Continual Learning via Function-Space Variational Inference: A Unifying View

Tim G. J. Rudner [1] **Freddie Bickford Smith** [1] **Qixuan Feng** [1] **Yee Whye Teh** [1] **Yarin Gal** [1]

Abstract

Continual learning is the process of developing new abilities while retaining existing ones. Sequential Bayesian inference is a natural framework for this, but applying it successfully to deep neural networks remains a challenge. We propose continual function-space variational inference (C-FSVI), in which the variational distribution over functions induced by stochastic model parameters is encouraged to match the variational distribution over functions induced by stochastic parameters inferred on previous tasks. Unlike approaches that explicitly penalize changes in the model parameters, function-space regularization allows parameters to vary widely during training, resulting in greater flexibility to fit new data. C-FSVI improves on existing approaches to function-space regularization by performing inference entirely in function space and without relying on carefully selected coreset points. We show that C-FSVI outperforms alternative methods based on parameter-space and function-space regularization on a range of task sequences.

1. Introduction

People effortlessly grow their set of abilities over time through continual learning. Enabling deep neural networks to successfully and scalably emulate this ability remains an open problem. Solving this would make possible a wide range of applications that require flexibly adapting to new tasks while maintaining reliable performance on past tasks.

A number of recent advances have been made in designing training objectives that support learning on the current task while retaining good performance on past tasks. However, despite progress, existing objective-based approaches to continual learning fall short in important ways. Many methods mitigate forgetting by placing an explicit penalty on undesired changes in a model's parameters from one task to another (Ahn et al., 2019; Aljundi et al., 2018; Chaudhry et al., 2018; Ebrahimi et al., 2020; Kirkpatrick et al., 2017; Lee et al., 2017; Liu et al., 2018; Loo et al., 2020; Nguyen et al., 2018; Park et al., 2019; Ritter et al., 2018; Schwarz et al., 2018; Swaroop et al., 2019; Yin et al., 2020a;b; Zenke et al., 2017a), but these methods are often brittle, reflecting the fact that parameters are only a proxy for a model's predictive function. More promising are approaches that explicitly penalize undesired changes in function space instead of parameter space (Benjamin et al., 2019; Jung et al., 2018; Kim et al., 2018; Li & Hoiem, 2018; Pan et al., 2020; Titsias et al., 2020). Unfortunately, although well-motivated, state-of-the-art function-space methods rely on restrictive variational families or are only applicable to a small set of model classes, like Bayesian linear models or non-parametric Gaussian processes (Pan et al., 2020; Titsias et al., 2020; Kapoor et al., 2021).

Contributions. We frame continual learning in terms of sequential Bayesian inference over stochastic functions and derive continual function-space variational inference (C-FSVI), a simple method that avoids the shortcomings of alternative approaches based on parameter-space and function-space regularization. Through empirical evaluation, we show that C-FSVI is effective at incorporating prior information while learning on the current task, does not require ad-hoc design modifications for implementation, and does not rely on careful coreset selection.

2. Preliminaries

Notation & Assumptions. Supervised learning typically consists of training a predictive model $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}; f)$ on a dataset $\mathcal{D} \doteq \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N = (\mathbf{X}, \mathbf{y})$ with inputs $\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^D$ and targets $\mathbf{y}_n \in \mathcal{Y}$ where $\mathcal{Y} \subseteq \mathbb{R}^Q$ for regression and $\mathcal{Y} \subseteq \{0, 1\}^Q$ for classification. In contrast to settings where the examples in $\mathcal{D}$ are assumed to be independently and identically distributed, in continual learning, the dataset $\mathcal{D}$ is typically split into T disjoint subsets $\{\mathcal{D}_t\}_{t=1}^T = \cup_{t=1}^T \{\mathbf{X}_t, \mathbf{y}_t\}$, with each of the subsets being associated with a distinct task. The fundamental challenge in continual learning is to obtain a predictive model that makes good predictions on examples drawn from any task while discarding all or almost all of the data in each task-specific dataset $\mathcal{D}_t$ when training on subsequent tasks.

Desiderata for Continual Learning. What constitutes desirable model behavior in continual-learning settings and how to evaluate such behaviors is a topic of ongoing dis-

[1]University of Oxford, Oxford, UK. Correspondence to: Tim G. J. Rudner <tim.rudner@cs.ox.ac.uk>. Preprint (July 24, 2021).

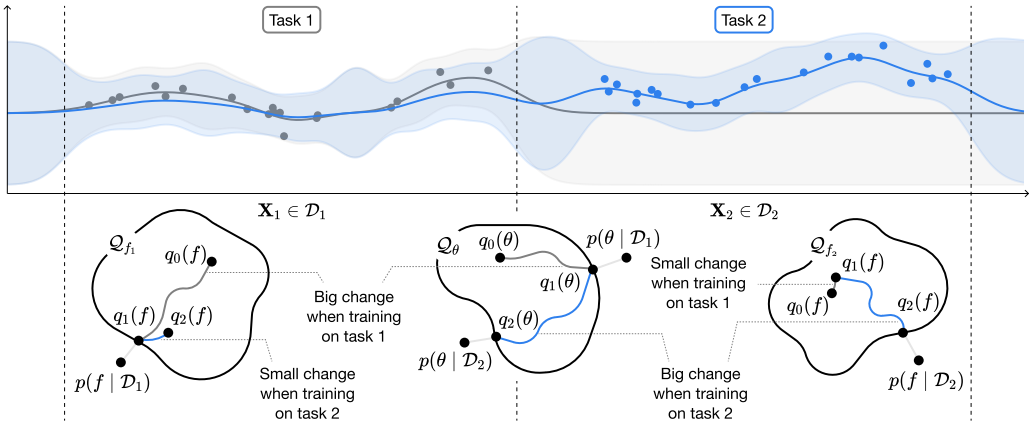


Figure 1. A schematic of how continual function-space variational inference (C-FSVI) allows a Bayesian neural network to learn on a new task while maintaining previously learned abilities. On task 1, the model fits dataset $\mathcal{D}_1$ by updating an initial distribution over parameters $q_0(\boldsymbol{\theta})$ to a variational posterior $q_1(\boldsymbol{\theta})$, which in turn induces a distribution over functions $q_1(f(\mathbf{X}_1); \boldsymbol{\theta})$, where $\mathbf{X}_1 \in \mathcal{D}_1$. On task 2, C-FSVI encourages the posterior distribution over functions to match $q_1(f(\mathbf{X}_1); \boldsymbol{\theta})$ on a small set of data points from task 1 while also fitting dataset $\mathcal{D}_2$. Meanwhile, the distribution over parameters may change significantly.

cussion (Farquhar & Gal, 2018). Informing the approach described in Section 3, we concentrate on three desiderata for continual learning. First, continual learning should avoid catastrophic forgetting of previously learned tasks. Second, the model should not rely on explicit information as to which task it is being tested on. For example, it should not require a task identifier in order to make a prediction on unseen data. This desideratum is violated in often-considered "multi-head" settings. Third, the storage of replay data from past tasks should be minimal. While some have suggested that no data at all should be stored, we argue this is an overly strong restriction and that storing minimal amounts of data (e.g., a single example per class) is scalable for a wide array of real-world continual-learning settings.

Function-Space Variational Inference. Instead of defining the probabilistic model implicitly in terms of the parameters, we define it explicitly in terms of the stochastic functions induced by the stochastic parameters $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. Considering a mapping f from parameters to functions $f(\mathbf{x}) : \boldsymbol{\Theta} \to \mathbb{R}^Q$ for $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$ and a distribution over parameters $p(\boldsymbol{\theta})$, we obtain a distribution over functions $p(f(\mathbf{x}; \boldsymbol{\theta}))$ indexed by $\mathcal{X}$.

For this model and an observed set of data points $\mathcal{D}$, we can then frame the problem of finding a posterior distribution over functions $p(f(\cdot\,; \boldsymbol{\theta}) \,|\, \mathcal{D})$ variationally as

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

where $q(f(\cdot\,; \boldsymbol{\theta}))$ is the variational distribution over functions induced by the variational distribution $q(\boldsymbol{\theta})$. As shown by de G. Matthews et al. (2016) and Rudner et al. (2021), for a likelihood function defined on a finite set of training targets $\mathbf{y}$, we can equivalently express this inference problem as maximizing a variational lower bound:

$$\max_{q(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\theta}}} \mathcal{F}(q(\boldsymbol{\theta}))) \doteq \max_{q(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\theta}}} \{\mathbb{E}_{q(f(\mathbf{X}_{\mathcal{D}}; \boldsymbol{\theta}))}[\log p(\mathbf{y} \,|\, f(\mathbf{X}_{\mathcal{D}}; \boldsymbol{\theta}))] - \mathbb{D}_{\mathrm{KL}}(q(f(\cdot\,; \boldsymbol{\theta})) \,\|\, p(f(\cdot\,; \boldsymbol{\theta})))\}, \tag{1}$$

where $\mathbb{D}_{\mathrm{KL}}(q(f(\cdot\,; \boldsymbol{\theta})) \,\|\, p(f(\cdot\,; \boldsymbol{\theta})))$ is a KL divergence between distributions over functions. For a measure-theoretic derivation of this result, see Rudner et al. (2021).

