FY2022/2022年度

Approximate Bayesian Inference Team

Mohammad Emtiyaz Khan

近似ベイズ推論チーム

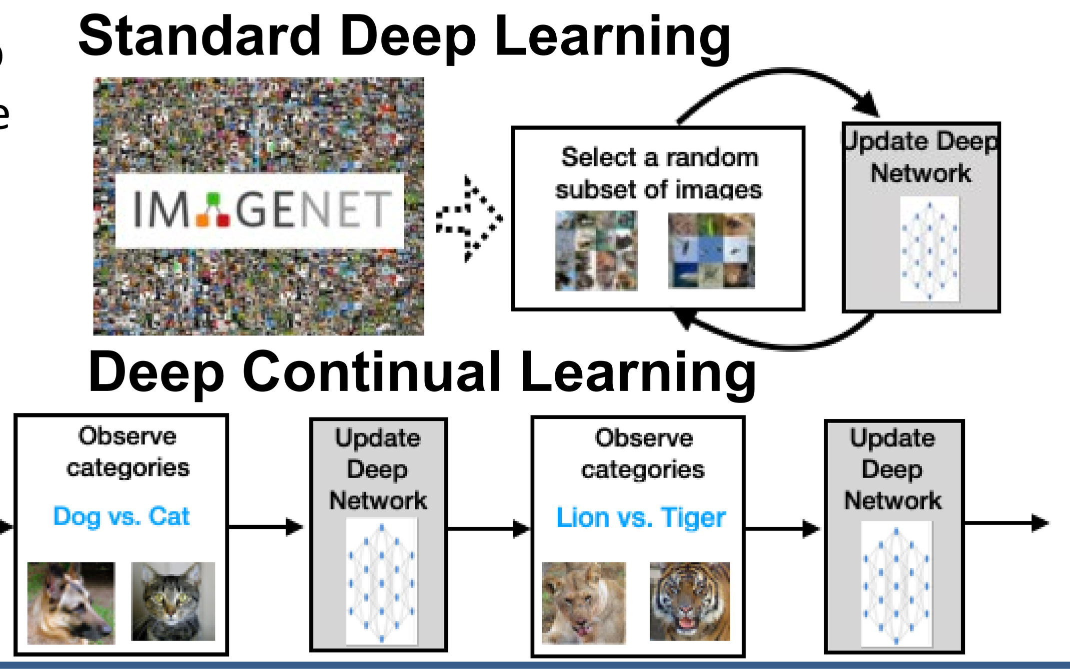
RIKEN AIP Center for Advanced Intelligence Project

Overview and Goals

Goal: AI systems that can learn and improve continually throughout their lives, just like humans and animals. Currently, deep learning (DL) requires a large amount of data which is costly and rigid, leading to a system that is unable to quickly adapt. We aim to fix this with a new learning paradigm based on Bayesian principles.

Summary of our research highlights in the year 2022 (paper number shown in green boxes)

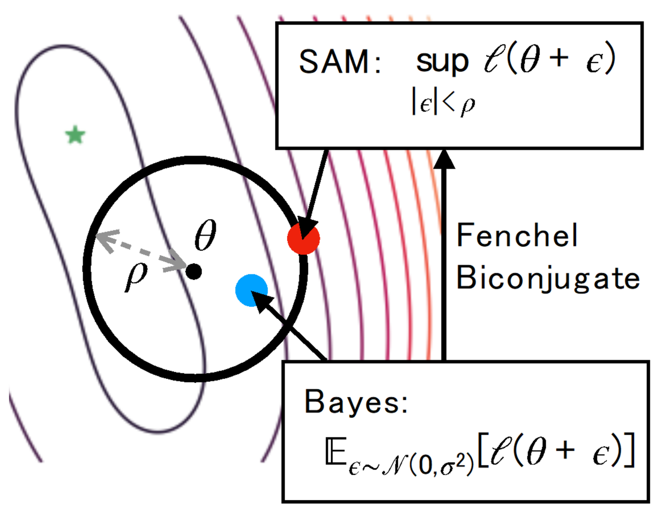
- 1 We show a robust deep-learning method is related to Bayes and add uncertainty to it (top 75/5000 papers at ICLR2023)
- 2 We extend the Bayesian learning rule using Lie-groups, simplifying gradient computations and eliminating retractions.
- 3 We design a generally unimprovable procedure for relaxation time estimation in non-reversible Markov chains.
- 4 We simplify momentum-based Riemannian optimization over positive-semi-definite matrix submanifold.
- 5 We use low-rank matrix completion techniques to reconstruct partially-observed high-dimensional time series.
- 6-10 Analysis of MMD estimation, Neural processes, A Dataset for African Languages, Deviation inequalities, and more …!

