Presentation at ISBA 2026, July 1, 2026

Adaptive Bayesian Intelligence

The Road to Sustainable AI

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo, Japan

TU Darmstadt and Hessian.AI, Germany

https://emtiyaz.github.io

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

Sustainable AI Training

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

Adaptive Intelligence [1,2]

Challenging for existing AI training due to "catastrophic interference" [3,4]

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

- Developing Adaptive Intelligence via Bayesian Principles
- Part 1: Bayesian Learning Rule [1]
  - Unifies many machine-learning algorithms
  - We use it to improve Deep Learning [2]
- Part 2: Posterior Correction [3]
  - Unifies many knowledge-adaptation methods
  - We use it to improve Continual learning [4], Variance reduction [5], and Distributed optimization [6]
- Bayes for the next-generation adaptive intelligence

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023)
2. Shen et al. Variational Learning is Effective for Large Deep Networks, ICML (2024)
3. Khan. Knowledge Adaptation as Posterior Correction, arXiv (2025)
4. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS (2021)
5. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)
6. Moellenhoff et al. Federated ADMM from Bayesian Duality. ICLR (2026)

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)

Deep Learning

Empirical Risk Minimization: $\theta_t \leftarrow \arg\min_{\theta} \sum_{j=0}^{t} \ell_j(\theta)$ Loss

SGD: $\theta \leftarrow \theta - \rho \nabla \ell_j(\theta)$

Adam [1,2]: $\theta \leftarrow \theta - \rho \, \mathrm{diag}(\hat{h})^{-1/2} \nabla \ell_j(\theta)$

$\hat{h} \leftarrow (1-\eta)\hat{h} + \eta [\nabla \ell_j(\theta)]^2$

Scales to billions of parameters and data examples, and works extremely well. Are there Bayesian versions that are equally scalable and accurate?

1. Tieleman and Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning 4 (2012)
2. Adam: A method for stochastic optimization, ICLR, 2014 (citation >247,000)

Variational Bayesian Posteriors

We can find such generalizations by introducing "posterior approximations" through variational Bayes.

Bayes for ERM [1]

$$p_t \propto p_0 \prod_{j=1}^{t} \exp(-\ell_j)$$

Gibbs Variational principle [2-4]

$$p_t = \arg\min_{q \in \mathcal{P}} \sum_{j=1}^{t} \mathbb{E}_q[\ell_j] + KL(q \| p_0)$$

Variational Bayesian (VB) learning

$$q_t = \arg\min_{q \in \mathcal{Q}} \sum_{j=1}^{t} \mathbb{E}_q[\ell_j] + KL(q \| p_0)$$

* This is not pure variational "inference" rather ERM-like "learning" variant

1. Zhang, Theoretical analysis of a class of randomized regularization methods, COLT (1999)
2. Donsker & Varadhan, Asymptotic evaluation of certain Markov process expectations for large time (1976-83)
3. Williams, Bayesian conditionalisation and the principle of minimum information (1980)
4. Zellner, Optimal Information Processing and Bayes' Theorem. The American Statistician (1988)

The Structure of VB Posteriors [1,2]

A VB solution q_t has an equivalent representation in terms of loss-surrogates (called sites) denoted by $\hat{\ell}_{j|t}$ [1]

$$p_t \propto p_0 \prod_{j=1}^{t} \exp(-\ell_j)$$

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

Posterior Sites

1. Khan and Nielsen. Fast yet simple natural-gradient descent for VI..,ISITA (2018)
2. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
3. Khan and Lin. Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations, ICML (2019)
4. Kiral et al. Lie-Group Bayesian Learning Rule, AISTATS (2023)

The Structure of VB Posteriors [1,2]

A VB solution q_t has an equivalent representation in terms of loss-surrogates (called sites) denoted by $\hat{\ell}_{j|t}$ [1]

$$p_t \propto p_0 \prod_{j=1}^{t} \exp(-\ell_j)$$

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

Posterior Sites

Exp-Fam: $q_t \propto \exp\left[T(\theta)^\top \lambda_t\right]$

Sufficient statistics Natural parameter

1. Khan and Nielsen. Fast yet simple natural-gradient descent for VI..,ISITA (2018)
2. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
3. Khan and Lin. Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations, ICML (2019)
4. Kiral et al. Lie-Group Bayesian Learning Rule, AISTATS (2023)

The Structure of VB Posteriors [1,2]

A VB solution q_t has an equivalent representation in terms of loss-surrogates (called sites) denoted by $\hat{\ell}_{j|t}$ [1]

$$p_t \propto p_0 \prod_{j=1}^{t} \exp(-\ell_j)$$

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

Posterior Sites

Exp-Fam: $q_t \propto \exp\left[T(\theta)^\top \lambda_t\right]$

Sufficient statistics Natural parameter

Sites: $\hat{\ell}_{j|t} = T(\theta)^\top \tilde{\nabla}_{\lambda_t} \mathbb{E}_{q_t}[\ell_j]$

Natural gradient

1. Khan and Nielsen. Fast yet simple natural-gradient descent for VI..,ISITA (2018)
2. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
3. Khan and Lin. Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations, ICML (2019)
4. Kiral et al. Lie-Group Bayesian Learning Rule, AISTATS (2023)

The Structure of VB Posteriors [1,2]

A VB solution q_t has an equivalent representation in terms of loss-surrogates (called sites) denoted by $\hat{\ell}_{j|t}$ [1]

$$p_t \propto p_0 \prod_{j=1}^{t} \exp(-\ell_j)$$

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

Posterior Sites

Exp-Fam: $q_t \propto \exp\left[T(\theta)^\top \lambda_t\right]$

Sufficient statistics Natural parameter

Sites: $\hat{\ell}_{j|t} = T(\theta)^\top \tilde{\nabla}_{\lambda_t} \mathbb{E}_{q_t}[\ell_j]$

Natural gradient

This result stems from the first-order optimality and extends to many cases [3,4] (but not always as elegant)

