Presentation at the 4th Bayes-Duality Workshop, June 25, 2026

A (Quick?) Summary of The Bayes-Duality Project

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo, Japan

TU Darmstadt and Hessian.AI, Germany

https://emtiyaz.github.io

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

A new Bayes-duality principle for adaptive, robust, & lifelong learning of AI

A JST CREST- French ANR project

October 2021 - March 2027

JPY 240M + EUR 500K

Webpage:https://bayesduality.github.io/

The Bayes-Duality Project

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

Emtiyaz Khan

Research director (Japan side)

Approx-Bayes team at RIKEN-AIP and OIST

Julyan Arbel

Research director (France side)

Statify-team, Inria Grenoble Rhône-Alpes

Kenichi Bannai

Co-PI (Japan side)

Math-Science Team at RIKEN-AIP and Keio University

Rio Yokota

Co-PI (Japan side)

Tokyo Institute of Technology

Received total funding of JPY 240M + EUR 500K through the CREST-ANR grant! Thanks to JST for their generous funding!

Bayes-Duality Workshop (June 25-27, 2025)

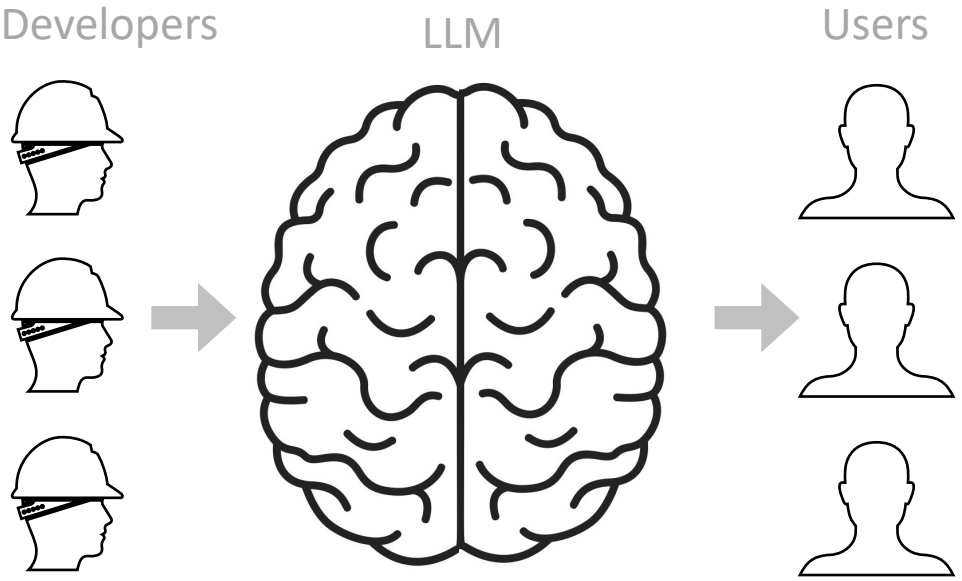
June 15-18, 2026

Sustainable AI Training

The immense data, compute, and infrastructure demands are increasingly unsustainable

How did we get here?

Unsustainability is largely due to the obsession to develop large, static, singular, general intelligence



We want to create a more adaptive AI marketplace that aims to “reduce, reuse, recycle, repair”, etc.

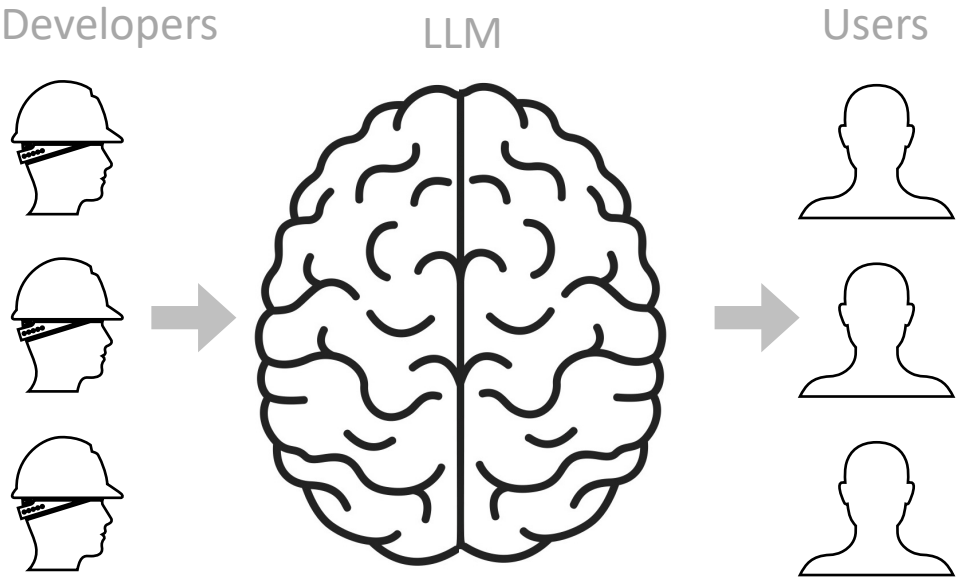
Adaptive Intelligence [1,2]

Challenging due to “catastrophic interference” [3,4]. We aim to address this issue and instill human and animal-like adaptivity in AI.

1. Sternberg. *A theory of adaptive intelligence and its relation to general intelligence.* Journal of Intelligence(2019)
2. Sternberg. *Adaptive intelligence.* New York: Cambridge University Press (2021)
3. Sutton. *Two Problems with Backpropagation and Other Steepest-Descent Learning…*, Cog. Sci. Society (1986)
4. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.

Adaptive Bayesian Intelligence

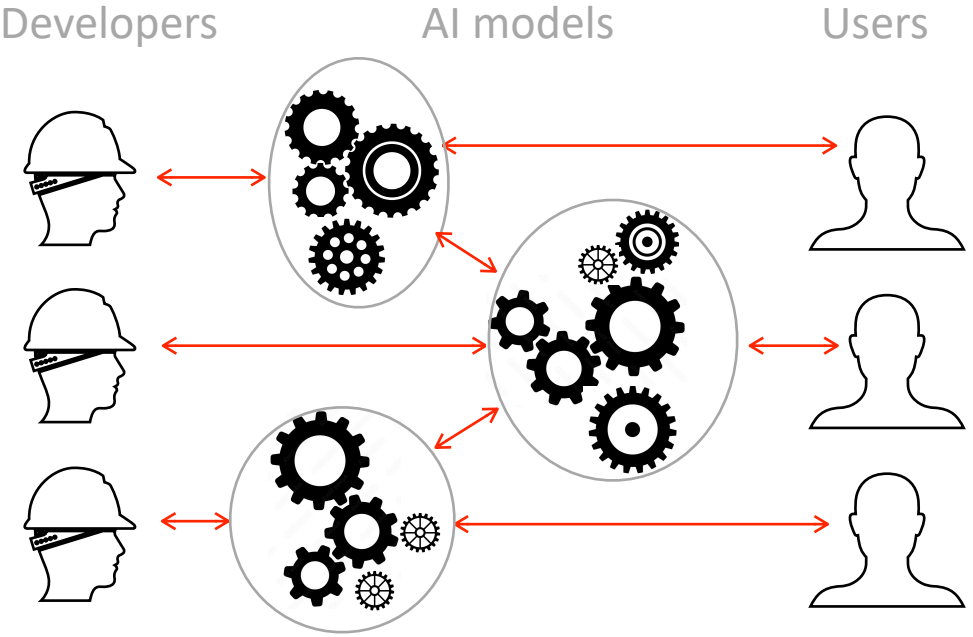
Use “Bayesian-duality” to open up the model



And use Bayesian updating to reuse knowledge.

Towards a Self-Adaptive Collective

We believe our work will lead a new learning paradigm: "self-adaptive collective" intelligence



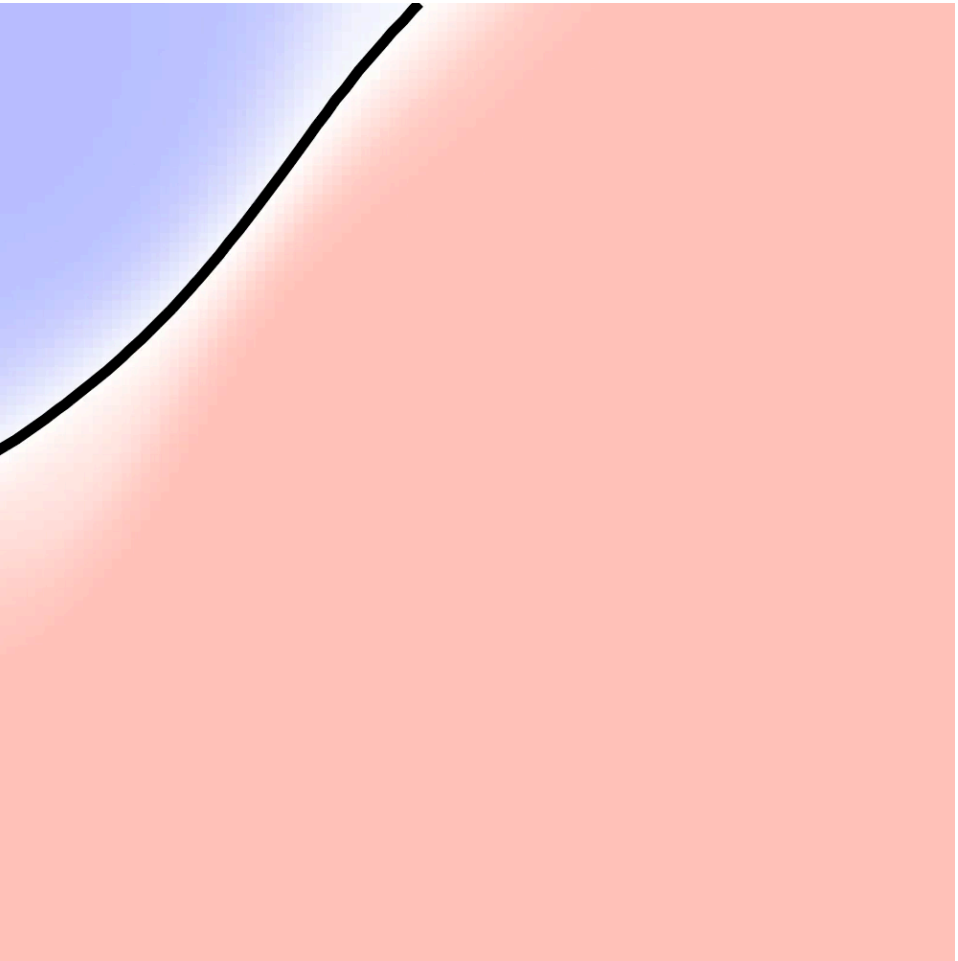
Members communicate and collaborate with each other to ensure well-being and welfare of the whole system

A Summary of Our Research

- Adaptive Intelligence (focus on being “quick”)
  - Unification of existing adaptation frameworks
- When is quick adaptation possible
  - Interference=mismatch & adaptation=correction
- New optimizers that naturally balance past and future knowledge and collect new information
  - New ways to define and compute information gain
- Main challenge: Memory/communication/curriculum
- New paradigm: Self-adaptive collective intelligence

Adaptative Intelligence: An Example

Big markers indicate a recent use of the example



Challenging due to “catastrophic interference” where new data interferes with the old decision boundary.

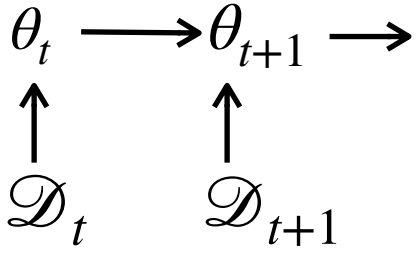
Balancing the past and future to avoid interference is extremely hard.