As shown by Burt et al. (2021), the above variational objective function is well-defined for suitably chosen prior distributions over functions. Specifically, the KL divergence between two distributions over functions generated from different distributions over parameters applied to the same mapping (e.g., the same neural network architecture) is finite if the KL divergence between the distributions over parameters is finite since, by the strong data-processing inequality (Polyanskiy & Wu, 2017),

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

As a result, if $\mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})) < \infty$, which is the case for finite-dimensional parameter vectors $\boldsymbol{\theta}$ if and only if $q(\boldsymbol{\theta})$ is absolutely continuous with respect to $p(\boldsymbol{\theta})$, then the function-space KL divergence is finite and thus well-defined as a variational objective.

3. Continual Learning via Function-Space Variational Inference

Sequential Bayesian inference over stochastic functions is a natural framework for continual learning. Using this framework, we derive a simple function-space variational objective for Bayesian deep neural networks.

3.1. Continual Function-Space Variational Inference via Empirical Prior Distributions over Functions

Suppose a variational distribution over functions $q_1(f(\cdot\,;\boldsymbol{\theta}))$ was learned by maximizing the variational objective in Equation (1) with respect to a variational distribution $q_1(\boldsymbol{\theta})$. To retain the distribution over functions $q_1(f(\cdot\,;\boldsymbol{\theta}))$ on input points living in subsets of the data space corresponding to task 1, we consider an empirical prior distribution over functions. Specifically, we define an empirical prior distribution by the previously inferred distribution over functions $q_1(f(\cdot\,;\boldsymbol{\theta}))$. To do so, we consider a variational distribution and a prior distribution over functions defined in terms of the same mapping, but induced by different distributions over parameters. That is, we define an empirical prior over functions as the distribution over functions induced by the previously inferred distribution over parameters $q_1(\boldsymbol{\theta})$. Letting $p_2(\boldsymbol{\theta}) \doteq q_1(\boldsymbol{\theta})$, we obtain an empirical prior distribution over functions $p_2(f(\cdot;\boldsymbol{\theta}))$ matching the previously inferred variational distribution and ensuring that, by the data-processing inequality, the resulting variational objective is well-defined.

Learning to solve a second task without forgetting the first then requires solving the following variational problem:

$$\max_{q_2(\boldsymbol{\theta})\in\mathcal{Q}_{\boldsymbol{\theta}}} \mathcal{F}(q_2(\boldsymbol{\theta})) \doteq \max_{q_2(\boldsymbol{\theta})\in\mathcal{Q}} \{\mathbb{E}_{q_2(f(\mathbf{X}_2;\boldsymbol{\theta}))}[\log p(\mathbf{y}_2 \mid f_{\mathbf{X}_2})] - \mathbb{D}_{\mathrm{KL}}(q_2(f(\cdot\,;\boldsymbol{\theta})) \,\|\, p_2(f(\cdot\,;\boldsymbol{\theta})))\}.$$

Note that the mapping that induces the distributions over functions on the first and second task remains the same but the prior distributions over parameters differ. Importantly, since the mapping from the distribution over parameters to the distribution over functions remains unchanged, the data-processing inequality guarantees that the variational objective remains well-defined as long as the variational distribution over parameters $q_2(\boldsymbol{\theta})$ is absolutely continuous with respect to the prior distributions over parameters $p_2(\boldsymbol{\theta}) = q_1(\boldsymbol{\theta})$.

In general, the function-space variational problem for a pair of sequential tasks t and $t-1$ is given by

$$\max_{q_t(\boldsymbol{\theta})\in\mathcal{Q}_{\boldsymbol{\theta}}} \mathcal{F}(q_t(\boldsymbol{\theta})) \doteq \max_{q_t(\boldsymbol{\theta})\in\mathcal{Q}} \{\mathbb{E}_{q_t(f(\mathbf{X}_t;\boldsymbol{\theta}))}[\log p(\mathbf{y}_t \mid f_{\mathbf{X}_t})] - \mathbb{D}_{\mathrm{KL}}(q_t(f(\cdot\,;\boldsymbol{\theta})) \,\|\, q_{t-1}(f(\cdot\,;\boldsymbol{\theta})))\}. \tag{3}$$

Intuitively, Equation (3) trades off fitting the predictive distribution to the current training data while also matching the empirical prior distribution as well as possible. For a sufficiently well-parameterized stochastic mapping $f(\cdot\,;\boldsymbol{\Theta})$, the objective promotes a variational distribution over parameters that induces a variational distribution that matches the empirical prior in parts of the input space not associated with the current task, while fitting the current training data. As such, the objective may result in a distribution over parameters that is significantly different from the empirical prior distribution over parameters, while matching the induced empirical distribution over functions for sets of input points in parts of the input space not associated with the current task. We demonstrate this process pictorially.

3.2. Continual Function-Space Variational Inference

The variational objective in Equation (3) is in general intractable for Bayesian neural networks, since the KL divergence between distributions over functions induced by a non-linear mapping on a distribution over parameters is analytically intractable. To make the objective analytically tractable, we follow Rudner et al. (2021) and employ a local approximation to the stochastic function $f(\cdot\,;\boldsymbol{\Theta})$ by a distribution over stochastic parameters $\boldsymbol{\Theta}$ by linearizing them about the means of the task-specific variational and prior distributions over parameters $q_t(\boldsymbol{\theta})$ and $p_t(\boldsymbol{\theta})$, respectively. Assuming $q_t(\boldsymbol{\theta})$ and $p_t(\boldsymbol{\theta})$ to be Gaussian, this approximation turns the distribution over linearized stochastic functions induced by the variational distribution over parameters $\tilde{q}_t(\tilde{f}(\cdot\,;\boldsymbol{\theta}))$ and the distribution over linearized stochastic functions induced by the prior distribution over parameters $\tilde{p}_t(\tilde{f}(\cdot\,;\boldsymbol{\theta}))$ into degenerate Gaussian processes. To evaluate the resulting KL divergence between distributions over linearized stochastic functions, we follow (Rudner et al., 2021) and make several variational assumptions, which result in a tractable KL divergence evaluated at a finite number of evaluation points. Instead of restating the full set of approximations outlined in (Rudner et al., 2021), we present the final variational objective and refer the reader to Appendix A.1 for the full set of approximations.

Proposition 1 (Continual Function-Space Variational Objective (adapted from Rudner et al. (2021))). *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*

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

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

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

respectively. Under the approximations in Appendix A.1, we obtain the variational objective

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

Proof. See Appendix A.2 □

To obtain a practically useful and scalable variational objective that can be estimated via Monte Carlo estimation, we derive the following estimator:

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 function,*

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

the objective can be optimized via stochastic variational inference on

$$\begin{aligned}\bar{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) &= \frac{1}{S}\sum_{i=1}^{S}\log p(\mathbf{y}_{\mathcal{B}_t} \mid f(\mathbf{X}_{\mathcal{B}_t}; h(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\epsilon}^{(i)}))) \\ &\quad - \sum_{k=1}^{Q_t}\sum_{j=1}^{|\mathbf{X}_\mathcal{I}|}\frac{1}{2}\Bigg(\log\frac{[K_{\mathcal{I}\mathcal{I}}^{p_t}]_{j,k}}{[K_{\mathcal{I}\mathcal{I}}^{q_t}]_{j,k}} + \frac{[K_{\mathcal{I}\mathcal{I}}^{q_t}]_{j,k}}{[K_{\mathcal{I}\mathcal{I}}^{p_t}]_{j,k}} - 1 \\ &\qquad + \frac{([f(\mathbf{X}_\mathcal{I}; \boldsymbol{\mu}_t)]_{j,k} - [f(\mathbf{X}_\mathcal{I}; \boldsymbol{\mu}_{t-1})]_{j,k})^2}{[K_{\mathcal{I}\mathcal{I}}^{p_t}]_{j,k}}\Bigg),\end{aligned} \tag{5}$$

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 used in tasks $t' \leq t-1$.

