Continual Learning via Sequential Function-Space Variational Inference

Tim G. J. Rudner
@timrudner

Freddie Bickford Smith
@fbickfordsmith

Qixuan Feng
@qixuan_feng

Yee Whye Teh
@yeewhye

Yarin Gal
@yaringal

International Conference on Machine Learning 2022

Correspondence to
tim.rudner@cs.ox.ac.uk

Paper: http://timrudner.com/sfsvi

Code: http://timrudner.com/sfsvi-code

OATML

UNIVERSITY OF OXFORD

Continual Learning

Goal:

- **Adapt** to new tasks **without forgetting** previously **attained abilities**.

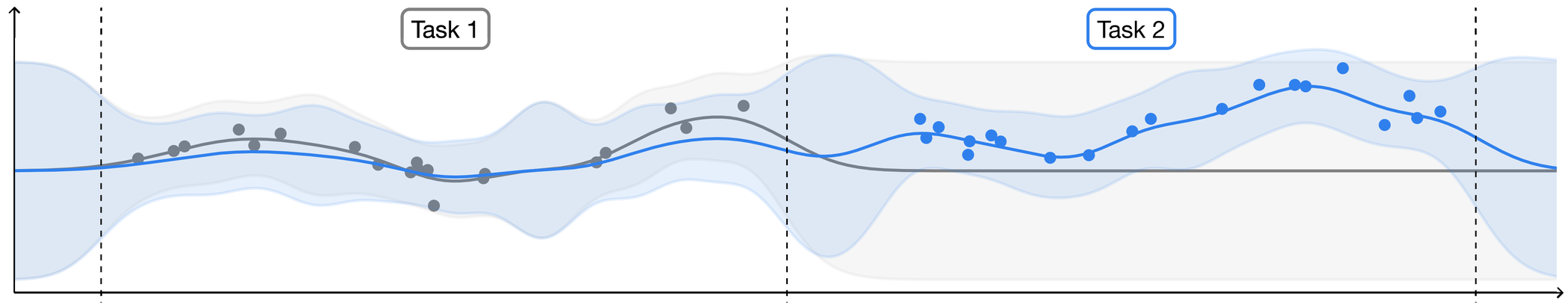
In practice:

- Learn to make accurate *predictions* for inputs associated with new tasks without "un-learning" to make accurate predictions on inputs associated with previous tasks.

Continual Learning via Function-Space Variational Inference

Key ideas

1. Learning new tasks by selecting "good" functions from a distribution over functions
2. Incorporating prior knowledge via **prior distributions over functions**



Sequential Bayesian Inference

- **Posterior distribution over stochastic functions:**

$$p(f|\mathcal{D}_1, \ldots, \mathcal{D}_t) \propto p(\mathcal{D}_t|f)p(f|\mathcal{D}_1, \ldots, \mathcal{D}_{t-1})$$

- **Sequential variational problem**:

$$\Longleftrightarrow \qquad \min_{q_t(f)\in\mathcal{Q}_f} \left\{\mathbb{D}_{\mathrm{KL}}\left(q_t(f)\|p_t\left(f \mid \mathcal{D}_1, \ldots, \mathcal{D}_t\right)\right)\right\}$$

$$\max_{q_t(f)\in\mathcal{Q}_f} \left\{\mathbb{E}_{q_t(f)}\left[\log p\left(\mathbf{y}_t \mid f\left(\mathbf{X}_t\right)\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q_t(f)\|p_t\left(f \mid \mathcal{D}_1, \ldots, \mathcal{D}_{t-1}\right)\right)\right\}$$

Sequential Bayesian Inference

- **Posterior distribution over stochastic functions:**

$$p(f|\mathcal{D}_1, \ldots, \mathcal{D}_t) \propto p(\mathcal{D}_t|f)p(f|\mathcal{D}_1, \ldots, \mathcal{D}_{t-1})$$

