Distributed Bayesian Estimation in Sensor Networks: Consensus on Marginal Densities

Parth Paritosh, *Member, IEEE,* Nikolay Atanasov, *Senior Member, IEEE,* and Sonia Martínez, *Fellow, IEEE*

Abstract—In this paper, we design and analyze distributed Bayesian estimation algorithms for sensor networks. We consider estimation problems, such as cooperative localization and federated learning, where the data collected at any agent depends on a subset of all variables of interest. We provide a unified formulation of centralized, distributed and marginal probabilistic estimation as a Bayesian density estimation problem using data from non-linear likelihoods at agent. We develop distributed estimation algorithms based on stochastic mirror descent with appropriate regularization to enforce distributed or marginal density constraints. We prove almost-sure convergence to the optimal set of probabilities at each agent in both the distributed and marginal settings. Finally, we present Gaussian density versions of these algorithms and compare them to belief propagation variants in a node localization problem with relative position measurements. We also demonstrate our algorithms in a multi-agent mapping problem using LiDAR data.

Index Terms—Network optimization and control, Statistical network models, Network inference.

1 Introduction

THE advent of low-cost computing, storage and communication devices has made large sensor networks integral to urban, transportion and power-grid infrastructure. Efficient inference algorithms are needed for automated monitoring of the underlying processes. Any centralized solution to this inference problem necessitates data aggregation which, while potentially more accurate, incurs prohibitive processing and communication costs, especially in real-time settings. Real-time inference is crucial for tasks such as indoor positioning [1], urban monitoring [2], and path planning for robotic networks [3]. Thus, modern sensor networks parallelize inference across nodes improving communication efficiency and robustness to node failures.

However, most distributed algorithms do not account for the relevance of the information shared among the nodes. Motivated by this, we design algorithms to simultaneously address the inherent commmunication network constraints while accounting for variable relevance at each node.

Literature review: To achieve online estimation in connected sensor networks, researchers have studied schemes to combine distributed estimates [4], notably classified as opinion pooling [5] and graph-based message-passing algorithms [6]. Message-passing algorithms, such as Gaussian, sigma-point and non-linear belief propagation (BP), are appropriate when the causal relationships between variables are known. For further insights, see [7] and references therein. In contrast, linear and geometric averages of probabilistic estimates are commonly used to pool opinions [8] in a network with communication across one-hop neighbors. The seminal work in [4] presents a local and computationally tractable consensus estimation algorithm as a two step process, consisting of a non-Bayesian pooling step followed by a Bayesian update with locally available data.

Distributed estimation algorithms can be analyzed as steps of gradient-based optimization methods [9] that minimize the divergence between the data generating process and the estimated model. This approach establishes consistency of the estimation task, with estimation quality as the objective. For the consensus step, this approach generates algorithms beyond linear and logarithmic pooling choices, see [10], [11]. Mirror descent methods [12], [13] generalize the first-order gradient methods via metric-space projections to exploit the inherent problem geometry. Past research on distributed estimation using partially informative observation models has relied on fusing observation likelihoods with individual agent's network sized estimates [14], [15], [16]. Doan et al. [17] apply mirror descent to the linear average of neighbor estimates for consistent estimation in discrete space. Another algorithm in [15] incorporates geometric averaging with stochastic mirror descent (SMD) to achieve consensus over the network. As centralized objective, one can select the divergence between true and estimated densities to derive linear regression updates, Kalman filter and particle filters as special cases. The work in [13] further extends the SMD algorithm for finding optimal continuous-space probability density functions (pdfs), although in a centralized setting with a variationally coherent objective. More recently, [18] studied convergence of variational estimates on compact subsets of hypotheses. However, all of these papers assume that agents estimate a common set of variables and neither one includes distributional convergence guarantees.

In this work, in addition to distributing the estimation process, we focus on distributing the storage by estimating only a subset of variables relevant to the local data generating process at each node. This significantly reduces the storage and communication requirements for distributed inference. One example of estimating relevant variable sub-

- *The authors are with the Contextual Robotics Institute, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093 (e-mails: {pparitos,natanasov,soniamd}@ucsd.edu). We gratefully acknowledge support from ONR N00014-19-1-2471, ARL DCIST CRA W911NF17-2-0181, and NSF FRR CAREER 2045945.*

Manuscript received ...

sets at different nodes is a sensor network using relative measurements for node localization [19]. In this problem, the measurement likelihoods are determined by the position of node i making the measurement and the positions of the measured neighbors $\mathcal{V}_i$. A practical example of relative-measurement localization is a beacon network deployed in underwater or indoor settings using range or acoustic measurements to estimate the node positions [19], [20]. Since we estimate marginal densities over the relevant variables at different nodes, we design and analyze algorithms to enforce consistent marginals of the network-sized joint pdf.

BP [21], [22] is a widely used algorithm for probabilistic estimation of marginal densities in a network with applications in error-correcting codes [23], computer vision, and robotics [24]. This method employs node-specific observation models and pairwise interaction models between agents, utilizing messages exchanged between neighboring nodes to compute the marginal probabilities of individual variables at each node. The convergence of BP in generally not guaranteed in graphs with loops [25]. Recently proposed variants, such as α-BP [26] and circular BP [27], obtain consistent estimates in arbitrary graphs but the convergence guarantees are limited to binary probabilities. Instead of learning marginals over local node variables only, we design an algorithm estimating the marginal probability density over a set of relevant variables at each node.

The key challenge to guarantee consistency in estimating different variables at different agents is the incompatibility of the variable domains due to the different number of neighbors at each agent. This sub-problem of combining partial estimates has been framed in terms of statistical matching [28] and minimum entropy coupling [29], aiming to find a joint pdf minimizing divergence to the relevant marginal densities. A recursive optimization approach is proposed in [30], [31] but it is computationally expensive for real-time inference. In the presence of streaming measurements, our prior work [32] addressed a discrete version of this problem. However, in various applications it is necessary to consider probability densities in continuous space.

Statement of Contributions:

This work proposes a distributed Bayesian estimation algorithm to obtain marginal densities over relevant variable subsets at each node. The contributions of this paper are summarized as follows. (i) We formulate the estimation problem as a stochastic optimization over the functional space of probability density functions, presenting a unified framework to express centralized, distributed and marginal estimation in a network. This formulation relates their solutions using gradient descent variants to Bayesian estimation algorithms. (ii) We develop two distributed estimation algorithms relying on one-hop neighbor communication, one estimating densities over all unknown variables and the other estimating marginal densities only over a relevant set of variables at each agent. Our distributed marginal density estimation algorithm reduces the storage, communication, and computation requirements compared to consensus-based distributed estimation algorithms [15], [33], [18], [8] (iii) We prove novel almost-sure convergence result for our distributed and marginal algorithms. Our results apply to continuous probability densities and hold in any connected network, in contrasts with message-passing and belief propagation methods [34], [35] that generally cannot provide convergence guarantees in graphs with cycles. (iv) We demonstrate that our algorithms achieve higher estimation accuracy than belief propagation in a distributed node localization problem using relative position measurements and significantly reduce the storage and communication load compared to full state estimation algorithms in a distributed mapping problem using LiDAR data.

This paper extends our prior work [36] on estimating marginal densities over the states of an agent and its neighbors to an arbitrary set of variables by introducing a marginal consensus constraint. Additionally, we analyze the convergence of the distributed and marginal algorithms and provide a new application to distributed mapping. We also extend a relative localization example from [36] by comparing the performance of our algorithms to new variants of the BP algorithm [37], [34], [38] in networks with different connectivity and observation noise.

In Section 2, we pose the distributed estimation problem as minimizing divergence between the data-generating density and an estimated likelihood, and recall relevant mathematical preliminaries in Section 3. An SMD-based solution to this problem is presented in Section 4. Next, we solve the distributed estimation problem in Section 5 where agents maintain equal network-scale estimates. Section 6 extends the estimation problem to a marginal density setting where agents maintain estimates on variables co-estimated with one-hop neighbors. Finally, Section 7 presents a distributed relative localization example comparing the proposed algorithms with BP variants and a distributed mapping application using the marginal estimation in conjunction with variational inference.

2 Problem Formulation: Distributed Partial Parameter Estimation

We consider an estimation problem with cooperative agents in the set $\mathcal{V} = \{1, \ldots, n\}$ communicating over a static connected network. The agents aim to infer m vector values collectively given as the d-dimensional vector $\mathcal{X}^\star = [\boldsymbol{x}_1^\star, \ldots, \boldsymbol{x}_m^\star]^\top$ with $\boldsymbol{x}_v^\star \in \mathbb{R}^{d_v}$ and $d = \sum_v d_v$. With a abuse of notation, we overload $\mathcal{X}^\star$ to also denote the set of m-vectors $\{\boldsymbol{x}_v^\star\}_{v=1}^m$. The terms $\boldsymbol{x}_v^\star$ may represent the value of model parameters in a mapping problem, or the agents' pose in a relative localization problem. Each agent receives measurements from a local probability density function dependent on a subset $\mathcal{X}_i^\star \subseteq \mathcal{X}^\star$ and shares its estimates with one-hop neighbors. The variables in the local subset $\mathcal{X}_i^\star$ could represent model parameters relevant to the agent's trajectory in a mapping problem, or the agent neighbors' poses in a localization problem. Relying on the subsets $\mathcal{X}_i^\star$ instead of $\mathcal{X}^\star$ reduces the storage and communication costs of distributed estimation at individual agents.

To set up the estimation problem formally, we define a vector $\mathcal{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$ with $\boldsymbol{x}_v \in \mathbb{R}^{d_v}$ corresponding to the variables of interest $\boldsymbol{x}_v^\star$. At time step t, the known likelihood of receiving measurement $z_{i,t} \in \mathbb{R}^{\ell_i}$ by agent i is given as $\mathrm{q}_i(z_{i,t}|\mathcal{X}_i)$, where $\mathcal{X}_i \subseteq \mathcal{X}$. Thus, the measurement generation at each agent i is determined by the unknown variables $\mathcal{X}_i^\star$ via the density model $\mathrm{q}_i^\star(z_{i,t}) = \mathrm{q}_i(z_{i,t}|\mathcal{X}_i = \mathcal{X}_i^\star) \in \mathcal{F}_{\ell_i}$, where the space $\mathcal{F}_\ell$ of pdfs is defined as:

$$\mathcal{F}_\ell = \left\{ g \in \mathrm{L}^1(\mathbb{R}^\ell) \mid \int g(\boldsymbol{x}) d\boldsymbol{x} = 1, g(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathbb{R}^\ell \right\}. \quad (1)$$

We assume that $\cup_i \mathcal{X}_i^\star = \mathcal{X}^\star$ to ensure that the combined agent network can jointly observe all variables of interest. Let z_t represent all observations $z_{i,t}$ collected by the multi-agent system at time t with combined likelihood model $\mathrm{q}(z_t|\mathcal{X}) \in \mathcal{F}_\ell$, where $\ell = \sum_{i=1}^n \ell_i$.

Assumption 1 (Independence)**.** Agent i samples observation $z_{i,t}$ at time t independently across time and agents as,

$$\mathrm{q}(z_1, \ldots, z_T|\mathcal{X}^\star) = \prod_{t=1}^T \mathrm{q}(z_t|\mathcal{X}^\star) = \prod_{t=1}^T \prod_{i\in\mathcal{V}} \mathrm{q}_i(z_{i,t}|\mathcal{X}_i^\star) \quad (2)$$

Since the agents need to reach consistent estimates, any two agents observing the same variable communicate their estimates over a connected digraph $\mathcal{G}$ [3], with node set $\mathcal{V}$ and edge set $\mathcal{E}$. The neighbors of agent i, including itself, are denoted as $\mathcal{V}_i$. The communication graph has an associated non-negative adjacency matrix $A \in \mathbb{R}^{n\times n}$ with entries $A_{ij} > 0$ iff $(i,j) \in \mathcal{E}$, including self-loops. Any such matrix A representing a connected network can be made symmetric and doubly stochastic, e.g., via the Sinkhorn's algorithm [39].

Assumption 2 (Graph adjacency)**.** The connected digraph $\mathcal{G}$ is represented by a symmetric, doubly stochastic adjacency matrix A with $A\mathbf{1}_n = \mathbf{1}_n, A = A^\top$, and diagonal entries $A_{ii} > 0\,, \forall i \in \{1,\ldots,n\}$, where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector of ones.

Next, we express the estimation problem using a pdf $p(\mathcal{X}) \in \mathcal{F}_d$ instead of a point estimate in $\mathbb{R}^d$ to capture the associated epistemic uncertainty. We aim to find the pdf $p \in \mathcal{F}_d$ minimizing the objective:

$$\min_{p\in\mathcal{F}_d} \left\{ \mathop{\mathbb{E}}_{\mathcal{X}\sim p} [\mathrm{KL}[\mathrm{q}(\cdot|\mathcal{X}^\star), \mathrm{q}(\cdot|\mathcal{X})]] \right\}, \quad (3)$$

where the expectation is defined over the KL-divergence term $\mathrm{KL}[\mathrm{q}^\star, \mathrm{q}(\cdot|\mathcal{X})] = \int_{\mathbb{R}^\ell} \mathrm{q}(z|\mathcal{X}^\star) \log(\frac{\mathrm{q}(z|\mathcal{X}^\star)}{\mathrm{q}(z|\mathcal{X})}) dz$ quantifying the discrepancy between the true likelihood pdf $\mathrm{q}^\star \triangleq \mathrm{q}(\cdot|\mathcal{X}^\star)$ and the agent likelihood models. Since the divergence is zero iff $\mathrm{q}^\star = \mathrm{q}(\cdot|\mathcal{X})$ almost everywhere (a.e.) w.r.t. the Lebesgue measure, the Dirac-delta function at $\mathcal{X} = \mathcal{X}^\star$ lies in this objective's minimizer set. Please note that the equality of measures is understood in this sense throughout the manuscript. Additional minimizers would satisfy the property of **observational equivalence**; i.e., any two values $\mathcal{X}_a, \mathcal{X}_b \in \mathbb{R}^d$ are observationally equivalent, if the corresponding likelihoods satisfy $\mathrm{q}(\cdot|\mathcal{X}_a) = \mathrm{q}(\cdot|\mathcal{X}_b)$. Observational equivalence relates the solutions in pdf space to the vector space of $\mathcal{X}$. Every point $\mathcal{X}_a$ observationally equivalent to $\mathcal{X}^\star$ is included in the set of minimizers.

