Scale-invariant Bayesian Neural Networks with Connectivity Tangent Kernel

Sungyub Kim, Sihwan Park, Kyungsu Kim, Eunho Yang

ICLR 2023
Graduate School of AI, KAIST

KAIST

MLILAB
Machine Learning & Intelligence

Background

- **Data-dependent PAC-Bayes bound**
  - **PAC-Bayes bound**: Generalization gap is bounded by **KL divergence between prior and posterior**

$$\mathrm{err}_{\mathcal{D}}(\mathbb{Q}) \leq \mathrm{err}_{\mathcal{S}}(\mathbb{Q}) + \sqrt{\frac{\mathrm{KL}[\mathbb{Q}\|\mathbb{P}] + \log(2\sqrt{N}/\delta)}{2N}}$$

  - **Data-dependent PAC-Bayes bound:** Use data-dependent PAC-Bayes prior
    - Partition training $\mathcal{S}$ into $\mathcal{S}_{\mathbb{P}}$ and $\mathcal{S}_{\mathbb{Q}}$
    - Pre-train a $\mathbb{P}_{\mathcal{D}}$ with $\mathcal{S}_{\mathbb{P}}$ (Therefore, $\mathbb{P}_{\mathcal{D}}$ and $\mathcal{S}_{\mathbb{Q}}$ are **independent**.)
    - Fine-tuning $\mathbb{Q}$ with entire dataset $\mathcal{S}$.

$$\mathrm{err}_{\mathcal{D}}(\mathbb{Q}) \leq \mathrm{err}_{\mathcal{S}_{\mathbb{Q}}}(\mathbb{Q}) + \sqrt{\frac{\mathrm{KL}[\mathbb{Q}\|\mathbb{P}_{\mathcal{D}}] + \log(2\sqrt{N_{\mathbb{Q}}}/\delta)}{2N_{\mathbb{Q}}}}$$

Independent

Background

- **Invariance of generalization bounds for function-preserving transformations**
  - **Function-preserving scaling transformations (FPST)**
    - Scale transformation (diagonal matrix) of parameters that preserves function

$$f(x, \mathcal{T}(\psi)) = f(x, \psi),\ \forall x \in \mathbb{R}^D, \forall \psi \in \mathbb{R}^P$$

    - Practical examples
      - 1) Rescaling transformation of positive homogeneous (e.g., ReLU) NNs

$$(\mathcal{R}_{\gamma,l,k}(\theta))_i = \begin{cases} \gamma \cdot \theta_i & , \text{ if } \theta_i \in \{\text{param. subset connecting as input edges to } k\text{-th activation at } l\text{-th layer}\} \\ \theta_i / \gamma & , \text{ if } \theta_i \in \{\text{param. subset connecting as output edges to } k\text{-th activation at } l\text{-th layer}\} \\ \theta_i & , \text{ for } \theta_i \text{ in the other cases} \end{cases}$$

      - 2) Weight decay (WD) with Batch Normalization (BN) layers

$$(\mathcal{S}_{\gamma,l,k}(\theta))_i = \begin{cases} \gamma_k \cdot \theta_i & , \text{ if } \theta_i \in \{\text{param. subset connecting as input edges to } k\text{-th activation at } l\text{-th layer}\} \\ \theta_i & , \text{for } \theta_i \text{ in the other cases} \end{cases}$$

Background

- **Invariance of generalization bounds for function-preserving transformations**
  - Existing sharpness metric are vulnerable to FPSTs.
    - Dinh et al. (2017) showed the vulnerability of sharpness to **rescaling transformations**.
      - Rescaling two successive layers can arbitrarily sharpens NNs with preserving functions.
    - Li et al. (2018) showed the vulnerability of sharpness to **weight decay**.
      - WD improves generalizability, but sharpens NNs.
  - Prior works focused only on the scale-invariance of sharpness metrics (and **not generalization bounds**)
    - G-bounds of Tsuzuku et al. (2020) and Kwon et al. (2021) include scale-dependent terms.
    - G-bounds of Petzka et al. (2021) only holds for single-layer NNs.