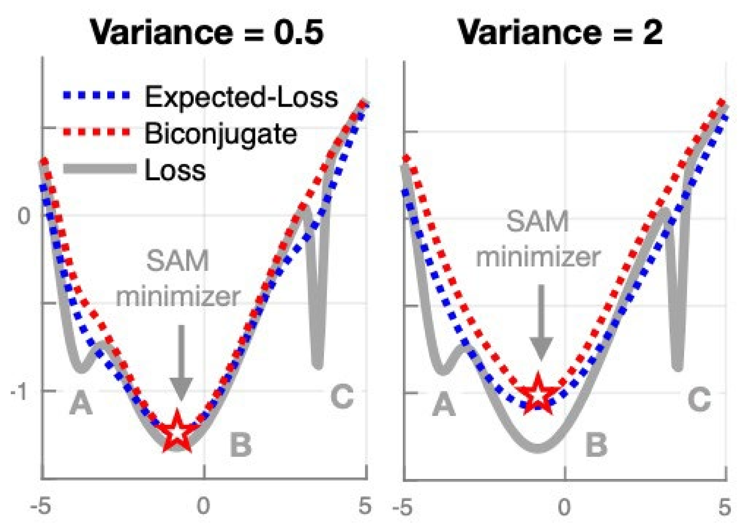


SAM as an Optimal Relaxation of Bayes

Problem: Sharpness-Aware Minimization (SAM) by Foret et al. improve significantly over SGD but the reasons behind its success are unclear.

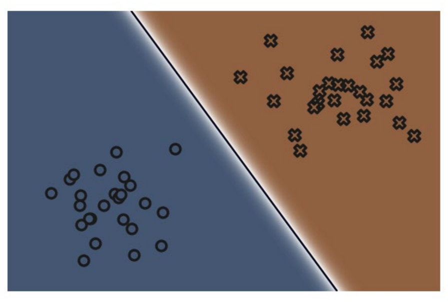
Solution: We show that SAM is equivalent to an optimal relaxation of Bayes obtained by using Fenchel biconjugate (left figure). SAM can be seen as "smoothing" the objective using a "posterior variance" and always upper-bounding Bayes (right).
Our paper [1] is among top-5% of all accepted papers (75 out of 5000 submissions)



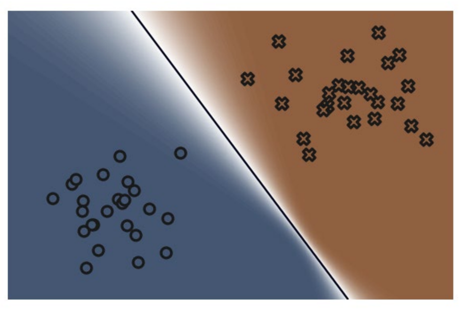


Contribution #2: Our Bayesian-SAM improves accuracy over SGD (by 8%), SAM-Adam (by 10%), and Adam (by 22%), but also improves AUROC (for CIFAR-100 using ResNet-20 with 270K params).

	Accuracy	AUROC
SGD	$55.82_{(0.97)}$ +8%	$0.811_{(0.004)}$
SAM-SGD	$58.58_{(0.59)}$	$0.827_{(0.003)}$
SWAG	$56.53_{(0.40)}$	$0.814_{(0.004)}$
VOGN	$59.83_{(0.75)}$	$0.830_{(0.002)}$
Adam	$39.73_{(0.97)}$ +10%	$0.775_{(0.004)}$
SAM-Adam	$53.25_{(0.80)}$ +22%	$0.818_{(0.005)}$
bSAM (ours)	$\mathbf{62.64}_{(0.33)}$	$\mathbf{0.841}_{(0.004)}$



SAM is overconfident (narrow white region)



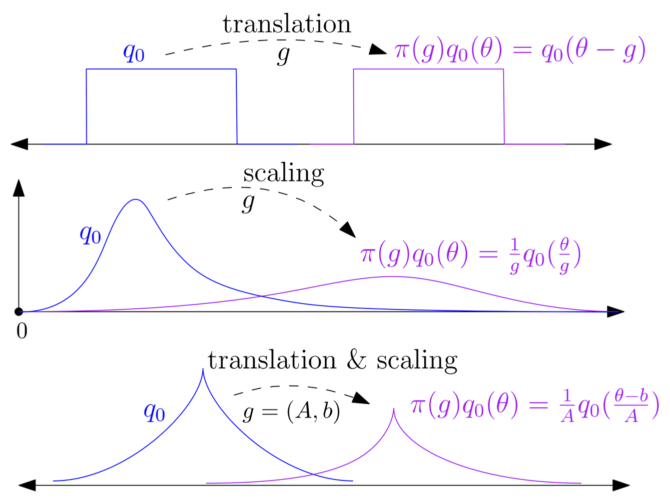
Bayesian-SAM fixes this (more blurry away from the data)