As we sequentially sample the true likelihood pdf $\mathrm{q}^\star$, we aim to find the minimizing argument p of the sample average approximation w.r.t. z_t as shown next. The optimization presented here follows stochastic programming [40], and we make use of the inner product notation $\langle p_1, p_2\rangle = \int p_1 p_2 dz$, for $p_1, p_2 \in \mathcal{F}_\ell$. From (3),

$$\begin{aligned} p^\star &\in \mathop{\arg\min}_{p\in\mathcal{F}_d} \left\{ \mathop{\mathbb{E}}_{\mathcal{X}\sim p} [\mathrm{KL}[\mathrm{q}^\star, \mathrm{q}(\cdot|\mathcal{X})]] \right\} \\ &= \mathop{\arg\min}_{p\in\mathcal{F}_d} \left\{ \mathop{\mathbb{E}}_{\mathcal{X}\sim p} [-\langle \mathrm{q}^\star, \log(\mathrm{q}(\cdot|\mathcal{X}))\rangle] \right\} \\ &= \mathop{\arg\min}_{p\in\mathcal{F}_d} \left\{ \mathop{\mathbb{E}}_{z_t\sim\mathrm{q}^\star} F_t[p] \right\} \equiv \mathcal{F}^\star, \end{aligned} \quad (4)$$

$$f[p] = \mathop{\mathbb{E}}_{z_t\sim\mathrm{q}^\star} F_t[p], \quad F_t[p] = \mathop{\mathbb{E}}_{\mathcal{X}\sim p} [-\log(\mathrm{q}(z_t|\mathcal{X}))], \quad (5)$$

where the first equality in (4), follows from the independence of the entropy term $\int \mathrm{q}^\star \log(\mathrm{q}^\star)$ w.r.t. $\mathcal{X}$. The set $\mathcal{F}^\star$ contains pdfs minimizing the objective function in (4). Using Fubini-Tonelli's theorem, we switch the data and state variable integrals to obtain the last equality of (4), defined using (5). Since $\mathrm{q}^\star$ is unknown, we approximate the expectation operator in the final equality of (4) in terms of sampled data in $\{z_t\}$, and state the estimation problem as follows.

Problem 1 (Centralized estimation)**.** Given observations $\{z_{i,t}\}_{i=1}^n$ and known agent likelihoods $\prod_{i=1}^n \mathrm{q}_i(z_{i,t}|\mathcal{X}_i)$ defined over the subsets of $\mathcal{X}$, find the pdf $p \in \mathcal{F}_d$ minimizing the approximation to the objective in (3):

$$\min_{p\in\mathcal{F}_d} \left\{ \frac{1}{T} \sum_{t=1}^T F_t[p] \right\}, \quad (6)$$

where the functional F_t is defined in (5).

Assuming that the estimate pdf $p \in \mathcal{F}_d$ lies in L^1, the inner product objective defined in (5) exists if the gradient of the objective is defined in the dual space L^∞. Given the gradient definition $\frac{\delta}{\delta p} F[p] = -\log(\mathrm{q}(z|\mathcal{X}))$, the dual-space norm is $\|\frac{\delta}{\delta p} F[p]\|_\infty = \sup_z[-\log(\mathrm{q}(z|\mathcal{X}))]$. Therefore, the gradient exists if the $[-\log(\mathrm{q}(z|\mathcal{X}))] < \infty$ for all choices of z. We highlight this requirement in the next assumption[1].

Assumption 3 (Bounded gradient)**.** The gradient of the objective functional $\left|\frac{\delta F}{\delta \pi}(\pi, z)\right| = |-\log(\mathrm{q}_i(z|\mathcal{X}))| \leq L$ is uniformly bounded for all $\pi \in \mathcal{F}_m$, $z \in \mathbb{R}^{d_z}$. This implies that $|\log(\mathrm{q}_i(\cdot|\mathcal{X}))|$ (resp. $\mathrm{q}_i(\cdot|\mathcal{X})$) are uniformly upper (resp. lower) bounded.

The uniform lower bound on the likelihood $0 < \alpha < \mathrm{q}_i(\cdot|\mathcal{X})$ has an 'expected data' interpretation, i.e., a strictly positive likelihood of receiving data $z_{i,t}$ at agent i.

The linearity of the objective function with respect to p and the independence assumptions on the data model are necessary to derive the algorithms in this work. The independence across time enables writing the sampling average, whereas the independence across agents allows us to obtain a distributed formulation in the following sections so that each agent i can estimate a copy or a marginal of a true pdf $p^\star$.

3 Convex Functionals and Sequences

This section reviews the stochastic mirror descent (SMD) algorithm, and relevant functional analysis and stochastic sequence results needed to apply it to functional spaces.

[1] The assumption makes use of a functional derivative defined in the following section.

3.1 The stochastic mirror descent algorithm

The SMD algorithm [41], [42] generalizes stochastic gradient descent (SGD) to non-Euclidean spaces for convex optimization problems via a divergence operator. Consider an arbitrary real-valued function $f(\boldsymbol{w}, \boldsymbol{v})$ that is convex in its first argument $\boldsymbol{w} \in \mathbb{R}^n$ for $\boldsymbol{v} \in \mathbb{R}^m$ in its second argument. We define an associated stochastic optimization problem as:

$$\min_{\boldsymbol{w}} \mathbb{E}[f(\boldsymbol{w}, \boldsymbol{v})] \approx \frac{1}{T} \sum_{t=1}^{T} f(\boldsymbol{w}, \boldsymbol{v}_t),$$

where $\{\boldsymbol{v}_t\}$ is a series of independent samples from a random variable whose distribution defines the expectation $\mathbb{E}$. Precisely computing gradient with extensive sampling is computationally expensive. Instead, the *SMD algorithm* optimizes iteratively using gradient samples $\nabla f(\boldsymbol{w}, \boldsymbol{v}_t)$ as,

$$\boldsymbol{w}_{t+1} \in \arg\min_{\boldsymbol{w}} \left\{ \langle \nabla f(\boldsymbol{w}_t, \boldsymbol{v}_t), \boldsymbol{w} \rangle + \frac{1}{\alpha_t} D_\phi(\boldsymbol{w}, \boldsymbol{w}_t) \right\}. \quad (7)$$

Here, $\langle \cdot, \cdot \rangle$ is the inner product on $\mathbb{R}^n$ and $D_\phi(\boldsymbol{w}, \boldsymbol{w}_t)$ is the *Bregman divergence* [43] between $\boldsymbol{w}$ and $\boldsymbol{w}_t$.

Definition 1 (Bregman divergence)**.** Consider a continuously differentiable and strictly convex function $\phi : \mathcal{W} \subseteq \mathbb{R}^n \to \mathbb{R}$. The *Bregman divergence* associated with ϕ for points $\boldsymbol{w}, \bar{\boldsymbol{w}} \in \mathcal{W}$ is $D_\phi(\boldsymbol{w}, \bar{\boldsymbol{w}}) := \phi(\boldsymbol{w}) - \phi(\bar{\boldsymbol{w}}) - \langle \nabla \phi(\bar{\boldsymbol{w}}), \boldsymbol{w} - \bar{\boldsymbol{w}} \rangle$.

The choice $\phi(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2$ makes D_ϕ the squared Euclidean distance and (7) the standard SGD algorithm. The convergence rate for the minimization of convex functions is $O(\frac{1}{\sqrt{T}})$, independently of the problem dimension [42].

3.2 Functional Bregman divergence and derivatives

The stochastic optimization in (6) is defined over the functional space of pdfs $\mathcal{F}_d$. Therefore, we generalize the terms in (7) to the pdf space $\mathcal{F}_d$ to apply the SMD from (6).

Consider functions $p, g \in \mathrm{L}^1(\mathbb{R}^d)$. As before, the inner product notation on $\mathrm{L}^1(\mathbb{R}^d)$ is defined as $\langle p, g \rangle := \int pg dx$, assuming the existence of this integral. A subset $\mathcal{A}$ of $\mathrm{L}^1(\mathbb{R}^d)$ is convex if and only if $\alpha p + (1-\alpha) g \in \mathcal{A}$ for any $p, g \in \mathcal{A}$ and $\alpha \in [0,1]$. Therefore, the set of pdfs $\mathcal{F}_d$ defined in (1) is a closed convex subset of $\mathrm{L}^1(\mathbb{R}^d)$. To define a divergence operator over $\mathcal{F}_d$, we consider the entropy functional $\Psi[p] = \int p \log(p) d\mu$ for $p \in \mathcal{F}_d$. Entropy is continuously differentiable and strictly convex as (i) $\mathcal{F}_d$ is convex, (ii) $x \log(x)$ is strictly convex over the positive real domain, and (iii) the integration operator is linear, so it holds that $\Psi[\alpha p + (1-\alpha) g] < \alpha \Psi[p] + (1-\alpha)\Psi[g]$ for all $p, g \in \mathcal{F}_d$, $p \neq g$ a.e.. The Bregman divergence associated with Ψ is the *Kullback-Leibler divergence* $\mathrm{KL}[p, g] := \int p \log(p/g) d\mu$. The KL-divergence inherits following properties from the Bregman divergence [43]:

- (Convexity) The functional $\mathrm{KL}[p, g]$ is convex w.r.t. the first argument $p \in \mathcal{F}_d$.
- (Generalized Pythagorean inequality) For pdf's $p_0, p_1, p_2 \in \mathcal{F}_d$, the divergence terms are related to the directional gradients of Ψ as,

$$\left\langle \frac{\delta \Psi}{\delta p}[p_2], p_0 - p_2 \right\rangle - \left\langle \frac{\delta \Psi}{\delta p}[p_1], p_0 - p_2 \right\rangle = \mathrm{KL}[p_0, p_1] - \mathrm{KL}[p_0, p_2] - \mathrm{KL}[p_2, p_1]. \quad (8)$$

The extension of SMD to pdfs in $\mathcal{F}_d$ requires a definition of the functional derivative. To evaluate how a functional F changes in the vicinity of $g \in \mathrm{L}^1(\mathbb{R}^d)$, we consider variations of g defined as $g + \epsilon\eta$, where $\eta \in \mathrm{L}^1(\mathbb{R}^d)$ and $\epsilon \geq 0$ is a small scalar. For fixed g, η, $F[g + \epsilon\eta]$ is a function of ϵ and limits can be evaluated in the usual sense.

Definition 2. ([44, p. 16]) Consider a functional $F : \mathrm{L}^1(\mathbb{R}^d) \to \mathbb{R}$ and an arbitrary function $g \in \mathrm{L}^1(\mathbb{R}^d)$. A linear functional $\frac{\delta F}{\delta g}[\eta]$ is called the *first variation* of F at g if for all $\eta \in \mathrm{L}^1(\mathbb{R}^d)$ and $\epsilon > 0$ we have

$$F[g + \epsilon\eta] = F[g] + \epsilon \frac{\delta F}{\delta g}[\eta] + o(\epsilon),$$

where $o(\epsilon)$ satisfies $\lim_{\epsilon \to 0} o(\epsilon)/\epsilon = 0$.

The first variation of a functional is related to the Gateaux derivative defined below.

Definition 3. ([45, p. 49]) A functional $F : \mathrm{L}^1(\mathbb{R}^d) \to \mathbb{R}$ is *Gateaux differentiable* at $g \in \mathrm{L}^1(\mathbb{R}^d)$, if the limit

$$F'[g, \eta] := \lim_{\epsilon \to 0+} \frac{F[g + \epsilon\eta] - F[g]}{\epsilon} \quad (9)$$

exists for any $\eta \in \mathrm{L}^1(\mathbb{R}^d)$ and there is an element $\frac{\delta F}{\delta g} \in \mathrm{L}^1(\mathbb{R}^d)$ such that $\int \frac{\delta F}{\delta g} \eta d\mu = F'[g; \eta]$. The element $\frac{\delta F}{\delta g}$ is the Gateaux derivative of functional F.

proposition]theorem For $p, g \in \mathcal{F}_d$, we have the following:

Proposition 0. 1) *If $\Lambda[p] = \langle p, g \rangle$, then $\frac{\delta \Lambda}{\delta p} = g$,*
2) *if $\Psi[p] = \langle p, \log(p) \rangle$, then $\frac{\delta \Psi}{\delta p} = 1 + \log p$,*
3) *if $\mathrm{KL}[p, g] = \langle p, \log(p/g) \rangle$, then $\frac{\delta\, \mathrm{KL}}{\delta p} = 1 + \log(p/g)$.*
Each of the above first variations allow the computation of the corresponding Gateaux derivatives following Definition 3.

Proof. See Appendix **??**. ■

Definition 4. ([46, Definition 2.4]) Let set $\mathcal{B}(\mathbb{R}^d)$ be the σ-algebra of the set $\mathbb{R}^d$. The total variation distance (TV) between two pdfs p_0, p_1 defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is,

$$\|p_0 - p_1\|_{TV} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |p_0(A) - p_1(A)|.$$

lemma]theorem The KL-divergence between pdfs $p, g \in \mathcal{F}_d$ satisfies $\mathrm{KL}[p, g] \geq 2\|p - g\|_{TV}^2$.

lemma]theorem Given functions $\Psi_0 \in \mathrm{L}^\infty$ and $p, g \in \mathrm{L}^1$, it holds that $\langle \Psi_0, p - g \rangle \leq 2\|\Psi_0\|_\infty \|p - g\|_{TV}$.

Proof. See Appendix **??**. ■

3.3 Convergent stochastic sequences

To aid with the convergence analysis of the proposed algorithms, we next introduce known sufficient conditions for convergence of sequences.

Definition 5. A filtration is an increasing nested sequence of σ-algebras, $\mathcal{Z}_1 \subseteq \mathcal{Z}_2 \subseteq \ldots$, where $\mathcal{Z}_t = \sigma(X_1, \ldots, X_t)$. If S_t is $\mathcal{Z}_t$-measurable, then $\{S_t\}$ is $\{\mathcal{Z}_t\}$-adapted.

