Flat Seeking Bayesian Neural Networks

Nguyen et al. 2023

Reviewer: Jihu Lee

IDEA lab
Department of Statistics
Seoul National University

January 8, 2024

Introduction

Objective of work

- Propose a sharpness-aware posterior
- Propose its variational approach

Problem Setting

Notations

- $f_\theta(x)$: model (NN) with $\theta \in \Theta$
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$: training set, where $(x_i, y_i) \sim \mathcal{D}$
- $\mathbb{P}, p(\theta)$: prior distribution, prior density
- $l(f_\theta(x), y)$: loss function

Standard Posterior

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

Likelihood ($\lambda \geq 0$: regularization parameter)

$$p(y|x, \mathcal{S}, \theta) \propto \exp\left\{-\frac{\lambda}{|\mathcal{S}|} l(f_\theta(x), y)\right\} = \exp\left\{-\frac{\lambda}{n} l(f_\theta(x), y)\right\} \tag{2}$$

Problem Setting

Population & Empirical losses

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(f_\theta(x), y)] \tag{3}$$

$$\mathcal{L}_{\mathcal{S}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{S}}[l(f_\theta(x), y)] = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i) \tag{4}$$

Standard Posterior (rewritten)

$$q(\theta|\mathcal{S}) \propto \exp\{-\lambda\mathcal{L}_{\mathcal{S}}(\theta)\}\, p(\theta) \tag{5}$$

Population & Empirical losses over $\mathbb{Q}$ (distribution over θ with density q)

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) = \int_{\Theta} \mathcal{L}_{\mathcal{D}}(\theta)d\mathbb{Q}(\theta) = \int_{\Theta} \mathcal{L}_{\mathcal{D}}(\theta)q(\theta)d\theta \tag{6}$$

$$\mathcal{L}_{\mathcal{S}}(\mathbb{Q}) = \int_{\Theta} \mathcal{L}_{\mathcal{S}}(\theta)d\mathbb{Q}(\theta) = \int_{\Theta} \mathcal{L}_{\mathcal{S}}(\theta)q(\theta)d\theta \tag{7}$$

Theorem (3.1)

Consider the following optimization problem:

$$\min_{\mathbb{Q} \ll \mathbb{P}} \{\lambda \mathcal{L}_{\mathcal{S}}(\mathbb{Q}) + KL(\mathbb{Q}, \mathbb{P})\} \tag{8}$$

This optimization has a closed-form optimal solution $\mathbb{Q}^*$ *with the density*

$$q^*(\theta) \propto \exp\{-\lambda \mathcal{L}_{\mathcal{S}}(\theta)\} \, p(\theta) \tag{9}$$

which is exactly the standard posterior $\mathbb{Q}_{\mathcal{S}}$ *with the density* $q(\theta|\mathcal{S})$.

Optimization Problem (population)

$$\min_{\mathbb{Q} \ll \mathbb{P}} \{\lambda \mathcal{L}_{\mathcal{D}}(\mathbb{Q}) + KL(\mathbb{Q}, \mathbb{P})\} \tag{10}$$

Theory

Theorem (3.2)

Assume that Θ *is a compact set. Under some mild conditions given any* $\delta \in [0,1]$*, with the probability at least* $1-\delta$ *over the choice of* $\mathcal{S} \sim \mathcal{D}^n$*, for any distribution* $\mathbb{Q}$*, we have*

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) \leq \mathbb{E}_{\theta \sim \mathbb{Q}}\left[\max_{\theta' : ||\theta' - \theta|| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta')\right] + f\left(\max_{\theta \in \Theta} ||\theta||^2, n\right) \tag{11}$$

where f *is a non-decreasing function w.r.t. the first variable and approaches 0 when the training size* n *approaches* ∞*.*

Upper bound (rewritten)

$$\mathcal{L}_{\mathcal{D}}(\mathbb{Q}) \leq \mathcal{L}_{\mathcal{S}}(\mathbb{Q}) \tag{12}$$

$$+ \mathbb{E}_{\theta \sim \mathbb{Q}} \left[\max_{\theta : ||\theta' - \theta|| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta') - \mathcal{L}_{\mathcal{S}}(\theta) \right] \tag{13}$$

$$+ f\left(\max_{\theta \in \Theta} ||\theta||^2, n \right) \tag{14}$$

- First term: empirical losses over $\mathbb{Q}$
- Second term: sharpness over $\mathbb{Q}$
- Last term: constant

Theory

Optimization Problem

$$\min_{\mathbb{Q}\ll\mathbb{P}} \left\{ \lambda \mathbb{E}_{\theta\sim\mathbb{Q}} \left[\max_{\theta':||\theta'-\theta||\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta') \right] + KL(\mathbb{Q},\mathbb{P}) \right\} \tag{15}$$

- Considering sharpness-aware loss $\max_{\theta':||\theta'-\theta||\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')$ makes generalization ability better

Theory

Theorem (3.3)

