Function-Space Regularization in Neural Networks: A Probabilistic Perspective

Tim G. J. Rudner*† Sanyam Kapoor† Shikai Qiu† Andrew Gordon Wilson†

† New York University * Corresponding author: tim.rudner@nyu.edu @timrudner

PAPER

Main Takeaways

- We propose **Function-Space Empirical Bayes** (FSEB) for training deterministic NNs and Bayesian NNs.
- FSEB leads to **significantly improved predictive uncertainty quantification** across a wide range of problems.
- FSEB yields a transparent and probabilistically principled function-space regularizer that is **easy to implement on top of existing methods**.

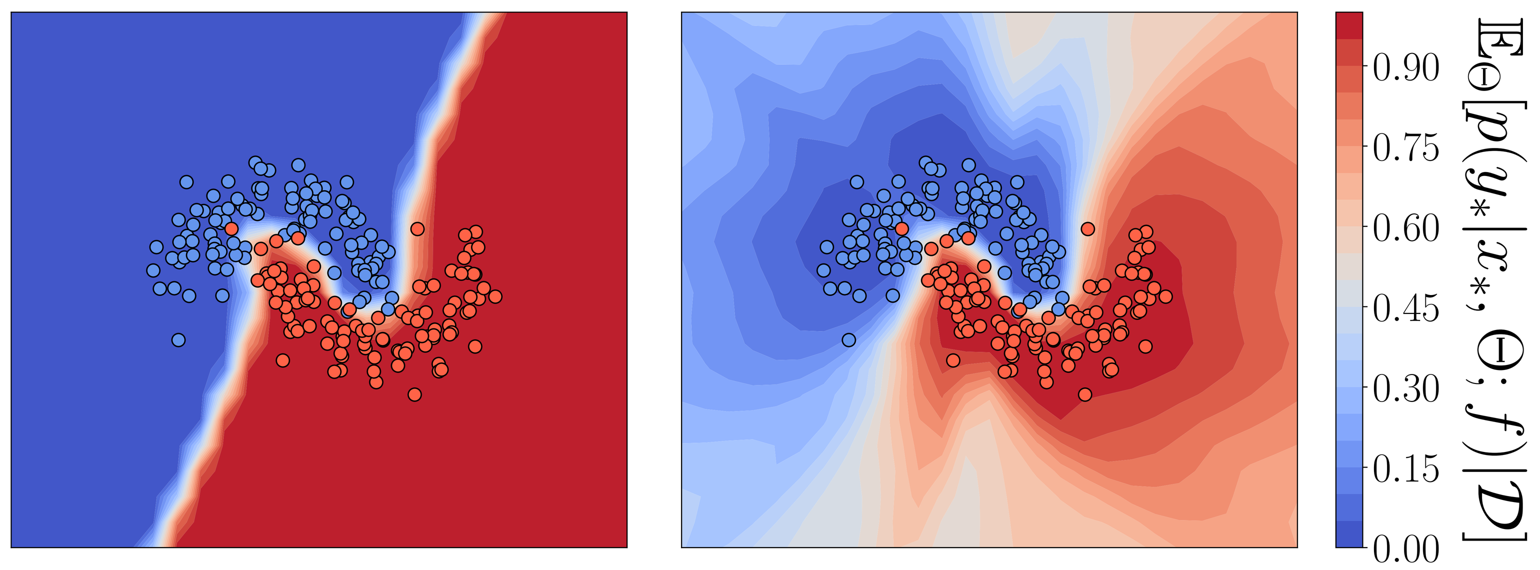

Figure 1: Predictive distributions obtained by training on the *Two Moons* datasets using standard parameter-space maximum a posteriori estimation (**Left**) and function-space empirical Bayes (FSEB) (**Right**) in a two-layer MLP.