1. *Moellenhoff, Khan*, SAM as an Optimal Relaxation of Bayes, *ICLR 2023* ***(oral, notable top-5%)***

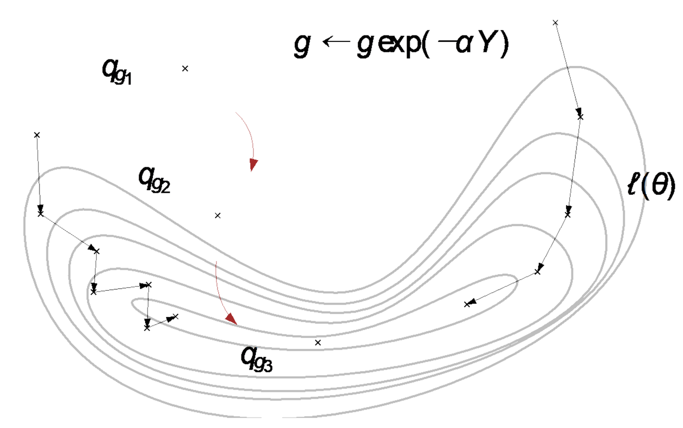
The Lie-Group Bayesian Learning Rule

Problem: Many popular algorithms can be derived from the Bayesian learning rule of Khan and Rue (2021) but the rule can be difficult to apply, e.g., gradients are difficult to compute, and steps can lead to invalid distributions (e.g., –ve variances).

Solution: We extend the rule by using **Lie-groups** which solves the above problems: gradients can always be obtained by reparameterizations, and steps always stay on the manifold. Fisher computation is also simplified and only need to be done once.



Conceptual algorithm:



Result: Reduces to the update [3] when applied to exp. fam. and linearized.

Additive

$$g \leftarrow g - \alpha \mathbb{E}_{q_g}[\nabla_\theta \ell].$$

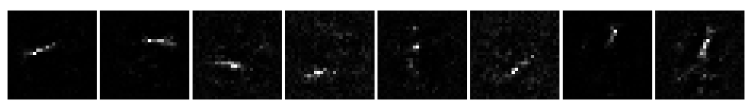
Multiplicative

$$g_i \leftarrow g_i \exp\left(-\alpha\left(\mathbb{E}_{q_g}[\theta_i \partial_i \ell] - \tau\right)\right)$$

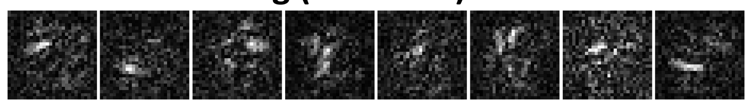
Affine

$$b_i \leftarrow b_i + \frac{c_X}{c_y} A_i \frac{\exp(-\alpha U) - 1}{U} V,$$
$$A_i \leftarrow A_i \exp(-\alpha U)$$

Multiplicative Learning (nll=0.058)



Additive Learning (nll=0.083)



Dataset	Metric	Affine (Alg. 3) gaussian	laplace	uniform	Additive ([27], Alg. 1) gaussian	laplace	uniform
CIFAR-10	Acc. (↑)	$91.53_{\pm 0.10}$	$91.87_{\pm 0.04}$	$91.60_{\pm 0.05}$	$91.28_{\pm 0.11}$	$91.14_{\pm 0.12}$	$91.07_{\pm 0.08}$
	NLL (↓)	$0.294_{\pm 0.004}$	$0.272_{\pm 0.002}$	$0.300_{\pm 0.002}$	$0.328_{\pm 0.008}$	$0.312_{\pm 0.005}$	$0.365_{\pm 0.003}$
	ECE (↓)	$0.036_{\pm 0.001}$	$0.029_{\pm 0.001}$	$0.040_{\pm 0.001}$	$0.045_{\pm 0.001}$	$0.039_{\pm 0.001}$	$0.052_{\pm 0.001}$
CIFAR-100	Acc. (↑)	$66.55_{\pm 0.10}$	$66.44_{\pm 0.10}$	$66.08_{\pm 0.12}$	$64.61_{\pm 0.20}$	$64.85_{\pm 0.13}$	$64.29_{\pm 0.09}$
	NLL (↓)	$1.255_{\pm 0.005}$	$1.247_{\pm 0.006}$	$1.288_{\pm 0.007}$	$1.390_{\pm 0.008}$	$1.359_{\pm 0.006}$	$1.437_{\pm 0.004}$
	ECE (↓)	$0.079_{\pm 0.002}$	$0.071_{\pm 0.001}$	$0.093_{\pm 0.002}$	$0.107_{\pm 0.001}$	$0.096_{\pm 0.001}$	$0.121_{\pm 0.001}$
TinyImgNet	Acc. (↑)	$51.13_{\pm 0.16}$	$51.36_{\pm 0.14}$	$51.19_{\pm 0.12}$	$49.62_{\pm 0.15}$	$49.73_{\pm 0.18}$	$49.34_{\pm 0.14}$
	NLL (↓)	$2.098_{\pm 0.004}$	$2.101_{\pm 0.009}$	$2.099_{\pm 0.005}$	$2.204_{\pm 0.003}$	$2.184_{\pm 0.007}$	$2.234_{\pm 0.008}$
	ECE (↓)	$0.070_{\pm 0.002}$	$0.065_{\pm 0.001}$	$0.076_{\pm 0.001}$	$0.099_{\pm 0.002}$	$0.089_{\pm 0.001}$	$0.107_{\pm 0.001}$