- **Sequential variational problem**:

$$\Longleftrightarrow \qquad \min_{q_t(f)\in\mathcal{Q}_f} \left\{\mathbb{D}_{\mathrm{KL}}\left(q_t(f)\|p_t\left(f \mid \mathcal{D}_1, \ldots, \mathcal{D}_t\right)\right)\right\}$$

$$\max_{q_t(f)\in\mathcal{Q}_f} \left\{\mathbb{E}_{q_t(f)}\left[\log p\left(\mathbf{y}_t \mid f\left(\mathbf{X}_t\right)\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q_t(f)\|q_{t-1}\left(f\right)\right)\right\}$$

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **Distributions over functions** induced by distributions over parameters:

$$p(\boldsymbol{\theta}) \Longrightarrow p(f(\cdot\,;\boldsymbol{\theta})) \qquad q(\boldsymbol{\theta}) \Longrightarrow q(f(\cdot\,;\boldsymbol{\theta}))$$

- **Variational problem**:

$$\Longleftrightarrow \qquad \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \mathbb{D}_{\mathrm{KL}}(q(f(\cdot;\boldsymbol{\theta}))\|p(f(\cdot;\boldsymbol{\theta}) \mid \mathcal{D}))$$

$$\max_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \left\{ \mathbb{E}_{q(f(\mathbf{X}_\mathcal{D};\boldsymbol{\theta}))} \left[\log p\left(\mathbf{y} \mid f\left(\mathbf{X}_\mathcal{D};\boldsymbol{\theta}\right)\right)\right] - \mathbb{D}_{\mathrm{KL}}(q(f(\cdot;\boldsymbol{\theta}))\|p(f(\cdot;\boldsymbol{\theta}))) \right\}$$

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **Distributions over functions** induced by distributions over parameters:

$$p(\boldsymbol{\theta}) \Longrightarrow p(f(\cdot\,;\boldsymbol{\theta})) \qquad q(\boldsymbol{\theta}) \Longrightarrow q(f(\cdot\,;\boldsymbol{\theta}))$$

- **Variational problem**:

$$\Longleftrightarrow \qquad \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \mathbb{D}_{\mathrm{KL}}(q(f(\cdot;\boldsymbol{\theta}))\|p(f(\cdot;\boldsymbol{\theta}) \mid \mathcal{D}))$$

$$\max_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \left\{ \mathbb{E}_{q(f(\mathbf{X}_\mathcal{D};\boldsymbol{\theta}))} \left[\log p\left(\mathbf{y} \mid f\left(\mathbf{X}_\mathcal{D};\boldsymbol{\theta}\right)\right)\right] - \sup_{\mathbf{X} \in \mathcal{X}_\mathbb{N}} \mathbb{D}_{\mathrm{KL}}\left(q\left(f\left(\mathbf{X};\boldsymbol{\theta}\right)\right) \| p\left(f\left(\mathbf{X};\boldsymbol{\theta}\right)\right)\right) \right\}$$

(Sun et al. [2019])

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **KL Approximation:**
  - Linearize mapping: $f(\cdot; \boldsymbol{\Theta}) \approx \tilde{f}(\cdot; \boldsymbol{\Theta}) \doteq f(\cdot; \mathbf{m}) + \mathcal{J}_{\mathbf{m}}(\cdot)(\boldsymbol{\Theta} - \mathbf{m})$

$$\text{with} \quad \mathcal{J}_{\mathbf{m}}(\cdot) \doteq \left. \frac{\partial f(\cdot; \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \right|_{\Theta = \mathbf{m}}$$

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **KL Approximation:**
  - Distributions over functions under linearized mapping

$$q(f(\cdot;\boldsymbol{\theta})) \approx \tilde{q}(\tilde{f}(\cdot;\boldsymbol{\theta})) \qquad p(f(\cdot;\boldsymbol{\theta})) \approx \tilde{p}(\tilde{f}(\cdot;\boldsymbol{\theta}))$$

