}_{ak}(\tau)$ is a noisy linear transformation of the state,

$$\dot{\mathbf{s}}_a(\tau) = \mathbf{A}_a\, \mathbf{s}_a(\tau) + \mathbf{u}_a(\tau), \tag{19}$$

$$\mathbf{x}_{ak}(\tau) = \mathbf{H}_{ak}\, \mathbf{s}_a(\tau) + \mathbf{n}_{ak}(\tau). \tag{20}$$

The driving noise $\mathbf{u}_a(\tau)$ in (19) is drawn from a zero-mean Wiener process with covariance matrix $\mathbf{Q}_a$, and the observation noise $\mathbf{n}_a(\tau)$ is drawn from a zero-mean Wiener process with covariance matrix $\mathbf{R}_{ak}$.

An equivalent discrete-time model tracks the state $\mathbf{s}^n = \mathbf{s}_a(nT_s)$ at times nT_s using sampled observations $\mathbf{x}_k^n = \mathbf{x}_{ak}(nT_s)$ [27, Chapter 4.9]. Solving the differential (19) between times $(n-1)\,T_s$ and nT_s with initial condition $\mathbf{s}^{n-1}$ we can relate subsequent state observations $\mathbf{s}^{n-1}$ and $\mathbf{s}^n$ as

$$\mathbf{s}^n = \exp(\mathbf{A}_a T_s)\mathbf{s}^{n-1} + \int_{(n-1)\,T_s}^{n\,T_s} \exp(\mathbf{A}_a(nT_s - \tau))\mathbf{u}_a(\tau)d\tau. \tag{21}$$

Upon defining the noise vector $\mathbf{u}^n := \int_{(n-1)\,T_s}^{n\,T_s} \exp(\mathbf{A}_a(n\,T_s - \tau))\mathbf{u}_a(\tau)d\tau$ as well as matrices $\mathbf{A} := \exp(\mathbf{A}_a\,T_s)$ and $\mathbf{H}_k := \mathbf{H}_{ak}$ we can write the equivalent discrete time AR process as

$$\mathbf{s}^n = \mathbf{A}\,\mathbf{s}^{n-1} + \mathbf{u}^{n-1}, \tag{22}$$

$$\mathbf{x}_k^n = \mathbf{H}_k\,\mathbf{s}^n + \mathbf{n}_k^n. \tag{23}$$

It follows from its definition that the driving input noise $\mathbf{u}^n$ is white Gaussian with covariance matrix

$$\mathbf{Q} := \mathbb{E}\left[\mathbf{u}^{n^T}\mathbf{u}^n\right] = \left(\frac{\mathbf{Q}_a}{2}\right)\mathbf{A}_a^{-1}(\exp(2\mathbf{A}_a\,T_s) - \mathbf{I}). \tag{24}$$

This observation combined with (22) implies that the transition probability distribution $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)$ is normal with mean $\mathbf{s}^{n-1}$ and covariance $\mathbf{Q}$. To sample observations with period T_s we need to pass $\mathbf{x}_{ak}(\tau)$ through a low pass filter with bandwidth $\frac{1}{T_s}$. Assuming this filter is perfect, the discrete time noise $\mathbf{n}_k^n$ is white Gaussian with covariance matrix

$$\mathbf{R}_k := \mathbb{E}\left[\mathbf{n}_k^{n^T}\mathbf{n}_k^n\right] = \frac{\mathbf{R}_{ak}}{T_s}. \tag{25}$$

This fact combined with (23) implies that $\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)$ is also normal with mean $\mathbf{H}_k\mathbf{s}^n$ and covariance $\mathbf{R}_k$. Given that both $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)$ and $\mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)$ are normal the log-likelihoods in (6) are quadratic and the centralized MAP estimator $\mathbf{s}_{\mathrm{MAP}}(t)$ in (6) reduces to the maximization of the quadratic form

$$\begin{aligned}\mathbf{s}_{\mathrm{MAP}}(t) &= \arg\max_{\mathbf{s}} -\sum_{n,k}\Big[\Big(\mathbf{x}_k^n - \mathbf{H}_k\mathbf{s}^n\Big)^T\mathbf{R}_k^{-1}\Big(\mathbf{x}_k^n - \mathbf{H}_k\mathbf{s}^n\Big)\\ &\quad + \frac{1}{K}\Big(\mathbf{s}^n - \mathbf{A}\mathbf{s}^{n-1}\Big)^T\mathbf{Q}^{-1}\Big(\mathbf{s}^n - \mathbf{A}\mathbf{s}^{n-1}\Big)\Big].\end{aligned} \tag{26}$$

For the D-MAP algorithm we need to specify the primal and dual iterations in (17) and (18). For the primal iteration just observe that the log-likelihoods are quadratic forms as in (26) to conclude that (17) takes the specific form

$$\begin{aligned}\mathbf{s}_k(t) = \arg\max_{\mathbf{s}_k} -\sum_n\Big[&\Big(\mathbf{x}_k^n - \mathbf{H}_k\mathbf{s}_k^n\Big)^T\mathbf{R}_k^{-1}\Big(\mathbf{x}_k^n - \mathbf{H}_k\mathbf{s}_k^n\Big)\\ &+ \Big(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1}\Big)^T\mathbf{Q}^{-1}\Big(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1}\Big)\\ &+ \sum_{l\in n_k}\mathbf{s}_k^{nT}\Big(\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)\Big)\Big].\end{aligned} \tag{27}$$

The dual iteration is given by (18) since its form does not depend on the specific signal model. Since the maximand in (27) is a quadratic form, the distributed estimate $\mathbf{s}_k(t)$ can be computed in closed form by determining the estimate that sets the gradient of the quadratic form to zero.

C. Quantized Observations

Consider a modification of the linear model in (22)–(23) in which 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 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 (23). The binary variable y_{kj}^n indicates whether the analog observation x_{kj}^n exceeds the threshold $\theta_{0,kj}$,

$$y_{kj}^n = \mathbb{I}\left\{x_{kj}^n \geq \theta_{0,kj}\right\}, \tag{28}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. For simplicity of exposition assume the observation noise is uncorrelated so that the covariance matrix $\mathbf{R}_k$ takes on the diagonal form $\mathbf{R}_k = \mathrm{diag}(r_{k1}, r_{k2}\ldots, r_{kJ})$. This assumption makes the observations x_{kj}^n, and as consequence the binary variables y_{kj}^n, independent of each other. It follows that to determine the log-likelihood $\ln \mathrm{P}\left(\mathbf{y}_k^n|\mathbf{s}^n\right)$ we just need to determine the probabilities $\mathrm{P}\left(y_{kj}^n = 1|\mathbf{s}^n\right)$ so as to compute

$$\begin{aligned}\ln \mathrm{P}\left(\mathbf{y}_k^n|\mathbf{s}^n\right) = \sum_{j=1}^J\Big(&y_{kj}^n \ln \mathrm{P}\left(y_{kj}^n = 1|\mathbf{s}^n\right)\\ &+(1 - y_{kj}^n)\ln\left(1 - \mathrm{P}\left(y_{kj}^n = 1|\mathbf{s}^n\right)\right)\Big).\end{aligned} \tag{29}$$

To determine $\mathrm{P}\left(y_{kj}^n = 1|\mathbf{s}^n\right)$, let $\mathbf{h}_k^T$ denote the k-th row of the observation matrix $\mathbf{H}_k$. According to (28), $y_{kj}^n = 1$ is equivalent to $x_{kj}^n \geq \theta_{0,kj}$. Since the pdf of x_{kj}^n is normal with mean $\mathbf{h}_k^T\mathbf{s}^n$ and variance r_{kj} we can write this probability as

$$\begin{aligned}&\mathrm{P}\left(y_{kj}^n = 1|\mathbf{s}^n\right)\\ &= \frac{1}{\sqrt{2\pi r_{kj}}}\int_{\theta_{0,kj}}^{\infty}\exp\left(-\frac{1}{2}(x - \mathbf{h}_k^T\mathbf{s}^n)r_{kj}^{-1}(x - \mathbf{h}_k^T\mathbf{s}^n)\right)dx.\end{aligned} \tag{30}$$

We remark that since the integrand in (30) is a log-concave function of x, the resulting integral $\mathrm{P}\left(y_{kj}^n = 1|\mathbf{s}^n\right)$ is also a log-concave function of $\mathbf{s}^n$[28, p. 106]. This implies that the log-likelihood $\ln \mathrm{P}\left(\mathbf{y}_k^n|\mathbf{s}^n\right)$ in (29) is a concave function of $\mathbf{s}^n$ and thus consistent with the assumptions in Section II.

To write the primal iteration for D-MAP we also need to determine the transition probability distribution $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)$. But since we have not changed the signal model, $\mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right)$ is normal with mean $\mathbf{s}^{n-1}$ and covariance $\mathbf{Q}$ as commented after (24). The primal iteration in (17) then takes the form

$$\begin{aligned}\mathbf{s}_k(t) = \arg\max_{\mathbf{s}_k}&\sum_n \ln \mathrm{P}\left(\mathbf{y}_k^n|\mathbf{s}^n\right)\\ &+ \frac{1}{K}(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1})^T\mathbf{Q}^{-1}(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1})\\ &+ \sum_{l\in n_k}\mathbf{s}_k^{nT}(\boldsymbol{\lambda}_{kl}^n(t) - \boldsymbol{\lambda}_{lk}^n(t)),\end{aligned} \tag{31}$$

with $\ln \mathrm{P}\left(\mathbf{y}_k^n|\mathbf{s}^n\right)$ as given in (29). The dual iteration is again given by (18) as in the case of the linear Gaussian AR model of Section II-B because its form is the same irrespective of the particular signal model. In this case it is not possible to get a closed form expression for the primal iteration. However, since $\ln \mathrm{P}\left(\mathbf{y}_k^n|\mathbf{s}^n\right)$ is a concave function of $\mathbf{s}_k^n$, the maximand in (31) is a concave function of $\mathbf{s}^n$. The maximum arguments $\mathbf{s}_k(t)$ can then be numerically determined using Newton's method as shown in Appendix A. We emphasize that Newton descent for (31) is implemented locally at each sensor. There is no coordination between neighboring sensors other than the exchange of Lagrange multipliers $\boldsymbol{\lambda}_{lk}(t)$ and primal iterates $\mathbf{s}_k(t)$.

The model in this section can be generalized to multilevel quantizers. Colored observation noise can be handled with the use of a whitening filter. See [29] for further details.

III. Convergence Properties

To determine the optimality of (17)–(18), we want to assess how the D-MAP algorithm compares to the centralized MAP. Therefore we want to compare the distance $\|\mathbf{s}_k(t) - \mathbf{s}_{\mathrm{MAP}}(t)\|$

between the primal iterate $\mathbf{s}_k(t)$ computed by sensor k at time t with the corresponding centralized MAP estimator $\mathbf{s}_{\rm MAP}(t)$. Given the equivalence of (6) and (7) we know that $\mathbf{s}_{\rm MAP}(t) = \mathbf{s}^*(t)$ from where it follows that the distance of interest satisfies

$$\|\mathbf{s}_k(t) - \mathbf{s}_{\rm MAP}(t)\| = \|\mathbf{s}_k(t) - \mathbf{s}_k^*(t)\| \leq \|\mathbf{s}(t) - \mathbf{s}^*(t)\|. \quad (32)$$

where the inequality is due to the fact that the vectors $\mathbf{s}_k(t)$ and $\mathbf{s}_k^*(t)$ are components of the vectors $\mathbf{s}(t)$ and $\mathbf{s}^*(t)$. The rightmost term in (32) 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 optimality gap $\|\mathbf{s}(t) - \mathbf{s}^*(t)\|$ in (32) is then bounded by the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$. The derivation of these results requires making the following assumptions on the edge incidence matrix $\mathbf{C}$, 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 $0 < \gamma_2 \leq \ldots \leq \gamma_K$. For future reference define $\gamma := \gamma_2$ and $\Gamma := \gamma_K$.

