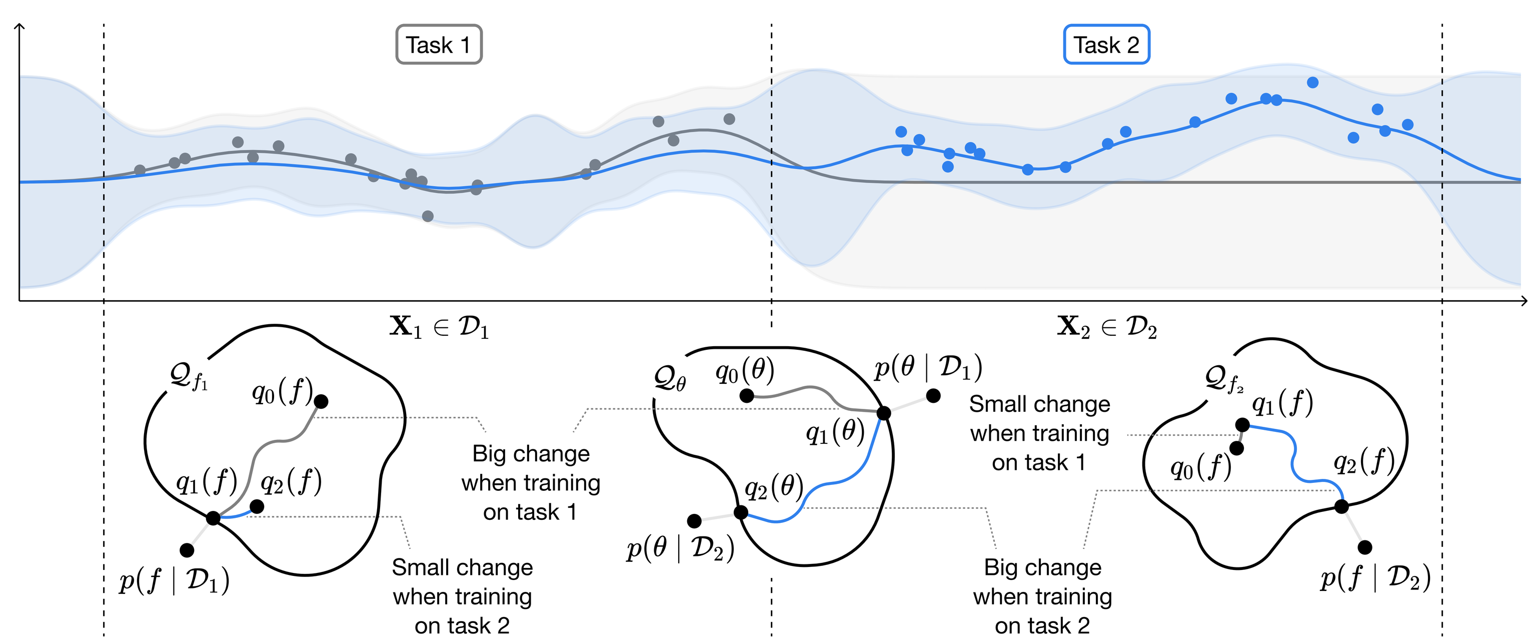
Continual Learning via Sequential Function-Space Variational Inference

Tim G. J. Rudner[†] [*] Freddie Bickford Smith[†] Qixuan Feng[†] Yee Whye Teh[†] Yarin Gal[†]

[†] University of Oxford Corresponding author: tim.rudner@cs.ox.ac.uk. @timrudner

TL;DR

- We show that continual learning can be formulated as function-space variational inference and propose a tractable variational objective for scalable and effective learning.
- We demonstrate that our method significantly outperforms related approaches on single- and multi-head tasks.



Background

- Consider a neural network $f(\mathbf{x};\boldsymbol{\theta})$ parameterized by stochastic parameters $\boldsymbol{\Theta} \in \mathbb{R}^P$ and define a conditional distribution of targets given function values $f(\mathbf{x};\boldsymbol{\theta})$: $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}; f)$.