Key Idea – Modification of Jacobian

- **Many sharpness metrics / generalization bounds are based on the Jacobian.**
  - Hessian, Fisher, and Gauss-Newton (GN) matrix: (Jacobian) (Jacobian)$^{\mathsf{T}}$

$$\mathbf{G} = \mathbf{J}_\theta^\top \mathbf{J}_\theta$$

  - Neural Tangent Kernel (NTK): (Jacobian)$^{\mathsf{T}}$ (Jacobian)

$$\Theta = \mathbf{J}_\theta \mathbf{J}_\theta^\top$$

- **However, Jacobian w.r.t. parameter is not invariant to FPST.**

$$\mathbf{J}_{\mathcal{T}(\theta)}(x, \mathcal{T}(\psi)) = \mathbf{J}_\theta(x, \psi)\mathcal{T}^{-1} \neq \mathbf{J}_\theta(x, \psi)$$

- **To mitigate this issue, we consider Jacobian w.r.t. connectivity.**
  - Note that we relax the binary constraints of connectivity for differentiable formulation.

$$\begin{aligned}\mathbf{J}_c(x, \mathcal{T}(\psi) \odot c)|_{c=\mathbf{1}_P} &= \mathbf{J}_\theta(x, \mathcal{T}(\psi))\mathrm{diag}(\mathcal{T}(\psi)) \\ &= \mathbf{J}_\theta(x, \psi)\mathcal{T}^{-1}\mathcal{T}\mathrm{diag}(\psi) \\ &= \mathbf{J}_c(x, \psi \odot c)|_{c=\mathbf{1}_P}\end{aligned}$$

Key Idea – Design of PAC-Bayes prior and posterior

- **Apply Bayesian Linear Regression for PAC-Bayes prior and posterior**
  - PAC-Bayes prior: Isotropic Gaussian distribution centered at **pre-trained parameter**

$$\mathbb{P}_{\theta^*}(\psi) := \mathcal{N}(\psi \,|\, \theta^*, \alpha^2 \mathrm{diag}(\theta^*)^2)$$

  - PAC-Bayes posterior: **Posterior of Bayesian Linear Regression** given PAC-Bayes prior

$$\mathbb{Q}_{\theta^*}(\psi) = \mathcal{N}(\psi | \theta^* + \theta^* \odot \mu_{\mathbb{Q}}, \mathrm{diag}(\theta^*) \Sigma_{\mathbb{Q}} \mathrm{diag}(\theta^*))$$

where

$$\mu_{\mathbb{Q}} := \frac{\Sigma_{\mathbb{Q}} \mathbf{J}_c^\top (\mathcal{Y} - f(\mathcal{X}, \theta^*))}{\sigma^2} = \frac{\Sigma_{\mathbb{Q}} \mathrm{diag}(\theta^*) \mathbf{J}_\theta^\top (\mathcal{Y} - f(\mathcal{X}, \theta^*))}{\sigma^2}$$

$$\Sigma_{\mathbb{Q}} := \left(\frac{\mathbf{I}_P}{\alpha^2} + \frac{\mathbf{J}_c^\top \mathbf{J}_c}{\sigma^2} \right)^{-1} = \left(\frac{\mathbf{I}_P}{\alpha^2} + \frac{\mathrm{diag}(\theta^*) \mathbf{J}_\theta^\top \mathbf{J}_\theta \mathrm{diag}(\theta^*)}{\sigma^2} \right)^{-1}$$

  - Check Appendix D for the detailed derivation!