(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

$$f_0(\mathbf{s}, t) \leq f_0(\mathbf{r}, t) + \nabla f_0(\mathbf{r}, t)^T(\mathbf{s} - \mathbf{r}) + \frac{1}{2m}\|\mathbf{s} - \mathbf{r}\|^2. \quad (33)$$

(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

$$f_0(\mathbf{s}, t) \geq f_0(\mathbf{r}, t) + \nabla f_0(\mathbf{r}, t)^T(\mathbf{s} - \mathbf{r}) + \frac{1}{2M}\|\mathbf{s} - \mathbf{r}\|^2. \quad (34)$$

(A4) The Lagrange multipliers are initialized at some value $\boldsymbol{\lambda}(0) \in \mathrm{Im}(\mathbf{C}^T)$ in the image of the transposed replicated 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

$$\lim_{T_s \to 0} \mathbb{E}\left[\left\|\nabla f_0\big(\mathbf{s}^*(t), t\big) - \nabla f_0\big(\mathbf{s}^*(t+1), t+1\big)\right\| \Big| \mathbf{x}(0:t)\right] \leq \delta(T_s), \quad (35)$$

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

Assumption (A1) is typical in distributed algorithms. Observe that the squares of the singular values of $\mathbf{C}$ are eigenvalues of the replicated Laplacian matrix $\mathbf{C}^T\mathbf{C}$. In particular γ^2 is the spectral gap of the network graph which is known to control 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 $g(\boldsymbol{\lambda}, t)$. Assumptions (A2) and (A3) can be translated into similar statements of the dual Hessian using the extremal singular values γ and Γ—see Lemma 1. We remark that the strong convexity Assumption (A3) requires, in particular, that the dimensionality of the observations $\mathbf{x}_k$ be equal or larger than the dimensionality of the signals $\mathbf{s}_k$. Assumption (A4) is a restriction in the initial multipliers which is easy to ensure as it suffices to make $\boldsymbol{\lambda}(0) = \mathbf{0}$. Selecting $\boldsymbol{\lambda}(0) \in \mathrm{Im}(\mathbf{C}^T)$ guarantees that $\boldsymbol{\lambda}(t) \in \mathrm{Im}(\mathbf{C}^T)$ for all times $t \geq 0$. This is true because $\nabla g(\boldsymbol{\lambda}, t) \in \mathrm{Im}(\mathbf{C}^T)$ as it follows from its expression in (14).

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. Observe however that (35) limits variability of the log-likelihood gradients $\nabla f_0(\mathbf{s}, t)$, which is a stronger requirement than limiting the variability of $f_0(\mathbf{s}, t)$. Further recall that functions $f_0(\mathbf{s}, t)$ are random as they depend on the observations $\mathbf{x}_k^n(t+1)$. The bound in (35) is weak in a stochastic sense as it only constrains the expected value of the difference between subsequent gradients. It is also important to note that in most cases of practical interest, the constant $\delta(T_s)$ vanishes as the sampling time $T_s \to 0$. For the linear Gaussian AR model of Section II-B $\delta(T_s) = o(\sqrt{T_s})$ vanishes as $\sqrt{T_s}$; see Appendix A. Note that $\nabla f_0\big(\mathbf{s}^*(t), t\big) \neq \mathbf{0}$ in general because $\mathbf{s}^*(t)$ solves a constrained optimization problem [cf. (7)].

To specify the sequence of given optimal dual variables $\boldsymbol{\lambda}_0^*(t) \in \boldsymbol{\Lambda}^*(t)$ let us characterize the optimal dual subspace $\boldsymbol{\Lambda}^*(t)$. According to the Karush-Kuhn-Tucker (KKT) conditions, any pair $\mathbf{s}^*(t)$, $\boldsymbol{\lambda}^*(t)$ of primal and dual optimum variables satisfies

$$\nabla_{\mathbf{s}} \mathcal{L}(\mathbf{s}^*(t), \boldsymbol{\lambda}^*(t), t) = \nabla f_0(\mathbf{s}^*(t), t) + \mathbf{C}^T\boldsymbol{\lambda}^*(t) = \mathbf{0}. \quad (36)$$

The optimal argument $\mathbf{s}^*(t)$ is unique because the log-likelihood $f_0(\mathbf{s}, t)$ is assumed strongly convex as per assumption (A3). However, solving (36) for $\boldsymbol{\lambda}^*(t)$ results in multiple possible solutions because the rank of the $KJ \times |\mathcal{E}|$ matrix $\mathbf{C}^T$ is $\mathrm{rank}(\mathbf{C}^T) = K - 1$. Then, if there is a solution to (36) we can describe the subspace $\boldsymbol{\Lambda}^*(t)$ by specifying a unique vector $\boldsymbol{\lambda}_0^*(t) \in \mathrm{Im}(\mathbf{C}^T)$ in the image of $\mathbf{C}^T$ from which all optimal multipliers $\boldsymbol{\lambda}^*(t)$ are obtained by adding a vector $\mathbf{w}(t) \in \mathrm{null}(\mathbf{C}^T)$ in the null space of $\mathbf{C}^T$. I.e., for any $\boldsymbol{\lambda}^*(t) \in \boldsymbol{\Lambda}^*(t)$ there exist $\mathbf{w}(t) \in \mathrm{null}(\mathbf{C}^T)$ such that

$$\boldsymbol{\lambda}^*(t) = \boldsymbol{\lambda}_0^*(t) + \mathbf{w}(t) \quad (37)$$

where $\boldsymbol{\lambda}_0^*(t) \in \mathrm{Im}(\mathbf{C}^T)$ is the unique vector in the image of $\mathbf{C}^T$ that satisfies (36). The vector $\boldsymbol{\lambda}_0^*(t) \in \mathrm{Im}(\mathbf{C}^T)$ can be written in terms of the Moore-Penrose pseudoinverse $\mathbf{C}^\dagger$ of $\mathbf{C}^T$ as

$$\boldsymbol{\lambda}_0^*(t) = -\mathbf{C}^\dagger \nabla f_0(\mathbf{s}^*(t), t). \quad (38)$$

Since the definition of $\mathbf{C}^\dagger$ includes the property $\mathbf{C}^T = \mathbf{C}^T\mathbf{C}^\dagger\mathbf{C}^T$ we can write $\mathbf{v}(t) = (\mathbf{I} - \mathbf{C}^\dagger\mathbf{C}^T)\mathbf{w}(t)$ for some vector $\mathbf{w}(t)$, but this is not relevant to subsequent derivations.

Observe that $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\lambda}_0^*(t)$ are both random because the observations $\mathbf{x}(t)$ are random. We therefore derive two asymptotic stochastic bounds on the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$. The first is a mean bound that holds across ensemble averages, and the second bound holds almost surely for individual realizations. Both of these bounds are parametric on the variation bound $\delta(T_s)$ as we state in the following theorem.

Theorem 1: Let $\boldsymbol{\lambda}(t)$ denote the vector with current dual iterates obtained at time t from (18) and $\boldsymbol{\lambda}_0^*(t) \in \mathrm{Im}(\mathbf{C}^T)$ 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 $\epsilon < 1/(M\Gamma^2)$. 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

$$\liminf_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|\right] \leq \frac{1 + \epsilon m\gamma^2}{\epsilon m\gamma^3}\delta(T_s). \tag{39}$$

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

$$\liminf_{t\to\infty} \|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\| \leq \frac{1 + \epsilon m\gamma^2}{\epsilon m\gamma^3}\delta(T_s) \quad a.s. \tag{40}$$

Proof: See Section III-A. ∎

The first result in (39) 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 (35) 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 $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$.

The bounds in (39) and (40) become large as the step size ϵ becomes small. This is not unreasonable because the optimization problem in (7) changes with each time step. Hence, while $\boldsymbol{\lambda}(t+1)$ becomes closer to $\boldsymbol{\lambda}_0^*(t)$, the optimal argument $\boldsymbol{\lambda}_0^*(t)$ drifts away to $\boldsymbol{\lambda}_0^*(t+1)$. As we reduce ϵ, D-MAP loses its ability to track these changes in $\boldsymbol{\lambda}_0^*(t)$. The optimal stepsize selection is $\epsilon = 1/(M\Gamma^2)$. This uncovers the dependence of (39) and (40) on the condition number m/M of the primal objective as is always the case in gradient descent algorithms. A final interesting observation is the dependence of (39) and (40) on the spectral radius γ which characterizes networks for which D-MAP performs poorly. The eigenvalue γ is small for networks that are sparsely connected and large for densely connected networks.

Coming back to the original goal we relate the suboptimality of primal iterates $\mathbf{s}(t)$ to the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$ whose asymptotic behavior is characterized in Theorem 1. This is a simple result that follows from the strong convexity of the primal objective stated in Assumption (A3) as we show in the following theorem.

Theorem 2: Let $\mathbf{s}(t)$ denote the current primal iterate at time t with components $\mathbf{s}_k(t)$ given as in (17) and let $\mathbf{s}^*(t)$ be the optimal argument of (7) with components $\mathbf{s}_k^*(t) = \mathbf{s}_{\mathrm{MAP}}(t)$. With the same definitions and assumptions of Theorem 1, the distance $\|\mathbf{s}(t) - \mathbf{s}^*(t)\|$ between $\mathbf{s}_k(t)$ and $\mathbf{s}^*(t)$ can be bounded as

$$\|\mathbf{s}(t) - \mathbf{s}^*(t)\| \leq \Gamma M\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|. \tag{41}$$

Proof: Consider the KKT condition $\nabla\mathcal{L}_{\mathbf{s}}(\mathbf{s}^*(t), \boldsymbol{\lambda}^*(t), t) = \mathbf{0}$ which we explicitly write as

$$\nabla f_0(\mathbf{s}^*(t), t) + \mathbf{C}^T\boldsymbol{\lambda}^*(t) = \mathbf{0}. \tag{42}$$

Furthermore, since $\mathbf{s}(t)$ is the primal variable that maximizes the Lagrangian $\mathcal{L}(\mathbf{s}, \boldsymbol{\lambda}(t), t)$ according to (13), it holds that

$$\nabla f_0(\mathbf{s}(t), t) + \mathbf{C}^T\boldsymbol{\lambda}(t) = \mathbf{0}. \tag{43}$$

Subtracting (43) from (42), rearranging terms and taking the norm of the resulting expressions yields

$$\begin{aligned}\|\nabla f_0(\mathbf{s}(t), t) - \nabla f_0(\mathbf{s}^*(t), t)\| &= \|\mathbf{C}^T(\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t))\| \\ &\leq \|\mathbf{C}^T\|\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|, \end{aligned}\tag{44}$$

where the inequality follows from Cauchy-Schwarz's inequality.

Consider now the strong convexity of the primal stated in Assumption (A3) and evaluate (34) for vectors $\mathbf{s} = \mathbf{s}(t)$ and $\mathbf{r} = \mathbf{s}^*(t)$ as well as for variables $\mathbf{s} = \mathbf{s}^*(t)$ and $\mathbf{r} = \mathbf{s}(t)$. Adding up the resulting expressions yields

$$0 \geq \left[\nabla f_0(\mathbf{r}, t) - \nabla f_0(\mathbf{s}, t)\right]^T(\mathbf{s} - \mathbf{r}) + \frac{1}{M}\|\mathbf{s} - \mathbf{r}\|^2, \tag{45}$$

after rearranging and canceling terms. Using Cauchy-Schwarz's inequality in (45) leads to

$$\|\mathbf{s}(t) - \mathbf{s}^*(t)\| \leq M\|\nabla f_0(\mathbf{s}(t), t) - \nabla f_0(\mathbf{s}^*(t), t)\|. \tag{46}$$

Substituting (44) into (46) the result in (41) follows after noting that $\|\mathbf{C}^T\| = \Gamma$ as per Assumption (A1). ∎

According to Theorem 2 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. (39)] and for almost all realizations of the dynamic system of interest D-MAP estimates are close to MAP estimates infinitely often [cf. (40)]. The bound depends on the condition number $\frac{m}{M}$ of the primal objective, the spectral radius γ of the edge incidence matrix, and the objective smoothness parameter $\delta(T_s)$. Moreover, the bound can be made arbitrarily small by reducing the sampling time T_s so that $\delta(T_s) \to 0$.

Proof of Theorem 1

The proof of Theorem 1 builds on a relationship between the distance $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_0^*(t)\|$ of the current dual iterate $\boldsymbol{\lambda}(t)$ to the optimal Lagrange multipliers $\boldsymbol{\lambda}_0^*(t)$ in the span of $\mathbf{C}^T$ and the expected value of the corresponding distance $\|\boldsymbol{\lambda}(t+1) - \boldsymbol{\lambda}_0^*(t+1)\|$ in the subsequent time slot. This relationship is similar to a

supermartingale contraction as stated in Theorem 3. We obtain (39) of Theorem 1 by applying this contraction recursively. We obtain (40) of Theorem 1 by modifying the sequence $\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\|$ to construct a proper supermartingale.

Before proceeding to the proof of Theorem 1 let us introduce a preliminary result to translate assumptions (A2) and (A3) into corresponding statements on the gradient Lipschitz continuity and strong convexity of the dual functions $g(\boldsymbol{\lambda},t)$.