The optimal solution of eq. (15) *is the sharpness-aware posterior distribution* $\mathbb{Q}_{\mathcal{S}}^{SA}$ *with the density funciton* $q^{SA}(\theta|\mathcal{S})$*:*

$$q^{SA}(\theta|\mathcal{S}) \propto \exp\left\{-\lambda \max_{\theta':||\theta'-\theta||\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')\right\} p(\theta) = \exp\left\{-\lambda\mathcal{L}_{\mathcal{S}}(s(\theta))\right\} p(\theta) \tag{16}$$

where $s(\theta) = \operatorname*{argmax}_{\theta':||\theta'-\theta||\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta')$*.*

- Closed form of the sharpness-aware posterior distribution $\mathbb{Q}_{\mathcal{S}}^{SA}$

Sharpness-aware Likelihood

$$p^{SA}(y|x,\mathcal{S},\theta) \propto \exp\left\{-\frac{\lambda}{n} l(f_{s(\theta)}(x), y)\right\} \tag{17}$$

where $s(\theta) = \underset{\theta':||\theta'-\theta||\leq\rho}{} \mathcal{L}_{\mathcal{S}}(\theta')$

Sharpness-aware Posterior

$$q^{SA}(\theta|\mathcal{S}) \propto \prod_{i=1}^{n} p^{SA}(y_i|x_i,\mathcal{S},\theta)p(\theta) \tag{18}$$

- $X = [x_1, \dots, x_n], Y = [y_1, \dots, y_n]$
- $\{q_\phi(\theta) : \phi \in \Phi\}$: approximate posterior family

ELBO

$$\max_{q_\phi} \left\{ \mathbb{E}_{q_\phi(\theta)} \left[\sum_{i=1}^{n} \log p^{SA}(y_i | x_i, \mathcal{S}, \theta) \right] - KL(q_\phi, p) \right\} \tag{19}$$

$$\Leftrightarrow \min_{q_\phi} \left\{ \lambda \mathbb{E}_{q_\phi(\theta)} \left[\max_{\theta' : ||\theta' - \theta|| \leq \rho} \mathcal{L}_{\mathcal{S}}(\theta') \right] + KL(q_\phi, p) \right\} \tag{20}$$

Table 1: Classification score on CIFAR-100 dataset.Each experiment is repeated three times with different random seeds and reports the mean and standard deviation.

	PreResNet-164			WideResNet28x10		
Method	ACC ↑	NLL ↓	ECE ↓	ACC ↑	NLL ↓	ECE ↓
Variational inference						
MC-Dropout	79.50 ± 0.37	0.9162 ± 0.0103	0.0993 ± 0.0033	82.30 ± 0.19	0.6500 ± 0.0049	0.0574 ± 0.0028
F-MC-Dropout	**81.06 ± 0.44**	**0.7027 ± 0.0049**	**0.0514 ± 0.0047**	**83.24 ± 0.11**	**0.6144 ± 0.0068**	**0.0250 ± 0.0027**
Deep-ens	82.08 ± 0.42	0.7189 ± 0.0108	0.0334 ± 0.0064	83.04 ± 0.15	0.6958 ± 0.0335	0.0483 ± 0.0017
F-Deep-ens	**82.54 ± 0.10**	**0.6286 ± 0.0022**	**0.0143 ± 0.0041**	**84.52 ± 0.03**	**0.5644 ± 0.0106**	**0.0191 ± 0.0039**
Markov chain Monte Carlo						
SGLD	80.13 ± 0.01	0.7604 ± 0.0010	0.1161 ± 0.0031	81.38 ± 0.10	0.7123 ± 0.0204	0.0958 ± 0.0004
F-SGLD	**80.82 ± 0.02**	**0.7276 ± 0.0012**	**0.1085 ± 0.0008**	**82.12 ± 0.16**	**0.6722 ± 0.0112**	**0.0820 ± 0.0021**
Sample						
SWAG-Diag	80.18 ± 0.50	0.6837 ± 0.0186	**0.0239 ± 0.0047**	82.40 ± 0.09	0.6150 ± 0.0029	0.0322 ± 0.0018
F-SWAG-Diag	**81.01 ± 0.29**	**0.6645 ± 0.0050**	0.0242 ± 0.0039	**83.50 ± 0.29**	**0.5763 ± 0.0120**	**0.0151 ± 0.0020**
SWAG	79.90 ± 0.50	**0.6595 ± 0.0019**	0.0587 ± 0.0048	82.23 ± 0.19	0.6078 ± 0.0006	**0.0113 ± 0.0020**
F-SWAG	**80.93 ± 0.27**	0.6704 ± 0.0049	**0.0350 ± 0.0025**	**83.57 ± 0.26**	**0.5757 ± 0.0136**	0.0196 ± 0.0015

References

Nguyen, Van-Anh et al. (2023). "Flat Seeking Bayesian Neural Networks". In: *arXiv preprint arXiv:2302.02713*.