Distributed Variational Inference for Online Supervised Learning

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

***Abstract*— This paper introduces a scalable distributed probabilistic inference algorithm for intelligent sensor networks, tackling challenges of continuous variables, intractable posteriors and large-scale real-time data. In a centralized setting, variational inference is a fundamental tool to extend the utility of Bayesian estimation, by approximating a parametrized form of an intractable posterior density. Our key contribution is deriving the distributed evidence lower bound (DELBO) from the centralized estimation objective, whose separable structure enables distributed inference with one-hop sensor communication. The DELBO consists of observation likelihood and divergences to prior estimates, and the gap to the measurement evidence is ascribed to consensus and modeling errors. For supervised learning, we design a DELBO-maximizing online distributed algorithm, and specialize it to Gaussian variational densities with non-linear likelihoods. We extend the resulting distributed Gaussian variational inference (DGVI) updates via diagonalized and 1-rank covariance inversions for high-dimensional estimates and apply it to multi-robot probabilistic mapping using indoor LiDAR data.**

I. Introduction

Modern cyber-physical networks composed of autonomous vehicles and IoT devices generate large volumes of data continuously. Estimating variables and parameters of interest from the data efficiently and accurately subject to the computation, communication, and storage constraints of the networked devices is a critical problem. For instance, multi-robot mapping [49] requires learning the map parameters by robots collecting occupancy data online. Low onboard storage and processing capabilities may limit the robot's ability to perform inference with exhaustive sampling. Networks with limited communication bandwidth may not transmit upto a million points generated each second by LiDARs. New methods are needed to handle the communication and processing restrictions in distributed estimation.

Bayesian inference is a probabilistic estimation method that accumulates observation likelihood information to compute the (posterior) distribution of the variables of interest conditioned on the observations. This is especially useful in prediction problems because the uncertainty quantification provided by the posterior distribution helps limit overconfidence about the best estimate. Yet, the Bayesian approach comes at a cost, which is computational intractability for general observation models. This has given rise to approximate inference rules, including expectation propagation and variational inference, which can provide more efficient posterior computations. This work investigates the design of a distributed variational inference algorithm that can handle continuous variables, intractable posteriors, and large datasets in sensor networks.

Related work: Variational inference (VI), a technique outlined in [22], is a method to approximate intractable posteriors in standard Bayesian inference. It finds application to diverse problems such as state estimation [16], learning from demonstrations [41], and simultaneous localization and mapping [2]. VI has also been used to train autoencoders and deep generative models [24], [40]. In VI [20], posterior probability density functions (pdf) are calculated to maximize a lower bound (ELBO) on measurement evidence containing divergence to the true posterior pdf. See the early work [15], which computes such updates for conjugate families of prior and likelihood distributions. However, many applications require non-linear log-likelihood models and non-conjugate priors. Posterior sampling techniques relying on sequential or Hamiltonian Monte Carlo sampling [8], [43] produce posterior approximations by collecting samples from a Markov chain model. Recent work [10] established that VI solutions achieve a non-asymptotic convergence rate under conditions such as concave log-likelihoods, if the samples represent the posterior well. However, obtaining enough representative samples becomes computationally prohibitive in high-dimensional problems. Instead, stochastic optimization algorithms [17] are applied to the ELBO objective to learn an approximate posterior density from noisy gradients. Under some assumptions, stochastic gradient descent can even be interpreted as a Markov chain to infer posteriors [29]. We rely on gradient descent to derive updates specialized to a class of parametric families for analytic computation.

A popular adaptation of stochastic optimization in VI takes the form of Gaussian variational inference (GVI), where a Gaussian posterior is estimated for non-linear data likelihoods. Barfoot et al. [2] estimate blocks of a sparse information matrix to develop an online GVI algorithm. However, none of these inference methods yield a distributed framework, needed to share computational load across the network, and avoid raw data transmission. Decentralized algorithms perform better even in practice [27] as they reduce the load on the busiest node and avoid single point failures. In what follows, we specialize our review in probabilistic inference to federated learning, and distributed optimization and estimation literature.

Federated learning was originally developed for learning

The authors are with the Contextual Robotics Institute, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093 {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.

models over data repositories [23] in server-client architectures, such as edge computing. Federated averaging was shown to perform accurate inference on non-IID data distributions over this architecture in [32], with posterior density averaging in [1]. There have been recent extensions to fully decentralized settings with non-IID data [4], [51], [52]. In Gaussian inference, the covariance matrix is updated from batches of data in federated settings [35]. More recently, model aggregation has been studied over arbitrary communication networks [46]. Their work draws from the social learning analysis to upper bound the error in the estimated pdf but the updates rely on sample-intensive Monte Carlo methods.

In contrast, distributed optimization problems such as distributed least squares require consistent solutions under arbitrary connectivity. The algorithmic solutions minimize a sum of separable objective functions subject to a consensus constraint; see the recent survey on distributed learning via parametric optimization [6]. Variants of stochastic gradient descent are widely used to obtain consistent solutions with inexact local gradient samples, but most are limited to finite dimensional point estimates [48]. Additionally, the guarantees for strongly convex objectives do not hold for the divergence terms in a VI objective. For these divergence objectives, we perform probabilistic inference in presence of noisy gradients evaluated at the data streamed over a connected network. This differs from prior work in [38], [45], that present a class of distributed Bayesian algorithms estimating entire pdfs.

Distributed Bayesian filters have been developed as a nonlinear extension to classical estimation techniques [7], such as the variants of extended Kalman and particle filters. Such approaches [12] are limited to low dimensional estimates, due to high computational cost. Distributed estimation commonly relies on linear and geometric averaging for locally pooling neighbor estimates [14]. The seminal work in [21] estimates a probability mass function by averaging neighbor estimates followed by a Bayesian update on local likelihood samples. But even for these algorithms, efficient implementations are restricted to conditionally conjugate families of distributions. To relax this assumption, we combine VI methods with such distributed Bayesian algorithms with noisy gradients for strongly connected directed networks. An existing VI algorithm [19] solves a similar distributed inference problem, but our solution avoids the reliance on computationally expensive sampling. We instead look at specific classification, regression, and filtering models to obtain analytical updates.

Contributions: This work designs a distributed variational inference algorithm to perform probabilistic supervised learning in a network of agents collecting data independently. Our contributions are the following. (i) We derive a distributed version of the evidence lower bound (ELBO) to approximate posterior densities via optimization. This approach allows the implementation of probabilistic updates even when the likelihoods are not conjugate to the prior densities. (ii) We design a separable form of the distributed ELBO (DELBO) with local objectives at each agent. This enables the design of a fully distributed iterative inference algorithm. (iii) By approximating posterior densities using Gaussian pdfs, we derive an associated distributed Gaussian variational inference (DGVI) algorithm, with an iterative update to handle any nonlinear likelihoods. The specialization to diagonal covariances improves computational efficiency to enable large-scale inference. (iv) Finally, we apply these algorithms to achieve distributed probabilistic classification in multi-robot mapping problems using streaming LiDAR data.

The rest of the manuscript is organized as follows. Section II formulates the distributed inference problem over the space of pdfs. Section III introduces variational inference and derives the ELBO. Section IV devises a distributed version of the evidence lower bound which leads to distributed variational inference. Tractable iterative update rules are presented in Section V for Gaussian family densities. These algorithms are demonstraed in multi-robot mapping problems in Section VI.

II. Problem formulation: Distributed inference

Consider n agents $\mathcal{V} = \{1, \dots, n\}$ aiming to estimate an unknown variable $\theta \in \mathbb{R}^l$ cooperatively. The variable θ may represent a measurement source in environmental monitoring, relative agent positions in a localization problem, or environment occupancy in a mapping problem. The agents need to address two main challenges: 1) observations are received online and are noisy and 2) the observations are partially informative about θ due to the agents' states and limited sensing capabilities. Therefore, the agents need to cooperate to learn an accurate and consistent estimate of θ. Suppose that agent i receives observation $z_{i,t} \in \mathbb{R}^d$, at each time t, according to a known likelihood model $\ell_i(z_{i,t}|\theta)$. We make the following assumptions.

Assumption 1 (Independence and differentiability)**.** *The observations $z_t = \{z_{i,t}\}_{i\in\mathcal{V}}$ received by the agents at any time t are independent samples of the likelihood $\ell(z_t|\theta) = \prod_{i\in\mathcal{V}} \ell_i(z_{i,t}|\theta)$. The log likelihood $\log \ell_i(z_{i,t}|\theta)$ is twice differentiable in terms of θ.*

To account for stochastic and partially informative observations, the agents are to cooperatively agree on a probability distribution $p(\theta)$ over the variable θ. This cooperation is enabled by communication over a strongly connected digraph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The edge $(i, j) \in \mathcal{E}$ implies that node j transmits information to node i. Recall that a graph is strongly connected [5] if there exist a directed path between any two nodes in the network, thus allowing flow of information across nodes. The allowable information flow is captured using a non-negative, irreducible weighted adjacency matrix A, such that with $A_{ij} > 0$ only if $(i, j) \in \mathcal{E}$. Using the Sinkhorn's algorithm [42], the adjacency matrix can be made doubly stochastic, i.e., $A\mathbf{1}_n = A^\top \mathbf{1}_n = \mathbf{1}_n$, where $\mathbf{1}_n$ is a vector of ones. Therefore, we assume the following.

Assumption 2 (Connectivity)**.** *The weighted adjacency matrix A representing the communication graph $\mathcal{G}$ is doubly stochastic $A\mathbf{1}_n = A^\top \mathbf{1}_n = \mathbf{1}_n$ with $A_{ii} > 0$ and strongly connected.*

The collaborative network thus aims to estimate the density $p(\theta|z_{\le t}) \in \mathcal{F} \subseteq \mathcal{P}(\mathbb{R}^l)$ at time t, where $z_{\le t}$ represents observations collected by all agents until time t, $\mathcal{P}(\mathbb{R}^l)$ is the set of all probability densities over $\mathbb{R}^l$ and $\mathcal{F}$ is some known family of pdfs. We assume that the selected agent priors

$p_i(\theta|z_{<0})$ are positive over the feasible domain in θ. Based on this, we formally state the problem.

Problem 1. *Given observations $\{z_{i,t}\}$ sampled from the agent observation models $\ell_i(z_{i,t}|\theta)$, and priors $\{p_i(\theta|z_{<0})\}$ over an unknown parameter θ, compute a posterior pdf $p_i(\theta|z_{\leq t}) \in \mathcal{F}$, where $\mathcal{F}$ is a known pdf family and subject to consensus constraint $p_i(\theta|z_{\leq t}) = p(\theta|z_{\leq t})$, for $i \in \mathcal{V}$ and any $t \geq 0$.*

There are three key challenges in this problem, namely online and private observations, consensus constraint on estimated densities with restricted communication and inference constrained to a known pdf family $\mathcal{F}$.

