FY2023/2023年度

Approximate Bayesian Inference Team

Mohammad Emtiyaz Khan
近似ベイズ推論チーム

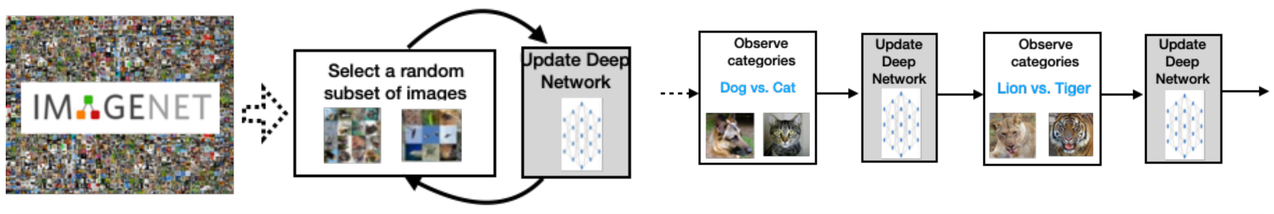
RIKEN Center for Advanced Intelligence Project

Overview and Goals

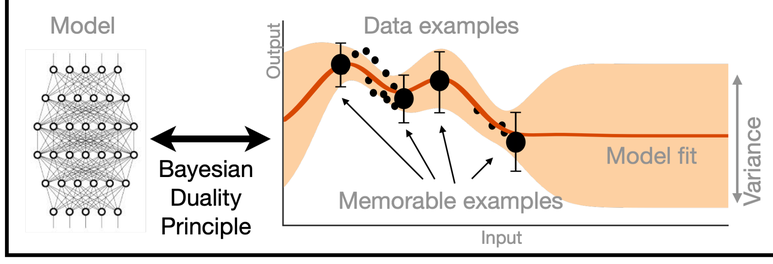
Goal: to design AI systems that can learn and improve continually throughout their lives, just like humans and animals. Currently, deep learning (DL) requires a large amount of data which is costly and rigid, leading to a system that is unable to quickly adapt. We aim to fix this with a new principle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

Summary for the year 2023 (paper number shown in green boxes)

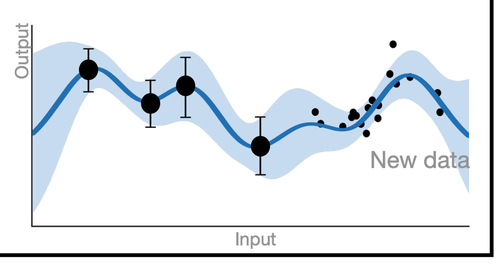
1. We propose memory-perturbation to understand sensitivity of algorithms to data.
2. We improve continual learning on ImageNet by adding memory (<10% of data).
3. We use memory to improve sequential learning in sparse Gaussian Process.
4. We give a simple recipe to derive variational Bayes updates.
5. We improve sample efficiency of RL by using a functional regularization.
6. We propose a structured inference method for Neural Processes.
7. We discuss challenges in annotating African languages.
8. We infer properties of Markov chains with sparse connection graphs.
9. We recover known sampling algorithms by means of information geometry.
10. We propose a reduction method for identity testing of Markov chains.



Bayes-duality relates *models parameters (left)* to the *data examples* (small dots at the right). The principle enables us to identify a few *memorable examples* (big black circle).



The memorable examples can be reused later during training with the new data. This avoids forgetting of the past.



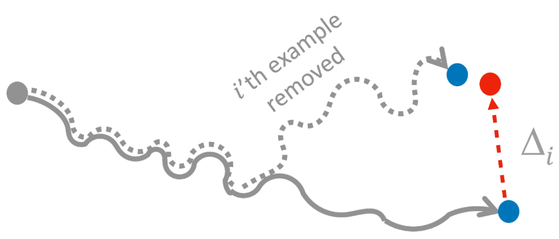
Memory-Perturbation Equation

Problem: Understanding model's sensitivity to training data is crucial, but can be challenging and costly, especially during training (including existing work on influence measures, for example, Koh and Liang (2017)).