Definition 6. A $\{\mathcal{Z}_t\}$-adapted sequence $\{X_t\}$ on the probability space $(\Omega, \{\mathcal{Z}_t\}, \mathbb{P})$ is a *martingale difference sequence* if $\mathbb{E}[|X_t|] < \infty$ and $\mathbb{E}[X_t | \mathcal{Z}_{t-1}] = 0$, a.s..

lemma]theorem Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of non-negative random variables such that $\mathbb{E}\left[X_1\right] < \infty$ and $\mathbb{E}\left[X_{t+1}|X_1,\ldots,X_t\right] \leq (1+\delta_t)X_t + \epsilon_t$, where δ_t, ϵ_t are non-negative deterministic sequences with $\sum_{t=1}^{\infty}\delta_t < \infty$, $\sum_{t=1}^{\infty}\epsilon_t < \infty$. Then, X_t converges almost surely to some random variable $X_\infty \geq 0$.

lemma]theorem Let $S_t := \sum_{\tau=1}^{t} X_\tau$ be a martingale with respect to the filtration $\mathcal{Z}_t$ on a probability space $(\Omega, \{\mathcal{Z}_t\}, \mathbb{P})$. Let $\{\beta_t\}_{t=1}^{\infty}$ be a non-decreasing sequence of positive numbers with $\lim_{t\to\infty}\beta_t = \infty$. If $\sum_{t=1}^{\infty}\beta_t^{-p}\mathbb{E}\left[|X_t|^p|\mathcal{Z}_{t-1}\right] < \infty$ a.s. for some $p \in [1,2]$, then $\lim_{t\to\infty}\beta_t^{-1}S_t = 0$ almost surely.

4 Centralized Estimation

We begin our discussion with designing and analyzing an estimation algorithm in the centralized setting, as this provides the necessary components for upcoming sections. To obtain an iterative update in $\mathcal{F}_d$, we apply the SMD algorithm to minimize the objective in (6). Then, we prove the convergence of this algorithm to the set $\mathcal{F}^\star$ composed of pdfs minimizing the objective defined in (3).

4.1 Centralized SMD algorithm

We define $\mathrm{KL}[p, p_t]$ as the KL-divergence between $p, p_t \in \mathcal{F}_d$ (c.f. Sec. 3). The generalized SMD algorithm iteratively minimizes the objective in (6) to generate pdf p_{t+1} as,

$$p_{t+1} \in \arg\min_{p\in\mathcal{F}_d}\left\{\alpha_t\left\langle\frac{\delta F_t}{\delta p}[p_t], p\right\rangle + \mathrm{KL}[p,p_t]\right\}. \quad (10)$$

Let us define the term $J_t[p,p_t] = \alpha_t\langle\frac{\delta F_t}{\delta p}[p_t], p\rangle + \mathrm{KL}[p,p_t]$ as the shorthand for the minimization objective at each iteration. The functional $J_t[p,p_t]$ is convex in pdf p as it is a linear combination of a convex entropy and linear functionals. The SMD algorithm is guaranteed to optimize any convex functional F using noisy gradients if the steps α_t satisfy the following condition:

Assumption 4 (Robbins-Monro condition)**.** The positive step-size sequence $\{\alpha_t\}$ is square-summable but not summable i.e. $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$.

proposition]theorem The closed-form solution to (10) is,

$$p_{t+1} = \frac{1}{Z_t}\exp\left(-\alpha_t\frac{\delta F_t}{\delta p}[p_t]\right)p_t, \quad (11)$$

where $Z_t = \int\exp\left(-\alpha_t\frac{\delta F_t}{\delta p}[p_t]\right)p_t$.

Proof. See Appendix **??**. ■

For our specific choice of $F_t[p] = -\langle\log \mathrm{q}(z_t|\mathcal{X}), p\rangle$,

$$\frac{\delta F_t}{\delta p}[p_t] = -\log \mathrm{q}(z_t|\,\mathcal{X}).$$

Applying Proposition leads to the following pdf update,

$$p_{t+1} = \mathrm{q}(z_t|\mathcal{X})^{\alpha_t}p_t/\left(\int \mathrm{q}(z_t|\mathcal{X})^{\alpha_t}p_t\right). \quad (12)$$

Assumption 5 (Positive initial probability)**.** The prior pdf at initial time step is strictly positive, i.e., $p_0 > 0, \forall\mathcal{X}$.

Assuming a positive initial pdf is sufficient to estimate any possible pdfs. A weaker assumption would require that the positive domain of pdf $p^\star$ is contained within the positive domain of the prior $p_0 > 0$.

4.2 Almost sure convergence with centralized SMD

In this subsection, we study the convergence properties of the estimated pdf p_t to the optimal set $\mathcal{F}^\star$ under the centralized SMD algorithm. The first theorem proves that the KL divergence between any optimal pdf $p^\star \in \mathcal{F}^\star$ and p_t converges to a constant, while the second result shows that this constant is zero. To begin, we introduce the divergence neighborhood of a set of pdfs as,

Definition 7 (ϵ-Divergence neighborhood)**.** The ϵ-neighborhood $\mathbb{B}(\mathcal{F}^\star,\epsilon)$ of the pdf set $\mathcal{F}^\star$ is given as,

$$\mathbb{B}(\mathcal{F}^\star,\epsilon) = \left\{p \in \mathcal{F}_d|\min_{p^\star\in\mathcal{F}^\star}\mathrm{KL}[p^\star,p] \leq \epsilon\right\}.$$

Here, we choose the order of the pdf arguments in the divergence term to match the unknowns in the objective function. This definition aids the upcoming analysis. The proofs to the following claims are in **Appendix ??**.

proposition]theorem Let pdf p_{t+1} in (11) minimize the optimization argument $J_t[p,p_t]$ with arbitrary pdf $p \in \mathcal{F}_d$ in (10), then the change in divergence in each update is upper bounded as,

$$\mathrm{KL}[p,p_{t+1}] - \mathrm{KL}[p,p_t] \leq \alpha_t\langle\frac{\delta F_t[p_t]}{\delta p}, p - p_t\rangle + 2\alpha_t^2L^2.$$

This previous result relies on the sampled gradient of the objective $\frac{\delta}{\delta p}F_t[p]$, that we next relate to its expected value.

lemma]theorem Under Assumption 3, the gradient of the expected value of objective functional defined in (5) is equal to the expectation of its gradient, i.e. $\frac{\delta f}{\delta p}[p_t] = \underset{z_t\sim \mathrm{q}^\star}{\mathbb{E}}\frac{\delta F_t}{\delta p}[p_t]$.

Next, we will employ Proposition to upper bound the divergence from the estimate to the optimal set $\mathcal{F}^\star$ to show convergence of this divergence term.

Theorem 1. *Under Assumptions 1-5, the KL-divergence $\mathrm{KL}[p^\star,p_t]$ between any minimizer $p^\star \in \mathcal{F}^\star$ and the estimate p_t generated by the SMD algorithm in (12) converges almost surely to some finite value.*

Next, we use Theorem 1 to prove almost sure convergence of the divergence terms arbitrarily close to zero.

Theorem 2. *Under Assumptions 1-5, the pdf sequence $\{p_t\}$ generated by the SMD algorithm in (12) converges almost surely to an ϵ-divergence neighborhood $\mathcal{B}(\mathcal{F}^\star,\epsilon)$ around the set of minimizers in $\mathcal{F}^\star$ for any $\epsilon > 0$.*

Theorem 2 establishes the convergence of the pdf iterates in centralized SMD algorithm to ϵ-divergence neighborhood of the optimal set $\mathcal{F}^\star$. We have shown this result for adaptive learning rate α_t satisfying Robbins-Monro condition. While this is sufficient to prove almost sure convergence of the centralized update in (12), we can leverage the existence of an adaptive learning rate to prove that the objective function converges at the rate $O(1/\sqrt{T})$.

Theorem 3. *For a natural filtration of observations $\mathcal{Z}_{t-1} = \sigma_t(z_1,\ldots,z_{t-1})$, and the adaptive step sizes $\alpha_t < (f[p_t] - f[p^\star])/2L^2$, the expected objective function satisfies,*

$$f[\bar{p}_t] - f[p^\star] \leq \sqrt{\frac{8L^2\,\mathrm{KL}[p^\star,p_0]}{t}}, \quad (13)$$

where $\bar{p}_t = \frac{1}{t}\sum_{k=1}^{t}p_k$ and $p^\star$ minimizes $f[p]$.

In this section, we have established the weak convergence of pdf estimates in a centralized setting for the proposed SMD algorithm with square summable step sizes. Additionally, we have shown existence of a decaying step size that achieves a $\mathcal{O}(1/\sqrt{t})$ convergence rate.

5 Distributed Estimation

In this section, we present and analyze a distributed estimation algorithm in which each agent updates a pdf for all variables and shares it with one-hop neighbors. While our proposed algorithm is similar to [18], [50], our novel analysis demonstrates almost sure convergence to a common pdf in a functional space. This analysis is integral for the subsequent analysis of the marginal distributed algorithm in Section 6.

5.1 Distributed estimation problem

We start by setting up a distributed estimation problem, noting the separability of the objective function F in (6) across agents. Since agents sample z_i independently, the likelihood and the data-generating density are separable across agents as,

$$\mathrm{q}(z|\mathcal{X}) = \prod_{i=1}^{n} \mathrm{q}_i(z_i|\mathcal{X}_i), \quad \mathrm{q}^\star(z_t) = \prod_{i\in\mathcal{V}} \mathrm{q}_i^\star(z_{i,t}). \tag{14}$$

Thus, each component of F can be expressed in terms of the likelihood of the agents' private observations. That is, the centralized objective in (5) separates across agents as $F_t[p] = \sum_{i=1}^{n} F_{i,t}[p_i]$, where,

$$F_{i,t}[p_i] = \mathop{\mathbb{E}}_{\mathcal{X}\sim p_i}[-\log(\mathrm{q}_i(z_{i,t}|\mathcal{X}_i))]. \tag{15}$$

Here, the expectation is computed using the variables in $\mathcal{X}_i$ even though the samples from p_i contain all variables in $\mathcal{X}$.

Problem 2 (Distributed Estimation)**.** Given observations $z_{i,t}$ and agent likelihoods $\mathrm{q}_i(z_{i,t}|\mathcal{X}_i)$, for each $i \in \mathcal{V}$, find the pdf $p_i \in \mathcal{F}_d$ minimizing the sample average approximation to the agent objective defined using F_i in (15) as:

$$\min_{p_i\in\mathcal{F}_d} \left\{ \frac{1}{T}\sum_{t=1}^{T} F_{i,t}[p_i] \right\}, \text{ s.t. } p_i = p_j, \forall i,j \in \mathcal{V}, \tag{16}$$

under the consensus constraint enforcing equal estimates.

5.2 Distributed SMD algorithm

For Problem 2, each agent i learns a copy p_i of the pdf solution $p \in \mathcal{F}^\star$. Taking inspiration from the centralized setting, we deploy the SMD algorithm at any time t to compute pdf $p_{i,t+1}$ based on agent i's local log-likelihood samples and a prior mixed with neighbor estimates as,

$$\min_{p\in\mathcal{F}_d} J_{i,t}[p, v_{i,t}], \;\; v_{i,t} = \prod_{j\in\mathcal{V}_i} (p_{j,t})^{A_{ij}}, \tag{17}$$
$$J_{i,t}[p, v_{i,t}] = -\langle \log \mathrm{q}_i(z_{i,t}|\mathcal{X}), p\rangle + \frac{1}{\alpha_t}\mathrm{KL}[p, v_{i,t}].$$

To achieve consensus, we substitute the prior $p_{i,t}$ with the mixed pdf $v_{i,t}$, a geometric average of neighbor estimates $p_{j,t}$ weighted by terms A_{ij} satisfying Assumption 2. Thus, the distributed update at agent i is,

$$p_{i,t+1} = \mathrm{q}_i(z_{i,t}|\mathcal{X}_i)^{\alpha_t} v_{i,t} / \left(\int \mathrm{q}_i(z_{i,t}|\mathcal{X}_i)^{\alpha_t} v_{i,t} \right). \tag{18}$$

The work in [18] makes use of geometrically averaged neighbor estimates to achieve consensus. They analyze the convergence of probabilities estimated by this algorithm over compact sets in the domain of variables $\mathcal{X}$. With this consensus update, [50] shows the convergence of the modes of estimated pdfs to the same optimizer as the centralized case. Instead of these probability concentration results to the optimal parameter, we prove almost sure convergence of the KL-divergence between the estimated and an optimal pdf in $\mathcal{F}^\star$ defined over the continuous domain.

Our **analysis strategy** first studies the relative change of the algorithm mixing-step with respect to the previous algorithm iterate with respect to a reference pdf (cf. Section 5.3), then provides summable upper-bounds for various sequential differences (cf. Section 5.4), then uses these to eventually prove convergence to the optimal probability density $p^\star$ (cf. Section 5.5). In what follows, the expected value of centralized and agent-specific objectives are,

$$f[p] = \mathop{\mathbb{E}}_{z_t\sim\mathrm{q}^\star(z_t)} F_t[p], \; f_i[p_i] = \mathop{\mathbb{E}}_{z_{i,t}\sim\mathrm{q}_i^\star(z_{i,t})} F_{i,t}[p_i],$$

and their derivatives as $\frac{\delta f}{\delta p}$ and $\frac{\delta f_i}{\delta p_i}$. By the linearity of the expectation operator, it follows that $f[p] = \sum_{i=1}^{n} f_i[p]$. The proofs to our claims are presented in **Appendix ??**.

5.3 Analysis of probability-mixing steps

We first analyze the convergence characteristics of the mixing step; that is the behavior of $v_{i,t}$ relative to $p_{i,t}$ for all t and i. This analysis entails the definition of a consensus manifold for the estimated pdfs.

Definition 8. The **consensus manifold** for a connected graph $\mathcal{G}$ satisfying Assumption 2 is a set $\mathcal{M}$ of pdfs that are a.e. equal to some pdf $\bar{p} \in \mathcal{F}_d$,

$$\mathcal{M} = \left\{ \{p_{i,t}\}_{i=1}^{n} \,|\, \sum_{i=1}^{n} \mathrm{KL}[\bar{p}, p_{i,t}] = 0, p_{i,t} \in \mathcal{F}_d, \bar{p} \in \mathcal{F}_d \right\}.$$

Note that the estimated pdfs lying on the consensus manifold are equal a.e. Now, we show that the divergence between any pdf $p \in \mathcal{F}_d$ to the estimated pdfs $\{p_{i,t}\}$ decreases under the mixing step in (17), unless the pdfs lie on the consensus manifold. This result is critical to work with ϵ-divergence neighborhoods around optimal pdfs.

proposition]theorem The sum of divergences between an arbitrary pdf $p \in \mathcal{F}_d$ to the estimates $p_{i,t} \in \mathcal{F}_d$ upper bounds the divergence sum to the agent geometric averages $v_{i,t} = \frac{1}{Z_{i,t}^v}\prod_{j=1}^{n} p_{j,t}^{A_{ij}}$ with normalization factor $Z_{i,t}^v = \int \left(\prod_{j=1}^{n} p_{j,t}^{A_{ij}}\right) d\mathcal{X}$ as,

$$\sum_{i=1}^{n} \mathrm{KL}[p, v_{i,t}] \leq \sum_{i=1}^{n} \mathrm{KL}[p, p_{i,t}],$$

with equality holding iff pdfs $\{p_{i,t}\}$ lie on the consensus manifold.