1. Khan and Nielsen. Fast yet simple natural-gradient descent for VI..,ISITA (2018)
2. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
3. Khan and Lin. Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations, ICML (2019)
4. Kiral et al. Lie-Group Bayesian Learning Rule, AISTATS (2023)

Bayes' vs Taylor's Approximations

For Gaussian posteriors, the sites closely resemble Taylor's approximations [1,2]

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

Posterior Sites

1. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
2. Khan. Knowledge adaptation as posterior correction. ArXiv (2025)

Bayes' vs Taylor's Approximations

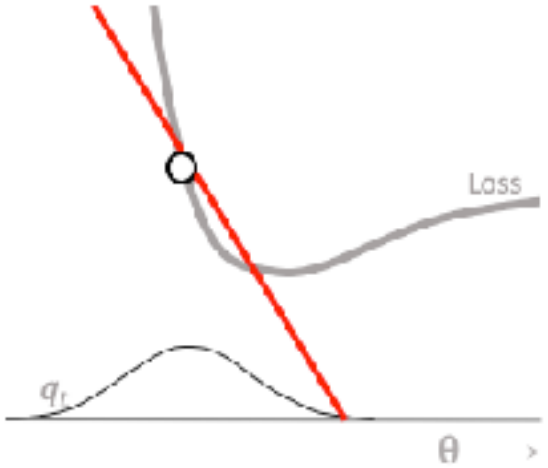
For Gaussian posteriors, the sites closely resemble Taylor's approximations [1,2]

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

Posterior — Sites

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

iso-Gauss — Gradients



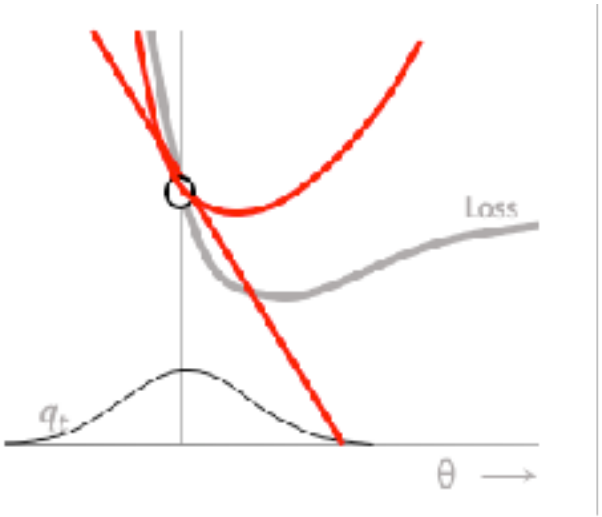
1. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
2. Khan. Knowledge adaptation as posterior correction. ArXiv (2025)

Bayes' vs Taylor's Approximations

For Gaussian posteriors, the sites closely resemble Taylor's approximations [1,2]

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

Posterior — Sites



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

iso-Gauss — Gradients

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

full-Gauss — Gradients — Hessians

1. Khan. Information Geometry of Variational Bayes, Information Geometry Journal (2025)
2. Khan. Knowledge adaptation as posterior correction. ArXiv (2025)

The Bayesian Learning Rule (BLR)

The structure suggests a natural algorithm to estimate VB posteriors, similarly to SGD or Adam.

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

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

The Bayesian Learning Rule (BLR)

The structure suggests a natural algorithm to estimate VB posteriors, similarly to SGD or Adam.

$$q_t = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t}) \qquad \text{Initialize } q \text{ and iterate}$$

$$\Longrightarrow \quad q \leftarrow q^{1-\rho} \left[p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_j) \right]^{\rho} \qquad \text{(Learning rate } \rho\text{)}$$

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

The Bayesian Learning Rule (BLR)

The structure suggests a natural algorithm to estimate VB posteriors, similarly to SGD or Adam.

$$q_t = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t}) \qquad \text{Initialize } q \text{ and iterate}$$

$$\Longrightarrow \quad q \leftarrow q^{1-\rho} \left[p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_j) \right]^{\rho} \qquad \text{(Learning rate } \rho\text{)}$$

We call this the Bayesian "Learning" Rule (which is also a proper natural-gradient/mirror descent algorithm)

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

The Bayesian Learning Rule (BLR)

The structure suggests a natural algorithm to estimate VB posteriors, similarly to SGD or Adam.

$$q_t = \frac{1}{\hat{Z}_t} p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t}) \qquad \text{Initialize } q \text{ and iterate}$$

$$\Longrightarrow \quad q \leftarrow q^{1-\rho} \left[p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_j) \right]^{\rho} \qquad \text{(Learning rate } \rho\text{)}$$

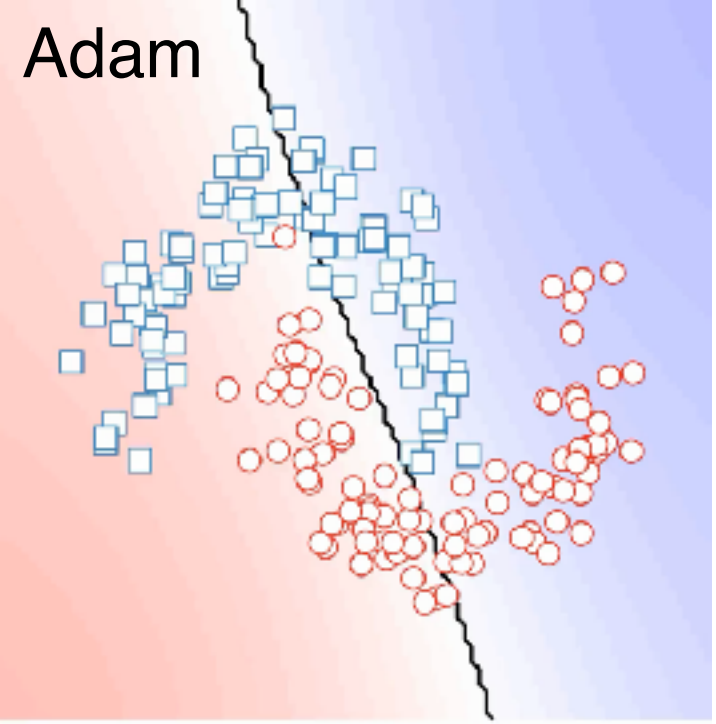
We call this the Bayesian "Learning" Rule (which is also a proper natural-gradient/mirror descent algorithm)

