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

Overview and Goals

Goal: To gain 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 (PROMP, 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.

Bayesian Learning Rule (BLR)
Problem: Is there a common principle behind “successful” algorithms (e.g., those in DL)?
\[ \min \quad \ell(\theta) \quad \text{vs} \quad \min E_{q(\theta)}[\ell(\theta)] - H(q) \quad \text{Entropy} \]
Generalized Posterior approx.
Solution: We propose the Bayesian Learning Rule [1]
\[ \lambda \approx (1 - \rho)L - \lambda \nabla \mathbb{E}_{q(\theta)}[\ell(\theta)] \quad \text{Old belief} \]
Revised using new information through natural gradients

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, Yueqiong Shen, Gian Maria Marconi, Peter Nickl, Emtiyaz Khan.

OGN
\[ \text{OGN, Online Gauss-Newton} \]

(i) iVON
\[ \text{iVON} \]

Bayesian Approximations of the Posterior and Natural-gradient. New algorithms are marked with “(New).”
Abbreviations: cov.

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 (e.g., recovering LBFGS/DPF style updates). This work uses Lie-Group structures.
- BayesBiNN [4] is a BLR variant for Binary Neural Networks which recovers the STE algorithm.

Conclusion of Large Experiments

2. Lin, Schmidt, Khan, Handling the Positive-Definite Constraint in the Bayesian Learning Rule, ICML 2020
3. Lin, Nielsen, Khan, Schmidt, Tractable structured natural-gradient descent using local parameterizations, ICCV 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.

Solution: We show that functional regularization of “memorable past” (PROMP) [5] gives better results.

Kernels weighs examples according to their memorability

In [6], we quantitatively “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.

In [8], we propose imitation learning for diverse kinds of data.

Weight-space

Function-space

A Summary of Other Works

General 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

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.

15. Tangkarat, Han, Sujiyono, 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].


Theoretical Results for Online Bayes

Problem: Theoretical analysis for online Bayesian learning hold under restrictive conditions.
Solution: We propose to relax these conditions, by using a generalized online Bayesian methods where arbitrary divergences can be used (instead of KL) [8].

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

We prove a regret bound that holds for below the usual bounded setting (less restrictive).
B. 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 (e.g., KFAC).

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

1. Jang, Poole, Barlow, CTRF-100
2. Wang et al., 32 layers, 5221 parameters
3. Pratapa, 15 layers, 151 parameters
4. Tramel et al., 1 layer, 151 parameters

A Summary of Other Works

Bayesian Approximations

Learning Algorithm

Posterior Approximation

Natural-Gradients Approximation

Sec. 2

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

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

Overview and Goals

Goal: To gain 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 (PROMP, 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.

Bayesian Learning Rule (BLR)
Problem: Is there a common principle behind “successful” algorithms (e.g., those in DL)?

\[ \min \quad \ell(\theta) \quad \text{vs} \quad \min E_{q(\theta)}[\ell(\theta)] - H(q) \quad \text{Entropy} \]
Generalized Posterior approx.
Solution: We propose the Bayesian Learning Rule [1]

\[ \lambda \approx (1 - \rho)L - \lambda \nabla \mathbb{E}_{q(\theta)}[\ell(\theta)] \quad \text{Old belief} \]
Revised using new information through natural gradients

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, Yueqiong Shen, Gian Maria Marconi, Peter Nickl, Emtiyaz Khan.

OGN
\[ \text{OGN, Online Gauss-Newton} \]

(i) iVON
\[ \text{iVON} \]

Bayesian Approximations of the Posterior and Natural-gradient. New algorithms are marked with “(New).”
Abbreviations: cov.

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 (e.g., recovering LBFGS/DPF style updates). This work uses Lie-Group structures.
- BayesBiNN [4] is a BLR variant for Binary Neural Networks which recovers the STE algorithm.

Conclusion of Large Experiments

2. Lin, Schmidt, Khan, Handling the Positive-Definite Constraint in the Bayesian Learning Rule, ICML 2020
3. Lin, Nielsen, Khan, Schmidt, Tractable structured natural-gradient descent using local parameterizations, ICCV 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.

Solution: We show that functional regularization of “memorable past” (PROMP) [5] gives better results.

Kernels weighs examples according to their memorability

In [6], we quantitatively “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.

In [8], we propose imitation learning for diverse kinds of data.

Weight-space

Function-space

A Summary of Other Works

General 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

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.

15. Tangkarat, Han, Sujiyono, 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].