Parameter-Space Variational Inference in BNNs

- Want to find the posterior over parameters: $p(\boldsymbol{\theta} \,|\, \mathcal{D})$
- Find variationally via $\min_{q(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\theta}}} \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \,|\, \mathcal{D}))$,

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

Function-Space Kullback-Leibler Divergence

- Want to find the posterior over functions: $p(f(\cdot\,;\boldsymbol{\theta}) \,|\, \mathcal{D})$
- Find variationally via

$$\min_{q(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\theta}}} \mathbb{D}_{\mathrm{KL}}(q(f(\cdot\,;\boldsymbol{\theta})) \,\|\, p(f(\cdot\,;\boldsymbol{\theta}) \,|\, \mathcal{D})) \quad (1)$$

- Data Processing Inequality (Polyanskiy and Wu, 2017):

$$\mathbb{D}_{\mathrm{KL}}(q(f(\cdot\,;\boldsymbol{\theta})) \,\|\, p(f(\cdot\,;\boldsymbol{\theta}))) \leq \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})) \quad (2)$$

- If $\mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})) < \infty$, then the function-space KL is well-defined.

Continual Learning via Function-Space VI

Proposition 1 (Continual Function-Space Variational Objective). *Let $q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and $p_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$, and let the linearization of the mapping f about parameters $\tilde{\boldsymbol{\theta}}$ be given by*

$$\tilde{f}(\cdot\,;\boldsymbol{\Theta}) \doteq f(\cdot\,;\tilde{\boldsymbol{\theta}}) + \mathcal{J}_{\tilde{\boldsymbol{\theta}}}(\cdot)(\boldsymbol{\Theta} - \tilde{\boldsymbol{\theta}}), \quad (3)$$

For $\boldsymbol{\Theta}$ distributed according to $q_t(\boldsymbol{\theta})$ and $p_t(\boldsymbol{\theta})$, the induced distributions under the linearized mapping $\tilde{f}$ evaluated at $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ are given by

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

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

Under certain variational assumptions and approximations (see paper), we obtain the variational objective

$$\tilde{\mathcal{F}}(q_t(\boldsymbol{\theta})) \doteq \mathbb{E}_{q_t(f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta}))}[\log p(\mathbf{y}_t \,|\, f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta})] - \mathbb{D}_{\mathrm{KL}}(\tilde{q}_t(\tilde{f}(\mathbf{X}_{\mathcal{I}};\boldsymbol{\theta})) \,\|\, \tilde{p}_t(\tilde{f}(\mathbf{X}_{\mathcal{I}};\boldsymbol{\theta}))). \quad (4)$$

Proposition 2 (Continual Function-Space Variational Inference (C-FSVI)). *For a mini-batch $(\mathbf{X}_{\mathcal{B}_t}, \mathbf{y}_{\mathcal{B}_t})$, and under diagonal approximations to the variational and prior covariance,*

$$K^{p_t}_{\mathcal{I}\mathcal{I}} \doteq \operatorname{diag}\left(\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\boldsymbol{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top\right)$$

$$K^{q_t}_{\mathcal{I}\mathcal{I}} \doteq \operatorname{diag}\left(\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X})\boldsymbol{\Sigma}_t\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X}')^\top\right)$$

the objective can be optimized via stochastic VI on

$$\bar{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \frac{1}{S}\sum_{i=1}^{S} \log p(\mathbf{y}_{\mathcal{B}_t} \,|\, f(\mathbf{X}_{\mathcal{B}_t}; h(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\epsilon}^{(i)})))$$
$$- \sum_{k=1}^{Q_t}\sum_{j=1}^{|\mathbf{X}_{\mathcal{I}}|}\frac{1}{2}\Bigg(\log\frac{[K^{p_t}_{\mathcal{I}\mathcal{I}}]_{j,k}}{[K^{q_t}_{\mathcal{I}\mathcal{I}}]_{j,k}} + \frac{[K^{q_t}_{\mathcal{I}\mathcal{I}}]_{j,k}}{[K^{p_t}_{\mathcal{I}\mathcal{I}}]_{j,k}} - 1$$
$$+ \frac{\left([f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_t)]_{j,k} - [f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_{t-1})]_{j,k}\right)^2}{[K^{p_t}_{\mathcal{I}\mathcal{I}}]_{j,k}}\Bigg),$$

where $h(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\epsilon}^{(i)}) \doteq \boldsymbol{\mu}_t + \boldsymbol{\Sigma}_t \odot \boldsymbol{\epsilon}^{(i)}$ is a reparameterization of $\boldsymbol{\Theta}$ with $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P)$ and Q_t is the number of model output dimensions over which the KL is being evaluated.

