Improving Generalization with Flat Hilbert Bayesian Inference

Tuan Truong [*, 1] Quyen Tran [*, 1] Quan Pham [1] Nhat Ho [2]
Dinh Phung [3] Trung Le [3]

[1]Qualcomm AI Research

[2]The University of Texas at Austin

[3]Monash University

ICML 2025

Introduction

Problem: Approximate Bayesian Inference

- Given some observations, how to estimate the underlying posterior distribution.

Introduction

Problem: Approximate Bayesian Inference

- Given some observations, how to estimate the underlying posterior distribution.
- Beneficial in quantifying and tackling uncertainty for deep learning models.

Introduction

Problem: Approximate Bayesian Inference

- Given some observations, how to estimate the underlying posterior distribution.
- Beneficial in quantifying and tackling uncertainty for deep learning models.
- **Contribution**: we propose a Bayesian Inference method with *improved generalization ability*

Background

- Consider a family of neural networks $f_{\boldsymbol{\theta}}(x)$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, a training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ sampled from a distribution $\mathcal{D}$

ICML
International Conference
On Machine Learning

Background

- Consider a family of neural networks $f_{\boldsymbol{\theta}}(x)$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, a training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ sampled from a distribution $\mathcal{D}$
- Prior works typically focus on approximating the *empirical posterior*

$$p(\boldsymbol{\theta}|\mathcal{S}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i|x_i, \mathcal{S}, \boldsymbol{\theta}).$$

$$p(\boldsymbol{\theta}|\mathcal{S}) = \exp\left(-\frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(x_i), y_i) \right) p(\boldsymbol{\theta})$$

Background

- Consider a family of neural networks $f_{\boldsymbol{\theta}}(x)$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, a training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ sampled from a distribution $\mathcal{D}$
- Prior works typically focus on approximating the *empirical posterior*

$$p(\boldsymbol{\theta}|\mathcal{S}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i|x_i, \mathcal{S}, \boldsymbol{\theta}).$$

$$p(\boldsymbol{\theta}|\mathcal{S}) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(x_i), y_i)\right) p(\boldsymbol{\theta})$$

- However, we may want to approximate $p(\boldsymbol{\theta}|\mathcal{D})$ instead to avoid overfitting

Theoretical Analysis

To avoid overfitting, it is preferable to sample the particle models $\theta_{1:m}$ from the *population posterior* $p(\boldsymbol{\theta}|\mathcal{D})$

Proposition 1: Consider the problem

$$\min_{\mathbb{Q} \ll \mathbb{P}_\theta} \left\{ \mathbb{E}_{\theta \sim \mathbb{Q}}[\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})] + D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}_\theta) \right\},$$

where we search over $\mathbb{Q}$ absolutely continuous w.r.t $\mathbb{P}_\theta$, and $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(x), y)]$ is the population loss. The closed-form solution to this problem is exactly the **population posterior** $p(\theta|\mathcal{D})$

Theoretical Analysis

- Objective: Approximate $p(\boldsymbol{\theta}|\mathcal{D})$ with a simpler distribution q^*

$$q^* = \arg\min_{q \in \mathcal{F}} D_{\mathrm{KL}}\Big(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}|\mathcal{D})\Big).$$

- We define $\mathcal{F}$ as the set of distributions for random variables of the form $\vartheta = \boldsymbol{T}(\boldsymbol{\theta})$, where $\boldsymbol{T} : \Theta \to \Theta$ is a smooth, bijective mapping.
- We restrict the set of $\boldsymbol{T}$ to the maps of the form $\boldsymbol{T}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{f}(\boldsymbol{\theta})$, where $\boldsymbol{f} \in \mathcal{H}^d$ is a **vector-valued RKHS**

Theoretical Analysis

The optimization problem becomes:

$$\boldsymbol{f}^* = \underset{\boldsymbol{f} \in \mathcal{H}^d, \|\boldsymbol{f}\|_{\mathcal{H}^d} \leq \epsilon}{\arg\min} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}|\mathcal{D}) \right).$$

where we have

$$q_{[\boldsymbol{T}]}(\vartheta) = q(\boldsymbol{T}^{-1}(\vartheta)) |\det(\nabla_\vartheta \boldsymbol{T}^{-1}(\vartheta))|.$$

Theoretical Analysis

Theorem (Informal)