Theoretical results

- **PAC-Bayes-CTK and its invariance**

Theorem 2.2 (PAC-Bayes-CTK and its invariance)**.** *Let us assume pre-trained parameter θ^* with data $\mathcal{S}_{\mathbb{P}}$. By applying $\mathbb{P}_{\theta^*}$ and $\mathbb{Q}_{\theta^*}$ to data-dependent PAC-Bayes bound (equation 2), we get*

$$\mathrm{err}_{\mathcal{D}}(\mathbb{Q}_{\theta^*}) \leq \mathrm{err}_{\mathcal{S}_{\mathbb{Q}}}(\mathbb{Q}_{\theta^*}) + \sqrt{\overbrace{\underbrace{\frac{\mu_{\mathbb{Q}}^{\top}\mu_{\mathbb{Q}}}{4\alpha^2 N_{\mathbb{Q}}}}_{\text{(average) perturbation}} + \underbrace{\sum_{i=1}^{P}\frac{h\left(\beta_i\right)}{4N_{\mathbb{Q}}}}_{\text{sharpness}}}^{\text{KL divergence}} + \frac{\log(2\sqrt{N_{\mathbb{Q}}}/\delta)}{2N_{\mathbb{Q}}}} \tag{8}$$

where $\{\beta_i\}_{i=1}^{P}$ are eigenvalues of $(\mathbf{I}_P + \frac{\alpha^2}{\sigma^2}\mathbf{J}_c^{\top}\mathbf{J}_c)^{-1}$ and $h(x) := x - \log(x) - 1$. ***This upper bound is invariant to*** *$\mathcal{T}$ for the function-preserving scale transformation by Proposition 2.1.*

- **Perturbation term** relates to the Complexity Measure of Data (CMD) term of Arora et al. (2019).
- **Sharpness term** is positively correlated to the eigenvalues of CTK.
- Each term in PAC-Bayes-CTK is invariant to FPST.

Empirical results

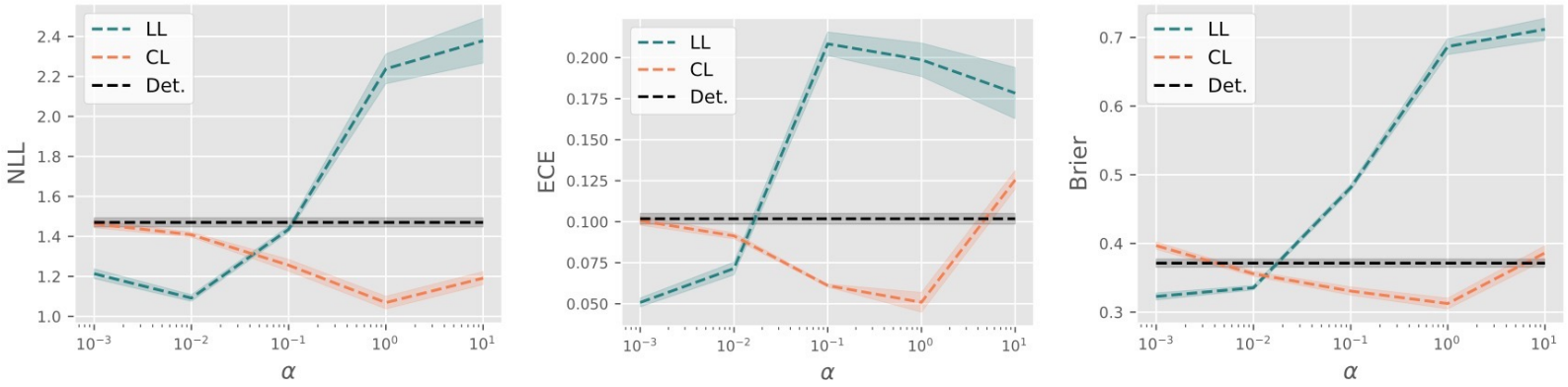
- **Invariance of PAC-Bayes-CTK**
  - The tightness of PAC-Bayes-CTK is not changed by FPSTs.

