TSUBAME 共同利用　令和年度　学術利用　成果報告書

# 利用課題名　22IAJ - Improving Sequential Bayesian Inference

利用課題責任者
Mohammad Emtiyaz Khan

所属
Approximate Bayesian Inference Team, RIKEN Center for Advanced Intelligence Project
http://team-approx-bayes.github.io

Bayesian principles offer a mathematically rigorous way to fundamentally improve and rethink continual learning and adaptive AI systems. However, performing approximate Bayesian inference for deep learning and large neural networks is still a challenging open question and requires large amounts of compute resources. In this project, we employed TSUBAME3.0 to develop new state-of-the-art Bayesian deep learning methods. These lay the foundation for our future work on developing robust practical AI systems which can continually learn and update their beliefs by performing sequential Bayesian inference.

*Keywords:* Approximate Bayesian Inference, Continual Learning, Deep Learning, Optimization, Numerical Algorithms

## 背景と目的

Current machine-learning and AI systems are very rigid and struggle to adapt to new situations. Sequential Bayesian inference offers a principled way to update the current beliefs based on newly observed information. However, performing exact Bayesian inference for large neural network models is computationally intractable. The goal of this project was to develop new state-of-the-art methods for approximate Bayesian inference. These algorithms are expected to enable AI systems to become more robust to perturbations and continually adapt like humans and animals.

## 概要

The report is structured in three sections. Each section corresponds to an algorithm for Bayesian deep learning we researched and developed using TSUBAME3.0.

1. Second-Order Optimization via Bayes (SOBA)
2. The Lie-Group Bayesian Learning Rule (LieGroup-BLR)
3. Sharpness-Aware Minimization as an Optimal Relaxation of Bayes (bSAM)

## 結果および考察

### 1. **Second-Order Optimization via Bayes (SOBA)**

Most of modern deep learning uses optimizers to train deep neural networks. The following equations show the update-rule of  the popular RMSprop / Adam deep learning optimizer:

$$g \leftarrow \nabla \ell(\theta)$$
$$s \leftarrow (1-\rho)s + \rho g^2$$
$$\theta \leftarrow \theta - \alpha(\sqrt{s}+\gamma)^{-1} g$$

To enable deep learning systems to adapt to new situations and continually learn, it is essential for these models to know what they know. However, these models should also understand what they don't know yet and should not be overconfident in such situations. Starting from a rigorous mathematical theory, we have developed a modification of RMSprop which can estimate uncertainty in deep learning by computing an approximate Bayesian posterior. The algorithm is shown in the following equations, where the changes to RMSprop/Adam are highlighted in red.

$$g \leftarrow \nabla \ell(\theta+\epsilon), \ \text{where } \epsilon \sim \mathcal{N}(0, s^{-1})$$
$$\theta \leftarrow \theta - \alpha\, (s)^{-1} g$$
$$s \leftarrow (1-\rho)s + \rho(s\epsilon g) + \rho^2/2s(\epsilon g - 1)^2$$

This method allows the deep learning model to be uncertain about its decisions in regions where there is no data, as shown in the following Figure 1.

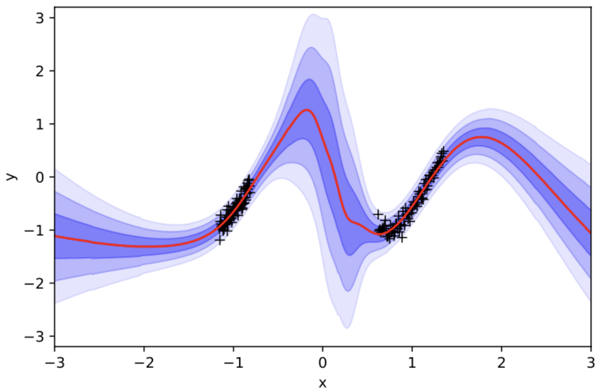



Figure 1. When training neural networks, using our SOBA optimizer instead of RMSprop allows the model to be uncertain in situations where there is little observed data. This uncertainty is essential to quickly adapt to new beliefs once more information is observed.

The SOBA optimizer offers state-of-the-art performance for Bayesian deep learning. A previous version was part of our winning solution to the NeurIPS 2021 Challenge on Approximate Inference in Bayesian Deep Learning ([https://izmailovpavel.github.io/neurips_bdl_competition/](https://izmailovpavel.github.io/neurips_bdl_competition/)).