III. Background

This section reviews the centralized variational inference (VI) approach, that we later extend to the proposed distributed VI setting. The classic Bayes approach calculates the posterior distribution of a parameter θ at time t as,

$$p(\theta|z_{\leq t}) = \frac{\ell(z_t|\theta)p(\theta|z_{<t})}{p(z_t|z_{<t})}, \tag{1}$$

by which the posterior $p(\theta|z_{\leq t})$ is proportional to the likelihood $\ell(z_t|\theta)$ and the prior $p(\theta|z_{<t})$. The posterior in (1) has an analytic expression only if the prior is conditionally conjugate to the likelihood [13]. For instance, combining a Gaussian prior with Gaussian linear likelihood results in a standard Gaussian posterior update. Yet, the exact calculation of (1) for general prior-likelihood pairs is not possible, as the computation of the normalization factor $p(z_t|z_{<t}) = \int \ell(z_t|\theta)p(\theta|z_{<t})d\theta$ is intractable.

The Bayesian inference rule (1) can be obtained as the solution to a maximization problem over the space $\mathcal{P}(\mathbb{R}^l)$ of probability distributions $q(\theta)$ on $\theta \in \mathbb{R}^l$. This maximization is performed over the so-called Evidence Lower Bound (ELBO). The VI approach specializes this problem to a family of finite-dimensional pdfs, $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^l)$, which often includes exponential densities [50]. Despite ELBO's ubiquity in the VI literature, we briefly reproduce it here for the sake of completeness and clarify the parallel with the proposed distributed version. To proceed, for pdfs $p, q \in \mathcal{F}$, we define KL-divergence $\mathrm{KL}[q(\theta)||p(\theta)] = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right]$.

Lemma 1. *Given a pdf $q(\theta)$, the log-normalization factor $\log p(z_t|z_{<t})$ in (1) is lower bounded by the ELBO,*

$$\mathop{\mathbb{E}}_{q(\theta)}[\log \ell(z_t|\theta) - \log(q(\theta)) + \log p(\theta|z_{<t})].$$

Proof. Using (1), we express the log-normalization factor in terms of the approximated posterior pdfs and drop the non-negative variational gap term $\mathrm{KL}[q(\theta)||p(\theta|z_{\leq t})] \geq 0$ to compute its lower bound as,

$$\begin{aligned}\log p(z_t|z_{<t}) &= \mathop{\mathbb{E}}_{q(\theta)}\left[\log \frac{\ell(z_t|\theta)p(\theta|z_{<t})q(\theta)}{p(\theta|z_{\leq t})q(\theta)}\right] \\ &\geq \mathop{\mathbb{E}}_{q(\theta)}[\log \ell(z_t|\theta)] - \mathrm{KL}[q(\theta)||p(\theta|z_{<t})] \\ &= \mathop{\mathbb{E}}_{q(\theta)}[\log \ell(z_t|\theta) - \log q(\theta) + \log p(\theta|z_{<t})]. \end{aligned} \tag{2}$$

Since θ in $q(\theta)$ is independent of the data $z_{\leq t}$, the expectation does not alter the value of the log-normalization. □

To continue iteratively in VI, we approximate the posterior $p(\theta|z_{\leq t})$ using pdf $q_t(\theta)$ in a family $\mathcal{F}$ for each time t. We replace the last posterior $p(\theta|z_{<t})$ by its known approximator $q_{t-1}(\theta)$ and maximize ELBO to select the next posterior $q_t(\theta)$,

$$q_t(\theta) \in \mathop{\arg\max}_{q(\theta)\in\mathcal{F}} \{\mathbb{E}_q[\log \ell(z_t|\theta)] - \mathrm{KL}[q||q_{t-1}]\}. \tag{3}$$

The lower bound explains the modeling error induced by the choice of the distributional family $\mathcal{F}$. Thus, VI can be interpreted as finding the best $\mathcal{F}$-constrained optimizer [25, Section 2.2]. Hereafter, p denotes the pdf of a random variable, and q is a variational approximation from the family $\mathcal{F}$.

IV. Distributed evidence lower bound

In this section, we derive a distributed version of the VI optimization problem in (3). In this setting, the n agents follow Assumption 1 to collect data independently. Each agent i maintains its local pdf $p_i(\theta|z_{<t})$ estimating the centralized density $p(\theta|z_{<t})$ over the parameter θ at time t. Since the agents have their own likelihood models, their estimated densities may not be equal. Therefore, we represent the centralized prior as $p(\theta|z_{<t}) \propto \prod_{i=1}^n p_i(\theta|z_{<t})^{1/n}$, using the geometric average of the agent pdfs. The geometric average is chosen for its mode preserving properties [30] when combining multiple pdfs. With this mean, we can rewrite Bayes' rule with corresponding normalization factor $p(z_t|z_{<t}) = \int \prod_{i\in\mathcal{V}} \ell_i(z_{i,t}|\theta)p_i(\theta|z_{<t})^{1/n}d\theta$ as,

$$p(\theta|z_{\leq t}) = \frac{\prod_{i\in\mathcal{V}} \ell_i(z_{i,t}|\theta)p_i(\theta|z_{<t})^{1/n}}{p(z_t|z_{<t})}. \tag{4}$$

To perform estimation using VI, we start by computing a lower bound on the log-normalization term analogous to the ELBO in (2). For distributed implementation, we express a separable version of the VI objective, summing over terms containing an agent's likelihood and neighbor estimates. To satisfy the consensus constraint while performing inference, we assume that each agent i estimates pdf $q_{i,t} \in \mathcal{F}$ that is equal to some q_t. Maximizing the separable components at each agent yields a distributed probabilistic inference algorithm, where each component contains the corresponding agent's private observations and one-hop neighbor estimates.

Theorem 1. *Given agent pdfs $q_{i,t}(\theta) = q_t(\theta)$ for some pdf $q_t(\theta)$ and agents $i \in \mathcal{V}$, the log-normalization factor $\log p(z_t|z_{<t})$ in (4) is lower bounded by the separable distributed evidence lower bound (DELBO),*

$$\sum_{i\in\mathcal{V}} \mathop{\mathbb{E}}_{q_{i,t}}[\log \ell_i(z_{i,t}|\theta) - \frac{1}{n}\log(q_{i,t}(\theta)) + \sum_{j\in\mathcal{V}} \frac{A_{ij}}{n} \log p_j(\theta|z_{<t})],$$

where A is the adjacency matrix satisfying Assumption 2.

Proof. Given the agent pdfs $p_i(\theta|z_{<t})$, the centralized estimate at time t is defined as their normalized geometric average $p(\theta|z_{<t}) = \frac{1}{K_{<t}} \prod_{i\in\mathcal{V}} (p_i(\theta|z_{<t}))^{1/n}$. The normalization factor $K_{<t} = \int \prod_{i\in\mathcal{V}} (p_i(\theta|z_{<t}))^{1/n} d\theta$ is the integral of the geometric average. Due to the column stochasticity of matrix A from Assumption 2, the geometric average satisfies $\prod_{i\in\mathcal{V}} (p_i(\theta|z_{<t}))^{1/n} = \prod_{i\in\mathcal{V}} (\prod_{j\in\mathcal{V}} p_j(\theta|z_{<t})^{A_{ij}})^{1/n}$. By definition of positive terms in A, this property relates the agent prior densities with those of its one-hop neighbors. Analogous

to the ELBO derivation, the normalization in (4) is expressed in terms of the agent log likelihoods, neighbor prior estimates and the posterior as,

$$p(z_t|z_{<t}) = \frac{p(z_t|\theta)p(\theta|z_{<t})}{p(\theta|z_{\leq t})} = \frac{1}{K_{<t}} \prod_{i\in\mathcal{V}} \frac{\ell_i(z_{i,t}|\theta)p_i(\theta|z_{<t})^{\frac{1}{n}}}{p(\theta|z_{\leq t})^{\frac{1}{n}}}$$

The geometric average of the non-negative pdfs is pointwise upper bounded by their arithmetic average, and, hence, its integral satisfies $K_{<t} \leq \int \sum_i (1/n) p_j(\theta|z_{<t}) d\theta = 1$. As a result, $\log K_{<t} \leq 0$. As in the centralized setting, since the argument in pdf $q_t(\theta)$ is independent of the observation $z_{\leq t}$, the expectation of the log-normalization factor does not alter its value. Assuming that $q_{i,t}(\theta) = q_t(\theta)$, we separate the expectation over the agent likelihoods and priors as follows,

$$\log p(z_t|z_{<t}) = -\mathop{\mathbb{E}}_{q_t(\theta)} \log K_{<t} \tag{5}$$
$$+ \mathop{\mathbb{E}}_{q_t(\theta)} \sum_{i\in\mathcal{V}} \left[\log \frac{\ell_i(z_{i,t}|\theta) \prod_{j\in\mathcal{V}} p_j(\theta|z_{<t})^{\frac{A_{ij}}{n}} q_{i,t}(\theta)^{\frac{1}{n}}}{q_{i,t}(\theta)^{1/n} p(\theta|z_{\leq t})^{1/n}} \right].$$

$$\log p(z_t|z_{<t}) \geq \sum_{i\in\mathcal{V}} \mathop{\mathbb{E}}_{q_{i,t}(\theta)} [\log \ell_i(z_{i,t}|\theta)] \tag{6}$$
$$+ \frac{1}{n} \mathrm{KL}[q_{i,t}(\theta)||p(\theta|z_{\leq t})] - \frac{1}{n} \mathrm{KL}[q_{i,t}(\theta)||p_i^g(\theta|z_{<t})],$$
$$\geq \sum_{i\in\mathcal{V}} \mathop{\mathbb{E}}_{q_{i,t}(\theta)} [\log \ell_i(z_{i,t}|\theta)] - \frac{1}{n} \mathrm{KL}[q_{i,t}(\theta)||p_i^g(\theta|z_{<t})],$$

where $p_i^g(\theta|z_{<t}) = \prod_{j\in\mathcal{V}} p_j(\theta|z_{<t})^{A_{ij}}$ in the weighted geometric average of the agent prior pdfs. Since the KL divergence term representing the modeling error between the approximation $q_{i,t}$ and the estimate $p(\theta|z_{\leq t})$ is non-negative, we can drop this term to obtain a separable lower bound of the log-normalization factor as,

$$\log p(z_t|z_{<t}) \geq \sum_{i\in\mathcal{V}} \left[\mathop{\mathbb{E}}_{q_{i,t}(\theta)} [\log \ell_i(z_{i,t}|\theta)] \right.$$
$$\left. - \frac{1}{n} \sum_{j\in\mathcal{V}} \mathop{\mathbb{E}}_{q_{i,t}(\theta)} A_{ij} [\log q_{i,t}(\theta) - \log p_j(\theta|z_{<t})] \right] \tag{7}$$

The separable terms contain only the agent's observation z_i and are thus analogous to the ELBO at each agent. □

While deriving the DELBO in Theorem 1, we observe that posterior approximation contains modeling and consensus error terms. The consensus error at time t is defined in (5) as $\log(1/K_{<t})$ where $K_{<t} = \int \prod_i p_i(\theta|z_{<t})^{1/n} d\theta$. Since

$$\log(1/K_{<t}) = 1/n \sum_{i\in\mathcal{V}} \mathrm{KL}[p_g||p_i(\theta|z_{<t})]$$

for $p_g = \prod_i p_i(\theta|z_{<t})^{1/n}/K_{<t}$, this error is zero only if the agent pdfs are equal almost everywhere. The modeling error is defined in (6) as the divergence $\sum_i \mathbb{E}_{q_{i,t}} \mathrm{KL}[q_{i,t}||p_i(\theta|z_{\leq t})]$. This error is zero only if the pdfs $q_{i,t}$ are computed in the family of accurate posterior densities. Replacing the accurate pdfs $p_i(\theta|z_{<t})$ with their last known approximations $q_{i,t-1}$ in family $\mathcal{F}$ in DELBO yields a separable functional $J_t[q_{1,t}, \ldots, q_{n,t}] = \frac{1}{n} \sum_{i\in\mathcal{V}} J_{i,t}[q_{i,t}]$ with,

$$J_{i,t}[q_{i,t}] = n \mathop{\mathbb{E}}_{q_{i,t}} \left[\log \left(\ell_i(z_{i,t}|\theta) \prod_{j\in\mathcal{V}} q_{j,t-1}^{\frac{A_{ij}}{n}} \right) - \log q_{i,t}^{\frac{1}{n}} \right] \tag{8}$$
$$= n \mathop{\mathbb{E}}_{q_{i,t}} [\log[\ell_i(z_{i,t}|\theta)]] - \mathrm{KL}[q_{i,t}|| \prod_{j\in\mathcal{V}} q_{j,t-1}^{A_{ij}}].$$

The weighted sum of KL-divergences in (8) penalizes deviation from consensus of the agent pdfs $q_{i,t-1}$. Sharing weighted pdfs with neighbors is key to reaching consistent estimates across the network. The positive terms in the matrix A enforce the communication links into the separable components. The assumption on posteriors $q_{i,t} = q_t$ merely aids the design of the DELBO, with its local solution stated next.

