Posterior Meta-Replay for Continual Learning

Christian Henning* **Maria R. Cervera*** Francesco d'Angelo, Johannes von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F. Grewe and João Sacramento

Institute of Neuroinformatics, UZH / ETH Zurich, Switzerland, *These authors contributed equally to this work.

Introduction

Continual learning (CL) typically refers to the problem of sequentially learning a set of tasks $\mathcal{D}_1 \dots \mathcal{D}_T$, where $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{n_t} \overset{iid}{\sim} p_t(x) p_t(y \mid x)$.

Bayesian CL approaches commonly adopt a *prior-focused* view [1, 2, 3] and rely on a recursive Bayesian update to incorporate new tasks:

$$p(\mathbf{W} \mid \mathcal{D}_1, \mathcal{D}_2) = \frac{p(\mathcal{D}_2 \mid \mathbf{W}) p(\mathbf{W} \mid \mathcal{D}_1)}{p(\mathcal{D}_2 \mid \mathcal{D}_1)} \quad (1)$$

However approximations $q_{\boldsymbol{\theta}}^{(1:t)}(\mathbf{W}) \approx p(\mathbf{W} \mid \mathcal{D}_1, \dots \mathcal{D}_t)$ are necessary and can lead to practical challenges.

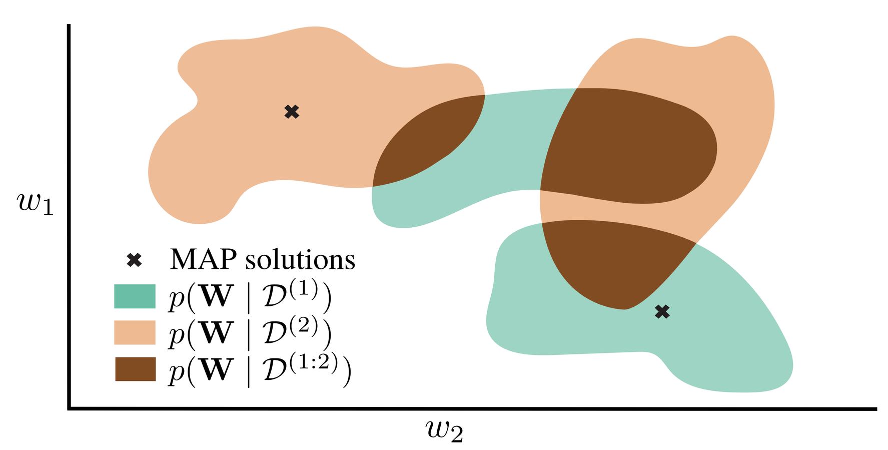

Figure 1: Bayesian CL approaches. While *prior-focused* CL is constrained to regions of overlap between task posteriors, *posterior meta-replay* can learn individual posteriors.

Motivation Can we overcome the limitations of *prior-focused* by learning task-specific posteriors?

Methods

To address this problem, we propose *posterior meta-replay*, a new Bayesian CL framework that compresses task-specific posteriors into a single shared meta-model.