Empirical Evaluation

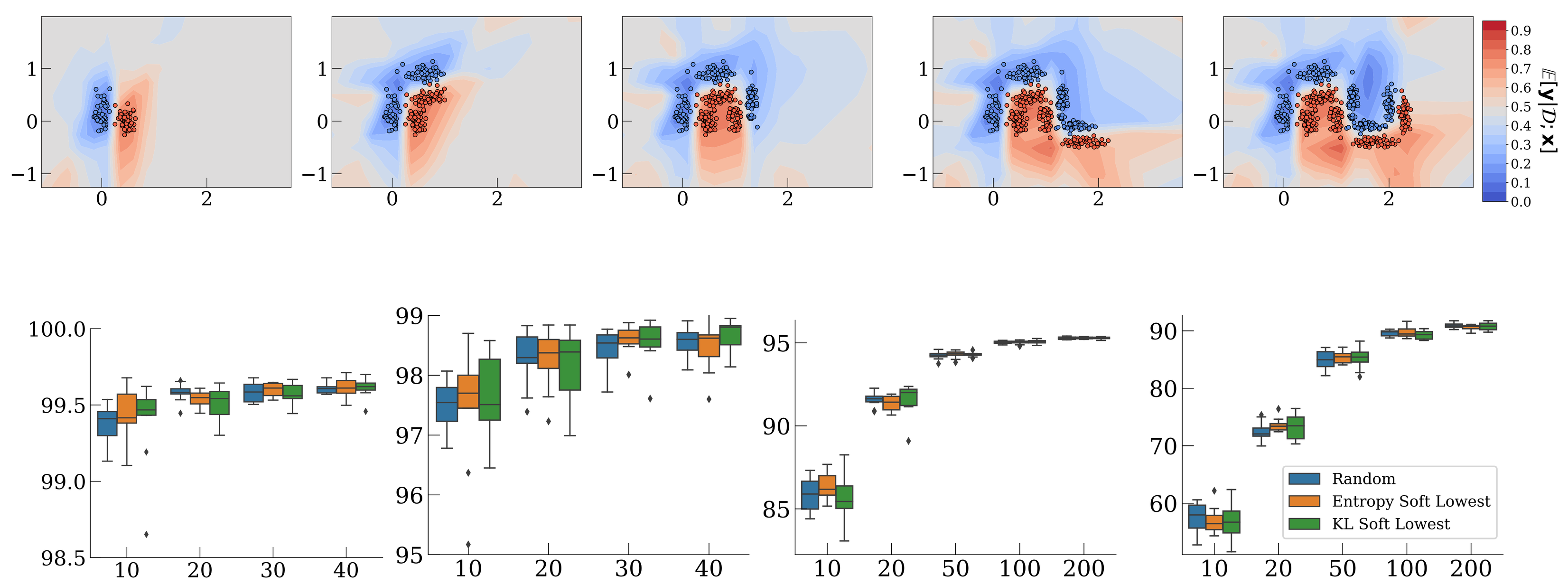


Figure 1: Comparison of coreset selection methods.