2. *Kiral, Moellenhoff, Khan*, The Lie-Group Bayesian Learning Rule, *AISTATS 2023*.

Estimation of Relaxation Time in Non-Reversible Markov Chains

Problem: Estimate the **relaxation time** of an unknown, ergodic, non-reversible Markov chain **from a single trajectory of length m started from an arbitrary state.**

Application: This is useful in reinforcement learning, Markov Chain Monte Carlo diagnostics, and deriving generalization bounds with Markov-dependent data.

Contributions:

- Prove there exists an estimator $\hat{t}_{\text{rel}}$ for t_{rel} such that for any ergodic chain, when
$$m = \tilde{\mathcal{O}}\left(t_{\text{rel}}/\pi_\star\right)$$
where $\pi_\star$ is the minimum stationary probability, it holds w.h.p. that $\left|\hat{t}_{\text{rel}}/t_{\text{rel}} - 1\right| < 5$.
Upper bound matches known lower bounds, thus **generally unimprovable**.
- Design algorithm with **fully empirical confidence intervals** that decay in $\tilde{\mathcal{O}}(1/\sqrt{m})$.

3. *Wolfer and Kontorovich*, Improved Estimation of Relaxation Time in Non-reversible Markov Chains, *arXiv 2022.*

Other Works

Simplify momentum-based Riemannian optimization: We consider the specific case of a submanifold containing symmetric positive-definite matrices. The method uses a generalized version of local coordinates which "trivializes" the Fisher matrix.

4. *Lin, Duruisseaux, Leok, Nielsen, Khan, Schmidt*, Practical Structured Riemannian Optimization with Momentum by using Generalized Normal Coordinates, NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representation

High-dimensional time series completion: We use low-rank matrix completion techniques to reconstruct partially observed high-dimensional time series and show that periodicity or smoothness can even lead to faster rates than in the independent setting.

5. Alquier, Marie, Rosier, Tight risk bound for high dimensional time series completion, EJS 2022

Finite sample properties of parametric MMD estimation: We tackle the problem of universal estimation using a minimum distance estimator based on Maximum Mean Discrepancy, and we show its robustness to both dependence and presence of outliers.

6. *Chérief-Abdellatif, Alquier,* Finite Sample Properties of Parametric MMD Estimation: Robustness to Misspecification and Dependence, Bernoulli 2022

Improving Neural Process: We improve test-time inference for Neural Processes by incorporating and exploiting graphical-model structure among context points.

7. *Tailor, Khan, Nalisnick*, Exploiting Inference Structure in Neural Processes, UAI 2022 Workshop on Tractable Probabilistic Modeling

Name-Entity Recognition (NER) dataset for Sub-Saharan African languages: We create the largest human-annotated dataset called MasakhaNER 2.0, and analyze features that contribute to cross-lingual transfer, giving large gains for 0-shot learning.

8. *Buzaaba with many others*, MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition, EMNLP 2022

Deviation inequalities for stochastic approximation by averaging: We establish deviation inequalities for separately Lipschitz functions of Markov chains belonging to a certain class we define, and which includes models of stochastic approximation by averaging and non-averaging.

9. Fan, Alquier, Doukhan, Deviation inequalities for stochastic approximation by averaging,SPA 2022

Prioritization of minibatches: We give empirical support for the hypothesis that improving calibration can help in prioritizing minibatches during training.

10. *Tata, Gudur, Chennupati, Khan*, Can calibration improve sample prioritization, NeurIPS 2022 Workshop on Has It Trained Yet