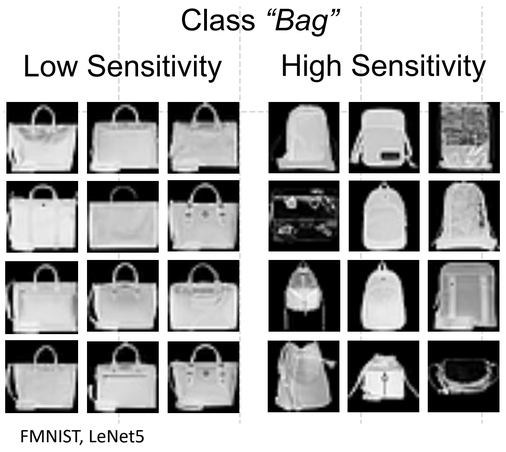
Contribution: We use Bayesian principles to derive the Memory Perturbation Equation (MPE) which unifies and generalizes sensitivity measures to a wide-variety of algorithms and unravels useful properties regarding their sensitivities.

Deviation in posterior parameters (Δ) = Natural gradient of the example removed

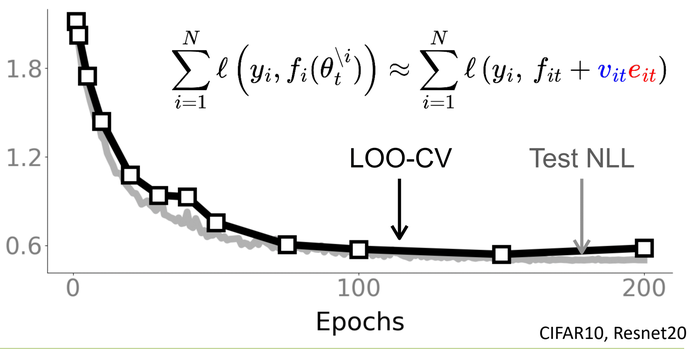
$\Delta \propto$ PredError x PredVariance



Algorithm	Sensitivity
Newton's method	$\mathbf{H}_{t-1}^{-1}\nabla\ell_i(\boldsymbol{\theta}_t)$
Online Newton (ON) [28]	$\mathbf{S}_t^{-1}\nabla\ell_i(\boldsymbol{\theta}_t)$
ON (diagonal+minibatch) [28]	$\mathbf{s}_t^{-1}\cdot\nabla\ell_i(\boldsymbol{\theta}_t)$
iBLR (diagonal+minibatch) [35]	$\mathbf{s}_t^{-1}\cdot\nabla\ell_i(\boldsymbol{\theta}_t)$
RMSprop/Adam [30]	$\mathbf{s}_t^{-\frac{1}{2}}\cdot\nabla\ell_i(\boldsymbol{\theta}_t)$
SGD	$\nabla\ell_i(\boldsymbol{\theta}_t)$



Sensitivities can be used to predict generalization error by using training data alone

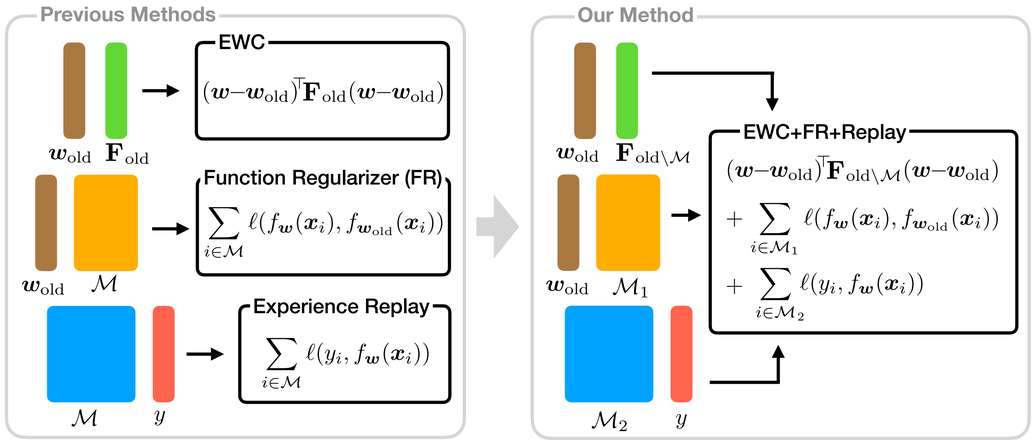


1. Nickl, Xu, Tailor*, Moellenhoff, Khan,* The Memory Perturbation Equation: Understanding Model's Sensitivity to Data*, NeurIPS 2023.*