Proof. See Appendix A.2 □

3.3. Relationship to Other Function-Space Objectives

Functional regularization for continual learning (FRCL; (Titsias et al., 2020)) and functional regularization of the memorable past (FROMP; (Pan et al., 2020)) use probabilistic objective functions conceptually similar to Equation (3) and mathematically similar to Equation (5). Here, we make the link between C-FSVI and FROMP explicit and show that FROMP and FRCL either rely on explicit parameter-space regularization or incomplete function-space regularization:

Proposition 3 (Correspondence between C-FSVI and and FROMP). *Let $\tilde{\mathcal{F}}(q_t(\boldsymbol{\theta}))$ be as defined. Then, up to a multiplicative constant, the FROMP objective (Pan et al., 2020) corresponds to the C-FSVI objective $\tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ under a block-diagonal approximation without inter-task dependence, the prior covariance given by the Laplace approximation about $\boldsymbol{\mu}_{t-1}$, and the variational distribution given by a Dirac delta distribution $q_t(\boldsymbol{\theta}) \doteq \delta(\boldsymbol{\theta} - \boldsymbol{\mu}_t)$. Letting the prior covariance under the Laplace approximation about $\boldsymbol{\mu}_{t-1}$ be denoted by $\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})$, the FROMP objective can be expressed in terms of the C-FSVI objective as*

$$\mathcal{L}^{\text{FROMP}}(\boldsymbol{\mu}_t) = \tilde{\mathcal{F}}(\delta(\boldsymbol{\theta} - \boldsymbol{\mu}_t); \hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})) - \mathcal{V}(\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})), \tag{6}$$

where

$$\mathcal{V}(\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})) \doteq -\frac{1}{2}\sum_{j,k}\left(\log\frac{[\bar{K}_{\mathcal{I}\mathcal{I}}^{\hat{p}_t}]_{j,k}}{[\bar{K}_{\mathcal{I}\mathcal{I}}^{q_t}]_{j,k}} + \frac{[\bar{K}_{\mathcal{I}\mathcal{I}}^{q_t}]_{j,k}}{[\bar{K}_{\mathcal{I}\mathcal{I}}^{\hat{p}_t}]_{j,k}} - 1\right), \tag{7}$$

$\bar{K}$ denotes the covariance under a block-diagonalization without inter-task dependence, and

$$\bar{K}_{\mathcal{I}\mathcal{I}}^{\hat{p}_t} \doteq \text{block-diag}\left(\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top\right). \tag{8}$$

That is, the FROMP objective is only a function of $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{t-1}$, as it is defined for a restrictive variational family without a separately parameterized (co-)variance over functions $\bar{K}_{\mathcal{I}\mathcal{I}}^{\hat{q}_t}$, and considers a prior distribution over functions induced by the Laplace approximation about the previous mean parameters $\boldsymbol{\mu}_{t-1}$.

Proof. See Appendix A.3 □

Similarly, we can relate C-FSVI to FRCL:

Proposition 4 (Correspondence between C-FSVI and FRCL). *Let $\tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ be as defined. Then the FRCL objective (Titsias et al., 2020) corresponds to the C-FSVI objective $\tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ applied to Bayesian linear models under a block-diagonal approximation without inter-task dependence and with an additional weight-space KL divergence penalty and C-FSVI, that is, for a Bayesian linear model $f(\cdot; \boldsymbol{\Theta}) = \Phi_\psi(\cdot)\boldsymbol{\Theta}$, where $\Phi_\psi(\cdot)$ is a deterministic feature map parameterized by parameters ψ, $p_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$, and $q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$,*

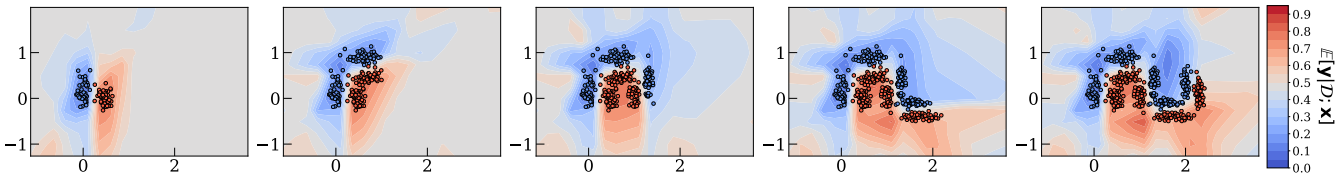
$$\mathcal{L}^{\text{FRCL}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) - \mathbb{D}_{\text{KL}}(q_t(\boldsymbol{\theta}) \,\|\, p_t(\boldsymbol{\theta})). \tag{9}$$

Proof. See Appendix A.3 □

4. Related Work

Approaches to continual learning can be viewed as belonging to three non-mutually-exclusive categories. Objective-based approaches modify the training objective used to optimize a predictive model. Replay-based approaches summarize past tasks using either stored data or freshly generated synthetic data. Architecture-based approaches change the model configuration from one task to another. For extensive reviews, see Lange et al. (2021) and Parisi et al. (2019). Our method centres around a new function-space variational objective, so we focus on objective-based approaches here.

Function-Space Regularization. Retaining a model's previously learned abilities requires maintaining a predictive function consistent with those from previous tasks. One way of achieving this is to modify the training objective so as to discourage changes in the model's predictions or internal representations from one task to another.



(a) Task 1 **(b)** Task 2 **(c)** Task 3 **(d)** Task 4 **(e)** Task 5

Figure 2. A visual demonstration of continual function-space variational inference (C-FSVI) on a task sequence based on synthetic 2D data. The model gradually infers the decision boundary between the two data clusters while maintaining high predictive uncertainty away from the data. The experiment setup is described in detail in Appendix B.

Learning without forgetting (Li & Hoiem, 2018) includes in the training objective a modified cross-entropy loss that penalizes the difference between the predictions of the current model on the current task data and the predictions of the previous model on the current task data. Less-forgetful learning (Jung et al., 2018) employs the same approach but uses squared Euclidean distance rather than the modified cross-entropy loss and applies it to the penultimate-layer representations rather than the model's predictions. Keep and learn (Kim et al., 2018) also uses internal representations as a basis for regularization. In contrast with these other methods, Benjamin et al. (2019) proposed comparing the current model with all previous versions of the model and on data from all past tasks instead of considering only the most recent model on data from the current task. In this method, pairs of models are compared by computing the Euclidean distance between the models' predictions. While these approaches mitigate catastrophic forgetting, they do not explicitly account for predictive uncertainty, which is an issue if a model is a poor fit to the data.

More recent work improved on past methods by taking a probabilistic approach to function-space regularization, constraining predictions to be consistent with prior distributions over functions instead of with exact predictions. Functional regularization for continual learning (FRCL; Titsias et al. (2020)) considers a Bayesian linear model with neural-network feature maps. Based on the duality between parameter space and function space in this model, the FRCL objective includes the KL divergence between predictive distributions at a selection of input points. This encourages similarity between the current predictive distribution and those from past tasks. FRCL is theoretically appealing, building on a well-understood method for stochastic variational inference using inducing points. But its regularization only acts on the final-layer parameters, not on the parameters of the neural-network feature maps. In contrast with FRCL, functional regularization of the memorable past (FROMP; Pan et al. (2020)) maintains a posterior distribution over all the parameters of a neural network. While FROMP achieves state-of-the-art performance on several continual-learning task sequences, it relies on a change in the underlying model and uses a surrogate objective for optimization, which divorces it from function-space variational objectives. As we show, this results in suboptimal performance compared to C-FSVI, which maintains a clear link to the underlying Bayesian approximation.

Parameter-Space Regularization. Historically more common than function-space approaches are those in which regularization is expressed more directly in terms of the parameters of a model. Among these, most relevant to our work are those that approximate Bayesian updating, in which the posterior from the previous task forms the prior for the current task. A key idea is shared between many of these methods: for each parameter, apply a penalty on the difference between its current setting and its prior setting, weighted by a measure of the parameter's importance.

Methods vary in how they measure importance. Variational continual learning (VCL; Nguyen et al. (2018); Swaroop et al. (2019)) uses the parameter covariance matrix of the model currently serving as the prior. Elastic weight consolidation (EWC; Kirkpatrick et al. (2017)) and its successors (Chaudhry et al., 2018; Lee et al., 2017; Liu et al., 2018; Schwarz et al., 2018)) use a Fisher information matrix computed on each task. Online structured Laplace (Ritter et al., 2018) and second-order loss approximation (Yin et al., 2020a) respectively use Kronecker-factored and low-rank Hessians. Synaptic intelligence (SI; Zenke et al. (2017a)) uses a cumulative sum of the gradient of the training objective with respect to the parameters. Memory-aware synapses (MAS; Aljundi et al. (2018)) use the gradient of the model output with respect to the parameters.

Other related work includes various modifications to VCL (Ahn et al., 2019; Kessler et al., 2019), VCL-inspired uncertainty-guided continual BNNs Ebrahimi et al. (2020), and a variation of SI known as asymmetric loss approximation with single-side overestimation (Park et al., 2019). There have also been efforts to conceptually unify some of the approaches outlined above. Loo et al. (2020) drew a link between VCL and online EWC. Chaudhry et al. (2018) combined EWC and SI in a single method. Yin et al. (2020b) generalized EWC, online structured Laplace, SI and MAS.

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. This accuracy is, where possible, based on multiple random seeds (10 for C-FSVI) with different random seeds, with both the average value and standard error reported. For each task sequence, all methods use the same architecture and coreset size unless explicitly indicated otherwise. For details about models and training procedures, see Appendix B. [1]Accuracies computed using best coreset selection method out of random selection and k-center selection. [2]Uses random coreset selection. [3]Uses task identifiers for permuted MNIST. [4]See Table 2 in Appendix B (MLPs were used in all experiments). [5]Evaluates KL divergence at points sampled from the empirical data distribution of the current task. [6]Uses one sample per class as coreset (see Appendix B for further details).

