Presentation at the Deep Learning Workshop (Tokyo) Mar 19 2024

The Bayesian Learning Rule

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo

http://emtiyaz.github.io

Summary of recent research at https://emtiyaz.github.io/papers/symposium_2023.pdf

Slides available at https://emtiyaz.github.io/

How to make AI that can adapt quickly?

And continue to do so throughout its life

Continual Lifelong Adaptation in Machine Learning

- Even a small change may need full retraining
  - Huge amount of resources only few can afford (costly & unsustainable) [1,2, 3]
  - Difficult to apply in “dynamic” settings (robotics, epidemiology, climate science etc)
- Fix and improve deep learning

1. Diethe et al. Continual learning in practice, arXiv, 2019.
2. Paleyes et al. Challenges in deploying machine learning: a survey of case studies, arXiv, 2021.
3. https://www.youtube.com/watch?v=hx7BXih7zx8&t=897s

Bayesian Learning Rule [1]

- Bridge DL & Bayesian learning [2-5]
  - SOTA on GPT-2 and ImageNet [5]
- Improve other aspects of DL [5-7]
  - Calibration, memory, lifelong learning
- Towards human-like quick adaptation

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Khan, et al. Fast and scalable Bayesian deep learning by weight-perturbation in Adam, ICML (2018).
3. Osawa et al. Practical Deep Learning with Bayesian Principles, NeurIPS (2019).
4. Lin et al. Handling the positive-definite constraints in the BLR, ICML (2020).
5. Shen et al. Variational Learning is Effective for Large Deep Networks, Under review.
6. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
7. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)

The Bayesian Learning Rule

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

Entropy

Posterior approximation (expo-family)

Bayesian Learning Rule [1,2] (natural-gradient descent)

Natural and Expectation parameters of q

$$\lambda \leftarrow \lambda - \rho F(\lambda)^{-1} \nabla_\lambda \left\{ \mathbb{E}_q[\ell(\theta)] - \mathcal{H}(q) \right\}$$

Many well-known algorithms are special-instances obtained by choosing approximation to q and natural-gradients.

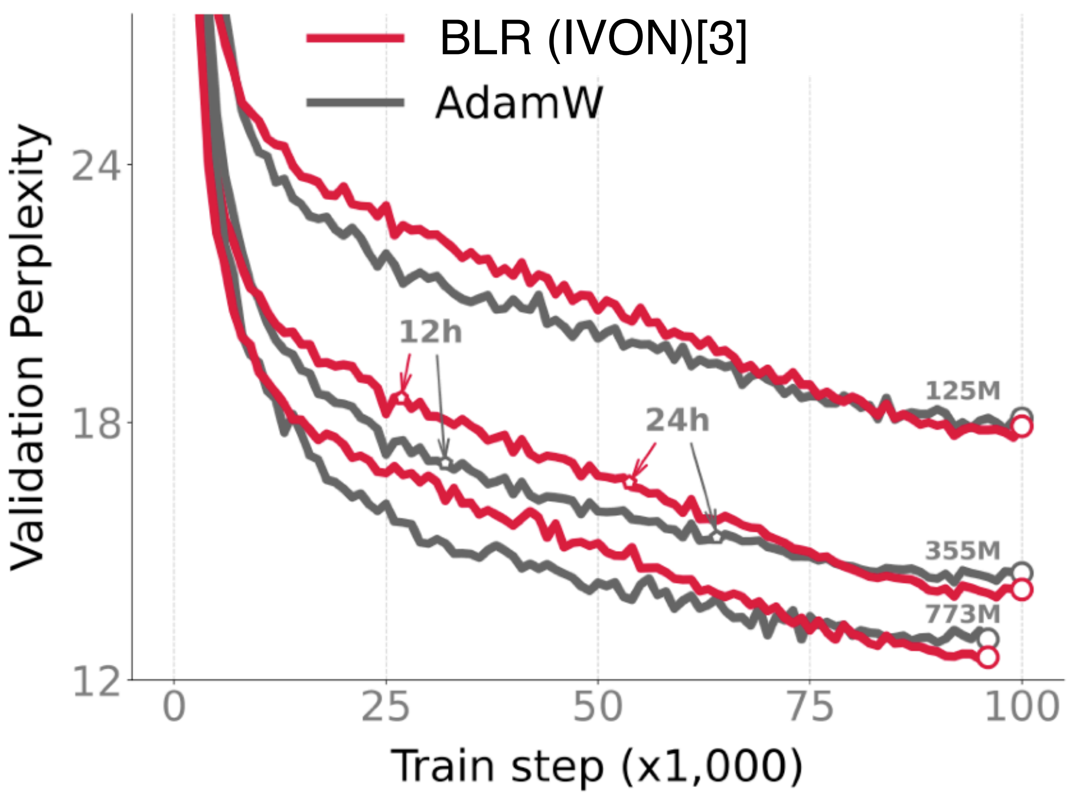
1. Khan and Rue, The Bayesian Learning Rule, JMLR, 2023
2. Khan and Lin. "Conjugate-computation variational inference…." AIstats, 2017

List of algorithms as a special case of the BLR

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

GPT-2 with Bayesian Learning Rule [1]

Better performance & uncertainty at the same cost [2]



Trained on OpenWebText data (49.2B tokens).

On 773M, we get a gain of 0.5 in perplexity.

On 355M, we get a gain of 0.4 in perplexity.

