Approximate Bayesian Inference Team
Mohammad Emtiyaz Khan
近似ベイズ推論チーム　カーン　エムティヤーズ

Center for Advanced Intelligence Project

RIKEN

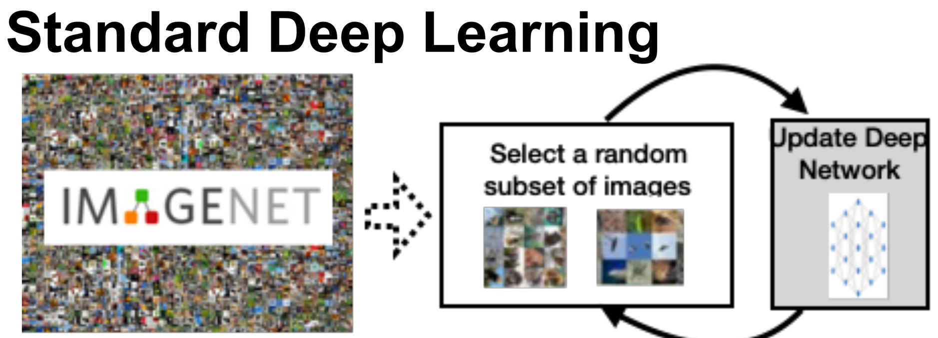
Overview and Goals

Goal: AI that can continue to learn and improve throughout their lives, just like humans and animals. Currently, deep learning (DL) requires a large amount of data which is costly and rigid (cannot quickly adapt). We aim to fix these issues with a new learning paradigm based on Bayesian principles.

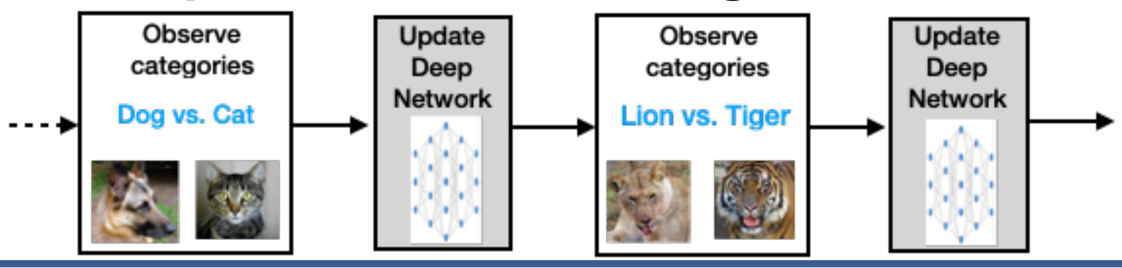
Summary of our research in the years 2020-2021:

A. Proposed Bayesian learning rule (BLR) yielding a wide-range of algorithms.
B. New BLR variants for DL, one of which won the NeurIPS-2021 Approximate Inference challenge.
C. Progress on adaptation and continual learning (FROMP, K-priors, Bayes-duality).
D. New theoretical results for online Bayes
E. Hyperparameter and architecture search using Bayesian methods.
F. A new paper on AI for social good in Nature communications.

Standard Deep Learning



Deep Continual Learning



Bayesian Learning Rule (BLR)

Problem: Is there a common principle behind "successful" algorithms (e.g., those in DL)?

$$\min_{\theta} \ \ell(\theta) \quad \text{vs} \quad \min_{q \in \mathcal{Q}} \ \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{H}(q)$$

Entropy

Generalized-Posterior approx.

Solution: we propose the Bayesian Learning Rule [1]

Natural and Expectation parameters of q

$$\lambda \leftarrow (1-\rho)\lambda - \rho \tilde{\nabla}_{\mu} \mathbb{E}_q[\ell(\theta)]$$

Old belief

Revise using new information through natural gradients

By choosing different approximations, we can derive a wide-variety of learning-algorithms. Better approximations lead to better algorithms.