Method	Split MNIST (MH)	Split FMNIST (MH)	Permuted MNIST (SH)	Split MNIST (SH)
EWC (Kirkpatrick et al., 2017)	63.10%	—	84.00%	—
SI (Zenke et al., 2017a)	98.90%	—	86.00%	—
VCL (Nguyen et al., 2018)[1]	98.40%	98.60%$\pm$0.04	93.00%	32.11%$\pm$1.16
VCL (no coreset)	97.00%	89.60%$\pm$1.75	—	—
FRCL (Titsias et al., 2020)[3]	97.80%$\pm$0.22	97.28%$\pm$0.17	94.30%$\pm$0.06	—
FROMP (Pan et al., 2020)	99.00%$\pm$0.04	99.00%$\pm$0.03	94.90%$\pm$0.04	35.29%$\pm$0.52
VAR-GP (Kapoor et al., 2021)	—	—	97.20%$\pm$0.08	90.57%$\pm$1.06
C-FSVI[2]	99.54%$\pm$0.04	99.19%$\pm$0.02	95.76%$\pm$0.02	92.87%$\pm$0.14
C-FSVI (larger networks)[4]	**99.77%**$\pm$0.00	99.16%$\pm$0.03	**97.50%**$\pm$0.01	**93.38%**$\pm$0.10
C-FSVI (no coreset)[5]	99.62%$\pm$0.02	**99.54%**$\pm$0.01	—	—
C-FSVI (minimal coreset)[6]	—	—	89.59%$\pm$0.30	51.44%$\pm$1.22

5. Empirical Evaluation

Through a systematic empirical study, we demonstrate how C-FSVI works in practice. Details regarding hyperparameter selection, experiment setups, optimization routines, etc. can be found in Appendix B.

5.1. Overall Performance

To highlight the performance benefits of C-FSVI, we compare it with alternative objective-based methods for continual learning. We do this on four task sequences of varying difficulty (see Appendix B for further details):

1. (Multi-Head) Split MNIST: five binary classification tasks; task identifier provided at test time;
2. (Multi-Head) Split FashionMNIST: five binary classification tasks; task identifier provided at test time;
3. (Single-Head) Permuted MNIST: ten ten-way classification tasks; task identifier not provided at test time;
4. (Single-Head) Split MNIST: five ten-way classification tasks; task identifier not provided at test time.

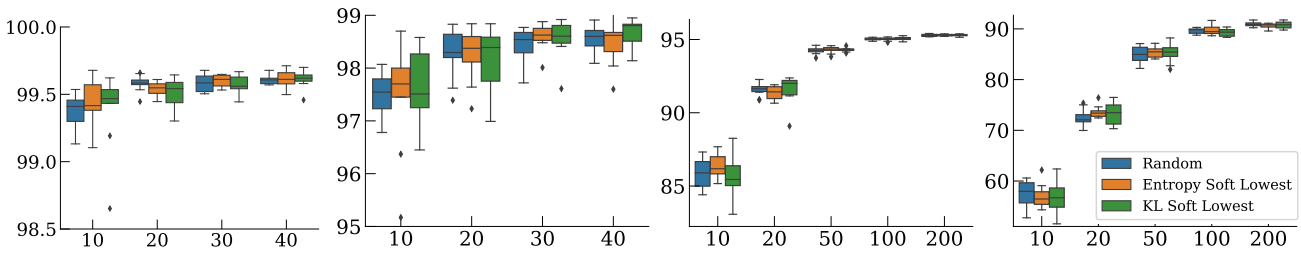
C-FSVI outperforms all related methods by a statistically significant margin on both multi-head task sequences (Table 1). On the single-head task sequences, it performs on par with VAR-GP, which uses a non-parametric model and is therefore not directly comparable to (parameterized) BNN models, whereas C-FSVI outperforms all related methods on the split MNIST without task identifiers.

We further note that C-FSVI achieves state-of-the-art predictive accuracies on settings with task identifiers *without* a coreset. In contrast, prior works such as FROMP and FRCL explicitly rely on careful coreset selection, while C-FSVI does not require a coreset at all.

Finally, we consider using a minimal coreset of only one data point per class for settings without task identifiers, finding that C-FSVI performs competitively on permuted MNIST, outperforming EWC and SI. On split MNIST without task identifiers, using a minimal coreset of one data point per class (i.e., two data points per task) results in a significantly lower performance than when using a coreset, but still significantly outperforms both VCL and FROMP, both with coresets of 40 points per task. This further illustrates C-FSVI's comparably low level of dependence on suitably chosen data points even for challenging task sequences.

5.2. Illustrative Example

To provide intuition for how C-FSVI allows learning on new tasks while maintaining previously acquired abilities, we demonstrate its use on a task sequence based on easy-to-visualize synthetic 2D data, originally proposed by Pan et al. (2020). In this scenario, each data point belongs to one of two classes, and more data points are revealed as the task sequence progresses. While there is some aleatoric uncertainty at the boundary of the two classes, the data-generating process is assumed to reveal data from mostly non-overlapping subsets of the input space. The continual-learning problem is then to infer the decision boundary around data points revealed up to and including the current task without forgetting the decision boundary inferred on previous tasks. We consider a single-head model that has to solve a binary classification problem at every task.

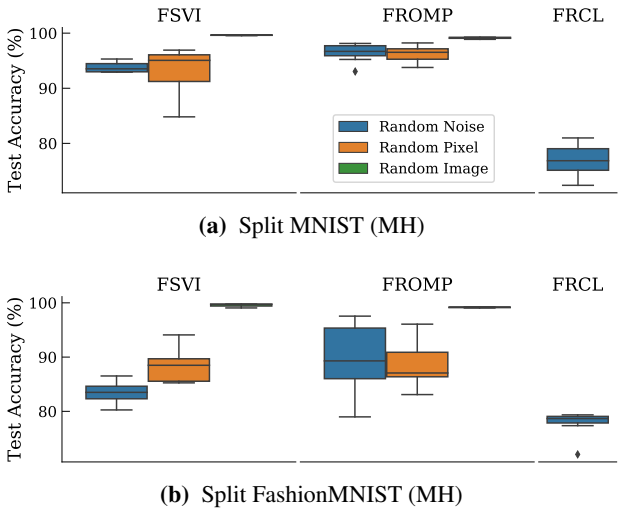


(a) Split MNIST (MH) **(b)** Split FashionMNIST (MH) **(c)** Permuted MNIST (SH) **(d)** Split MNIST (SH)

Figure 3. Comparison showing the effect of varying the coreset size (horizontal axis) and changing the coreset-selection method (color) on the predictive accuracy of C-FSVI (vertical axis). We consider three coreset-selection methods: sampling data points with uniform probability; sampling with probability proportional to the entropy of the model's posterior predictive distribution; sampling with probability proportional to the KL divergence between the posterior predictive distribution and the prior predictive distribution. The number of inducing points is ten in all cases. No coreset-selection method consistently yields higher accuracy.

In Figure 2, we plot the model's posterior predictive distribution after training on each of five tasks. As can be seen in Figure 2a, after training on task 1, the posterior predictive distribution exhibits low uncertainty on the two data manifolds and high uncertainty (class probabilities around 0.5) everywhere else. On task 2, the C-FSVI seeks to match the variational distribution over functions inferred in the previous task while attaining a posterior predictive distribution that fits the new set of data points. As can be seen in Figure 2b, C-FSVI achieves this and has expanded the area in input space where the model is confident in its predictions.

As more tasks and data are revealed, C-FSVI allows the model to continually explore the data space and infer the decision boundary while maintaining accurate, high-confidence predictions on data points in parts of the inputs space where it was previously trained on observed data. Finally, as shown in Figure 2e, after training on five tasks, the model has inferred the decision boundary between the two data sets, while maintaining high predictive uncertainty in parts of the input space, where no data points have been observed yet. Unlike deterministic neural networks, which tend to make highly confident predictions in parts of the inputs space where no data has been observed, or on data points that lie outside of the distribution of the training data, the model maintains high predictive uncertainty away from the data, which makes makes it easier for the optimization procedure to adopt to new tasks.



(a) Split MNIST (MH)

(b) Split FashionMNIST (MH)

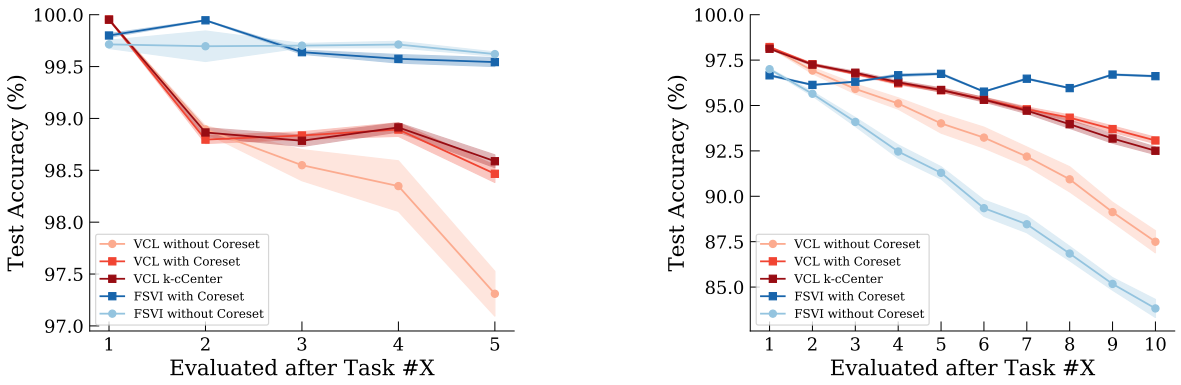
Figure 4. Comparison of predictive accuracies of C-FSVI, FROMP and FRCL on multi-head tasks *without* using coresets. Evaluation points are sampled according to three different sampling schemes derived from the current task's empirical data distribution.