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **KL Approximation:**
  - Approximate KL divergence:

$$\sup_{\mathbf{X} \in \mathcal{X}_{\mathbb{N}}} \mathbb{D}_{\mathrm{KL}}(q(f(\mathbf{X}; \boldsymbol{\theta})) \| p(f(\mathbf{X}; \boldsymbol{\theta}))) \approx \sup_{\mathbf{X} \in \mathcal{X}_{\mathbb{N}}} \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}(\mathbf{X}; \boldsymbol{\theta})) \| \tilde{p}(\tilde{f}(\mathbf{X}; \boldsymbol{\theta})))$$

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **KL Approximation:**
  - Approximate KL divergence:

$$\sup_{\mathbf{X}\in\mathcal{X}_{\mathbb{N}}} \mathbb{D}_{\mathrm{KL}}(q(f(\mathbf{X};\boldsymbol{\theta}))\|p(f(\mathbf{X};\boldsymbol{\theta}))) \approx \sup_{\mathbf{X}\in\mathcal{X}_{\mathbb{N}}} \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))\|\tilde{p}(\tilde{f}(\mathbf{X};\boldsymbol{\theta})))$$

  - Supremum Estimation:

$$\sup_{\mathbf{X}\in\mathcal{X}_{\mathrm{N}}} \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))\|\tilde{p}(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))) \approx \max_{\mathbf{X}\in\mathcal{X}_{\mathcal{C}}^{S}} \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))\|\tilde{p}(\tilde{f}(\mathbf{X};\boldsymbol{\theta})))$$

    where $\mathcal{X}_{\mathcal{C}}^{S} = \{\mathbf{X}^{(i)}\}_{i=1}^{S}$ with $\mathbf{X}^{(i)} = \{\mathbf{x}^{(j)}\}_{j=1}^{K}$ sampled from $p_{\mathcal{X}_{\mathcal{C}}}$

Background: Function-Space Variational Inference

Function-Space Variational Inference (Rudner et al. [2021])

- **Approximate variational problem:**

$$\max_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \left\{ \mathbb{E}_{q(f(\mathbf{X}_\mathcal{D}; \boldsymbol{\theta}))} \left[\log p\left(\mathbf{y} \mid f\left(\mathbf{X}_\mathcal{D}; \boldsymbol{\theta}\right)\right)\right] - \max_{\mathbf{X} \in \mathcal{X}_\mathcal{C}^S} \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}(\mathbf{X}; \boldsymbol{\theta})) \| \tilde{p}(\tilde{f}(\mathbf{X}; \boldsymbol{\theta}))) \right\}$$

where $\mathcal{X}_\mathcal{C}^S = \{\mathbf{X}^{(i)}\}_{i=1}^S$ with $\mathbf{X}^{(i)} = \{\mathbf{x}^{(j)}\}_{j=1}^K$ sampled from $p_{\mathcal{X}_\mathcal{C}}$

Sequential Function-Space Variational Inference

Empirical prior and variational distributions over functions

- Define prior and variational distributions over parameters:

$$p_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$$
$$q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

- Tractable prior and variational distributions over functions:

$$\tilde{p}_t(\tilde{f}(\mathbf{X};\boldsymbol{\theta})) = \tilde{q}_{t-1}(\tilde{f}(\mathbf{X};\boldsymbol{\theta})) = \mathcal{N}(f(\mathbf{X};\boldsymbol{\mu}_{t-1}), \mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\boldsymbol{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top)$$

$$\tilde{q}_t(\tilde{f}(\mathbf{X};\boldsymbol{\theta})) = \mathcal{N}(f(\mathbf{X};\boldsymbol{\mu}_t), \mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X})\boldsymbol{\Sigma}_t\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X}')^\top)$$

Sequential Function-Space Variational Inference (S-FSVI)

- Approximate function-space variational objective under linearization:

$$\mathbb{E}_{q(f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta}))}\left[\log p\left(\mathbf{y}_{\mathcal{D}_t} \mid f\left(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta}\right)\right)\right] - \max_{\mathbf{X}\in\mathcal{X}_{\mathcal{C}}^{S}} \mathbb{D}_{\mathrm{KL}}(\tilde{q}_t(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))\|\tilde{q}_{t-1}(\tilde{f}(\mathbf{X};\boldsymbol{\theta})))$$