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Shen et al. “Variational Learning is Effective for Large Deep Networks.” Under review (2024)

BLR for large deep networks

RMSprop/Adam

$\hat{g} \leftarrow \hat{\nabla}\ell(\theta)$

$\hat{h} \leftarrow \hat{g}^2$

$h \leftarrow (1-\rho)h + \rho\hat{h}$

$\theta \leftarrow \theta - \alpha(\hat{g} + \delta m)/(\sqrt{h} + \delta)$

BLR [1] variant called IVON [5]
(Improved Variational Online Newton)

$\hat{g} \leftarrow \hat{\nabla}\ell(\theta)$ where $\theta \sim \mathcal{N}(m, \sigma^2)$

$\hat{h} \leftarrow \hat{g} \cdot (\theta - m)/\sigma^2$

$h \leftarrow (1-\rho)h + \rho\hat{h} \;\; +\rho^2(h-\hat{h})^2/(2(h+\delta))$

$m \leftarrow m - \alpha(\hat{g} + \delta m)/(h + \delta)$

$\sigma^2 \leftarrow 1/(N(h+\delta))$

Only tune initial value of h (a scalar)
Check out the blog: https://team-approx-bayes.github.io/blog/ivon/

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
3. Osawa et al. “Practical Deep Learning with Bayesian Principles.” NeurIPS (2019).
4. Lin et al. “Handling the positive-definite constraints in the BLR.” ICML (2020).
5. Shen et al. “Variational Learning is Effective for Large Deep Networks.” Under review (2024)

Drop-in replacement of Adam

https://github.com/team-approx-bayes/ivon

```
import torch
+import ivon

train_loader = torch.utils.data.DataLoader(train_dataset)
test_loader = torch.utils.data.DataLoader(test_dataset)
model = MLP()

-optimizer = torch.optim.Adam(model.parameters())
+optimizer = ivon.IVON(model.parameters())

for X, y in train_loader:

+    for _ in range(train_samples):
+       with optimizer.sampled_params(train=True)
            optimizer.zero_grad()
            logit = model(X)
            loss = torch.nn.CrossEntropyLoss(logit, y)
            loss.backward()

    optimizer.step()
```

IVON [3] got 1st prize in NeurIPS 2021 Approximate Inference Challenge

Watch Thomas Moellenhoff's talk at https://www.youtube.com/watch?v=LQInIN5EU7E.

Mixture-of-Gaussian Posteriors with an Improved Bayesian Learning Rule

Thomas Möllenhoff[1], Yuesong Shen[2], Gian Maria Marconi[1]
Peter Nickl[1], Mohammad Emtiyaz Khan[1]

1 Approximate Bayesian Inference Team
RIKEN Center for AI Project, Tokyo, Japan

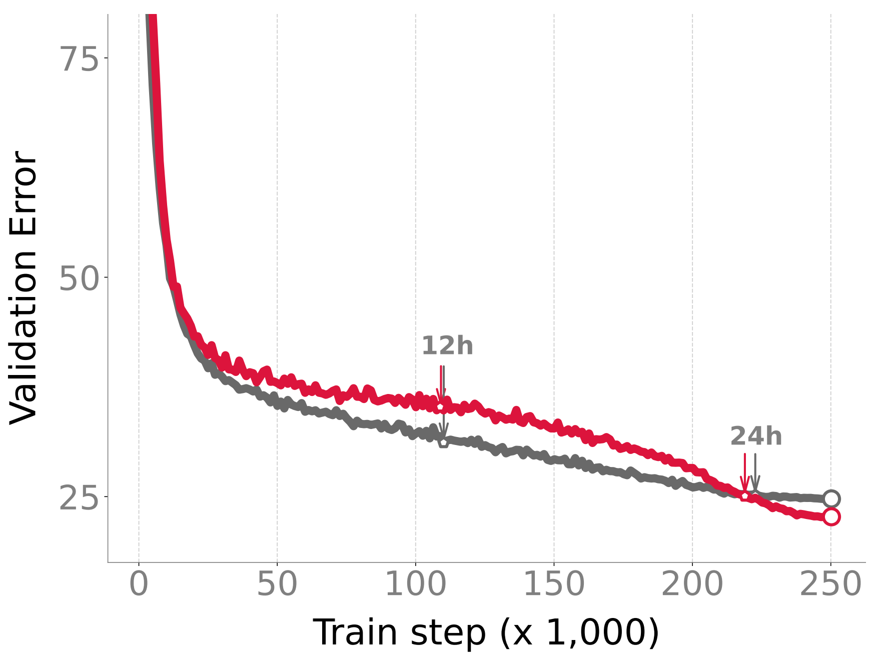
2 Computer Vision Group
Technical University of Munich, Germany

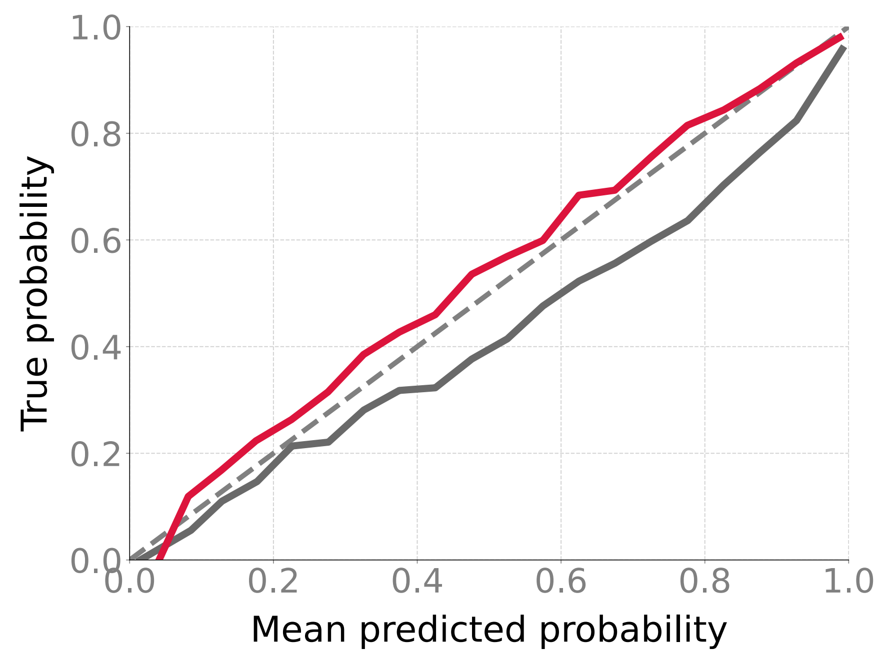
Dec 14th, 2021 — NeurIPS Workshop on Bayesian Deep Learning

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. "Practical Deep Learning with Bayesian Principles." NeurIPS (2019).
3. Lin et al. "Handling the positive-definite constraints in the BLR." ICML (2020).