5.3. Coreset

Similar to existing methods such as FROMP and FRCL, C-FSVI includes in the training objective a function-space regularization term that encourages matching between the posterior and prior distributions over functions at a selection of data points. Whereas many alternative methods require an expensive process to select data points from previous tasks, C-FSVI achieves a strong performance even with random coreset selection. In some settings, C-FSVI works well with an extremely small coreset or even without using a any points from previous tasks at all.

C-FSVI with Different Methods for Coreset Selection. Figure 3 shows the predictive accuracy of C-FSVI on different tasks, different coreset-selection methods, and varying coreset sizes. As can be seen, uniform-random selection of coreset points performs as well as other methods, indicating that C-FSVI is not sensitive to the choice of coreset.

C-FSVI without a Coreset. We evaluate how well C-FSVI performs without selecting a coreset and instead only sam-



(a) Split MNIST (MH) **(b)** Permuted MNIST (SH)

Figure 5. Comparison of predictive accuracies of function-space variational inference (C-FSVI; shades of blue) and parameter-space variational inference (VCL; shades of red). One standard error around the mean across 10 seeds is plotted as shaded area. With a coreset, C-FSVI outperforms VCL on both task sequences. Without a coreset, C-FSVI exhibits weak performance on permuted MNIST (SH).

pling white noise or images derived from the empirical data distribution of the current task. Figure 4 shows the predictive accuracies of C-FSVI, FROMP, and FRCL for different sampling schemes for Split MNIST (MH) and Split FashionMNIST (MH) and demonstrates that C-FSVI with KL divergence evaluation points sampled exclusively from the current task outperforms other methods (also see Table 1).

5.4. Function-Space vs. Parameter-Space Inference

Finally, we perform a direct comparison between function- and parameter-space variational inference. A direct comparison between function- and parameter-space views is important to elucidate the importance of informative prior distributions over functions for successful continual learning. Specifically, by the data-processing inequality, we would expect an information loss when applying a non-injective mapping to the distributions over parameters, which may be reflected in the usefulness of the KL divergence for continual learning.

As can be seen in Figure 5, C-FSVI consistently outperforms VCL on Split MNIST (MH) and Split Permuted MNIST (SH) and, crucially, maintains its predictive performance over time, whereas the predictive performance of VCL steadily degrades. This result, while not general, suggests that the information loss in the KL divergence between distributions over functions compared to the KL divergence between distributions over parameters is not sufficiently significant to offset the advantage of an explicitly-defined prior distribution over functions. Figure 5a, however, does expose a potential failure mode in C-FSVI's reliance on random sampling methods when not using a coreset (light-blue line). We hope to address this deficiency in follow-up work.

6. Discussion

We demonstrated that C-FSVI is able to achieve high predictive accuracy across task sequences. However, in our empirical evaluation, we found that the impact of maintaining a set of data points from past tasks varies significantly between single- and multi-head settings. Specifically, we observed that single-head Split MNIST, where the data points associated with different tasks have little to no overlap in image space (e.g., the appearance of the digit "2" is distinct from the appearance of the digit "8"), is challenging for all methods—even when maintaining a coreset—and C-FSVI only achieved an accuracy of approximately 51% with a minimal coreset of one data point per class per task. While task sequences like multi-head Split MNIST/FashionMNIST with predictive accuracies of 99.77%/99.54% under C-FSVI without a coreset may be considered solved, single-head settings without coresets remain challenging.

7. Conclusion

We presented continual function-space variational inference (C-FSVI), a method for continual learning in Bayesian deep neural networks and established a precise mathematical correspondence between C-FSVI and existing function-space regularization methods. We showed that C-FSVI yields significantly improved and more consistently high predictive performance than alternative approaches based on parameter- and function-space regularization. Lastly, we demonstrated that C-FSVI does not rely on careful coreset selection, as is the case for existing function-space regularization methods, and in multi-head settings can achieve state-of-the-art performance even without coresets.

Acknowledgements

Tim G. J. Rudner is funded by the Rhodes Trust and the Engineering and Physical Sciences Research Council (EPSRC). We gratefully acknowledge donations of computing resources by the Alan Turing Institute.

References

Ahn, H., Cha, S., Lee, D., and Moon, T. Uncertainty-based continual learning with adaptive regularization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/2c3ddf4bf13852db711dd1901fb517fa-Paper.pdf.

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision – ECCV 2018*, pp. 144–161, Cham, 2018. Springer International Publishing.

Benjamin, A., Rolnick, D., and Kording, K. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkMwpiR9Y7.

Burt, D. R., Ober, S. W., Garriga-Alonso, A., and van der Wilk, M. Understanding variational inference in function-space. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL https://openreview.net/forum?id=7P9y3sRa5Mk.

Chaudhry, A., Dokania, P., Ajanthan, T., and Torr, P. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018.

de G. Matthews, A. G., Hensman, J., Turner, R., and Ghahramani, Z. On sparse variational methods and the kullback-leibler divergence between stochastic processes. volume 51 of *Proceedings of Machine Learning Research*, pp. 231–239, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/matthews16.html.

Ebrahimi, S., Elhoseiny, M., Darrell, T., and Rohrbach, M. Uncertainty-guided continual learning with bayesian neural networks. *International Conference on Learning Representations*, 2020.

Farquhar, S. and Gal, Y. Towards Robust Evaluations of Continual Learning. *Lifelong Learning: A Reinforcement Learning Approach Workshop at ICML*, 2018.

Jung, H., Ju, J., Jung, M., and Kim, J. Less-forgetful learning for domain expansion in deep neural networks. In *AAAI*, 2018.

Kapoor, S., Karaletsos, T., and Bui, T. D. Variational auto-regressive gaussian processes for continual learning, 2021.

Kessler, S., Nguyen, V., Zohren, S., and Roberts, S. Hierarchical indian buffet neural networks for bayesian continual learning. *arXiv preprint arXiv:1912.02290*, 2019.

Kim, H.-E., Kim, S., and Lee, J. Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G. (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 520–528, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00928-1.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2017.

Lange, M. D., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021.

Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., and Zhang, B.-T. Overcoming catastrophic forgetting by incremental moment matching. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2935–2947, 2018.

Liu, X., Masana, M., Herranz, L., van de Weijer, J., López, A. M., and Bagdanov, A. D. Rotate your networks: Better weight consolidation and less catastrophic forgetting. *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2262–2268, 2018.

Loo, N., Swaroop, S., and Turner, R. E. Generalized variational continual learning. *International Conference on Learning Representations*, 2020.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational continual learning. *International Conference on Learning Representations*, 2018.

Pan, P., Swaroop, S., Immer, A., Eschenhagen, R., Turner, R., and Khan, M. E. E. Continual deep learning by functional regularisation of memorable past. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4453–4464. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/2f3bbb9730639e9ea48f309d9a79ff01-Paper.pdf.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural networks : the official journal of the International Neural Network Society*, 113:54–71, 2019.

Park, D., Hong, S., Han, B., and Lee, K. M. Continual learning by asymmetric loss approximation with single-side overestimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3334–3343, 2019.

Polyanskiy, Y. and Wu, Y. Strong data-processing inequalities for channels and bayesian networks. In Carlen, E., Madiman, M., and Werner, E. M. (eds.), *Convexity and Concentration*, pp. 211–249, New York, NY, 2017. Springer New York. ISBN 978-1-4939-7005-6.

Ritter, H., Botev, A., and Barber, D. Online structured laplace approximations for overcoming catastrophic forgetting. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/f31b20466ae89669f9741e047487eb37-Paper.pdf.

Rudner, T. G. J., Chen, Z., and Gal, Y. Rethinking Function-Space Variational Inference in Bayesian Neural Networks. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/schwarz18a.html.

Swaroop, S., Nguyen, C. V., Bui, T. D., and Turner, R. E. Improving and understanding variational continual learning. *Continual Learning Workshop, Neural Information Processing Systems*, 2019.

Titsias, M. K. Variational learning of inducing variables in sparse gaussian processes. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/titsias09a.html.

Titsias, M. K., Schwarz, J., de G. Matthews, A. G., Pascanu, R., and Teh, Y. W. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxCzeHFDB.

Yin, D., Farajtabar, M., and Li, A. Sola: Continual learning with second-order loss approximation. *ArXiv*, abs/2006.10974, 2020a.

Yin, D., Farajtabar, M., Li, A., Levine, N., and Mott, A. Optimization and generalization of regularization-based continual learning: a loss approximation viewpoint. *arXiv: Learning*, 2020b.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 06–11 Aug 2017a. URL http://proceedings.mlr.press/v70/zenke17a.html.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017b.

Supplementary Materials

A. Assumptions & Proofs

A.1. Approximations for Function-Space Variational Inference

Below, we restate the approximations made in Rudner et al. (2021) to derive the variational objective in Equation (4).

To obtain a tractable distribution over functions and a tractable KL divergence, we make two variational assumptions and two approximations to obtain a tractable estimator for the function-space variational objective.