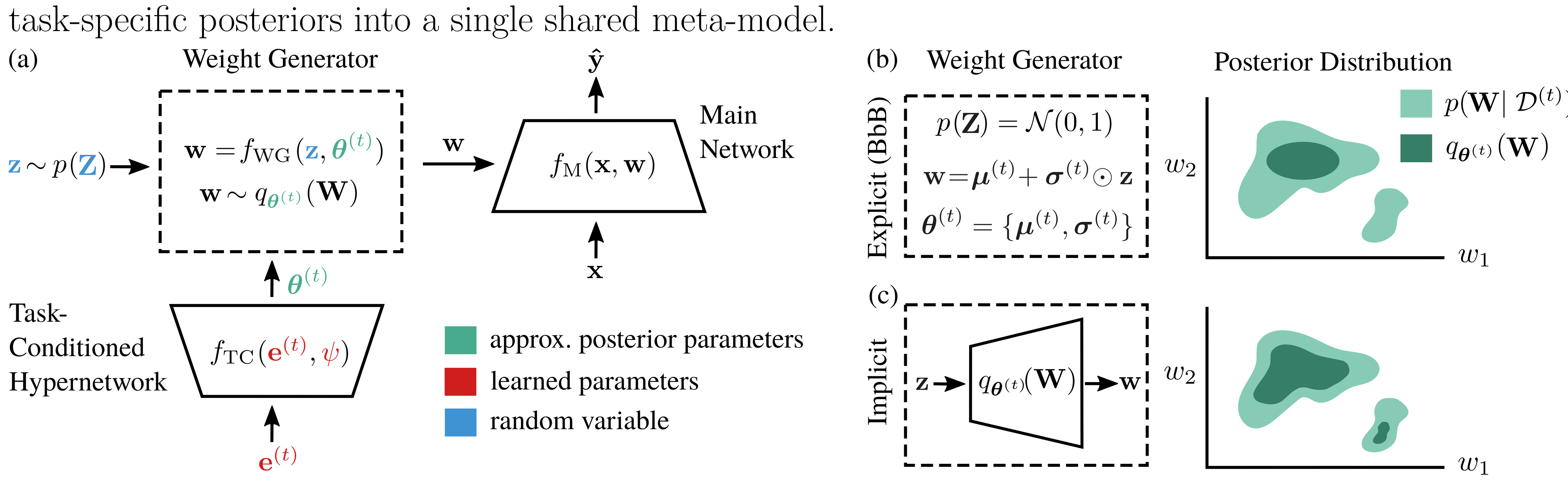

Figure 2: The (a) *posterior meta-replay* framework for CL with (b) explicit or (c) implicit approximate posterior distributions.

Task-specific posteriors are learned within a shared task-conditioned hypernetwork [4] which generates posterior parameters $\boldsymbol{\theta}^{(t)}$ upon conditioning by the task-embedding $\mathbf{e}^{(t)}$. By design, the number of trainable parameters does not increase (i.e., $\dim(\boldsymbol{\psi}) + \Sigma_t \dim(\mathbf{e}^{(t)}) < \dim(\mathbf{W})$).

The choice of approximate posterior remains flexible and depends on a weight generator (WG) parametrized by $\boldsymbol{\theta}^{(t)}$. The WG applies the reparametrization trick to sample from the approximate, which can be, for instance, a simple mean-field Gaussian or an implicit distribution defined by a neural network.

Forgetting at the meta-level is prevented with the use of a meta-regularizer that ensures that previously learned posteriors $q_{\boldsymbol{\theta}^{(t',*)}}(\mathbf{W})$ are not changed. The loss for task t thus becomes:

$$\mathcal{L}^{(t)}(\boldsymbol{\psi}, \mathcal{E}, \mathcal{D}^{(t)}) = \mathcal{L}_{\text{task}}(\boldsymbol{\psi}, \mathbf{e}^{(t)}, \mathcal{D}^{(t)}) + \beta \Sigma_{t'<t} \mathrm{D}(q_{\boldsymbol{\theta}^{(t',*)}}(\mathbf{W}) || q_{\boldsymbol{\theta}^{(t')}}(\mathbf{W})) \quad (2)$$

The task with lowest predictive uncertainty is selected when processing unseen inputs.

Experiments

Simple 1D regression illustrates the pitfalls of prior-focused learning. While task-specific posteriors are easily learned with our approach (Fig. 3a), *prior-focused* approaches struggle to find a single trade-off solution that successfully fits all three tasks (Fig. 3b).

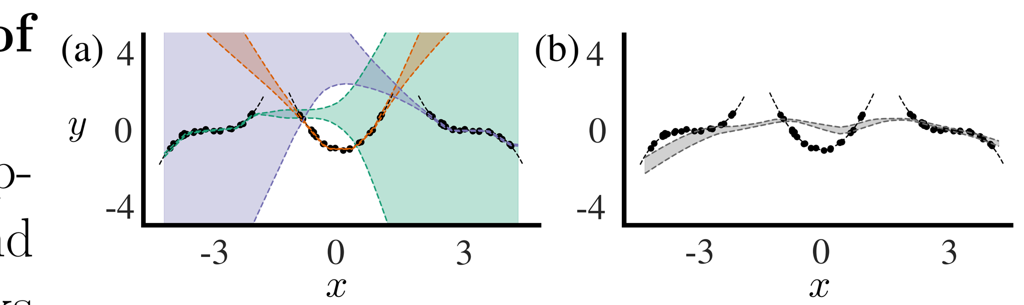

Figure 3: 1D regression problem with (a) *posterior meta-replay* and (b) *prior-focused* methods.

