Function-Space Regularization in Neural Networks: A Probabilistic Perspective

Tim G. J. Rudner
@timrudner

Sanyam Kapoor
@psiyumm

Shikai Qiu
@shikaiqiu

Andrew Gordon Wilson
@andrewgwils

International Conference on Machine Learning 2023

NYU

Correspondence to
tim.rudner@nyu.edu

Paper: timrudner.com/fseb
Code: timrudner.com/fseb-code

Setting & Problem Statement

Uncertainty quantification

- A key ingredient to making neural networks **reliably safe**
- **Goal:** Obtain **reliable predictive uncertainty** for neural networks

Setting & Problem Statement

Uncertainty quantification

- A key ingredient to making neural networks **reliably safe**
- **Goal:** Obtain **reliable predictive uncertainty** for neural networks

Bayesian deep learning

- Infer **posterior distribution** over neural network parameters
- **Problem:** State-of-the-art methods underperform deterministic models

Key Ideas

Function-Space Empirical Bayes

- Goal: **Match or outperform predictive accuracy** of standard neural network training while **improving predictive uncertainty estimation**.

Key Ideas

Function-Space Empirical Bayes

- Goal: **Match or outperform predictive accuracy** of standard neural network training while **improving predictive uncertainty estimation**.
- Define **empirical prior** that reflects beliefs about desired functions **and** parameters

Key Ideas

Function-Space Empirical Bayes

- Goal: **Match or outperform predictive accuracy** of standard neural network training while **improving predictive uncertainty estimation**.
- Define **empirical prior** that reflects beliefs about desired functions **and** parameters
- Use empirical prior to derive inference method that yields **function-** and **parameter-space regularization**

Empirical Bayes Auxiliary Model

- Empirical prior: $\hat{p}(\theta \mid \hat{y}, \hat{x}) \propto \hat{p}(\hat{y} \mid \hat{x}, \theta; f)p(\theta)$
- How to specify auxiliary likelihood and how to specify $\hat{x} = \{x_1, \ldots, x_M\}$?

Empirical Priors via Distributions over Functions

Empirical Bayes Auxiliary Model

- Empirical prior: $\hat{p}(\theta \mid \hat{y}, \hat{x}) \propto \hat{p}(\hat{y} \mid \hat{x}, \theta; f) p(\theta)$
- How to specify auxiliary likelihood and how to specify $\hat{x} = \{x_1, \ldots, x_M\}$?

Goal: Match Desired Function Evaluations

- Consider the model

$$Z_k(x) \doteq h(x; \phi_0) \Psi_k + \varepsilon \quad \text{with} \quad \Psi_k \sim \mathcal{N}(\psi; \mu, \tau_f^{-1} I) \quad \text{and} \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \tau_f^{-1} I)$$

- Induced distribution over functions:

$$\mathcal{N}(z_k(\hat{x}); h(\hat{x}; \phi_0) \mu_k, \tau_f^{-1} K(\hat{x}, \hat{x}; \phi_0)) \quad \text{with} \quad K(\hat{x}, \hat{x}; \phi_0) \doteq h(\hat{x}; \phi_0) h(\hat{x}; \phi_0)^\top + I$$

Empirical Bayes Auxiliary Likelihood

- Induced distribution over functions $\mathcal{N}(z_k(\hat{x}); h(\hat{x}; \phi_0)\, \mu_k, \tau_f^{-1} K(\hat{x}, \hat{x}; \phi_0))$
- View as likelihood: $\hat{p}(\hat{y}_k \mid \hat{x}, \theta; f) \doteq \mathcal{N}(\hat{y}_k; f(\hat{x}; \theta)_k, \tau_f^{-1} K(\hat{x}, \hat{x}; \phi_0))$
- With zero-mean: $\hat{y} \doteq \{\mathbf{0}, \ldots, \mathbf{0}\}$
- Factorization across dimensions: $\hat{p}(\hat{y} \mid \hat{x}, \theta; f) \doteq \prod_{k=1}^{K} \hat{p}(\hat{y}_k \mid \hat{x}, \theta; f)$