Let q be any distribution and d_{VC} denotes the VC dimension of he hypothesis space $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. For any $\rho > 0$, with probability of $1 - \delta$ over the training set $\mathcal{S}$ generated by distribution $\mathcal{D}$, we have:

$$D_{\mathrm{KL}}\Big(q_{[\boldsymbol{I}+\boldsymbol{f}]} || p(\boldsymbol{\theta}|\mathcal{D})\Big) \leq \max_{\boldsymbol{f}' \in \mathcal{H}^d, \|\boldsymbol{f}' - \boldsymbol{f}\| \leq \rho} D_{\mathrm{KL}}\Big(q_{[\boldsymbol{I}+\boldsymbol{f}']} || p(\boldsymbol{\theta}|\mathcal{S})\Big)$$
$$+ \mathcal{O}\left(\sqrt{\frac{\log(1 + \frac{1}{\rho^2}) + \log\left(\frac{n}{\delta}\right)}{n-1}} + \frac{\sqrt{d_{VC} \log \frac{2en}{d_{VC}}}}{\delta\sqrt{2n}}\right).$$

Theoretical Analysis

Goal: find a sequence of transportation functions $\{\boldsymbol{f}_k\}_k$ that converges to the optimal $\boldsymbol{f}^*$, we can obtain the flow of distributions $q^{(k)} = q_{[\boldsymbol{I}+\boldsymbol{f}]}$.

$$\underset{\|\boldsymbol{f}'-\boldsymbol{f}\|_{\mathcal{H}^d}\leq\rho}{\arg\max} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}']}||p(\boldsymbol{\theta}|\mathcal{S})\right) \approx \underset{\|\hat{\boldsymbol{f}}\|_{\mathcal{H}^d}\leq 1}{\arg\max} \left\langle \hat{\boldsymbol{f}}, \nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]}\|p(\boldsymbol{\theta}|\mathcal{S})\right)\right\rangle_{\mathcal{H}^d}.$$

$$\hat{\boldsymbol{f}}^* = \frac{\nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]}\|p(\cdot|\mathcal{S})\right)}{\left\|\nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]}\|p(\cdot|\mathcal{S})\right)\right\|_{\mathcal{H}^d}}.$$

Theoretical Analysis

Functional sharpness-aware procedure

$$\hat{\boldsymbol{f}}_k^* = \rho \frac{\nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S})\right)\Big|_{\boldsymbol{f}=\boldsymbol{f}_k}}{\left\|\nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S})\right)\Big|_{\boldsymbol{f}=\boldsymbol{f}_k}\right\|_{\mathcal{H}^d}}$$

$$\boldsymbol{f}_{k+1} = \boldsymbol{f}_k - \epsilon \nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S})\right)\Big|_{\boldsymbol{f}=\boldsymbol{f}_k+\hat{\boldsymbol{f}}_k^*}$$

$$q^{(k+1)} = q_{[\boldsymbol{I}+\boldsymbol{f}_{k+1}]}.$$

Theoretical Analysis

Functional sharpness-aware procedure

$$\hat{\boldsymbol{f}}_k^* = \rho \frac{\nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S})\right)\Big|_{\boldsymbol{f}=\boldsymbol{f}_k}}{\left\|\nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S})\right)\Big|_{\boldsymbol{f}=\boldsymbol{f}_k}\right\|_{\mathcal{H}^d}}$$

$$\boldsymbol{f}_{k+1} = \boldsymbol{f}_k - \epsilon \nabla_{\boldsymbol{f}} D_{\mathrm{KL}}\left(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S})\right)\Big|_{\boldsymbol{f}=\boldsymbol{f}_k+\hat{\boldsymbol{f}}_k^*}$$

$$q^{(k+1)} = q_{[\boldsymbol{I}+\boldsymbol{f}_{k+1}]}.$$

Lemma

Let $\boldsymbol{F}[\boldsymbol{f}] = D_{\mathrm{KL}}(q_{[\boldsymbol{I}+\boldsymbol{f}]} \| p(\cdot|\mathcal{S}))$. *When* $\|\boldsymbol{f}\|$ *is sufficiently small,*

$$\nabla_{\boldsymbol{f}} \boldsymbol{F}[\boldsymbol{f}] \approx -\mathbb{E}_q[\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} + \boldsymbol{f}(\boldsymbol{\theta})|\mathcal{S}) k(\boldsymbol{\theta}, \cdot) + \nabla_{\boldsymbol{\theta}} k(\boldsymbol{\theta}, \cdot)].$$