Lemma 1: Consider the dual function $g(\boldsymbol{\lambda},t)$ as defined in (10). Assume that the primal objective $f_0(\mathbf{s},t)$ satisfies assumptions (A2) and (A3) and that the edge incidence matrix complies with Assumption (A1). The dual function $g(\boldsymbol{\lambda},t)$ has Lipschitz gradients with dual Lipschitz parameter $M\Gamma^2$,

$$g(\boldsymbol{\mu},t) \leq g(\boldsymbol{\lambda},t)+\nabla g(\boldsymbol{\lambda},t)^T(\boldsymbol{\lambda}-\boldsymbol{\mu})+\frac{M\Gamma^2}{2}\|\boldsymbol{\lambda}-\boldsymbol{\mu}\|^2. \quad (47)$$

The dual function $g(\boldsymbol{\lambda},t)$ is strongly convex in the span of $\mathbf{C}^T$ with dual strong convexity constant $m\gamma^2$. I.e., for any pair of vectors $\boldsymbol{\lambda},\boldsymbol{\mu}\in \mathrm{Im}(\mathbf{C}^T)$ it holds

$$g(\boldsymbol{\mu},t) \geq g(\boldsymbol{\lambda},t)+\nabla g(\boldsymbol{\lambda},t)^T(\boldsymbol{\lambda}-\boldsymbol{\mu})+\frac{m\gamma^2}{2}\|\boldsymbol{\lambda}-\boldsymbol{\mu}\|^2. \quad (48)$$

Proof: Observe that both (33) and (34) combine the mean value theorem with the corresponding eigenvalue bound for the primal function. The mean value theorem for the dual function states that for all dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ there exists a vector $\boldsymbol{\nu}$ in the segment $\boldsymbol{\lambda}-\boldsymbol{\mu}$ for which

$$g(\boldsymbol{\mu},t) = g(\boldsymbol{\lambda},t)+\nabla g(\boldsymbol{\lambda},t)^T(\boldsymbol{\lambda}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\mu})^T\nabla^2 g(\boldsymbol{\nu},t)(\boldsymbol{\lambda}-\boldsymbol{\mu}). \quad (49)$$

Recall that Hessians $\nabla^2 g(\boldsymbol{\lambda},t)$ of dual functions $g(\boldsymbol{\lambda},t)$ can be computed by producing a second order Taylor approximation of (7) and considering the quadratic dual of the resulting quadratic program. This procedure leads to

$$\nabla^2 g(\boldsymbol{\lambda},t) = \mathbf{C}(-\nabla^2 f_0(\mathbf{s}(\boldsymbol{\lambda}),t)^{-1})\mathbf{C}^T \quad (50)$$

where $\mathbf{s}(\boldsymbol{\lambda})$ is the primal Lagrangian maximizer as defined in (13)—see, e.g., [30, Eqs. (8)–(10)]. Using this fact we can bound the largest eigenvalue of $\nabla^2 g(\boldsymbol{\lambda},t)$ by $M\Gamma^2$ due to the strong convexity assumption in (34) and the edge incidence matrix largest eigenvalue bound. This latter observation substituted in (49) yields the Lipschitz gradient statement in (47). Observe that strong convexity of the primal translates into Lipschitz gradient continuity because the dual Hessian is a linear transformation of the inverse $(-\nabla^2 f_0(\mathbf{s}(\boldsymbol{\lambda}),t)^{-1})$ of the Hessian of the primal objective.

In the same way in which the primal strong convexity assumption in (34) is translated into the dual Lipschitz gradient property in (47), the primal Lipschitz gradient assumption in (33) can be translated into the dual strong convexity property in (48). For that matter consider the mean value theorem statement in (49), the dual Hessian expression in (50), and the primal Hessian eigenvalue bound in Assumption (A2) to conclude that for any pair of vectors $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ it holds

$$g(\boldsymbol{\mu},t) \geq g(\boldsymbol{\lambda},t)+\nabla g(\boldsymbol{\lambda},t)^T(\boldsymbol{\lambda}-\boldsymbol{\mu})+m(\boldsymbol{\lambda}-\boldsymbol{\mu})^T\mathbf{C}\mathbf{C}^T(\boldsymbol{\lambda}-\boldsymbol{\mu}). \quad (51)$$

If we restrict our attention to vectors $\boldsymbol{\lambda},\boldsymbol{\mu}\in\mathrm{Im}(\mathbf{C}^T)$ in the image of the transposed replicated edge incidence matrix $\mathbf{C}^T$ we can bound $(\boldsymbol{\lambda}-\boldsymbol{\mu})^T\mathbf{C}\mathbf{C}^T(\boldsymbol{\lambda}-\boldsymbol{\mu})\geq\gamma^2\|\boldsymbol{\lambda}-\boldsymbol{\mu}\|^2$. Substituting this bound in (51) yields the dual strong convexity statement in (48). ■

Lemma 1 produces Lipschitz gradient and strong convexity statements for the dual functions $g(\boldsymbol{\lambda},t)$ that are needed to study gradient descent on these functions. Observe that (48) is a statement of strong convexity of the dual function restricted to vectors in the image of $\mathbf{C}^T$. The dual function is not strongly convex in general. This weaker statement is sufficient for the convergence analysis that we perform in the remainder of this section—see Lemma 2.

The following theorem relates the distances $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ between the current dual iterate $\boldsymbol{\lambda}(t)$ and the unique current optimal multiplier $\boldsymbol{\lambda}_0^*(t)$ in the image of $\mathbf{C}^T$ at subsequent times.

Theorem 3: Let $\boldsymbol{\lambda}(t)$ denote a sequence of dual variables obtained through recursive application of (17)–(18) and $\boldsymbol{\lambda}_0^*(t)\in \mathrm{Im}\mathbf{C}^T$ denote the unique optimal argument of the dual function $g(\boldsymbol{\lambda},t)$ that lies in the image of the tranposed replicated edge incidence matrix $\mathbf{C}^T$. Assume the stepsize in (18) satisfies $\epsilon\leq\frac{1}{(M\Gamma^2)}$ and that assumptions (A1)–(A5) hold. Then, subsequent distances $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ satisfy

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\|\,|\mathbf{x}(0:t)\right] \leq \frac{1}{1+\epsilon m\gamma^2}\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|+\frac{1}{\gamma}\delta(T_s). \quad (52)$$

Proof: Consider a triangle with vertices $\boldsymbol{\lambda}(t+1)$, $\boldsymbol{\lambda}^*(t)$, and $\boldsymbol{\lambda}_0^*(t+1)$. The triangle inequality for $\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\|$ yields

$$\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\| \leq \|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|+\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\|. \quad (53)$$

The first term in the right hand side of (53) is the distance $\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|$ between the optimal multiplier $\boldsymbol{\lambda}_0^*(t)$ and the subsequent dual iterate $\boldsymbol{\lambda}(t+1)$. Since we are descending along the gradient $\nabla g(\boldsymbol{\lambda}(t),t)$ of the dual function at time t we expect this distance to be smaller than the distance $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ between $\boldsymbol{\lambda}_0^*(t)$ and the current dual iterate $\boldsymbol{\lambda}(t)$. This is proved true in Lemma 2. The second term in the right hand side of (53) is the distance $\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\|$ between subsequent optimal multipliers $\boldsymbol{\lambda}_0^*(t)$ and $\boldsymbol{\lambda}_0^*(t+1)$. We expect this distance to depend on the variation between log-likelihood functions $f_0(\mathbf{s},t)$ and $f_0(\mathbf{s},t+1)$ at subsequent time slots as bounded in assumption (A5). The relationship between these quantities is established in Lemma 3.

Lemma 2: Assume the same hypotheses and definitions of Theorem 3. The distances $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ and $\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|$ of subsequent iterates $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\lambda}(t+1)$ to the optimal dual variable $\boldsymbol{\lambda}_0^*(t)\in\mathrm{Im}(\mathbf{C}^T)$ satisfies

$$\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\| \leq \frac{1}{1+\epsilon m\gamma^2}\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|. \quad (54)$$

Proof: Since the optimal dual variable $\boldsymbol{\lambda}_0^*(t)$ is the same on both sides, (54) is just a statement on the optimality improvement of subsequent gradient descent iterates. The proof is therefore analog to the convergence analysis for gradient descent al-

gorithms—see e.g., [28, p. 466]—modified to use the milder strong convexity assumption in (48) of Lemma 1. The proof is also adapted to relate subsequent distances $\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|$ and $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ between iterates and the optimal set instead of the corresponding suboptimalities $g(\boldsymbol{\lambda}(t+1),t)-g(\boldsymbol{\lambda}_0^*(t),t)$ and $g(\boldsymbol{\lambda}(t),t)-g(\boldsymbol{\lambda}_0^*(t),t)$.

Begin the proof by writing the dual Lipschitz gradient statement in (47) of Lemma 1 for $\boldsymbol{\lambda}=\boldsymbol{\lambda}(t)$ and $\boldsymbol{\mu}=\boldsymbol{\lambda}(t+1)$

$$g(\boldsymbol{\lambda}(t+1),t) \le g(\boldsymbol{\lambda}(t),t)+\nabla g(\boldsymbol{\lambda}(t),t)^T(\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}(t)) + \frac{M\Gamma^2}{2}\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}(t)\|^2. \quad (55)$$

Further recall that as per the dual update in (18) we have $\boldsymbol{\lambda}(t+1)=\boldsymbol{\lambda}(t)-\epsilon\nabla g(\boldsymbol{\lambda}(t),t)$ because according to (15) the constraint slacks $\mathbf{s}_k^n(t)-\mathbf{s}_l^n(t)$ are components of the dual function gradient. Substituting this equality in (55) yields

$$g(\boldsymbol{\lambda}(t+1),t) \le g(\boldsymbol{\lambda}(t),t)-\epsilon\nabla g(\boldsymbol{\lambda}(t),t)^T\nabla g(\boldsymbol{\lambda}(t),t) + \frac{M\Gamma^2\epsilon^2}{2}\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2. \quad (56)$$

The second term in the right hand side of (56) simplifies to $\epsilon\nabla g(\boldsymbol{\lambda}(t),t)^T\nabla g(\boldsymbol{\lambda}(t),t) = \epsilon\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2$. Making this substitution in (56) and pulling common factors yields

$$g(\boldsymbol{\lambda}(t+1),t) \le g(\boldsymbol{\lambda}(t),t)+\epsilon\left(\frac{M\Gamma^2\epsilon}{2}-1\right)\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2. \quad (57)$$

The hypotheses include the stepsize bound $\epsilon \le 1/(M\Gamma^2)$ which is equivalent to $(M\Gamma^2\epsilon)/2-1 \le -1/2$. Using this fact and subtracting the optimal value $g(\boldsymbol{\lambda}_0^*(t),t)$ from both sides of (57) yields

$$g(\boldsymbol{\lambda}(t+1),t)-g(\boldsymbol{\lambda}_0^*(t),t) \le g(\boldsymbol{\lambda}(t),t)-g(\boldsymbol{\lambda}_0^*(t),t)-\frac{\epsilon}{2}\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2. \quad (58)$$

We are now interested in rewriting the right-hand side of (58). To do so, expand the squared distance between multipliers $\boldsymbol{\lambda}(t+1)$ at time $t+1$ and optimal multipliers $\boldsymbol{\lambda}_0^*(t)$ to write

$$\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|^2 = \|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|^2 - 2\epsilon\nabla g(\boldsymbol{\lambda}(t),t)^T(\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)) + \epsilon^2\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2. \quad (59)$$

Due to the fact that the dual function is (strictly) convex we can use the bound

$$g(\boldsymbol{\lambda}(t),t)-g(\boldsymbol{\lambda}_0^*(t),t) \le \nabla g(\boldsymbol{\lambda}(t),t)^T(\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)), \quad (60)$$

whose right hand side contains the term in the middle of the right hand side of (59). Using this observation to combine (59) and (60) gives, after rearranging terms,

$$g(\boldsymbol{\lambda}(t),t)-g(\boldsymbol{\lambda}_0^*(t),t) \le \frac{1}{2\epsilon}(-\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|^2 + \|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|^2) + \frac{\epsilon}{2}\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2. \quad (61)$$

After substituting the bound in (61) for the term $g(\boldsymbol{\lambda}(t),t)-g(\boldsymbol{\lambda}_0^*(t),t)$ of (58) the terms $\pm(\epsilon/2)\|\nabla g(\boldsymbol{\lambda}(t),t)\|^2$ of (58) and (61) cancel out each other leading to

$$g(\boldsymbol{\lambda}(t+1),t)-g(\boldsymbol{\lambda}_0^*(t),t) \le -\frac{1}{2\epsilon}\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|^2+\frac{1}{2\epsilon}\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|^2. \quad (62)$$

Having found an upper bound on $g(\boldsymbol{\lambda}(t+1),t)-g(\boldsymbol{\lambda}_0^*(t),t)$ we set to find a lower bound in the same quantity. Given Assumption (A4) and the fact that gradients $\nabla g(\boldsymbol{\lambda}(t),t)\in \mathrm{Im}(\mathbf{C}^T)$ for all times t [cf. (14)] it follows that $\boldsymbol{\lambda}(t)\in\mathrm{Im}(\mathbf{C}^T)$ for all t. Since it is also true that $\boldsymbol{\lambda}_0^*(t)\in\mathrm{Im}(\mathbf{C}^T)$, we can write the restricted dual strong convexity statement in (48) of Lemma 1 for $\boldsymbol{\lambda}=\boldsymbol{\lambda}_0^*(t)$ and $\boldsymbol{\mu}=\boldsymbol{\lambda}(t+1)$ to obtain the bound

$$g(\boldsymbol{\lambda}(t+1),t) \ge g(\boldsymbol{\lambda}_0^*(t),t) + \nabla g(\boldsymbol{\lambda}_0^*(t),t)^T(\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)) + \frac{m\gamma^2}{2}\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|^2. \quad (63)$$