ImageNet on ResNet-50 (25.6M)

2% better accuracy over AdamW and 1% over SGD. Better calibration (ECE of 0.022 vs 0.066)



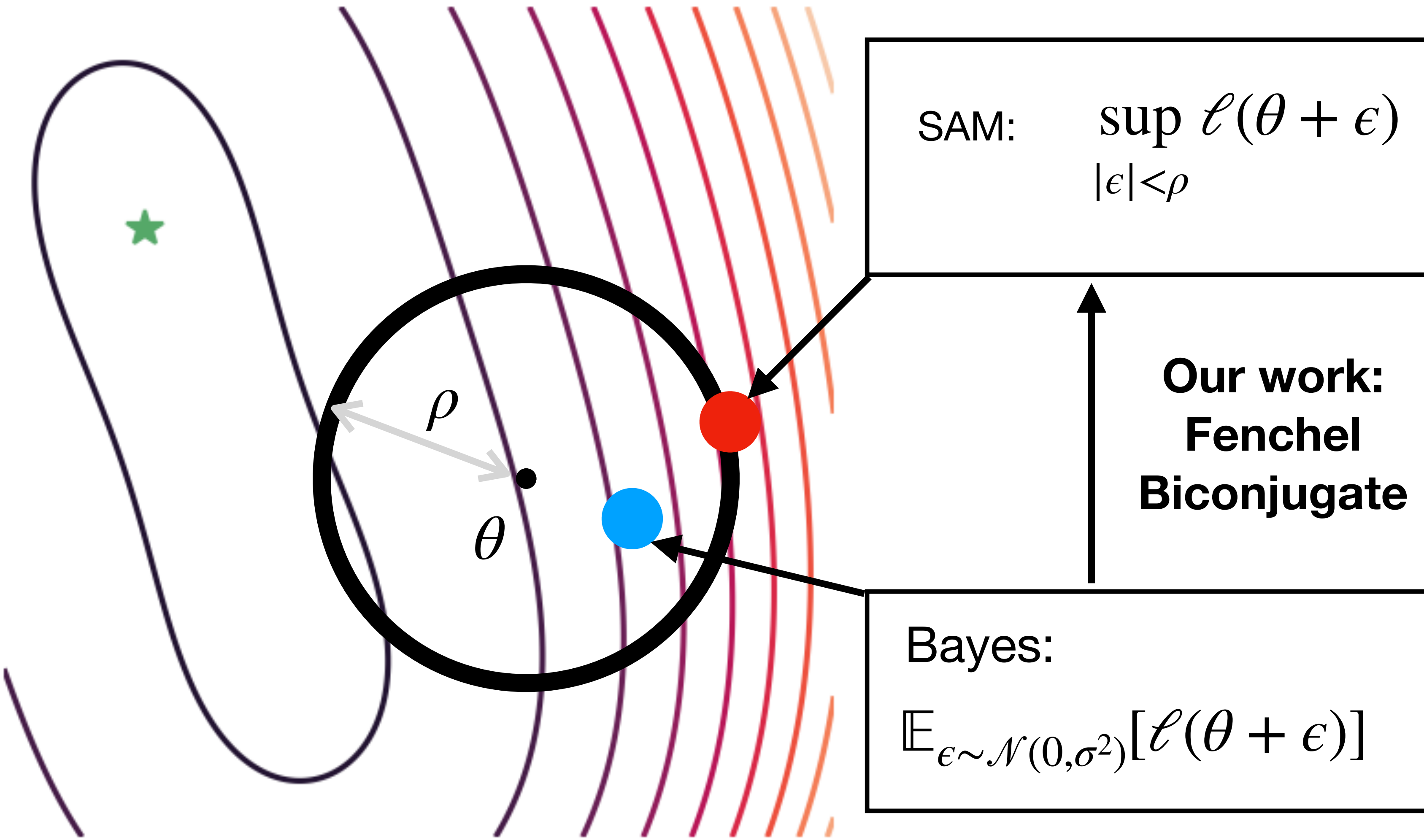


ImageNet on ResNet-50 (25.6M)

No severe overfitting like AdamW while improving accuracy over SGD consistently & better uncertainty