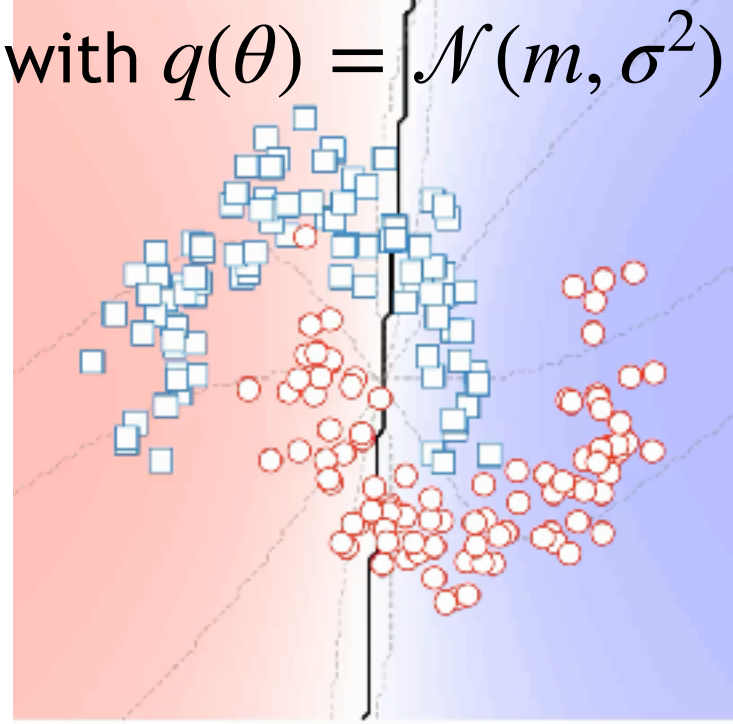
$$q \leftarrow q^{1-\rho} \exp(-\hat{\ell}_j)^{\rho} \qquad \text{(mini-batch of size 1 and } p_0 \propto \exp(-\ell_0)\text{)}$$

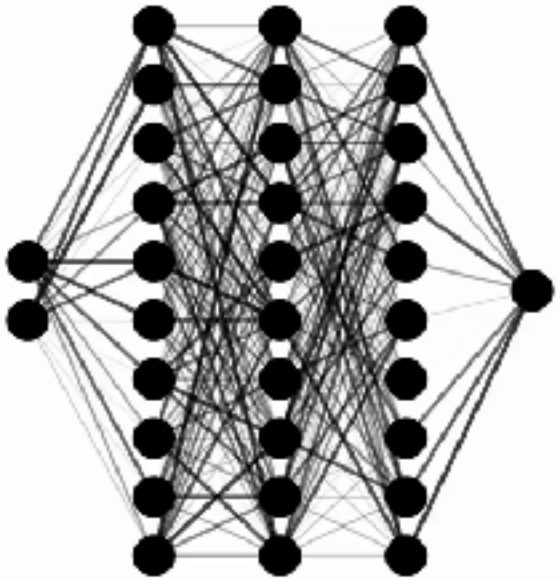
1. Khan and Rue. The Bayesian Learning Rule. JMLR (2023)
2. Khan and Lin. Conjugate-Compute Variational Inference..... AISTATS (2017)

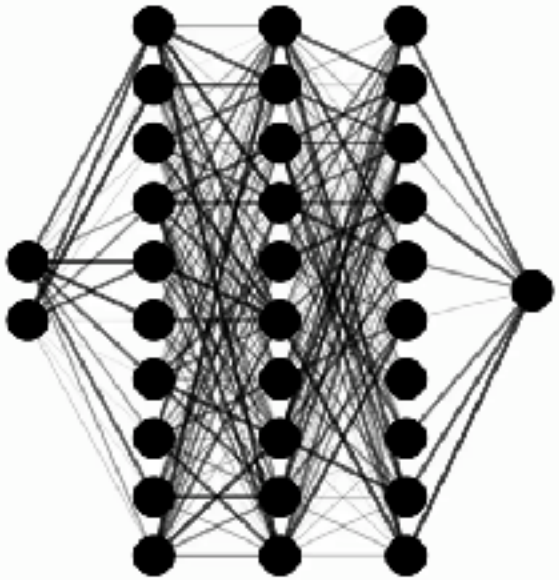
Adam

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









Adam from the BLR

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

RMSprop/Adam

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

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)

Adam from the BLR

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

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)

IVON won 1st prize in NeurIPS 2021 Approximate Inference Challenge

Watch Thomas Moellenhoff's talk at https://www.youtube.com/watch?v=LQInlN5EU7E.