Corollary 1. *The pdf $q_{i,t}(\theta)$ maximizing the DELBO component $J_{i,t}$ in (8) is given as,*

$$q_{i,t}(\theta) = \ell_i(z_{i,t}|\theta)^n q_i^g(\theta) \Big/ \int \ell_i(z_{i,t}|\theta)^n q_i^g(\theta) d\theta, \tag{9}$$

where the mixed pdf at agent i is $q_i^g(\theta) = \prod_{j\in\mathcal{V}_i} q_{j,t-1}(\theta)^{A_{ij}}$.

Proof. We follow the proof in [37, Proposition 2] using the Gateaux derivative $\frac{\delta}{\delta q_i} \mathrm{KL}[q_i||q_j] = 1 + \log(q_i/q_j)$. The constraint $\int q_{i,t} = 1$ is used to construct the Lagrangian with multiplier λ as $\mathcal{L}(q_{i,t}, \lambda) = J_{i,t}[q_{i,t}] + \lambda(\int q_{i,t} - 1)$. Its variation with respect to $q_{i,t}$ is,

$$\frac{\delta \mathcal{L}}{\delta q_{i,t}} = n \log[\ell_i(z_{i,t}|\theta)] - \sum_{j\in\mathcal{V}} (1 + \log q_{i,t} - A_{ij} \log q_{j,t-1}) + \lambda.$$

Setting the variation to zero and solving for $q_{i,t}$ leads to:

$$q_{i,t}(\theta) = e^{\lambda - 1} \ \ell_i(z_{i,t}|\theta)^n \prod_{j\in\mathcal{V}_i} q_{j,t-1}(\theta)^{A_{ij}}.$$

The value of λ can be obtained from the constraint $\int q_{i,t}(\theta) = 1$ yielding the result in (9). □

For consensus, the asymptotic averaging properties $\lim_{t\to\infty} A^t = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ of matrix A generate agent estimates eventually consistent with the centralized one $q_t(\theta) = q_{i,t}(\theta)$. To observe the impact of matrix A on guaranteeing consensus in distributed estimation problems, please refer to the convergence analysis in [33], [36], [38].

Remark 1 (Distributed estimation). With conjugate agent likelihoods $\ell_i(z_i|\theta)$ weighted by factor n, the distributed updates in [38] match the DELBO updates, thus guaranteeing probabilistic convergence for accurate posterior computations.

The posterior $p(\theta)$ in (9) can be approximated for arbitrary likelihood pdfs using black-box VI [39] in the variational message passing framework [47]. We employ this approach in the next example to show the impact of sampling on accuracy.

Example 1 (Estimating geometric mixing of Gaussians). In this example, we examine a sampled version of the update in (9) from the perspective of agent 1 and time $t = 0$ with neighbor weights $A_{1j} = 1/n$ in a network of $n = 4$ agents with Gaussian priors and likelihoods. Because of the dependence sampling, we observe that the VI solution may not match the analytical solution. Assume that the Gaussian prior for any agent $j \in \mathcal{V}$ at time $t = 0$ is $q_{j,0}(\theta) = \mathcal{N}(\mu_{j,0}, (\Omega_{j,0})^{-1})$ with mean $\mu_{j,0}$, and informa-

tion matrix $\Omega_{j,0}$. Suppose that the local observation likelihoods $\ell_j(z_{j,t}|\theta) = \mathcal{N}(H\theta, (\Omega_j^z)^{-1})$ are Gaussian as well. Since the geometric average of the priors is conditionally conjugate to the likelihood, the posterior at agent 1 is $\mathcal{N}(\Omega_{1,1}^{-1}(bH^\top\Omega_1^z z_{1,1} + \sum_{j=1}^n A_{1j}\Omega_{j,0}\mu_{j,0}), \Omega_{1,1}^{-1})$, with information matrix $\Omega_{1,1} = nH^\top\Omega_1^z H + \sum_{j=1}^n A_{1j}\Omega_{j,0}$.

To compare, we estimate this Gaussian posterior using VI with sampling [28]. Let the agent estimate an expressive pdf $p(\theta) = \mathcal{N}(\theta|\mu, \Omega^{-1})\, p_\mu\, p_\Omega$ using observation $z_{1,1}$ and prior normal distribution $p_\mu = \mathcal{N}(\mu_p, \Sigma_p)$ on the mean μ and Wishart distribution $p_\Omega = W(\lambda, V)$ on the precision matrix. To correctly estimate $q_{1,1}(\theta)$, the proposed samples must span the support of the unknown posterior. Therefore, we consider the component pdfs $q_{j,0}$ as the proposal for generating samples of θ and weigh each sample with $\ell_1(z_{1,1}|\theta)^n \prod_{j\in\mathcal{V}_i} q_{j,0}(\theta)^{A_{1j}}$ from the update in (9). Upon normalization, stratified resampling generates proposal samples representing the posterior which is then used to obtain $q(\theta)$. The number and distribution of these posterior samples is crucial of accurate inference.

Due to the need of high number of proposal samples, sample intensive VI approaches are computationally expensive in high-frequency online estimation settings such as filtering. Further, even significant number of proposal samples produce a good estimate only if they represent the posterior well. We see this issue with the mean and covariance of the inferred density in Fig. 1. For the sampled particles inside a 3 unit radius circle centered at $(0, 0)$, we observe that the estimated mean is 0.5 units away from the analytical value. Over multiple time steps, the sampling error may accumulate. Therefore, we will develop approximate analytical updates to perform computationally efficient and accurate online inference.

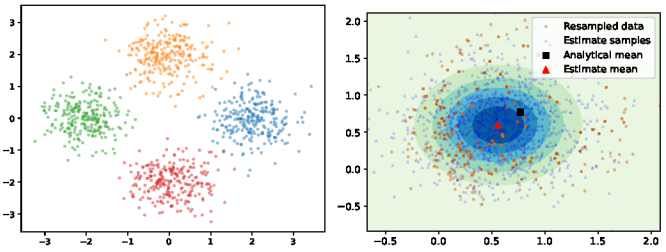

Fig. 1. (a) Samples from Gaussian priors $\boldsymbol{p_{i,0}}$ with unit covariance and means on a circle of radius $\mathbf{1}$. (b) Particles resampled w.r.t. probability weights in (9) for data $\boldsymbol{z_{1,1}} = [\mathbf{1}, \mathbf{1}]$, estimated pdf and analytical mean.

In this section, we derived a distributed variational inference algorithm in (9) requiring costly computation of the normalization factor. To enable efficient implementation, we further develop this algorithm to use stochastic gradients of log-likelihood terms and compute their analytical approximations.

V. Distributed Gaussian variational inference

This section derives agent specific iterative updates for variational inference with Gaussian variational densities and twice differentiable log-likelihood functions. Appropriate approximations to the expected log-likelihood derivatives are devised to generate analytical Gaussian updates for distributed classification and regression problems. Further, rank-correcting inverse and diagonalized covariance updates are presented to support efficient real-time implementation.

A. Distributed Gaussian variational inference (DGVI)

We assume that the agents collect observations from individual likelihoods that may not be Gaussian but estimate variational pdfs $q_{i,t}(\theta)$ restricted to a Gaussian pdf family $\mathcal{F}$. The solution to the ELBO optimization in (3) for a Gaussian pdf family $\mathcal{F}$ is stated in the next proposition.

Proposition 1 (Gaussian variational inference)**.** *Assuming that the known prior density $q_{t-1}(\theta)$ is a Gaussian $\mathcal{N}(\theta|\mu_{t-1}, \Omega_{t-1}^{-1})$ with mean μ_{t-1} and information matrix Ω_{t-1}, the Gaussian pdf q_t maximizing the ELBO in (3) is,*

$$\begin{aligned}\Omega_t &= \Omega_{t-1} - \mathbb{E}_{q_{t-1}}[\nabla_\theta^2 \log \ell(z_t|\theta)],\\ \mu_t &= \mu_{t-1} + \Omega_t^{-1}\mathbb{E}_{q_{t-1}}[\nabla_\theta \log \ell(z_t|\theta)].\end{aligned} \tag{10}$$

Proof. The proof is presented in Appendix A. We pose the ELBO objective as the loss functional in [2, Eqn. 25], avoiding the implicit expectation $\mathbb{E}_{q_t}[\nabla_\theta \log \ell(z_t|\theta)]$ as in [26]. □

Proposition 1 has an online update maximizing the ELBO objective over the set of Gaussian densities in $\mathcal{F}$. The DELBO in Theorem 1 admits separable objectives for each agent. Each DELBO component contains only the agent's current observation likelihood that is based on sampled data and the last received estimates from neighbors, in alignment with iterative sampling and synchronous communication structure. When the DELBO component is optimized locally at agent i, the resulting inference update is online and distributed owing to the access to current likelihood and last neighbor estimates respectively. The following proposition optimizes the agent components of the distributed objective in (8) over Gaussians.

Proposition 2 (Distributed Gaussian variational inference)**.** *Let agent i in an n-node network observe $z_{i,t}$ with likelihood $\ell(z_{i,t}|\theta)$ and receive neighbor estimates $q_{j,t-1}(\theta) = \phi(\theta|\mu_{j,t-1}, \Omega_{j,t-1}^{-1})$ at time t. Weighing the neighbor estimates with matrix A, the mean and information matrix of the pdf $\phi(\theta|\mu_{i,t}, \Omega_{i,t}^{-1})$ maximizing DELBO in (8) are,*

$$\begin{aligned}\Omega_{i,t}^g &= \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}, \Omega_{i,t}^g\mu_{i,t}^g = \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}\mu_{j,t-1}\\ \Omega_{i,t} &= \Omega_{i,t}^g - n\mathbb{E}_{q_{i,t}^g}[\nabla_\theta^2 \log \ell(z_{i,t}|\theta)],\\ \mu_{i,t} &= \mu_{i,t}^g + n(\Omega_{i,t}^g)^{-1}\mathbb{E}_{q_{i,t}^g}[\nabla_\theta \log \ell(z_{i,t}|\theta)],\end{aligned} \tag{11}$$

where $q_{i,t}^g(\mu_{i,t}^g, \Omega_{i,t}^g)$ is the geometric average of prior pdfs.