Empirical Priors via Distributions over Functions

Empirical Bayes Auxiliary Likelihood

- Induced distribution over functions $\mathcal{N}(z_k(\hat{x}); h(\hat{x}; \phi_0)\, \mu_k, \tau_f^{-1} K(\hat{x}, \hat{x}; \phi_0))$
- View as likelihood: $\hat{p}(\hat{y}_k \mid \hat{x}, \theta; f) \doteq \mathcal{N}(\hat{y}_k; f(\hat{x}; \theta)_k, \tau_f^{-1} K(\hat{x}, \hat{x}; \phi_0))$
- With zero-mean: $\hat{y} \doteq \{\mathbf{0}, \ldots, \mathbf{0}\}$
- Factorization across dimensions: $\hat{p}(\hat{y} \mid \hat{x}, \theta; f) \doteq \prod_{k=1}^{K} \hat{p}(\hat{y}_k \mid \hat{x}, \theta; f)$

Empirical Bayes Auxiliary Prior

- Standard prior over parameters, e.g.: $p(\theta) = \mathcal{N}(\theta; \mathbf{0}, \tau_\theta^{-1})$

Unnormalized Empirical Prior Density Function

- Analytically tractable unnormalized log joint density:

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

- Distance measure in function and parameter space

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

where $d_M^2(v, K) \doteq v^\top K^{-1} v$ is the squared Mahalanobis distance from 0

Maximum A Posteriori (MAP) Estimation

- Find parameters that maximize the posterior distribution

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

- That is:

$$\max_{\theta} \log p_{\Theta|Y,X}\left(\theta \mid y_{\mathcal{D}}, x_{\mathcal{D}}\right) \Leftrightarrow \max_{\theta} \log p_{Y|X,\Theta}\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta\right) + \log p_{\Theta}(\theta)$$

- Optimization objective:

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

- Gaussian prior: L2 regularization

Function-Space Empirical Bayes Regularizer

- Empirical Bayes log joint distribution:

$$\log p\left(\theta \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}, \hat{x}\right) \propto \log p\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta\right) + \log \hat{p}(\theta \mid \hat{y}, \hat{x})$$

where

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

Empirical Bayes Maximum A Posteriori Estimation

Function-Space Empirical Bayes Regularizer

- Empirical Bayes log joint distribution:

$$\log p\left(\theta \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}, \hat{x}\right) \propto \log p\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta\right) + \log \hat{p}(\theta \mid \hat{y}, \hat{x})$$

  where

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

Empirical Bayes Maximum A Posteriori

- Optimization objective:

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

Making Auxiliary Inputs Stochastic

- Extended model: $p\,(\theta', \hat{x} \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}) \propto p\,(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta')\,\hat{p}\,(\theta' \mid \hat{y}, \hat{x})\,p(\hat{x})$

  with empirical prior $\hat{p}\,(\theta' \mid \hat{y}, \hat{x}) \propto \hat{p}\,(\hat{y} \mid \hat{x}, \theta'; f)\,p\,(\theta')$

Making Auxiliary Inputs Stochastic

- Extended model: $p\left(\theta', \hat{x} \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}\right) \propto p\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta'\right) \hat{p}\left(\theta' \mid \hat{y}, \hat{x}\right) p(\hat{x})$

  with empirical prior $\hat{p}\left(\theta' \mid \hat{y}, \hat{x}\right) \propto \hat{p}\left(\hat{y} \mid \hat{x}, \theta'; f\right) p\left(\theta'\right)$

Variational Problem

- Variational distribution: $q\left(\theta', \hat{x}\right) \doteq q\left(\theta'\right) q(\hat{x})$
- Inference problem: $\min_{q_{\Theta', \hat{X}} \in \mathcal{Q}} D_{\mathrm{KL}}\left(q_{\Theta', \hat{X}} \| p_{\Theta', \hat{X} \mid Y_{\mathcal{D}}, X_{\mathcal{D}}, \hat{Y}}\right)$

Empirical Bayes Variational Inference

Making Auxiliary Inputs Stochastic

- Extended model: $p\left(\theta', \hat{x} \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}\right) \propto p\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta'\right) \hat{p}\left(\theta' \mid \hat{y}, \hat{x}\right) p(\hat{x})$

  with empirical prior $\hat{p}\left(\theta' \mid \hat{y}, \hat{x}\right) \propto \hat{p}\left(\hat{y} \mid \hat{x}, \theta'; f\right) p\left(\theta'\right)$