The Goal of Adaptive Intelligence

Update knowledge "quickly". For instance, in continual learning, we want to update the model with new data (although focus on quick adaptation is often missing)

Old model: $\theta_t \leftarrow \arg\min \sum_{j=0}^{t} \ell_j$

Retraining: $\theta_{t+1} \leftarrow \arg\min \ell_{t+1} + \sum_{j=0}^{t} \ell_j$



Quick Adaptation: $\theta_{t+1} \leftarrow Adapt(\mathscr{D}_{t+1}, \theta_t, \mathscr{D}_{1:t})$

EWC [1]: $\hat{\theta}_{t+1} \leftarrow \arg\min \ell_{t+1} + \rho_t \|\theta - \theta_t\|^2$

Other case: model merging, federated/distribution learning, unlearning, model editing, local learning

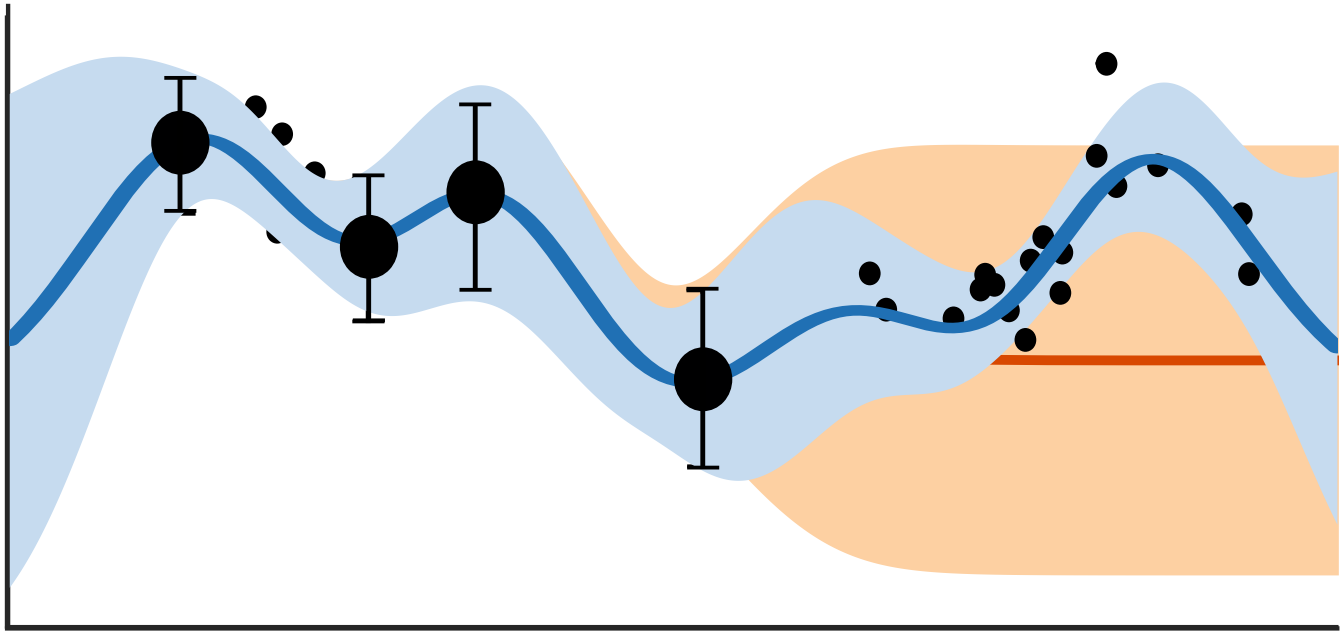
1. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.

Adaptation via Bayes

Bayes' Rule: $p_{t+1} \propto p_0 \prod_{j=1}^{t+1} \overbrace{\exp(-\ell_j)}^{=lik_j} = p_t \times \exp(-\ell_{t+1})$

Recursive Aggregation

What if we do not have the full posterior, rather only approximate information about it?



The Bayesian Duality Principle

Bayesian generalization of the representer theorem, applicable to non-convex cases

Knowledge Representation

Duality [1] provides one way to represent the past

Fixed point: $$\theta_t = \sum_{i=1}^{t} x_i \underbrace{(x_i^\top \theta - y_i)}_{=\alpha_{i|t}} = \sum_{i=1}^{t} \underbrace{\nabla \ell_i(\theta_t)}_{=g_{i|t}}$$

Dual pair: $(\theta_t, \boldsymbol{\alpha}_t)$ Vector of length t $(\theta_t, \boldsymbol{g}_t)$

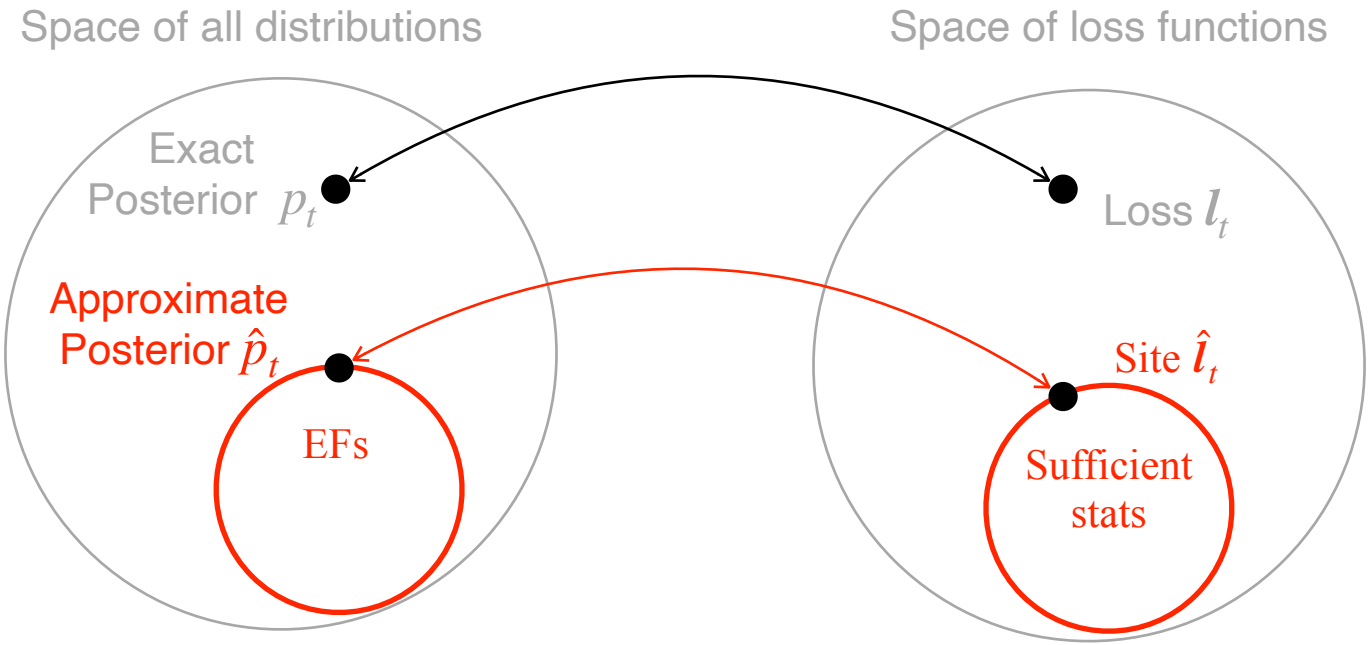
These are solutions of equivalent problems defined in two dual spaces [2].

Is there a more general formulation (for Bayes & for neural networks)? Can we use similar ideas to "compactly" represent uncertainty and facilitate quick adaptation?

1. Atiyah, Duality in Mathematics and Physics, Lecture (Dec 18, 2007)
2. Scholkopf et al. A generalized representer theorem. COLT (2001)

The Bayesian Duality Principle

Every (variational) posterior has a dual pairing



Sites are parameterized by “natural” gradients.

Duality of (Exact) Bayes

Bayesian generalization with dual-pair $(p_t, \boldsymbol{l}_t)$; see [3]

$$p_t = \frac{1}{Z_t} p_0 \prod_{j=1}^{t} \exp(-\ell_j) \qquad = \arg\min_{q \in \mathscr{P}} \sum_{j=1}^{t} \mathbb{E}_q[\ell_j] + KL(q \| p_0)$$

$$\boldsymbol{l}_t = \log\left(Z_t \frac{p_t}{p_0}\right) \qquad = \arg\min_{\boldsymbol{f} \in \mathscr{P}^*} \sum_{j=1}^{t} \mathbb{E}_{p_t}[f_j] + \log \int dp_0 \prod_{j=1}^{t} \exp(-f_j)$$

Do similar representations exist for approximations?

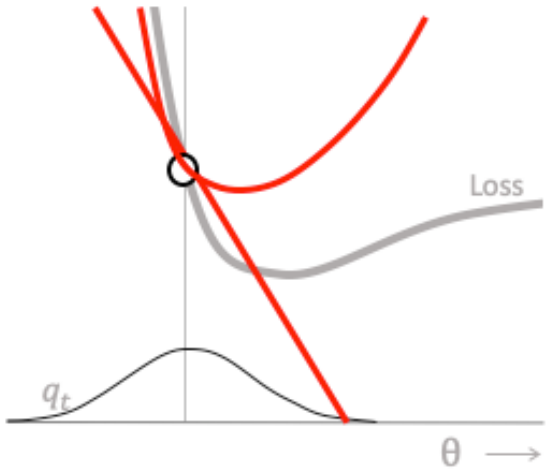
1. Kimeldorf & Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* (1970)
2. Csató & Opper. Sparse on-line Gaussian processes. *Neural computation* (2002)
3. Zhu et al. Bayesian inference with posterior regularization and applications to infinite latent SVMs.JMLR (2014)

Duality of Variational Bayes [1]

An (optimal) posterior $\hat{p}_t = q_t$ has a dual "site" $\hat{\boldsymbol{l}}_t$

Exp-Fam posterior: $q_t = \frac{1}{Z}\exp(T(\theta)^\top \lambda)$ Sites: $\hat{\ell}_{j|t} = T(\theta)^\top \tilde{\nabla}_\lambda \mathbb{E}_{q_t}[\ell_j]$

$$q_t = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})$$



$$N(m_t, I) = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\theta^\top \mathbb{E}_{q_t}[\nabla \ell_j])$$

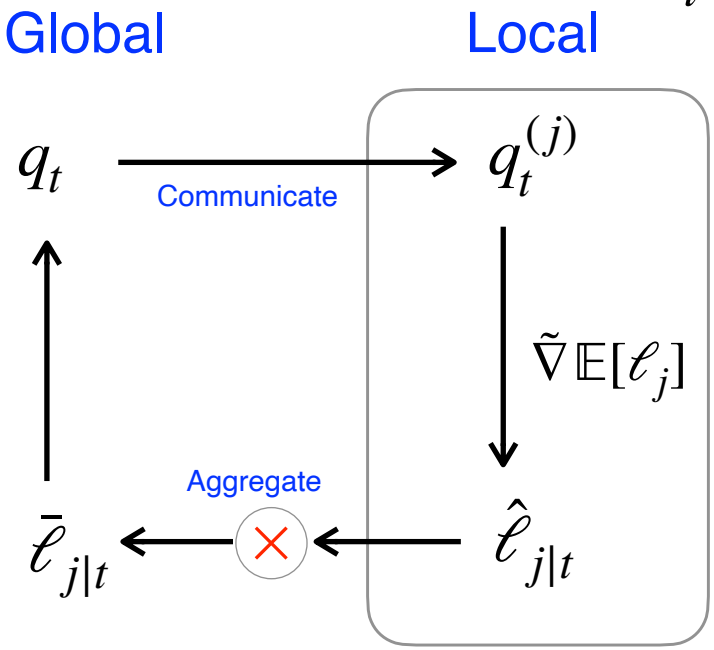
$$N(m_t, \Sigma_t) = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\theta^\top \mathbb{E}_{q_t}[\nabla \ell_j] - \frac{1}{2}(\theta - m_t)^\top \mathbb{E}_{q_t}[\nabla^2 \ell_i](\theta - m_t))$$

1. Khan and Nielsen. Fast yet Simple Natural-Gradient Descent for VI… ISITA (2018)

The Bayesian-Duality Structure

The flow of information suggests how to compute optimal posterior through local computations

$$q_t = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})$$



Natural-gradient descent to implement this leads to the Bayesian Learning Rule [2]

$$q_t \leftarrow q_t^{1-\rho} \left[p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t}) \right]^{\rho}$$

1. Khan et al. Fast Dual Variational Inference for Non-Conjugate LGMs. ICML (2013)
2. Khan and Rue. The Bayesian Learning Rule. JMLR (2023)

Optimization

Gradient Descent
Newton's Method
Multimodal Optimization

Deep-Learning

SGD, RMSprop and Adam
Sharpness-Aware Minimization
Dropout, STE, Label Smoothing
SOAP….

