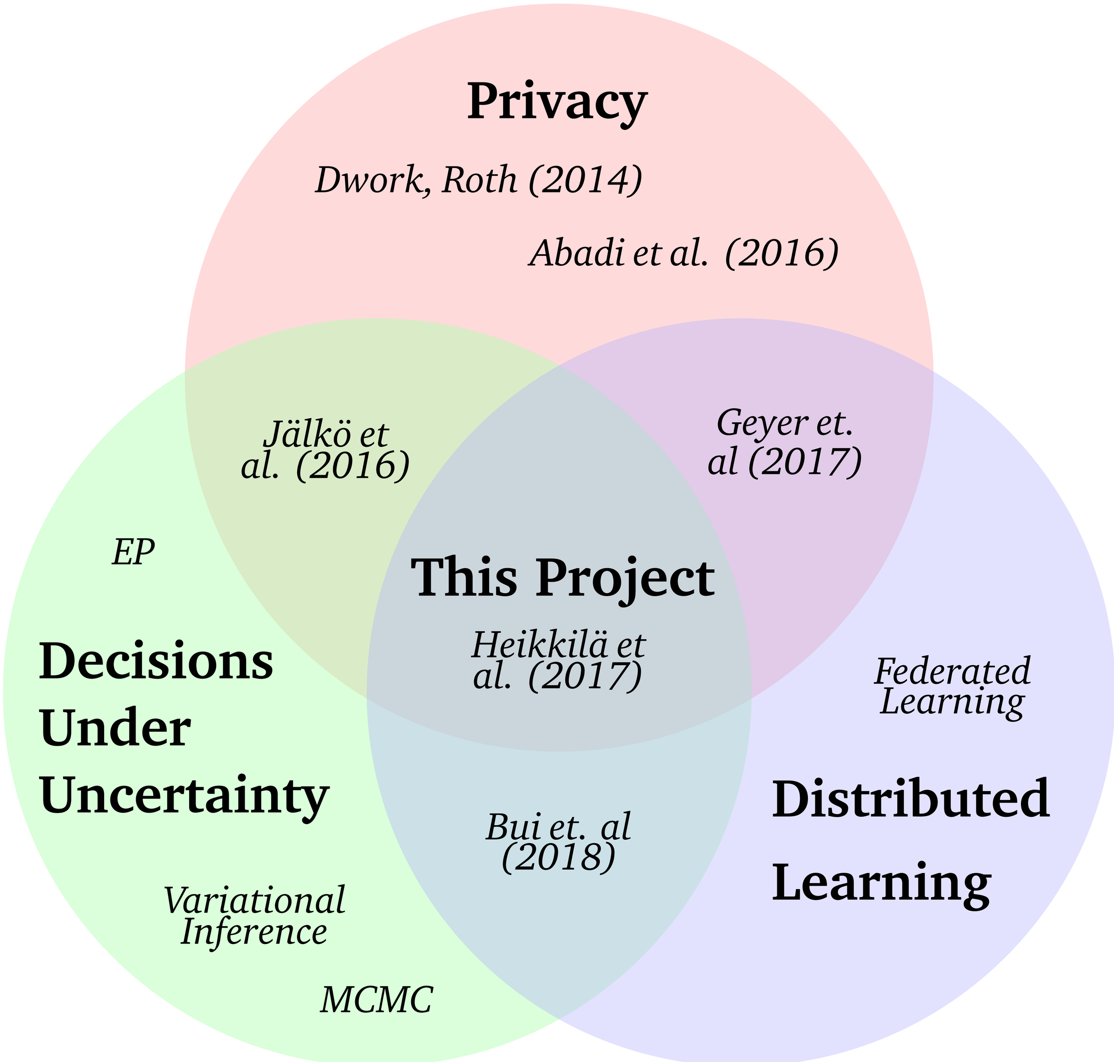
Differentially Private Federated Variational Inference

Mrinank Sharma,[1*] Michael Hutchinson,[1*] Siddharth Swaroop,[2] Antti Honkela,[3] Richard E. Turner[2]
[1] University of Oxford, UK [2] University of Cambridge, UK [3] University of Helsinki, Finland
[*] Work done whilst at the Univeristy of Cambridge. Equal Contribution. Correspondence to Mrinank Sharma <mrinank@robots.ox.ac.uk>.