Dataset & Model	Epochs	Method	Top-1 Acc. ↑	Top-5 Acc. ↑	NLL ↓	ECE ↓	Brier ↓
ImageNet-1k **ResNet-50** (25.6M params)	100	AdamW	74.56$_{\pm 0.24}$	92.05$_{\pm 0.17}$	1.018$_{\pm 0.012}$	0.043$_{\pm 0.001}$	0.352$_{\pm 0.003}$
		SGD	**76.18**$_{\pm 0.09}$	**92.94**$_{\pm 0.05}$	**0.928**$_{\pm 0.003}$	0.019$_{\pm 0.001}$	**0.330**$_{\pm 0.001}$
		IVON@mean	**76.14**$_{\pm 0.11}$	92.83$_{\pm 0.04}$	0.934$_{\pm 0.002}$	0.025$_{\pm 0.001}$	**0.330**$_{\pm 0.001}$
		IVON	**76.24**$_{\pm 0.09}$	**92.90**$_{\pm 0.04}$	**0.925**$_{\pm 0.002}$	**0.015**$_{\pm 0.001}$	**0.330**$_{\pm 0.001}$
	200	AdamW	+2% 75.16$_{\pm 0.14}$	92.37$_{\pm 0.03}$	1.018$_{\pm 0.003}$	0.066$_{\pm 0.002}$	0.349$_{\pm 0.002}$
		SGD	+1% 76.63$_{\pm 0.45}$	93.21$_{\pm 0.25}$	0.917$_{\pm 0.026}$	0.038$_{\pm 0.009}$	0.326$_{\pm 0.006}$
		IVON@mean	77.30$_{\pm 0.08}$	93.58$_{\pm 0.05}$	0.884$_{\pm 0.002}$	0.035$_{\pm 0.002}$	**0.316**$_{\pm 0.001}$
		IVON	**77.46**$_{\pm 0.07}$	**93.68**$_{\pm 0.04}$	**0.869**$_{\pm 0.002}$	**0.022**$_{\pm 0.002}$	**0.315**$_{\pm 0.001}$
TinyImageNet **ResNet-18** (11M params, wide)	200	AdamW	+15% 47.33$_{\pm 0.90}$	71.54$_{\pm 0.95}$	6.823$_{\pm 0.235}$	0.421$_{\pm 0.008}$	0.913$_{\pm 0.018}$
		SGD	+1% 61.39$_{\pm 0.18}$	82.30$_{\pm 0.22}$	1.811$_{\pm 0.010}$	0.138$_{\pm 0.002}$	0.536$_{\pm 0.002}$
		IVON@mean	**62.41**$_{\pm 0.15}$	**83.77**$_{\pm 0.18}$	1.776$_{\pm 0.018}$	0.150$_{\pm 0.005}$	0.532$_{\pm 0.002}$
		IVON	**62.68**$_{\pm 0.16}$	**84.12**$_{\pm 0.24}$	**1.528**$_{\pm 0.010}$	**0.019**$_{\pm 0.004}$	**0.491**$_{\pm 0.001}$
TinyImageNet **PreResNet-110** (4M params, deep)	200	AdamW	+10% 50.65$_{\pm 0.0*}$	74.94$_{\pm 0.0*}$	4.487$_{\pm 0.0*}$	0.357$_{\pm 0.0*}$	0.812$_{\pm 0.0*}$
		AdaHessian	55.03$_{\pm 0.53}$	78.49$_{\pm 0.34}$	2.971$_{\pm 0.064}$	0.272$_{\pm 0.005}$	0.690$_{\pm 0.008}$
		SGD	+2% 59.39$_{\pm 0.50}$	81.34$_{\pm 0.30}$	2.040$_{\pm 0.040}$	0.176$_{\pm 0.006}$	0.577$_{\pm 0.007}$
		IVON @mean	**60.85**$_{\pm 0.39}$	**83.89**$_{\pm 0.14}$	1.584$_{\pm 0.009}$	0.053$_{\pm 0.002}$	**0.514**$_{\pm 0.003}$
		IVON	**61.25**$_{\pm 0.48}$	**84.13**$_{\pm 0.17}$	**1.550**$_{\pm 0.009}$	**0.049**$_{\pm 0.002}$	**0.511**$_{\pm 0.003}$
CIFAR-100 **ResNet-18** (11M params, wide)	200	AdamW	+11% 64.12$_{\pm 0.43}$	86.85$_{\pm 0.51}$	3.357$_{\pm 0.071}$	0.278$_{\pm 0.005}$	0.615$_{\pm 0.008}$
		SGD	+.7% 74.46$_{\pm 0.17}$	92.66$_{\pm 0.06}$	1.083$_{\pm 0.007}$	0.113$_{\pm 0.001}$	0.376$_{\pm 0.001}$
		IVON@mean	74.51$_{\pm 0.24}$	92.74$_{\pm 0.19}$	1.284$_{\pm 0.013}$	0.152$_{\pm 0.003}$	0.399$_{\pm 0.002}$
		IVON	**75.14**$_{\pm 0.34}$	**93.30**$_{\pm 0.19}$	**0.912**$_{\pm 0.009}$	**0.021**$_{\pm 0.003}$	**0.344**$_{\pm 0.003}$

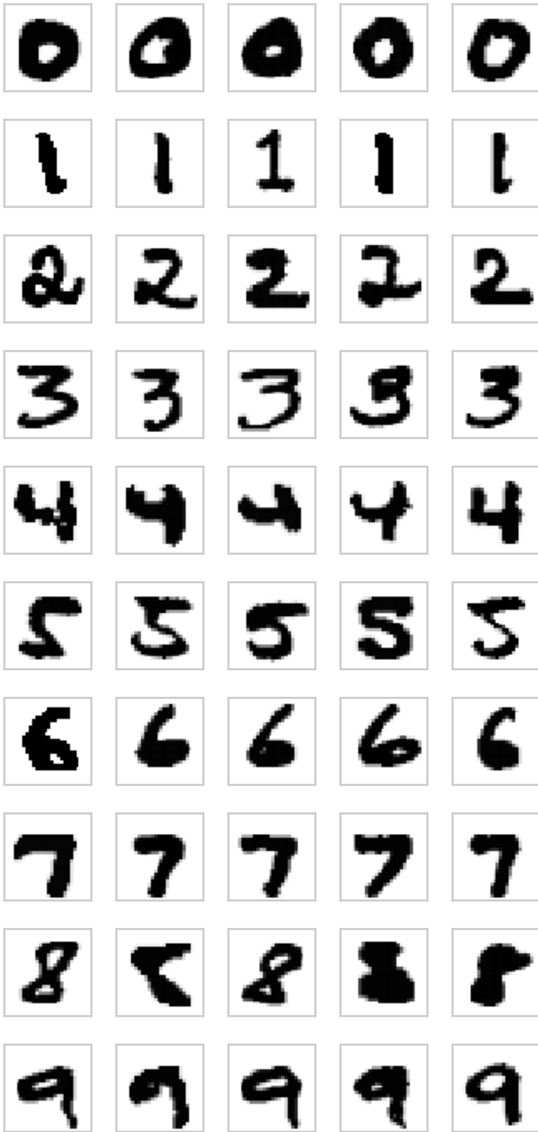
Sharpness-Aware Minimization (SAM) from BLR