Proof. Please refer to Appendix A. □

Both the centralized and distributed Gaussian variational update rules in Propositions 1 and 2 contain the expected log-likelihood gradient and Hessian terms. Estimating the expectations using Monte Carlo methods is computationally expensive, especially for high-dimensional parameters. Therefore, we derive analytic approximations of the gradient and Hessian expectations for classification and regression problems in the next two subsections.

B. DGVI for classification

We consider a kernel-based observation likelihood model for probabilistic classification. The kernel parameters consist of a set of known fixed feature points and corresponding weights. The data $z = (x, y)$ is embedded in feature space by a transformation $\Phi_x = [1, k_1(x), \dots, k_l(x)]$ with elements $k_s(x) = \gamma_1 \exp(-\gamma_2 \|x - x^{(s)}\|^2)$ where $x^{(s)}$ are the known kernel centers and (γ_1, γ_2) are kernel scaling parameters chosen to suit the domain and regularity of the model. The likelihood of an observation $z = (x, y)$ with input $x \in \mathbb{R}^d$, feature $\Phi_x \in \mathbb{R}^{l+1}$, and label $y \in \{0, 1\}$ is modeled as,

$$\ell(z|\theta) = \sigma(\Phi_x^\top \theta)^y (1 - \sigma(\Phi_x^\top \theta))^{1-y}, \tag{12}$$

defined with model parameters θ and the sigmoid function σ.

To estimate the distribution of the parameters θ using the GVI algorithm in Proposition 1, we would need to estimate the expectation over the log-likelihood gradient, $\nabla_\theta \log p(z|\theta)$, and Hessian, $\nabla^2_\theta \log p(z|\theta)$. We derive an analytical approximation to these terms. With $\nabla_\theta \sigma(\Phi_x^\top \theta) = \sigma(\Phi_x^\top \theta)(1 - \sigma(\Phi_x^\top \theta))\Phi_x^\top$, the log-likelihood derivatives are,

$$\log \ell(z|\theta) = y \log \sigma(\Phi_x^\top \theta) + (1 - y) \log(1 - \sigma(\Phi_x^\top \theta)),$$
$$\nabla_\theta \log \ell(z|\theta) = (y - \sigma(\Phi_x^\top \theta))\Phi_x^\top, \tag{13}$$
$$\nabla^2_\theta \log \ell(z|\theta) = -\sigma(\Phi_x^\top \theta)(1 - \sigma(\Phi_x^\top \theta))\Phi_x \Phi_x^\top. \tag{14}$$

To analytically compute the expectation of gradient, Hessian and their derivative terms with respect to a Gaussian density, we approximate the sigmoid function $\sigma(x)$ with an inverse probit function $\Gamma(\xi x) = \int_{-\infty}^{\xi x} \phi(\alpha|0, 1) d\alpha$ for $\xi = 0.61$ according to [9]. Fortunately, the expectation of the inverse probit function with respect to a Gaussian density is an inverse probit. For the second derivative, the derivative of the sigmoid function is approximated via a Gaussian probability density function ϕ with zero mean and unit covariance. Using $\sigma(\Phi_x^\top \theta) \approx \Gamma(\xi \Phi_x^\top \theta)$, the Hessian becomes,

$$\begin{aligned}\nabla^2_\theta \log \ell(z|\theta) &= -\nabla_\theta \sigma(\Phi_x^\top \theta)\Phi_x^\top \approx -\nabla_\theta \Gamma(\xi \Phi_x^\top \theta)\Phi_x^\top \\ &= -\xi \phi(\xi \Phi_x^\top \theta|0, 1)\Phi_x \Phi_x^\top. \end{aligned} \tag{15}$$

To specialize the DGVI algorithm in Proposition 2 to the classification objective, we next find analytic approximations of the expectation over gradient and Hessian terms.

Proposition 3 (Expected log-likelihood gradient and Hessian)**.** *For probabilisitic classification with a kernel-based observation likelihood model in (12), the expected gradient and Hessian of the log-likelihood in (13) with respect to a Gaussian density $q_t(\theta) = \phi(\theta|\mu_t, \Omega_t^{-1})$ satisfy,*

$$\mathbb{E}_{q_t}[\nabla_\theta \log \ell(z|\theta)] \approx \left(y - \Gamma\left(\xi \Phi_x^\top \mu_t \Big/ \sqrt{\beta}\right)\right) \Phi_x^\top,$$

$$\mathop{\mathbb{E}}_{q_t}[\nabla^2_\theta \log \ell(z|\theta)] \approx -\left(\xi \Big/ \sqrt{2\pi\beta}\right) e^{\left(-\frac{1}{2}[\frac{\xi^2}{\beta} \mu_t^\top \Phi_x \Phi_x^\top \mu_t]\right)} \Phi_x \Phi_x^\top,$$

where $\beta = 1 + \xi^2 \Phi_x^\top \Omega_t^{-1} \Phi_x$.

Proof. Please refer to Appendix B. □

Methods to estimate Gaussian variational posteriors are surveyed in [34], and the expectation propagation method is recommended for its accuracy. However, the associated computational complexity may not allow real-time implementation. Our approximations of the log-likelihood gradient and Hessian expectations can be substituted in Proposition 2 to obtain analytical updates for approximate distributed Gaussian VI. In the distributed setting, each agent knows the fixed kernel centers $\{x^{(s)}\}$ and scale parameters γ_1, γ_2, receives private observations $z_{i,t}$, and estimates a pdf over the weights θ.

Proposition 4 (DGVI for kernel classification)**.** *For observation $z = (x, y)$ received at agent i in an n node network, the classification likelihood defined in (12), and neighbor estimates $\phi(\theta|\mu_{j,t-1}, \Omega^{-1}_{j,t-1})$, the DELBO maximizing Gaussian $\mathcal{N}(\theta|\mu_{i,t}, \Omega^{-1}_{i,t})$ is,*

$$\Omega^g_{i,t} = \sum_{j \in \mathcal{V}} A_{ij} \Omega_{j,t-1},\ \Omega^g_{i,t} \mu^g_{i,t} = \sum_{j \in \mathcal{V}} A_{ij} \Omega_{j,t-1} \mu_{j,t-1},$$
$$\Omega_{i,t} = \Omega^g_{i,t} + \gamma_t \Phi_x \Phi_x^\top, \tag{16}$$
$$\Omega^{-1}_{i,t} = (\Omega^g_{i,t})^{-1} - \gamma_t/\gamma_{1,t} (\Omega^g_{i,t})^{-1} \Phi_x \Phi_x^\top (\Omega^g_{i,t})^{-1} \tag{17}$$
$$\mu_{i,t} = \mu^g_{i,t} + n \left(y - \Gamma\left(\xi \Phi_x^\top \mu^g_{i,t} \Big/ \sqrt{\beta}\right)\right) \Omega^{-1}_{i,t} \Phi_x \tag{18}$$

with $\beta = 1 + \xi^2 \Phi_x^\top (\Omega^g_{i,t})^{-1} \Phi_x$, $\gamma_{1,t} = 1 + \gamma_t \Phi_x^\top (\Omega^g_{i,t})^{-1} \Phi_x$ and $\gamma_t = n \sqrt{\frac{\xi^2}{2\pi\beta}} \exp\left(-0.5[\frac{\xi^2}{\beta} (\mu^g_{i,t})^\top \Phi_x \Phi_x^\top \mu^g_{i,t}]\right)$.

Proof. The mean $\mu^g_{i,t}$ and information matrix $\Omega^g_{i,t}$ describe the weighted geometric average of prior Gaussians. Then, the steps for Proposition 1 lead to the Gaussian maximizing the agent separable DELBO. The expected log-likelihood derivatives in in Proposition 2 are substituted with the analytic approximations in Proposition 3. This is followed by the steps reducing matrix inversion computations in Appendix B. □

The DGVI updates in Proposition 4 include two linear system solutions $(\Omega^g_{i,t})^{-1}(\sum_{j \in \mathcal{V}} A_{ij} \Omega_{j,t-1} \mu_{j,t-1})$ and $(\Omega^g_{i,t})^{-1} \Phi_x$. In a centralized setting, the matrix inversion needs to be performed only at the first step to compute Ω_0^{-1}, and any following inverses may be computed iteratively via (17). The costly matrix inversion can be avoided by using Gaussian pdfs with diagonal covariances, discussed next.

Proposition 5 (Diagonalized GVI for kernel classification)**.** *For observation $z = (x, y)$ received at agent i in an n node network, classification likelihood defined in (12), and neighbor estimates $\phi(\theta|\mu_{j,t-1}, D^{-1}_{j,t-1})$ with diagonal information matrices $D_{j,t}$, the DGVI update to Gaussian pdf $q_t(\theta) = \phi(\theta|\mu_{i,t}, D^{-1}_{i,t})$ with diagonal information matrix $D_{i,t}$ is,*

$$D^g_{i,t} = \sum_{j \in \mathcal{V}} A_{ij} D_{j,t-1},\ D^g_{i,t} \mu^g_{i,t} = \sum_{j \in \mathcal{V}} A_{ij} D_{j,t-1} \mu_{j,t-1},$$
$$D_{i,t} = D^g_{i,t} + \gamma \sqrt{\xi^2/2\pi\beta}\, \mathrm{diag}(\Phi_x \Phi_x^\top), \tag{19}$$
$$\mu_{i,t} = \mu^g_{i,t} + n (D^g_{i,t})^{-1} \left(y - \Gamma\left(\xi \Phi_x^\top \mu^g_{i,t} \Big/ \sqrt{\beta}\right)\right) \Phi_x^\top,$$

where $\gamma = n \exp\left(-0.5[\frac{\xi^2}{\beta} (\mu^g_{i,t})^\top \Phi_x \Phi_x^\top \mu^g_{i,t}]\right)$, and $\beta = 1 + \xi^2 \Phi_x^\top (D^g_{i,t})^{-1} \Phi_x$.

Proof. Please refer to Appendix C. □

For the classification likelihood introduced in (12), we have presented approximate analytic updates for inferring Gaussian densities over the unknown parameters. The updates consist of geometric average of Gaussian pdfs and likelihood updates with efficient inverse and diagonal covariance computations.

C. Distributed Gaussian variational inference for regression

In this section, we derive distributed Gaussian VI updates for regression. Consider a linear model $y = \Phi_x^\top \theta$ defined using a feature vector $\Phi_x = [1, k_1(x), \ldots, k_l(x)]$ with elements $k_m(x)$ defined as in Sec. V-B and parameters θ. Assume that agent i receives observation $z_i = (x, y)$ sampled from $\ell_i(z_i|\theta) \propto \exp(-0.5(y-\Phi_x^\top\theta)^\top S_i(y-\Phi_x^\top\theta))$ with symmetric and positive definite $S_i = S_i^\top$.