We consider the distribution over parameters that gives rise to the distribution over functions $q(f(\cdot\,;\boldsymbol{\theta}))$ and assume that the variational distribution over functions is induced by a mean-field Gaussian distribution over parameters:

Approximation 1 (Gaussian Mean-Field Variational Distribution over Parameters)**.** *Assume a factorized Gaussian variational distribution over parameters, $q(\boldsymbol{\theta}) \doteq \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Define a variational distribution over functions $q(f(\cdot\,;\boldsymbol{\theta}))$ as the distribution induced by the variational distribution over parameters $q(\boldsymbol{\theta})$ under the mapping f.*

We further make an assumption about how the distribution over functions $q(f(\cdot\,;\boldsymbol{\theta})) = q(f(\mathbf{X}_*;\boldsymbol{\theta}), f(\mathbf{X}_\mathcal{I};\boldsymbol{\theta}))$ factorizes, where $\mathbf{X}_\mathcal{I}$ is a finite set of so-called inducing points and $\mathbf{X}_* \doteq \mathcal{X}\backslash\mathbf{X}_\mathcal{I}$ is an infinite set of evaluation points containing all points in the data space except for $\mathbf{X}_\mathcal{I}$. Specifically, we assume prior conditional matching, that is:

Approximation 2 (Prior Conditional Matching (Titsias, 2009; de G. Matthews et al., 2016))**.** *Let the variational distribution over functions factorize as*

$$q(f(\mathbf{X}_*;\boldsymbol{\theta}), f(\mathbf{X}_\mathcal{I};\boldsymbol{\theta})) \doteq p(f(\mathbf{X}_*) \mid f(\mathbf{X}_\mathcal{I}))q(f(\mathbf{X}_\mathcal{I};\boldsymbol{\theta})),$$

where $p(f(\mathbf{X}_*) \mid f(\mathbf{X}_\mathcal{I}))$ is the conditional prior distribution over functions under the mapping f and some prior distribution over parameters $p(\boldsymbol{\theta})$.

To apply prior conditional matching to the function-space variational objective and maintain marginal consistency, we make the following approximation:

Approximation 3 (Marginal Consistency on Observed Data)**.** *Assume that*

$$q(f(\mathbf{X}_\mathcal{D}\,;\boldsymbol{\theta})) = \int p(f(\mathbf{X}_\mathcal{D}\,;\boldsymbol{\theta}), f(\mathbf{X}^\mathsf{c}\,;\boldsymbol{\theta}) \mid f(\mathbf{X}_\mathcal{I};\boldsymbol{\theta}))\, q(f(\mathbf{X}_\mathcal{I};\boldsymbol{\theta}))\, df(\mathbf{X}^\mathsf{c};\boldsymbol{\theta})\, df(\mathbf{X}_\mathcal{I};\boldsymbol{\theta}).$$

Finally, we consider a linearization of the mapping f, which we will use to obtain a tractable estimator of the function-space KL divergence:

Approximation 4 (Linearization about Parameters)**.** *For mapping a f, stochastic parameters $\boldsymbol{\Theta}$ with mean $\mathbf{m} \doteq \mathbb{E}[\boldsymbol{\Theta}]$, and Jacobian $\mathcal{J}_\mathbf{m}(\cdot) \doteq \frac{\partial f(\cdot\,;\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}|_{\boldsymbol{\Theta}=\mathbf{m}}$, define the linearization of the stochastic function $f(\cdot\,;\boldsymbol{\Theta})$ about $\mathbf{m}$ by*

$$f(\cdot\,;\boldsymbol{\Theta}) \approx \tilde{f}(\cdot\,;\boldsymbol{\Theta}) \doteq f(\cdot\,;\mathbf{m}) + \mathcal{J}_\mathbf{m}(\cdot)(\boldsymbol{\Theta} - \mathbf{m}).$$

Due to local linearity, the approximation $\tilde{f}(\cdot\,;\boldsymbol{\Theta})$ will be accurate for realizations $\boldsymbol{\theta}$ close to $\boldsymbol{\mu}$, and hence, the distribution over the linearized stochastic function $\tilde{f}(\cdot\,;\boldsymbol{\Theta})$ (induced by a distribution over the parameters $\boldsymbol{\Theta}$), denoted by $\tilde{q}(\tilde{f}(\cdot\,;\boldsymbol{\theta}))$ will be close to the distribution over $f(\cdot\,;\boldsymbol{\Theta})$ for small variance parameters $\boldsymbol{\Sigma}$.

A.2. Variational Objective and Distributions over Linearized Mapping

Proposition 1 (Continual Function-Space Variational Objective (adapted from Rudner et al. (2021)))**.** *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 its mean parameters be given by*

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

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

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

respectively. Under the approximations in Appendix A.1, we obtain the variational objective

$$\bar{\mathcal{F}}(q_t(\boldsymbol{\theta})) \doteq \mathbb{E}_{q_t(f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta}))}[\log p(\mathbf{y}_{\mathcal{D}_t} \mid f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta})] - \mathbb{D}_{\mathrm{KL}}(\tilde{q}_t(\tilde{f}(\mathbf{X}_\mathcal{I};\boldsymbol{\theta})) \,\|\, \tilde{p}_t(\tilde{f}(\mathbf{X}_\mathcal{I};\boldsymbol{\theta}))). \tag{A.1}$$

Proof. Rudner et al. (2021) show that for $\boldsymbol{\Theta}$, $f(\cdot\,;\boldsymbol{\theta})$, $\tilde{f}(\cdot\,;\boldsymbol{\theta})$, as defined above, the mean and variance of the distribution over $\tilde{f}(\cdot\,;\boldsymbol{\Theta})$ under $q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ evaluated at $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ are given by

$$\mathbb{E}_{\tilde{q}_t(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))}[\tilde{f}(\mathbf{X};\boldsymbol{\theta})] = f(\mathbf{X};\boldsymbol{\mu}_t) \quad \text{and} \quad \mathbb{V}(\tilde{f}(\mathbf{X};\boldsymbol{\theta})) = \mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X})\boldsymbol{\Sigma}_t\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X}')^\top, \tag{A.2}$$

and the distribution $\tilde{q}$ over $\tilde{f}(\mathbf{X}\,;\boldsymbol{\Theta})$ is a multivariate Gaussian distribution:

$$\tilde{q}_t(\tilde{f}(\mathbf{X}\,;\boldsymbol{\theta})) = \mathcal{N}(f(\mathbf{X}\,;\boldsymbol{\mu}_t), \mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X})\boldsymbol{\Sigma}_t\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X}')^\top). \tag{A.3}$$

Similarly, the mean and variance of the distribution over $\tilde{f}(\cdot\,;\boldsymbol{\Theta})$ under $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ evaluated at $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ are given by

$$\mathbb{E}_{\tilde{p}_t(\tilde{f}(\mathbf{X};\boldsymbol{\theta}))}[\tilde{f}(\mathbf{X};\boldsymbol{\theta})] = f(\mathbf{X};\boldsymbol{\mu}_{t-1}) \quad \text{and} \quad \mathbb{V}(\tilde{f}(\mathbf{X};\boldsymbol{\theta})) = \mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\boldsymbol{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top, \tag{A.4}$$

and the distribution $\tilde{p}$ over $\tilde{f}(\mathbf{X}\,;\boldsymbol{\Theta})$ is a multivariate Gaussian distribution:

$$\tilde{p}_t(\tilde{f}(\mathbf{X}\,;\boldsymbol{\theta})) = \mathcal{N}(f(\mathbf{X}\,;\boldsymbol{\mu}_{t-1}), \mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\boldsymbol{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top). \tag{A.5}$$

Furthermore, Rudner et al. (2021) show that, under Approximations 1, 2, 3, and 4, we obtain the function-space variational objective

$$\bar{\mathcal{F}}(q_t(\boldsymbol{\theta})) \doteq \mathbb{E}_{q_t(f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta}))}[\log p(\mathbf{y}_{\mathcal{D}_t} \mid f(\mathbf{X}_{\mathcal{D}_t};\boldsymbol{\theta})] - \mathbb{D}_{\mathrm{KL}}(\tilde{q}_t(\tilde{f}(\mathbf{X}_{\mathcal{I}};\boldsymbol{\theta})) \,\|\, \tilde{p}_t(\tilde{f}(\mathbf{X}_{\mathcal{I}};\boldsymbol{\theta}))), \tag{A.6}$$

where $\mathbb{D}_{\mathrm{KL}}(\tilde{q}_t(\tilde{f}(\mathbf{X}_{\mathcal{I}};\boldsymbol{\theta})) \,\|\, \tilde{p}_t(\tilde{f}(\mathbf{X}_{\mathcal{I}})))$ is analytically tractable. This concludes the proof. □

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 function,*

$$K_{\mathcal{I}\mathcal{I}}^{p_t} \doteq \mathrm{diag}\left(\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\boldsymbol{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top\right) \qquad K_{\mathcal{I}\mathcal{I}}^{q_t} \doteq \mathrm{diag}\left(\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X})\boldsymbol{\Sigma}_t\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X}')^\top\right)$$

the objective can be optimized via stochastic variational inference on

$$\begin{aligned}\bar{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \frac{1}{S}\sum_{i=1}^{S} &\log p(\mathbf{y}_{\mathcal{B}_t} \mid f(\mathbf{X}_{\mathcal{B}_t}; h(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\epsilon}^{(i)}))) \\ &- \sum_{k=1}^{Q_t}\sum_{j=1}^{|\mathbf{X}_{\mathcal{I}}|}\frac{1}{2}\left(\log\frac{[K_{\mathcal{I}\mathcal{I}}^{p_t}]_{j,k}}{[K_{\mathcal{I}\mathcal{I}}^{q_t}]_{j,k}} + \frac{[K_{\mathcal{I}\mathcal{I}}^{q_t}]_{j,k}}{[K_{\mathcal{I}\mathcal{I}}^{p_t}]_{j,k}} - 1 + \frac{([f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_t)]_{j,k} - [f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_{t-1})]_{j,k})^2}{[K_{\mathcal{I}\mathcal{I}}^{p_t}]_{j,k}}\right),\end{aligned} \tag{A.7}$$

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 used in tasks $t' \leq t-1$.