Learning Algorithm	Posterior Approx.	Natural-Gradient Approx.	Sec.
Optimization Algorithms			
Gradient Descent	Gaussian (fixed cov.)	Delta method	1.3
Newton's method	Gaussian	—"—	1.3
Multimodal optimization (New)	Mixture of Gaussians	—"—	3.2
Deep-Learning Algorithms			
Stochastic Gradient Descent	Gaussian (fixed cov.)	Delta method, stochastic approx.	4.1
RMSprop/Adam	Gaussian (diagonal cov.)	Delta method, stochastic approx., Hessian approx., square-root scaling, slow-moving scale vectors	4.2
Dropout	Mixture of Gaussians	Delta method, stochastic approx., responsibility approx.	4.3
STE	Bernoulli	Delta method, stochastic approx.	4.5
Online Gauss-Newton (OGN) (New)	Gaussian (diagonal cov.)	Gauss-Newton Hessian approx. in Adam & no square-root scaling	4.4
Variational OGN (New)	—"—	Remove delta method from OGN	4.4
BayesBiNN (New)	Bernoulli	Remove delta method from STE	4.5
Approximate Bayesian Inference Algorithms			
Conjugate Bayes	Exp-family	Set learning rate $\rho_t = 1$	5.1
Laplace's method	Gaussian	Delta method	4.4
Expectation-Maximization	Exp-Family + Gaussian	Delta method for the parameters	5.2
Stochastic VI (SVI)	Exp-family (mean-field)	Stochastic approx., local $\rho_t = 1$	5.3
VMP	—"—	$\rho_t = 1$ for all nodes	5.3
Non-Conjugate VMP	—"—	—"—	5.3
Non-Conjugate VI (New)	Mixture of Exp-family	None	5.4

1. *Khan and Rue,* The Bayesian Learning Rule, *arXiv, 2021*

1st Place in NeurIPS 2021 Challenge

Problem: Approximate the expensive, exact Bayesian posterior (computed over several weeks on 512 TPUs) but don't exceed ~10x the cost of standard training.

Solution: A BLR variant, called iVON [2], uses mixture-of-Gaussian posterior approximation. Won first prize! Team consisted of Thomas Möllenhoff, Yuesong Shen, Gian Maria Marconi, Peter Nickl, Emtiyaz Khan.

VOGN

$$g \leftarrow \nabla \ell(\theta), \ \text{where} \ \theta \sim \mathcal{N}(m, (s+\gamma)^{-1})$$
$$s \leftarrow (1-\rho)s + \rho \Sigma_i g_i^2$$
$$m \leftarrow m - \alpha(s+\gamma)^{-1} g$$

iVON[1]

$$g \leftarrow \nabla \ell(\theta), \ \text{where} \ \theta \sim \mathcal{N}(m, s^{-1})$$
$$g_s \leftarrow [s(\theta - m)]g + \gamma - s$$
$$m \leftarrow m - \alpha s^{-1} g$$
$$s \leftarrow s + (1-\rho)g_s + 0.5(1-\rho)^2 s^{-1} g_s^2$$

Team	Method	Rank (Light Track)	Rank (Ext. Track)	CIFAR Agree	CIFAR TVD	Med-MNIST Agree	Med-MNIST TVD	UCI-Gap W2
RIKEN Team ABI	Bayesian Learning Rule	**1**	**1.67**	**0.787**	**0.197**	0.884	0.0994	**-0.094**
École Polytechnique	MultiSWAG	2.5	2.5	0.777	0.218	**0.8905**	**0.0983**	-0.166
University of Liège	Seq Anchored Ensembles	2.5	3	0.773	0.210	0.8745	0.1066	-0.115

More BLR variants:

- iVON [2] is proposed to ensure the steps of BLR always lead to positive covariances.
- New generalizations in [3] for "structured" covariances allow low-rank and sparse structures (eg, recovering LBFGS/DFP style updates). This work uses Lie-Group structures.
- BayesBiNN [4] is a BLR variant for Binary Neural Networks which recovers the STE algorithm

2. *Lin, Schmidt, Khan,* Handling the Positive-Definite Constraint in the Batesian Learning Rule, *ICML 2020*
3. *Lin, Nielsen, Khan, Schmidt,* Tractable structured natural-gradient descent using local parameterizations, *ICML 2021*
4. *Meng, Bachman, Khan,* Training Binary Neural Networks using the Bayesian Learning Rule, *ICML 2020*

Continual Learning and Adaptation

Problem: Reduce catastrophic forgetting of the past. A popular method is to use quadratic weight regularizers.

$$q_{new}(\theta) = \min_{q \in \mathcal{Q}} \mathbb{E}_{q(\theta)}[\ell_{new}(\theta)] - \mathcal{H}(q) - \mathbb{E}_{q(\theta)}[\log q_{old}(\theta)]$$