Maintaining parameter uncertainty is crucial for robust task inference. A 2D classification problem highlights that deterministic solutions display arbitrary uncertainty away from the training data of the corresponding task (Fig. 4b), while introducing parameter uncertainty can lead to high uncertainty out-of-distribution (Fig. 4c), and enable more robust task inference.

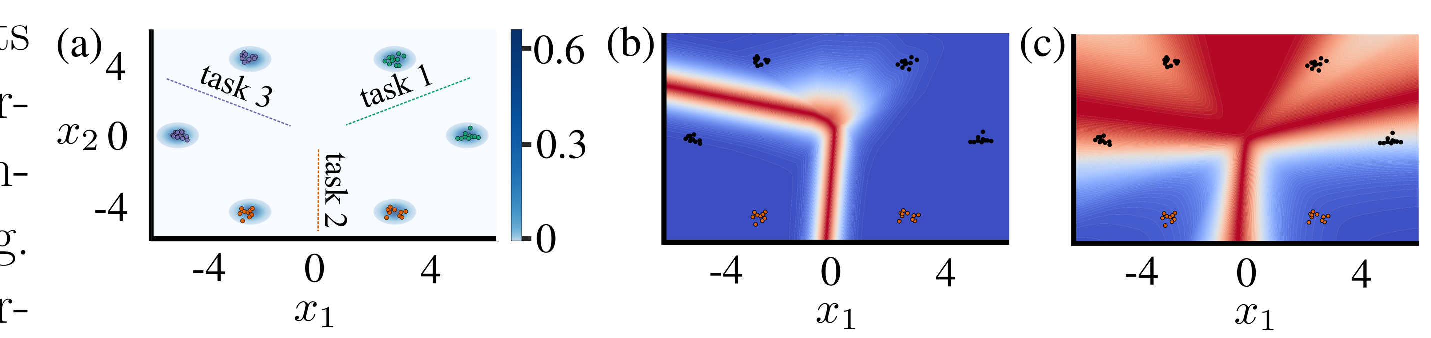

Figure 4: 2D binary classification problem. Input density map (a), and entropy of the posterior distribution of the second task with *posterior meta-replay* for (b) a Dirac distribution and (b) an implicit posterior.

Posterior meta-replay scales to CIFAR-10 We perform SplitCIFAR-10 experiments with a Resnet-32. We observe improvements through the incorporation of epistemic uncertainty (i.e., PR-Dirac vs. PR-Explicit). Compared to *prior-focused* methods, our approach exhibits very little forgetting and improved final accuracy. Also compared to competing approaches like experience-replay, our approach shows performance gains in task-agnostic settings. Performance can be further improved through several extensions (BW and CS).

Table 1: Accuracies of SplitCIFAR-10 experiments (Mean ± SEM in %, $n = 10$), during (*TGiven-During*) and at the end of training when the task is given (*TGiven-Final*) or inferred (*TInfer-Final*). *PR* denotes *posterior meta-replay*.

	TGiven-During	TGiven-Final	TInfer-Final
EWC-growing	N/A	N/A	20.40 ± 0.95
PR-Dirac	94.59 ± 0.10	93.77 ± 0.31	54.83 ± 0.79
PR-Explicit	95.59 ± 0.08	95.43 ± 0.11	61.90 ± 0.66
PR-Implicit	94.25 ± 0.07	92.83 ± 0.16	51.95 ± 0.53
PR-Explicit-BW	95.59 ± 0.08	95.43 ± 0.11	92.94 ± 1.04
PR-Explicit-CS	95.15 ± 0.11	92.48 ± 0.13	64.76 ± 0.34
Exp-Replay	N/A	N/A	41.38 ± 2.80

Conclusion

Bayesian statistics provide a theoretical basis for continual learning algorithms. However, practical challenges arise through the necessary use of approximate inference. When learning a sequence of tasks, this can be solved by having task-specific posteriors that are learned within a single shared meta-model. This approach has much more flexibility, and performance can further benefit from improved task-inference.

References

[1] Farquhar et al. A Unifying Bayesian View of Continual Learning. *Bayesian Deep Learning Workshop at NeurIPS*, 2018.

[2] Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017.

[3] Loo et al. Generalized Variational Continual Learning. In *International Conference on Learning Representations*, 2021.

[4] von Oswald et al. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2020.

Universität Zürich[UZH]

ETH zürich

NEURAL INFORMATION PROCESSING SYSTEMS