Table 1: Comparison between PAC-Bayes-CTK and PAC-Bayes-NTK for ResNet-18

CIFAR-10	PAC-Bayes-CTK				PAC-Bayes-NTK			
Parameter scale	0.5	1.0	2.0	4.0	0.5	1.0	2.0	4.0
Trace ($\times 10^{-4}$)	1.91 ± 0.04	1.91 ± 0.04	1.91 ± 0.04	1.91 ± 0.04	8793.18 ± 227.31	2590.97 ± 62.10	1107.58 ± 20.64	766.26 ± 11.80
Perturbation	6.26 ± 0.15	5.72 ± 0.09	5.77 ± 0.08	5.84 ± 0.05	636.55 ± 12.16	564.38 ± 7.56	438.21 ± 10.73	288.24 ± 6.38
Sharpness	28.92 ± 0.21	28.96 ± 0.22	28.95 ± 0.22	28.95 ± 0.20	728.80 ± 2.69	602.17 ± 2.77	502.32 ± 2.32	441.91 ± 2.08
KL	17.59 ± 0.17	17.34 ± 0.15	17.36 ± 0.15	17.39 ± 0.13	682.68 ± 4.74	583.27 ± 2.88	470.27 ± 4.32	365.07 ± 2.96
Test err. ($\times 10^2$)	4.82 ± 0.12	4.78 ± 0.11	4.78 ± 0.11	4.77 ± 0.12	13.00 ± 0.53	8.26 ± 0.17	6.94 ± 0.09	6.25 ± 0.08
Bound ($\times 10^2$)	9.21 ± 0.04	9.21 ± 0.02	9.21 ± 0.02	9.21 ± 0.03	39.00 ± 0.61	32.07 ± 0,25	28.24 ± 0.06	24.84 ± 0.11
CIFAR-100	PAC-Bayes-CTK				PAC-Bayes-NTK			
Parameter scale	0.5	1.0	2.0	4.0	0.5	1.0	2.0	4.0
Trace ($\times 10^{-4}$)	2.33 ± 0.37	2.34 ± 0.37	2.33 ± 0.37	2.34 ± 0.37	5830.55 ± 532.26	1913.90 ± 244.05	1089.53 ± 104.06	955.05 ± 66.16
Perturbation	14.54 ± 0.25	14.32 ± 0.24	14.08 ± 0.20	13.84 ± 0.14	620.18 ± 6.83	569.16 ± 6.94	459.16 ± 2.86	329.29 ± 3.34
Sharpness	42.52 ± 5.26	42.53 ± 5.26	42.52 ± 5.27	42.53 ± 5.26	694.45 ± 8.67	580.20 ± 12.22	519.78 ± 7.89	504.74 ± 6.10
KL	28.53 ± 2.51	28.42 ± 2.52	28.30 ± 2.55	28.19 ± 2.56	657.31 ± 7.21	574.68 ± 8.95	489.47 ± 3.57	417.02 ± 3.98
Test err. ($\times 10^2$)	21.78 ± 0.14	21.82 ± 00.18	21.84 ± 0.19	21.86 ± 0.21	43.39 ± 0.64	37.06 ± 0.26	32.32 ± 0.32	28.37 ± 0.13
Bound ($\times 10^2$)	27.74 ± 0.37	27.76 ± 0.40	27.75 ± 0.42	27.75 ± 0.42	68.44 ± 0.82	59.96 ± 0.19	51.90 ± 0.04	44.80 ± 0.25

Empirical results

- **Robustness to the selection of prior scale.**
  - Standard **Linearized Laplace (LL)** is sensitive to the prior scale (α) and has different optimal sharpness for each metric.
  - Our PAC-Bayes posterior, called **Connecitivity Laplace (CL)**, was robust to the prior scale and consistent between metrics.



(a) NLL (b) ECE (c) Brier Score

Conclusion

- **Our contribution is threefold:**

  - We introduced **a novel PAC-Bayes bound** guarantees **invariance for FPSTs** with a broad class of networks. We empirically verify this bound gives tight results for ResNet with 11M parameters.

  - Based on the sharpness term of our bound, we provided a low-complexity sharpness metric, called **Connectivity Sharpness**.

  - To prevent overconfident predictions, we showed how our PAC-Bayes posterior can be used to solve **pitfalls of WD with BNs**, proving its practicality.