TSUBAME3.0 was essential in developing and benchmarking this method. Large computational resources are required when training neural networks on big datasets such as ImageNet or simply to train modern neural network architectures. In future projects on sequential Bayesian inference for large neural networks, SOBA will be one of the key ingredients to obtain accurate models with well-calibrated uncertainty.

## 2. The Lie-Group Bayesian Learning Rule (LieGroup-BLR)

Due to its mathematical formulation, the SOBA optimizer appears to be restricted to perturb the weights with Gaussian noise (see the Algorithm before Figure 1). In a follow-up work, we found a way to generalize SOBA from Gaussian noise to arbitrary noise distributions. This is achieved by using the mathematical framework of Lie groups (see the following Figure 2).

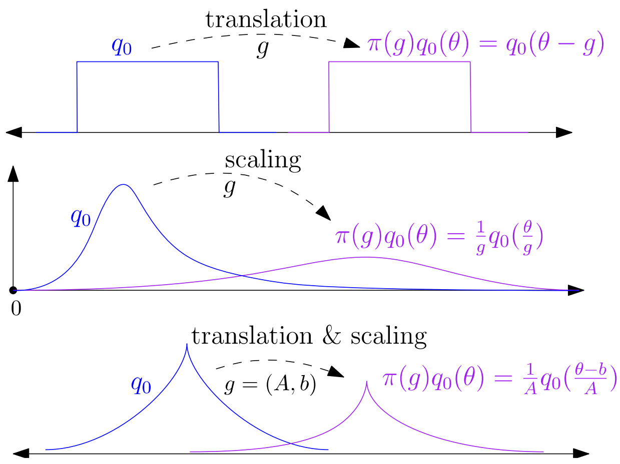



Figure 2. In approximate Bayesian inference, a probability distribution (called the Bayesian posterior) is estimated. SOBA is restricted to Gaussian distributions, but using the Lie-group formalism, we were able to generalize to other distributions. The figure shows Uniform, Rayleigh and Laplace distributions.

In Figure 3, we compare the LieGroup-BLR to SGD and the SOBA optimizer (here called "iVON"). We observe that the additional flexibility to choose a non-Gaussian noise distribution can sometimes improve the results.

| Method | Family $Q$ | CIFAR-10<br>Acc. ↑<br>(higher is better) | NLL ↓<br>(lower is better) | ECE ↓<br>(lower is better) |
|---|---|---|---|---|
| Additive (Alg. 1) | Uniform | $91.07_{\pm 0.08}$ | $0.365_{\pm 0.003}$ | $0.052_{\pm 0.001}$ |
| | Gaussian | $91.28_{\pm 0.11}$ | $0.328_{\pm 0.008}$ | $0.045_{\pm 0.001}$ |
| | Laplace | $91.14_{\pm 0.12}$ | $0.312_{\pm 0.005}$ | $0.039_{\pm 0.001}$ |
| Affine (Alg. 3) | Uniform | $91.60_{\pm 0.05}$ | $0.300_{\pm 0.002}$ | $0.040_{\pm 0.001}$ |
| | Gaussian | $91.53_{\pm 0.10}$ | $0.294_{\pm 0.004}$ | $0.036_{\pm 0.001}$ |
| | Laplace | $91.87_{\pm 0.04}$ | $0.272_{\pm 0.002}$ | $0.029_{\pm 0.001}$ |
| SGD | – | $91.22_{\pm 0.07}$ | $0.354_{\pm 0.006}$ | $0.050_{\pm 0.001}$ |
| iVON | Gaussian | $91.80_{\pm 0.05}$ | $0.288_{\pm 0.003}$ | $0.038_{\pm 0.001}$ |
| VOGN | Gaussian | $91.32_{\pm 0.09}$ | $0.264_{\pm 0.003}$ | $0.011_{\pm 0.000}$ |

Figure 3. The developed Lie-Group optimization methods improve the performance over traditional SGD optimization in terms of accuracy and uncertainty on image classification datasets with neural network models.

## 3. Sharpness-Aware Minimization as an Optimal Relaxation of Bayes

Sharpness-Aware Minimization (SAM) is a recent method proposed by Google researchers to improve generalization and robustness in deep learning. The method has been very influential over the last years, accumulating over 500 citations in a short timespan.

