DISTRIBUTED VARIATIONAL INFERENCE FOR ONLINE SUPERVISED LEARNING

Parth Paritosh[*], **Nikolay Atanasov**[†] **& Sonia Martínez**[†]
[*]DEVCOM US Army Research Laboratory, [†]Contextual Robotics Institute, UC San Diego
Funded by NSF, ARL & ONR

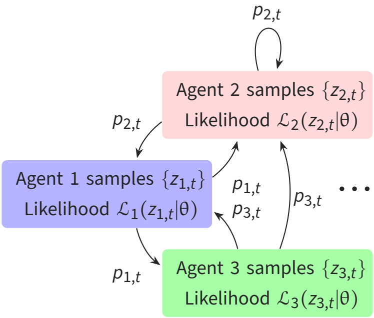
IEEE CDC 2025 – Rio de Janeiro

Available at `pptx.github.io/pparitosh/publications.html`

THE NEED FOR REAL-TIME DISTRIBUTED INFERENCE

- Heterogeneous sensing
- Distributed, private data collection
- Data sharing is expensive
- Online inference
- Quantified uncertainty in inference
- Computation is spread out

REAL-TIME DISTRIBUTED BAYESIAN INFERENCE



- Non-linear heterogeneous likelihoods
- Distributed communication
- Online probabilistic inference

Goal: Design a distributed real-time approximate inference algorithm for learning probability density function $p(\theta)$ over unknown θ.

VARIATIONAL INFERENCE: BACKGROUND

- Bayes' rule: Posterior on θ satisfies $p(\theta|z_{\leq t}) = \dfrac{\overbrace{\mathcal{L}(z_t|\theta)}^{\text{Likelihood}}\ \overbrace{p(\theta|z_{<t})}^{\text{Prior}}}{\underbrace{p(z_t|z_{<t})}_{\text{Normalization factor}}}$
- Computing normalization factor is intractable (unless conditionally conjugate)
- Approximate posterior via a variational family of distributions $q(\theta) \in \mathcal{F}$
- Maximize Evidence Lower Bound (ELBO) on the normalization factor,

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

- In recursive settings, replace prior $p(\theta|z_{<t})$ with $q_{t-1}(\theta)$

VARIATIONAL INFERENCE: BACKGROUND

- Bayes' rule: Posterior on θ satisfies $p(\theta|z_{\leq t}) = \dfrac{\overbrace{\mathcal{L}(z_t|\theta)}^{\text{Likelihood}}\ \overbrace{p(\theta|z_{<t})}^{\text{Prior}}}{\underbrace{p(z_t|z_{<t})}_{\text{Normalization factor}}}$
- Computing normalization factor is intractable (unless conditionally conjugate)
- Approximate posterior via a variational family of distributions $q(\theta) \in \mathcal{F}$
- Maximize Evidence Lower Bound (ELBO) on the normalization factor,

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

- In recursive settings, replace prior $p(\theta|z_{<t})$ with $q_{t-1}(\theta)$

DISTRIBUTED VARIATIONAL INFERENCE

THEOREM: DISTRIBUTED EVIDENCE LOWER BOUND (DELBO)

Assuming:

- Connected network,
- Independent observations z_i from likelihoods $\mathcal{L}_i(z_i|\theta)$,
- Approximate agent PDFs $q_i(\theta) = q_t(\theta)$ for some PDF $q_t(\theta)$,

the separable distributed evidence lower bound (DELBO) on the normalization factor is,

$$p(z_t|z_{<t}) \geq J_t[q_1, \ldots, q_n] = \sum_{i \in \mathcal{V}} J_{i,t}[q_i],$$

$$J_{i,t}[q_i] = \mathop{\mathbb{E}}_{q_i(\theta)}\left[\mathcal{L}_i(z_{i,t}|\theta) - \frac{1}{n}\log(q_i(\theta)) + \sum_{j \in \mathcal{V}} \frac{A_{ij}}{n} \log p_j(\theta|z_{<t})\right],$$

where A is the adjacency matrix representing connected networks.