The previous proposition establishes that the sum of divergences from an arbitrary pdf to agent estimates decreases with the mixing step. The next proposition establishes a geometric contraction rate for the consensus step of the algorithm to the network wide average $p_t \propto \prod_{i=1}^{n} p_{i,t}^{1/n}$.

proposition]theorem(See [51, Theorem 5]) Under Assumption 2, we have $\|v_{i,t}(\boldsymbol{x}) - p_t(\boldsymbol{x})\|_{TV} \leq \sigma(A)\|p_{i,t}(\boldsymbol{x}) - p_t(\boldsymbol{x})\|_{TV}$ with $\sigma(A) < 1$.

This allows us to later prove distributed estimation guarantees similar to Theorem 1. Based on the consensus results, we continue to analyzing objective functional evaluated at probability estimates and their geometric average.

5.4 Probability-mixing and algorithm iterate gaps

In this subsection, we prove the sequence of total variation (TV) distance between terms after likelihood updates are summable. Summability of positive sequences [52] implies vanishing terms, and this property aids our convergence results in the next Subsection 5.5. More specifically, we upper bound TV distances between the mixed pdf $v_{i,t}$, agents' next estimate $p_{i,t+1}$, and network wide-averages p_t, p_{t+1}. Next, we upper bound the TV distance between the mixed prior $v_{i,t}$ and estimate $p_{i,t+1}$.

proposition]theorem Under Assumption 3, the pdf $p_{i,t+1}$ minimizing the distributed objective $J_{i,t}[p, v_{i,t}]$ in (17) satisfies,

$$\alpha_t L\|v_{i,t} - p_{i,t+1}\|_{TV} \leq \frac{\alpha_t^2 L^2}{2}.$$

Note that the upper bound in Proposition relies on the boundedness of log-likelihood from the Assumption 3. We show that a similar bound exists for the geometric average $p_t \propto \prod_{i=1}^n p_{i,t}^{1/n}$, a proxy for centralized estimate.

proposition]theorem Let Assumptions 2-3 hold. Following the distributed SMD algorithm in (17), the update to the geometric average $p_t = \prod_{i=1}^n p_{i,t}^{1/n}/Z_t$ for normalization factor $Z_t = \left(\int \prod_{i=1}^n p_{i,t}^{1/n} d\mathcal{X}\right)$ satisfies $\|p_t - p_{t+1}\|_{TV} \leq \alpha_t L/2$.

The presence of α_t in the upper bound limits the relative error between network estimates at each time step. Now, we study the convergence of the TV distances between the agent estimates $p_{i,t}$ to the geometric average p_t and the true pdf $p^\star$. To establish vanishing distances, we bypass the need for a geometric rate of contraction like Proposition by showing the summability of this sequence with distance terms. The following technical result relates the difference between objective functions at these pdfs to the TV distance.

proposition]theorem For the pdf estimates in (18), the sum of objectives is upper bounded as $\alpha_t \sum_{i=1}^n (f_i[p^\star] - f_i[v_{i,t}]) \leq 2\sigma\alpha_t L \sum_{i=1}^n \|p_t - p_{i,t}\|_{TV}$ for $\sigma < 1$.

Now, we show that the upper bounding distance between the average p_t and estimate $p_{i,t}$ in Proposition is summable. With decaying step-size α_t, this implies that the individual estimates would converge to their geometric average. In comparison to the last subsection, here the averages include the likelihood updates across time.

proposition]theorem Under Assumptions 2-3, the updates in (17) lead to a summable sequence of distance terms $\alpha_t L \sum_{i=1}^n \|p_t - p_{i,t}\|_{TV}$ between the geometric average p_t and agent estimates.

5.5 Almost sure convergence with distributed SMD

Aided by the preliminary results, we prove the convergence of the distributed estimation algorithm with the next two theorems. The first theorem shows almost sure convergence of the KL-divergence between the estimated and true pdf to a finite positive value, and the next one proves existence of a subsequence of pdf estimates to the optimal set.

Theorem 4. *Under Assumptions 1-5, the divergence functional $\sum_{i=1}^n \mathrm{KL}[p^\star, v_{i,t}]$ of the mixed pdf sequence $\{v_{i,t}\}_{i\in\mathcal{V}}$ generated via distributed SMD algorithm in (17) almost surely converges to some non-negative value.*

Next, we show that the divergence sum in Theorem 4 converges arbitrarily close to zero.

Theorem 5. *Under Assumptions 1-5, the sequence $\{v_{i,t}\}$ generated by applying distributed SMD algorithm in (17) converges almost surely to ϵ-divergence neighborhood $\mathcal{B}(\mathcal{F}^\star, \epsilon)$ around optimal pdf set $\mathcal{F}^\star$ for any $\epsilon > 0$.*

This proves that the pdf estimates generated by the proposed algorithm in a connected network almost surely converge to the set of optimal pdfs. Based on the proposed distributed estimation algorithm and its analysis, we will extend our discussion to estimating marginal pdfs over subset of variables $\mathcal{X}$ in connected networks.

6 Distributed Marginal Estimation

In several inference problems over networks, the data likelihood at a node depends on the state of that node and its one-hop neighbors, rather than the entire network. Motivated by this, this section extends the distributed SMD algorithm to find marginal densities defined over a relevant subset of variables at each node. First, we derive a distributed estimation objective, then modify the algorithm to store and update pdf over node-specific variable sets, and finally discuss the convergence properties.

6.1 Distributed Marginal Estimation Problem

We aim to estimate the marginal density of local subsets of variables $\mathcal{X}_i$ at each agent i. This is enabled by Assumption 1 that establishes the independence among the observations $z_{i,t}$ generated using likelihoods $\mathrm{q}_i(z_{i,t}|\mathcal{X}_i)$ at agents $i \in \mathcal{V}$. Let us denote the set of variables common to agents i, j as $\mathcal{X}_{ij} = \mathcal{X}_i \cap \mathcal{X}_j$. For a well-posed estimation problem, we assume the existence of a communication pathway between agents i, j estimating any common variables in $\mathcal{X}_{ij}$.

Assumption 6 (Marginal consensus)**.** The set of agents $\mathcal{V}(\boldsymbol{x}_i) \subseteq \mathcal{V}$ estimating the same variable $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$ induces a connected subgraph $\mathcal{G}(\boldsymbol{x}_i)$ of $\mathcal{G}$ with edge set $\mathcal{E}(\boldsymbol{x}_i) = \{(j,k) \in \mathcal{E} | \forall j,k \in \mathcal{V}(\boldsymbol{x}_i)\}$.

For a given communication network, the problem of assigning connected subgraphs to estimate particular variables is NP-hard, with a feasible solution presented in [32]. We will leverage this assumption to design our marginal estimation algorithm, and show that it achieves consistent estimates on the relevant subspaces.

We follow the distributed SMD derivation in Section 5 to distribute the centralized estimation objective in (3) along the agents' independent observations. We first drop the entropy term unrelated to the optimization argument of the objective in (3). Then, the observational independence in (14) allows us to define objective functionals of marginal pdfs $p_i(\mathcal{X}_i)$ integrated along individual observations as,

$$\begin{aligned}&\min_p \mathop{\mathbb{E}}_{\mathcal{X}\sim p}[\mathrm{KL}[\mathrm{q}^\star(z_{1:n}), \mathrm{q}(z_{1:n}|\mathcal{X})]]\\ &= \min_p \mathop{\mathbb{E}}_{\mathcal{X}\sim p} \sum_{i\in\mathcal{V}} \left[\int_{z_{1:n}} -\mathrm{q}^\star(z_{1:n})\log(\mathrm{q}_i(z_i|\mathcal{X}_i))\right]\\ &= \min_p \sum_{i\in\mathcal{V}} \mathop{\mathbb{E}}_{\mathcal{X}_i\sim p_i}\left[\int_{z_i} -\mathrm{q}_i^\star(z_i)\log(\mathrm{q}_i(z_i|\mathcal{X}_i))\right]\\ &= \sum_{i\in\mathcal{V}}\min_{p_i} \mathop{\mathbb{E}}_{\mathcal{X}_i\sim p_i}\mathop{\mathbb{E}}_{z_i\sim q_i^\star}[-\log(\mathrm{q}_i(z_i|\mathcal{X}_i))] = \min_p f[p],\end{aligned}$$

where each pdf $p_i(\mathcal{X}_i) \in \mathcal{F}_{\mathfrak{d}_i}$ is a marginal of the joint pdf $p(\mathcal{X}) \in \mathcal{F}_d$ and $\mathfrak{d}_i$ is the dimension of $\mathcal{X}_i$. Making the objective $f[p]$ distributed along marginals $p_i(\mathcal{X}_i)$ is possible with additional equality constraints on the shared states $\mathcal{X}_{ij}$. These constraints are represented as agreement on marginal pdfs $p_i, \forall i \in \mathcal{V}$ over shared variables as,

$$\int p_i(\mathcal{X}_i)d\boldsymbol{x}|_{\boldsymbol{x}\in\mathcal{X}_i\backslash\mathcal{X}_{ij}} = \int p_j(\mathcal{X}_j)d\boldsymbol{x}|_{\boldsymbol{x}\in\mathcal{X}_j\backslash\mathcal{X}_{ij}}, \forall(i,j)\in\mathcal{E},$$

where $\int p_j d\boldsymbol{x}|_{\boldsymbol{x}\in\mathcal{X}_j\backslash\mathcal{X}_{ij}}$ defines an integral over all variables in the set $\mathcal{X}_j\backslash\mathcal{X}_{ij}$. As before, a finite objective allows using Fubini-Tonelli's theorem to switch the order of expectations. Along with a sample-average approximation of the integral over data in $\{z_{i,t}\}$, the online objective is expressed as,

$$\begin{aligned}\min_p f[p] &= \sum_{i\in\mathcal{V}}\min_{p_i} f_i[p_i],\\ f_i[p_i] &= \mathop{\mathbb{E}}_{z_i\sim q_i^\star}\mathop{\mathbb{E}}_{\mathcal{X}_i\sim p_i}[-\log(\mathrm{q}_i(z_i|\mathcal{X}_i))]\\ &\approx \sum_{i\in\mathcal{V}}\min_{p_i}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{X}_i\sim p_i}[-\log(\mathrm{q}_i(z_{i,t}|\mathcal{X}_i))].\end{aligned}$$

Thus, the distributed objective at time t becomes,

$$F_{i,t}[p_i] = \mathop{\mathbb{E}}_{\mathcal{X}_i\sim p_i}[-\log(\mathrm{q}_i(z_{i,t}|\mathcal{X}_i))]. \tag{19}$$

Problem 3 (Distributed marginal estimation)**.** Given observations $z_{i,t}$ and agent likelihoods $\mathrm{q}_i(z_{i,t}|\mathcal{X}_i)$ at any agent $i \in \mathcal{V}$, find pdf $p_i \in \mathcal{F}_{\mathfrak{d}_i}$ minimizing:

$$\min_{p_i\in\mathcal{F}_{\mathfrak{d}_i}}\left\{\frac{1}{T}\sum_{t=1}^{T}F_{i,t}[p_i]\right\}, \text{s.t. } p_i(\mathcal{X}_{ij}) = p_j(\mathcal{X}_{ij}), \tag{20}$$

for all agents $i,j \in \mathcal{V}$ over the marginal pdfs $p_i(\mathcal{X}_{ij}) = \int p_i(\mathcal{X}_i)d\boldsymbol{x}|_{\boldsymbol{x}\in\mathcal{X}_i\backslash\mathcal{X}_{ij}}$.

6.2 Distributed Marginal SMD Algorithm (DMSMD)

Similar to Sec. V, each agent i applies the SMD algorithm to its local objective in (20), with two exceptions. Firstly, the agents locally estimate a pdf over relevant variables $p_{i,t}(\mathcal{X}_i)$, and secondly, they enforce marginal consensus constraint equating agent i's marginal $p_{ij} = \int_{\mathcal{X}_i\backslash\mathcal{X}_{ij}} p_i$ to agent j's marginal p_{ji}. As before, the likelihood update follows from the Gateaux derivative $\frac{\delta}{\delta p_i}F_{i,t}[p_i] = -\log(\mathrm{q}_i(z_{i,t}|\mathcal{X}_i))$ as computed for linear functional in Proposition [0 .

Each agent i co-estimates some variables with its one-hop neighbors. Therefore, it merges neighbor j's information over shared variables $\mathcal{X}_{ij}$ to own estimate on distinct variables $\mathcal{X}_i\backslash\mathcal{X}_{ij}$. The incoming density over the shared variables is $p_{ji,t}(\mathcal{X}_{ij})$ and the self-conditional density at agent i over distinct variables w.r.t. neighbor j is given by $p_{i,t}(\mathcal{X}_i\backslash\mathcal{X}_{ij}|\mathcal{X}_{ij})$. The marginal agreement is enforced with geometric averaging on self-conditional and neighbor-marginals product $\tilde{p}_{ji,t}$ as,

$$v_{i,t} = \frac{1}{Z^v_{i,t}}\prod_{j\in\mathcal{V}_i}(\tilde{p}_{ji,t})^{A_{ij}},\; Z^v_{i,t} = \int\prod_{j\in\mathcal{V}_i}(\tilde{p}_{ji,t})^{A_{ij}}, \tag{21}$$

$$\tilde{p}_{ji,t} = p_{i,t}(\mathcal{X}_i\backslash\mathcal{X}_{ij}|\mathcal{X}_{ij})p_{ji,t}(\mathcal{X}_{ij}), \tag{22}$$

$$p_{ji,t}(\mathcal{X}_{ij}) = \int p_{j,t}(\mathcal{X}_j)d\boldsymbol{x}|_{\boldsymbol{x}\in\mathcal{X}_j\backslash\mathcal{X}_{ij}}.$$

Now, applying the SMD algorithm with the gradient defined as negative log-likelihood sample in Section 5, and the mixed pdf $v_{i,t}$ in (21), the marginal consensus estimation is performed as follows,

$$\begin{aligned}p_{i,t+1}(\mathcal{X}_i) &\in \mathop{\arg\min}_{p\in\mathcal{F}_{\mathfrak{d}_i}} J_{i,t}[p, v_{i,t}], \qquad (23)\\ J_{i,t}[p, v_{i,t}] &= \left\{\alpha_t\left\langle\frac{\delta F_{i,t}}{\delta p}[p_{i,t}], p\right\rangle + \mathrm{KL}[p, v_{i,t}]\right\}.\end{aligned}$$

We summarize the updates for agent i at time t in **Algorithm 1**. The algorithm consists of **edge merging**, **geometric pooling**, **likelihood update** and **message generation**. At each agent, these steps correspond to self-conditional and neighbor-marginal products, their weighted average, Bayesian likelihood update, and generation of marginal densities for its neighbors.