New data

Weight-regularizer

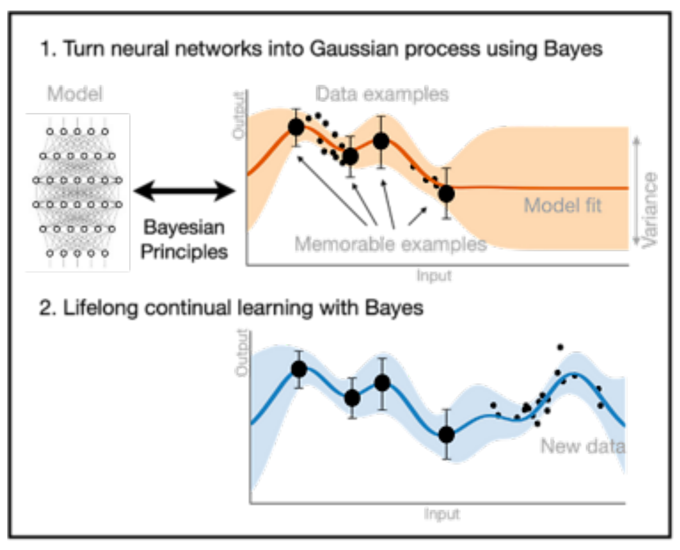
Solution: We show that functional regularization of "memorable past" (FROMP) [5] gives better results

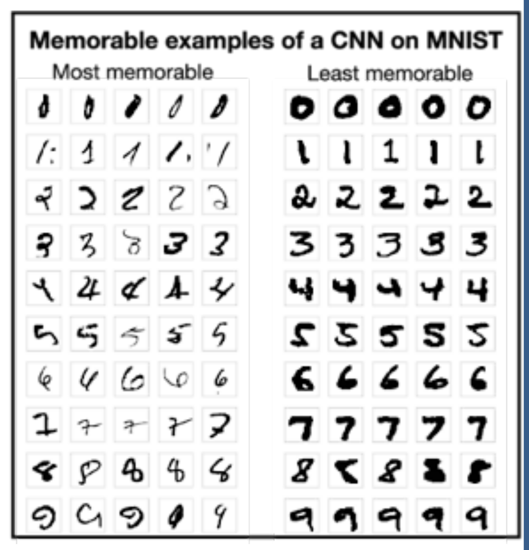
$$\approx \mathbb{E}_{\tilde{q}_\theta(\mathbf{f})}[\log \tilde{q}_{\theta_{old}}(\mathbf{f})]$$

$$[\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]^\top K_{old}^{-1} [\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]$$

Kernels weighs examples according to their memorability

Forces network-outputs to be similar





In [6], we quantify "forgetting" in terms of past memory represented via principal components analysis.
In [7], we present a generalization called K-priors to unify such adaptation methods. We show that these methods faithfully reconstruct the gradient of the past.

Weight-space　　Function-space

$$\mathcal{K}(\theta) = \tau \mathbb{D}_\theta(\theta \| \theta_{\text{old}}) + \mathbb{D}_f(\mathbf{f}(\theta) \| \mathbf{f}(\theta_{\text{old}}))$$

5. *Pan, Swaroop, Immer, Eschenhagen, Turner, Khan,* Continual Deep Learning by Functional Regularisation of Memorable Past, *NeurIPS 2020*
6. *Doan, Abbana Bennani, Mazoure, Rabusseau, Alquier,* A Theoretical Analysis of Catastrophic Forgetting through the NTK Overlap Matrix, AIStats 2021
7. *Khan & Swaroop,* Knowledge-Adaptation Priors, *NeurIPS 2021*

Theoretical Results for Online Bayes

$$\rho^t = \underset{\rho \in \mathcal{P}(\Theta)}{\text{argmin}} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim \rho}[\ell_s(\theta)] + \frac{\text{KL}(\rho \| \pi)}{\eta} \right\}$$

Problem: Theoretical analysis for online Bayesian learning hold under restrictive conditions.

Solution: We propose to relax these conditions, by using a generalize online Bayesian methods where arbitrary divergences can be used (instead of KL) [8]

$$\rho^t = \underset{\rho \in \mathcal{P}(\Theta)}{\text{argmin}} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim \rho}[\ell_s(\theta)] + \frac{D_\phi(\rho \| \pi)}{\eta} \right\}$$

We derive an explicit formula for the updates which we call generalized Bayes rule.

$$\rho^t(d\theta) = \nabla \tilde{\phi}^* \left(\lambda_t - \eta \sum_{s=1}^{t-1} \ell_s(\theta) \right) \pi(d\theta)$$

We prove a regret bound that holds for **below the usual bounded setting** (less restrictive).