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

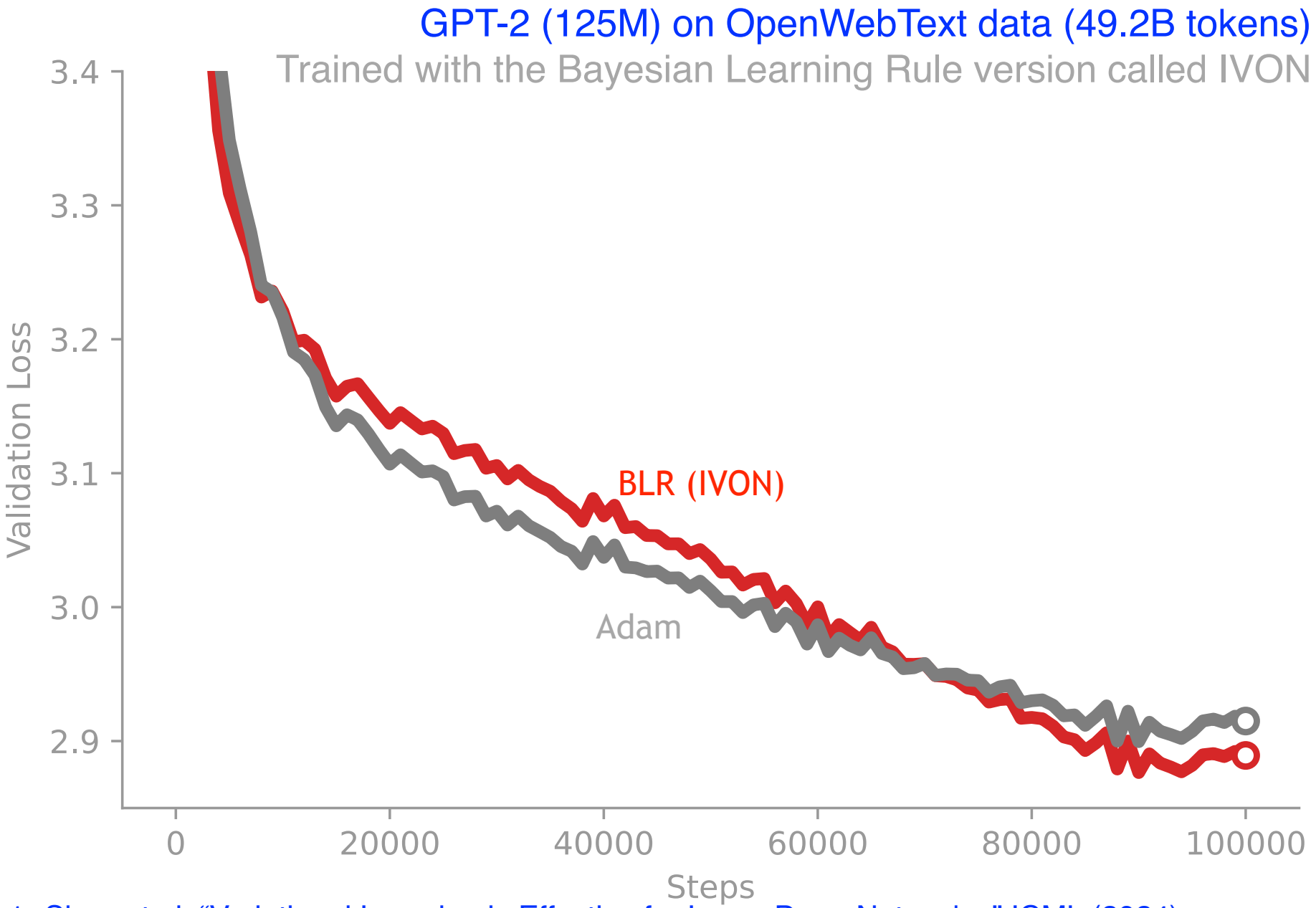
Thomas Möllenhoff[1], Yuesong Shen[2], Gian Maria Marconi[1]
Peter Nickl[1], Mohammad Emtiyaz Khan[1]

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

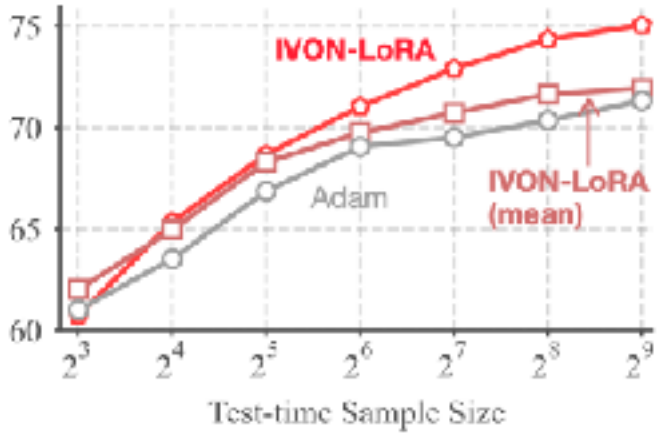
Bayesian Learning Rule is as cheap and accurate as Adam



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

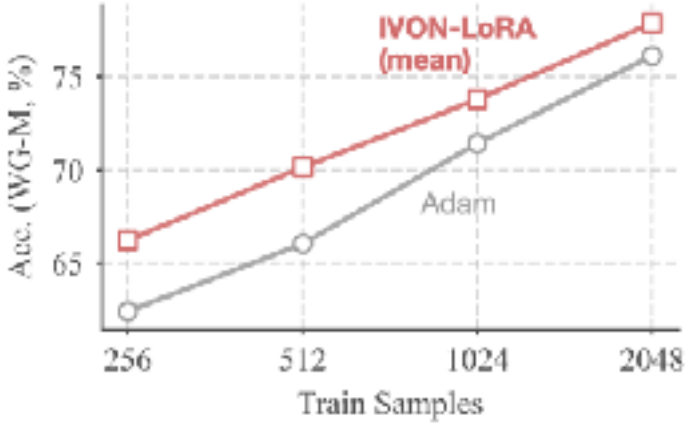
IVON Improves LoRA Finetuning (7B)

Posterior averages improve accuracy

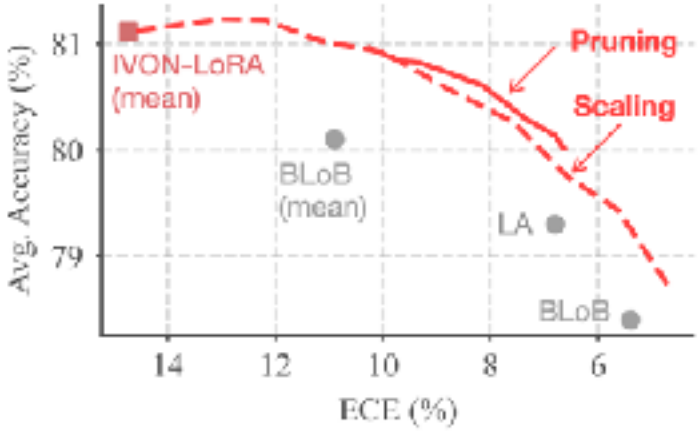


Train longer, generalize better

Get more out of less (data)