Bayesian Learning Rule [1]

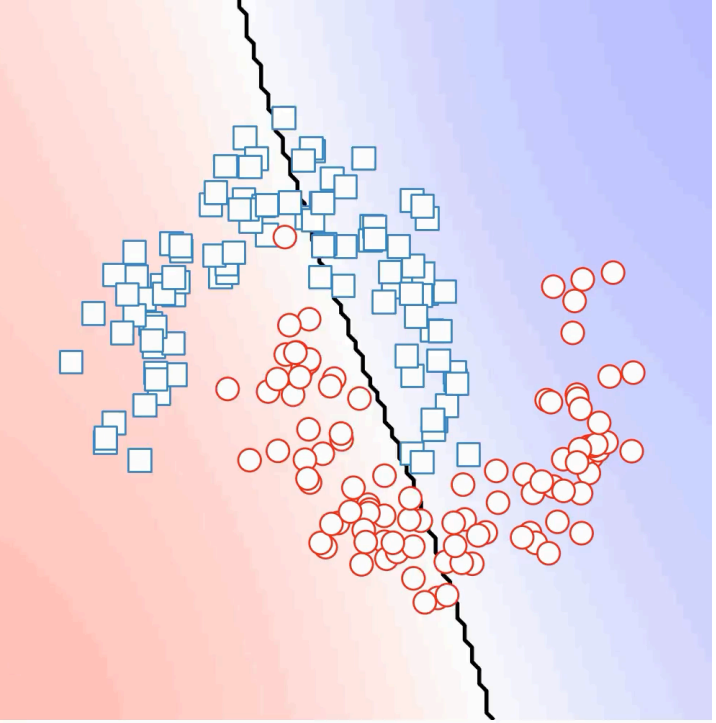
Approximate Inference

Conjugate Bayes
Laplace's Method
Expectation Maximization
Stochastic Variational Inference
Variational Message Passing

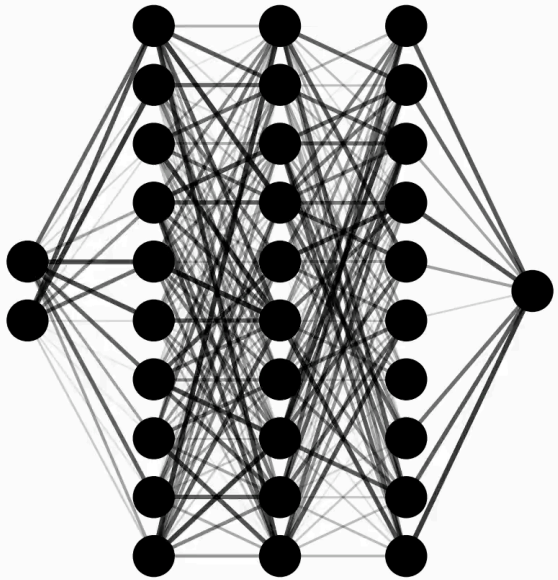
Global-Optimization

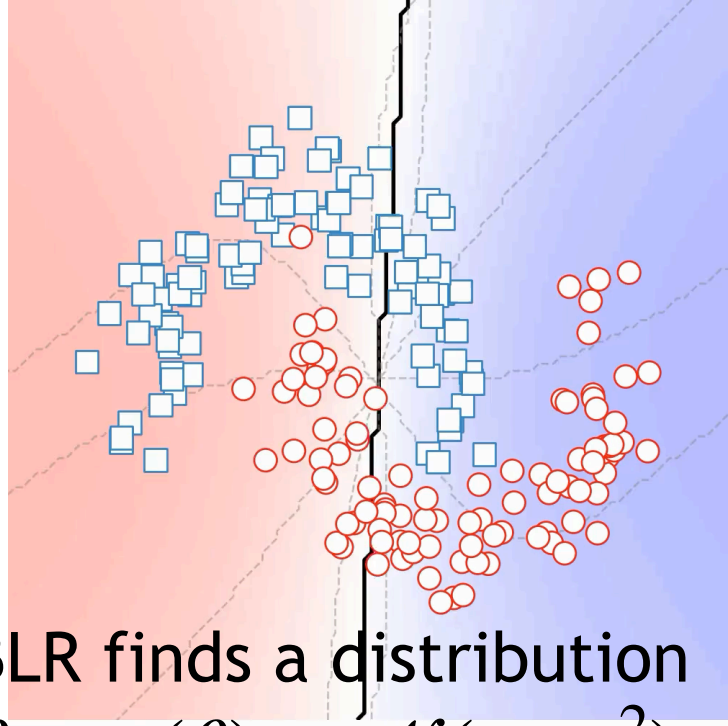
Exponential-Weight Aggregation
Natural Evolution Strategy
Gaussian Homotopy
Smoothed Optimization
Weight-perturbed Optimization
Stochastic Search (annealing)
Stochastic Relaxation

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023)
2. Khan and Lin. Conjugate-Compute Variational Inference, AISTATS (2017)



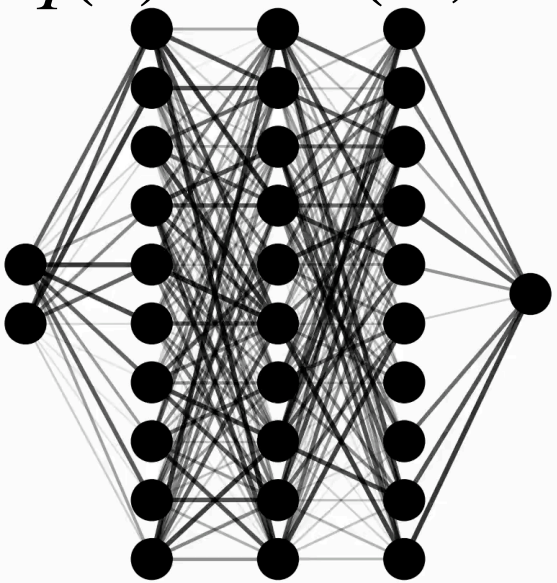
Adam finds θ





BLR finds a distribution

$\theta \sim q(\theta) = \mathcal{N}(m, \sigma^2)$



BLR via Adam-Like Training

For $q = \mathcal{N}(m, \sigma^2)$,
BLR resembles Adam.

$$q_t \leftarrow q_t^{1-\rho} \prod_{j \in \mathcal{B}} \exp(-\rho\, \hat{\ell}_{j|t})$$

RMSprop/Adam

$$\hat{g} \leftarrow \hat{\nabla} \ell(\theta)$$
$$\hat{h} \leftarrow \hat{g}^2$$
$$h \leftarrow (1-\rho)h + \rho\hat{h}$$
$$\theta \leftarrow \theta - \alpha(\hat{g} + \delta m)/(\sqrt{h} + \delta)$$

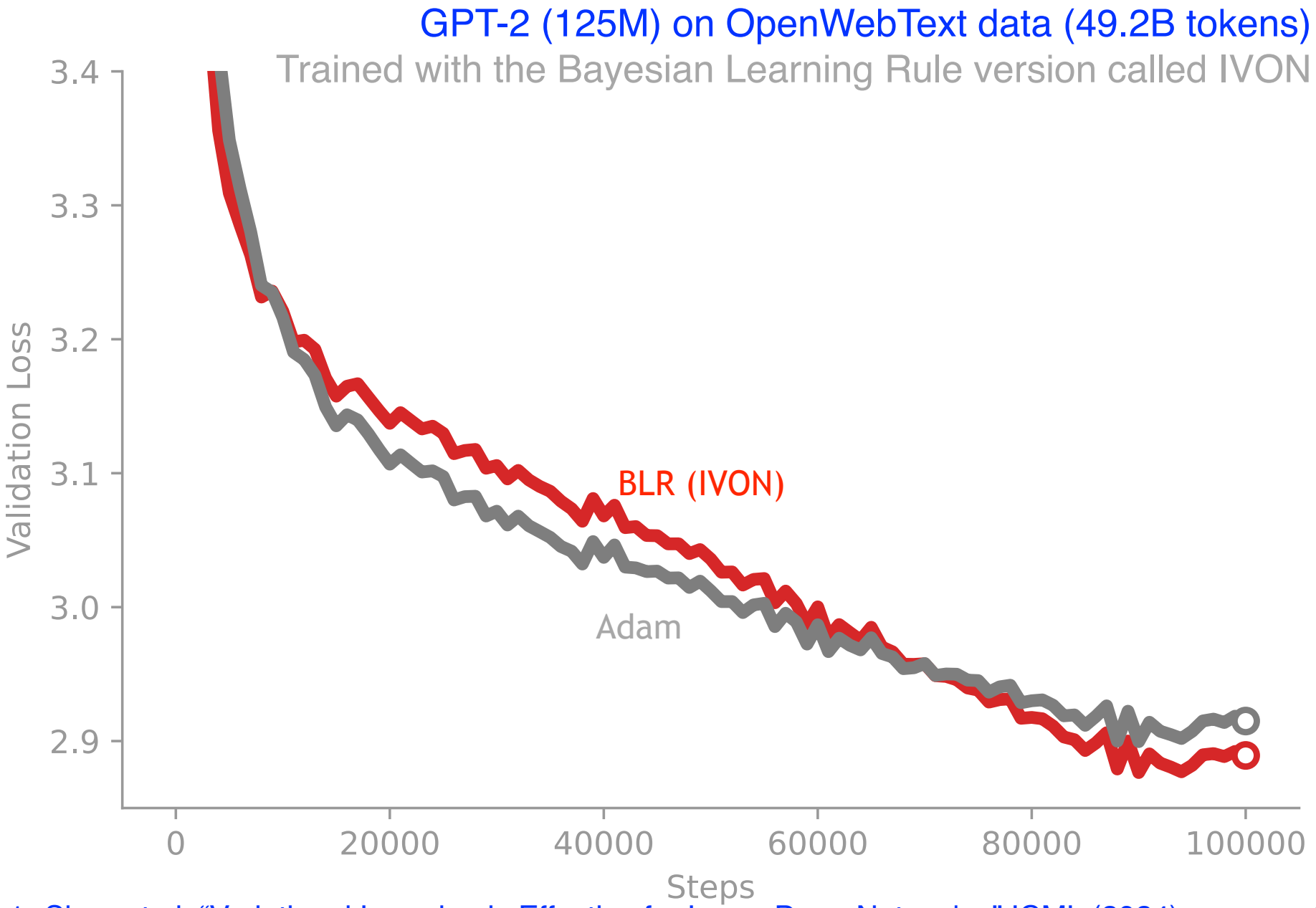
Improved Variational Online Newton (IVON) [4]

$$\hat{g} \leftarrow \hat{\nabla} \ell(\theta) \text{ where } \theta \sim \mathcal{N}(m, \sigma^2)$$
$$\hat{h} \leftarrow \hat{g} \cdot (\theta - m)/\sigma^2$$
$$h \leftarrow (1-\rho)h + \rho\hat{h} + \rho^2(h - \hat{h})^2/(2(h+\delta))$$
$$m \leftarrow m - \alpha(\hat{g} + \delta m)/(h + \delta)$$
$$\sigma^2 \leftarrow 1/(N(h+\delta))$$

```
pip install ivon-opt
```

1. Khan, et al. "Fast and scalable Bayesian deep learning by…." 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).
4. Shen et al. "Variational Learning is Effective for Large Deep Networks." ICML (2024)