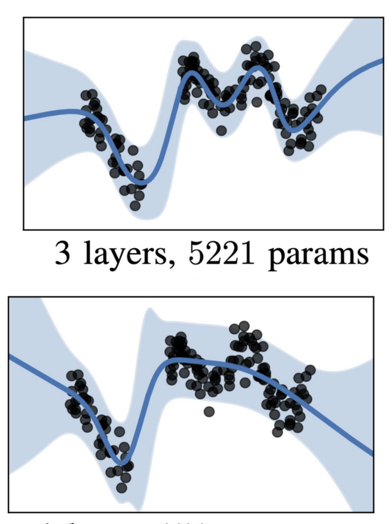
8. *Alquier,* Non-exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions, *ICML 2021*

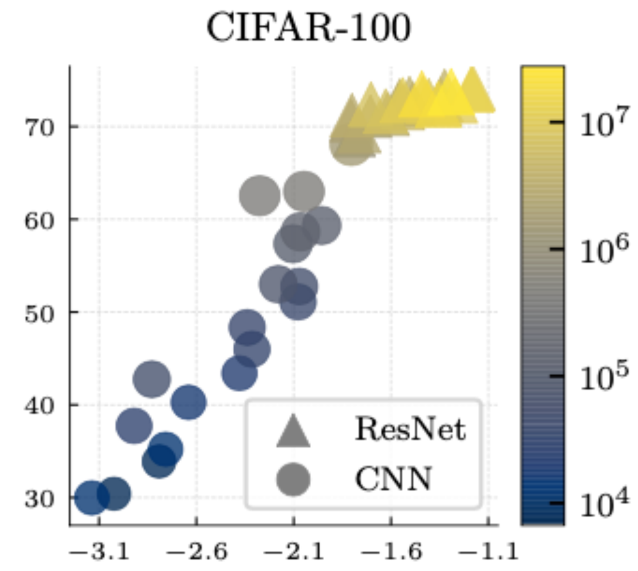
Architecture Selection for Deep Networks

Problem: Existing methods require validation data to select architecture and hyperparameters.
Solution: A method based on marginal likelihood using only training data. Uses Laplace approximation[9.10] with scalable Hessian approx (eg, KFAC).

$$\log p(\mathcal{D} \mid \mathcal{M}) \approx \underbrace{\log p(\mathcal{D} \mid \theta_*, \mathcal{M})}_{\text{Training data fit}} + \underbrace{\log p(\theta_* \mid \mathcal{M}) - \frac{1}{2}\log\left|\frac{1}{2\pi}\mathbf{H}_{\theta_*}\right|}_{\text{complexity penalty}}$$

Larger models, which give better test error, also generally have higher marginal likelihoods.





9. *Immer, Bauer, Fortuin, Ratsch, Khan,* Scalable marginal likelihood for model selection in deep learning, *ICML 2021*
10. *Immer, Korpeza, Bauer,* Improving predictions of Bayesian neural networks via local linearization, Aistats 2021

A Summary of Other Works

Gaussian Process: Using BLR, we derive a fast algorithm for state-space GP [11]. We also show that a dual parameterization useful for sparse GPs [12]. We derive a sparse representation using subset of data [13]

11. *Chang, Adam, Khan, Solin,* Dual Parameterization of Sparse Variational Gaussian Processes, *ICML 2021*
12. *Chang, Wilkinson, Khan, Solin,* Fast Variational Learning in State-Space Gaussian Process Models, MLSP, 2020
13. *Jain, PK, Khan,* Subset-of-Data Variational Inference for Deep Gaussian-Process Regression, UAI 2021

Reinforcement Learning: We propose a replacement of "target networks" by functional regularization [14]. In [15], we propose imitation learning for diverse kinds of feedback, appropriately re-weighting them.

14. *Piche, Thomas, Marino, Marconi, Pal, Khan.,* Beyond Target Networks: Improving Deep Q-learning with Functional Regularization, *arXiv 2021*
15. *Tangkaratt, Han, Khan, Sugiyama,* VILD: Variational Imitation Learning with Diverse-quality Demonstrations, ICML 2020

AI for Social Good: We outline a few guidelines on how to align AI systems for social good applications [16].

16. *Tomasev et al.,* AI for Social Good: Unlocking the Opportunity for Positive Impact, *Nature communications 2020*