Summary

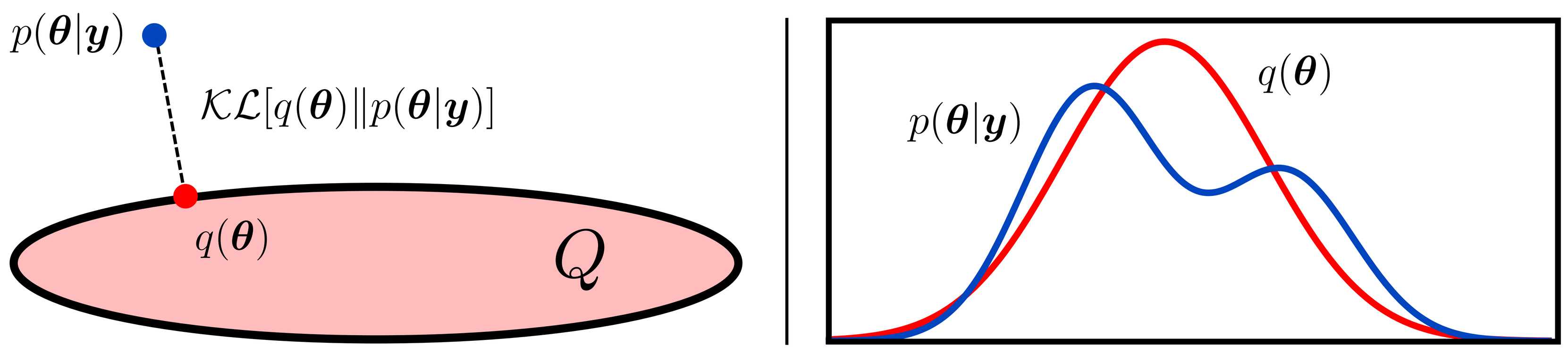
Problem. Perform (approximate) probabilistic inference on distributed data whilst respecting the privacy of individual clients.
Proposal. Combine *Partitioned Variational Inference* (PVI) with *Differentially Private* (DP) client side optimisation.
Results. Learn strongly private logistic regression models in the federated setting which achieves similar performance to non-private centralized training.



Variational Inference

Uncertainty is essential for optimal decision making, but often performing inference is intractable. *Variational Inference (VI)* **approximates the posterior with a simpler variational distribution**, $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, with $\boldsymbol{\lambda}$ chosen to maximise $\mathcal{F}(\boldsymbol{\theta})$, the **Free Energy**.

$$\begin{aligned}\mathcal{F}(\boldsymbol{\theta}) &= \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \log p(\boldsymbol{y}) - \mathcal{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\boldsymbol{y}))\end{aligned} \tag{1}$$



Partitoned Variational Inference (PVI)

The data is now partitioned across M clients i.e., $\boldsymbol{y} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_M\}$.
We change our **Variational Distribution** to match this:

$$q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{m=1}^{M} t_m(\boldsymbol{\theta}) \simeq \underbrace{\frac{p(\boldsymbol{\theta})}{\mathcal{Z}} \prod_{m=1}^{M} p(\boldsymbol{y}_m|\boldsymbol{\theta})}_{p(\boldsymbol{\theta}|\boldsymbol{y})}. \tag{2}$$

We minimise the ***Local*** **Free Energy:**

$$\begin{aligned}\mathcal{F}_m^{(i)}(q(\boldsymbol{\theta})) &= \int q(\boldsymbol{\theta}) \log \frac{1}{q(\boldsymbol{\theta})} \overbrace{\frac{q^{(i-1)}(\boldsymbol{\theta}) p(\boldsymbol{y}_m|\boldsymbol{\theta})}{t_m^{(i-1)}(\boldsymbol{\theta})}}^{\mathcal{Z}'\hat{p}(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \log \mathcal{Z}' - \mathcal{KL}(q(\boldsymbol{\theta}) || \hat{p}(\boldsymbol{\theta})),\end{aligned} \tag{3}$$

$\hat{p}(\boldsymbol{\theta})$ is known as the *titled distribution*. At each iteration, we update each client.

$$q_m^{(i)}(\boldsymbol{\theta}) = \arg\min \mathcal{F}_m^{(i)}(q(\boldsymbol{\theta})), \quad t_m^{(i)}(\boldsymbol{\theta}) = \frac{q_m^i(\boldsymbol{\theta})}{q^{(i-1)}(\boldsymbol{\theta})} t_m^{(i-1)}(\boldsymbol{\theta})$$

Any fixed point of PVI is a fixed point of global VI.

DP-PVI

1: **Input:** Clients $\{\boldsymbol{y}_m\}_{m=1}^M$, where $\boldsymbol{y}_m = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{N_m}$.
2: **Parameters:** minibatch size L, gradient norm bound C, noise scale σ.
3: Within each client, having received $q^{\text{old}}(\boldsymbol{\theta})$ from the server, optimize:

$$q_m^{\text{new}}(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{Q}}{\arg\min} \ \mathcal{KL}\Big(q(\boldsymbol{\theta}) || \frac{1}{\mathcal{Z}'} \frac{q^{\text{old}}(\boldsymbol{\theta})}{t_m^{\text{old}}(\boldsymbol{\theta})} p(\boldsymbol{y}_m|\boldsymbol{\theta})\Big). \tag{4}$$

This optimisation is done via Adagrad. At each iteration t, use the Gaussian Mechanism on the minibatch gradient, subsampling a minibatch of size L (denoted as $\mathcal{L}$):

$$\tilde{\boldsymbol{g}}_t = \frac{1}{L}\left[\sum_{i \in \mathcal{L}} \frac{\boldsymbol{g}(\boldsymbol{x}_i)}{\max\left(1, \frac{\|\boldsymbol{g}(\boldsymbol{x}_i)\|_2}{C}\right)} + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})\right]. \tag{5}$$