THEOREM: DISTRIBUTED EVIDENCE LOWER BOUND (DELBO)

Assuming:

- Connected network,
- Independent observations z_i from likelihoods $\mathcal{L}_i(z_i|\theta)$,
- Approximate agent PDFs $q_i(\theta) = q_t(\theta)$ for some PDF $q_t(\theta)$,

the separable distributed evidence lower bound (DELBO) on the normalization factor is,

$$p(z_t|z_{<t}) \geq J_t[q_1, \ldots, q_n] = \sum_{i \in \mathcal{V}} J_{i,t}[q_i],$$

$$J_{i,t}[q_i] = \mathop{\mathbb{E}}_{q_i(\theta)}\left[\mathcal{L}_i(z_{i,t}|\theta) - \frac{1}{n}\log(q_i(\theta)) + \sum_{j \in \mathcal{V}} \frac{A_{ij}}{n} \log p_j(\theta|z_{<t})\right],$$

where A is the adjacency matrix representing connected networks.

GAP BETWEEN NORMALIZATION FACTOR AND DELBO

The gap between $p(z_t|z_{<t})$ and $J_t[q_1, \ldots, q_n]$ decomposes into:

1. Distributed model error:

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}[q_{i,t}(\theta) \| p(\theta|z_{\leq t})]$$

   Divergence of the approximated local posterior $q_{i,t}$ from the truth.

2. Consensus error:

$$\frac{1}{n}\sum_{i\in\mathcal{V}} \mathrm{KL}[p_g \| p_i(\theta|z_{<t})], \quad p_g = \frac{\prod_i p_i(\theta|z_{<t})^{1/n}}{\int \prod_i p_i(\theta|z_{<t})^{1/n} d\theta}$$

   Disagreement between true local posteriors and their geometric average.

OPTIMIZING DELBO TO COMPUTE VARIATIONAL DENSITIES

- Replace neighbor priors $p_j(\theta|z_{<t})$ with approximations $q_{j,t-1}(\theta)$
- Optimize each component of the separable objective $J_{i,t}[q_i]$,

$$q_{i,t}(\theta) \in \arg\max_{q_i} \mathbb{E}_{q_i}\left[n\,\mathcal{L}_i(z_{i,t}|\theta) - \log(q_i(\theta)) + \sum_{j\in\mathcal{V}} A_{ij} \log q_{j,t-1}(\theta)\right]$$

- Optimal PDF for agent i is $q_{i,t}(\theta) \propto \mathcal{L}_i(z_{i,t}|\theta)^n q_i^g(\theta) \in \arg\max_{q_i} J_{i,t}[q_i]$
  - Mixed PDF $q_i^g(\theta) \propto \prod_{j\in\mathcal{V}_i} q_{j,t-1}(\theta)^{A_{ij}}$ with likelihood exponent n.

How to handle non-conditionally conjugate likelihoods?

Approximate Gaussian variational densities for differentiable likelihoods

OPTIMIZING DELBO TO COMPUTE VARIATIONAL DENSITIES

- Replace neighbor priors $p_j(\theta|z_{<t})$ with approximations $q_{j,t-1}(\theta)$
- Optimize each component of the separable objective $J_{i,t}[q_i]$,

$$q_{i,t}(\theta) \in \arg\max_{q_i} \mathop{\mathbb{E}}_{q_i}\left[n\,\mathcal{L}_i(z_{i,t}|\theta) - \log(q_i(\theta)) + \sum_{j\in\mathcal{V}} A_{ij} \log q_{j,t-1}(\theta) \right]$$

- Optimal PDF for agent i is $q_{i,t}(\theta) \propto \mathcal{L}_i(z_{i,t}|\theta)^n q_i^g(\theta) \in \arg\max_{q_i} J_{i,t}[q_i]$
  - Mixed PDF $q_i^g(\theta) \propto \prod_{j\in\mathcal{V}_i} q_{j,t-1}(\theta)^{A_{ij}}$ with likelihood exponent n.

How to handle non-conditionally conjugate likelihoods?