Because the dual function $g(\boldsymbol{\lambda},t)$ is convex it holds that $\nabla g(\boldsymbol{\lambda}_0^*(t),t)^T(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0^*(t))\ge 0$ for arbitrary dual variable $\boldsymbol{\lambda}$. Hence, we can drop the term in the middle of the right-hand side of (63) and rearrange terms to get

$$\frac{m\gamma^2}{2}\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\|^2 \le g(\boldsymbol{\lambda}(t+1),t)-g(\boldsymbol{\lambda}_0^*(t),t). \quad (64)$$

Combining the bounds in (62) and (64) to eliminate the terms $g(\boldsymbol{\lambda}(t+1),t)-g(\boldsymbol{\lambda}_0^*(t),t)$ yields the result in (54) after reordering terms. ■

Lemma 3: Assume the same hypotheses and definitions of Theorem 3 and consider the set of all past observations $\mathbf{x}(0:t)$ given. On average, the optimal multiplier $\boldsymbol{\lambda}_0^*(t+1)$ at the next time step deviates from the current optimal dual variable $\boldsymbol{\lambda}_0^*(t)$ by no more than $\delta(T_s)/\gamma$,

$$\mathbb{E}\left[\left\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\right\| \,\middle|\, \mathbf{x}(0:t)\right] \le \frac{1}{\gamma}\delta(T_s). \quad (65)$$

Proof: With $\mathbf{x}(t)$ given the log-likelihood function $f_0(\mathbf{s},t)$ is also given. As a consequence so is the dual function $g(\boldsymbol{\lambda},t)$ and the optimal dual variable $\boldsymbol{\lambda}_0^*(t)$. The optimal variable $\boldsymbol{\lambda}_0^*(t+1)$ depends on the random observation $\mathbf{x}(t+1)$. The expectation in (65) is with respect to the distribution of $\mathbf{x}(t+1)$ given $\mathbf{x}(t)$. Recall that $\boldsymbol{\lambda}_0^*(t)$ and $\boldsymbol{\lambda}_0^*(t+1)$ are the unique optimal dual arguments that lie in the image of the transpose of the replicated edge incidence matrix $\mathbf{C}^T$. Thus, using their explicit expressions in (38) we can write

$$\left\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\right\| = \left\|-\mathbf{C}^\dagger\nabla f_0(\mathbf{s}^*(t+1),t+1)+\mathbf{C}^\dagger\nabla f_0(\mathbf{s}^*(t),t)\right\|. \quad (66)$$

Applying Cauchy-Schwarz's inequality we can extract the norm of the pseudoinverse matrix $\mathbf{C}^\dagger$ to obtain

$$\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\| \le \|\mathbf{C}^\dagger\|\,\|\nabla f_0(\mathbf{s}^*(t+1),t+1)-\nabla f_0(\mathbf{s}^*(t),t)\|, \quad (67)$$

The norm of $\|\mathbf{C}^{\dagger}\|$ is its largest singular value. According to the properties of Moore-Penrose pseudoinverses, it is also the inverse of the smallest nonzero singular value of $\mathbf{C}$. Combining this observation with the definition of γ in assumption (A1) (67) simplifies to

$$\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\| \leq \frac{1}{\gamma}\|\nabla f_0(\mathbf{s}^*(t+1),t+1)-\nabla f_0(\mathbf{s}^*(t),t)\|. \quad (68)$$

Taking expectations on both sides of (68) conditional on the past observations $\mathbf{x}(0:t)$ yields

$$\mathbb{E}\left[\left\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\right\| \,\middle|\, \mathbf{x}(0:t)\right] \leq \frac{1}{\gamma}\mathbb{E}\left[\|\nabla f_0(\mathbf{s}^*(t+1),t+1)+\nabla f_0(\mathbf{s}^*(t),t)\| \,\middle|\, \mathbf{x}(0:t)\right]. \quad (69)$$

The result in (65) follows from substituting the bound (35) of assumption (A5) for the right hand side of (69). ∎

To complete the proof of (52) consider the expectation of the inequality in (53) with $\mathbf{x}(0:t)$ given. Since $\boldsymbol{\lambda}(t+1)$ is uniquely determined by $\mathbf{x}(0:t)$ taking this expectation yields

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\| \,\middle|\, \mathbf{x}(0:t)\right] \leq \|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t)\| + \mathbb{E}\left[\|\boldsymbol{\lambda}_0^*(t+1)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t)\right]. \quad (70)$$

Substituting the result (54) of Lemma 2 and the result (65) of Lemma 3 into (70) yields the bound in (52). ∎

Returning to the proof of Theorem 1 it leaves to be shown that the result from Theorem 3 can be used to characterize the convergence of the sequence $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$. For the mean convergence result in (39) consider the expectation of the statement in (52) of Theorem 3 with respect to the past observations $\mathbf{x}(0:t_0)$ up to a certain time $t_0 < t$

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\| \,\middle|\, \mathbf{x}(0:t_0)\right] \leq \frac{1}{1+\epsilon m\gamma^2}\mathbb{E}\left[\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right] + \frac{1}{\gamma}\delta(T_s). \quad (71)$$

We use (71) to show that the expectation $\mathbb{E}\left[\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right]$ eventually approaches the near optimality region in which $\mathbb{E}\left[\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right] \leq \delta(T_s)(1+\epsilon m\gamma^2)/(\epsilon m\gamma^3)$. To prove that this is true suppose that it is false and that it therefore holds

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right] \geq \frac{\delta(T_s)(1+\epsilon m\gamma^2)}{\epsilon m\gamma^3(1-\beta)}, \quad (72)$$

for some constant $0<\beta<1$ and all times $t \geq t_0$. We use this hypothetical conclusion to remove the term $\delta(T_s)/\gamma$ from (71) and group and reorder terms to obtain

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\| \,\middle|\, \mathbf{x}(0:t_0)\right] \leq \left(1-\beta\frac{\epsilon m\gamma^2}{1+\epsilon m\gamma^2}\right)\mathbb{E}\left[\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right]. \quad (73)$$

Since the bound in (73) is assumed to hold for all times we can use it recursively between times t_0 and t to write

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\| \,\middle|\, \mathbf{x}(0:t_0)\right] \leq \left(1-\beta\frac{\epsilon m\gamma^2}{1+\epsilon m\gamma^2}\right)^{t-t_0}\mathbb{E}\left[\|\boldsymbol{\lambda}(t_0)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right]. \quad (74)$$

But this implies that the expectation $\mathbb{E}\left[\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| \,\middle|\, \mathbf{x}(0:t_0)\right] \to 0$ as time $t \to \infty$ thereby contradicting (72). It follows that (72) is absurd and that as consequence we must have

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(t')-\boldsymbol{\lambda}_0^*(t')\| \,\middle|\, \mathbf{x}(0:t_0)\right] \leq \frac{(1+\epsilon m\gamma^2)}{\epsilon m\gamma^3(1-\beta)}\delta(T_s) \quad (75)$$

for at least some time $t' \geq t_0$ and *all* constants $0<\beta<1$. The limit infimum statement in Theorem 1 follows because (75) is true for arbitrary time t_0 and arbitrary constant $0<\beta<1$.

For the almost sure convergence result in (40) we construct a supermartingale based on the values of the distances $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$. To do so consider a given time t_0 and fix an arbitrary constant $0<\beta<1$ to define the stopping time τ_s as the first time $t \geq t_0$ at which the process $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ enters the near optimal region $\delta(T_s)(1+\epsilon m\gamma^2)/\epsilon m\gamma^3(1-\beta)$

$$\tau_s := \inf\left\{t > t_0 : \left\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\right\| \leq \frac{\delta(T_s)(1+\epsilon m\gamma^2)}{\epsilon m\gamma^3(1-\beta)}\right\}. \quad (76)$$

Based on the stopping time τ_s in (76) define the process $\alpha(t)$ with realizations

$$\alpha(t) = \|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|\mathbb{I}\{t<\tau_s\}, \quad (77)$$

for all $t \geq t_0$. The sequence $\alpha(t)$ follows the sequence of distances to optimality $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\|$ until this distance becomes smaller than $\delta(T_s)(1+\epsilon m\gamma^2)/\epsilon m\gamma^3(1-\beta)$ at time $t=\tau_s$. Thereafter, $\alpha(t)=0$ for all subsequent times $t \geq \tau_s$.

The process $\alpha(t)$ is a supermartingale. Indeed, consider the expected value of $\alpha(t+1)$ conditional on $\mathbf{x}(0:t)$—recall that all randomness in the system is measured if we are given $\mathbf{x}(0:t)$. There are two different cases $\alpha(t)=0$ and $\alpha(t)\neq 0$. When $\alpha(t)=0$, it must be that $\tau_s \leq t$ which implies $\tau_s \leq t+1$ yielding $\alpha(t+1)=0$ according to (77). A particular conclusion of this observation is that

$$\mathbb{E}\left[\alpha(t+1) \,\middle|\, \mathbf{x}(0:t), \alpha(t)=0\right] = \alpha(t) = 0. \quad (78)$$

When $\alpha(t)\neq 0$ use the definition of $\alpha(t)$ in (77) and the fact that $\mathbb{I}\{t<\tau_s\} \leq 1$ to write

$$\mathbb{E}\left[\alpha(t+1) \,\middle|\, \mathbf{x}(0:t), \alpha(t)\neq 0\right] \leq \mathbb{E}\left[\|\boldsymbol{\lambda}(t+1)-\boldsymbol{\lambda}_0^*(t+1)\| \,\middle|\, \mathbf{x}(0:t)\right]. \quad (79)$$

The right hand side on (79) is the expected distance bounded in Theorem 3. Thus, we can combine (52) of Theorem 3 with (79) to write

$$\mathbb{E}\left[\alpha(t+1) \,\middle|\, \mathbf{x}(0:t), \alpha(t)\neq 0\right] \leq \frac{1}{1+\epsilon m\gamma^2}\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| + \frac{1}{\gamma}\delta(T_s). \quad (80)$$

If $\alpha(t)\neq 0$ it must be that $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}_0^*(t)\| > \delta(T_s)(1+\epsilon m\gamma^2)/\epsilon m\gamma^3(1-\beta)$. Otherwise the stopping time would satisfy $\tau_s \leq t$ [cf. (76)] and $\alpha(t)=0$ as

a consequence [cf. (77)]. Using this inequality to eliminate $\delta(T_s)/\gamma$ from (80) yields

$$\mathbb{E}\left[\alpha(t+1)\,\middle|\,\mathbf{x}(0:t),\alpha(t)\neq 0\right] \leq \left(1-\beta\frac{\epsilon m\gamma^2}{1+\epsilon m\gamma^2}\right)\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}^*(t)\|. \quad (81)$$

When $\alpha(t)\neq 0$ we must have $\mathbb{I}\{t<\tau_s\}=1$ which means $\alpha(t)=\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}^*(t)\|$. Making this substitution in (79) yields

$$\mathbb{E}\left[\alpha(t+1)\,\middle|\,\mathbf{x}(0:t),\alpha(t)\neq 0\right] \leq \left(1-\beta\frac{\epsilon m\gamma^2}{1+\epsilon m\gamma^2}\right)\alpha(t). \quad (82)$$

Combining the claims in (78) and (82) it follows that $\alpha(t)$ is a non-negative supermartingale as we had claimed. It follows from the supermartingale convergence theorem [31, p. 352] that $\lim_{t\to\infty}\alpha(t)$ exists for almost all realizations of the process.