Background

- Consider data $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N} = (X_\mathcal{D}, y_\mathcal{D})$ with inputs $x_n \in \mathcal{X} \subseteq \mathbb{R}^D$ and targets $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^Q$ for regression and $\mathcal{Y} \subseteq \{0,1\}^Q$ for classification tasks
- Consider a parametric observation model $p_{Y|X,\Theta}(y \mid x, \theta; f)$ with mapping $f(\cdot\,; \theta) \doteq h(\cdot\,; \theta_h)\theta_L$ and a *prior* distribution over the parameters, $p_\Theta(\theta)$
- Probabilistic model:

$$p_{\Theta|Y,X}(\theta \mid y_\mathcal{D}, x_\mathcal{D}) \propto p_{Y|X,\Theta}(y_\mathcal{D} \mid x_\mathcal{D}, \theta) p_\Theta(\theta) \qquad (1)$$

- Likelihood factorization:

$$p(y_\mathcal{D} \mid x_\mathcal{D}, \theta) \doteq \prod_{n=1}^{N} p(y_\mathcal{D}^{(n)} \mid x_\mathcal{D}^{(n)}, \theta)$$

- MAP objective:

$$\mathcal{L}^{\text{MAP}}(\theta) = \sum_{n=1}^{N} \log p_{Y|X,\Theta}(y_\mathcal{D}^{(n)} \mid x_\mathcal{D}^{(n)}, \theta) + \log p_\Theta(\theta)$$

- $p(\theta) = \mathcal{N}(\theta; \mathbf{0}, \sigma_0^2 I) \Rightarrow$ standard L_2-norm regularization

Function-Space Empirical Bayes

Empirical Priors via Function-Space Regularization

- Auxiliary model: $\hat{p}(\theta \mid \hat{y}, \hat{x}) \propto \hat{p}(\hat{y} \mid \hat{x}, \theta; f) p(\theta)$
- Auxiliary likelihood:

$$\hat{p}(\hat{y}_k \mid \hat{x}, \theta; f) \doteq \mathcal{N}(\hat{y}_k; f(\hat{x}; \theta)_k, \tau_f^{-1} K(\hat{x}, \hat{x}; \phi_0)), \qquad (2)$$

with $K(\hat{x}, \hat{x}; \phi_0) \doteq h(\hat{x}; \phi_0) h(\hat{x}; \phi_0)^\top + I$

- For $p(\theta) = \mathcal{N}(\theta; \mathbf{0}, \tau_\theta^{-1})$, then

$$\log \hat{p}(\hat{y} \mid \hat{x}, \theta; f) + \log p(\theta) \qquad (3)$$

$$\propto -\sum_{k=1}^{K} \frac{\tau_f}{2} f(\hat{x}; \theta)_k^\top K(\hat{x}, \hat{x}; \phi_0)^{-1} f(\hat{x}; \theta)_k - \frac{\tau_\theta}{2} \|\theta\|_2^2,$$

- Function-space empirical Bayes regularizer:

$$\mathcal{J}(\theta, \hat{x}) \doteq -\sum_{k=1}^{K} \frac{\tau_f}{2} d_M^2(f(\hat{x}; \theta)_k, K(\hat{x}, \hat{x}; \phi_0)) - \frac{\tau_\theta}{2} \|\theta\|_2^2$$

Empirical Bayes Maximum A Posteriori Estimation

- Function-space empirical Bayes model:

$$p(\theta \mid y_\mathcal{D}, x_\mathcal{D}) \propto p(y_\mathcal{D} \mid x_\mathcal{D}, \theta) \hat{p}(\theta \mid \hat{y}, \hat{x}) \qquad (4)$$

- Function-space empirical Bayes MAP objective:

$$\mathcal{L}^{\text{EB-MAP}}(\theta) \doteq \sum_{n=1}^{N} \log p(y_\mathcal{D}^{(n)} \mid x_\mathcal{D}^{(n)}, \theta) + \mathcal{J}(\theta, \hat{x}) \qquad (5)$$

Empirical Bayes Variational Inference

- Extended probabilistic model

$$p(\theta', \hat{x} \mid y_\mathcal{D}, x_\mathcal{D}) \propto p(y_\mathcal{D} \mid x_\mathcal{D}, \theta') \hat{p}(\theta' \mid \hat{y}, \hat{x}) p(\hat{x}), \qquad (6)$$

