Supplementary Material

Robust Bayesian Neural Networks by Spectral Expectation Bound Regularization

Jiaru Zhang[1] Yang Hua[2] Zhengui Xue[1] Tao Song[1] Chengyu Zheng[1] Ruhui Ma[1] Haibing Guan[1]
[1]Shanghai Jiao Tong University [2]Queen's University Belfast
{jiaruzhang, zhenguixue, songt333, zhengcy, ruhuima, hbguan}@sjtu.edu.cn, Y.Hua@qub.ac.uk

1. Theorem Proofs

Theorem 1. *Consider function $f_{\mathbf{W}}(\mathbf{x}) = f(W\mathbf{x}+\mathbf{b})$, where the activation function $f(\cdot)$ is Lipschitz continuous with Lipschitz constant $Lip(f)$. For any perturbation $\boldsymbol{\xi}$ with norm $\|\boldsymbol{\xi}\|$, we have*

$$\mathbb{E}_{\mathbf{W}} \|f_{\mathbf{W}}(\mathbf{x}+\boldsymbol{\xi}) - f_{\mathbf{W}}(\mathbf{x})\| \leq Lip(f) \cdot \mathbb{E}\|W\|_2 \cdot \|\boldsymbol{\xi}\|, \tag{1}$$

where $\|W\|_2$ represents the spectral norm of matrix W, and it is defined as

$$\|W\|_2 = \max_{\boldsymbol{\xi}\in\mathbb{R}^n, \boldsymbol{\xi}\neq\mathbf{0}} \frac{\|W\boldsymbol{\xi}\|}{\|\boldsymbol{\xi}\|}. \tag{2}$$

Proof. As $f(\cdot)$ is Lipschitz continuous with Lipschitz constant $Lip(f)$, for $i = 1, 2, \ldots, m$, we have

$$\begin{aligned} & \mathbb{E}_{W,b}(f(W_{i,:}\mathbf{x} + W_{i,:}\boldsymbol{\xi} + b_i) - f(W_{i,:}\mathbf{x} + b_i))^2 \\ \leq\ & \mathbb{E}_W(Lip(f) \cdot W_{i,:}\boldsymbol{\xi})^2 \end{aligned} \tag{3}$$

Take a sum for Equation (3) in each i, we have

$$\begin{aligned} & \mathbb{E}_{W,b}\|f(W(\mathbf{x}+\boldsymbol{\xi})+\mathbf{b}) - f(W\mathbf{x}+\mathbf{b})\| \\ \leq\ & \mathbb{E}_W\|Lip(f)\cdot W\boldsymbol{\xi}\| \\ =\ & Lip(f)\cdot\mathbb{E}\|W\boldsymbol{\xi}\| \end{aligned} \tag{4}$$

Based on the definition of spectral norm, we have

$$\|W\boldsymbol{\xi}\| \leq \|W\|_2 \cdot \|\boldsymbol{\xi}\|. \tag{5}$$

By combining Equations (4) and (5), we can obtain the result of (1).

□

Theorem 2. *Consider a Gaussian random matrix $W \in \mathbb{R}^{m\times n}$, where $W_{ij} \sim N(M_{ij}, A_{ij}^2)$ with $M, A \in \mathbb{R}^{m\times n}$. Suppose $G \in \mathbb{R}^{m\times n}$ is a zero-mean Gaussian random matrix with the same variance,* i.e., *$G_{ij} \sim N(0, A_{ij}^2)$. We have*

$$\begin{aligned} & \mathbb{E}\|W\|_2 \\ \leq & \|M\|_2 + c\left(\max_i \|A_{i,:}\| + \max_j \|A_{:,j}\| + \mathbb{E}\max_{i,j}|G_{ij}|\right), \end{aligned} \tag{6}$$

where c is a constant independent of W.