Bayesian Learning Rule is as cheap and accurate as Adam



1. Shen et al. “Variational Learning is Effective for Large Deep Networks.” ICML (2024)

SOAP-Bubbles ∘❍
Structured Weight Uncertainty for Neural Networks

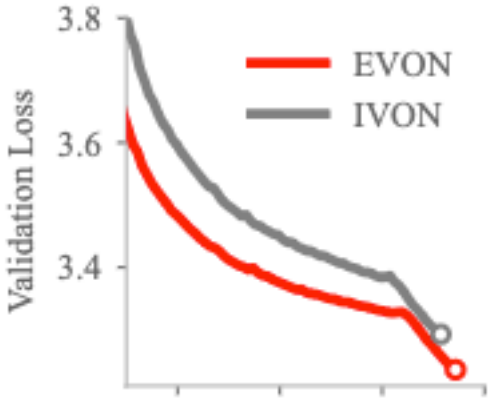
Adrian R. Minut[1*] **Nico Daheim**[2] **Marco Miani**[3]
Mohammad Emtiyaz Khan[4,5] **Wu Lin**[6] **Thomas Möllenhoff**[5†]

[1]Sapienza University of Rome, Rome, Italy
[2]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science, Technical University of Darmstadt
National Research Center for Applied Cybersecurity ATHENE, Germany
[3]Technical University of Denmark, Lyngby, Denmark
[4]TU Darmstadt & Hessian Center for AI (hessian.AI), Darmstadt, Germany
[5]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[6]University of Central Florida, Orlando, United States

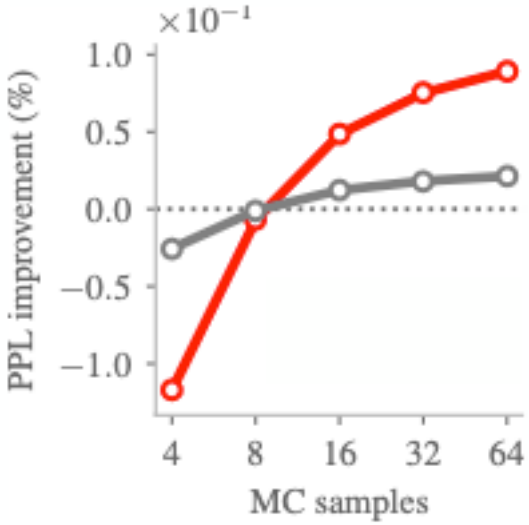
Abstract

Structured weight-uncertainty can improve many aspects of deep learning, but it remains costly to estimate and difficult to implement. Here, we show that these issues can be addressed by adapting the SOAP optimizer. Our key idea is to run IVON, an existing diagonal-covariance variational method, in the eigenspace of SOAP's preconditioner and then use the preconditioner to transform the diagonal estimate into a non-diagonal covariance. The resulting method has costs similar to those of SOAP and requires no drastic changes to training pipelines. We call the posteriors obtained in this way SOAP-Bubbles and our new optimizer Eigenspace-VON (EVON). We show that, for logistic regression, EVON recovers the exact Gaussian covariance and that, for language model pretraining, it yields significantly better results than existing diagonal-covariance methods. Our work makes it easier to estimate more expressive posterior distributions for deep learning at scale.

NanoGPT-123M-FineWeb1B



Test-Time Posterior Sampling



Using SOAP within the BLR

Algorithm 1 SOAP (Vyas et al., 2025)

No Sampling

$\mathbf{G} \leftarrow \hat{\nabla}\ell(\mathbf{\Theta}), \mathbf{G}^\circ \leftarrow \mathbf{Q}_L^\top \mathbf{G}\mathbf{Q}_R$

$\widehat{\mathbf{H}} \leftarrow \mathbf{G}^\circ \odot \mathbf{G}^\circ$

Update Preconditioner and Momentum

$\bar{\mathbf{G}} \leftarrow \beta_1 \bar{\mathbf{G}} + (1-\beta_1)\mathbf{G}^\circ$

$\mathbf{H} \leftarrow \beta_2 \mathbf{H} + (1-\beta_2)\widehat{\mathbf{H}}$

Update Parameters

$\mathbf{U} \leftarrow \bar{\mathbf{G}}/(\sqrt{\mathbf{H}} + \epsilon)$

$\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \alpha \mathbf{Q}_L \mathbf{U} \mathbf{Q}_R^\top$

Update Eigenbasis

$\mathbf{L} \leftarrow \beta_3 \mathbf{L} + (1-\beta_3)\mathbf{G}\mathbf{G}^\top$

$\mathbf{R} \leftarrow \beta_3 \mathbf{R} + (1-\beta_3)\mathbf{G}^\top \mathbf{G}$

Every T steps:

$\mathbf{Q}_L, \mathbf{Q}_R \leftarrow \mathrm{Eig}(\mathbf{L}), \mathrm{Eig}(\mathbf{R})$

Update momentum $\bar{\mathbf{G}}$ (new basis)

Algorithm 2 EVON (Proposed Method)

Sample Parameters

$V_{i,j} \leftarrow 1/(\zeta(H_{i,j} + \delta))$

$\mathbf{\Theta} \leftarrow \mathbf{M} + \mathbf{Q}_L \mathbf{E} \mathbf{Q}_R^\top, E_{i,j} \sim \mathcal{N}(0, V_{i,j})$

$\mathbf{G} \leftarrow \hat{\nabla}\ell(\mathbf{\Theta}), \mathbf{G}^\circ \leftarrow \mathbf{Q}_L^\top \mathbf{G}\mathbf{Q}_R$

$\widehat{\mathbf{H}} \leftarrow \mathrm{clip}(\mathbf{E} \odot \mathbf{G}^\circ / \mathbf{V})$

Update Hessian and Momentum

$\bar{\mathbf{G}} \leftarrow \beta_1 \bar{\mathbf{G}} + (1-\beta_1)\mathbf{G}^\circ$

$\mathbf{H} \leftarrow \beta_2 \mathbf{H} + (1-\beta_2)\widehat{\mathbf{H}} + \frac{1}{2}(1-\beta_2)^2(\widehat{\mathbf{H}} - \mathbf{H})^{\odot 2}/(\mathbf{H} + \delta)$

Update Posterior Mean

$\mathbf{U} \leftarrow (\bar{\mathbf{G}} + \delta \mathbf{Q}_L^\top \mathbf{M} \mathbf{Q}_R)/(\mathbf{H} + \delta)$

$\mathbf{M} \leftarrow \mathbf{M} - \alpha\, \mathrm{clip}(\mathbf{Q}_L \mathbf{U} \mathbf{Q}_R^\top)$

Update Eigenbasis

$\mathbf{L} \leftarrow \beta_3 \mathbf{L} + (1-\beta_3)\mathbf{G}\mathbf{G}^\top$

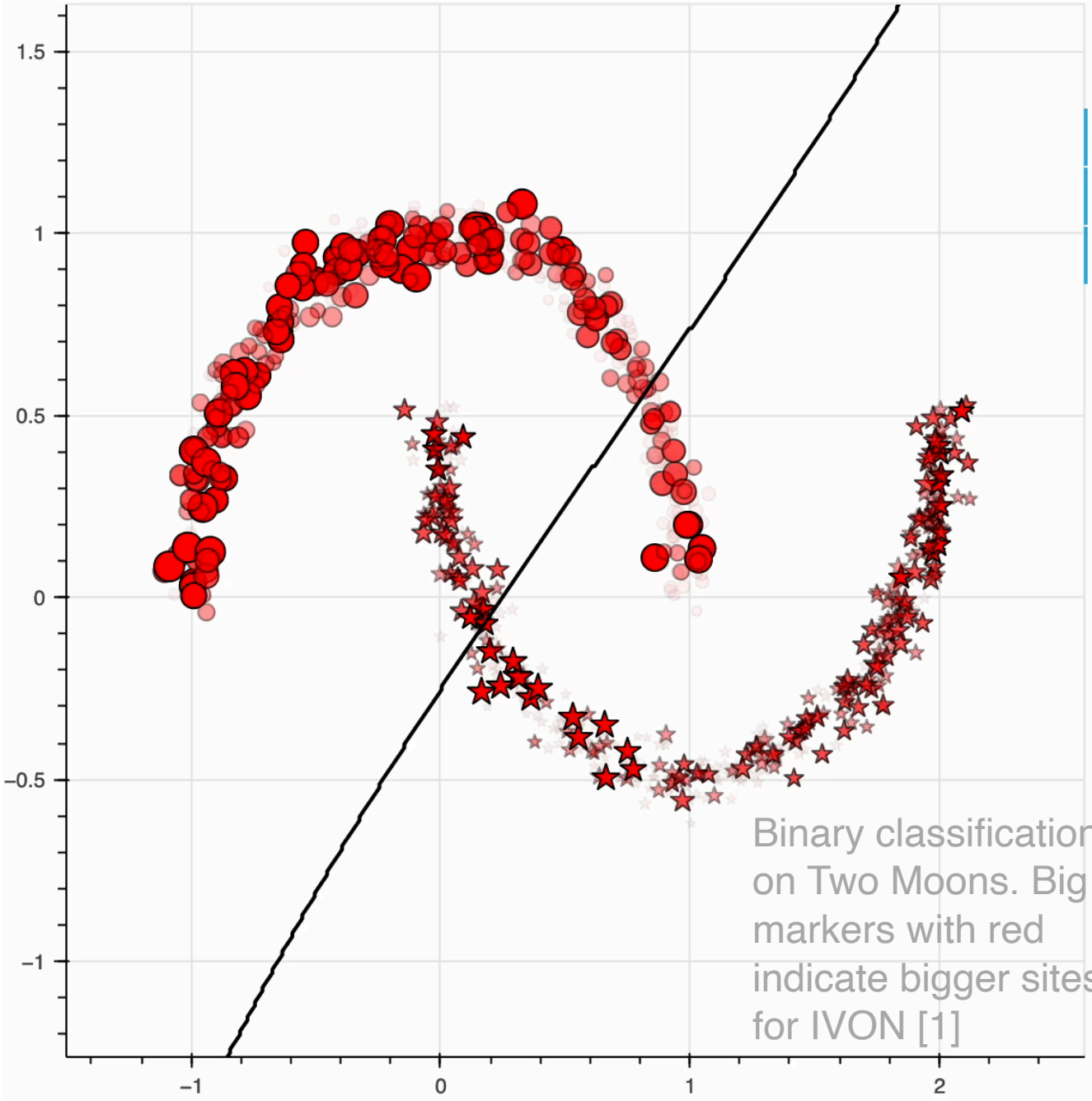
$\mathbf{R} \leftarrow \beta_3 \mathbf{R} + (1-\beta_3)\mathbf{G}^\top \mathbf{G}$

Every T steps:

$\mathbf{Q}_L, \mathbf{Q}_R \leftarrow \mathrm{Eig}(\mathbf{L}), \mathrm{Eig}(\mathbf{R})$

Update momentum $\bar{\mathbf{G}}$ (new basis)

1. Minut et al. ,SOAP-Bubbles ○❍ Structured Weight Uncertainty for Neural Networks, arXiv (2026)



Binary classification on Two Moons. Big markers with red indicate bigger sites for IVON [1]

The dual representation give rise to a new way to derive influence as “information gain”.

This is much more widely applicable than the standard “influence function” valid only for post-training.