In comparison to the distributed algorithm in Section 5, estimating the marginals reduces the set of stored variables at agent i to $\mathcal{X}_i$ with dimensions $\mathfrak{d}_i < d$. The size of the communicated messages reduces from a pdf in $\mathcal{F}_d$ over all network variables to a partial set $\mathcal{X}_{ij}$ shared between sensors i,j. Although, each node additionally computes the conditional density. The trade-off between memory and computation depends on the average degree in the network.

Following the previous section on distributed algorithm, our **analysis strategy** first discusses the monotonic convergence of estimates under marginal mixing step to an invariant consensus manifold defined later (cf. Section 6.3). , and then presents a specific independent variable setting for similar results in terms of total variation distances (cf. Section 6.4). We use them to establish summable upper-bounds for sequential differences between marginal estimates, and eventually prove convergence to the marginals of the optimal probability density $p^\star$ (cf. Section 6.5). All proofs to the claims in this section are in **Appendix ??**.

6.3 Marginal Consensus Analysis

In this subsection, we establish the invariance and convergence properties of the marginal consensus steps defined in (21). We define a marginal consensus manifold and analyze convergence of the consensus steps to this manifold.

Definition 9. The marginal consensus manifold for a graph $\mathcal{G}$ that satisfies Assumption 6 is a set $\mathcal{M} = \{\{p_{i,t}\}_{i=1}^n \,|\, \sum_{i=1}^n \mathrm{KL}[\bar{p}_i, p_{i,t}] = 0, p_{i,t}\in\mathcal{F}_{\mathfrak{d}_i}, \bar{p}\in\mathcal{F}\}$ of marginal pdfs consistent with some joint pdf $\bar{p} \in \mathcal{F}$.

The manifold consists of coherent marginals of some joint pdf $\bar{p}$ with $p_{i,t} = \bar{p}_i \in \mathcal{F}_{\mathfrak{d}_i}$ for all agents. The following technical result shows that the product of normalization

```
Inputs: estimate $p_{i,t}(\mathcal{X}_i)$, weights $\{A_{ij}\}_{j\in\mathcal{V}_i}$, neighbor
        messages $p_{ji,t}(\mathcal{X}_{ij})$, measurement $z_{i,t}$,
        measurement model $\mathrm{q}_i(z_{i,t}|\mathcal{X}_i)$
// Receive neighbor messages.
for $j \in \mathcal{V}_i$ do
    Common marginals at neighbors
    $p_{ji,t}(\mathcal{X}_{ij}) = \int_{\mathcal{X}_j\backslash\mathcal{X}_{ij}} p_{j,t}(\mathcal{X}_j)$
// Combine neighbor estimates.
for $j \in \mathcal{V}_i$ do
    Product of $j$'s marginal and $i$'s conditional:
    $\tilde{p}_{ji,t} = p_{i,t}(\mathcal{X}_i\backslash\mathcal{X}_{ij}|\mathcal{X}_{ij})p_{ji,t}(\mathcal{X}_{ij})$
Weighted average: $v_{i,t}(\mathcal{X}_i) := \prod_{j\in\mathcal{V}_i}\tilde{p}_{ji,t}(\mathcal{X}_i)^{A_{ij}}$
// Bayesian update.
$p_{i,t+1}(\mathcal{X}_i) = \mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)v_{i,t}(\mathcal{X}_i)$
```

Algorithm 1: Marginal density averaging at agent i

factors of mixed pdfs obtained after applying (21) to pdfs in the marginal consensus manifold $\mathcal{M}$ is 1.

proposition]theorem The product of normalization factors of mixed marginals satisfies $\prod_{i=1}^n Z_{i,t}^v = 1$, where $Z_{i,t}^v = \int \prod_{j=1}^n (\tilde{p}_{ji,t})^{A_{ij}} \, d\mathcal{X}_i$, if and only if the original pdfs $\{p_{i,t}\}$ lie on the marginal consensus manifold $\mathcal{M}$.

Next, we establish that the sum of KL divergences decreases strictly due to marginal mixing step if the agent pdfs are not on the marginal consensus manifold.

proposition]theorem For any pdf $p \in \mathcal{F}$, the mixed and original pdfs $\{v_{i,t}\}, \{p_{i,t}\}$, defined in the mixing step (23), satisfy

$$\sum_{i=1}^n \mathrm{KL}[p_i, v_{i,t}] \leq \sum_{i=1}^n \mathrm{KL}[p_i, p_{i,t}],$$

with equality if and only if the original pdfs $\{p_{i,t}\}$ lie on the marginal consensus manifold $\mathcal{M}$ in Definition 9.

To study convergence properties of marginal consensus manifold, denote $p_{i,t}^{(k)}$ as the pdf computed at agent i after the k-step marginal mixing from (21) on estimated pdfs $\{p_{i,t}\}$. For instance, mixed pdf $v_{i,t} = p_{i,t}^{(1)}$. Based on the consensus properties established in Propositions -, we show that the pdfs $p_{i,t}^{(k)}$ converge to the marginal pdfs $\bar{p}_{i,t}$ in the marginal consensus manifold $\mathcal{M}$ of Definition 9.

proposition]theorem Repeated application of the marginal consensus steps in (21) to pdfs $\{p_{i,t}\}$ leads to a limit pdf $\lim_{k\to\infty} p_{i,t}^{(k)}$ that lies in the marginal consensus manifold in Definition 9.

As a consequence of Proposition , the estimates after marginal mixing converge to marginals $\bar{p}_{i,t}$ on the manifold $\mathcal{M}$ consistent with some joint pdf $\bar{p}_t$,

$$\bar{p}_{i,t}(\mathcal{X}_i) = \int_{\mathcal{X}\backslash\mathcal{X}_i} \bar{p}_t(\mathcal{X}), \forall i \in \mathcal{V}. \tag{24}$$

Since we do not have an explicit form for the pdf $\bar{p}_t$, we study its properties in a specific case, where the pdf is independent w.r.t. the variables in $\mathcal{X}$.

6.4 Marginal Consensus with Independent Variables

We begin by recalling the mixing properties established for the distributed setting in Propositions -. We list the desired properties for $\bar{p}_t$ in the following conjecture and prove them for a special case with independence over the variables in $\mathcal{X}$.

Conjecture 1. For $v_{i,t}$ defined in (23) and arbitrary joint pdf $\bar{p}_t$, $\|v_{i,t} - \bar{p}_{i,t}\|_{TV} \leq \sigma(A)\|p_{i,t} - \bar{p}_{i,t}\|_{TV}$ for $\sigma(A) \in (0,1)$ and $\|\bar{p}_t - \bar{p}_{t+1}\|_{TV} \leq (c-1)\alpha_t L/2$ for some $c > 1$.

We consider the following special case where the estimated probabilities $p_{i,t}$ are independent w.r.t. each variable $\boldsymbol{x} \in \mathcal{X}_i$, the set of variables estimated by agent i as,

$$p_{i,t}(\mathcal{X}_i) = \prod_{\boldsymbol{x}\in\mathcal{X}_i} p_{i,t}(\boldsymbol{x}). \tag{25}$$

Since Assumption 6 assigns a connected subgraph $\mathcal{G}(\boldsymbol{x})$ to any variable $\boldsymbol{x}$, the resulting mixed pdf is expressed in terms of independent pdf components at $\boldsymbol{x}$ as,

$$v_{i,t}(\boldsymbol{x}) \propto \prod_{j\in\mathcal{V}\backslash\mathcal{V}(\boldsymbol{x})} p_{i,t}(\boldsymbol{x})^{A_{ij}} \prod_{j\in\mathcal{V}(\boldsymbol{x})} p_{j,t}(\boldsymbol{x})^{A_{ij}}.$$

Next, we will use this form to show that computing an independent component of agent estimates $p_{i,t+1}(\boldsymbol{x})$ involves multiplying the mixed pdf component with a bounded likelihood similar to the Assumption 3.

lemma]theorem Assuming that the mixed pdfs $v_{i,t}$ are independent w.r.t. variable $\boldsymbol{x} \in \mathcal{X}_i$, we can represent agent i's update w.r.t. any variable at time t as,

$$p_{i,t+1}(\boldsymbol{x}) \propto \mathrm{q}_i(z_{i,t}|\boldsymbol{x})^{\alpha_t} v_{i,t}(\boldsymbol{x}),$$

with the agent-variable likelihood,

$$\mathrm{q}_{i,t}(z_{i,t}|\boldsymbol{x})^{\alpha_t} = \int \mathrm{q}_i(z_{i,t}|\mathcal{X}_i)^{\alpha_t} \prod_{\boldsymbol{y}\in\mathcal{X}_i\backslash\boldsymbol{x}} v_{i,t}(\boldsymbol{y}) d\mathcal{X}_i\backslash\boldsymbol{x}$$

satisfying $\mathrm{q}_{i,t}(z_{i,t}|\boldsymbol{x})^{\alpha_t} \in [e^{-\alpha_t L}, e^{\alpha_t L}]$.

Since the estimates $p_{i,t}$ converge to consensus manifold $\mathcal{M}$, we now prove a geometric convergence bound for the independent form of $\bar{p}_t$ specified as follows,

$$\bar{p}_t(\mathcal{X}) = \prod_{\boldsymbol{x}\in\mathcal{X}} \bar{p}_t(\boldsymbol{x}), \bar{p}_t(\boldsymbol{x}) \propto \prod_{j\in\mathcal{V}(\boldsymbol{x})} p_{j,t}(\boldsymbol{x})^{\frac{1}{|\mathcal{V}(\boldsymbol{x})|}}.$$

lemma]theorem For $v_{i,t}$ defined in (23) with connectivity requirements in Assumption 2, additional variable independence assumption, and geometric average $\bar{p}_t$ in (25), we have the TV distance $\|v_{i,t}(\boldsymbol{x}) - \bar{p}_t(\boldsymbol{x})\|_{TV} \leq \sigma\|p_{i,t}(\boldsymbol{x}) - \bar{p}_t(\boldsymbol{x})\|_{TV}$ with $\sigma < 1$ and $\|\bar{p}_t - \bar{p}_{t+1}\|_{TV} \leq (c-1)\alpha_t L/2$ with $c = 1 + 2m$.

6.5 Almost Sure Convergence of DMSMD

Using the upper bounds computed for independent densities, we guarantee almost-sure convergence of the iterates to the marginal pdfs. The presentation here borrows from the distributed SMD algorithm analysis, with the following propositions establishing bounded iterate gaps similar to Section 5.4 and the final two theorems proving almost sure convergence as Section 5.5.

As discussed in Section 5.4, summability of positive upper bounds on the iterate gaps implies their asymptotic convergence to zero. To this end, the next proposition upper bounds the TV distance between estimates across the likelihood update.

proposition]theorem The pdf $p_{i,t+1}$ minimizing $J_{i,t}[p, v_{i,t}]$ defined in (23) satisfies, $\|v_{i,t} - p_{i,t+1}\|_{TV} \leq \alpha_t L/2$.

For the following analysis, we consider the marginals of the optimal pdf $p^\star(\mathcal{X})$ defined as,

$$p_i^\star(\mathcal{X}_i) = \int_{\mathcal{X}_i \backslash \mathcal{X}_i} p^\star(\mathcal{X}). \tag{26}$$

We next produce an upper bound similar to Proposition , but for the gap between the objective function evaluated at mixed estimate to true marginal.

proposition]theorem The term $\sum_{i=1}^n (f_i[p_i^\star] - f_i[v_{i,t}])$ is upper bounded by the distances $\sigma(A)\sum_{i=1}^n L\|\bar{p}_{i,t} - p_{i,t}\|_{TV}$.

Now, we show summability of the upper bound in Proposition containing the TV distance between marginal average $\bar{p}_{i,t}$ to the agent estimate $p_{i,t}$. With square summable α_t [52], this implies asymptotic convergence of the two pdfs.

proposition]theorem With Proposition and Conjecture 1, the sequence with terms $a_t = \sigma(A)\alpha_t L \sum_{i=1}^n \|\bar{p}_{i,t} - p_{i,t}\|_{TV}$ is summable.

proposition]theorem Assuming Conjecture 1 holds, the sequence $\alpha_t L\|\bar{p}_t - \bar{p}_t(\mathcal{X}|\mathcal{X}_i)v_{i,t}\|_{TV}$ is summable for any $i \in \mathcal{V}$.

Since the estimated pdfs are defined over distinct spaces, we define a neighborhood-based divergence metric relating marginal densities at any agent to the complete pdf.

Definition 10. Define the ϵ-neighborhood of a marginal $p_i^\star$ of $p^\star \in \mathcal{F}^\star$ as:

$$\mathbb{B}_i(\mathcal{F}^\star, \epsilon) = \left\{ p_i \in \mathcal{F}_{\mathfrak{d}_i} | \min_{p^\star \in \mathcal{F}^\star} \mathrm{KL}[p_i^\star, p_i] \leq \epsilon, p_i^\star = \int_{\mathcal{X}\backslash\mathcal{X}_i} p^\star \right\}.$$

As seen in prior sections, we employ the preliminary results to prove the convergence of the DMSMD algorithm with the next two theorems. The first theorem shows almost sure convergence of the KL-divergence between the estimated and marginals of the true pdf to a finite positive value, and the next one proves that the finite value is arbitrarily close to zero.