We will further show that all of these limits must be $\lim_{t\to\infty}\alpha(t)=0$. For doing so just notice that (78) and (82) can be combined to write

$$\mathbb{E}\left[\alpha(t+1)\,\middle|\,\mathbf{x}(0:t)\right] \leq \left(1-\beta\frac{\epsilon m\gamma^2}{1+\epsilon m\gamma^2}\right)\alpha(t). \quad (83)$$

Taking the expectation of both sides of (83) with respect to the distribution of $\mathbf{x}(t)$ given $\mathbf{x}(0:t_0)$ and applying the resulting inequality recursively between times t and t_0 yields

$$\mathbb{E}\left[\alpha(t+1)\,\middle|\,\mathbf{x}(0:t_0)\right] \leq \left(1-\beta\frac{\epsilon m\gamma^2}{1+\epsilon m\gamma^2}\right)^{t-t_0}\alpha(t_0). \quad (84)$$

Since the constant β satisfies $0<\beta<1$ and the process $\alpha(t)$ is nonnegative it follows from (84) that

$$\lim_{t\to\infty}\mathbb{E}\left[\alpha(t)\,\middle|\,\mathbf{x}(0:t_0)\right]=0. \quad (85)$$

Since we already observed that $\lim_{t\to\infty}\alpha(t)$ exists almost surely in order for (85) to be true all of these limits must be null because the sequence $\alpha(t)$ is nonnegative. Therefore

$$\lim_{t\to\infty}\alpha(t)=0 \quad \text{a.s..} \quad (86)$$

According to the definition of $\alpha(t)$ in (77) the result in (86) implies that the stopping time τ_s is almost surely finite, which means that for almost all processes there exists a time $t'\geq t_0$ such that

$$\left\|\boldsymbol{\lambda}(t')-\boldsymbol{\lambda}_0^*(t')\right\| \leq \frac{\delta(T_s)(1+\epsilon m\gamma^2)}{\epsilon m\gamma^3(1-\beta)}. \quad (87)$$

The limit infimum statement in Theorem 1 follows because (87) is almost surely true for arbitrary time t_0 and arbitrary constant $0<\beta<1$.

IV. Simulation Results

We implement the D-MAP algorithm in (17)–(18) for the linear Gaussian AR model introduced in Section II-B and the quantized observations model of Section II-C. In both cases we compare performance of D-MAP estimates $\mathbf{s}_k(t)$ to the centralized MAP estimator $\mathbf{s}_{\rm MAP}(t)$ in (6) which would be computed if all observations were available at a common location. We also compare D-MAP and local (L-) MAP estimates $\hat{\mathbf{s}}_k(t)$ computed using local observations only,

$$\hat{\mathbf{s}}_k(t)=\arg\max_{\mathbf{s}}\sum_n \ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}^n\right)+\ln \mathrm{P}\left(\mathbf{s}^n|\mathbf{s}^{n-1}\right). \quad (88)$$

As mentioned in Remark 1, the term $[\mathbf{s}_k^{t-W+1}-A\mathbf{s}_k^{t-W}(t-1)]^T\mathbf{Q}^{-1}[\mathbf{s}_k^{t-W+1}-\mathbf{A}\mathbf{s}_k^{t-W}(t-1)]$ is added to the distributed log-likelihood in (6). To maintain fair comparison benchmarks the term $\left[\mathbf{s}^{t-W+1}-\mathbf{A}\mathbf{s}_{\rm MAP}^{t-W}(t-1)\right]^T$ is added to the centralized log-likelihood in (6) and the term $[\mathbf{s}^{t-W+1}-\mathbf{A}\hat{\mathbf{s}}_k^{t-W}(t-1)]^T\mathbf{Q}^{-1}[\mathbf{s}^{t-W+1}-\mathbf{A}\hat{\mathbf{s}}_k^{t-W}(t-1)]$ is added to the local likelihood in (88).

In both subsequent numerical studies we consider a WSN with $K=8$ sensors and edges between any two sensors k and l present with probability $\frac{1}{2}$.

A. Linear Gaussian AR Model

Consider a two-dimensional signal vector $\mathbf{s}(\tau)=[s_1(\tau),s_2(\tau)]^T$ containing temperature values at two points in space. The dynamical model has state transition matrix $\mathbf{A}_a=[0.99,-0.10;0.10;0.99]/s$ and driving input covariance matrix $\mathbf{Q}_a=\mathrm{diag}\left(0.25^\circ\mathrm{C}^2/\mathrm{s}^2,\ 0.25^\circ\mathrm{C}^2/\mathrm{s}^2\right)$. The sensors observe the temperatures directly implying that the observation matrices are identities $\mathbf{H}_{ak}=\mathbf{I}$ for all sensors k. The noise covariance matrices are equal for all sensors k as well and given by $\mathbf{R}_{ak}=\mathrm{diag}(0.5^\circ\mathrm{C}^2,\ 0.5^\circ\mathrm{C}^2)$. The sampling time is $T_s=0.166$ s and the system is simulated for 96 observation slots corresponding to a total elapsed time of 16 s. The estimation window is set to $W=3$. The signal is initialized to $\mathbf{s}(0)=[2^\circ\mathrm{C},\ 2^\circ\mathrm{C}]^T$. The Lagrange multipliers promoting equality of local estimates for D-MAP are initialized to $\boldsymbol{\lambda}_{kl}(0)=0$ for all links (k,l). With $\boldsymbol{\lambda}_{kl}(0)=\boldsymbol{\lambda}_{lk}(0)$ for all pairs of links (k,l) the initial D-MAP estimates $\mathbf{s}_k(t)$ and initial L-MAP estimates $\hat{\mathbf{s}}_k(t)$ coincide [cf. (17) and (88)]. Multiplier updates follow (18) and subsequent D-MAP estimates are computed according to (27). The stepsize for D-MAP for each edge (k,l) and signal j varies across sensors and is chosen as 0.1 times the inverse of its respective diagonal entry in the dual Hessian. These values can be computed locally at each sensor.

Simulation results are shown in Figs. 1 and 2. Fig. 1 shows the signal trajectory for a sample run along with MAP estimates, D-MAP estimates (left), and L-MAP estimates (right). D-MAP estimates are closer to the centralized MAP which provides the best possible MSE performance with the given observations. D-MAP also exhibits a better transient behavior. Both of these observations are clearer in Fig. 2 which compares the empirical MSE of D-MAP (left) and L-MAP (right) with the MSE of centralized MAP for times $\tau\in[0,16\ \mathrm{s}]$ averaged over 10^3 simulation runs. For the given parameters the steady state MSE of the centralized MAP is $0.32^\circ\mathrm{C}^2$. The steady state MSE of L-MAP averaged over all sensors is $0.34^\circ\mathrm{C}^2$ whereas the average steady state MSE of D-MAP is reduced to $0.325^\circ\mathrm{C}^2$. Note also that it

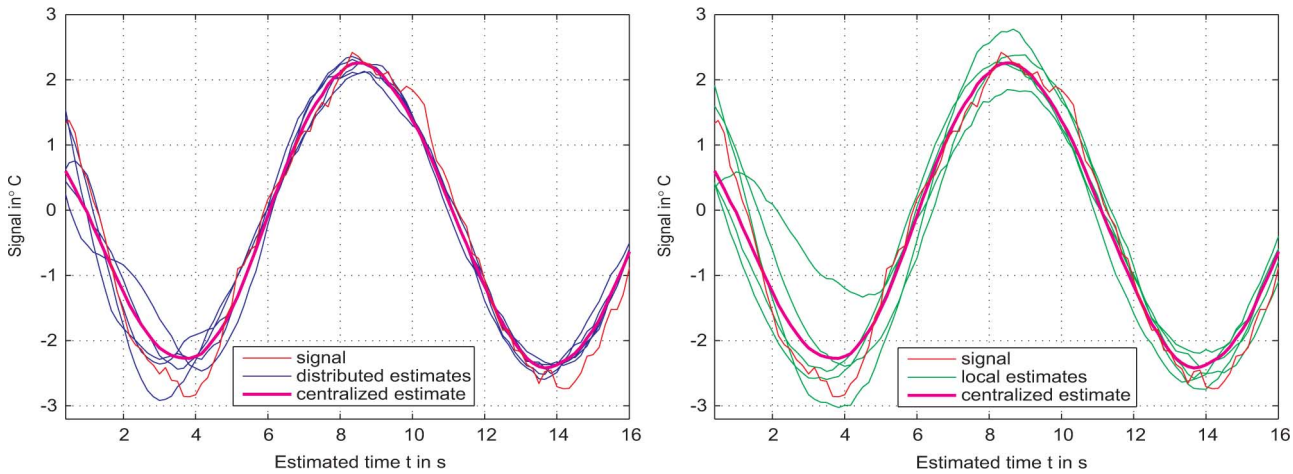

Fig. 1. Example run of D-MAP, L-MAP, and centralized MAP for a linear Gaussian autoregressive model. Signal values are shown along with centralized MAP estimates, D-MAP estimates (left), and L-MAP estimates (right) for times $\tau \in [0, 16\ \mathrm{s}]$. Steady state D-MAP estimates are closer than L-MAP estimates to the centralized MAP. D-MAP also exhibits better transient behavior than L-MAP.

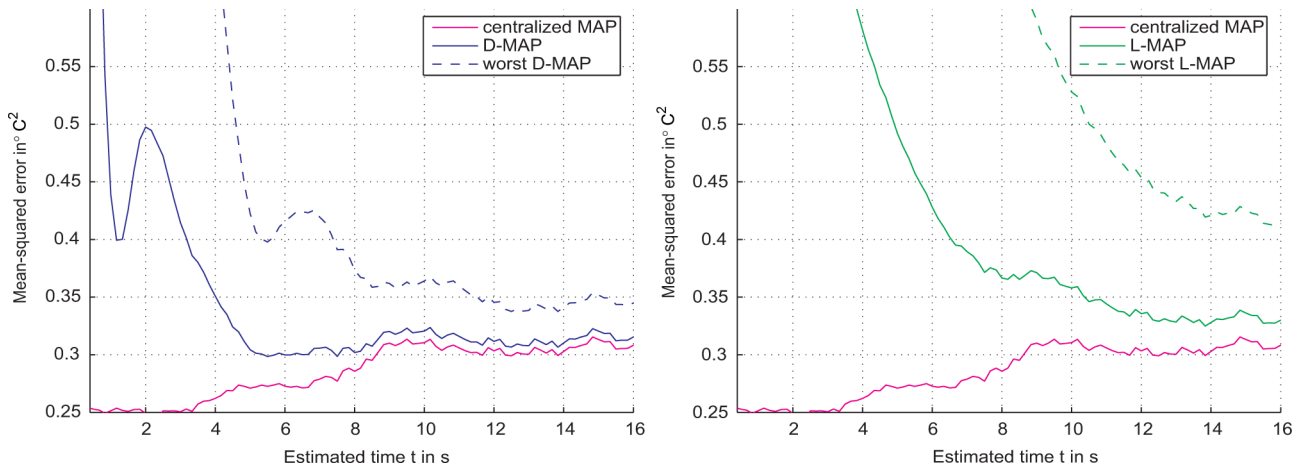

Fig. 2. Average and worst empirical mean squared error (MSE) attained by centralized MAP, D-MAP, and L-MAP for a linear Gaussian autoregressive model. Empirical MSEs are obtained as an average over 10^3 simulation runs and shown for times $\tau \in [0, 16]$. The worst empirical MSE averages the largest squared error across all sensors for each simulation realization. Empirical MSE of D-MAP is smaller than empirical MSE of L-MAP. The performance gain is more pronounced if we compare worst empirical MSEs.

takes L-MAP about 9 s to reach its steady state MSE whereas the D-MAP steady state MSE is reached after 6 s. The differences are more pronounced if we look at the worst squared error across all sensors as a function of time. The empirical average of this maximal squared error yields a measure of the worst MSE across all sensors that we also depict in Fig. 2. The worst MSE for D-MAP attains a steady-state value of $0.35°\mathrm{C}^2$ whereas for the L-MAP the worst empirical MSE approaches $0.42°\mathrm{C}^2$.

B. Quantized Model

Consider now the case in which sensors collect quantized binary observations as dictated by the model in Section II-C. 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 $\mathbf{Q}_a = q_a = 0.5°\mathrm{C}^2/\mathrm{s}^2$, observation noise covariances $\mathbf{R}_{ak} = r_{ak} = 1°\mathrm{C}^2$ for every sensor k, and observation matrices $\mathbf{H}_{ak} = h_{ak} = 1$ for all sensors k. We set the sampling time to $T_s = 0.166$ s and the initial temperature to $s(0) = 20°\mathrm{C}$. Quantization thresholds are set to $\theta_{0,k1} = \theta_{0,k} = 20°\mathrm{C}$ 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). Multiplier updates follow (18) as in the simulations in the linear model of Section IV-A. D-MAP estimates are computed according to (31). The stepsize for D-MAP for each edge (k, l) and signal j is chosen differently for each sensor k and set to 0.1 times the inverse of its respective diagonal entry in the dual Hessian of the corresponding Gaussian linear model.