By Rin Intachuen (RIKEN AIP)

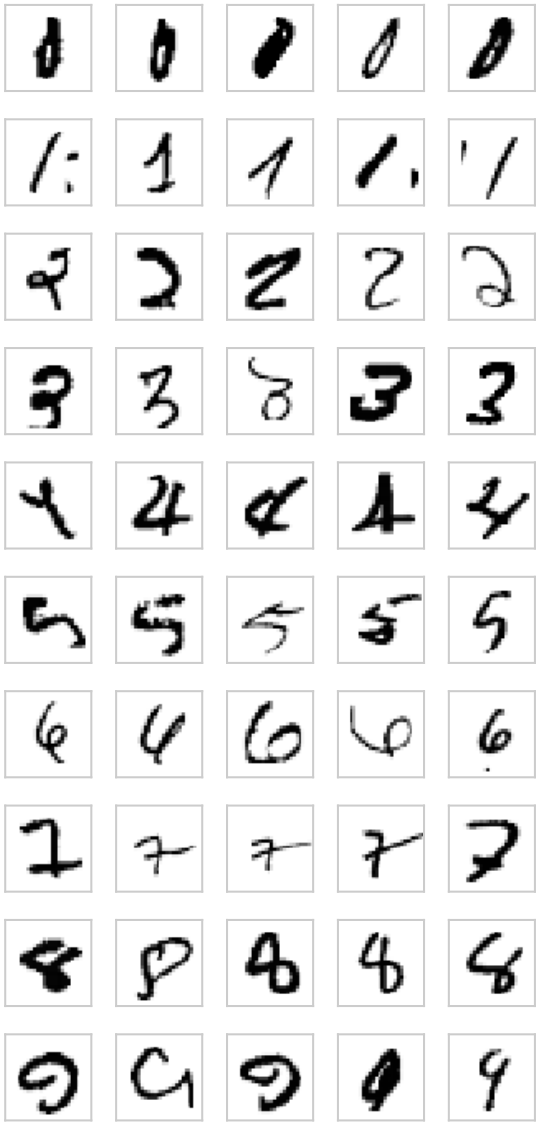
Epoch: 0

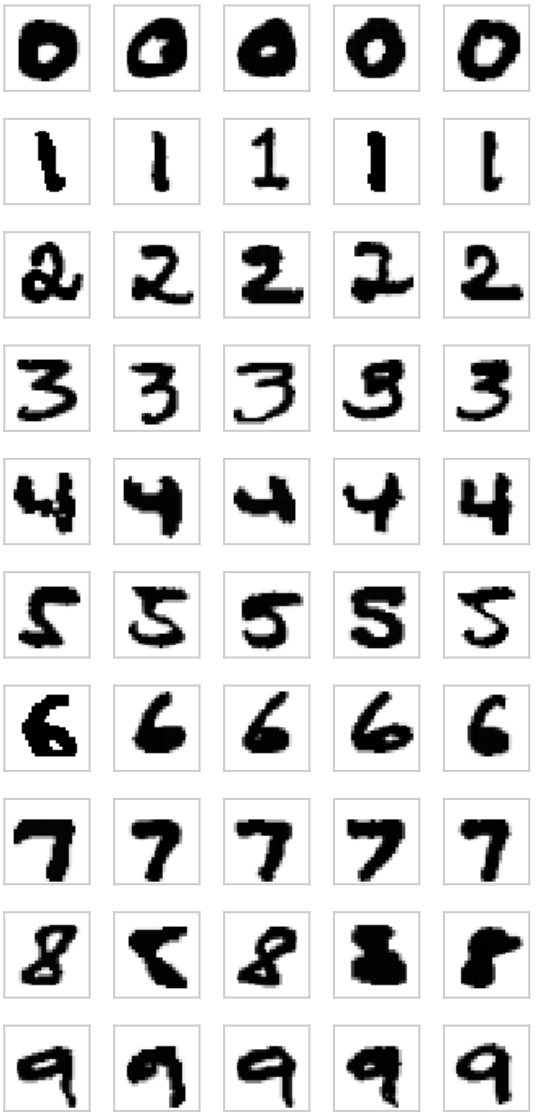
1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)

High Influence

MNIST with CNN

Low Influence





1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)
2. Pan et al. Continual Learning by Functional Regularisation of Memorable Past,NeurIPS (2020)
3. Khan et al. Approximate Inference Turns Deep Networks into GP, NeurIPS (2019)

High Influence

Traffic light (ImageNet)

What class is this?

Low Influence

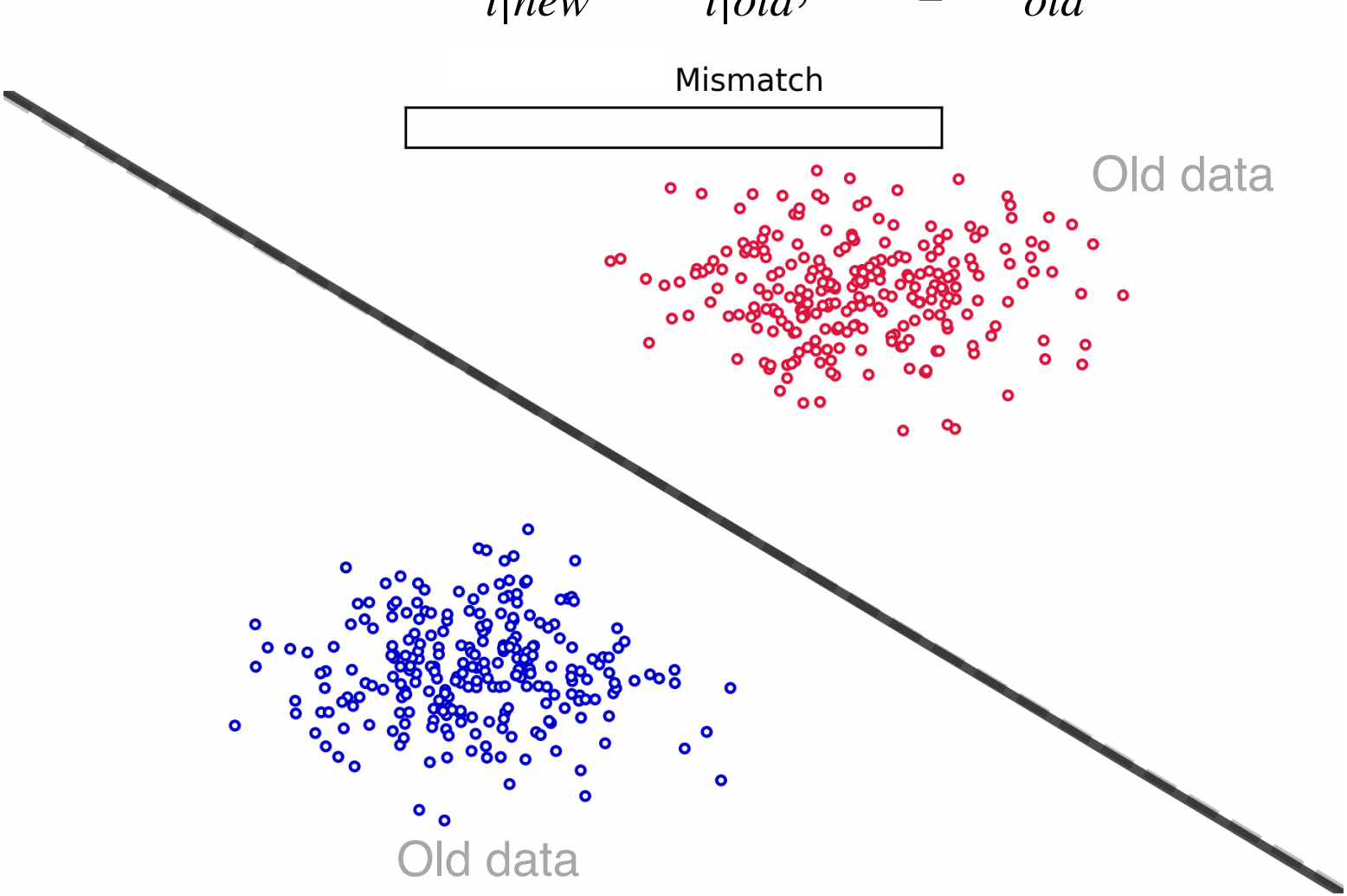
High Influence

Chihuahua class (ImageNet)

Low Influence

Interference as Mismatch

$$\hat{\ell}_{i|new} - \hat{\ell}_{i|old}, \forall i \in \mathcal{D}_{old}$$



Continual Learning (Sec 3.1)

Elastic Weight Consolidation
Variational Continual Learning
Memory Replay Methods
Functional Regularization
Knowledge Adaptation Prior

Model Merging (Sec 3.3)

Task Arithmetic
Fisher/Hessian-Based Merging
Ensembles Methods

Posterior Correction [1]

Unlearning and Influence (Sec 3.2)

Student-Teacher Learning (Sec 4.4)

Knowledge Distillation
Learning with Privileged information
Incremental SVMs

Variance Reduction [2]

SVRG, SAG, SARAH,…

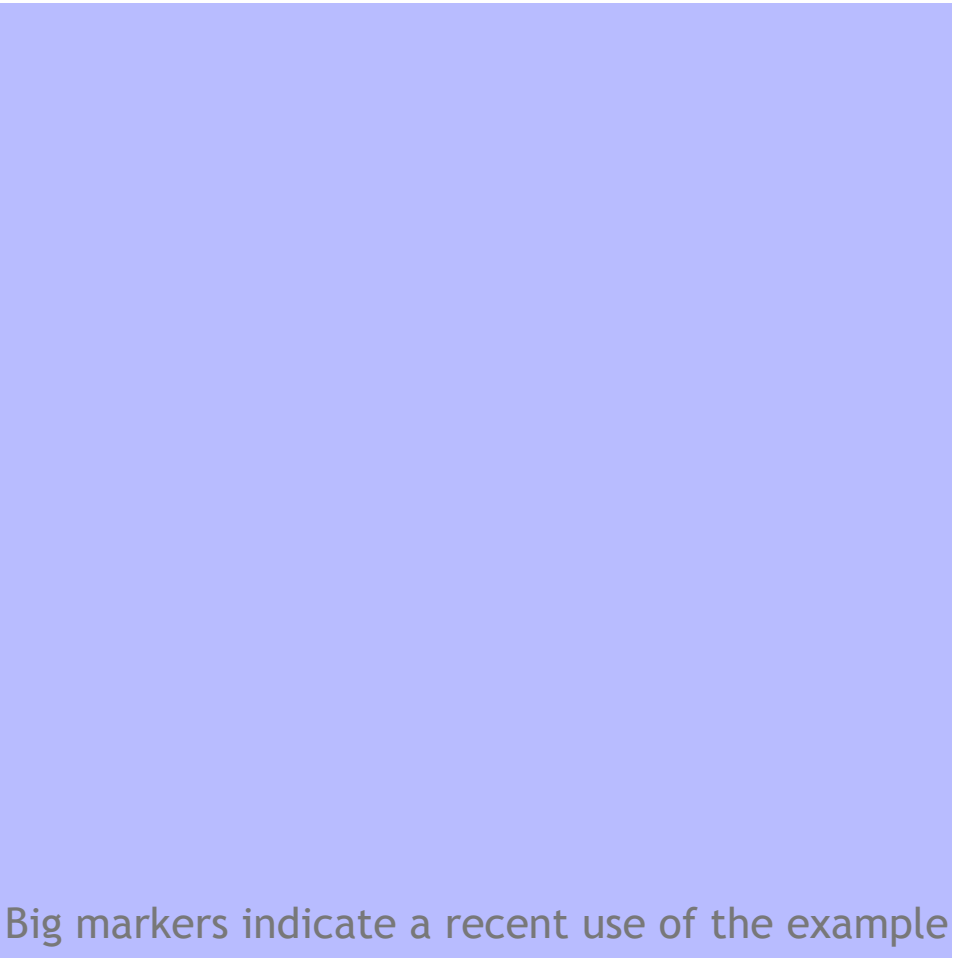
Federated Learning (Sec 3.4)

FedAvg, FedDyn
Alternating Direction Method of Multipliers (ADMM)
Alternating Minimization Algorithm (AMA)
Partitioned Variational Inference

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Daheim et al. SVRG and Beyond with Posterior Correction, arXiv (2025)