Variational Problem (simplified)

- Variational distribution: $q\left(\theta', \hat{x}\right) \doteq q\left(\theta'\right) p(\hat{x})$
- Inference problem: $\min_{q_{\Theta'} \in \mathcal{Q}} \mathbb{E}_{p_{\dot{X}}} \left[D_{\mathrm{KL}} \left(q_{\Theta'} \| p_{\Theta' \mid Y_{\mathcal{D}}, X_{\mathcal{D}}, \hat{Y}, \hat{X}} \right) \right]$

Empirical Bayes Variational Inference

Making Auxiliary Inputs Stochastic

- Extended model: $p\left(\theta', \hat{x} \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}\right) \propto p\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta'\right) \hat{p}\left(\theta' \mid \hat{y}, \hat{x}\right) p(\hat{x})$

  with empirical prior $\hat{p}\left(\theta' \mid \hat{y}, \hat{x}\right) \propto \hat{p}\left(\hat{y} \mid \hat{x}, \theta'; f\right) p\left(\theta'\right)$

Variational Problem (simplified)

- Variational distribution: $q\left(\theta', \hat{x}\right) \doteq q\left(\theta'\right) p(\hat{x})$
- Inference problem:

$$\max_{q_{\Theta'} \in \mathcal{Q}} \mathbb{E}_{q_{\Theta'}}\left[\log p\left(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \Theta'; f\right)\right] - \mathbb{E}_{p_{\hat{X}}}\left[D_{\mathrm{KL}}\left(q_{\Theta'} \| p_{\Theta' \mid \hat{Y}, \hat{X}}\right)\right]$$

Making Auxiliary Inputs Stochastic

- Extended model: $p\,(\theta', \hat{x} \mid y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{y}) \propto p\,(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta')\,\hat{p}\,(\theta' \mid \hat{y}, \hat{x})\,p(\hat{x})$

  with empirical prior $\hat{p}\,(\theta' \mid \hat{y}, \hat{x}) \propto \hat{p}\,(\hat{y} \mid \hat{x}, \theta'; f)\,p\,(\theta')$

Variational Problem (simplified)

- Variational distribution: $q\,(\theta', \hat{x}) \doteq q\,(\theta')\,p(\hat{x})$
- Inference problem:

$$\max_{q_{\Theta'} \in \mathcal{Q}} \mathbb{E}_{q_{\Theta'}}\left[\log p\,(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \Theta'; f)\right] - \mathbb{E}_{p_{\hat{X}}}\left[D_{\mathrm{KL}}\left(q_{\Theta'} \| p_{\Theta' \mid \hat{Y}, \hat{X}}\right)\right]$$

Function-Space Empirical Bayes Regularization Estimator

- KL estimator:

$$\mathbb{E}_{p_{\hat{X}}}\left[D_{\text{KL}}\left(q_{\Theta'} \| p_{\Theta' | \hat{Y}, \hat{X}}\right)\right] \approx \mathcal{F}(\theta) \doteq -\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mathcal{J}\left(\theta + \sigma\epsilon^{(j)}, \hat{X}^{(i)}\right) + C$$

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

Function-Space Regularization via Empirical Priors

Function-Space Empirical Bayes Regularization Estimator

- KL estimator:

$$\mathbb{E}_{p_{\hat{X}}}\left[D_{\mathrm{KL}}\left(q_{\Theta'} \| p_{\Theta' \mid \hat{Y}, \hat{X}}\right)\right] \approx \mathcal{F}(\theta) \doteq -\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mathcal{J}\left(\theta + \sigma\epsilon^{(j)}, \hat{X}^{(i)}\right) + C$$

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

Empirical Bayes Variational Inference

- Variational objective

$$\mathcal{L}^{\mathrm{EB-}VI}(\theta) = \frac{1}{S}\sum_{n=1}^{N}\sum_{s=1}^{S}\log p(y_{\mathcal{D}}^{(n)} \mid x_{\mathcal{D}}^{(n)}, \theta + \sigma\epsilon^{(s)}) - \mathcal{F}(\theta) \quad \text{with} \quad \epsilon^{(s)} \sim \mathcal{N}(\mathbf{0}, I)$$