Approximate Gaussian variational densities for differentiable likelihoods

DISTRIBUTED GAUSSIAN VARIATIONAL INFERENCE

DISTRIBUTED GAUSSIAN VARIATIONAL INFERENCE (DGVI)

At agent i and time t, given:

- observation $z_{i,t}$ with likelihood $\mathcal{L}_i(z_{i,t}|\theta)$,
- neighbor estimates $q_{j,t-1}(\theta) = \mathcal{N}(\theta|\mu_{j,t-1}, \Omega^{-1}_{j,t-1})$,
- Neighbor weights in communication matrix A,

Mean $\mu_{i,t}$ and information matrix $\Omega_{i,t}$ of the DELBO minimizing PDF $q_{i,t}$ are,

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

$$\Omega_{i,t} = \Omega^g_{i,t} - n\,\mathbb{E}_{q^g_{i,t}}[\nabla^2_\theta \log \mathcal{L}_i(z_{i,t}|\theta)],$$

$$\mu_{i,t} = \mu^g_{i,t} + n(\Omega^g_{i,t})^{-1}\mathbb{E}_{q^g_{i,t}}[\nabla_\theta \log \mathcal{L}_i(z_{i,t}|\theta)].$$

ADAPTING DGVI TO SUPERVISED LEARNING

Problem: Approximate $\mathbb{E}_{q^g_{i,t}}\left[\nabla_\theta \log \mathcal{L}(z_{i,t}|\theta)\right]$ for real-time computation:

- Agent likelihoods generated by kernel-based classifiers/regressors
- Expectation approximated w.r.t. the mixed Gaussian PDF:

$$q^g_{i,t} = \phi(\theta|\mu^g_{i,t}, (\Omega^g_{i,t})^{-1})$$

CLASSIFICATION MODEL

- Observed data $z = (x, y)$ with input $x \in \mathbb{R}^d$ and label $y \in \{0, 1\}$
- Model features $\Phi_x \in \mathbb{R}^{l+1}$ with kernel elements:
  $\Phi_x = [1, k_1(x), \ldots, k_l(x)], k_s(x) = \exp(-\gamma \|x - x^{(s)}\|^2)$
- Agent likelihood model with parameters θ and sigmoid function σ:

$$\mathcal{L}(z|\theta) = \sigma(\Phi_x^\top \theta)^y (1 - \sigma(\Phi_x^\top \theta))^{1-y}$$

DGVI FOR KERNEL CLASSIFICATION

For agent i's observation $z = (x, y)$ with classification likelihood, and neighbor estimates $\phi(\theta|\mu_{j,t}, \Omega_{j,t}^{-1})$,

the mean and information matrix $\mu_{i,t}, \Omega_{i,t}$ of DELBO maximizing PDF $q_{i,t}$ is,

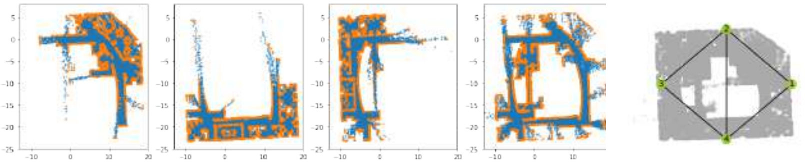
$$\Omega_{i,t}^{g} = \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1},\ \Omega_{i,t}^{g}\mu_{i,t}^{g} = \sum_{j\in\mathcal{V}} A_{ij}\Omega_{j,t-1}\mu_{j,t-1}, \Sigma_{i,t}^{g} = (\Omega_{i,t}^{g})^{-1}$$

$$\Omega_{i,t} = \Omega_{i,t}^{g} + \gamma\Phi_x\Phi_x^{\top}, \Omega_{i,t}^{-1} = \Sigma_{i,t}^{g} - \frac{\gamma}{\gamma_1}\Sigma_{i,t}^{g}\Phi_x\Phi_x^{\top}\Sigma_{i,t}^{g}$$

$$\mu_{i,t} = \mu_{i,t}^{g} + \left(y - \Gamma\left(\xi\Phi_x^{\top}\mu_{i,t}^{g}/\sqrt{\beta}\right)\right)\Omega_{i,t}^{-1}\Phi_x$$