Proposition 6 (DGVI for kernel regression)**.** *For data (x, y) and neighbor estimates $\phi(\theta|\mu_{j,t-1}, \Omega_{j,t-1}^{-1})$ received by agent i at time t in an n node network, the Gaussian density $q_{i,t}(\theta) = \phi(\theta|\mu_{i,t}, \Omega_{i,t}^{-1})$ maximizing DELBO for regression is,*

$$\Omega_{i,t}^g = \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}, \Omega_{i,t}^g\mu_{i,t}^g = \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}\mu_{j,t-1}$$

$$\Omega_{i,t} = \Omega_{i,t}^g + n\Phi_x S_i \Phi_x^\top, \Sigma_{i,t}^g = (\Omega_{i,t}^g)^{-1} \tag{20}$$

$$\Omega_{i,t}^{-1} = \Sigma_{i,t}^g - \Sigma_{i,t}^g\Phi_x((nS_i)^{-1} + \Phi_x^\top\Sigma_{i,t}^g\Phi_x)^{-1}\Phi_x^\top\Sigma_{i,t}^g$$

$$\mu_{i,t} = \mu_{i,t}^g + n(\Omega_{i,t})^{-1}(\Phi_x S_i^\top y - \Phi_x S_i \Phi_x^\top \mu_{i,t}^g) \tag{21}$$

Proof. Please refer to Appendix D. □

This section derives distributed variational inference algorithms to estimate optimal Gaussian densities using derivatives of the sampled log-likelihoods. For specific classification and regression likelihoods, we present an efficient version of the Gaussian inference algorithm that approximates the expected values of these sampled log-likelihood derivatives.

VI. Results

In this section, we evaluate our distributed inference algorithms on classification and mapping datasets. For mapping, the functions Φ_x in (12) are kernel functions rooted around the spatial point $x^{(i)}$, and corresponding θ_i represent the weight on the corresponding occupancy kernel. We explain the inference model setup for centralized binary classification on a toy dataset, followed by distributed inference over synthetic and real LiDAR data to generate probabilistic occupancy maps [1].

Toy data: We consider the Banana dataset [3], which consists of 5300 points with binary labels, visualized in Fig. 2. The probability of each point belonging to the first class, estimated by centralized version of our VI algorithm with matrix $A = 1$ in Proposition 4, is visualized in Fig. 2. We pick 50 feature points at random, with scale $\gamma_1 = 1$ and lengthscale $\gamma_2 = 0.3$ to construct feature functions Φ_x as defined prior to (12). We select 50% data for training, and run the single-agent version of the algorithm in Proposition 2 updating the mean and covariance of the weights θ over the feature points. With $20k$ steps, the algorithm achieves 88% classification accuracy on test set.

Intel LiDAR dataset [18]: In a cooperative mapping problem, robots follow their own trajectories and cooperate to infer a common map of the environment. A LiDAR sensor uses time of flight information to compute the distance to obstacles in each direction. To construct an occupancy dataset, the points

[1]Source code available at https://github.com/pptx/distributed-mapping.

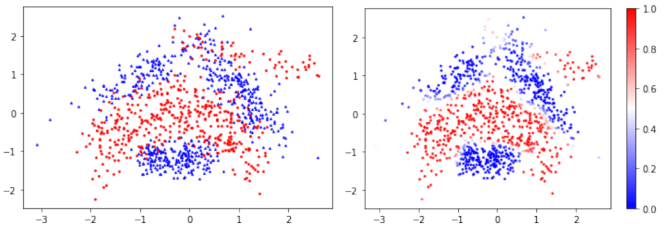

Fig. 2. True classes in Banana dataset (left) and predicted probability $\mathbb{E}_{q(\theta_T)}p(x, y|\theta)$ of point (x, y) belonging to the red class (right).

along the rays connecting the robot to obstacles are sorted into free and occupied points [11]. We assume that each networked robot extracts binary occupancy data from the LiDAR scans along its trajectory. To reduce the mapping effort, the robot trajectories may cover disjoint sections of the observed space, generating local data with different distributions.

Fig. 3 presents the results for single agent version of the algorithm in Proposition 4. We use 90% of the dataset for training. The remainder forms the test set with a small subset of 1000 samples forming the verification set for calculating the runtime error. The model is generated using 1200 feature points selected randomly from the testing set, with scale $\gamma_1 = 1$ and lengthscale $\gamma_2 = 0.5$. The diagonalized version of the algorithm in Proposition 5 runs for $400k$ steps to achieve 87% accuracy on the test set.

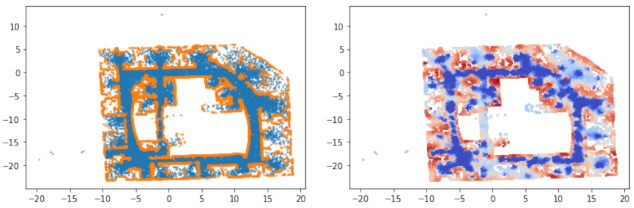

Fig. 3. True point classes from LiDAR scans with occupied spaces in orange (left). Predicted occupancy probability $\mathbb{E}_{q(\theta_T)}p(x, y|\theta)$ at position (x, y) in the test set. The darker red colors represent high occupancy probability, whereas blue represents the free space.

Fig. 4 presents the mean and diagonal covariance value at the individual feature points selected in the map, indicating uncertainty at the boundary of the free and occupied spaces.

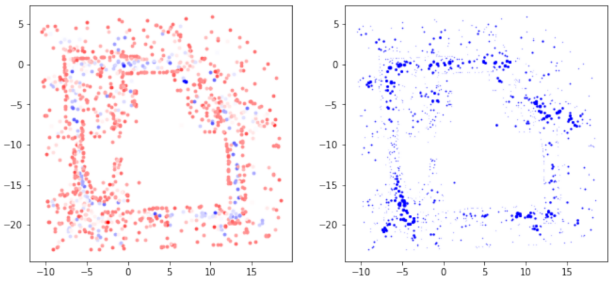

Fig. 4. Estimated mean μ_T and variance Σ_T of the parameter θ on **1200** feature points, representing the predictive impact of the kernel rooted at the spatial point.

Fig. 5 compares the accuracy achieved with the full and diagonalized covariance estimates for varied feature point counts. With the same feature points, the full covariance

updates are more accurate than the diagonalized ones. But, the computational time with full covariance updates is an order of magnitude longer. Therefore, we recommend increasing the number of feature points over performing full covariance estimates for increasing predictive accuracy.

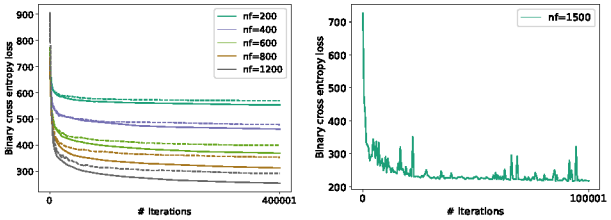

Fig. 5. Verification error during training process for increasing number of feature points with centralized full and diagonalized implementations. represented using solid and dashed lines respectively. Verification error in distributed diagonalized algorithm with **1500** feature points.

As seen in Fig. 6, we distribute a reduced dataset with 290k (out of 380k) sequential points across four agents, such that only their combined dataset has the complete map information. The agents communicate over a static connected graph in bottom-left of Fig. 6. The 1500 feature points and lengthscales $\gamma_2 = 0.5$ are selected at random from the test set as in the centralized setting, and these points are common across the agents. We achieve approximately 87% predictive accuracy on the same test set. Due to the presence of several agents, a quarter of iterations were sufficient to achieve this binary cross-entropy error as the centralized setting. The agents estimate similar mean values but their variances are lower for points close to the data collected.

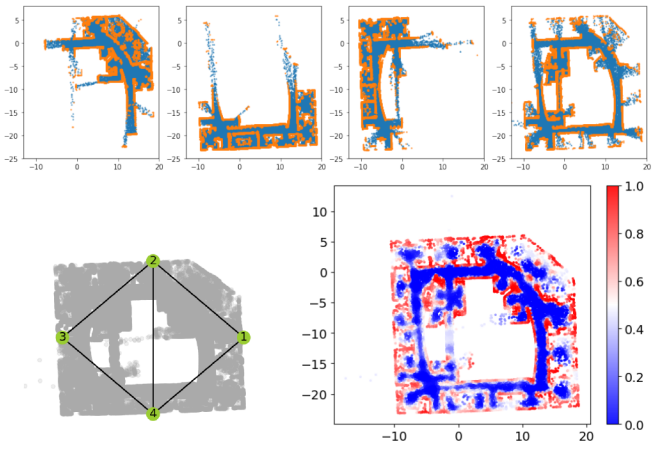

Fig. 6. Training data distributed among **4** agents sharing their inferences (top), Communication network, Occupancy probability indicating free and occupied spaces in blue and orange color respectively with a **1500** feature point model.

DiNNO dataset [49]: This dataset simulates LiDAR samples collected by multiple robots following independent trajectories with some overlap in observed environment. In contrast to Intel dataset where we separated the data into four sets, here the robots have pre-determined trajectories with minimal overlap in indoor space. The LiDAR distance data is converted to five free and occupied points as shown at the top of Fig. 7. The training set consists of a third of the dataset, an-eleventh for test set and an-eightieth for verification, chosen by slicing them along the trajectory. Each of the seven robots has roughly $90k$ training points, with $175k$ points in the test set. This dataset is challenging due to the low number of occupied points (10%) in comparison to the ones in free space. Therefore, we choose 300 feature points from the occupied space and remaining 700 randomly. Each kernel is defined with lengthscales γ_2 in $\{0.3, 3.\}$ depending on whether the data was chosen from occupied or free spaces respectively. The reconstruction of the indoor space using the diagonalized version of GVI from Proposition 5 is shown in Fig. 7.

The consensus error on the mean value of the parameters is computed as the deviation of the means $|\mu_{i,t}(\theta) - \frac{1}{n}\sum_{i=1}^{n}\mu_{i,t}(\theta)|$. We can see that this error decreases with the number of iterations, implying that agents learn a common estimate. During the training phase, prediction error is computed every 500 iterations on the verification set with $23k$ instances. The prediction error reaches a floor value over the $100k$ iterations for all agents.

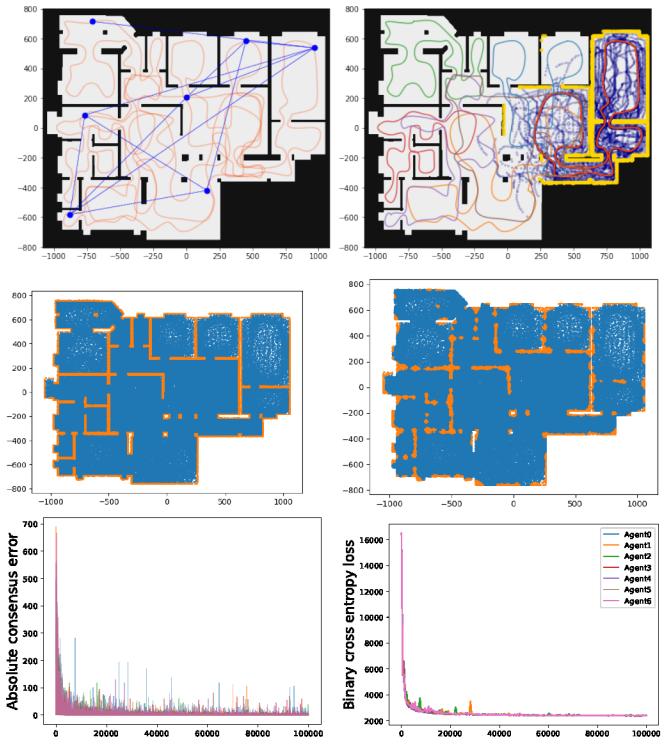

Fig. 7. Data obtained from the work in [49]. Top left: Communication network laid over the trajectories of **7** robots. Top right: Agent **1**'s LiDAR data with free and occupied point measurements highlighted in blue and yellow, respectively. Middle row: True and predicted point classes with a **1000** feature point model, with blue and orange dots corresponding to free and occupied spaces. Bottom left: Consensus error summed over parameters for each agent. Bottom right: Verification set error for each agent during training.