Proof. Under the diagonalization assumptions about the distributions over the linearized mappings under the prior and variational distributions over parameters, the KL divergence can be expressed as the sum of KL divergences over inducing inputs and output dimensions. The result then follows immediately from the closed-form expression for the KL divergence between two univariate Gaussian distributions. The reparameterization of the stochastic parameters $\boldsymbol{\Theta}$ allows gradient-based optimization via the reparameterization gradient estimator. This concludes the proof. □

A.3. Derivation of Correspondence to Other Function-Space Objectives

Proposition 3 (Correspondence between C-FSVI and and FROMP)**.** *Let $\tilde{\mathcal{F}}(q_t(\boldsymbol{\theta}))$ be as defined. Then, up to a multiplicative constant, the* FROMP *objective (Pan et al., 2020) corresponds to the* C-FSVI *objective $\tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ under a block-diagonal approximation without inter-task dependence, the prior covariance given by the Laplace approximation about $\boldsymbol{\mu}_{t-1}$, and the variational distribution given by a Dirac delta distribution $q_t(\boldsymbol{\theta}) \doteq \delta(\boldsymbol{\theta} - \boldsymbol{\mu}_t)$. Letting the prior covariance under the Laplace approximation about $\boldsymbol{\mu}_{t-1}$ be denoted by $\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})$, the* FROMP *objective can be expressed in terms of the* C-FSVI *objective as*

$$\mathcal{L}^{\text{FROMP}}(\boldsymbol{\mu}_t) = \tilde{\mathcal{F}}(\delta(\boldsymbol{\theta} - \boldsymbol{\mu}_t); \hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})) - \mathcal{V}(\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})), \tag{A.8}$$

where

$$\mathcal{V}(\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})) \doteq -\frac{1}{2}\sum_{j,k}\left(\log\frac{[\bar{K}^{\hat{p}_t}_{\mathcal{I}\mathcal{I}}]_{j,k}}{[\bar{K}^{q_t}_{\mathcal{I}\mathcal{I}}]_{j,k}} + \frac{[\bar{K}^{q_t}_{\mathcal{I}\mathcal{I}}]_{j,k}}{[\bar{K}^{\hat{p}_t}_{\mathcal{I}\mathcal{I}}]_{j,k}} - 1\right), \tag{A.9}$$

$\bar{K}$ denotes the covariance under a block-diagonalization without inter-task dependence, and

$$\bar{K}^{\hat{p}_t}_{\mathcal{I}\mathcal{I}} \doteq \text{block-diag}\left(\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{\mu}_{t-1})\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top\right). \tag{A.10}$$

That is, the FROMP *objective is only a function of $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{t-1}$, as it is defined for a restrictive variational family without a separately parameterized (co-)variance over functions $\bar{K}^{\hat{q}_t}_{\mathcal{I}\mathcal{I}}$, and considers a prior distribution over functions induced by the Laplace approximation about the previous mean parameters $\boldsymbol{\mu}_{t-1}$.*

Proof. By Equation (8) in (Pan et al., 2020), the FROMP objective function is given by

$$\begin{aligned}\mathcal{L}^{\text{FROMP}}(\boldsymbol{\mu}_t) = \frac{1}{S}\sum_{i=1}^{S}\log p([\mathbf{y}_{\mathcal{B}_t}]_k \mid [f(\mathbf{X}_{\mathcal{B}_t};\boldsymbol{\mu}_t)]_k) \\ + \sum_{k=1}^{t-1}\frac{\tau}{2}\left([f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_t)]_k - [f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_{t-1})]_k\right)^\top [K^{\hat{p}_t}_{\mathcal{I}\mathcal{I}}]^{-1}_k \left([f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_t)]_k - [f(\mathbf{X}_{\mathcal{I}};\boldsymbol{\mu}_{t-1})]_k\right),\end{aligned} \tag{A.11}$$

with temperature parameter τ. The result follows directly from the definition of $\bar{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and $\tau = 1$. □

Proposition 4 (Correspondence between C-FSVI and FRCL)**.** *Let $\tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ be as defined. Then the* FRCL *objective (Titsias et al., 2020) corresponds to the* C-FSVI *objective $\tilde{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ applied to Bayesian linear models under a block-diagonal approximation without inter-task dependence and an additional weight-space KL divergence penalty and* C-FSVI*, that is, for a Bayesian linear model $f(\cdot\,;\boldsymbol{\Theta}) = \Phi_\psi(\cdot)\boldsymbol{\Theta}$, where $\Phi_\psi(\cdot)$ is a deterministic feature map parameterized by parameters ψ, $p_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$, and $q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$,*

$$\mathcal{L}^{\text{FRCL}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \bar{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) - \mathbb{D}_{\text{KL}}(q_t(\boldsymbol{\theta}) \,\|\, p_t(\boldsymbol{\theta})). \tag{A.12}$$

Proof. By Section 2.3 in Titsias et al. (2020), the FRCL objective function is given by

$$\mathcal{L}^{\text{FRCL}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \doteq \mathbb{E}_{q_t(\boldsymbol{\theta})}[\log p(\mathbf{y}_{\mathcal{B}_t} \mid \Phi(\mathbf{X}_{\mathcal{B}_t})\boldsymbol{\theta})] - \sum_{t'=1}^{t-1}\mathbb{D}_{\text{KL}}(\tilde{q}_{t'}(\tilde{f}(\mathbf{X}_{\mathcal{I}_{t'}};\boldsymbol{\theta})) \,\|\, \tilde{p}_{t'}(\tilde{f}(\mathbf{X}_{\mathcal{I}_{t'}};\boldsymbol{\theta}))) - \mathbb{D}_{\text{KL}}(q_t(\boldsymbol{\theta}) \,\|\, p_t(\boldsymbol{\theta})), \tag{A.13}$$

while the C-FSVI objective for a Bayesian linear model is

$$\tilde{\mathcal{F}}(q_t(\boldsymbol{\theta})) \doteq \mathbb{E}_{q_t(\boldsymbol{\theta})}[\log p(\mathbf{y}_{\mathcal{B}_t} \mid \Phi(\mathbf{X}_{\mathcal{B}_t})\boldsymbol{\theta})] - \mathbb{D}_{\text{KL}}(\tilde{q}_{t'}(\tilde{f}(\mathbf{X}_{\mathcal{I}_{t'}};\boldsymbol{\theta})) \,\|\, \tilde{p}_{t'}(\tilde{f}(\mathbf{X}_{\mathcal{I}_{t'}};\boldsymbol{\theta}))). \tag{A.14}$$

Letting

$$K^{p_t}_{\mathcal{I}\mathcal{I}} \doteq \text{block-diag}\left(\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X})\boldsymbol{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{X}')^\top\right) \qquad K^{q_t}_{\mathcal{I}\mathcal{I}} \doteq \text{block-diag}\left(\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X})\boldsymbol{\Sigma}_t\mathcal{J}_{\boldsymbol{\mu}_t}(\mathbf{X}')^\top\right) \tag{A.15}$$

be block diagonal matrices without inter-task dependence, with diagonal entries $\{K^{q_1}_{\mathcal{I}_1\mathcal{I}_1}, ..., K^{q_1}_{\mathcal{I}_{t-1}\mathcal{I}_{t-1}}\}$ and $\{K^{q_1}_{\mathcal{I}_1\mathcal{I}_1}, ..., K^{q_1}_{\mathcal{I}_{t-1}\mathcal{I}_{t-1}}\}$, respectively, computed from task-specific inducing inputs $\mathcal{I}_{t'}$. Then, since in general for any block diagonal matrix $A \in \mathbb{R}^{Jm\times Jm}$ with diagonal entries $\{A_1, ..., A_J\}$ and $A_j \in \mathbb{R}^{m\times m}$, the determinant can be expressed as $\det(A) = \prod_{1=j}^{J}\det(A_j)$ and for any $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_J]$, with $\mathbf{x}_j \in \mathbb{R}^m$, the square form $\mathbf{x}^\top A\mathbf{x}$ can be expressed as $\mathbf{x}^\top A\mathbf{x} = \sum_{j=1}^{J}\mathbf{x}_j^\top A_j\mathbf{x}_j$, we can write the C-FSVI objective as

$$\tilde{\mathcal{F}}(q_t(\boldsymbol{\theta})) \doteq \mathbb{E}_{q_t(\boldsymbol{\theta})}[\log p(\mathbf{y}_{\mathcal{B}_t} \mid \Phi(\mathbf{X}_{\mathcal{B}_t})\boldsymbol{\theta})] - \sum_{t'=1}^{t-1}\mathbb{D}_{\text{KL}}(\tilde{q}_{t'}(\tilde{f}(\mathbf{X}_{\mathcal{I}_{t'}};\boldsymbol{\theta})) \,\|\, \tilde{p}_{t'}(\tilde{f}(\mathbf{X}_{\mathcal{I}_{t'}};\boldsymbol{\theta}))), \tag{A.16}$$

since the KL divergence between multivariate Gaussians is a sum of log-determinants, traces, and a square form. The result follows immediately. □

B. Experiment Details

B.1. Task Sequences

We consider four continual-learning task sequences: split MNIST multi-head (MH), split MNIST single-head (SH), permuted MNIST single-head (SH), and split FashionMNIST multi-head (MH).