Better acc-calibration trade-off



1. Cong et al. Improving LoRA with Variational Learning, ArXiv (2025).

IVON reduces Hallucinations

LLM decoding: I LOVE BAYES' _____

Standard approach: $\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathcal{V}^*} p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$

IVON-based: $\mathbf{y}^{\Theta} = \arg\max_{\mathbf{y}' \in \mathcal{V}^*} \sum_{\mathbf{y} \in \mathcal{V}^*} \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim q} [p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})] u(\mathbf{y}, \mathbf{y}')$

Utility

Posterior sampling

Posterior averaging improves the standard approach

IWSLT17 En-De			WMT18 Tr-En			XSUM			SAMSum		E2E NLG		STS-B
BLEU	COMET	LaBSE	BLEU	COMET	LaBSE	R-1	R-L	FactCC	R-1	R-L	R-1	R-L	RMSE
19.93	76.62	83.47	14.75	78.20	76.02	**33.63**	25.67	27.50	46.47	36.21	67.88	44.41	0.330
19.73	76.60	83.51	15.27	78.44	77.12	33.04	25.19	23.56	46.17	35.98	68.74	45.16	0.284
20.89	77.42	84.01	15.66	79.01	**77.79**	33.39	**25.73**	26.07	46.40	36.51	69.36	45.57	0.271
21.24	**77.94**	**84.20**	15.63	79.01	77.60	33.37	25.68	27.40	**46.71**	**36.87**	**69.56**	**45.77**	**0.269**

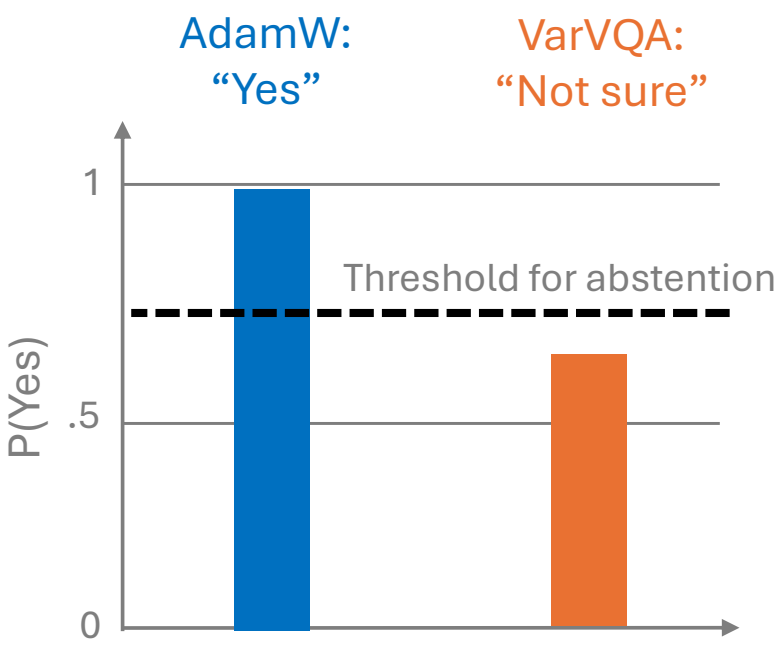
1. Daheim et al. Uncertainty-Aware Decoding with Minimum Bayes Risk, *ICLR (*2025)

Variational Visual-Question Answering

Abstain from answering when unsure (by using IVON)

Question: Does the pedestrian light say walk?

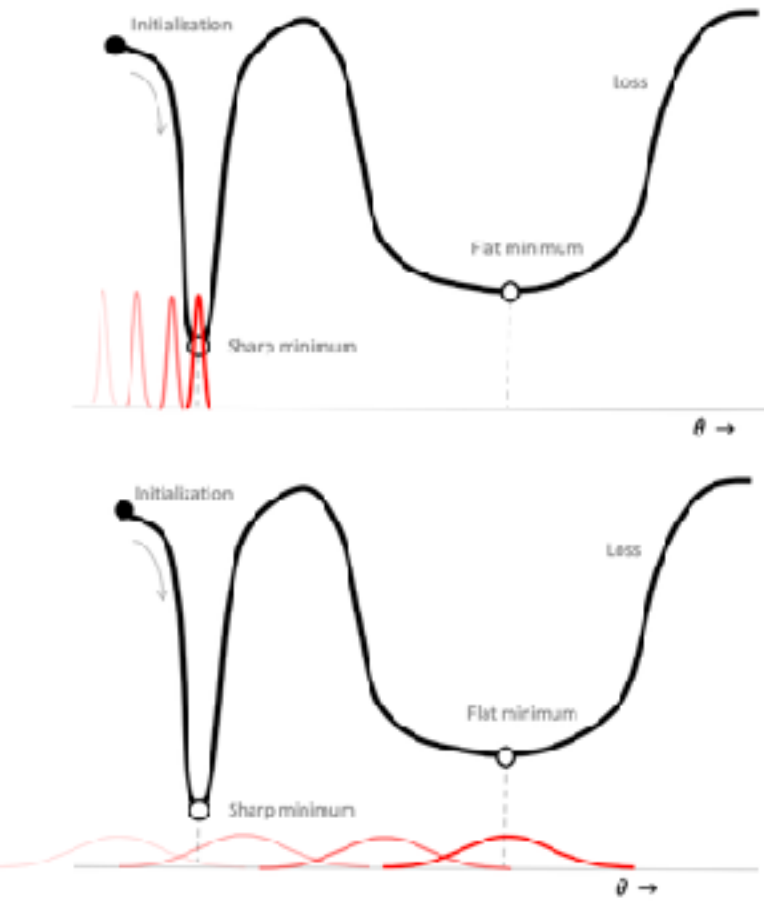
Correct answer: “No”