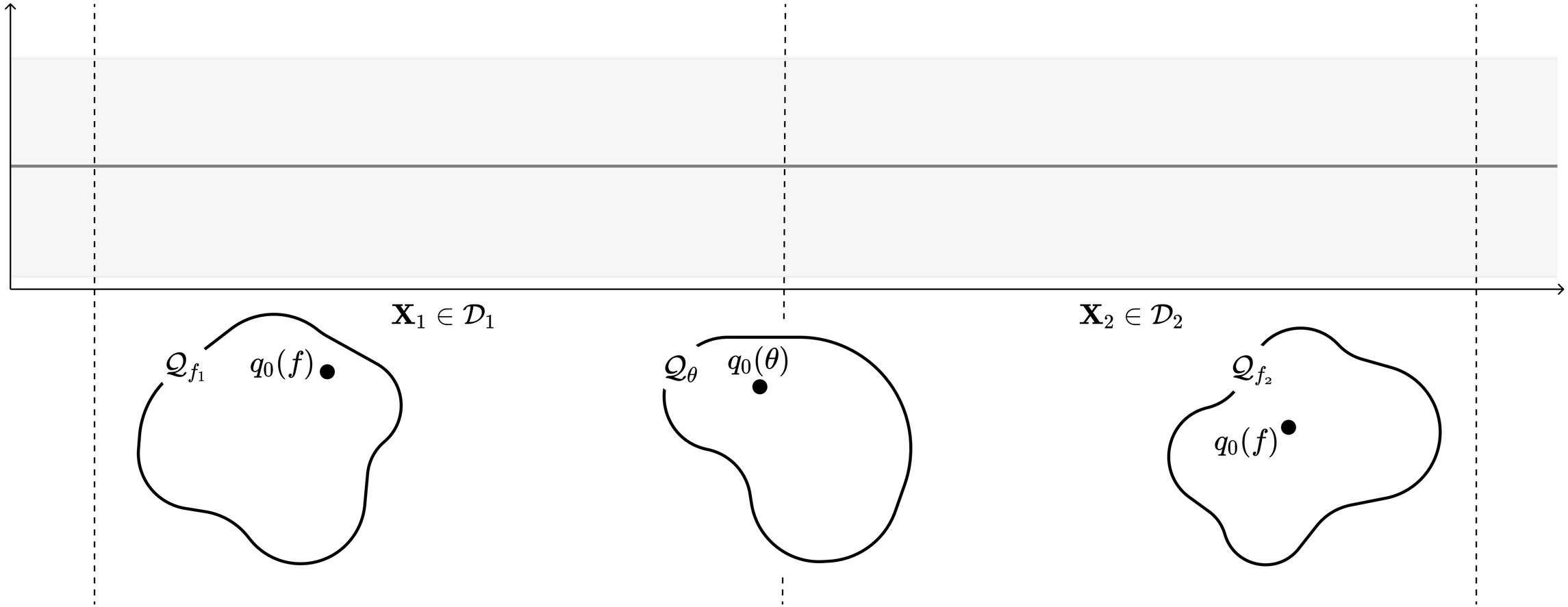
  where $\mathcal{X}_{\mathcal{C}}^{S} = \{\mathbf{X}^{(i)}\}_{i=1}^{S}$ with $\mathbf{X}^{(i)} = \{\mathbf{x}^{(j)}\}_{j=1}^{K}$ sampled from $p_{\mathcal{X}_{\mathcal{C}}}$

- Crucially, $p_{\mathcal{X}_{\mathcal{C}}}$ is defined over a set of **task-specific coresets**