Proof. From our hypothesis we have

$$\mathbb{E}\|W\|_2 = \mathbb{E}\|M+G\|_2. \tag{7}$$

From the triangle inequality of spectral norm,

$$\|M+G\|_2 \leq \|M\|_2 + \|G\|_2. \tag{8}$$

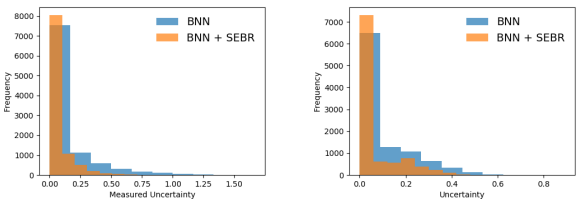
As matrix M is a constant matrix and matrix G is a Gaussian random matrix,

$$\mathbb{E}(\|M\|_2 + \|G\|_2) = \|M\|_2 + \mathbb{E}\|G\|_2. \tag{9}$$

According to the Conjecture 1.2 in [1], for our mean-zero Gaussian random matrix G, we have

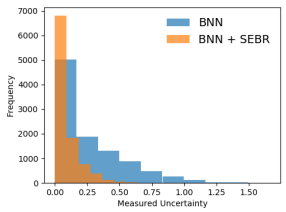
$$\mathbb{E}\|G\|_2 \leq c\left(\max\|A_{i,:}\| + \max\|A_{:,j}\| + \mathbb{E}\max_{i,j}|G_{ij}|\right) \tag{10}$$

Combining Equation (7)-(10) above, we prove the proposition. □

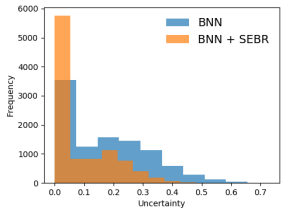


(a) Aleatoric Uncertainty (b) Epistemic Uncertainty

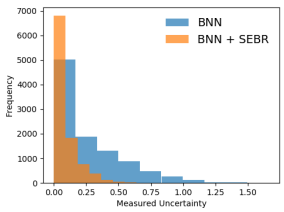
Figure S1. Uncertainties measured by Bayesian neural networks on data without noise. Models trained with SEBR have lower uncertainties on the predictions. *Best viewed in color.*



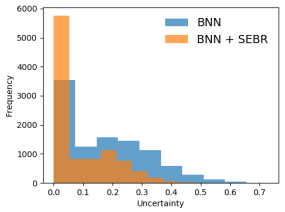
(a) Aleatoric, FGSM ($\ell_\infty = 0.05$)



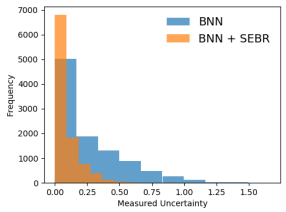
(b) Epistemic, FGSM ($\ell_\infty = 0.05$)



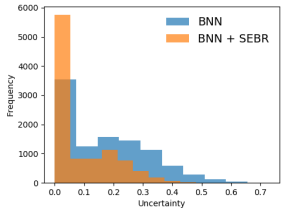
(c) Aleatoric, FGSM ($\ell_\infty = 0.2$)



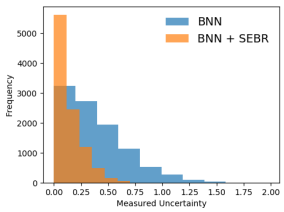
(d) Epistemic, FGSM ($\ell_\infty = 0.2$)



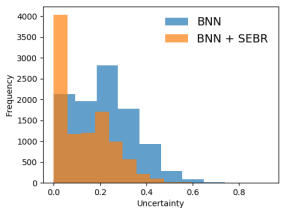
(e) Aleatoric, FGSM ($\ell_\infty = 0.3$)



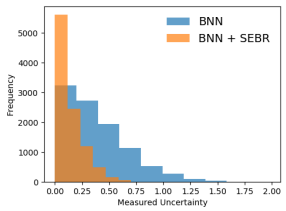
(f) Epistemic, FGSM ($\ell_\infty = 0.3$)



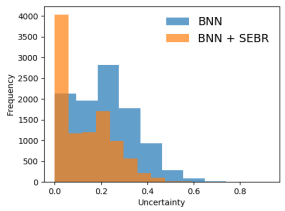
(g) Aleatoric, PGD ($\ell_\infty = 0.05$)