Empirical Evaluation

Improved Uncertainty-Aware Image Classification

- Setup: Training with Function-Space Empirical Regularizer
- Result 1: Match or outperforms predictive accuracy of standard training
- Result 2: Consistently improved uncertainty quantification

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

Method	Acc. ↑	Sel. Pred. ↑	NLL ↓	ECE ↓
PS-MAP	94.9%$\pm$0.2	99.3%$\pm$0.0	0.21$\pm$0.01	3.0%$\pm$0.1
FS-EB	**95.1%**$\pm$0.1	**99.4%**$\pm$0.0	**0.20**$\pm$0.00	**2.1%**$\pm$0.1
FS-VI	92.9%$\pm$0.1	98.0%$\pm$0.0	0.31$\pm$0.00	4.0%$\pm$0.1

Highly-Accurate Semantic Shift Detection

- Setup: Train on FMNIST/CIFAR-10 & OOD Detection on MNIST/SVHN
- Result 1: Near-perfect semantic shift detection (best-in-class)
- Alternative context distribution: corrupted/augmented training data

DATASET	METHOD	OOD AUROC ↑
FMNIST	PS-MAP	94.9%$\pm$0.4
	FS-EB (x_C = KMNIST)	**99.9%**$\pm$0.0
	FS-VI	98.0%$\pm$0.4

DATASET	METHOD	OOD AUROC ↑
CIFAR-10	PS-MAP	93.0%$\pm$0.4
	FS-EB (x_C = CIFAR100)	**99.4%**$\pm$0.1
	FS-VI	99.0%$\pm$0.1

Improved Transfer Learning with Pretrained Models

- Setup: Fine-tune model pertained on ImageNet 1K on CIFAR-10
- Result: Consistently improved uncertainty quantification

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

Main takeaways

Function-Space Empirical Bayes

- is **probabilistically principled** and **transparent**;
- yields both **parameter-** and **function-space regularization**;
- is **computationally cheap**;
- performs **on par with or better than standard training**;
- leads to **significantly improved predictive uncertainty quantification**.

THANK YOU!

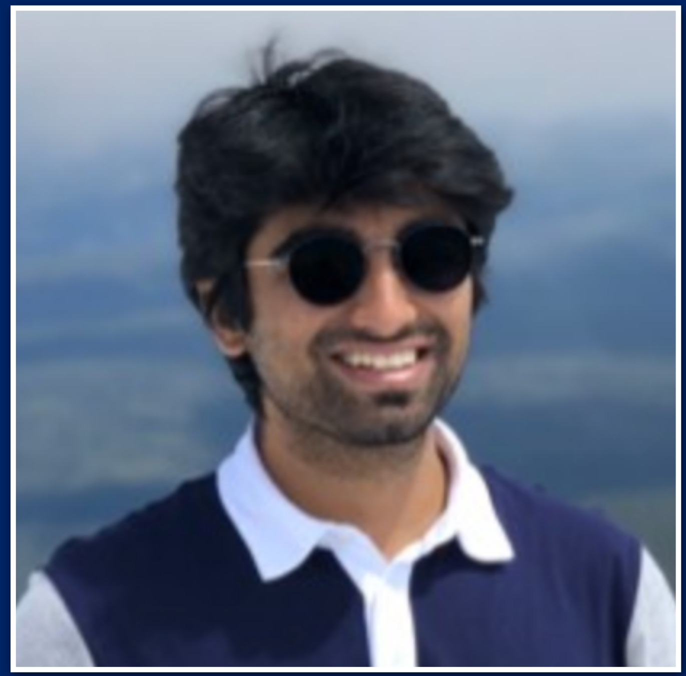
Correspondence to

tim.rudner@nyu.edu

Paper: timrudner.com/fseb

Code: timrudner.com/fseb-code

TIM G. J. RUDNER
@timrudner



SANYAM KAPOOR
@psiyumm

SHIKAI QIU
@shikaiqiu

ANDREW GORDON WILSON
@andrewgwils