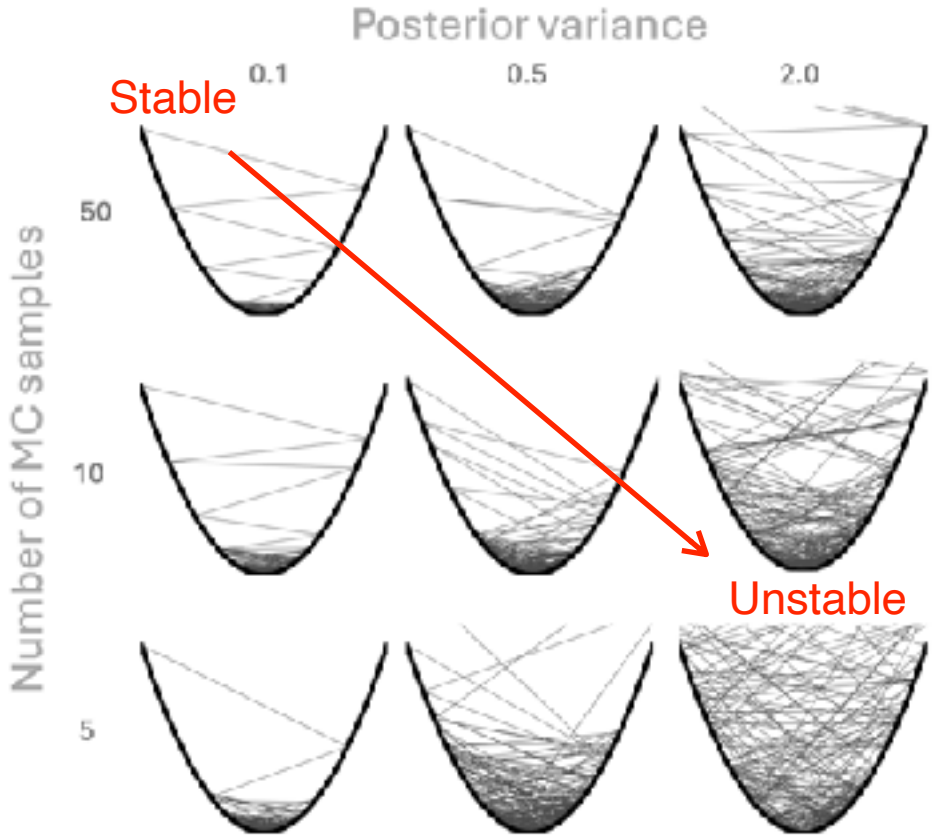


1. Wieczorek et al. Variational Visual Question Answering for Uncertainty-Aware Selective Prediction, *TMLR* (2025)

IVON Finds More Flat Solutions than GD at the Edge of Stability (EoS)