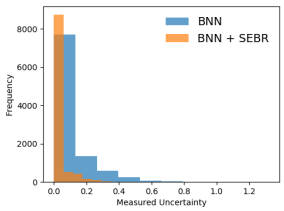
(h) Epistemic, PGD ($\ell_\infty = 0.05$)



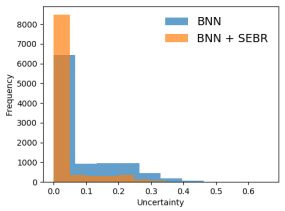
(i) Aleatoric, PGD ($\ell_\infty = 0.1$)



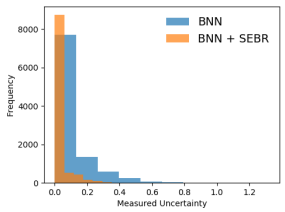
(j) Epistemic, PGD ($\ell_\infty = 0.1$)



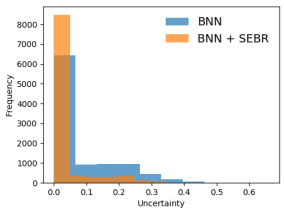
(k) Aleatoric, PGD ($\ell_\infty = 0.2$)



(l) Epistemic, PGD ($\ell_\infty = 0.2$)



(m) Aleatoric, PGD ($\ell_\infty = 0.3$)



(n) Epistemic, PGD ($\ell_\infty = 0.3$)

Figure S2. Uncertainties measured by Bayesian neural networks on data with adversarial noises. Models trained with SEBR have lower uncertainties on the predictions. *Best viewed in color.*

2. Additional Experiments

Implementation Details. We use both the Bayesian MLP model and the Bayesian CNN model in our experiments. The Bayesian MLP model contains three fully-connected layers, with neurons $784-1200-1200-10$. The Bayesian CNN structure on MNIST and Fashion-MNIST is LeNet-5 [2]. The Bayesian CNN structure on CIFAR-10 and CIFAR-100 is VGG16 [3]. The adversarial training on MNIST dataset is implemented with FGSM noises, with $\ell_\infty = 0.04$.

Hyper-parameter Selection. As we show in Section 6.2 in the paper, the selection of the hyper-parameter λ is important. To find a suitable λ for each task, we search from a series of $\ldots, 0.005, 0.01, 0.02, 0.05, 0.1, \ldots$. The settings used in our comparison is $\lambda = 0.02$ for Bayesian MLP model in MNIST dataset, $\lambda = 0.01$ for Bayesian CNN model with LeNet architecture in MNIST dataset, $\lambda = 0.05$ for both the two models in Fashion-MNIST dataset, $\lambda = 0.2$ for Bayesian CNN with VGG architecture in CIFAR-10 dataset, and $\lambda = 0.1$ for Bayesian CNN with VGG architecture in CIFAR-100 dataset.

Dataset	Attack	noise ℓ_∞	w/o. SEBR	w. SEBR
CIFAR10	/	0	91.65	**92.09**
	FGSM	0.005	58.65	**65.74**
		0.01	42.70	**54.78**
		0.02	32.73	**43.76**
	PGD	0.005	46.33	**50.40**
		0.01	9.73	**16.11**
		0.02	2.31	**2.95**
CIFAR100	/	0	**66.94**	66.56
	FGSM	0.002	45.96	**47.67**
		0.01	17.08	**21.18**
		0.02	12.52	**15.97**
	PGD	0.002	44.72	**46.85**
		0.01	2.91	**5.04**
		0.02	0.95	**1.95**

Table S1. Experiments on Bayesian CNN with VGG architecture.

More Experiment Results. Further experiments about SEBR of VGG architecture on CIFAR10 and CIFAR100 datasets are shown in Table S1. SEBR keeps effective on the larger diverse datasets and more complex network architecture.

More experiment results about the measured uncertainties on models with SEBR and without SEBR are presented in Figure S1 and S2. All results show that the models trained with SEBR have lower uncertainties, including both aleatoric uncertainties and the epistemic uncertainties, and support our proposal.

References

[1] Olivier Guédon, Aicke Hinrichs, Alexander E Litvak, and Joscha Prochno. On the expectation of operator norms of random matrices. In *Geometric aspects of functional analysis*, pages 151–162. Springer, 2017. 1

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2