Theorem 6. *Under Assumptions 1-6 and Conjecture 1, the divergence functional $\sum_{i=1}^n \mathrm{KL}[p_i^\star, v_{i,t}]$ of pdf sequences $\{v_{i,t}\}_{i\in\mathcal{V}}$ generated by applying the distributed SMD algorithm in (23) almost surely converges to some finite value.*

Theorem 7. *Under Assumptions 1-6 and Conjecture 1, the marginal pdfs $v_{i,t}$ generated by the distributed marginal algorithm in (23) for any agent $i \in \mathcal{V}$ converge almost surely to the partial neighborhood $\mathcal{B}_i(\mathcal{F}^\star, \epsilon)$ around optimal set $\mathcal{F}^\star$ for any $\epsilon > 0$.*

7 Distributed Marginal Gaussian Variational Inference

In this section, we specialize the distributed algorithms in Sections 5 and 6 for Gaussian estimates. At each agent, implementing the proposed algorithms is a two-step process: mixing the neighbor priors, and updating the likelihood.

Marginal mixing requires computing the Gaussian conditionals and marginals, and their product and geometric averages. Algorithm 2 computes this mixed Gaussian pdf $v_{i,t}(\mathcal{X}_i)$ using the derivations in our prior work [36]. This algorithm trivially holds for the standard distributed setting with conditional-marginal product equal to the neighbor estimate, i.e. $\tilde{p}_{ji,t} = p_{j,t}$. Here, we represent a Gaussian random variable with mean μ and information matrix Ω as $\mathcal{N}(\mu, \Omega^{-1})$, and its density function as $\phi(\cdot|\mu, \Omega^{-1})$.

```
Inputs: estimate $p_{i,t} = \phi(\mathcal{X}_i|\mu, \Omega^{-1})$, weights
    $\{A_{ij}\}_{j\in\mathcal{V}_i}$, neighbor estimates $p_{j,t}(\mathcal{X}_j)$
// Receive marginals from neighbors.
for $j \in \mathcal{V}_i$ do
    Compute marginal $p_{ji,t}$ using [36, Lemma 1] over
      $\mathcal{V}_{ij}$
// Combine neighbor estimates.
for $j \in \mathcal{V}_i$ do
    Use [36, Lemma 2] to compute conditional pdf
      $p_{i,t}(X_1|X_2)$ with separate variables $X_1 = \mathcal{X}_i \backslash \mathcal{X}_{ij}$
      and shared variables $X_2 = \mathcal{X}_{ij}$
    Compute $\tilde{p}_{ji,t}(\mathcal{X}_i)$ by multiplying $i$'s conditional
      with marginal $p_{ji,t}$ using [36, Proposition 3]
Compute mixed pdf $v_{i,t}(\mathcal{X}_i)$ using [36, Lemma 3] over
  $\tilde{p}_{ji,t}(\mathcal{X}_i)$
```

Algorithm 2: Marginal density mixing at agent i

Next, we express an analytic form of the likelihood update step in Algorithm 1 assuming that the prior mixed pdf $v_{i,t}$ and posterior $p_{i,t+1}$ are Gaussian. The analytic updates associated with the linear log-likelihood setting was presented in [36] is given as,

lemma]theorem Let the likelihood density be $q_i(z_{i,t}|\mathcal{X}_i) = \phi(z_{i,t}|H_i\mathcal{X}_i, V_i)$. Then, the posterior obtained as the product of the likelihood and prior $\phi(z_{i,t}|H_i\mathcal{X}_i, V_i)\phi(\mathcal{X}_i; \mu, \Omega_i^{-1})$ is a Gaussian distribution:

$$\mathcal{N}\left((H_i^T V_i H_i + \Omega_i)^{-1}(H_i^T V_i z_{i,t} + \Omega_i \mu_i), (H_i^T V_i H_i + \Omega_i)^{-1}\right).$$

For the non-linear log-likelihood $\mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)$ that does not yield an analytic update, one can approximate the likelihood update using distributed Gaussian variational inference [53] on the mixed pdf $p_{i,t}^v = \phi(\cdot|\mu_{i,t}^v, \Omega_{i,t}^v)$ as,

$$\Omega_{i,t+1} = \Omega_{i,t}^v - \mathbb{E}_{p_{i,t}^v}[\nabla_{\mathcal{X}_i}^2 \log \mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)],$$
$$\mu_{i,t+1} = \mu_{i,t}^v + (\Omega_{i,t}^v)^{-1}\mathbb{E}_{p_{i,t}^v}[\nabla_{\mathcal{X}_i} \log \mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)].$$

In the partial distributed mapping example explained later, we implement this algorithm to estimate Gaussians with diagonal covariance matrices. Therefore, we present a modified mixing step for the marginal distributed estimation algorithm in the following lemma.

lemma]theorem Assume that agent i receives observation $z_{i,t+1}$ with likelihood $\mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)$ and neighbor estimates $p_{j,t}(\mathcal{X}_j) = \mathcal{N}(\mathcal{X}_j|\mu_{j,t}, \Omega_{j,t}^{-1})$ at time t. Upon weighing neighbor opinions with elements of matrix A, the mean $\mu_{i,t+1}$ and information matrix $\Omega_{i,t+1}$ of the pdf $p_{i,t+1}$ is,

$$\tilde{\Omega}_{ji,t} = R_{ij}\Omega_{j,t} + S_{ij}\Omega_{i,t}, \tilde{\mu}_{ji,t} = R_{ij}\mu_{j,t} + S_{ij}\mu_{i,t} \tag{27}$$
$$\Omega_{i,t}^v = \sum_{j\in\mathcal{V}} A_{ij}\tilde{\Omega}_{ji,t}, \Omega_{i,t}^v\mu_{i,t}^v = \sum_{j\in\mathcal{V}} A_{ij}\tilde{\Omega}_{ji,t}\tilde{\mu}_{ji,t}$$
$$\Omega_{i,t+1} = \Omega_{i,t}^v - \mathbb{E}_{v_{i,t}}[\nabla_{\mathcal{X}}^2 \log \mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)],$$
$$\mu_{i,t+1} = \mu_{i,t}^v + (\Omega_{i,t}^v)^{-1}\mathbb{E}_{v_{i,t}}[\nabla_{\mathcal{X}} \log \mathrm{q}_i(z_{i,t+1}|\mathcal{X}_i)],$$

where mixed pdf $v_{i,t} = \phi(\mathcal{X}_i|\mu_{i,t}^v, \Omega_{i,t}^v)$, and matrices $R_{ij} \in \{0,1\}^{\mathfrak{d}_i \times \mathfrak{d}_j}$ and $S_{ij} \in \{0,1\}^{\mathfrak{d}_i \times \mathfrak{d}_i}$. Here, $R_{ij}[s_i, s_j] = 1$ where s_i, s_j are indices in agents i, j corresponding to a common variable. The matrix S_{ij} is a diagonal matrix with 1 at variable index distinct from agent j.

Proof. The updates on marginals and distributed consensus follow from prior discussion. The matrices S, R

match the indices between the agents and hypotheses to compute the diagonal information matrices. ■

Distributed Relative Localization: An Example

We consider a network of $n = 8$ agents aiming to estimate their positions $\boldsymbol{x}_i \in \mathbb{R}^2$ using noisy relative position measurements. To ensure a unique solution, we assume the presence of an anchor agent with known position at $(0,0)$. Each agent i observes the relative position of its neighbor j sampled as $\boldsymbol{z}_{ij} \sim \mathcal{N}(\boldsymbol{x}_i - \boldsymbol{x}_j, \Omega_{ij}^{-1})$. The relevant set of variables at agent i is thus given by $\mathcal{X}_i = \{\boldsymbol{x}_j\}_{j \in \mathcal{V}_i}$. The combined observation model at agent i for the observations relative to its neighbors $\boldsymbol{z}_i = \{\boldsymbol{z}_{ij}\}_{j \in \mathcal{V}_i}$ is,

$$\mathrm{q}_i(z_i|\mathcal{X}_i) = \prod_{j \in \mathcal{V}_i} \mathrm{q}_i(z_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{28}$$

The doubly stochastic matrix A represents agent communication as described in Assumption 2. We first mention the application of our distributed and marginal estimation algorithms, followed by standard and circular BP algorithms.

In the distributed setting, each agent i maintains a Gaussian distribution $\mathcal{N}(\mu_{i,t}, \Omega_{i,t}^{-1})$ with pdf $p_{i,t}(\mathcal{X})$ at time step t over all unknown variables $\mathcal{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top$. The corresponding observation model in (28) is expressed in terms of the variable $\mathcal{X}$ as $\mathrm{q}_i(z_{i,t}|\mathcal{X}) = \mathcal{N}(z_{i,t}|H_i^{(d)}\mathcal{X}, V_i^{(d)})$ where $H_i^{(d)} \in \mathbb{R}^{d \times nd}$. Each step in the distributed SMD algorithm in (18) at agent i uses data likelihood $\mathrm{q}_i(z_{i,t}|\mathcal{X})$, and neighbor pdfs $p_{j,t}(\mathcal{X})$ and weights A_{ij} for neighbors $j \in \mathcal{V}_i$, to obtain the mixed pdf $v_{i,t}(\mathcal{X})$ as:

$$\mathcal{N}((\Omega_{i,t+1}^g)^{-1}(\sum_{j \in \mathcal{V}_i} A_{ij}\Omega_{j,t}\mu_{j,t}), (\Omega_{i,t+1}^g)^{-1}),$$

where $\Omega_{i,t+1}^g = \sum_{j \in \mathcal{V}_i} A_{ij}\Omega_{j,t}$. This is followed by the Gaussian likelihood update in Lemma using the mixed pdf $v_{i,t}(\mathcal{X})$ and the Gaussian likelihood $\mathcal{N}(z_{i,t}|H_i^{(d)}\mathcal{X}, V_i^{(d)})$.

Next, we consider the marginal estimation setting, where each agent i estimates a pdf over the set of relevant variables $\mathcal{X}_i$, given by the vectorized version of $\{\boldsymbol{x}_j\}_{j \in \mathcal{V}_i}$. For this setting, we express the observation model given in (28) as $\mathrm{q}_i(z_{i,t}|\mathcal{X}_i) = \mathcal{N}(H_i^{(m)}\mathcal{X}_i, V_i^{(m)})$. We implement the Gaussian version of the marginal estimation using the mixed pdf update in Algorithm 2 followed by the likelihood update defined via the update in Lemma .

Next, we will describe the BP algorithm and a recent circular BP version [38], with further details in [27]. The BP algorithm allows the network to estimate a density of the form $\prod_{i \in \mathcal{V}} p_{i,t}(\boldsymbol{x}_i)$, such that agent i estimates the pdf $p_{i,t}(\boldsymbol{x}_i)$. In an undirected network, each agent i generates a message $m_{ij,t}(\boldsymbol{x}_j)$ for its neighbor j at time t, and vice-versa. Then, agent i merges the neighbor messages to form its own belief, and computes their marginal to generate the next set of messages as follows,

$$\begin{aligned} m_{ij,t+1}(\boldsymbol{x}_j) &= \int_{\boldsymbol{x}_i} \mathrm{q}_i(\boldsymbol{z}_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)p_{i,t}(\boldsymbol{x}_i) \prod_{k \in \mathcal{V}_i \setminus \{j\}} m_{ki,t}(\boldsymbol{x}_i) \\ p_{i,t+1}(\boldsymbol{x}_i) &\propto p_{i,t}(\boldsymbol{x}_i) \prod_{k \in \mathcal{V}_i} m_{ki,t}(\boldsymbol{x}_i) \end{aligned} \tag{29}$$

A recent version named circular BP [38] relies on scaling the message $m_{ji,t-1}(\boldsymbol{x}_j)$ with a symmetric pair-specific coefficients dependent on (j,i):

$$\begin{aligned} m_{ij,t+1}(x_j) \propto \int_{x_i} &\mathrm{q}_i(\boldsymbol{z}_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)^{\beta_{ij}} \\ &\Big(p_{i,t}(x_i)^{\gamma_i} m_{ji,t}(x_i)^{1-\frac{\alpha_{ij}}{\kappa_i}} \prod_{k \in \mathcal{V}_i \setminus \{j\}} m_{ki,t}(x_i)\Big)^{\kappa_i}. \end{aligned} \tag{30}$$

With $\alpha_{ij} = \beta_{ij} = \kappa_i = \gamma_i = 1$, this algorithm reduces to the standard BP. There exists a sufficiently small $\alpha_{ij} = \alpha_{ji} = \alpha \in (0,1)$ and the rest of the terms equal to one satisfying the convergence criterion in [38, Theorem 5.2], and further details in [27]. The theoretical fixed-point analysis in this work, however, remains limited to estimating binary probabilities. The Gaussian version of the update rule is derived in the following lemma.

lemma]theoremGiven data $\boldsymbol{z}_{ij}$ sampled by agent i from the likelihood $\phi(\boldsymbol{z}_{ij}|\boldsymbol{x}_j - \boldsymbol{x}_i, \Omega_{ij}^{-1})$, prior self and neighbor messages $\phi(\boldsymbol{x}_i; \mu_{ji,t}^{(m)}, (\Omega_{ji,t}^{(m)})^{-1})$ for $j \in \mathcal{V}_i$, the circular BP message with $\alpha_{ij} = \alpha \in (0,1)$ and $\beta_{ij} = \gamma_i = \kappa_i = 1$ to agent j is,

$$\begin{aligned} \Omega_{ij,t+1}^{(m)} &= \Omega_{ij} - \Omega_{ij}(\Omega_{ij,t}^g + \Omega_{ij})^{-1}\Omega_{ij} \\ \mu_{ij,t+1}^{(m)} &= \boldsymbol{z}_{ij} + (\Omega_{ij,t+1}^{(m)})^{-1}\Omega_{ij}(\Omega_{ij,t}^g + \Omega_{ij})^{-1}\Omega_{ij,t}^g\mu_{ij,t}^g \end{aligned}$$

where the information matrix is $\Omega_{ij,t+1}^g = \Omega_{i,t} + (1-\alpha)\Omega_{ji,t}^{(m)} + \sum_{k \in \mathcal{V}_i \setminus \{j\}} \Omega_{ki,t}^{(m)}$ and the mean is $\mu_{ij,t+1}^g = (\Omega_{ij,t+1}^g)^{-1}(\Omega_{i,t}\mu_{i,t} + (1-\alpha)\Omega_{ji,t}^{(m)}\mu_{ji,t}^{(m)} + \sum_{k \in \mathcal{V}_i \setminus \{j\}} \Omega_{ki,t}^{(m)}\mu_{ki,t}^{(m)})$.