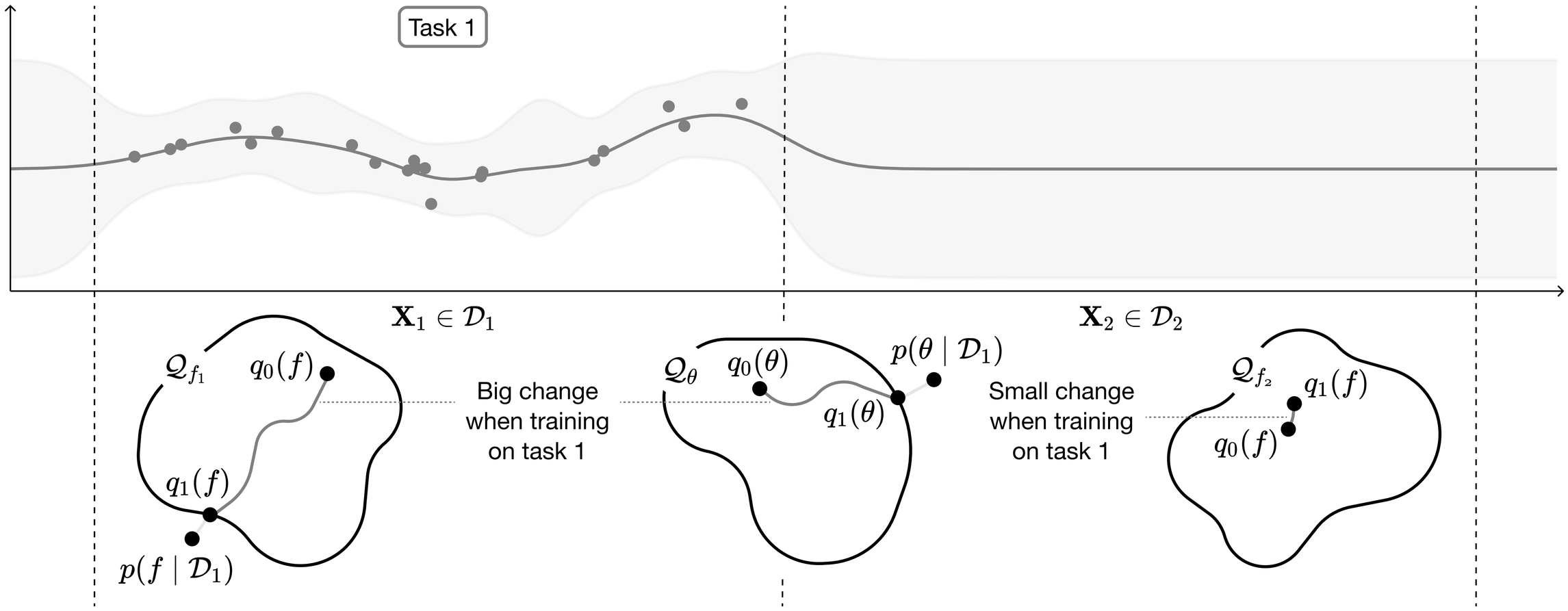
$$\mathcal{X}_{1:t} \subseteq \mathcal{P}\left(\{\mathbf{x}_t^{(n)}\}_{n=1}^{N_t} \cup \bigcup_{t'=1}^{t-1} \{\mathbf{x}_{t'}^{(n)}\}_{n=1}^{S_{t'}}\right)$$