with unit normal cdf Γ, $\beta = 1 + \xi^2\Phi_x^{\top}(\Omega_{i,t}^{g})^{-1}\Phi_x$, $\gamma_1 = 1 + \gamma\Phi_x^{\top}(\Omega_{i,t}^{g})^{-1}\Phi_x$ and $\gamma = \sqrt{\frac{\xi^2}{2\pi\beta}}\exp\left(-0.5[\frac{\xi^2}{\beta}(\mu_{i,t}^{g})^{\top}\Phi_x\Phi_x^{\top}\mu_{i,t}^{g}]\right)$.

DISTRIBUTED MAPPING WITH INTEL LIDAR DATASET [1]



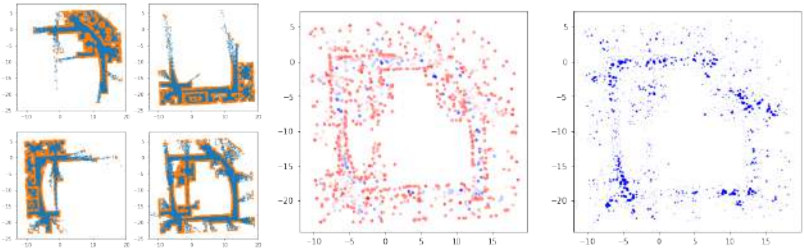
Training data distributed among 4 agents sharing their inferences, Communication network.

- Observed data $z = (x, y)$ with position $x \in \mathbb{R}^2$ and occupancy label $y \in \{0, 1\}$
- Model features $\Phi_x \in \mathbb{R}^{l+1}$ with kernels: $\Phi_x = [1, k_1(x), \ldots, k_l(x)]$
- Kernel $k_s(x) = \exp(-\gamma \|x - x^{(s)}\|^2)$ centered at $x^{(s)}$ with lengthscale γ
- Agent likelihood model with parameters θ and sigmoid function σ:

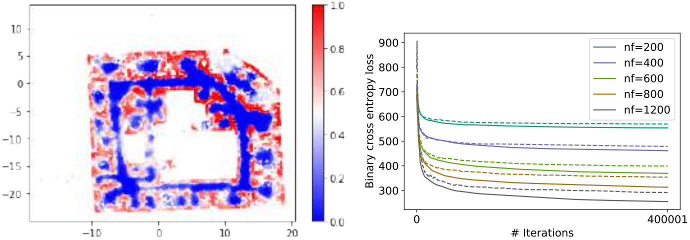
$$\mathcal{L}(z|\theta) = \sigma(\Phi_x^\top \theta)^y (1 - \sigma(\Phi_x^\top \theta))^{1-y}$$

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

DISTRIBUTED MAPPING WITH INTEL LIDAR DATASET

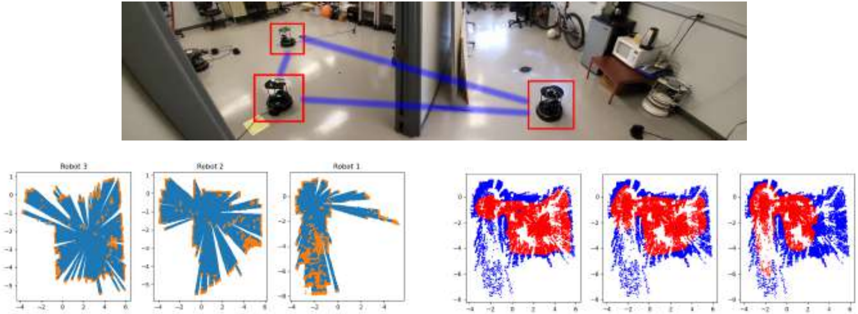


Training data sampled by 4 cooperative mapping agents, Estimated mean μ_T and variance Σ_T of the parameter θ on 1200 feature points $x^{(s)}$.



Free and occupied spaces with a 1500 features model. Verification loss with diagonalized covariances.