Multi-Head Setups. In a multi-head setup, a model performs a sequence of k-way classification tasks, using a different output head for each task. At test time, task identifiers are provided to the model to determine which output head to use for prediction. In split MNIST (MH) (Zenke et al., 2017b), each task is binary classification on a pair of MNIST classes. In split FashionMNIST (MH), each task is binary classification on a pair of FashionMNIST classes.

Single-Head Setups. In a single-head setup, a model performs k-way classification on a sequence of tasks, using the same output head throughout. In split MNIST (SH), each task is ten-way classification where the data from only two unique labels are provided (e.g. 0 and 1 for the first task, 2 and 3 for the second task, etc.). In permuted MNIST (SH), each task is 10-way classification on MNIST where the pixels of the input images undergo a fixed random permutation. We use ten tasks in total for permuted MNIST.

B.2. Training Details

Unless specified otherwise, the following setups apply to Figures 3 to 6 and **??** and Table 1.

Dataset. In all cases, 60,000 data samples are used for training and 10,000 data samples are used for testing. The input images are converted to float values in the range $[0, 1]$.

Network Size & Coreset Size. To ensure fair comparison, all methods in Table 1 (unless where explicitly indicated otherwise) use the same network size and coreset size (where applicable). Following prior works (Pan et al., 2020; Titsias et al., 2020), we use two-layer fully-connected neural networks with 100 hidden units per layer for permuted MNIST, and 256 units per layer for split settings (split MNIST MH and SH, split FashionMNIST MH). In all cases, the ReLU activation function is applied to non-output units. We use 200 coreset points for single-head settings (split MNIST SH, permuted MNIST), and 40 coreset points for multi-head settings (split MNIST, split Fashion MNIST).

Coreset Selection. For C-FSVI with a coreset, when training on the first task, 40 inducing points are generated by sampling each pixel uniformly from the range $[0, 1]$; during training on subsequent tasks, 40 inducing points are chosen randomly from the coreset. For C-FSVI without a coreset, when training on each task, 40 inducing inputs are chosen uniformly randomly from the training data of current task (corresponding to the "Random Image" label in Figure 3).

Prior Distribution. For the first task, C-FSVI uses a prior distribution over functions with fixed mean and diagonal covariance. When using a coreset, the prior distribution is assumed to be Gaussian with zero mean and a diagonal covariance of magnitude 0.001. When not using a coreset, the prior distribution is assumed to be Gaussian with zero mean and a diagonal covariance of magnitude 100. The prior variance is optimized via hyperparameter selection on a validation set.

Optimization. We use the Adam optimizer with an initial learning rate of 0.0005 ($\beta_1 = 0.9, \beta_2 = 0.999$). The number of epochs on each task for split MNIST (MH), split Fashion MNIST (MH), permuted MNIST (SH), split MNIST (SH) is 60, 60, 10, 80, respectively. The batch size is 128.

Prediction. The predictive distribution, used for computing the expected log-likelihood, is estimated using five Monte Carlo samples.

B.3. Hyperparameter Selection

For **C-FSVI** (optimized) in Table 1, we used the optimized hyperparameters chosen on a validation set after exploring the configurations shown in Table 2. For cases where no configuration is significantly better than the rest, the default value given in Appendix B.2 is used.

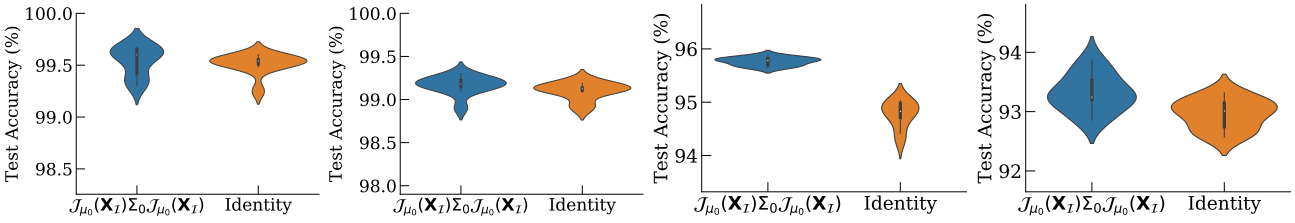
Table 2. Hyperparameter selection. Optimal values (in boldface) were chosen based on validation set accuracy. Standard errors were computed from ten random seeds.

Task Sequences	Number of Layers & Units	Magnitude of Prior Variance	Number of Epochs
Split MNIST (MH)	{**1**, 2} * {100, 200, 300, **400**}	{**0.001**, 0.01, 0.1, 1, 10, 100}	{40, **60**, 80, 120, 160}
Split Fashion MNIST (MH)	{**2**} * {50, **200**, 300, 400}	{**0.001**, 0.01, 0.1, 1, 10, 100}	{40, **60**, 80, 120, 160}
Permuted MNIST (SH)	{**2**} * {100, 200, 400, **500**}	{**0.001**, 0.01, 0.1, 1, 10, 100}	{10, **20**, 40, 60, 80}
Split MNIST (SH)	{**1**, 2} * {100, 200, 300, **400**}	{**0.001**, 0.01, 0.1, 1, 10, 100}	{**60**, 80, 120, 160, 240}

B.4. Visualization on Synthetic 2D Data

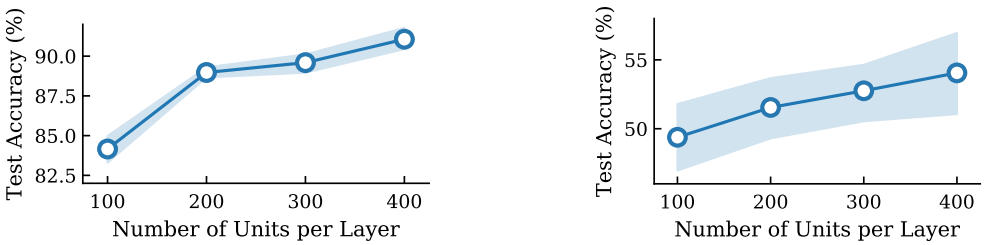
For the synthetic 2D dataset shown in Figure 2, we use the same setup as Pan et al. (2020). The dataset consists of five binary classification tasks with 2D inputs that are geometrically arranged for ease of interpretation. Each task consists of 3,600 training examples. Following Pan et al. (2020), we use a two-layer fully-connected neural network with 20 hidden units per layer and a two-dimensional output layer. For C-FSVI, the prior covariance is set to be $\mathbf{\Sigma}_0 = 0.1$. The model is trained for 250 epochs on each task. We use the Adam optimizer with an initial learning rate of 0.0005 ($\beta_1 = 0.9, \beta_2 = 0.999$) and a batch size of 128. The coreset is constructed by choosing 40 samples from the training data for each task. To evaluate the KL divergence between the variational and the prior distributions over functions, for each previous task, we sample 20 input points from the coreset and generate another 30 samples by sampling each pixel uniformly from the range $[-4, 4]$. For example, when we train the model on task $k \in \{1, 2, 3, \ldots\}$, we use $20(k-1)$ samples chosen from the coreset and $30k$ random-noise samples. The random-noise samples encourage the model to preserve high uncertainty in areas that are far from the training data.

C. Further Empirical Results



(a) Split MNIST (MH) **(b)** Split FashionMNIST (MH) **(c)** Permuted MNIST (SH) **(d)** Split MNIST (SH)

Figure 6. Comparison of predictive performance under the induced prior covariance function $K^{p_t}_{\mathcal{I}\mathcal{I}} = \text{diag}\left(\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{x})\mathbf{\Sigma}_{t-1}\mathcal{J}_{\boldsymbol{\mu}_{t-1}}(\mathbf{x}')^\top\right)$ (**left**) vs. an identity covariance function (**right**).



(a) Permuted MNIST (SH) **(b)** Split MNIST (SH)

Figure 7. Predictive accuracy under C-FSVI on permuted MNIST (SH) and split MNIST (SH) using only a minimal coreset of one sample per class, selected randomly, as a function of network width.