Figs. 3 and 4 show simulation results for the described setup. Fig. 3 shows an example of a signal trajectory for a sample run comparing D-MAP estimates (left) and L-MAP estimates (right). Upon reaching steady state, D-MAP estimates are closer to the centralized MAP than L-MAP estimates. Although an overshooting effect can be noted for the D-MAP until time $\tau = 6$ s, it still exhibits a better transient behavior than the L-MAP. Fig. 4 quantifies these observations by looking at the corresponding MSEs over 10^3 simulation runs. The empirical MSE

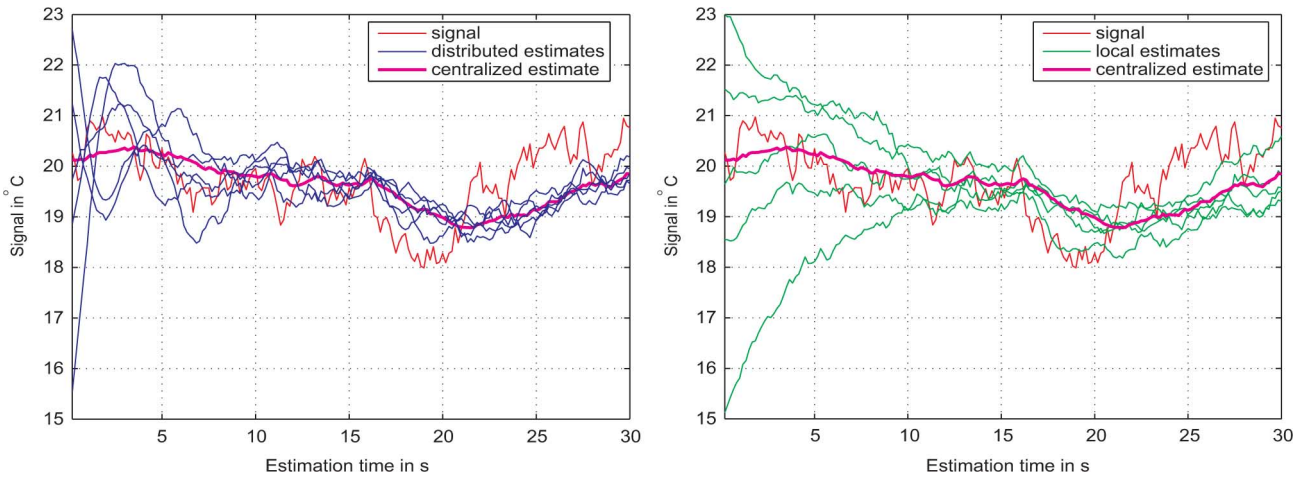

Fig. 3. Example run of D-MAP, L-MAP, and centralized MAP for a binary quantized model. Signal values are shown along with centralized MAP estimates, D-MAP estimates (left), and L-MAP estimates (right) for times $\tau \in [0, 30\ \mathrm{s}]$. The D-MAP displays steady-state behavior at time $\tau \approx 9$ s whereas the L-MAP only reaches steady state at $\tau \approx 12$ s. Steady state D-MAP estimates are closer than L-MAP estimates to the centralized MAP.

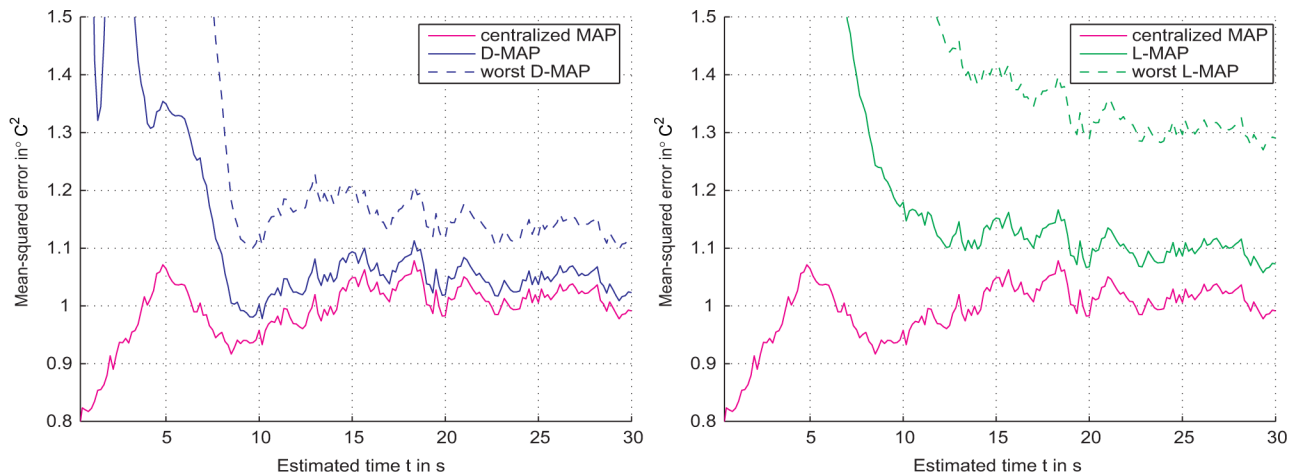

Fig. 4. Empirical mean squared error (MSE) for centralized MAP, D-MAP, and L-MAP and worst empirical MSE for times $\tau \in [0, 30\ \mathrm{s}]$, averaged over 10^3 simulation runs for a binary quantized model. The empirical MSE of D-MAP estimates is closer than that of L-MAP estimates to the empirical MSE of centralized MAP estimates.

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 $1.00°\mathrm{C}^2$. The steady state MSE of D-MAP is $1.05°\mathrm{C}^2$ on average whereas it is $1.1°\mathrm{C}^2$ for the L-MAP. The 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 $1.15°\mathrm{C}^2$ after time $\tau = 10$ s, the L-MAP takes $\tau = 20$ s to approach a worst steady-state MSE of $1.3°\mathrm{C}^2$.

V. Conclusion

This paper developed the distributed (D-) maximum a posteriori probability (MAP) estimation algorithm for the estimation of time-varying signals with a sensor network collecting noisy observations of a distributed nature. The algorithm incorporates information from neighboring sensors by communicating Lagrange multipliers which penalize the disagreement between neighbors. Lagrange multipliers are updated based on the differences between neighboring estimates as dictated by a dual gradient descent algorithm. We assess the tracking ability of D-MAP by studying the difference between distributed estimates and centralized estimates that would be computed if all the observations were available at a central location. This difference can be related to the suboptimality of the dual variables which is the main characterization presented in this paper. In particular, we proved 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. The size of the optimality neighborhood was characterized in terms of the condition number of the log-likelihood function, the Laplacian eigenvalue describing the connectedness of the sensor network, and a parameter describing the smoothness of the log-likelihood as a function of time. This latter parameter is a bound in the difference between the log-likelihood gradients at subsequent points in time. For linear models this parameter vanishes with decreasing sampling time at a rate proportional to the square

root of the sampling time. This smoothness condition is stronger than setting such a condition on the log-likelihoods themselves. Nevertheless, most log-likelihood functions are smooth in the sense that the difference between subsequent log-likelihood gradients vanishes with decreasing sampling time. It follows from this observation that the difference between D-MAP and centralized MAP estimates can be made arbitrarily small by reducing the sampling time of the process of interest.

Numerical results for a linear Gaussian auto regressive model and a nonlinear model with binary quantized observations corroborate the performance gains of D-MAP. Mean squared errors of D-MAP are lower than that of local MAP estimates in steady state operation and also exhibit better transient behavior. The advantage is most noticeable when comparing the worst mean squared error across different sensors in a given realization.

Appendix A
Assumption 5 for Linear Gaussian Autoregressive Model

Assumption 5 refers to a smoothness characteristic of the primal problem in (7). Specifically, it implies that the difference between subsequent gradients of the objective of the optimization problem in (7) evaluated at the optimal primal variables vanish with sampling time T_s. We show in this appendix that the assumption holds for the linear Gaussian AR model described in Section II-B. Consider the generic expression for the D-MAP estimate in (8). To specify the generic objective $f_0(\mathbf{s},t) = \sum_k f_k(\mathbf{s}_k,t) = \sum_{n,k} \ln \mathrm{P}\left(\mathbf{x}_k^n|\mathbf{s}_k^n\right) + (1/K)\ \ln \mathrm{P}\left(\mathbf{s}_k^n|\mathbf{s}_k^{n-1}\right)$ to the linear Gaussian AR model refer to the expression for the primal iteration in (27) and compare it with the generic primal iteration in (17) to conclude that

$$\begin{aligned} f_0(\mathbf{s},t) &= \sum_k f_k(\mathbf{s}_k,t) \\ &= -\sum_{n,k}\left(\mathbf{x}_k^n - \mathbf{H}_k\mathbf{s}_k^n\right)^T \mathbf{R}_k^{-1}\left(\mathbf{x}_k^n - \mathbf{H}_k\mathbf{s}_k^n\right) \\ &\quad + \left(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1}\right)^T \mathbf{Q}^{-1}\left(\mathbf{s}_k^n - \mathbf{A}\mathbf{s}_k^{n-1}\right). \end{aligned} \tag{89}$$

To simplify notation, we define the $WK \times WK$ matrices $\mathbf{H}_k'$, $\mathbf{Q}'$ and $\mathbf{R}_k'$ which are block-diagonal matrices stacking the matrices $\mathbf{H}_k$, $\mathbf{Q}$ and $\mathbf{R}_k$ for all times $n \in [t-W+1,t]$, i.e., $\mathbf{H}_k' = \mathrm{diag}(\mathbf{H}_k^{t-W+1}, \mathbf{H}_k^{t-W+2}, \ldots, \mathbf{H}_k^t)$, $\mathbf{Q}' = \mathrm{diag}(\mathbf{Q}^{t-W+1}, \mathbf{Q}^{t-W+2}, \ldots, \mathbf{Q}^t)$ and $\mathbf{R}_k = \mathrm{diag}(\mathbf{R}_k^{t-W+1}, \mathbf{R}_k^{t-W+2}, \ldots, \mathbf{R}_k^t)$. Furthermore, we define the $WK \times WK$ matrix $\mathbf{B}$ with blocks $\mathbf{A}$ corresponding to the transition matrix $\mathbf{A}$ in (22)

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A} & \mathbf{0} \end{bmatrix}. \tag{90}$$

Then the sum over time $n \in [t-W+1,t]$ in (89) can be rewritten in matrix form as

$$\begin{aligned} f_0(\mathbf{s},t) = \sum_k (\mathbf{x}_k(t) - \mathbf{H}_k'\,\mathbf{s}_k)^T \mathbf{R}_k'^{-1}(\mathbf{x}_k(t) - \mathbf{H}_k'\,\mathbf{s}_k) \\ + \frac{1}{K}(\mathbf{s}_k - \mathbf{B}\,\mathbf{s}_k)\mathbf{Q}'^{-1}(\mathbf{s}_k - \mathbf{B}\,\mathbf{s}_k). \end{aligned} \tag{91}$$

to express the $\mathbf{s}^n - \mathbf{s}^{n-1}$ terms in the last sum in (26) in matrix operations with the vector $(\mathbf{s} - \mathbf{B}\mathbf{s})$.

Due to the equivalence between the centralized MAP problem in (6) and the reformulation (7), we can compute a closed-form solution for $\mathbf{s}_k^*(t) = \mathbf{s}_{\mathrm{MAP}}(t)$ for all k using the centralized MAP formulation. Using the same notation, we can equivalently write the objective for the centralized MAP estimator in (26) as

$$\begin{aligned} f_{\mathrm{MAP}}(\mathbf{s},t) = \sum_k (\mathbf{x}(t) - \mathbf{H}_k'\,\mathbf{s})^T \mathbf{R}_k'^{-1}(\mathbf{x}_k(t) - \mathbf{H}_k'\,\mathbf{s}) \\ + \frac{1}{K}(\mathbf{s} - \mathbf{B}\,\mathbf{s})\mathbf{Q}'^{-1}(\mathbf{s} - \mathbf{B}\,\mathbf{s}). \end{aligned} \tag{92}$$

It is possible to find a closed-form solution for the centralized MAP estimate $\mathbf{s}_{\mathrm{MAP}}(t)$ corresponding to the centralized log-likelihood in (92). To do so compute the gradient of $f_{\mathrm{MAP}}(\mathbf{s},t)$ in (92) which can be written as

$$\begin{aligned} \nabla f_{\mathrm{MAP}}(\mathbf{s}_{\mathrm{MAP}}(t)) &= \sum_k {\mathbf{H}_k'}^T {\mathbf{R}_k'}^{-1}(\mathbf{x}_k(t) - \mathbf{H}_k'\,\mathbf{s}_{\mathrm{MAP}}(t)) \\ &\quad + \frac{1}{K}(\mathbf{I}-\mathbf{B})^T\mathbf{Q}^{-1}(\mathbf{I}-\mathbf{B})\mathbf{s}_{\mathrm{MAP}}(t) \\ &= \sum_k {\mathbf{H}_k'}^T {\mathbf{R}_k'}^{-1}\mathbf{x}_k(t) \\ &\quad - \sum_k ({\mathbf{H}_k'}^T {\mathbf{R}_k'}^{-1}\mathbf{H}_k' + \frac{1}{K}\tilde{\mathbf{Q}}^{-1})\mathbf{s}_{\mathrm{MAP}}(t), \end{aligned} \tag{93}$$

where the second equality is obtained by rearranging terms and defining $\tilde{\mathbf{Q}}^{-1} = (\mathbf{I}-\mathbf{B})^T\mathbf{Q}^{-1}(\mathbf{I}-\mathbf{B})$ to simplify notation. Setting the gradient in (93) to 0 yields the MAP estimate

$$\begin{aligned} \mathbf{s}_{\mathrm{MAP}}(t) = \left(\sum_k ({\mathbf{H}_k'}^T {\mathbf{R}_k'}^{-1}\mathbf{H}_k' + \frac{1}{K}\tilde{\mathbf{Q}}^{-1})\right)^{-1} \\ \times \left(\sum_k {\mathbf{H}_k'}^T {\mathbf{R}_k'}^{-1}\mathbf{x}_k(t)\right). \end{aligned} \tag{94}$$