By tweaking posterior variance, we can find flatter solutions

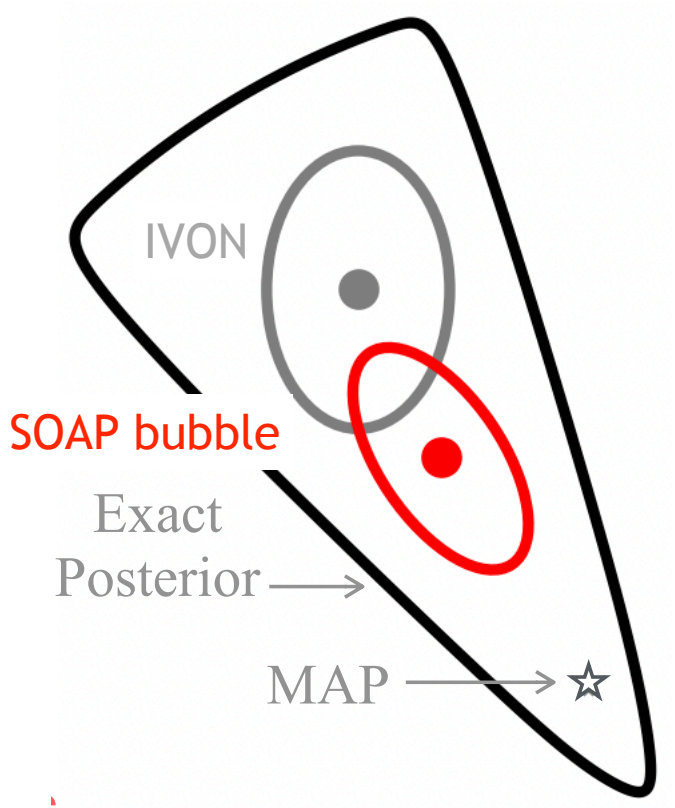


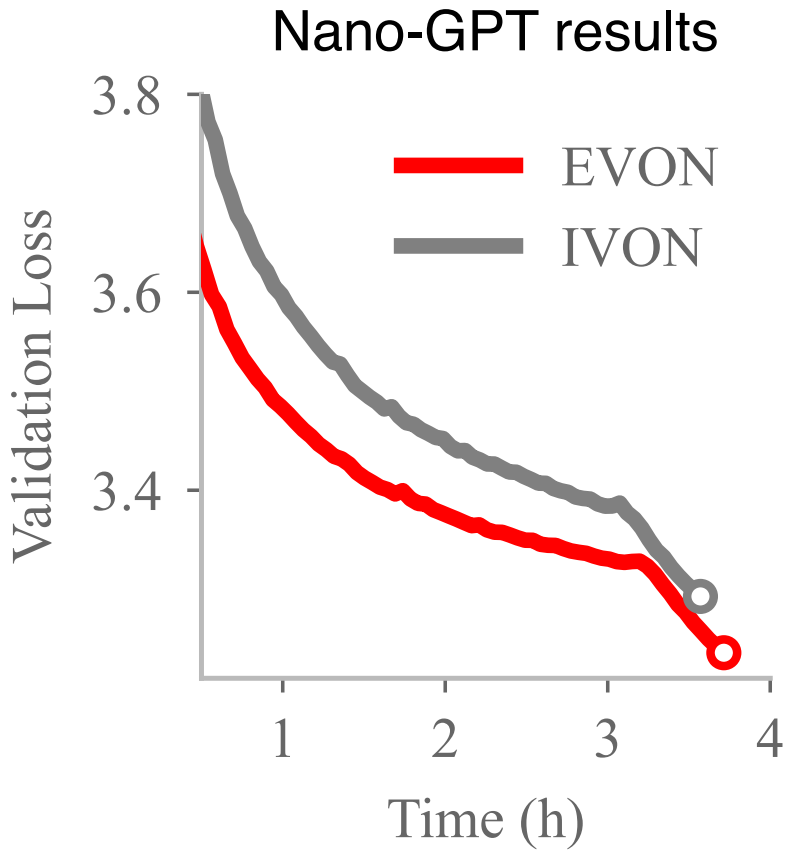


1. Ghosh et al. Variational Learning finds Flatter Solutions at the Edge of Stability, NeurIPS (2025)

SOAP-Bubbles ○❍
Structured Weight Uncertainty for Neural Networks

Get structured covariances for free by using SOAP

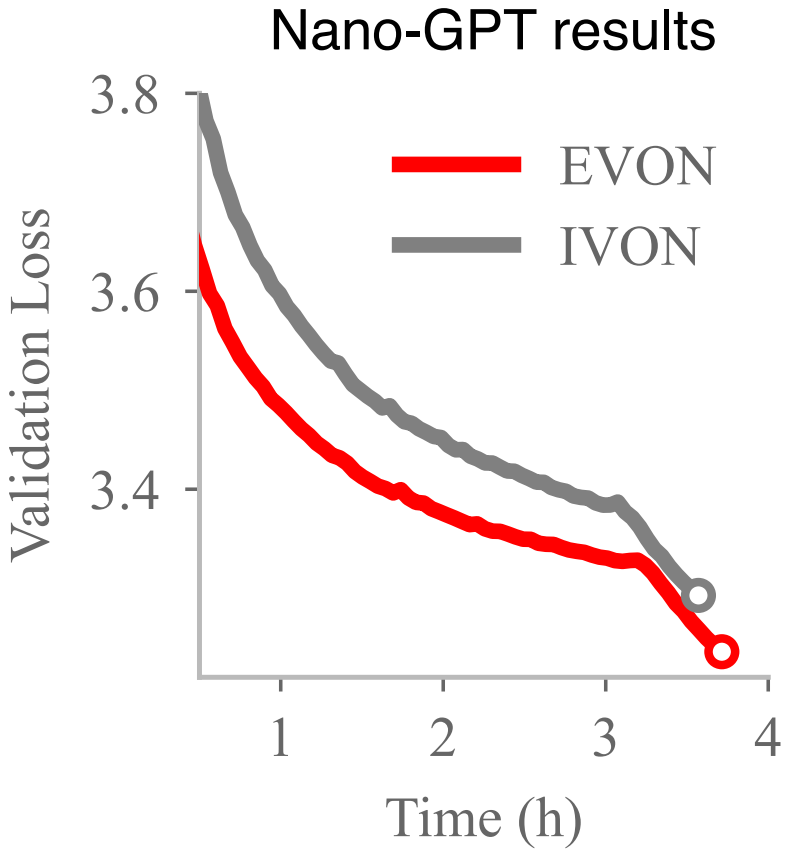




1. Minut et al. ,SOAP-Bubbles ○❍ Structured Weight Uncertainty for Neural Networks, arXiv (2026)

SOAP-Bubbles ○❍
Structured Weight Uncertainty for Neural Networks

Get structured covariances for free by using SOAP



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

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)

Adaptive Intelligence

Challenging due to “catastrophic interference” [1,2]. We address it by exploiting the structure of the VB posteriors to adapt quickly

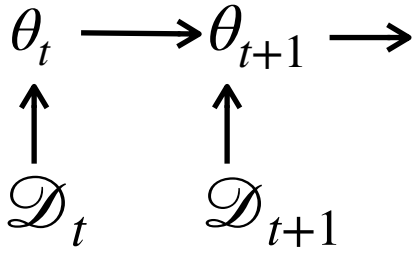
1. Sutton. *Two Problems with Backpropagation and Other Steepest-Descent Learning…*, Cog. Sci. Society (1986)
2. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.

The Goal of Adaptive Intelligence

Update knowledge "quickly". For instance, in continual learning, we want to update the model with new data