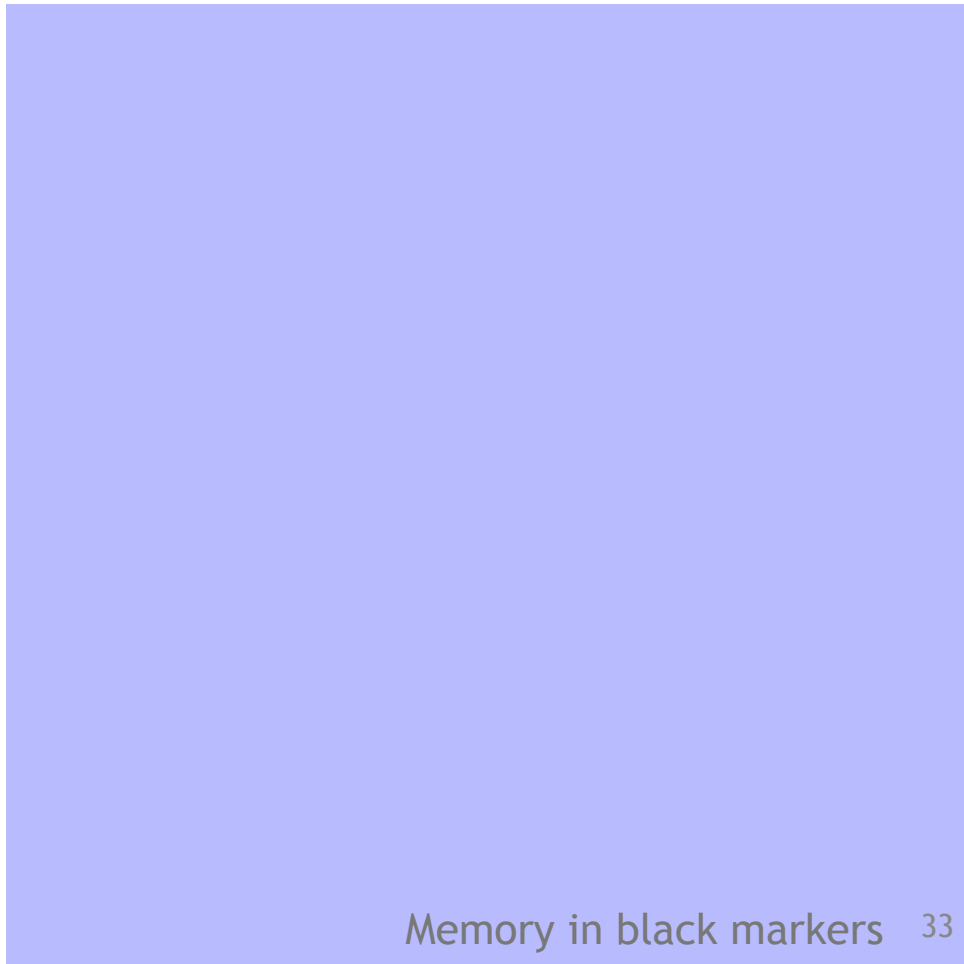
Towards an optimizer that naturally adapts continually

The Adam optimizer



Big markers indicate a recent use of the example

The PoCo optimizer



Memory in black markers

Speeding up training with memories

Given an older checkpoint(s) q_{old}, we can build a prior

$$q_{old} \leftarrow \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|old})$$

In the future, we aim to "correct" this as new data arrives

$$q \leftarrow q^{1-\rho} \prod_{j=1}^{t} \exp(-\rho\, \hat{\ell}_j) \quad \times \frac{q_{old}^{\rho}}{\prod_{j=1}^{t} \exp(-\rho \hat{\ell}_{j|old}))}$$

$$q \leftarrow q^{1-\rho} q_{old}^{\rho} \prod_{j \in \mathcal{B}} \exp(-\rho \underbrace{(\hat{\ell}_j - \hat{\ell}_{j|old})}_{correction})$$

Essentially, we replace old duals by new ones.
Surprisingly, this is also a form of variance reduction [1,2]

1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)
2. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
3. Paul's talk next will use this for fast and slow continual learning

SVRG and Beyond via Posterior Correction

Nico Daheim [1] **Thomas Möllenhoff** [2] **Ming Liang Ang** [3] **Mohammad Emtiyaz Khan** [2]

Abstract

Stochastic Variance Reduced Gradient (SVRG) and its variants aim to speed-up training by using gradient corrections. Originally proposed over a decade ago, these methods have never been connected to any Bayesian method at a fundamental level. Here, we fill this gap and derive surprising new connections of SVRG to a recently proposed Bayesian method called 'posterior correction'. Our main contribution is to show that SVRG can be recovered as a special case of posterior-correction over isotropic-Gaussian posteriors. Novel extensions of SVRG are automatically obtained by using more flexible exponential-family posteriors. We derive two new such extensions by using Gaussian families: a Newton-like variant with novel Hessian corrections, and an Adam-like extension that scales to large problems. Our work is the first to connect SVRG to Bayes and use it to speed-up training.

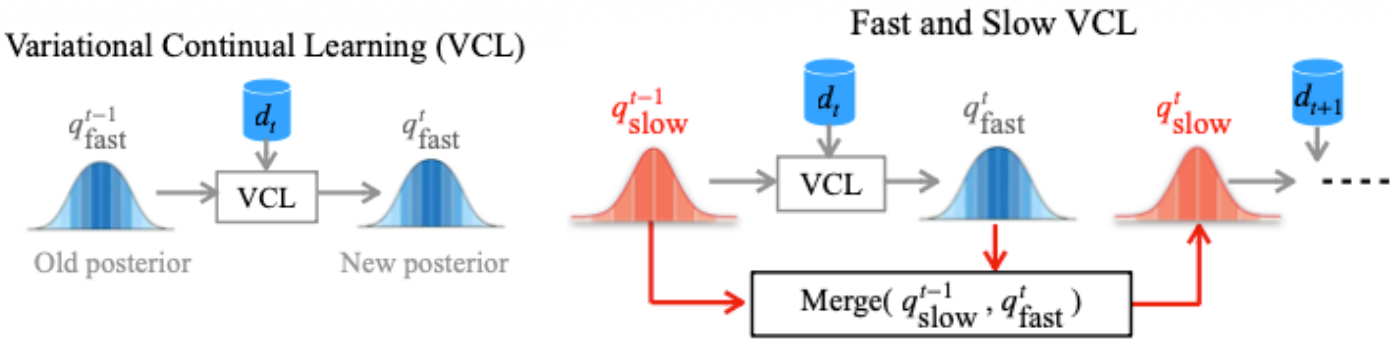
have never been connected to any Bayesian method. This is not because variance reduction is not useful for Bayesian methods. In fact, some works have attempted to use it to speed up Bayesian procedures (Mandt & Blei, 2014; Zhang et al., 2019). However, a deeper and more fundamental connection does not exist.

In this paper, we fill this gap and show a previously unknown connection between SVRG and Bayes. Our first contribution is to show that SVRG can be derived as a special case of a recently proposed Bayesian method called posterior-correction (PoCo) (Khan, 2025); see Fig. 1. The new connection is surprising because PoCo is a unifying mechanism for knowledge adaptation methods such as continual learning and model merging, and is not directly related to variance-reduction. Our result provides the first direct connection between such adaptation methods and variance reduction. It offers a new perspective where gradient-corrections in SVRG can be seen as a mechanism of knowledge-transfer between old and new gradients.

Our second contribution is to derive new extensions

Fast and Slow Variational Continual Learning

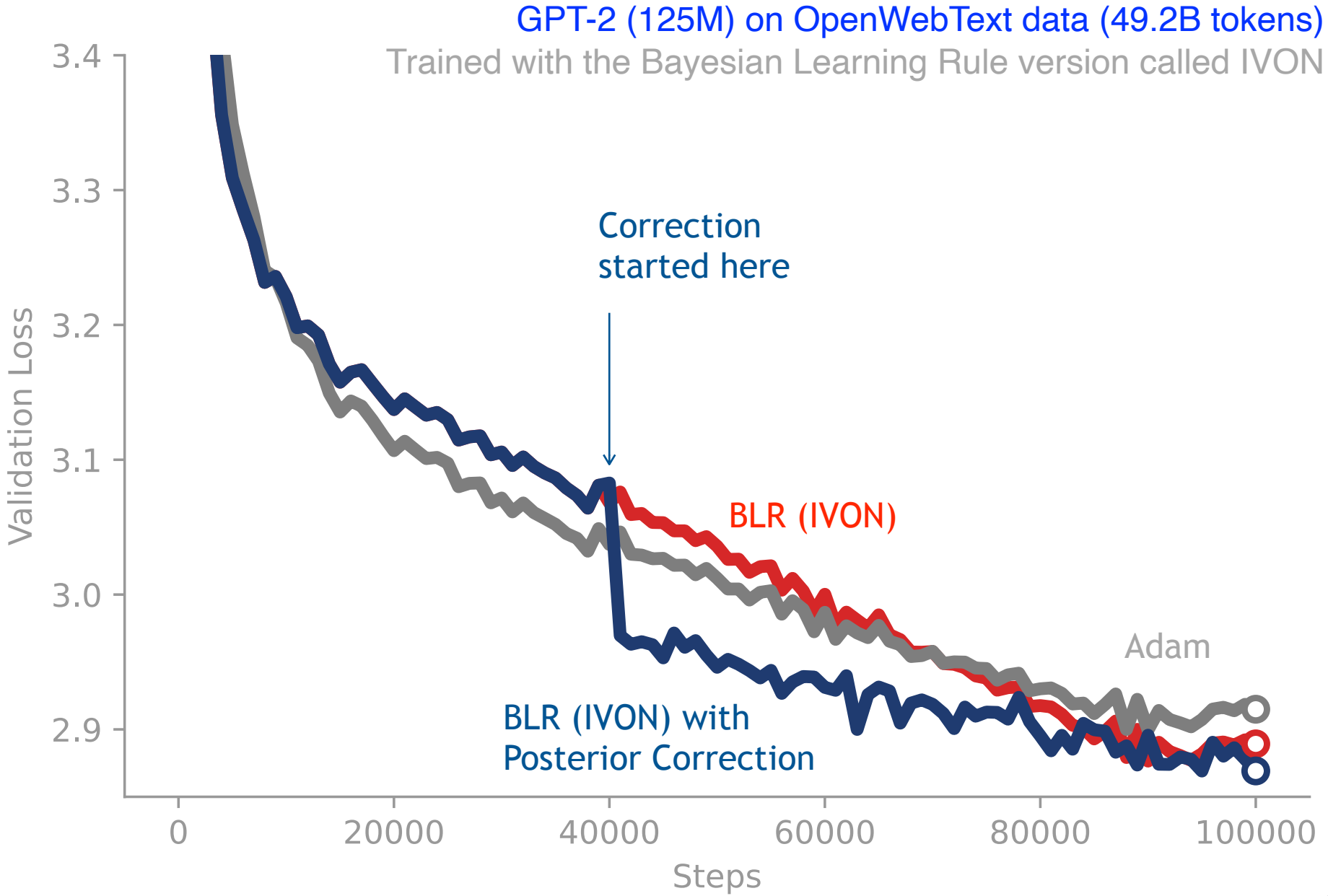
Subarnaduti Paul[1], **Bai Cong**[2,3], **Yohan Jung**[4], **Nico Daheim**[5], **Mohammad Emtiyaz Khan**[2,5,6], **Siddharth Swaroop**[7], **Thomas Möllenhoff**[2], **Martin Mundt**[1]



Abstract

Continual learning remains a major challenge for modern deep networks, partly because commonly used optimizers lack inherent mechanisms for continual adaptation. One such natural mechanism is 'fast and slow adaptation' to balance stability and plasticity. This mechanism has deep roots in neuroscience and biology but there is no consensus how to best incorporate it in commonly used optimizers. Here, we show that this can be easily done via the VCL framework where past posteriors are used as priors in the future. Our key idea is to incorporate slow adaptation via merging of past posteriors to slow down the drift in the knowledge as learning progresses. The merged posterior is then used as the prior in the VCL update to implement the fast-weight updates. These steps can be seamlessly implemented in the IVON optimizer whose form and costs are nearly identical to that of Adam. We call this new optimizer the Continual IVON (CoVON) optimizer and show that it not only consistently improves over existing VCL optimizers, but also performs better than other weight-regularization strategies across domain-incremental learning, continual pre-training, and fine-tuning of large language models.