Consider now the gradients of the distributed log-likelihood in (91) for times t and $t+1$. Use the notation $[\mathbf{v}]_k$ to represent a vector formed by the elements of $\mathbf{v}$ associated with sensor k. To determine the bound $\delta(T_s)$ in Assumption 5 we use the triangle inequality to separate the norm difference $\left\|\nabla f_0\left(\mathbf{s}^*(t),t\right) - \nabla f_0\left(\mathbf{s}^*(t+1),t+1\right)\right\|$ into its K per sensor components

$$\begin{aligned} &\left\|\nabla f_0\left(\mathbf{s}^*(t),t\right) - \nabla f_0\left(\mathbf{s}^*(t+1),t+1\right)\right\| \\ &\quad \le \sum_k \left[\left\|\nabla f_0\left(\mathbf{s}^*(t),t\right) - \nabla f_0\left(\mathbf{s}^*(t+1),t+1\right)\right\|\right]_k. \end{aligned} \tag{95}$$

Recalling the definition $f_0(\mathbf{s},t) = \sum_k f_k(\mathbf{s}_k,t)$ in (8) and the equivalence $\mathbf{s}_k^*(t) = \mathbf{s}_{\mathrm{MAP}}(t)$, which follows from the equivalence between (6) and (7), it follows that

$$\begin{aligned} &\left[\left\|\nabla f_0\left(\mathbf{s}^*(t),t\right) - \nabla f_0\left(\mathbf{s}^*(t+1),t+1\right)\right\|\right]_k \\ &\quad = \left\|\nabla f_k\left(\mathbf{s}_{\mathrm{MAP}}(t),t\right) - \nabla f_k\left(\mathbf{s}_{\mathrm{MAP}}^*(t+1),t+1\right)\right\|. \end{aligned} \tag{96}$$

The gradients in the right hand side of (96) can be written explicitly using the expression in (91) to write

$$\begin{aligned}&\Big[\big\|\nabla f_0\big(\mathbf{s}^*(t),t\big)-\nabla f_0\big(\mathbf{s}^*(t+1),t+1\big)\big\|\Big]_k\\&\quad=\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\big(\mathbf{x}_k(t+1)-\mathbf{H}'_k\mathbf{s}_{\mathrm{MAP}}(t+1)-\mathbf{x}_k(t)\\&\qquad-\mathbf{H}'_k\mathbf{s}_{\mathrm{MAP}}(t)\big)+\frac{1}{K}\tilde{\mathbf{Q}}^{-1}\big(\mathbf{s}_{\mathrm{MAP}}(t+1)-\mathbf{s}_{\mathrm{MAP}}(t)\big)\Big\|.\end{aligned}\tag{97}$$

Substituting the expression in (94) for the MAP estimates $\mathbf{s}_{\mathrm{MAP}}(t)$ and $\mathbf{s}_{\mathrm{MAP}}(t+1)$ in (97) and rearranging terms yields

$$\begin{aligned}&\Big[\big\|\nabla f_0\big(\mathbf{s}^*(t),t\big)-\nabla f_0\big(\mathbf{s}^*(t+1),t+1\big)\big\|\Big]_k\\&\quad=\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\big(\mathbf{x}_k(t+1)-\mathbf{x}_k(t)\big)\\&\qquad+\left({\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\mathbf{H}'_k+\frac{1}{K}\tilde{\mathbf{Q}}^{-1}\right)\\&\qquad\times\left(\sum_k{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\mathbf{H}'_k+\frac{1}{K}\tilde{\mathbf{Q}}^{-1}\right)^{-1}\\&\qquad\times\left(\sum_k{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\big(\mathbf{x}_k(t)-\mathbf{x}_k(t+1)\big)\right)\Big\|.\end{aligned}\tag{98}$$

Consider now the limit of the norm difference in (98) with vanishing sampling time $T_s\to 0$. We will show that this limit is of order $o(\sqrt{T_s})$. For that purpose notice that for some constant α we must have

$$\begin{aligned}&\lim_{T_s\to 0}\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\mathbf{H}'_k+\frac{1}{K}\tilde{\mathbf{Q}}^{-1}\Big\|\\&\qquad\Big\|\sum_k\Big({\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\mathbf{H}'_k+\frac{1}{K}\tilde{\mathbf{Q}}^{-1}\Big)\Big\|^{-1}=\alpha.\end{aligned}\tag{99}$$

Indeed, the second factor is the inverse of a per sensor sum of terms having the same form as the first factor. It follows that the rates at which these factors vanish are inverses of each other implying that (99) must be true.

Use the triangle and Cauchy-Schwarz inequalities in the right hand side of (98) and consider the limit of the expectation of both sides of the resulting expression as $T_s\to 0$. Further combining the result with the limit in (99) yields

$$\begin{aligned}&\lim_{T_s\to 0}\mathbb{E}\left[\Big[\big\|\nabla f_0\big(\mathbf{s}^*(t),t\big)\nabla f_0\big(\mathbf{s}^*(t+1),t+1\big)\big\|\Big]_k\,\Big|\,\mathbf{x}(0:t)\right]\\&\quad\le(\alpha+1)\lim_{T_s\to 0}\\&\qquad\sum_k\mathbb{E}\left[\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\big(\mathbf{x}_k(t)-\mathbf{x}_k(t+1)\big)\Big\|\,\Big|\,\mathbf{x}(0:t)\right].\end{aligned}\tag{100}$$

According to (111) the rate of $\mathbb{E}\left[\left[\big\|\nabla f_0\big(\mathbf{s}^*(t),t\big)-\nabla f_0\big(\mathbf{s}^*(t+1),t+1\big)\big\|\right]_k\,\Big|\,\mathbf{x}(0:t)\right]$ as $T_s\to 0$ is bounded by the rate of $\mathbb{E}\left[\big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\big(\mathbf{x}_k(t+1)-\mathbf{x}_k(t)\big)\big\|\,\big|\,\mathbf{x}(0:t)\right]$. To bound the rate of this latter term begin by noticing that due to the Cauchy-Schwarz inequality we have

$$\begin{aligned}&\mathbb{E}\left[\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\big(\mathbf{x}_k(t)-\mathbf{x}_k(t+1)\big)\Big\|\,\Big|\,\mathbf{x}(0:t)\right]\\&\quad\le\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\Big\|\mathbb{E}\left[\big\|\mathbf{x}_k(t+1)-\mathbf{x}_k(t)\big\|\,\big|\,\mathbf{x}(0:t)\right].\end{aligned}\tag{101}$$

We determine the order of $\Big\|{\mathbf{H}'_k}^T{\mathbf{R}'_k}^{-1}\Big\|$ and of $\mathbb{E}\left[\big\|\mathbf{x}_k(t+1)-\mathbf{x}_k(t)\big\|\,\big|\,\mathbf{x}(0:t)\right]$ separatedly. Starting the analysis with the last term on the right-hand side of (101), we first observe that the difference between $\mathbf{x}_k(t+1)$ and $\mathbf{x}_k(t)$ can be rewritten using the realization of the true signals $\tilde{\mathbf{s}}(t)$ and $\tilde{\mathbf{s}}(t+1)$ as well as the observation noise $\mathbf{n}_k(t)$ and $\mathbf{n}_k(t+1)$ at times t and $t+1$ as

$$\begin{aligned}&\mathbb{E}\left[\big\|\mathbf{x}_k(t+1)-\mathbf{x}_k(t)\big\|\,\big|\,\mathbf{x}(0:t)\right]\\&\quad=\mathbb{E}\left[\big\|\mathbf{H}'_k(\tilde{\mathbf{s}}(t+1)-\tilde{\mathbf{s}}(t))\mathbf{n}_k(t+1)+\mathbf{n}_k(t)\big\|\,\big|\,\mathbf{x}(0:t)\right]\\&\quad=\|\mathbf{H}'_k\|\mathbb{E}\left[\big\|(\tilde{\mathbf{s}}(t+1)-\tilde{\mathbf{s}}(t))\big\|\,\big|\,\mathbf{x}(0:t)\right]\\&\qquad+\mathbb{E}\left[\big\|\mathbf{n}_k(t+1)+\mathbf{n}_k(t)\big\|\,\big|\,\mathbf{x}(0:t)\right].\end{aligned}\tag{102}\tag{103}$$

Using the triangle and the Cauchy-Schwarz inequality yielding (103), we want to find bounds for the two summands separately. To find a bound for the first summand in (103), we define $\mathbf{A}'=\mathrm{diag}(\mathbf{A},\mathbf{A},\dots\mathbf{A})$ in a similar way as $\mathbf{H}'_k$, $\mathbf{Q}'$ and $\mathbf{R}'_k$, stacking for all times in the time window, and introduce $\mathbf{u}(t)$ to write the realization of the signal noise for times $t-W+1\dots t$. Using the system model for $\tilde{\mathbf{s}}_k(t+1)$ [cf. (22)] we get

$$\mathbb{E}\left[\|\tilde{\mathbf{s}}_k(t+1)-\tilde{\mathbf{s}}_k(t)\|\right]=\mathbb{E}\left[\|\mathbf{A}'\tilde{\mathbf{s}}_k(t)+\mathbf{u}(t)-\tilde{\mathbf{s}}_k(t)\|\right].\tag{104}$$

Recall the definition $\mathbf{A}:=\exp(\mathbf{A}_a\,T_s)$, we note that for T_s small, $\mathbf{A}'$ tends to the identity matrix,

$$\lim_{T_s\to 0}\mathbf{A}'\tilde{\mathbf{s}}_k(t)-\tilde{\mathbf{s}}_k(t)=\mathbf{I}\,\tilde{\mathbf{s}}_k(t)-\tilde{\mathbf{s}}_k(t)=0.\tag{105}$$

Combining (104) and (105) we find that the difference $\|\tilde{\mathbf{s}}_k(t+1)-\tilde{\mathbf{s}}_k(t)\|$ tends to the norm of the signal noise for vanishing sampling time. From the definition of the signal noise [cf. (24)] we can bound its expected norm using the signal noise covariance matrix,

$$\begin{aligned}\lim_{T_s\to 0}\mathbb{E}\left[\|\tilde{\mathbf{s}}_k(t+1)-\tilde{\mathbf{s}}_k(t)\|\right]&=\mathbb{E}\left[\|\mathbf{u}(t)\|\right]\\&\le\sqrt{\|\mathbf{Q}\|}=o(\sqrt{T_s}).\end{aligned}\tag{106}$$

From the same definition in (24), the norm of the signal noise covariance matrix is of order $\|\mathbf{Q}\|=o(T_s)$, leading to the last equality in (106).