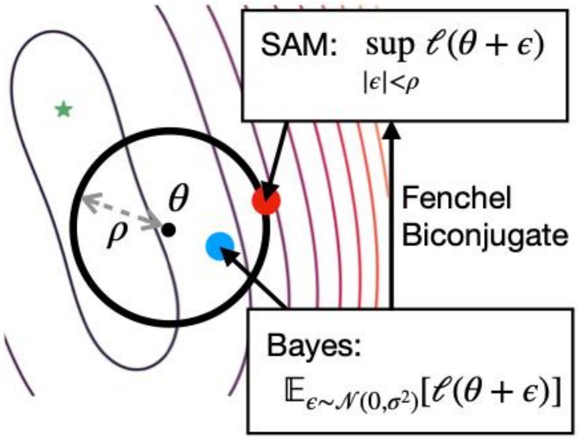
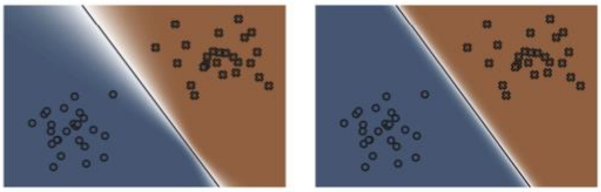



Figure 4. Sharpness-Aware Minimization considers the worst-case perturbation in a neighborhood, whereas Bayesian approaches average over many random perturbations.

In our recent work we discovered a connection of SAM to Bayesian approaches, as illustrated in the above Figure 4. This connection enabled us to propose a Bayesian-SAM method (called bSAM) which extends SAM to obtain uncertainty estimates.

We illustrate the uncertainty estimation in Figure 5. The decision boundary obtained by bSAM is more "blurred out", especially in regions where there is no data. In these regions, the model is uncertain about the correct prediction.



(b) bSAM (c) SAM (point est.)

Figure 5. bSAM extends SAM to obtain uncertainty estimates without computational overhead.

| Model / Dataset | Method | Accuracy ↑ (higher is better) | NLL ↓ (lower is better) | ECE ↓ (lower is better) | AUROC ↑ (higher is better) |
|---|---|---|---|---|---|
| ResNet-20-FRN / CIFAR-10 | SGD | $91.68_{(0.26)}$ | $0.29_{(0.008)}$ | $0.0397_{(0.002)}$ | $0.915_{(0.002)}$ |
| | SAM-SGD | $\mathbf{92.29}_{(0.39)}$ | $0.25_{(0.004)}$ | $0.0266_{(0.003)}$ | $0.920_{(0.003)}$ |
| | Adam | $89.97_{(0.27)}$ | $0.41_{(0.021)}$ | $0.0610_{(0.003)}$ | $0.900_{(0.003)}$ |
| | SAM-Adam | $91.57_{(0.21)}$ | $0.26_{(0.004)}$ | $0.0329_{(0.002)}$ | $0.918_{(0.001)}$ |
| | bSAM (ours) | $\mathbf{92.16}_{(0.16)}$ | $\mathbf{0.23}_{(0.003)}$ | $\mathbf{0.0057}_{(0.002)}$ | $\mathbf{0.925}_{(0.001)}$ |
| ResNet-20-FRN / CIFAR-100 | SGD | $66.48_{(0.10)}$ | $1.20_{(0.007)}$ | $0.0524_{(0.004)}$ | $0.846_{(0.002)}$ |
| | SAM-SGD | $67.27_{(0.22)}$ | $1.19_{(0.011)}$ | $0.0481_{(0.001)}$ | $0.848_{(0.002)}$ |
| | Adam | $61.76_{(0.67)}$ | $1.66_{(0.049)}$ | $0.1582_{(0.006)}$ | $0.826_{(0.003)}$ |
| | SAM-Adam | $65.34_{(0.32)}$ | $1.23_{(0.012)}$ | $\mathbf{0.0166}_{(0.003)}$ | $0.847_{(0.002)}$ |
| | bSAM (ours) | $\mathbf{68.22}_{(0.44)}$ | $\mathbf{1.10}_{(0.013)}$ | $0.0258_{(0.003)}$ | $\mathbf{0.857}_{(0.004)}$ |

Figure 6. The proposed bSAM method improves accuracy and uncertainty estimation when compared to SAM.

Using TSUBAME3.0 was essential in developing and researching our bSAM method on large deep learning models.

まとめ、今後の課題

In this project, we have developed new state-of-the-art methods for approximate Bayesian inference (SOBA, LieGroup-BLR and bSAM). Our next step is to use these algorithms to improve sequential and continual learning of AI systems, making them more adaptive and robust. Furthermore, future effort will be spent on applying the approximate Bayesian inference algorithms developed in this project on larger problems and machine learning models. For instance, we want to show that our methods can be useful in large-scale settings like large language models.