Proof. We start by noting that for $\alpha_{ij} = \alpha$, the product of the densities $\Big(p_{i,t}(x_i)m_{ji,t}(x_i)^{1-\alpha}\prod_{k \in \mathcal{V}_i \setminus \{j\}} m_{ki,t}(x_i)\Big)$ is given by the Gaussian with parameters $p_{ij,t}^g(\boldsymbol{x}_i) = \phi(\mu_{ij,t+1}^g, \Omega_{ij,t+1}^g)$. Next, we define $\bar{\boldsymbol{x}}_j = \boldsymbol{x}_j - \boldsymbol{z}_{ij}$ and start with expressing the integral coefficient in terms of $\boldsymbol{x}_i$ as,

$$\begin{aligned} &\int \mathrm{q}_i(\boldsymbol{z}_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)p_{ij,t}^g(\boldsymbol{x}_i)d\boldsymbol{x}_i \\ &\propto \int \exp\Big(-\frac{1}{2}[\boldsymbol{x}_i^\top(\Omega_{ij,t}^g + \Omega_{ij})\boldsymbol{x}_i - 2\boldsymbol{x}_i^\top(\Omega_{ij,t}^g\mu_{ij,t}^g + \Omega_{ij}\bar{\boldsymbol{x}}_j) \\ &\quad +(\mu_{ij,t}^g)^\top\Omega_{ij,t}^g\mu_{ij,t}^g + \bar{\boldsymbol{x}}_j^\top\Omega_{ij}\bar{\boldsymbol{x}}_j]\Big)d\boldsymbol{x}_i. \end{aligned}$$

Next, we recall from [54, Fact 14.12.1] $\int \exp(-\frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{c}^\top\boldsymbol{x} + a) = \sqrt{2\pi A^{-1}}\exp\left[\frac{1}{2}\boldsymbol{c}^\top A^{-1}\boldsymbol{c} + a\right]$ for a symmetric matrix $A \in \mathbb{R}^{d \times d}$, $\boldsymbol{c} \in \mathbb{R}^d, a \in \mathbb{R}$. We can compute the mean and information matrix of the marginal by setting $A = \Omega_{ij,t}^g + \Omega_{ij}$, $\boldsymbol{c} = \Omega_{ij,t}^g\mu_{ij,t}^g + \Omega_{ij}\bar{\boldsymbol{x}}_j$ and $a = (\mu_{ij,t}^g)^\top\Omega_{ij,t}^g\mu_{ij,t}^g + \bar{\boldsymbol{x}}_j^\top\Omega_{ij}\bar{\boldsymbol{x}}_j$. The terms containing $\bar{\boldsymbol{x}}_j$ in $\boldsymbol{c}^\top A^{-1}\boldsymbol{c} + a$ are,

$$\begin{aligned} -\bar{\boldsymbol{x}}_j^\top(\Omega_{ij} - \Omega_{ij}(\Omega_{ij,t}^g + \Omega_{ij})^{-1}\Omega_{ij})\bar{\boldsymbol{x}}_j \\ + 2\bar{\boldsymbol{x}}_j^\top\Omega_{ij}(\Omega_{ij,t}^g + \Omega_{ij})^{-1}\Omega_{ij,t}^g\mu_{ij,t}^g, \end{aligned}$$

which yields the final result. ■

We compared the distributed, marginal, BP, and circular BP algorithms in estimating the agent positions in an 8-agent network. Each agent collects data from the model with $\Omega_{ij} = \mathbb{I}_2$ and initializes their mean $\mu_{i,0}$ at $(0,0)$. The evolution of position means $\mu_{i,t}$ and their error with respect to the true positions $\boldsymbol{x}_i$ are shown in Fig. 1. The BP

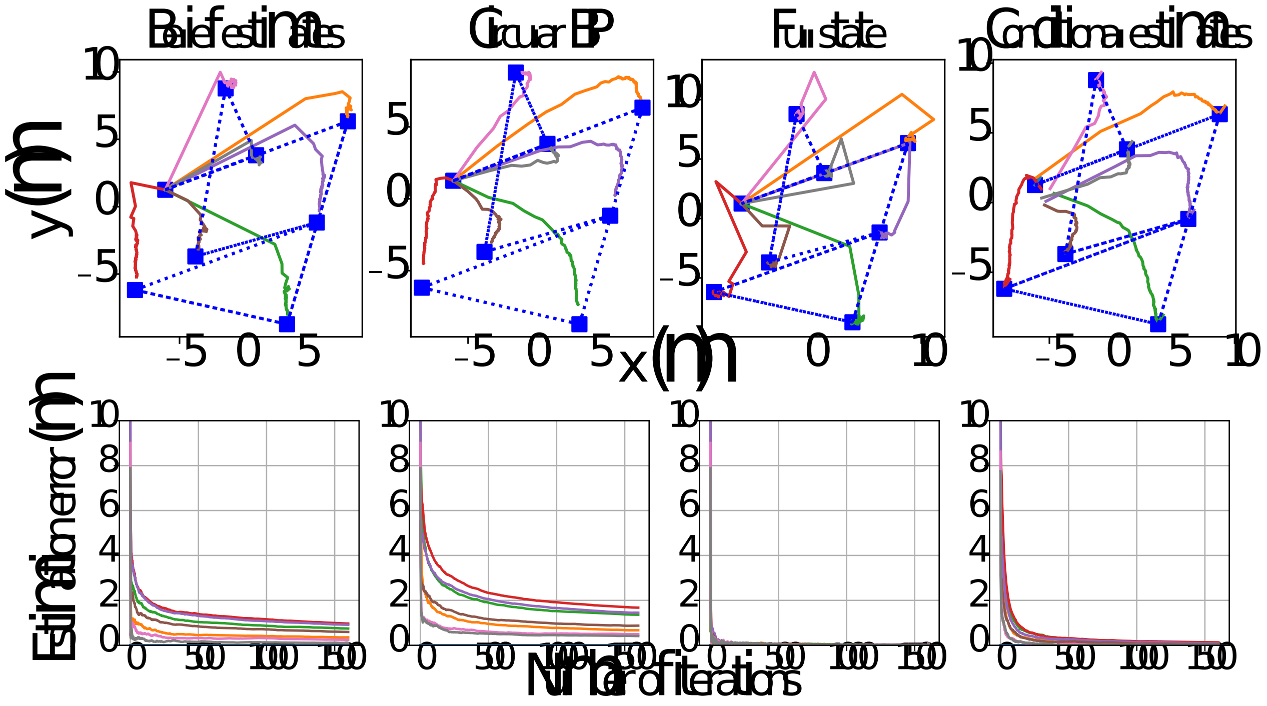

Figure 1. Trajectories of estimated node positions $\mu_{i,t}$ in an 8 agent ring network with true positions shown as blue squares (top). Estimation error $\|\mu_{i,t} - \boldsymbol{x}_i\|$ over 1600 time steps (bottom).

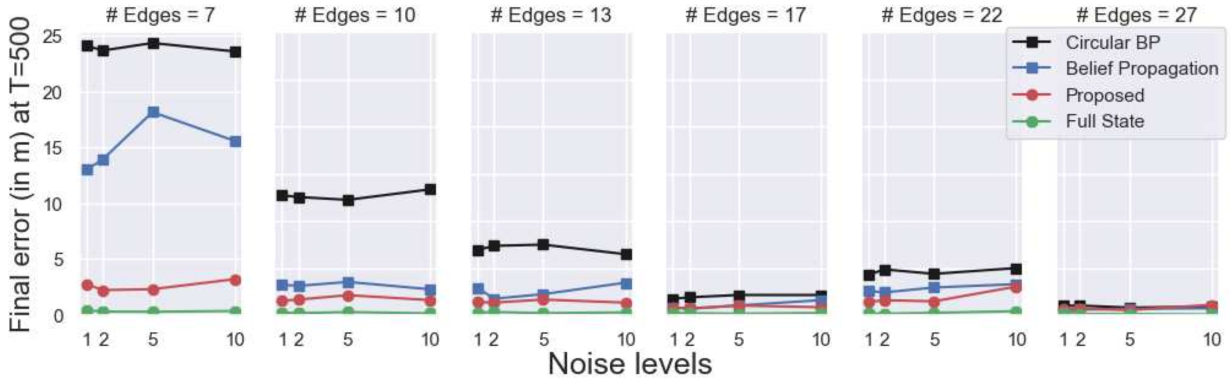

Figure 2. Plots of the 500-step average localization error, given by $1/n \sum_{i\in\mathcal{V}} \|\mu_{i,t} - \boldsymbol{x}_i\|$, using belief propagation, circular belief propagation, the proposed marginal estimation, and full state estimation algorithms in an 8 node network. The comparisons span measurement noise variances $\Sigma_{ij} = b\mathbb{I}_2$ for $b \in \{1, 2, 5, 10\}$ and network connectivities ranging from a line graph with 7-edges to a 27-edge fully connected one.

algorithms converge slower than the proposed distributed and marginal SMD algorithms.

Fig. 2 compares the performance of various algorithms as the noise levels and graph connectivity vary. The chosen performance metric is the estimation error of each algorithm at time step $T = 500$, after all algorithms have converged. Each of the six subplots represents a different graph with 8 nodes, ranging from a line graph (7 edges, leftmost subplot) to a fully connected graph (27 edges, rightmost subplot). In each subplot, estimation error (y axis) is plotted for algorithms implemented using noisy data sampled with information matrix value (x axis) , $b\mathbb{I}_2$, with magnitudes $b = 1, 2, 5, 10$. We present the circular BP algorithm results with $\alpha_{ij} = 0.8$ for all $i, j \in \mathcal{V}$.

From the plots, we note that the best performing algorithm across the board is the full state estimation algorithm, showing negligible error for all graphs and error levels. This is ascribed to the tracking and sharing of individual agent probabilities defined over all unknown variables. Taking this as a baseline, we can observe that the proposed algorithm follows closely to this, and provides lower error values over sparser graphs (3 left subplots) than other algorithms for all noise levels. The error of the proposed algorithm increases as the graph becomes more dense and the noise increases (values for $b = 10$ on the 3 right subplots.) In this case the performance of the belief propagation algorithm surpasses the proposed algorithm's; however, this performance difference is small and comparable.

Further, we see that circular BP is the least accurate on sparse graphs as we increase observation noise magnitude, owing to insufficient countering of the loop effects in circular BP algorithm. In denser graphs, the errors remain too close to compare.

Distributed Mapping: An Example

In this section, we apply the marginal estimation algorithm to distributed mapping. Please see [36] for a simpler example solving relative localization problem with linear observation model, where both the agent observation models and their estimates depend on self and neighbor states. In this multi-robot setting, each robot follows their own trajectory allowing them to gather data describing a portion of the map. Here, the challenge arises from the ability to achieve consensus over common areas by sharing partial information relevant to another robot's map. With the knowledge of observation models describing gathered data, the agents thus share a subset of the model parameters to collectively create a map of the entire space. Here, we use LiDAR post-processed distance data to obstacles for generating points in the free and occupied spaces.

Consider $n = 7$ robots collecting data of the form $z = (x, y)$ where x is a point in the observed space and y is a binary variable indicating free or occupied status. The point x can be embedded into the feature space using kernel functions $k_s(x) = \gamma_1 \exp(-\gamma_2 \|x - x^{(s)}\|^2)$ centered at $x^{(s)}$ and rescaled with parameters $\gamma_1, \gamma_2 > 0$ chosen to suit the domain and regularity of the model. In the partial distributed setting, this vector embedding at agent i is $\Phi_i(x) = [1, k_{i_1}(x), \ldots, k_{i_f}(x)] \in \mathbb{R}^{m_i+1}$. Since some of the kernel functions are shared with neighboring agents, the number of kernels is $m < \sum_i m_i$. The modeled likelihood of an observation $z = (x, y)$ at agent i with input $x \in \mathbb{R}^{\ell_i - 1}$, feature $\Phi_i(x)$, and label $y \in \{0, 1\}$ is,

$$\mathrm{q}(z|\mathcal{X}_i) = \sigma(\Phi_i(x)^\top \mathcal{X}_i)^y (1 - \sigma(\Phi_i(x)^\top \mathcal{X}_i))^{1-y}, \tag{31}$$

where $\mathcal{X}_i$ are the agent relevant weights and σ is the sigmoid function. The consensus constraint enforces equality of the weights assigned to common kernel functions in the agent models. To understand the role of any element i_ℓ in parameter $\mathcal{X}_i$ for constructing a map, note that its positivity emphasizes the confidence in occupancy prediction around feature point x^{i_ℓ} and vice-versa.

In a marginal distributed setting, agent i models the spatial occupancy in terms of kernels centered at relevant feature points $\left\{x^{(s)}\right\}_{s=i_1}^{i_f}$ out of a fixed set of 1000 such points across the entire map. We construct these subsets by selecting feature points whose distance to agent i's trajectory are under a threshold. For a distance threshold of 50-units, the number of parameters observed by the seven agents is $(208, 195, 247, 188, 180, 224, 216)$, thus bringing the number of variables across agents down from 7K to 1458 parameters. Out of the 216 parameters at the last agent, the number of parameters common with others is $(62, 66, 88, 41, 11, 42)$. The agent training datasets at each agent contain 80K-100K points and the verification sets consist of 3K-3.7K points approximately. If any two agent likelihood models contain the same feature point $x^{(s)}$, then they communicate through the network A to consent over common weight parameters.

In Figure 3, we present the robot trajectories for data collection, the training set, and the distinct and shared feature points embedded in the relevant space at two of the robots. For generating the map, we use Lemma in conjunction with [53, Lemma 4] to simplify the expected gradient and Hessian terms. The predictions on the verification set is presented in Figure 4, with maps estimated by individual agents in center figure, with error on agent-specific verification sets on the right of Figure 4.