Practical Algorithm

Input: Initial particles $\{\boldsymbol{\theta}_i^{(0)}\}_{i=1}^m$, number of epochs N, step size $\rho > 0$
Output: A set of particles $\{\boldsymbol{\theta}_i\}_{i=1}^m$ that approximates the population posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$
for iteration k **do**
$\hat{\varepsilon}_i^{(k)} \leftarrow \rho \frac{\phi(\boldsymbol{\theta}_i^{(k)})}{\|\phi(\boldsymbol{\theta}_i^{(k)})\|}$ where
$\phi(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{j=1}^m [k(\boldsymbol{\theta}, \boldsymbol{\theta}_j^{(k)})\nabla_{\boldsymbol{\theta}_j^{(k)}} \log p(\boldsymbol{\theta}_j^{(k)}|\mathcal{S}) + \nabla_{\boldsymbol{\theta}_j^{(k)}} k(\boldsymbol{\theta}, \boldsymbol{\theta}_j^{(k)})]$
$\boldsymbol{\theta}_i^{(k+1)} \leftarrow \boldsymbol{\theta}_i^{(k)} - \epsilon_i \psi(\boldsymbol{\theta}_i^{(k)}, \hat{\varepsilon}_i^{(k)})$
where
$\psi(\boldsymbol{\theta}, \varepsilon) = -\frac{1}{n}\sum_{j=1}^m [k(\boldsymbol{\theta}, \boldsymbol{\theta}_j^{(k)})\nabla_{\boldsymbol{\theta}_j^{(k)}} \log p(\boldsymbol{\theta}_j^{(k)} + \varepsilon|\mathcal{S}) + \nabla_{\boldsymbol{\theta}_j^{(k)}} k(\boldsymbol{\theta}, \boldsymbol{\theta}_j^{(k)})]$.
end for

Experimental settings:

- Problem: Fine-tune the Vision Transformer architecture ViT-B/16
- Dataset: VTAB-1K, consisting of 19 datasets on three domains: Natural, Specialized, Structured
- Baselines: full fine-tune, AdamW, SAM, BayesTune, SADA-JEM, SGLD, Sharpness-Aware Bayesian Neural Network, SVGD, Bayesian Deep Ensemble

Experimental Results

Table 1. `VTAB-1K` results evaluated on Top-1 accuracy. All methods are applied to finetune the same set of LoRA parameters on ViT-B/16 pre-trained with `ImageNet-21K` dataset.

	Natural							Specialized				Structured								
Method	CIFAR100	Caltech101	DTD	Flower102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI	dSpr-Loc	dSpr-Ori	sNORB-Azi	sNORB-Ele	**AVG**
FFT	68.9	87.7	64.3	97.2	86.9	**87.4**	38.8	79.7	95.7	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	65.6
AdamW	67.1	90.7	68.9	98.1	90.1	84.5	54.2	84.1	94.9	84.4	73.6	82.9	69.2	49.8	78.5	**75.7**	47.1	31.0	**44.0**	72.0
SAM	72.7	90.3	71.4	99.0	90.2	84.4	52.4	82.0	92.6	84.1	74.0	76.7	68.3	47.9	74.3	71.6	43.4	26.9	39.1	70.5
DeepEns	69.1	88.9	67.7	98.9	90.7	85.1	54.5	82.6	94.8	82.7	75.3	46.6	47.1	47.4	68.2	71.1	36.6	30.1	35.6	67.0
BayesTune	67.2	91.7	69.5	99.0	90.7	86.4	54.7	84.9	**95.3**	84.1	75.1	**82.8**	68.9	49.7	79.3	74.3	46.6	30.3	42.8	72.2
SGLD	68.7	91.0	67.0	98.6	89.3	83.0	51.6	81.2	93.7	83.2	76.4	80.0	70.1	48.2	76.2	71.1	39.3	31.2	38.4	70.4
SADA-JEM	70.3	91.9	70.2	98.2	91.2	85.6	54.7	84.3	94.1	83.4	77.0	79.9	72.1	51.6	79.4	70.7	45.3	29.6	40.1	72.1
SA-BNN	65.1	91.5	71.0	98.9	89.4	89.3	55.2	83.2	94.5	86.4	75.2	61.4	63.2	40.0	71.3	64.5	34.5	27.2	31.2	68.1
SVGD	71.3	90.2	71.0	98.7	90.2	84.3	52.7	83.4	93.2	86.7	75.1	75.8	70.7	49.6	79.9	69.1	41.2	30.6	33.1	70.9
FHBI	**74.1**	**93.0**	**74.3**	**99.1**	**92.4**	87.3	**56.5**	**85.3**	95.0	**87.2**	**79.6**	80.1	**72.3**	**52.2**	**80.4**	72.8	**51.2**	**31.9**	41.3	**73.7**
	(.17)	(.42)	(.15)	(0.20)	(0.21)	(.52)	(.12)	(.31)	(.57)	(.21)	(.20)	(.16)	(.27)	(.47)	(.31)	(.50)	(.32)	(.36)	(.59)	