- Explicitly discourages divergence from prior distribution over functions

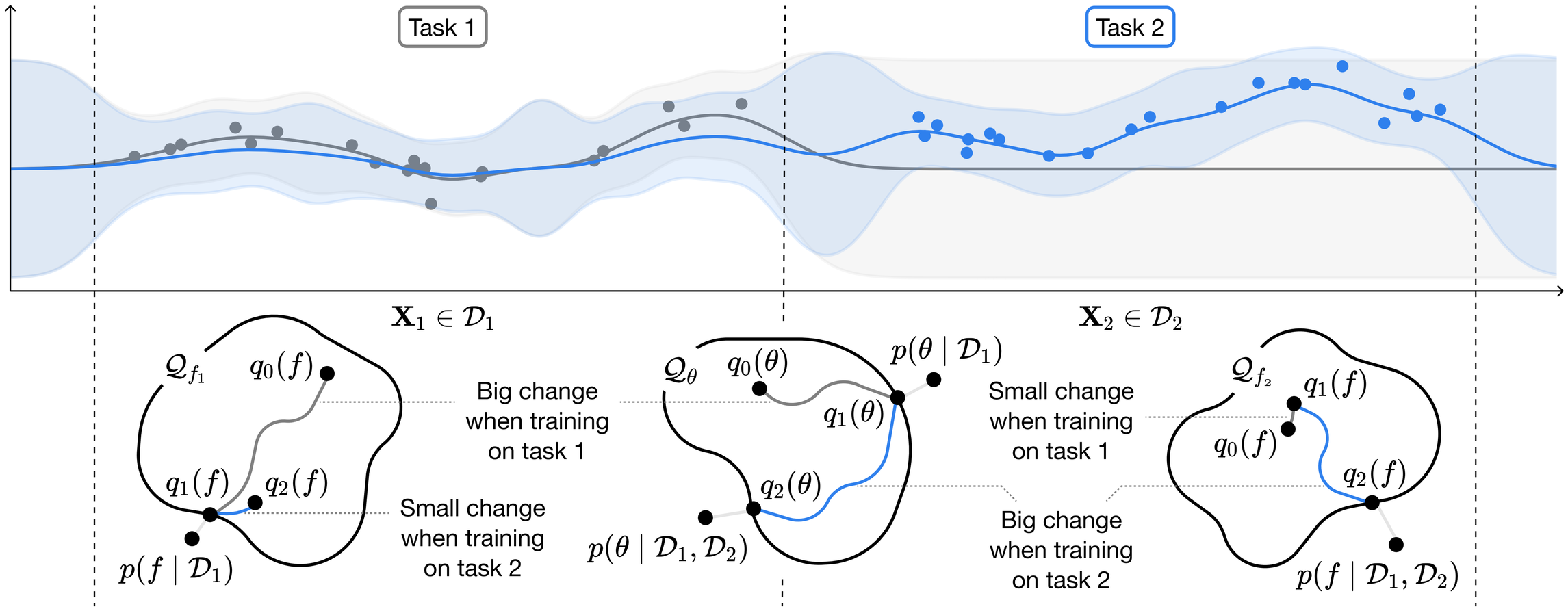
Sequential Function-Space Variational Inference