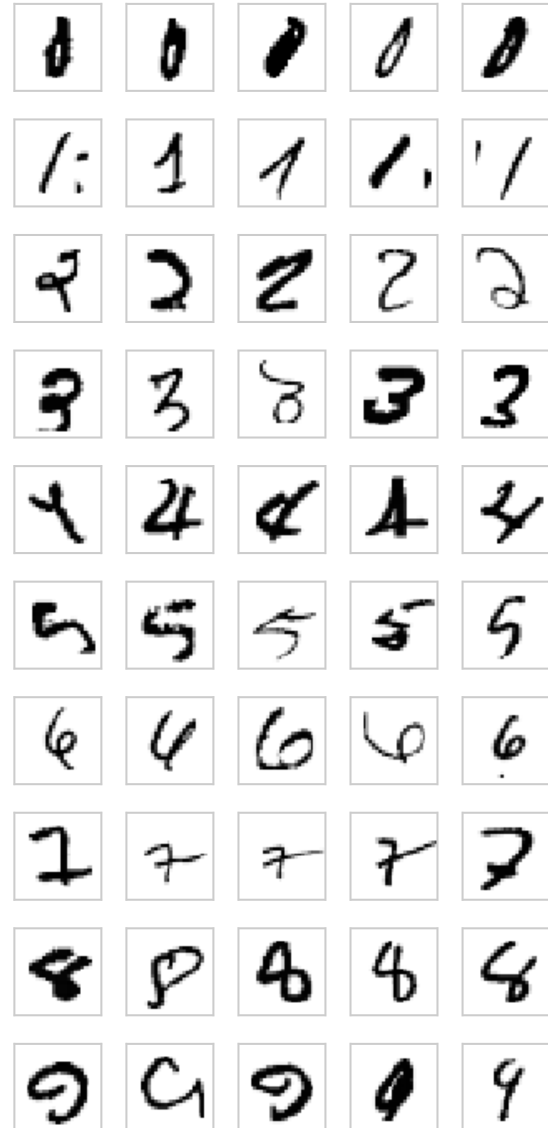
1. Foret et al. Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR, 2021
2. Moellenhoff and Khan, SAM as an Optimal Relaxation of Bayes, Under review, 2022

Characterizing memory through sensitivity

Low Sensitivity



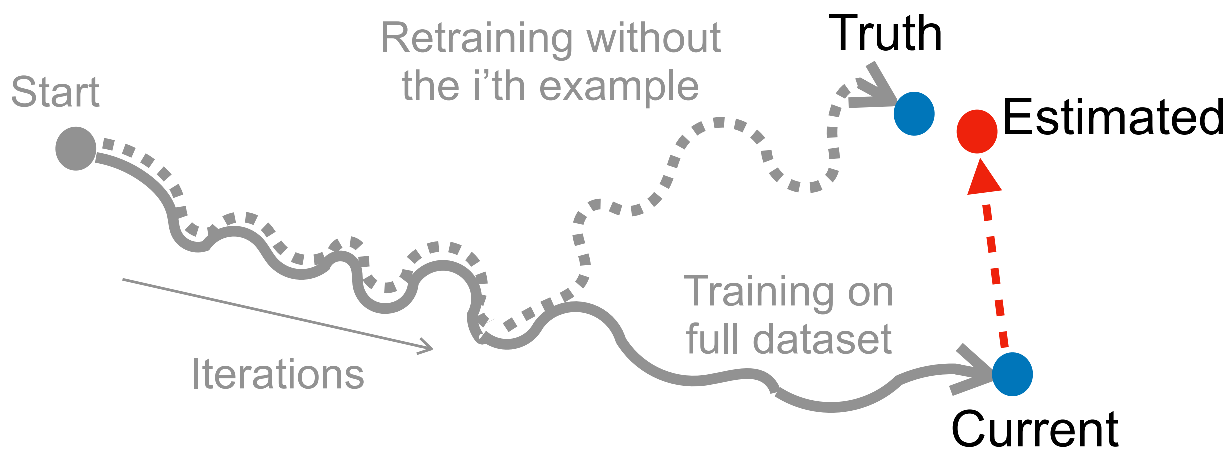
High Sensitivity



1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS, 2023

Memory Perturbation Equation

Past that has the most influence on the present



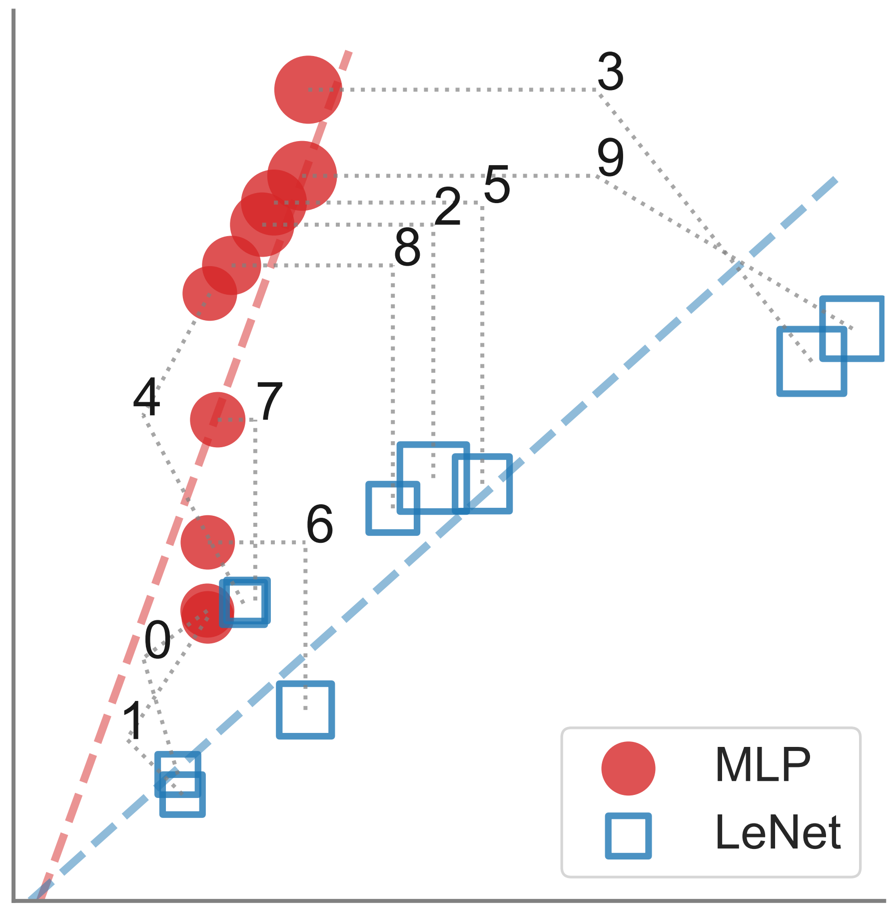
Estimating it without retraining: Using the BLR, we can recover all sorts of influence criteria used in literature.