IMPLEMENTATION: DISTRIBUTED MAPPING WITH TURTLEBOTS



Indoor lab space with directed communication (top), Collected training data and predicted maps by the 3 Turtlebots (bottom). [2]

[2] Source code available at github.com/pptx/distributed-mapping

CONTRIBUTIONS

- Devise a separable version of evidence lower bound for inference
- Distributed Gaussian updates with tractable expectation terms in supervised learning setting
- Simulation and implementation for distributed robot mapping

Publications:

- Parth Paritosh, Nikolay Atanasov, and Sonia Martínez, “Distributed Variational Inference for Online Supervised Learning,” IEEE Transactions on Control of Network Systems, vol. 12, no. 3, pp. 1843–1855, 2025.
- P. Paritosh, S. Lau, N. Atanasov and S. Martinez. Distributed Variational Inference for Online Estimation: A Distributed Mapping Implementation on Turtlebot4s. Poster at Southern California Robotics Symposium 2023.

CONTRIBUTIONS

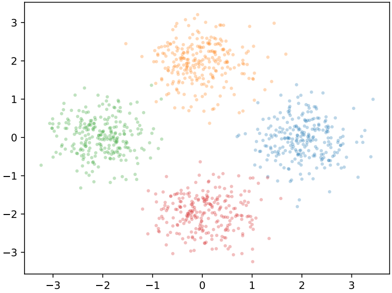
- Devise a separable version of evidence lower bound for inference
- Distributed Gaussian updates with tractable expectation terms in supervised learning setting
- Simulation and implementation for distributed robot mapping

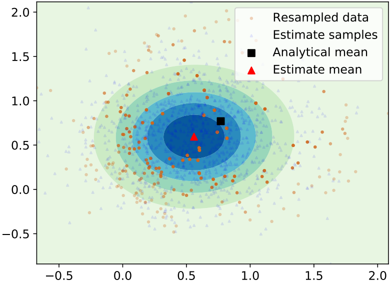
Publications:

- Parth Paritosh, Nikolay Atanasov, and Sonia Martínez, “Distributed Variational Inference for Online Supervised Learning,” IEEE Transactions on Control of Network Systems, vol. 12, no. 3, pp. 1843–1855, 2025.
- P. Paritosh, S. Lau, N. Atanasov and S. Martinez. Distributed Variational Inference for Online Estimation: A Distributed Mapping Implementation on Turtlebot4s. Poster at Southern California Robotics Symposium 2023.

Thank You!

A CASE STUDY: GEOMETRIC MIXING VIA SAMPLING





(a) Samples from Gaussian priors $p_{i,0}$ with unit covariance and means on a circle of radius 1.

(b) Comparing analytical mean and PDF estimated via particles resampled w.r.t. probability weights A_{ij} for data $z_{1,1} = [1, 1]$, .

- Bayesian update with mixed PDF $q_i^g(\theta) = \prod_{j \in \mathcal{V}_i} q_{j,t-1}(\theta)^{A_{ij}}$:

$$q_{i,t}(\theta) = \mathcal{L}_i(z_{i,t}|\theta)^n q_i^g(\theta) \Big/ \int \mathcal{L}_i(z_{i,t}|\theta)^n q_i^g(\theta) d\theta$$

Return to main slide

DISTRIBUTED REGRESSION MODEL

- Agent i samples observation $z_i = (x, y)$ from:

$$\mathcal{L}_i(z_i|\theta) \propto \exp(-0.5(y - \Phi_x^\top \theta)^\top S_i (y - \Phi_x^\top \theta)),$$

- Assuming a linear model $y = \Phi_x^\top \theta$ with feature vector:

$$\Phi_x = [1, k_1(x), \ldots, k_l(x)]$$

with elements $k_m(x)$ defined in Classification model (‣ slide).

DISTRIBUTED REGRESSION VIA VARIATIONAL INFERENCE

Proposition (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,*

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

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

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

$$\mu_{i,t} = \mu_{i,t}^{g} + n(\Omega_{i,t})^{-1}(\Phi_x S_i^{\top} y - \Phi_x S_i \Phi_x^{\top}\mu_{i,t}^{g}) \tag{2}$$

‣ Return to main slide