To find a bound on the second term on the right-hand side of (103) $\|\mathbf{n}_k(t+1)-\mathbf{n}_k(t)\|$, note that $\mathbf{n}_k(t)$ and $\mathbf{n}_k(t+1)$ are i.i.d. Then recall the definition of $\mathbf{R}_k$ [cf. (25)] from which it follows that the expected norm of the observation noise can be bounded by the observation noise covariance matrix,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{n}_k(t+1)-\mathbf{n}_k(t)\|\right]&=2\mathbb{E}\left[\|\mathbf{n}_k(t)\|\right]\\&\le\sqrt{\|\mathbf{R}_k\|}=o\left(\frac{1}{\sqrt{T_s}}\right).\end{aligned}\tag{107}$$

The last equality in (107) also follows from the definition of $\mathbf{R}_k$ [cf. (25)]. Plugging the results from (106) and (107) back into (103), the order of the left-hand side of (102) is

$$\lim_{T_s\to 0}\mathbb{E}\left[\big\|\mathbf{x}_k(t+1)-\mathbf{x}_k(t)\big\|\,\big|\,\mathbf{x}(0:t)\right]=o\left(\sqrt{T_s}+\frac{1}{\sqrt{T_s}}\right).\tag{108}$$

To find the order of the left-hand side of (101), it is left to show the order of the first term on the right-hand side of (101). Since $\mathbf{H}'_k$ is a constant with respect to T_s, it holds that $\|\mathbf{H}\mathbf{R}'^{-1}_k\| \leq \alpha_2 \|\mathbf{R}'^{-1}_k\|$ for some positive constant α_2 for any T_s. By the definition of $\mathbf{R}_k$ [cf. (25)], it holds that $\|\mathbf{R}'^{-1}_k\| = T_s \|\mathbf{R}^{-1}_{ak}\|$. Since $\|\mathbf{R}^{-1}_{ak}\|$ is a constant, it follows that the order of the first term on the right-hand side of (101) is of order $o(T_s)$,

$$\|\mathbf{H}\mathbf{R}'^{-1}_k\| \leq \alpha_2 \|\mathbf{R}'^{-1}_k\| = o(T_s), \tag{109}$$

Combining the results from (108) and (109), it follows that

$$\lim_{T_s \to 0} \mathbb{E}\left[\left\| \mathbf{H}'^{T}_k \mathbf{R}'^{-1}_k \left(\{\mathbf{x}_k(t+1) - \mathbf{x}_k(t)\} \right) \right\| \,\middle|\, \mathbf{x}(0:t) \right] = o\left(T_s \left(\sqrt{T_s} + \sqrt{\frac{1}{T_s}} \right) \right) \leq o(\sqrt{T_s}). \tag{110}$$

Using the result in (110), we can now express the order of the left-hand side of (111) as

$$\lim_{T_s \to 0} \mathbb{E}\left[\left[\left\| \nabla f_0\big(\mathbf{s}^*(t), t\big) - \nabla f_0\big(\mathbf{s}^*(t+1), t+1\big) \right\| \right]_k \,\middle|\, \mathbf{x}(0:t) \right] = o(\sqrt{T_s}). \tag{111}$$

The result of (111) finally leads to a bound on the original desired expression in (35). Recall that the expression in the expectation from (35) can be bounded using the triangle inequality by a sum of per sensor terms [cf. (95)] whose order is $o(\sqrt{T_s})$ according to the result in (111). It follows that the original expected value from (35) is also of order $o(\sqrt{T_s})$, i.e.,

$$\lim_{T_s \to 0} \mathbb{E}\left[\left\| \nabla f_0\big(\mathbf{s}^*(t), t\big) - \nabla f_0\big(\mathbf{s}^*(t+1), t+1\big) \right\| \,\middle|\, \mathbf{x}(0:t) \right] = o(\sqrt{T_s}). \tag{112}$$

This is tantamount to Assumption 5 for $\delta(T_s) = o(\sqrt{T_s})$.

Appendix B
Log-Likelihood Gradient and Hessian for Quantized Signal Models

The computation of the Lagrangian maximizers in (31) can be performed through Newton's method. For given sensor k and time t let i denote a Newton iteration index and $\mathbf{s}^{(i)}_k(t)$ the corresponding signal determined by the algorithm in that iteration. Denote as $\nabla_{\mathbf{s}_k} \mathcal{L}_k(\mathbf{s}^{(i)}_k(t), \boldsymbol{\lambda}(t), t)$ and $\nabla^2_{\mathbf{s}_k} \mathcal{L}_k(\mathbf{s}^{(i)}_k(t), \boldsymbol{\lambda}(t), t)$ the gradient and Hessian of the local Lagrangian $\mathcal{L}_k(\mathbf{s}_k, \boldsymbol{\lambda}(t), t)$ evaluated at $\mathbf{s}_k = \mathbf{s}^{(i)}_k(t)$. Newton's descent algorithm is then defined as the iteration

$$\mathbf{s}^{(i+1)}_k(t) = \mathbf{s}^{(i)}_k(t) - \left[\nabla^2_{\mathbf{s}_k} \mathcal{L}_k\left(\mathbf{s}^{(i)}_k(t), \boldsymbol{\lambda}(t), t \right) \right]^{-1} \times \nabla_{\mathbf{s}_k} \mathcal{L}_k\left(\mathbf{s}^{(i)}_k(t), \boldsymbol{\lambda}(t), t \right), \tag{113}$$

where $\mathbf{s}^{(0)}_k(t)$ is initialized as $\mathbf{s}_k(t-1)$. To write (113) explicitly for the quantized model of Section II-C denote $p^n_{kj}(\mathbf{s}^n) := \mathrm{P}\big(y^n_{kj} = 1 \,\big|\, \mathbf{s}^n\big)$. Using this notation and taking the gradient of the maximand in (31) yields

$$\nabla_{\mathbf{s}_k} \mathcal{L}_k(\mathbf{s}^{(i)}_k(t), \boldsymbol{\lambda}(t), t) = \left\{ \sum_n \sum_{j=1}^{J} \left(y^n_{kj} \frac{\nabla p^n_{kj}(\mathbf{s}^n)}{p^n_{kj}(\mathbf{s}^n)} - (1 - y^n_{kj}) \frac{\nabla p^n_{kj}(\mathbf{s}^n)}{(1 - p^n_{kj}(\mathbf{s}^n))} \right) - \frac{1}{K} (\mathbf{I} - \mathbf{B})^T \mathbf{Q}^{-1} (\mathbf{I} - \mathbf{B}) \mathbf{s}^{(i)}_k(t) + \sum_{l \in n_k} (\boldsymbol{\lambda}^n_{kl}(t) - \boldsymbol{\lambda}^n_{lk}(t)) \right\} \tag{114}$$

where we used the definition of the matrix $\mathbf{B}$ in (90). Similarly, the Hessian is given by the following expression,

$$\nabla^2_{\mathbf{s}_k} \mathcal{L}_k(\mathbf{s}^{(i)}_k(t), \boldsymbol{\lambda}(t), t) = \left\{ \sum_n \sum_{j=1}^{J} \left(y^n_{kj} \left(- \frac{\nabla p^n_{kj}(\mathbf{s}^n) \nabla p^{nT}_{kj}(\mathbf{s}^n)}{(p^n_{kj}(\mathbf{s}^n))^2} + \frac{\nabla^2 p^n_{kj}(\mathbf{s}^n)}{p^n_{kj}(\mathbf{s}^n)} \right) - (1 - y^n_{kj}) \left(- \frac{\nabla p^n_{kj}(\mathbf{s}^n) \nabla p^{nT}_{kj}(\mathbf{s}^n)}{(1 - p^n_{kj}(\mathbf{s}^n))^2} + \frac{\nabla^2 p^n_{kj}(\mathbf{s}^n)}{1 - p^n_{kj}(\mathbf{s}^n)} \right) \right) + \frac{1}{K} (\mathbf{I} - \mathbf{B})^T \mathbf{Q}^{-1} (\mathbf{I} - \mathbf{B}) \right\}. \tag{115}$$

To complete the derivation we need to compute the derivatives of $p^n_{kj}(\mathbf{s}^n)$. Denote by $\nabla p^n_{kj}(\mathbf{s}^n)$ the gradient vector of $p^n_{kj}(\mathbf{s}^n)$, which can be found by deriving the expression in (30) to obtain

$$\nabla p^n_{kj}(\mathbf{s}^n) = \frac{1}{\sqrt{2\pi}} \mathbf{h}^T_k \times \exp\left(-\frac{1}{2} (y^n_{kj} - \mathbf{h}^T_k \mathbf{s}^n_k) r^{-1}_{kj} (y^n_{kj} - \mathbf{h}^T_k \mathbf{s}^n_k) \right). \tag{116}$$

The Hessian $\nabla^2 p^n_{kj}(\mathbf{s}^n)$ can be found by taking derivatives in the gradient expression in (116) and is given by

$$\nabla^2 p^n_{kj}(\mathbf{s}^n) = \frac{1}{\sqrt{2\pi}} \mathbf{h}_k \mathbf{h}^T_k (x^n_{kj} - \mathbf{h}^T_k \mathbf{s}^n_k) \times \exp\left(-\frac{1}{2} (y^n_{kj} - \mathbf{h}^T_k \mathbf{s}^n) r^{-1}_{kj} (y^n_{kj} - \mathbf{h}^T_k \mathbf{s}^n_k) \right). \tag{117}$$

The gradients and Hessians in (116) and (117) can be substituted into the local Lagrangian gradient and Hessian expressions in (114) and (115). The results can then be substituted into (113) to implement Newton's descent algorithm.

References

[1] F. Jakubiec and A. Ribeiro, "Distributed maximum a posteriori probability estimation of dynamic systems with wireless sensor networks," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012.

[2] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[3] L. Xiao, S. Boyd, and S. J. Kim, "Distributed average consensus with least-mean-square deviation," *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.

[4] T. C. Aysal, M. J. Coates, and M. G. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4905–4918, 2008.

[5] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, 2010.

[6] U. A. Khan and J. M. F. Moura, "Distributing the Kalman filter for large-scale systems," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4919–4935, 2008.
[7] R. Olfati-Saber and J. S. Shamma, "Consensus filters for sensor networks and distributed sensor fusion," *Proc. IEEE CDC*, pp. 6698–6703, 2005.
[8] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
[9] P. Braca, S. Marano, V. Matta, and P. Willett, "Asymptotic optimality of running consensus in testing binary hypotheses," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 814–825, Feb. 2010.
[10] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4381–4396, 2011.
[11] S. Boyd, A. Ghosh, and B. P. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.
[12] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.
[13] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering based on consensus strategies," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 622–633, 2008.
[14] S. Kar and J. M. F. Moura, "Gossip and distributed Kalman filtering: Weak consensus under weak detectability," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1766–1784, 2011.
[15] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, 2008.
[16] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, 2010.
[17] S. Y. Tu and A. H. Sayed, "Mobile adaptive networks," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 649–664, 2011.
[18] S. I. Roumeliotis and G. A. Bekey, "Distributed multirobot localization," *IEEE Trans. Robot. Autom.*, vol. 18, no. 5, pp. 781–795, 2002.
[19] K. Zhou and S. I. Roumeliotis, "Multirobot active target tracking with combinations of relative observations," *IEEE Trans. Robot.*, vol. 27, no. 4, pp. 678–695, Aug. 2010.
[20] U. A. Khan, S. Kar, and J. M. F. Moura, "Diland: An algorithm for distributed sensor localization with noisy distance measurements," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1940–1947, 2010.
[21] M. G. Rabbat, R. D. Nowak, and J. A. Bucklew, "Generalized consensus computation in networked systems with erasure links," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. Process.*, Jun. 2005, pp. 1088–1092.
[22] 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 Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, 2008.
[23] I. D. Schizas, G. B. Giannakis, S. I. Roumeliotis, and A. Ribeiro, "Consensus in ad hoc WSNs with noisy links—Part: Distributed estimation and smoothing of random signals," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1650–1666, 2008.
[24] E. J. Msechu, S. I. Roumeliotis, A. Ribeiro, and G. B. Giannakis, "Decentralized quantized Kalman filtering with scalable communication cost," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3727–3741, 2008.
[25] A. Ribeiro, G. B. Giannakis, and S. I. Roumeliotis, "SOI-KF: Distributed Kalman filtering with low-cost communications using the sign of innovations," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4782–4795, 2006.
[26] A. Ribeiro, I. D. Schizas, S. I. Roumeliotis, and G. B. Giannakis, "Kalman filtering in wireless sensor networks," *IEEE Control Syst. Mag.*, vol. 30, no. 2, pp. 66–86, 2010.
[27] P. S. Maybeck, *Stochastic Models, Estimation and Control, Volume I*. New York: Academic, 1979.
[28] S. Boyd and L. Vanderberghe, *Convex Programming*. New York: Wiley, 2004.
[29] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks—Part I: Gaussian case," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 1131–1143, 2006.
[30] M. Zargham, A. Ribeiro, A. Jadbabaie, and A. Ozdaglar, "Accelerated dual descent for network optimization," *IEEE Trans. Autom. Control*, 2011.
[31] K. L. Chung, *A Course in Probability Theory*, 3rd ed. New York: Academic, 2001.

Felicia Y. Jakubiec received the Diploma degree in electrical engineering and management from the Technical University of Berlin, Berlin, Germany, in 2009 and the Master's degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2010.

She is currently working toward the University of Pennsylvania, Philadelphia. Her research interests include statistical signal processing, signal estimation, and stochastic optimization. Her current research focuses on the theory of distributed signal processing.

Alejandro Ribeiro received the B.Sc. degree in electrical engineering from the Universidad de la Republica Oriental del Uruguay, Montevideo, in 1998 and the M.Sc. and Ph.D. degree in electrical engineering from the Department of Electrical and Computer Engineering, the University of Minnesota, Minneapolis, in 2005 and 2007, respectively.

From 1998 to 2003, he was a member of the technical staff at Bellsouth Montevideo. After his M.Sc. and Ph.D studies, in 2008 he joined the University of Pennsylvania (Penn), Philadelphia, where he is currently an Assistant Professor at the Department of Electrical and Systems Engineering. His research interests lie in the areas of communication, signal processing, and networking. His current research focuses on network and wireless communication theory.

Dr. Ribeiro received the 2012 S. Reid Warren, Jr. Award presented by Penn's undergraduate student body for outstanding teaching and the NSF CAREER Award in 2010. He is also a Fulbright scholar and the recipient of student paper awards at ICASSP 2005 and ICASSP 2006.