Posterior Correction can boost LLM training



FEDERATED ADMM FROM BAYESIAN DUALITY

Thomas Möllenhoff[*]
RIKEN Center for AI Project
Tokyo, Japan
thomas.moellenhoff@riken.jp

Siddharth Swaroop[*]
University College London
London, United Kingdom
s.swaroop@ucl.ac.uk

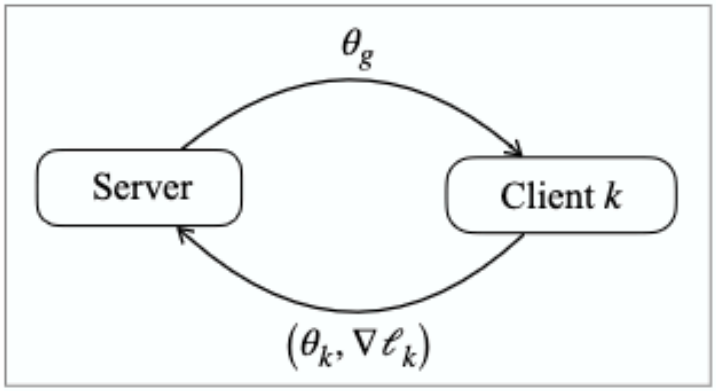
Finale Doshi-Velez
Harvard University
Cambridge, United States
finale@seas.harvard.edu

Mohammad Emtiyaz Khan[†]
RIKEN Center for AI Project
Tokyo, Japan
emtiyaz.khan@riken.jp

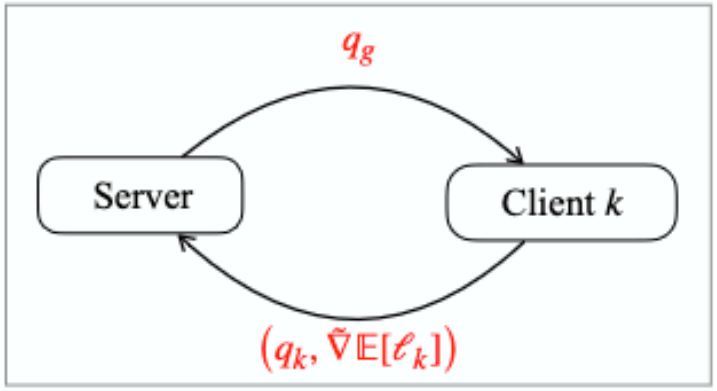
ABSTRACT

We propose a new Bayesian approach to generalize the federated Alternating Direction Method of Multipliers (ADMM). We show that the solutions of variational-Bayesian (VB) objectives are associated with a duality structure that not only resembles the structure of ADMM's fixed-points but also generalizes it. For example, ADMM-like updates are recovered when the VB objective is optimized over the isotropic-Gaussian family, and new non-trivial extensions are obtained for other exponential-family distributions. These extensions include a Newton-like variant that converges in one step on quadratic objectives and an Adam-like variant that yields up to 7% accuracy boosts for deep heterogeneous cases. Our work opens a new Bayesian way to generalize ADMM and other primal-dual methods.

ADMM

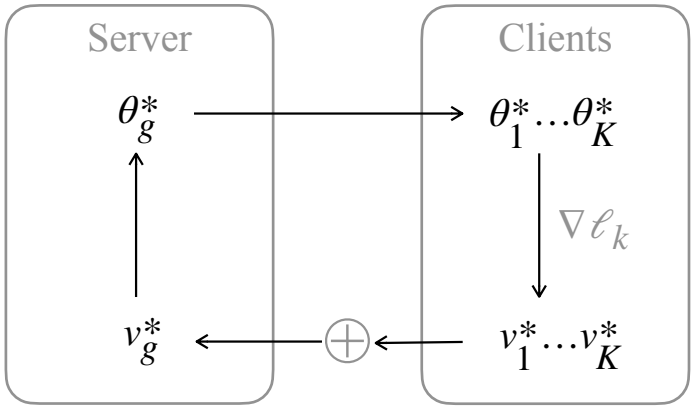


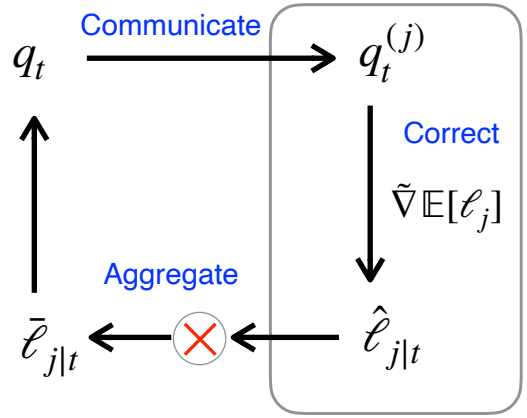
BayesADMM



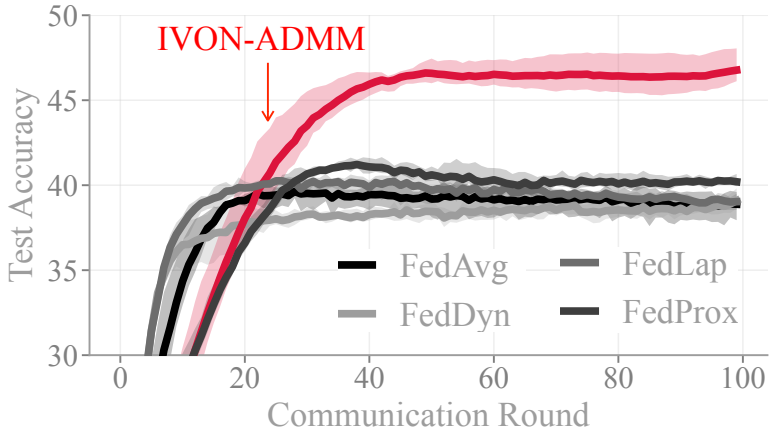
Distributed Training

ADMM's Duality

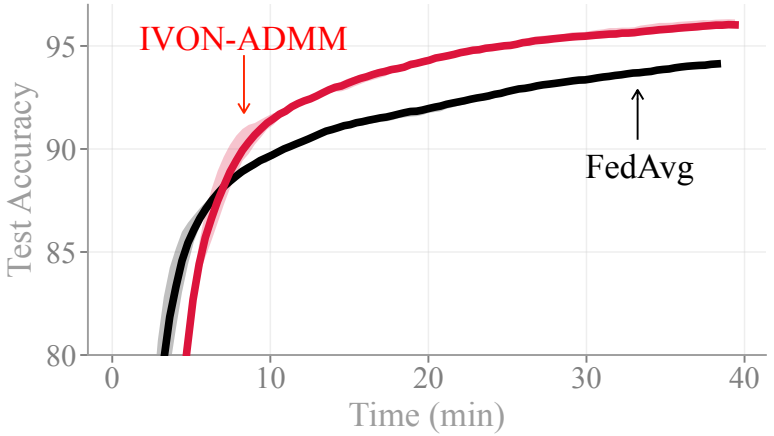




ResNet-20 on CIFAR-100 with 10 clients



MLP on MNIST with 100 clients



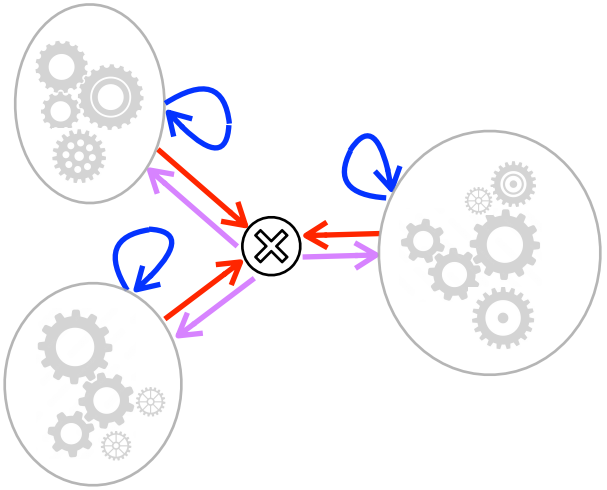
1. Moellenhoff et al. Federated ADMM from Bayesian Duality. ICLR (2026)

A framework for Self-Adaptive Collectives

Bayesian-duality to self-adapt via

1. Communication
2. Aggregation &
3. Correction

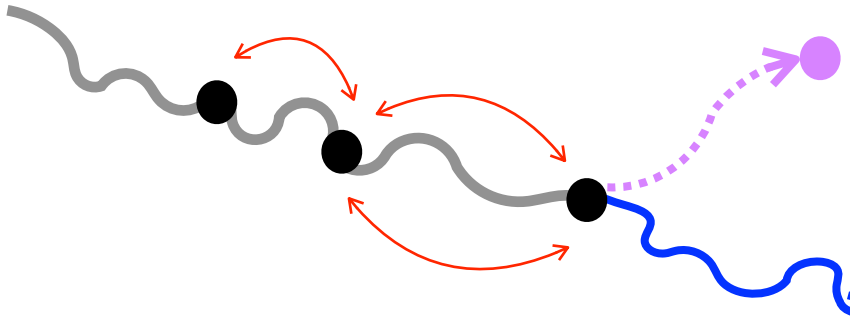
of local beliefs.



Prediction and Control of AI training

I. Communicate with other models

II. Predict the future behavior



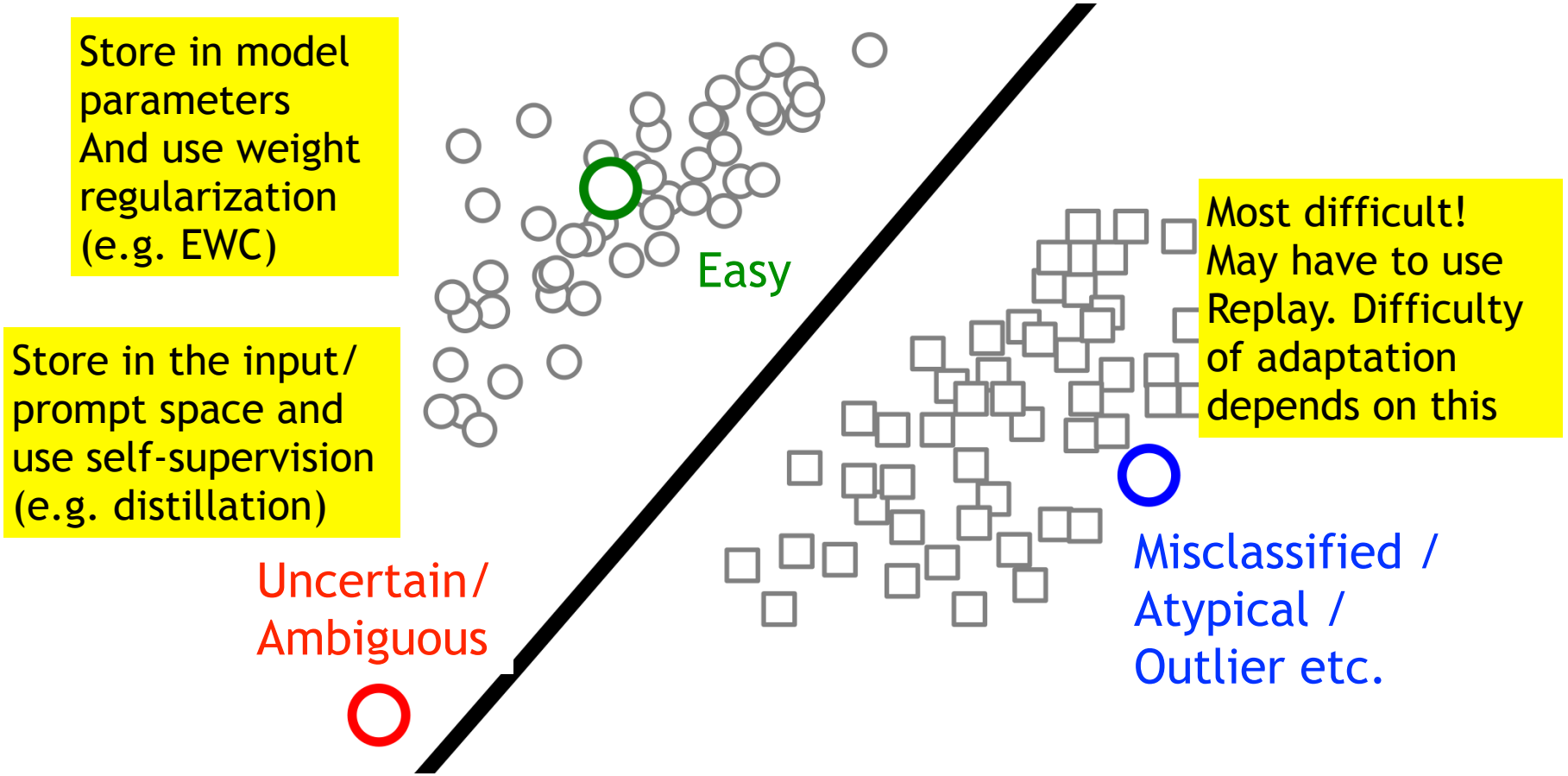
III. Control and change course

How to Build and Update Compact Memory?