Continual Learning via Function-Space Variational Inference



Continual Learning via Function-Space Variational Inference

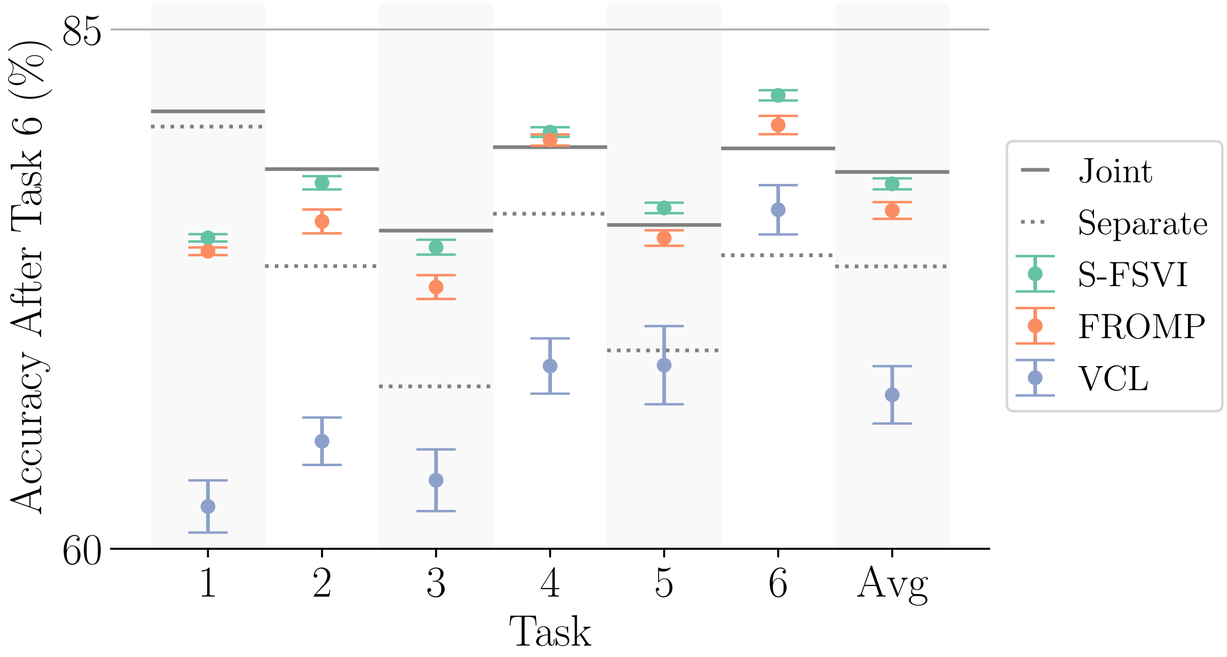


Predictive Performance: MNIST

Method	S-MNIST (MH)	S-FMNIST (MH)	P-MNIST (SH)	S-MNIST (SH)
EWC (Kirkpatrick et al., 2017)	63.10%	—	84.00%	—
SI (Zenke et al., 2017)	98.90%	—	86.00%	—
VCL (Nguyen et al., 2018)[1]	98.40%	98.60%±0.04	93.00%	32.11%±1.16
VCL (no coreset)	97.00%	89.60%±1.75	87.50%±0.61	17.74%±1.20
FRCL (Titsias et al., 2020)[3]	97.80%±0.22	97.28%±0.17	94.30%±0.06	—
FROMP (Pan et al., 2020)	99.00%±0.04	99.00%±0.03	94.90%±0.04	35.29%±0.52
VAR-GP (Kapoor et al., 2021)	—	—	**97.20%**±0.08	90.57%±1.06
S-FSVI (**ours**)[2]	**99.54%**±0.04	**99.19%**±0.02	95.76%±0.02	**92.87%**±0.14
S-FSVI Ablation Study:				
S-FSVI (larger networks)[4]	**99.76%**±0.00	99.16%±0.03	**97.50%**±0.01	**93.38%**±0.10
S-FSVI (no coreset)[5]	99.62%±0.02	**99.54%**±0.01	84.06%±0.46	20.15%±0.52
S-FSVI (minimal coreset)[6]	—	—	89.59%±0.30	51.44%±1.22

- Multi-head settings: virtually solved (with and without coresets)
- Single-head settings: still require coresets

Predictive Performance: Split CIFAR & Sequential Omniglot



Split CIFAR

Method	Test Accuracy
Learning Without Forgetting[1]	62.06%±2.0
EWC	67.32%±4.7
Online EWC[2]	69.99%±3.2
Progress & Compress[3]	70.32%±3.3
FRCL[4]	81.47%±1.6
S-FSVI (**ours**)[5]	**83.29%**±1.2

Sequential Omniglot

Summary

Sequential Function-Space Variational Inference

- Is based on **Bayesian inference over stochastic functions**
- **Outperforms related** objective- and replay-based **methods**
- Can be scaled to
  - **large network architectures**
  - **long task sequences**

Thank You!

Correspondence to

tim.rudner@cs.ox.ac.uk

Paper: http://timrudner.com/sfsvi
Code: http://timrudner.com/sfsvi-code

Tim G. J. Rudner
@timrudner

Freddie Bickford Smith
@fbickfordsmith

Qixuan Feng
@qixuan_feng

Yee Whye Teh
@yeewhye

Yarin Gal
@yaringal