Influence = predictError x predictVariance

1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS, 2023

Answering “What-If” Questions

What if we removed a class from MNIST?

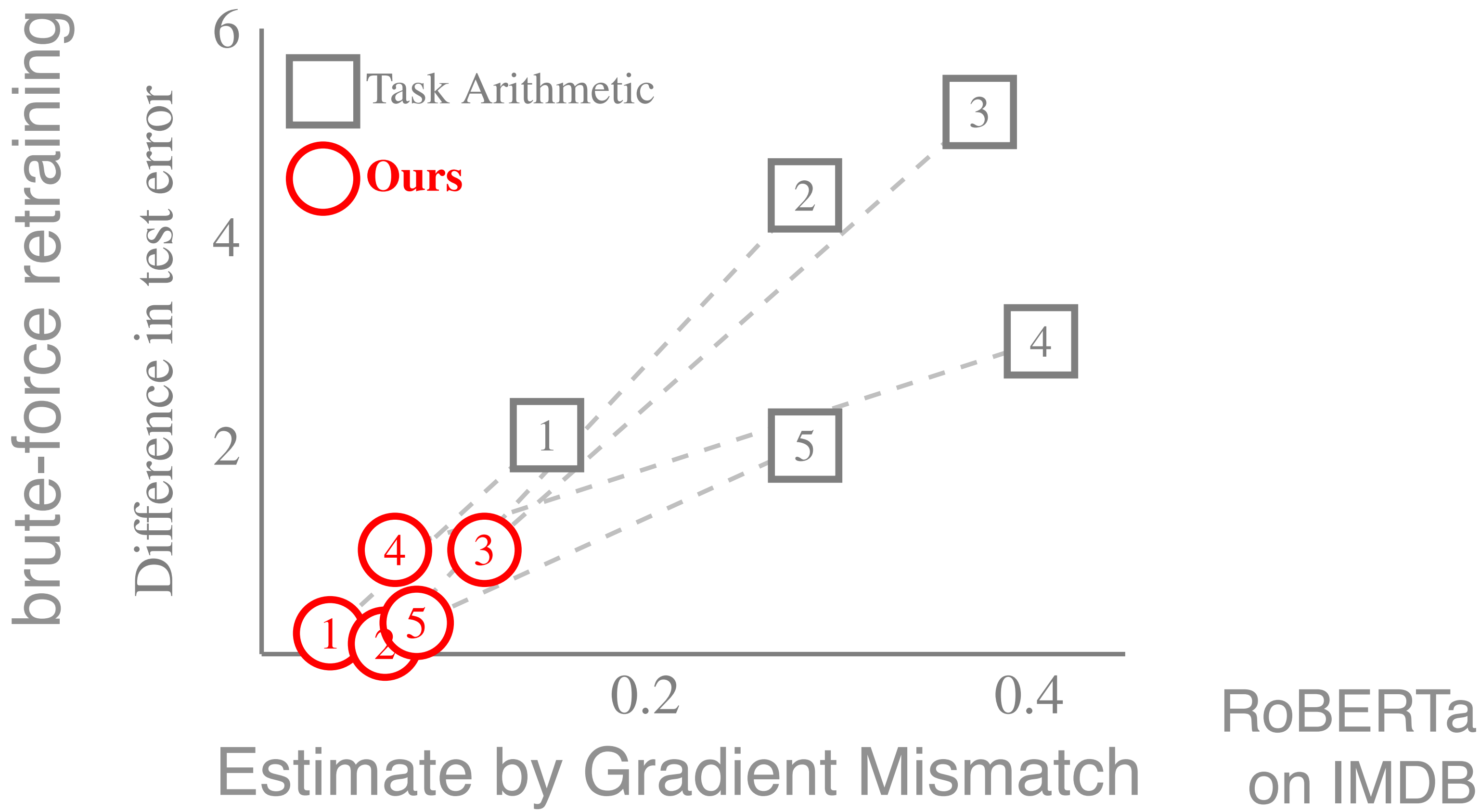


Estimates on training data (no retraining)