This is the core challenge but we have made good progress

How to Represent Past Knowledge

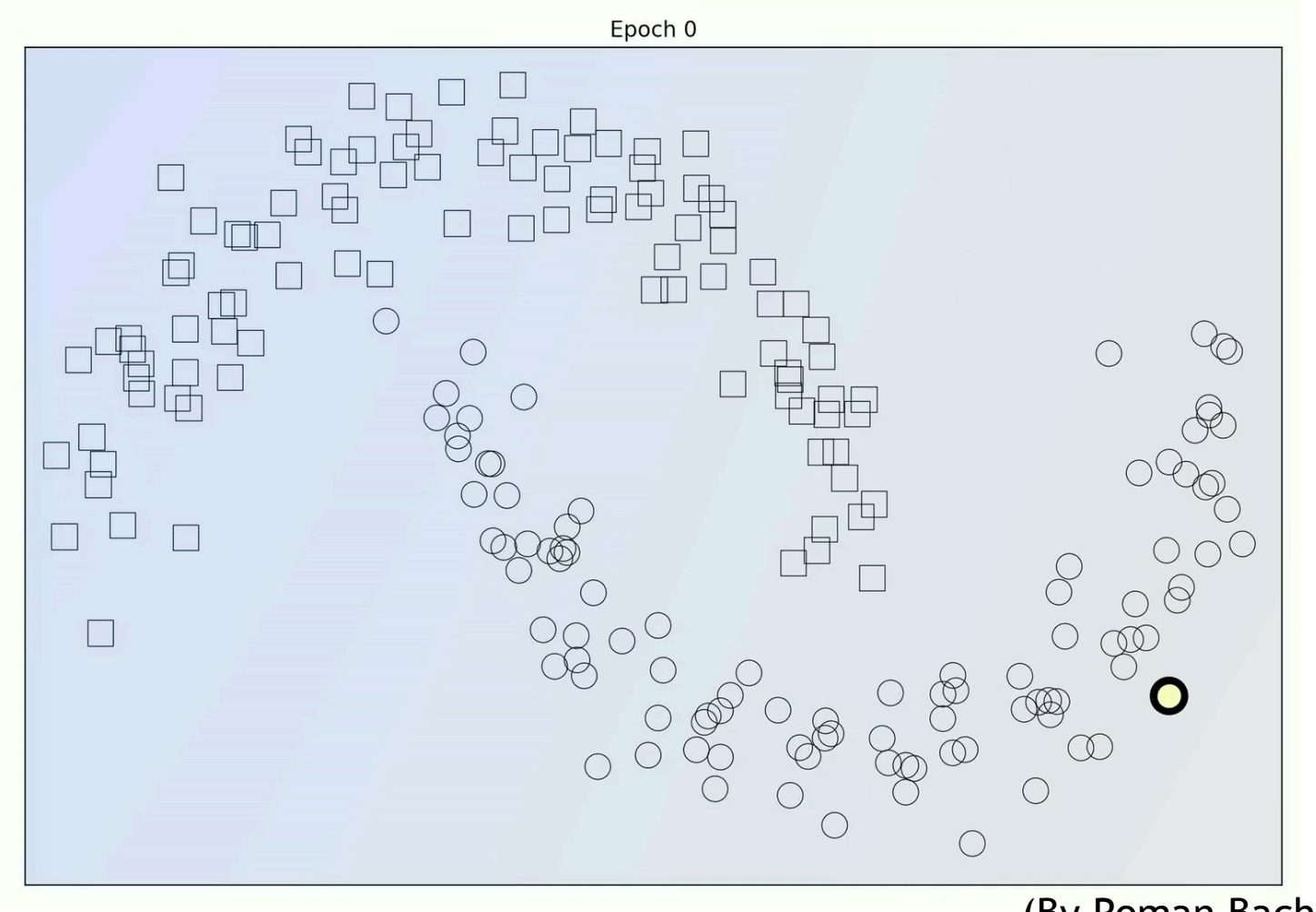
Three kinds of knowledge sources



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

But, it is difficult because “journey Matters, but not just the destination”

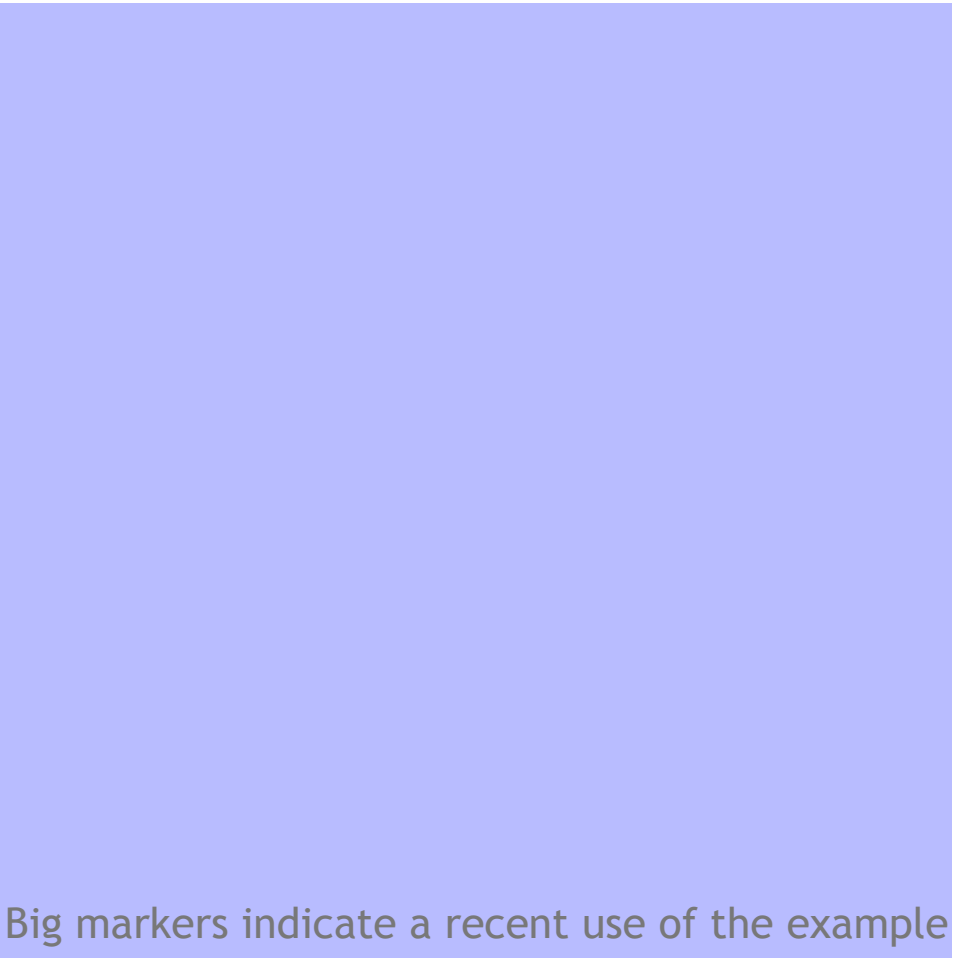
This is a slide from my NeurIPS tutorial in 2019!



(By Roman Bachmann)

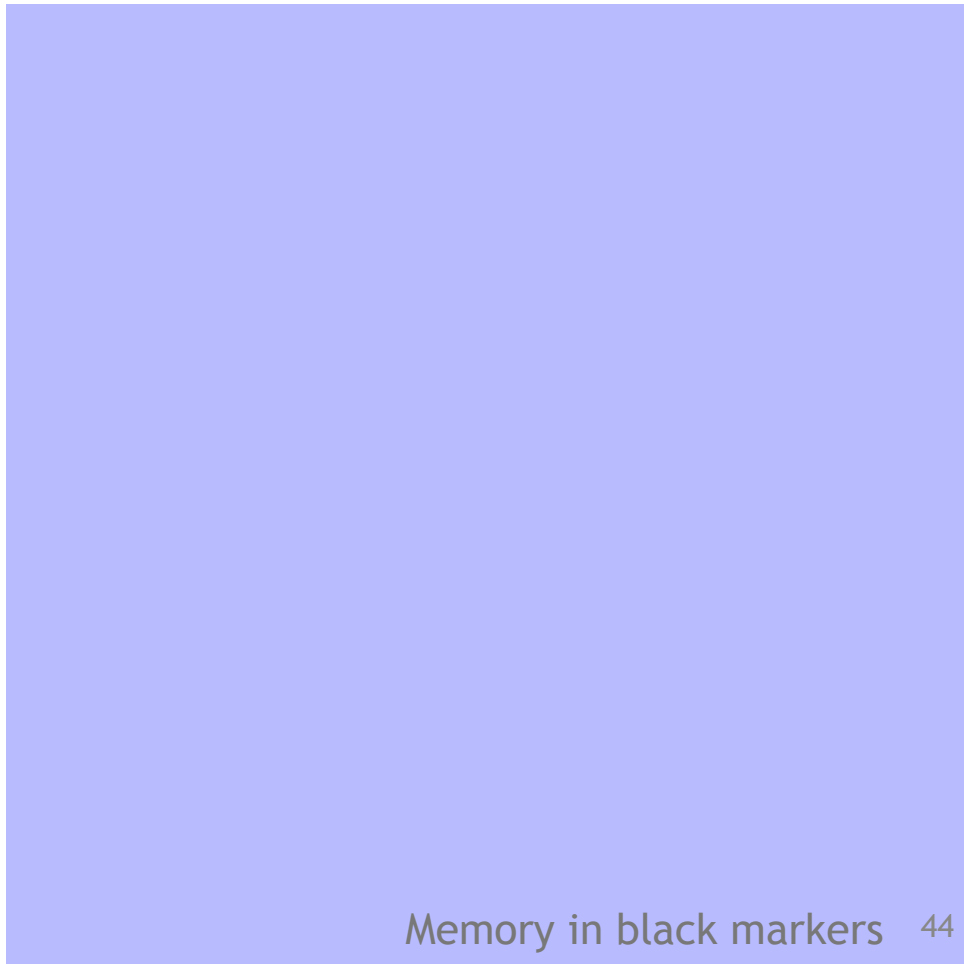
We can use “corrections” as “information gain” to build memory as we train the model

The Adam optimizer



Big markers indicate a recent use of the example

The PoCo optimizer



Memory in black markers

See the discussion in Sec 5 in Khan (2025)

Path to Adaptive Intelligence [2]

- How can we reduce the cost of training AI?
  - What should the algorithm remember and what new experiences it should seek?
  - Build a memory of the past, inject prior knowledge, and design a curriculum to slowly explore the future
- To truly reduce the cost, we also need
  1. Encourage parsimony in data and parameters
  2. Use local learning
  3. Perform active self-guidance
- Towards brain-like learning!
- And also a path towards sustainable & transparent AI

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023)
2. Khan. Knowledge Adaptation as Posterior Correction, arXiv (2025)

Thanks to all of my collaborators over the last 10 years