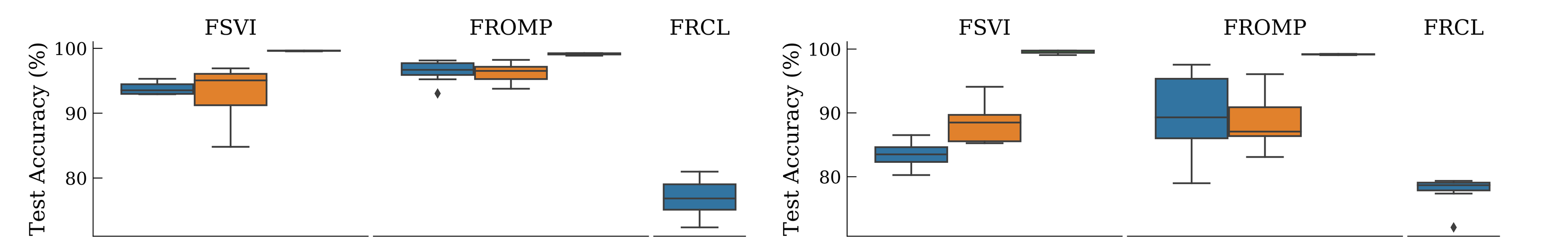


Figure 2: Comparison of C-FSVI to state-of-the-art functional regularization methods.

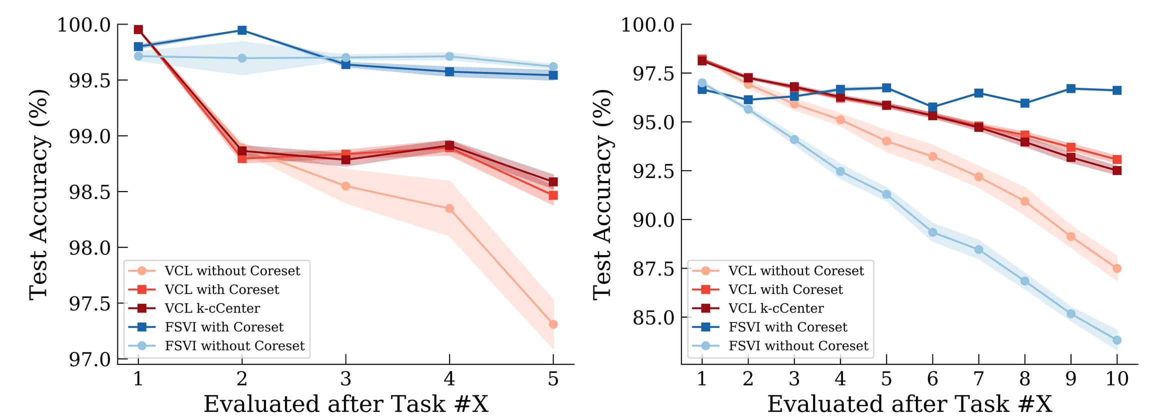


Figure 3: Comparison of function-space and parameter-space variational inference.

Table 1. Comparison of predictive performance of a selection of continual-learning methods on four task sequences, each with either a multi-head (MH) or single-head (SH) setup. Each numerical entry denotes the mean accuracy across tasks at the end of training (over ten random seeds for C-FSVI). For each task sequence, all methods use the same architecture and coreset size unless explicitly indicated otherwise.

Method	sMNIST (MH)	sFMNIST (MH)	pMNIST (SH)	sMNIST (SH)
EWC	63.10%	—	84.00%	—
SI	98.90%	—	86.00%	—
VCL	98.40%	98.60%±0.04	93.00%	32.11%±1.16
VCL (no coreset)	97.00%	89.60%±1.75	—	—
FRCL	97.80%±0.22	97.28%±0.17	94.30%±0.06	—
FROMP	99.00%±0.04	99.00%±0.03	94.90%±0.04	35.29%±0.52
VAR-GP	—	—	97.20%±0.08	90.57%±1.06
C-FSVI [2]	99.54%±0.04	99.19%±0.02	95.76%±0.02	92.87%±0.14
C-FSVI (larger networks)	**99.77%**±0.00	99.16%±0.03	**97.50%**±0.01	**93.38%**±0.10
C-FSVI (no coreset)	99.62%±0.02	**99.54%**±0.01	—	—
C-FSVI (minimal coreset)	—	—	89.59%±0.30	51.44%±1.22

Full paper: https://timrudner.com/cfsvi