Successful training and deployment: The theoretical derivation of DELBO assumes independent observations at each agent. This does not hold for mapping data generated from robot trajectories. Therefore, we use independent samples from a replay buffer storing data over a short window. When generating points in free and occupied space from distance

measurements, one should balance the points in each class while covering the entire space. We maintain a $80-20$ ratio for the DiNNO dataset, more skewed than the Intel dataset.

Another key to building a good map is appropriate selection of feature points and lengthscales. The order of selected lengthscales should match the represented features. For instance, the occupied spaces in the map should be represented with lengthscales matching the expected obstacle width. In maps with several obstacle sizes, one could choose multiple kernels with varying lengthscales at the same feature points. Greater density of feature points allow a detailed representation of geometric map features. Selecting them from both occupied and free spaces allows better representation of each set. We selected 40% of feature points in the occupied set to afford a better predictive resolution for DiNNO dataset.

VII. Conclusion

Analogous to the evidence lower bound (ELBO) in variational inference, this paper derived a distributed evidence lower bound (DELBO) on the observation evidence in multi-agent estimation problems. Gaussian constrained optimization of the DELBO components across the agents led to a distributed variational inference algorithm. We derived a version of the algorithm with Gaussian variational distributions and applied it to multi-robot mapping problems using streaming range measurements. Our distributed VI algorithm handles any differentiable non-linear log-likelihoods modeling agent observations, making it a promising efficient approach to solving networked estimation problems with various machine learning models. A potential avenue for future work is to improve the communication efficiency of the algorithm by limiting the number of communication rounds and the number of actively communicating agents or by allowing agents to share relevant subsets of their local parameter estimates.

References

[1] M. Al-Shedivat, J. Gillenwater, E. Xing, and A. Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. *arXiv preprint arXiv:2010.05273*, 2020.

[2] T. D. Barfoot, J. R. Forbes, and D. J. Yoon. Exactly sparse Gaussian variational inference with application to derivative-free batch nonlinear state estimation. *Int. J. Rob. Res.*, 39(13):1473–1502, 2020.

[3] A. Bordes, S. Ertekin, J. Weston, L. Botton, and N. Cristianini. Fast kernel classifiers with online and active learning. *J. Mach. Learn. Res.*, 6(9), 2005.

[4] T. D. Bui, C. V. Nguyen, S. Swaroop, and R. E. Turner. Partitioned variational inference: A unified framework encompassing federated and continual learning. *arXiv preprint arXiv:1811.11206*, 2018.

[5] F. Bullo, J. Cortés, and S. Martínez. *Distributed Control of Robotic Networks*. Applied Mathematics Series. Princeton University Press, 2009. Electronically available at http://coordinationbook.info.

[6] X. Cao, T. Başar, S. Diggavi, Y. C. Eldar, K. B. Letaief, H. V. Poor, and J. Zhang. Communication-efficient distributed learning: An overview. *IEEE J. Sel. Areas Commun.*, 2023.

[7] Z. Chen et al. Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, 182(1):1–69, 2003.

[8] H. Dai, Y. Zhang, and J. Liu. Structured variational methods for distributed inference in networked systems: Design and analysis. *IEEE Trans. Signal Process.*, 61(15):3827–3839, 2013.

[9] J. Daunizeau. Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. *arXiv preprint arXiv:1703.00091*, 2017.

[10] J. Domke, R. Gower, and G. Garrigos. Provable convergence guarantees for black-box variational inference. *dv. Neural Inf. Process Syst.*, 36, 2024.

[11] T. Duong, M. Yip, and N. Atanasov. Autonomous navigation in unknown environments with sparse Bayesian kernel-based occupancy mapping. *IEEE Trans. Robot.*, 38(6):3694–3712, 2022.

[12] H. Fang, N. Tian, Y. Wang, M. Zhou, and M. A. Haile. Nonlinear bayesian estimation: From kalman filtering to a broader horizon. *IEEE/-CAA Journal of Automatica Sinica*, 5(2):401–417, 2018.

[13] D. Fink. A compendium of conjugate priors. Technical report, 1997.

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

[15] Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. *Adv. Neural Inf. Process Syst.*, 13, 2000.

[16] S. Gultekin and J. Paisley. Nonlinear Kalman filtering with divergence minimization. *IEEE Trans. Signal Process.*, 65(23):6319–6331, 2017.

[17] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 2013.

[18] A. Howard and N. Roy. The robotics data set repository (radish), 2003.

[19] J. Hua and C. Li. Distributed variational Bayesian algorithms over sensor networks. *IEEE Trans. Signal Process.*, 64(3):783–798, 2015.

[20] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Stat. Comput.*, 10(1):25–37, 2000.

[21] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, Sept. 2012.

[22] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

[23] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1–2):1–210, 2021.

[24] D. P. Kingma, M. Welling, et al. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019.

[25] J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *J. Mach. Learn. Res.*, 23(132):1–109, 2022.

[26] M. Lambert, S. Bonnabel, and F. Bach. The recursive variational Gaussian approximation (R-VGA). *Stat. Comput.*, 32(1):10, 2022.

[27] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Adv. Neural Inf. Process Syst.*, 30, 2017.

[28] J. Luttinen. Bayesian python: Bayesian inference tools for python, 2021.

[29] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *J. Mach. Learn. Res.*, 18:1–35, 2017.

[30] H. Mangesius, D. Xue, and S. Hirche. Consensus driven by the geometric mean. *IEEE Trans. Control Netw. Syst.*, 5(1):251–261, 2016.

[31] A. W. Max. Inverting modified matrices. In *Memorandum Rept. 42, Statistical Research Group*, page 4. Princeton Univ., 1950.

[32] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In A. Singh and J. Zhu, editors, *Proc. of the 20th Int. Conf. on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282. PMLR, 20–22 Apr 2017.

[33] A. Nedić, A. Olshevsky, and C. A. Uribe. Fast convergence rates for distributed non-Bayesian learning. *IEEE Trans. Autom. Control*, 62(11):5538–5553, 2017.

[34] H. Nickisch and C. E. Rasmussen. Approximations for binary Gaussian process classification. *J. Mach. Learn. Res.*, 9(Oct):2035–2078, 2008.

[35] V. M.-H. Ong, D. J. Nott, and M. S. Smith. Gaussian variational approximation with a factor covariance structure. *J. Comput. Graph. Stat.*, 27(3):465–478, 2018.

[36] P. Paritosh, N. Atanasov, and S. Martínez. Hypothesis assignment and partial likelihood averaging for cooperative estimation. In *IEEE Int. Conf. on Decision and Control*, pages 7850–7856, 2019.

[37] P. Paritosh, N. Atanasov, and S. Martinez. Marginal density averaging for distributed node localization from local edge measurements. In *IEEE Int. Conf. on Decision and Control*, pages 2404–2410. IEEE, 2020.

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

[39] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Int. Conf. on Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

[40] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. Int. Conf. Mach. Learn.*, pages 1278–1286. PMLR, 2014.

[41] T. Shankar and A. Gupta. Learning robot skills with temporal variational inference. In *Proc. Int. Conf. Mach. Learn.*, pages 8624–8633. PMLR, 2020.

[42] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pac. J. Math.*, 21(2):343–348, 1967.

[43] V. Smidl and A. Quinn. Variational Bayesian filtering. *IEEE Trans. Signal Process.*, 56(10):5020–5030, 2008.

[44] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

[45] C. A. Uribe, A. Olshevsky, and A. Nedić. Nonasymptotic concentration rates in cooperative learning–part i: Variational non-bayesian social learning. *IEEE Trans. Control Netw. Syst.*, 9(3):1128–1140, 2022.

[46] X. Wang, A. Lalitha, T. Javidi, and F. Koushanfar. Peer-to-peer variational federated learning over arbitrary graphs. *IEEE J. Sel. Areas Inf. Theory*, 2022.

[47] J. Winn, C. M. Bishop, and T. Jaakkola. Variational message passing. *J. Mach. Learn. Res.*, 6(4), 2005.

[48] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson. A survey of distributed optimization. *Annu. Rev. Control*, 47:278–305, 2019.

[49] J. Yu, J. A. Vincent, and M. Schwager. DiNNO: Distributed neural network optimization for multi-robot collaborative learning. *IEEE Robot. Autom. Lett.*, 7(2):1896–1903, 2022.

[50] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):2008–2026, 2018.

[51] X. Zhang, Y. Li, W. Li, K. Guo, and Y. Shao. Personalized federated learning via variational Bayesian inference. In *Proc. Int. Conf. Mach. Learn.*, pages 26293–26310. PMLR, 2022.

[52] S. Zhou and G. Y. Li. FedGiA: An efficient hybrid algorithm for federated learning. *IEEE Trans. Signal Process.*, 2023.

APPENDIX

A. Gaussian variational inference

Proposition 1. First, we discuss the derivation of the variational inference algorithm from the gradient descent steps in [2]. We start by defining the objective function τ based on the known Gaussian pdf $q_{t-1}(\theta) = \phi(\theta|\mu_{t-1}, \Omega_{t-1}^{-1})$ as,

$$\tau(\theta) = -\log \ell(z_t|\theta) - \log(q_{t-1}(\theta)).$$

Thus, the variational objective $V(q) = \mathbb{E}_q[\tau(\theta) + \log q(\theta)]$ is the negative of the ELBO defined in (2). Enforcing the first order optimality condition in [2, Eqn. (25)] to minimize $V(q)$,

$$\Omega_t = \mathop{\mathbb{E}}_{q_{t-1}}\left[\frac{\partial}{\partial\theta^\top\partial\theta}\tau(\theta)\right], \delta\mu = -\Omega_t^{-1}\mathop{\mathbb{E}}_{q_{t-1}}\left[\frac{\partial}{\partial\theta^\top}\tau(\theta)\right], \quad (22)$$

where $\delta\mu = \mu - \mu_{t-1}$. The derivative w.r.t. θ and their expectations w.r.t. the prior q_{t-1} becomes,

$$\begin{aligned}
&\frac{\partial}{\partial\theta^\top}\tau(\theta) = -\frac{\partial}{\partial\theta^\top}[\log \ell(z_t|\theta)] + (\theta - \mu_{t-1})^\top\Omega_{t-1},\\
&\mathop{\mathbb{E}}_{q_{t-1}}\left[\frac{\partial}{\partial\theta^\top}\tau(\theta)\right] = -\mathop{\mathbb{E}}_{q_{t-1}}\frac{\partial}{\partial\theta^\top}[\log \ell(z_t|\theta)]. \quad (23)\\
&\frac{\partial}{\partial\theta^\top\partial\theta}\tau(\theta) = -\frac{\partial}{\partial\theta^\top\partial\theta}[\log \ell(z_t|\theta)] + \Omega_{t-1}, \quad (24)\\
&\mathop{\mathbb{E}}_{q_{t-1}}\left[\frac{\partial}{\partial\theta^\top\partial\theta}\tau(\theta)\right] = \Omega_{t-1} - \mathop{\mathbb{E}}_{q_{t-1}}\frac{\partial}{\partial\theta^\top\partial\theta}[\log \ell(z_t|\theta)].
\end{aligned}$$