8 Conclusion

This work designs and analyzes a novel distributed estimation algorithm for estimating marginal densities over relevant variables at each agent in an inference network. The Bayes-like distributed algorithm is designed from a stochastic mirror descent perspective, with almost sure convergence guarantees. Based on our analysis, we claim that any consensus rule with a geometric convergence rate can be coupled to stochastic mirror descent to convergence almost surely to the optimal pdf. This insight has far-reaching implications for developing distributed estimation algorithms in several metric spaces. The distributed mapping implementation demonstrates the vast storage savings due to the proposed algorithm. This algorithm can reduce storage and communication costs in networked estimation problems, based on computation-communication trade-offs.

References

[1] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, pp. 2568–2599, 2019.

[2] S. Kumar, A. Deshpande, S. S. Ho, J. S. Ku, and S. E. Sarma, "Urban street lighting infrastructure monitoring using a mobile sensor platform," *IEEE Sens. J.*, vol. 16, no. 12, pp. 4981–4994, 2016.

[3] F. Bullo, J. Cortés, and S. Martínez, *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Princeton University Press, 2009.

[4] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games. Econ. Behav.*, vol. 76, no. 1, pp. 210–225, Sep. 2012.

[5] R. T. Clemen and R. L. Winkler, "Combining probability distributions from experts in risk analysis," *Risk Anal.*, vol. 19, no. 2, pp. 187–203, 1999.

[6] T. Minka, "Divergence measures and message passing," Tech. Rep., 2005, Microsoft Research MSR-TR-2005-173.

[7] F. Meyer, O. Hlinka, and F. Hlawatsch, "Sigma point belief propagation," *IEEE Trans. Signal Process.*, vol. 21, no. 2, pp. 145–149, 2013.

[8] M. Kayaalp, Y. Inan, E. Telatar, and A. H. Sayed, "On the arithmetic and geometric fusion of beliefs for distributed inference," *IEEE Transactions on Automatic Control*, 2023.

[9] M. E. Khan and H. Rue, "The bayesian learning rule," *J. Mach. Learn. Res.*, vol. 24, no. 281, pp. 1–46, 2023.

[10] A. Garg, T. S. Jayram, S. Vaithyanathan, and H. Zhu, "Generalized opinion pooling," *Ann. Math. Artif. Intell.*, 2004.

[11] G. Koliander, Y. El-Laham, P. M. Djurić, and F. Hlawatsch, "Fusion of probability density functions," *Proceedings of the IEEE*, vol. 110, no. 4, pp. 404–453, 2022.

[12] A. Nemirovski, "Tutorial: Mirror descent algorithms for large-scale deterministic and stochastic convex optimization," in *Conference on Learning Theory*, 2012.

[13] Z. Zhou, P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn, "On the convergence of mirror descent beyond stochastic convex programming," *SIAM J. optim.*, vol. 30, no. 1, pp. 687–716, 2020.

[14] K. R. Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *IEEE Conf. on Decision and Control*. IEEE, 2010, pp. 5050–5055.

[15] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Trans. Autom. Contr.*, vol. 62, no. 11, pp. 5538–5553, 2017.

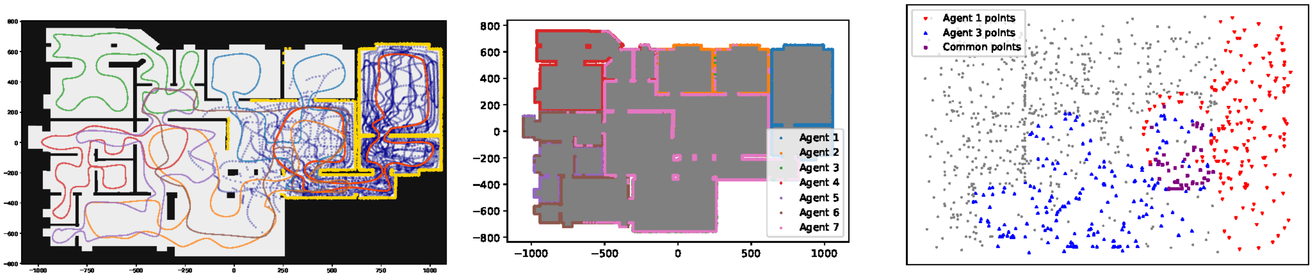

Figure 3. Agent trajectories with training samples collected by agent 1 with free and occupied points in blue and yellow respectively (left). Binary data in all training sets with gray free and labeled occupied points (center). Distinct and shared feature points in the likelihoods of agents 1 and 3 (right).

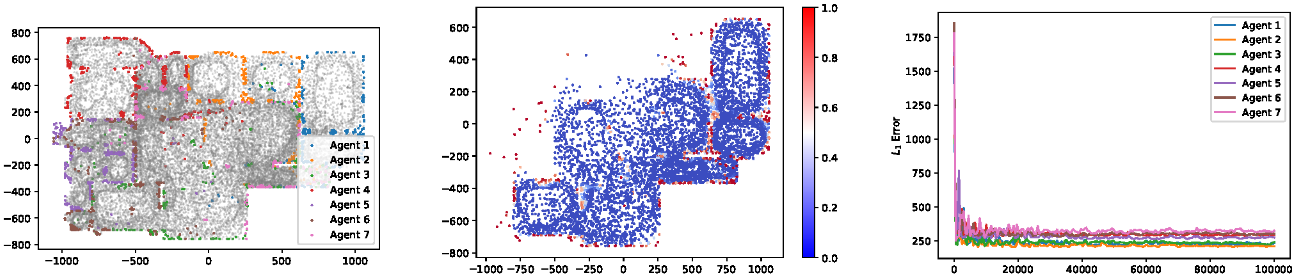

Figure 4. Predicted classes by the agents over the verification set with gray free points and labeled occupied ones (left). Occupancy probability for points in verification sets for agents 1 and 3 (center). L^1 error over the verification set during the 100k training steps.

[16] A. Lalitha, A. Sarwate, and T. Javidi, “Social learning and distributed hypothesis testing,” in *IEEE Int. Symp. on Info. Theory*. IEEE, 2014, pp. 551–555.

[17] T. T. Doan, S. Bose, D. H. Nguyen, and C. L. Beck, “Convergence of the iterates in mirror descent methods,” *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 114–119, 2018.

[18] C. A. Uribe, A. Olshevsky, and A. Nedich, “Non-asymptotic concentration rates in cooperative learning part I: Variational non-Bayesian social learning,” *IEEE Trans. Control. Netw. Syst.*, 2022.

[19] N. Atanasov, R. Tron, V. M. Preciado, and G. J. Pappas, “Joint estimation and localization in sensor networks,” in *IEEE Conf. on Decision and Control*, 2014, pp. 6875–6882.

[20] G. Piovan, I. Shames, B. Fidan, F. Bullo, and B. D. Anderson, “On frame and orientation localization for relative sensing networks,” *Automatica*, vol. 49, no. 1, pp. 206–213, 2013.

[21] J. Pearl, “Reverend Bayes on inference engines: A distributed hierarchical approach,” in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, 1982, pp. 133–136.

[22] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Understanding belief propagation and its generalizations.”

[23] Y.-H. Liu and D. Poulin, “Neural belief-propagation decoders for quantum error-correcting codes,” *Physical review letters*, vol. 122, no. 20, p. 200501, 2019.

[24] K. Desingh, S. Lu, A. Opipari, and O. C. Jenkins, “Efficient nonparametric belief propagation for pose estimation and manipulation of articulated objects,” *Science Robotics*, vol. 4, no. 30, p. eaaw4523, 2019.

[25] T. Heskes, “On the uniqueness of loopy belief propagation fixed points,” *Neural Computation*, vol. 16, no. 11, pp. 2379–2413, 2004.

[26] D. Liu, “Perspectives on probabilistic graphical models,” Ph.D. dissertation, KTH Royal Institute of Technology, 2020.

[27] V. Bouttier, “Circular belief propagation as a model for optimal and suboptimal inference in the brain : extending the algorithm and proposing a neural implementation,” Theses, Université Paris Cité, Dec. 2021. [Online]. Available: https://theses.hal.science/tel-04530051

[28] B. Vantaggi, “Statistical matching of multiple sources: A look through coherence,” *Int. J. Approx. Reason.*, vol. 49, no. 3, pp. 701–711, 2008.

[29] Y. Y. Shkel and A. K. Yadav, “Information spectrum converse for minimum entropy couplings and functional representations,” in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 66–71.

[30] J. Kracík, “Combining marginal probability distributions via minimization of weighted sum of kullback–leibler divergences,” *Int. J. Approx. Reason.*, vol. 52, no. 6, pp. 659–671, 2011.

[31] F. Cicalese, L. Gargano, and U. Vaccaro, “Minimum-entropy couplings and their applications,” *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3436–3451, 2019.

[32] P. Paritosh, N. Atanasov, and S. Martinez, “Hypothesis assignment and partial likelihood averaging for cooperative estimation,” in *IEEE Int. Conf. on Decision and Control*. IEEE, 2019, pp. 7850–7856.

[33] R. Parasnis, M. Franceschetti, and B. Touri, “Non-bayesian social learning on random digraphs with aperiodically varying network connectivity,” *IEEE Transactions on Control of Network Systems*, vol. 9, no. 3, pp. 1202–1214, 2022.

[34] D. Bickson, “Gaussian belief propagation: Theory and application,” *arXiv preprint arXiv:0811.2518*, 2008.

[35] J. Pearl, “Fusion, propagation, and structuring in belief networks,” in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 139–188.

[36] P. Paritosh, N. Atanasov, and S. Martínez, “Marginal density averaging for distributed node localization from local edge measurements,” in *IEEE Int. Conf. on Decision and Control*. IEEE, 2020, pp. 2404–2410.

[37] A. T. Ihler, A. S. Willsky *et al.*, “Loopy belief propagation: Convergence and effects of message errors,” *J. Mach. Learn. Res.*, vol. 6, no. May, pp. 905–936, 2005.

[38] V. Bouttier, R. Jardri, and S. Deneve, “Circular belief propagation for approximate probabilistic inference,” *arXiv preprint arXiv:2403.12106*, 2024.

[39] R. Sinkhorn and P. Knopp, “Concerning nonnegative matrices and doubly stochastic matrices,” *Pac. J. Math.*, vol. 21, no. 2, pp. 343–348, 1967.

[40] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

[41] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, 2015.

[42] A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[43] B. A. Frigyik, S. Srivastava, and M. R. Gupta, “Functional Bregman

divergence and Bayesian estimation of distributions," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5130–5139, 2008.

[44] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press, 2011.

[45] W. Cheney, *Analysis for applied mathematics*. Springer Science & Business Media, 2001, vol. 208.

[46] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[47] M. Pinsker, "Information and information stability of random variables and processes (in russian)," 1960.

[48] B. T. Polyak, "Introduction to optimization," *Inc., Publications Division, New York*, vol. 1, p. 49, 1987.

[49] P. Hall and C. C. Heyde, *Martingale limit theory and its application*. Academic press, 2014.

[50] P. Paritosh, N. Atanasov, and S. Martinez, "Distributed Bayesian estimation of continuous variables over time-varying directed networks," *IEEE Control Syst. Lett.*, vol. 6, pp. 2545–2550, 2022.

[51] S. Bandyopadhyay and S.-J. Chung, "Distributed estimation using Bayesian consensus filtering," in *2014 American control conference*. IEEE, 2014, pp. 634–641.

[52] B. Franci and S. Grammatico, "Convergence of sequences: A survey," *Annu. Rev. Control*, vol. 53, pp. 161–186, 2022.

[53] P. Paritosh, N. Atanasov, and S. Martinez, "Distributed variational inference for online supervised learning," *arXiv preprint arXiv:2309.02606*, 2023.

[54] D. S. Bernstein, *Scalar, Vector, and Matrix Mathematics: Theory, Facts, and Formulas - Revised and Expanded Edition*. Princeton University Press, 2018.

Parth Paritosh is currently a Postdoctoral Fellow with Research Associateship Program at the U.S. Army Combat Capabilities DEVCOM Army Research Laboratory. He earned his Ph.D. from the Mechanical and Aerospace Engineering Department at the University of California, San Diego (UCSD). He obtained his M.S. in Mechanical Engineering from Purdue University in May 2017 and his B.Tech. in Mechanical Engineering with a minor in Computer Science and Engineering in May 2015. His research focuses on enhancing robotic localization and inference capabilities, particularly in multi-agent autonomous systems.

Nikolay Atanasov (S'07-M'16-SM'23) is an Assistant Professor of Electrical and Computer Engineering at the University of California San Diego, La Jolla, CA, USA. He obtained a B.S. degree in Electrical Engineering from Trinity College, Hartford, CT, USA in 2008 and M.S. and Ph.D. degrees in Electrical and Systems Engineering from the University of Pennsylvania, Philadelphia, PA, USA in 2012 and 2015, respectively. Dr. Atanasov's research focuses on robotics, control theory, and machine learning, applied to active perception problems for autonomous mobile robots. He works on probabilistic models that unify geometric and semantic information in simultaneous localization and mapping (SLAM) and on optimal control and reinforcement learning algorithms for minimizing probabilistic model uncertainty. Dr. Atanasov's work has been recognized by the Joseph and Rosaline Wolf award for the best Ph.D. dissertation in Electrical and Systems Engineering at the University of Pennsylvania in 2015, the Best Conference Paper Award at the IEEE International Conference on Robotics and Automation (ICRA) in 2017, the NSF CAREER Award in 2021, and the IEEE RAS Early Academic Career Award in Robotics and Automation in 2023.

Sonia Martínez (M'02-SM'07-F'18) is a Professor of Mechanical and Aerospace Engineering at the University of California, San Diego, CA, USA. She received her Ph.D. degree in Engineering Mathematics from the Universidad Carlos III de Madrid, Spain, in May 2002. She was a Visiting Assistant Professor of Applied Mathematics at the Technical University of Catalonia, Spain (2002-2003), a Postdoctoral Fulbright Fellow at the Coordinated Science Laboratory of the University of Illinois, Urbana-Champaign (2003-2004) and the Center for Control, Dynamical systems and Computation of the University of California, Santa Barbara (2004-2005). Her research interests include the control of networked systems, multi-agent systems, nonlinear control theory, and planning algorithms in robotics. She is a Fellow of IEEE. She is a co-author (together with F. Bullo and J. Cortés) of "Distributed Control of Robotic Networks" (Princeton University Press, 2009). She is a co-author (together with M. Zhu) of "Distributed Optimization-based Control of Multi-agent Networks in Complex Environments" (Springer, 2015). She is the Editor in Chief of the recently launched *CSS IEEE Open Journal of Control Systems.*