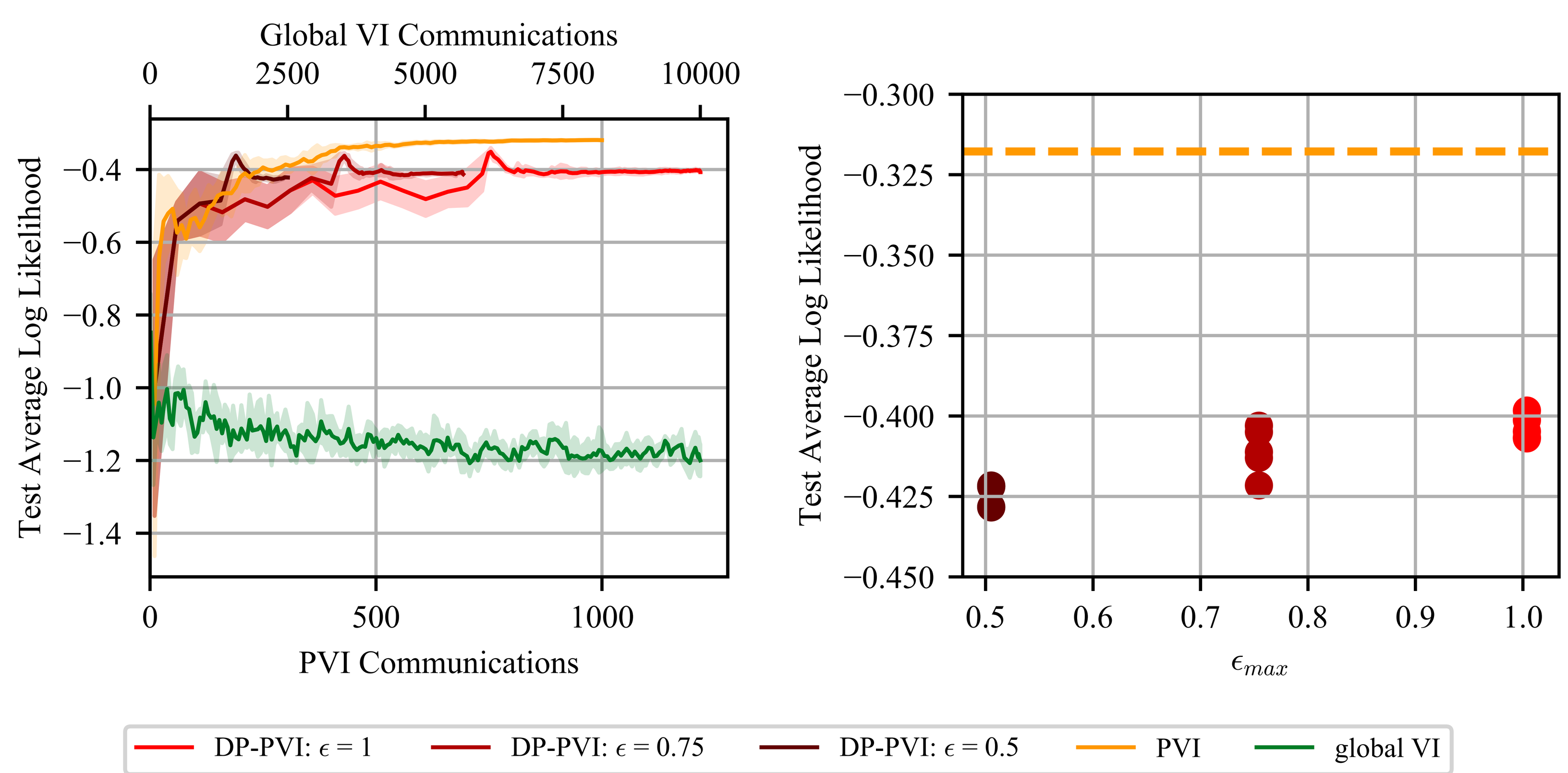
4: After optimisation, communicate to the global server:

$$\Delta t_m(\boldsymbol{\theta}) = \frac{t_m^{\text{new}}(\boldsymbol{\theta})}{t_m^{\text{old}}(\boldsymbol{\theta})} = \frac{q_m^{\text{new}}(\boldsymbol{\theta})}{q^{\text{old}}(\boldsymbol{\theta})}. \tag{6}$$

5: The global server updates $q(\boldsymbol{\theta}) \leftarrow q^{old}(\boldsymbol{\theta}) \Delta t_m(\boldsymbol{\theta})$.

Results

Mean-field Bayesian logistic regression, $M = 10$ clients on UCI Adult. **Imbalanced client data-set sizes and class imbalance** on the dataset distribution. **Asynchronous Setting**.



Conclusions & Future Work

- First-of-its-kind method for **private, federated, Bayesian ML**.
- Similar performance to PVI whilst achieving strong privacy guarantees.
- Significantly outperforms non-private VI.

Client Level Privacy

Often clients hold data about themselves only. This setting requires *client level differential privacy*, where neighbouring datasets are those which differ by an entire client.