Thus, the updated mean and information matrix are given as,

$$\begin{aligned}
\mu_t &= \mu_{t-1} + \Omega_t^{-1}\mathop{\mathbb{E}}_{q_{t-1}}\left[\frac{\partial}{\partial\theta^\top}[\log \ell(z_t|\theta)]\right],\\
\Omega_t &= \Omega_{t-1} - \mathop{\mathbb{E}}_{q_{t-1}}\left[\frac{\partial}{\partial\theta^\top\partial\theta}[\log \ell(z_t|\theta)]\right].
\end{aligned} \quad (25)$$

This relates mean and covariance updates to the gradient and Hessian of the log-likelihood samples. □

Proposition 2. The proof to DGVI algorithm proceeds as Proposition 1, but derives the optimal variations in the agent hyperparameters. Our presentation begins with a concise description of the relevant results in [2], which apply to a centralized setting. We first define the function $\tau_i(\theta)$ based on the sampled likelihood and known neighbor prior pdfs $q_{j,t-1}(\theta) = \phi(\theta|\mu_{j,t-1}, \Omega_{j,t-1}^{-1})$ available at agent i as,

$$\tau_i(\theta) = -n\log \ell_i(z_{i,t}|\theta) - \log q_{i,t}^g(\theta),\ q_{i,t}^g = \prod_{j=1}^n q_{j,t-1}^{A_{ij}}. \quad (26)$$

Agent i minimizes the variational objective $V_i(q) = \mathbb{E}_q[\tau_i(\theta) + \log q(\theta)]$, that matches its negative DELBO component $-J_{i,t}$ defined in (8), to compute the optimal pdf $q_{i,t}$ at time t. The geometric average $q_{i,t}^g$ has mean and information matrix,

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

Next, we compute the derivatives of the variational objective to identify its optimizers. For any integrable function $\tau_i(\theta)$, the derivatives of $V_i(q)$ in terms of the mean in $q(\theta|\mu, \Omega^{-1})$ are,

$$\begin{aligned}
\frac{\partial V_i(q)}{\partial\mu^T} &= \Omega\mathbb{E}_q[(\theta - \mu)\tau_i(\theta)], \quad (27)\\
\frac{\partial^2 V_i(q)}{\partial\mu^T\partial\mu} &= \Omega\mathbb{E}_q[(\theta - \mu)(\theta - \mu)^T\tau_i(\theta)]\Omega - \Omega\mathbb{E}_q[\tau_i(\theta)].
\end{aligned}$$

The information matrix derivative $\frac{\partial V_i(q)}{\partial\Omega} = -\frac{1}{2}\mathbb{E}_q[(\theta-\mu)(\theta-\mu)^T\tau_i(\theta)]\Omega + \frac{1}{2}\Omega^{-1}\mathbb{E}_q[\tau_i(\theta)] + \frac{1}{2}\Omega^{-1}$ is given in terms of the second order mean derivative as,

$$\frac{\partial^2 V_i(q)}{\partial\mu^T\partial\mu} = \Omega - 2\Omega\frac{\partial V_i(q)}{\partial\Omega}\Omega. \quad (28)$$

Since setting these derivatives to zero does not yield closed form solutions for (μ, Ω), [2] performs a Taylor expansion of the variational objective at the prior and selects non-trivial variations $\delta\mu, \delta\Omega$ to ensure a locally decreasing objective. Similarly, we perform the approximation at the geometric average $q_{i,t}^g(\theta)$ of the neighbor priors, leading to,

$$\begin{aligned}
V_i(q) \approx V_i(q_{i,t}^g) &+ \left(\frac{\partial V_i(q)}{\partial\mu^T}\bigg|_{q_{i,t}^g}\right)\delta\mu \quad (29)\\
&+ \frac{1}{2}\delta\mu^T\left(\frac{\partial^2 V_i(q)}{\partial\mu^T\partial\mu}\bigg|_{q_{i,t}^g}\right)\delta\mu + \mathrm{tr}\left(\frac{\partial V_i(q)}{\partial\Omega}\bigg|_{q_{i,t}^g}\delta\Omega\right),
\end{aligned}$$

for $\delta\mu = \mu - \mu_{i,t}^g, \delta\Omega = \Omega - \Omega_{i,t}^g$. Setting the derivative w.r.t. Ω in (28) to zero, and solving the quadratic for $\delta\mu$ in (29) generates explicit values,

$$\Omega_{i,t} = \frac{\partial^2 V_i(q)}{\partial\mu^T\partial\mu}\bigg|_{q_{i,t}^g}, \left(\frac{\partial^2 V_i(q)}{\partial\mu^T\partial\mu}\bigg|_{q_{i,t}^g}\right)\delta\mu = -\frac{\partial V_i(q)}{\partial\mu^T}\bigg|_{q_{i,t}^g}. \quad (30)$$

Finally, we transform the gradients in terms of (μ, Ω) in (27) into those w.r.t. variables θ using Stein's lemma [44] as,

$$\begin{aligned}
\mathop{\mathbb{E}}_q[(\theta - \mu)\tau_i(\theta)] &\equiv \Omega\mathop{\mathbb{E}}_q\left[\frac{\partial\tau_i(\theta)}{\partial\theta^\top}\right] = \Omega\frac{\partial}{\partial\mu^\top}V_i(q), \quad (31)\\
\mathop{\mathbb{E}}_q[(\theta - \mu)(\theta - \mu)^\top\tau_i(\theta)] &\equiv \Omega\mathop{\mathbb{E}}_q\left[\frac{\partial^2\tau_i(\theta)}{\partial\theta^\top\partial\theta}\right]\Omega + \Omega\mathop{\mathbb{E}}_q[\tau_i(\theta)].
\end{aligned}$$

Based on their relation to gradients w.r.t. the mean μ in (27), we substitute them into (30) to obtain the optimal $(\mu_{i,t}, \Omega_{i,t})$

locally minimizing $V_i(q)$ for $\delta\mu = \mu_{i,t} - \mu^g_{i,t}$ as,

$$\mu_{i,t} - \mu^g_{i,t} = -\Omega_t^{-1} \mathop{\mathbb{E}}_{q^g_{i,t}} \left[\frac{\partial}{\partial \theta^\top} \tau_i(\theta) \right], \Omega_{i,t} = \mathop{\mathbb{E}}_{q^g_{i,t}} \left[\frac{\partial}{\partial \theta^\top \partial \theta} \tau_i(\theta) \right].$$

Since these equations mirror (22) in the proof to Proposition 1, the rest of the proof follows exactly with prior $q^g_{i,t}$ to yield,

$$\mu_{i,t} = \mu^g_{i,t} + n\Omega^{-1}_{i,t} \mathop{\mathbb{E}}_{q^g_{i,t}} \left[\frac{\partial}{\partial \theta^\top} [\log \ell(z_{i,t}|\theta)] \right],$$

$$\Omega_{i,t} = \Omega^g_{i,t} - n \mathop{\mathbb{E}}_{q^g_{i,t}} \left[\frac{\partial}{\partial \theta^\top \partial \theta} [\log \ell(z_{i,t}|\theta)] \right],$$

with additional multiple n on the log-likelihood. □

B. Gaussian Expectation of classification Model

Expected gradient in Proposition 3. From (13), the gradient of sigmoid function is, $\nabla_\theta \log \ell(z|\theta) = (y - \sigma(\Phi_x^\top \theta))\Phi_x^\top$. Its expected value with $q(\theta) \sim \mathcal{N}(\mu, \Sigma)$ follows from the expectation of the term $\sigma(\Phi_x^\top \theta)$. For this computation, we recall that the inverse probit function, or a cumulative distribution function defined as $\Gamma(\theta) = \int_{\alpha=-\infty}^{\theta} \phi(\alpha) d\alpha$. The cdf approximates the sigmoid function with the relationship $\sigma(\theta) = \Gamma(\xi\theta)$ for $\xi = 0.61$ [9]. To compute the approximation $\mathbb{E}_{q(\theta)}[\Gamma(\xi \Phi_x^\top \theta)]$, we substitute $u = \xi \Phi_x^\top \theta$ and express the cdf at u in terms of standard normal random variable Z as $\Gamma(u) = \mathbb{P}(Z \leq U | U = u)$. Therefore,

$$\mathop{\mathbb{E}}_{q(\theta)}[\Gamma(U)] = \mathop{\mathbb{E}}_{q(\theta)}[\mathbb{P}(Z \leq U | U = u)] = \mathbb{P}(Z - U \leq 0).$$

Since the variables Z, U are jointly Gaussian, and U is an affine transformation of θ, their pdf can be expressed as $Z - U = \phi(\cdot | -\xi \Phi_x^\top \mu, 1 + \xi^2 \Phi_x^\top \Sigma \Phi_x)$,

$$\mathbb{P}(Z - U \leq 0) = \Gamma\left((\xi \Phi_x^\top \mu) \Big/ \sqrt{1 + \xi^2 \Phi_x^\top \Sigma \Phi_x} \right)$$

With $\beta = 1 + \xi^2 \Phi_x^\top \Sigma \Phi_x$, the approximate expected value of the sigmoid function in the gradient defined in (13) is,

$$\mathop{\mathbb{E}}_{q_t(\theta)}[\sigma(\Phi_x^\top \theta)] \approx \int \Gamma(\xi \Phi_x^\top \theta) q_t(\theta) d\theta = \Gamma\left(\xi \Phi_x^\top \mu_t \Big/ \sqrt{\beta} \right).$$

Thus, the expected gradient of the log-likelihood is,

$$\mathop{\mathbb{E}}_{q_t}[(y - \sigma(\Phi_x^\top \theta))\Phi_x^\top] = \left(y - \Gamma\left(\xi \Phi_x^\top \mu_t \Big/ \sqrt{\beta} \right) \right) \Phi_x^\top.$$

□

Expected Hessian in Proposition 3. To find a tractable analytical expression for the new covariance matrix Ω^{-1}_{t+1}, We start by computing the expectation from (15),

$$\mathbb{E}_{q_t}[\phi(\xi \Phi_x^\top \theta | 0, 1)] = \sqrt{|\Omega_t| / (2\pi)^{l+1}} \exp(-0.5 \mu_t^\top \Omega_t \mu_t)$$
$$\int_\theta \exp(-0.5[\theta^\top (\Omega_t + \xi^2 \Phi_x \Phi_x^\top)\theta - 2\theta^\top \Omega_t \mu_t]) d\theta.$$

Proceeding with the sum of squares technique on the quadratic exponential argument,

$$\mathbb{E}_{q_t}[\phi(\xi \Phi_x^\top \theta | 0, 1)] = \sqrt{|\Omega_t| / (2\pi |\Omega_t + \xi^2 \Phi_x \Phi_x^\top|)}$$
$$\exp\left(-\frac{1}{2} [-\mu_t^\top \Omega_t^\top (\Omega_t + \xi^2 \Phi_x \Phi_x^\top)^{-1} \Omega_t \mu_t + \mu_t^\top \Omega_t \mu_t] \right).$$