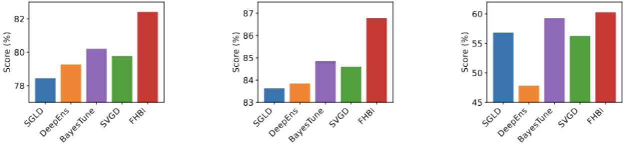


Figure 2. Domain-wise average scores on Natural (left), Specialized (middle), and Structured (right) datasets. FHBI performs best in all three domains compared to the Bayesian inference baselines.

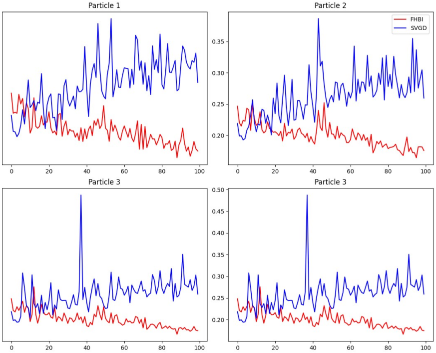
ICML
International Conference On Machine Learning

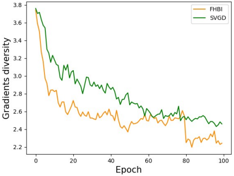
Experimental Results

Table 2. VTAB-1K results evaluated on the Expected Calibration Error (ECE) metric. All methods are applied to finetune the same set of LoRA parameters on ViT-B/16 pre-trained with ImageNet-21K dataset.

	Natural							Specialized				Structured								
Method	CIFAR100	Caltech101	DTD	Flower102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI	dSpr-Loc	dSpr-Ori	sNORB-Azi	sNORB-Ele	**AVG**
FFT	0.29	0.23	0.20	0.13	0.27	0.19	0.45	0.21	0.13	0.18	0.17	0.41	0.44	0.42	0.22	**0.14**	0.23	0.24	0.40	0.26
AdamW	0.38	0.19	0.18	0.05	0.09	0.10	0.14	0.11	0.09	0.12	0.11	**0.12**	0.19	0.34	0.18	0.14	0.21	0.18	0.31	0.17
SAM	0.21	0.25	0.20	0.11	0.12	0.15	**0.14**	0.17	0.16	0.14	0.09	0.12	0.17	0.24	0.16	0.21	0.19	0.13	0.16	0.16
DeepEns	0.24	0.12	0.22	0.04	0.10	0.13	0.23	0.16	0.07	0.15	0.21	0.31	0.32	0.36	**0.13**	0.32	0.31	0.16	0.29	0.20
BayesTune	0.32	**0.08**	0.20	**0.03**	0.85	0.12	0.22	0.13	0.07	0.13	0.22	0.12	0.23	0.30	0.24	0.28	0.28	0.31	0.26	0.23
SGLD	0.26	0.20	0.17	0.05	0.18	0.14	0.23	0.18	0.09	**0.12**	0.32	0.26	0.29	0.21	0.26	0.42	0.39	**0.11**	0.24	0.22
SADA-JEM	0.22	0.11	0.20	0.05	0.13	0.16	0.18	0.15	0.21	0.23	0.26	0.19	0.20	0.25	0.27	0.35	0.20	0.14	0.13	0.19
SA-BNN	0.22	0.08	0.19	0.15	0.12	0.12	0.24	0.13	0.06	0.12	0.18	0.14	0.21	0.22	0.24	0.25	0.41	0.46	0.34	0.20
SVGD	0.20	0.13	0.19	0.04	0.16	0.09	0.20	0.15	0.11	0.13	0.12	0.17	0.21	0.30	0.18	0.21	0.25	0.14	0.26	0.18
FHBI	**0.19**	0.10	**0.16**	0.06	**0.06**	**0.09**	0.16	**0.09**	**0.05**	0.12	**0.08**	0.14	**0.15**	**0.21**	0.15	0.16	**0.18**	0.11	**0.07**	**0.12**

FHBI **reduces the sharpness of every particle** and **promotes ensemble diversity**.





ICML
International Conference
On Machine Learning

Conclusion

- We presented a framework that strengthens prior generalization bounds from Euclidean spaces to the reproducing kernel Hilbert spaces (RKHS).
- We translated this framework to the context of Bayesian inference.
- We presented Flat Hilbert Bayesian Inference (FHBI), which improves generalization ability upon prior works.

Thank you for your attention.