Test Performance (NLL) by brute-force retraining

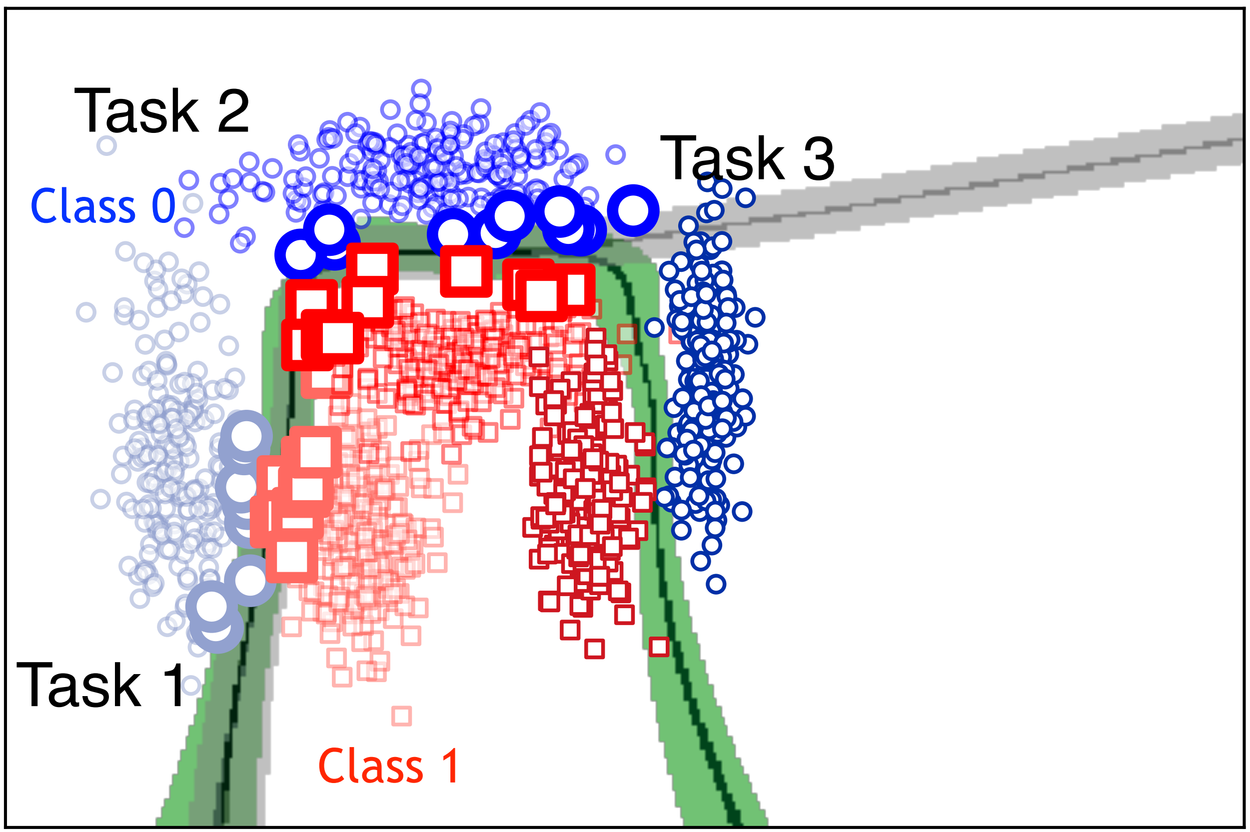
Answering "What-If" Questions

What if we merge fine-tuned large-language models?



1. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).

Functional Regularization of Memorable Examples [2]



1. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
2. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

Functional Regularization of Memorable Past (FROMP)

Weight-regularizer (EWC) [1]

$$(\theta - \theta_{\text{old}})^\top \underbrace{F_{\text{old}}}_{\text{Weight uncertainty}}(\theta - \theta_{\text{old}})$$

Functional regularizer (FROMP) [2]

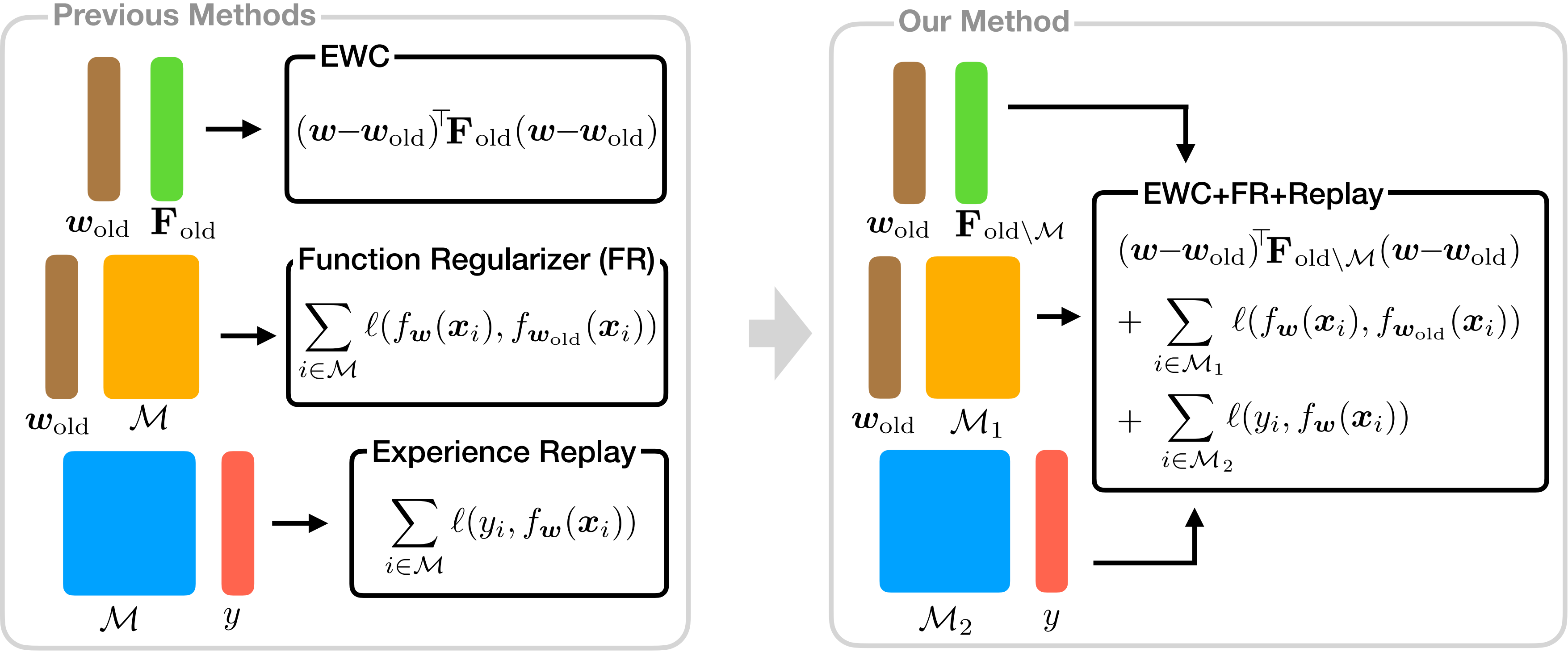
$$[\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]^\top \underbrace{K_{old}^{-1}}_{\text{Uncertainty}}[\sigma(\mathbf{f}(\theta)) - \underbrace{\sigma(\mathbf{f}_{old})}_{\text{Predictions}}]$$

Why does this work? It is a way to replay past gradients, which leads to the idea of K-priors.

1. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS* 2017
2. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

How to combine EWC + FR + Replay

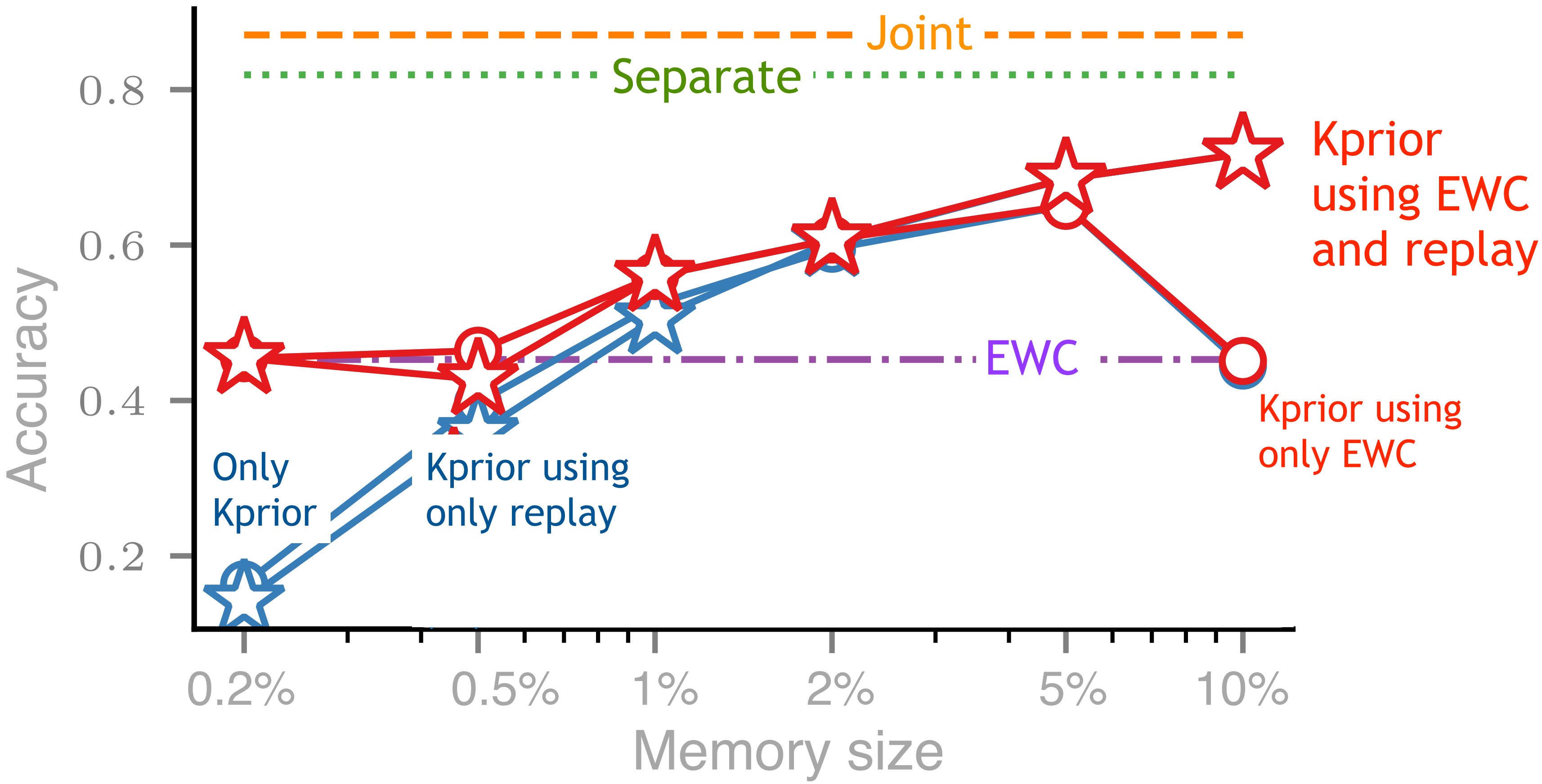
Combine to reduce grad-replay error



1. Daxberger et al. Improving CL by Accurate Gradient Reconstruction of the Past, TMLR 2023.

Continual Learning on ImageNet

K-prior allows us to optimally combine model and data to get good accuracy with little memory.



1. Daxberger et al. Improving CL by Accurate Gradient Reconstruction of the Past, TMLR 2023.

Bayesian Learning Rule [1]

- Bridge DL & Bayesian learning [2-5]
  - SOTA on GPT-2 and ImageNet [5]
- Improve DL [5-7]
  - Calibration, uncertainty, memory etc.
  - Understand and fix model behavior
- Towards human-like quick adaptation

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Khan, et al. Fast and scalable Bayesian deep learning by weight-perturbation in Adam, ICML (2018).
3. Osawa et al. Practical Deep Learning with Bayesian Principles, NeurIPS (2019).
4. Lin et al. Handling the positive-definite constraints in the BLR, ICML (2020).
5. Shen et al. Variational Learning is Effective for Large Deep Networks, Under review.
6. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
7. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)

The webpage is available at https://bayesduality.github.io/, and Twitter account @BayesDuality

The Bayes-Duality Project

Toward AI that learns adaptively, robustly, and continuously, like humans

Emtiyaz Khan	Julyan Arbel	Kenichi Bannai	Rio Yokota
Research director (Japan side)	Research director (France side)	Co-PI (Japan side)	Co-PI (Japan side)
Approx-Bayes team at RIKEN-AIP and OIST	Statify-team, Inria Grenoble Rhône-Alpes	Math-Science Team at RIKEN-AIP and Keio University	Tokyo Institute of Technology

Received total funding of around USD 3 million through JST's CREST-ANR (2021-2027) and Kakenhi Grants (2019-2021).

Team Approx-Bayes

https://team-approx-bayes.github.io/

Emtiyaz Khan
Team Leader

Thomas Möllenhoff
Research Scientist

Geoffrey Wolfer
Special Postdoctoral Resesarcher

Hugo Monzón Maldonado
Postdoctoral Researcher

Many thanks to our group members and collaborators (many not on this slide).

We are always looking for new collaborations.

Keigo Nishida
Postdoctoral Researcher
RIKEN BDR

Zhedong Liu
Postdoctoral Researcher

Peter Nickl
Research Assistant

Joseph Austerweil
Visiting Scientist
University of Winsconsin-Madison

Pierre Alquier
Visiting Scientist
ESSEC Business School

Dharmesh Tailor
Remote Collaborator
University of Amsterdam