Since computing the determinant and the inverse in the previous formula is expensive, we employ the matrix determinant lemma stating that $|\Omega_t + \xi^2 \Phi_x \Phi_x^\top| = (1 + \xi^2 \Phi_x^\top \Omega_t^{-1} \Phi_x) |\Omega_t|$.

$$\sqrt{|\Omega_t| / (2\pi |\Omega_t + \xi^2 \Phi_x \Phi_x^\top|)} = (2\pi (1 + \xi^2 \Phi_x^\top \Omega_t^{-1} \Phi_x))^{-0.5}.$$

The inverse of the dense matrix $(\Omega_t + \xi^2 \Phi_x \Phi_x^\top)^{-1}$ can be simplified using Woodbury's formula [31] such that we use the precomputed covariance matrix Ω_t^{-1} along with a scalar inverse. In batch settings, this inverse is over low dimensions in comparison to number of feature points l.

$$(\Omega_t + \xi^2 \Phi_x \Phi_x^\top)^{-1}$$
$$= \Omega_t^{-1} - \xi^2 \Omega_t^{-1} \Phi_x (1 + \xi^2 \Phi_x^\top \Omega_t^{-1} \Phi_x)^{-1} \Phi_x^\top \Omega_t^{-1}.$$

Substituting $\beta = 1 + \xi^2 \Phi_x^\top \Omega_t^{-1} \Phi_x$, the expected second order derivative is thus simplified as,

$$\mathop{\mathbb{E}}_{q_t}[\nabla^2_\theta \log p(z_t|\theta)] = -\sqrt{\xi^2/(2\pi\beta)} \mathrm{e}^{\left(-\frac{1}{2} [\frac{\xi^2}{\beta} \mu_t^\top \Phi_x \Phi_x^\top \mu_t] \right)} \Phi_x \Phi_x^\top.$$

Thus, the information matrix update will be linear. □

Proposition 4. The mean and covariance updates at any agent i follow from gradient and Hessians of the likelihood w.r.t. the mixed pdf $q^g_{i,t} = \prod_j q^{A_{ij}}_{j,t-1}$. A computationally cheap method to compute the inverse of information matrix Ω_t in the expression of the next mean value in (25) is derived from the matrix inversion lemma [31] as,

$$\Omega_t^{-1} = \Omega_{t-1}^{-1} - \gamma \Omega_{t-1}^{-1} \Phi_x (I + \gamma \Phi_x^\top \Omega_{t-1}^{-1} \Phi_x)^{-1} \Phi_x^\top \Omega_{t-1}^{-1}.$$

In a single agent setting, this avoids performing any matrix inverse after the initial step. □

C. Diagonal Gaussian derivation

Proof for Proposition 5. This proof mirrors the optimization of the agent-objective outlined in Proposition 2, with two key differences: (i) the derivatives are expressed in terms of the diagonal elements of the information matrix, and, (ii) a diagonal approximation is applied to the second-order Taylor expansion of the objective. Assume that the Gaussians $q(\mu, D)$ and $q^g_{i,t}(\mu^g_{i,t}, D^g_{i,t})$ have diagonalized information matrices with diagonal vectors $\Delta, \Delta^g_{i,t}$ whose k-th elements are $\Delta[k], \Delta^g_{i,t}[k]$. The geometric average $q^g_{i,t}$ is expressed in terms of prior neighbor estimates $q_{j,t-1}(\mu_{j,t-1}, D_{j,t-1})$ with elements of the mean $\mu^g_{i,t}[k] = (\Delta^g_{i,t}[k])^{-1} (\sum_{j \in \mathcal{V}} A_{ij} \Delta_{j,t-1}[k] \mu_{j,t-1}[k])$ and covariance $\Delta^g_{i,t} = \sum_{j \in \mathcal{V}} A_{ij} \Delta_{j,t-1}[k]$. With $\tau_i(\theta) = -n \log \ell_i(z|\theta) - \log q^g_{i,t}(\theta)$, the variational objective is,

$$V_i(q) = \mathop{\mathbb{E}}_q[\tau_i(\theta) + \log q(\theta)] = \frac{1}{2} \sum_{k=1}^{l} \log \Delta[k]$$
$$+ \int_\theta \tau_i(\theta) \prod_{k=1}^{l} \left(\frac{2\pi}{\Delta[k]} \right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \sum_{k=1}^{l} \Delta[k] (\theta[k] - \mu[k])^2 \right).$$

The elementwise derivatives of DELBO w.r.t. the mean and information matrix follow from (27), which relates the terms in Hessian w.r.t. the mean to that of the information matrix as,

$$\frac{\partial^2}{\partial \mu[k]^2} V_i(q) = -2(\Delta[k])^2 \frac{\partial}{\partial \Delta[k]} V_i(q) + \Delta[k].$$

Since $\frac{\partial}{\partial \Delta[k]} V_i(q) = 0$ for all k at the local optimum, the optimal information matrix $D_{i.t}$'s elements are,

$$\Delta_{i,t}[k] = \frac{\partial^2}{\partial \mu[k]^2} V_i(q) \Big|_{q^g_{i,t}}, \forall k \in \{1, \ldots, l\}. \tag{32}$$

As shown in [2], we express the Taylor approximation of function V_i at the geometric average $q^g_{i,t}$ in terms of vector differentials on mean $\delta\mu = \mu - \mu^g_{i,t}$ and information diagonal $\delta\Delta = \Delta - \Delta^g_{i,t}$ as,

$$V(q_{i,t}) \approx V_i(q^g_{i,t}) + \frac{\partial}{\partial\mu}V_i(q)\Big|_{q^g_{i,t}}\delta\mu + \frac{\partial}{\partial\Delta}V_i(q)\Big|_{q^g_{i,t}}\delta\Delta$$
$$+ \frac{1}{2}\delta\mu^\top \text{diag}\,\frac{\partial^2}{\partial\mu^\top\partial\mu}V_i(q)\Big|_{q^g_{i,t}}\delta\mu, \qquad \text{(Diagonal Hessian)}$$

where we approximate the quadratic coefficient $\frac{\partial^2}{\partial\mu^\top\partial\mu}V_i(q)\Big|_{q^g_{i,t}}$ with its diagonal matrix. The diagonal approximation of the Hessian matrix is appropriate if the underlying log-likelihood model $\log\ell_i(z|\theta)$ is almost linear in terms of parameters θ. Since the approximation is locally quadratic in $\delta\mu$, we find the optimal mean $\mu_{i,t}$ by setting its derivative in terms of $\delta\mu$ to zero. Recalling that $D_{i,t} = \text{diag}\,\frac{\partial^2 V_i(q)}{\partial\mu^\top\partial\mu}\Big|_{q^g_{i,t}}$ in (32), we obtain the linear system, $\delta\mu = \mu_{i,t} - \mu^g_{i,t} = D^{-1}_{i,t}(\frac{\partial V_i(q)}{\partial\mu^\top}\Big|_{q^g_{i,t}})$.

Similar to (31), we apply Stein's lemma [44] for the diagonalized covariance Guassian $q(\theta)$ to relate derivatives in terms of μ, Ω to that of θ, yielding the update rules,

$$D_{i,t} = \text{diag}\left(\mathop{\mathbb{E}}_{q^g_{i,t}}\left[\frac{\partial^2\tau_i(\theta)}{\partial\theta^\top\partial\theta}\right]\right),\ \mu_{i,t} - \mu^g_{i,t} = D^{-1}_{i,t}\mathop{\mathbb{E}}_{q^g_{i,t}}\left[\frac{\partial\tau_i(\theta)}{\partial\theta^\top}\right].$$

Using the simplification in Appendix A followed by expectation of the classification model in Appendix B and diagonalized $D^g_{i,t}$, we obtain the updates,

$$D_{i,t} = \text{diag}\left(D^g_{i,t} + \gamma\Phi_x\Phi_x^\top\right) = D^g_{i,t} + \gamma\text{diag}\left(\Phi_x\Phi_x^\top\right),$$
$$\mu_{i,t} - \mu^g_{i,t} \approx nD^{-1}_{i,t}(y - \Gamma\left(\xi\Phi_x^\top\mu^g_{i,t}/\sqrt{\beta}\right))\Phi_x^\top,$$

where, $\gamma = \sqrt{\frac{\xi^2}{2\pi\beta}}\exp\left(-\frac{1}{2}[\frac{\xi^2}{\beta}(\mu^g_{i,t})^\top\Phi_x\Phi_x^\top\mu^g_{i,t}]\right)$, with $\beta = 1 + \xi^2\Phi_x^\top(D^g_{i,t})^{-1}\Phi_x$ over data $z = (x, y)$. □

D. Distributed regression in Gaussian models

Let the linear regression model with parameters θ describe the relationship between input-output pairs $z = (x, y)$ at agent i be specified as the likelihood $\ell_i(z|\theta) \propto \exp(-0.5(y - \Phi_x^\top\theta)^\top S_i(y - \Phi_x^\top\theta))$, where S_i is positive definite. Following the steps for the classification problem, the log likelihood gradient and Hessian terms are,

$$\nabla_\theta \log p(z_i|\theta) = \Phi_x S_i(y - \Phi_x^\top\theta), \nabla^2_\theta \log p(z_i|\theta) = -\Phi_x S_i\Phi_x^\top.$$

The mixed Gaussian pdf $q^g_{i,t} = \mathcal{N}(\theta|\mu^g_{i,t}, \Sigma^g_{i,t})$ for regression follows from Proposition 4 with $\Sigma^g_{i,t} = (\Omega^g_{i,t})^{-1}$,

$$\Omega^g_{i,t} = \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}, \mu^g_{i,t} = (\Omega^g_{i,t})^{-1}\sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}\mu_{j,t-1}.$$

Then, we can follow (25) and Woodbury's matrix inversion lemma [31] w.r.t. $q^g_{i,t}$,

$$\Omega_{i,t} = \Omega^g_{i,t} - n\mathbb{E}_{q^g_{i,t}}[\nabla^2_\theta \log\ell_i(z|\theta)] = \Omega^g_{i,t} + n\Phi_x S_i\Phi_x^\top,$$
$$\Omega^{-1}_{i,t} = \Sigma^g_{i,t} - \Sigma^g_{i,t}\Phi_x((nS_i)^{-1} + \Phi_x^\top\Sigma^g_{i,t}\Phi_x)^{-1}\Phi_x^\top\Sigma^g_{i,t},$$
$$\mu_{i,t} = \mu^g_{i,t} + n(\Omega_{i,t})^{-1}(\Phi_x S_i^\top y - \Phi_x S_i\Phi_x^\top\mu^g_{i,t}).$$

Thus, we have distributed probabilistic updates on the parameters of the linear regression model.

Parth Paritosh is a Postdoctoral fellow at U.S. Army Combat Capabilities Development Command Army Research Laboratory (DEVCOM ARL). His doctoral research in distributed estimation algorithms was completed at the Mechanical and Aerospace Engineering Department at the University of California San Diego (UCSD). He received his M.S. degree in Mechanical Engineering from the Purdue University, United States in May 2017, and his B.Tech. degree in Mechanical Engineering (major) and Computer Science and Engineering (Minor) in May 2015. His research aims to advance robotic localization and inference capabilities, with a focus on 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.*