- Empirical prior: $\hat{p}(\theta' \mid \hat{y}, \hat{x}) \propto \hat{p}(\hat{y} \mid \hat{x}, \theta'; f) p(\theta')$
- Variational distribution: $q(\theta', \hat{x}) \doteq q(\theta') p(\hat{x})$
- Variational objective: $\min_{q_{\Theta'} \in \mathcal{Q}} \mathbb{E}_{p_{\hat{X}}} \left[D_{\text{KL}}(q_{\Theta'} \,\|\, p_{\Theta' \mid Y_\mathcal{D}, X_\mathcal{D}}) \right]$
- Function-space empirical Bayes regularization estimator:

$$\mathcal{F}(\theta) \doteq -\frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mathcal{J}(\theta + \sigma\epsilon^{(j)}, \hat{X}^{(i)}) + C \qquad (7)$$

with $\hat{X}^{(i)} \sim p_{\hat{X}}$ and $\epsilon^{(j)} \sim \mathcal{N}(\mathbf{0}, I)$

- Function-space empirical Bayes variational objective:

$$\mathcal{L}^{\text{EB-VI}}(\theta) = \frac{1}{S} \sum_{n=1}^{N} \sum_{s=1}^{S} \log p(y_\mathcal{D}^{(n)} \mid x_\mathcal{D}^{(n)}, \theta + \sigma\epsilon^{(s)}) - \mathcal{F}(\theta)$$

Empirical Evaluation

Accuracy, Calibration, & Selective Prediction

Table 1: **FashionMNIST**.

Method	Acc. ↑	Sel. Pred. ↑	NLL ↓	ECE ↓
PS-MAP	93.8%±0.0	**98.9%**±0.0	0.26±0.00	3.6%±0.0
FS-EB	**94.1%**±0.1	98.8%±0.0	**0.19**±0.00	**1.8%**±0.1
FS-VI	**94.1%**±0.0	98.4%±0.0	0.24±0.00	2.6%±0.1

Table 2: **CIFAR-10**.

Method	Acc. ↑	Sel. Pred. ↑	NLL ↓	ECE ↓
PS-MAP	93.8%±0.0	**98.9%**±0.0	0.26±0.00	3.6%±0.0
FS-EB	**94.1%**±0.1	98.8%±0.0	**0.19**±0.00	**1.8%**±0.1
FS-VI	**94.1%**±0.0	98.4%±0.0	0.24±0.00	2.6%±0.1

→ FSEB leads to improved uncertainty quantification and leads to better NLL, ECE, and selective prediction accuracy.

Generalization under Covariate Shift

Figure 2: (a) Accuracy

Figure 3: (b) Selective Prediction AUC

→ FSEB leads to improved generalization and selective prediction under most corrupted CIFAR-10 covariate shifts.

Transfer Learning & Semantic Shift Detection

Table 3: Transfer from a ResNet-18 pretrained on ImageNet to CIFAR-10.

Method	Acc. ↑	Sel. Pred. ↑	NLL ↓	ECE ↓	OOD ↑
PS-MAP	96.2%±0.1	99.6%±0.0	0.13±0.01	3.2%±0.2	96.3%±0.7
FS-EB	96.2%±0.1	99.6%±0.0	**0.11**±0.00	**1.3%**±0.1	**98.9%**±0.1

Table 4: Semantic shift detection

Dataset	Method	OOD AUROC ↑
FMNIST	PS-MAP	94.9%±0.4
	FS-EB (x_C = KMNIST)	**99.9%**±0.0
	FS-VI	98.0%±0.4
CIFAR-10	PS-MAP	93.0%±0.4
	FS-EB (x_C = CIFAR100)	**99.4%**±0.1
	FS-VI	99.0%±0.1

→ FSEB leads to improved NLL, ECE, and selective prediction for pretrained models.

→ FSEB significantly improves semantic shift detection for models trained from scratch and pretrained models.

Our code is available on GitHub (link in paper)!

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