Improving Continual Learning by Gradient Reconstructions

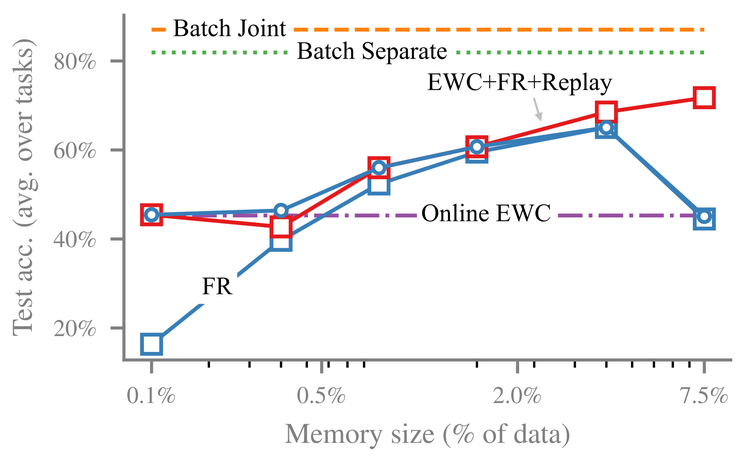
Problem: There are many well-known continual learning algorithms that work well, but is there a principled way to combine them?

Contribution: Following Khan & Swaroop (2021), we design a regularizer that faithfully reconstruct gradients of past data by combining Memory Replay, Elastic-Weight Consolidation, and Functional Regularization.



Result: We often obtain > 80% of the batch performance by using a memory of < 10% of the past data.

On ImageNet-1000, by using 7.5% past data, we achieve around 76% accuracy, 11% less than batch accuracy 87%.



2. Daxberger, Swaroop, Osawa, Yokota, Turner, Hernandez-Lobato, Khan, Improving Continual Learning by Accurate Gradient Reconstruction of the Past, TMLR *2023*.

Other Works

Sequential learning with Gaussian processes: We present a method to avoid accumulation of errors during sequential learning by using the recently proposed dual sparse variational GP. The method enables accurate inference for generic likelihoods and improves learning by actively building and updating a memory of past data.

3. Chang, Verma, John, Solin, Khan, Memory-based dual GP for sequential learning. ICML 2023

A three-step Recipe for Variational Bayes: We give a simple recipe to identify the posterior form by explicitly looking for linearity with respect to expectations of well-known distributions. This enables us to directly write the update by simply "reading-off" the terms in front of those expectations.

4. Khan ME. Variational Bayes Made Easy, AABI 2023

Replacing Target Networks by Functional Regularization: We propose an explicit Functional Regularization to replace target networks and empirically demonstrate it leads to better sample efficiency and performance improvements.

5. Piche, Thomas, Marino, Pardinas, Marconi, Pal, Khan, Bridging the Gap Between Target Networks and Functional Regularization, TLMR 2023

Structured inference networks for Neural Processes: We enrich the latent variable of Neural Processes with structured priors (e.g. with multiple modes, heavy-tails, etc.) and provide a framework that directly translates such distributional assumptions into an aggregation strategy for the context set.

6. Tailor, Khan, Nalisnick, Exploiting inferential structure in neural processes, UAI 2023

Tagging for Diverse African Languages: We discuss the challenges in annotating POS for African languages using the Universal Dependencies guidelines and conduct extensive part-of-speech baseline experiments using various models.

7. Dione, Adelani, Nabende, Alabi, Sindane, Buzaaba, Muhammad, Emezue, Ogayo, Aremu, Gitau, MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African Languages, ACL 2023

Estimation of Mixing Properties of Markov Chains: We derive confidence intervals to learn transition matrices of Markov chains, yielding improved estimators for some of its mixing properties. Notably, our analysis goes beyond the worst-case scenario by leveraging the sparsity of the connection graph.

8. Wolfer, Empirical and Instance-Dependent Estimation of Markov Chain and Mixing Time, Scandinavian Journal of Statistics 2023

Reversiblizations of Markov Chains: We propose systematic projection schemes to *reversiblize* Markov chains. Various divergences lead to many popular Markov Chain Monte Carlo algorithms, including a recently popularized Barker dynamics algorithm.

9. Choi, Wolfer, Systematic Approaches to Generate Reversiblizations of Markov Chains, IEEE Transactions on Information Theory 2023

Identity Testing of Markov Chains: We show that the problem of identity testing for reversible Markov chains can be reduced to the same for symmetric Markov chains. Our streamlined approach not only recovers the state-of-the-art sample complexity for the problem but also extends to related problems.

10. Wolfer, Watanabe, Geometric Reduction for Identity Testing of Reversible Markov Chains, Geometric Science of Information 2023