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

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



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

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

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

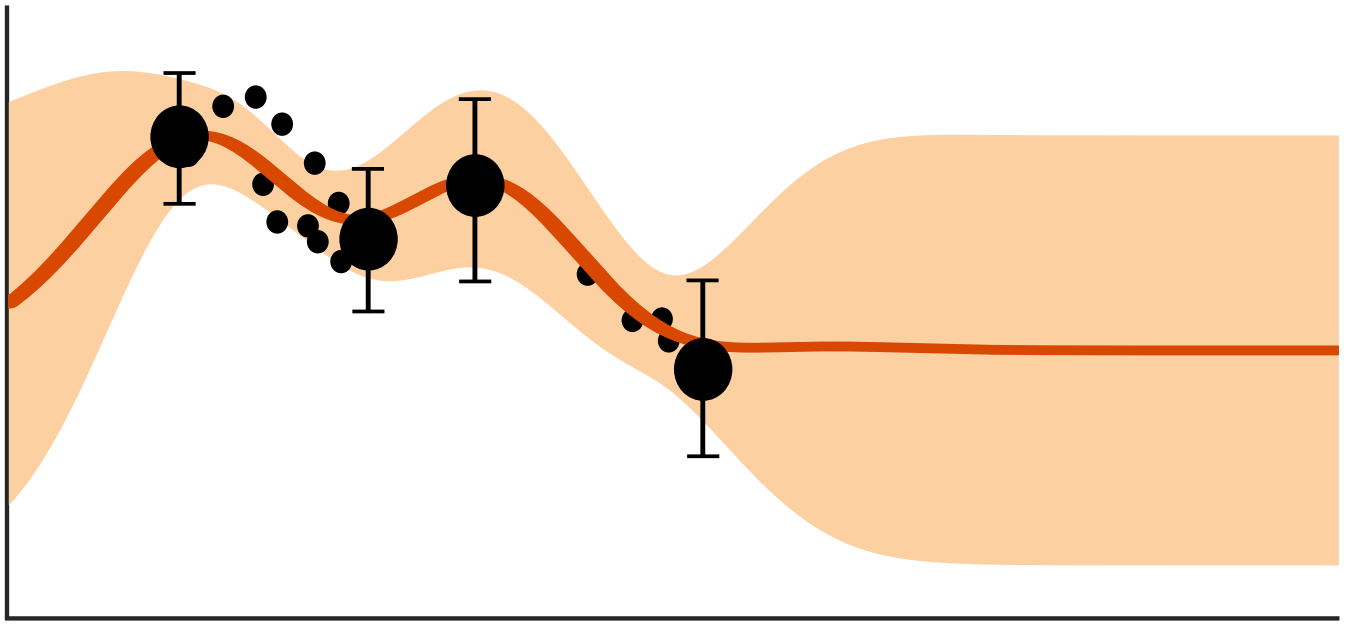
1. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.

Adaptation via Bayes

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

Recursive Update

What if we do not have p_t, but approximate q_t? Can we still control the accumulation of error at t increases?

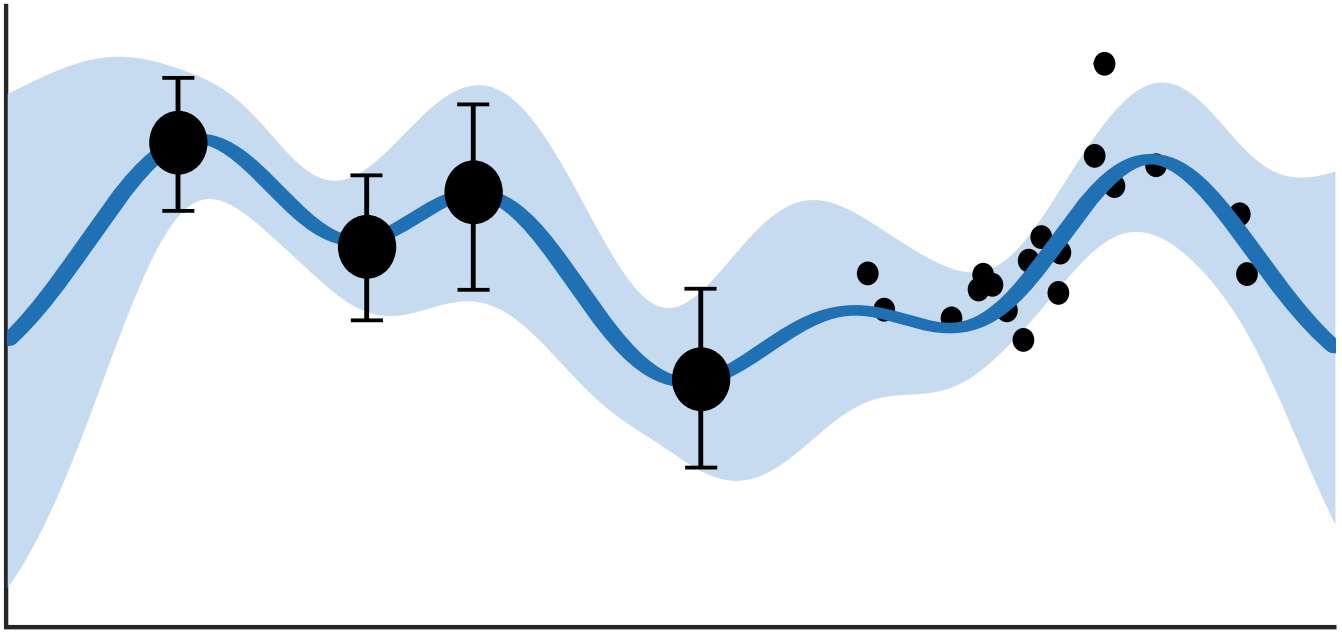


Adaptation via Bayes

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

Recursive Update

What if we do not have p_t, but approximate q_t? Can we still control the accumulation of error at t increases?



Eq. 4 in Khan (2025)

Posterior Correction for CL [1]

Quantify the error by writing q_{t+1} in terms of q_t

Batch: $$q_{t+1} = \arg\min \mathbb{E}_q[\ell_{t+1}] + \sum_{i=1}^{t} \mathbb{E}_q[\ell_i] + KL(q\|p_0)$$

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Nguyen et al. Variational Continual Learning, ICLR (2018)

Eq. 4 in Khan (2025)

Posterior Correction for CL [1]

Quantify the error by writing q_{t+1} in terms of q_t

Batch: $$q_{t+1} = \arg\min \mathbb{E}_q[\ell_{t+1}] + \sum_{i=1}^{t} \mathbb{E}_q[\ell_i] + KL(q\|p_0)$$

$$q_{t+1} \propto p_0 \prod_{j=1}^{t+1} \exp(-\hat{\ell}_{j|t+1})$$

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Nguyen et al. Variational Continual Learning, ICLR (2018)

Eq. 4 in Khan (2025)

Posterior Correction for CL [1]

Quantify the error by writing q_{t+1} in terms of q_t

Batch: $$q_{t+1} = \arg\min \mathbb{E}_q[\ell_{t+1}] + \sum_{i=1}^{t} \mathbb{E}_q[\ell_i] + KL(q\|p_0)$$

$$q_{t+1} \propto p_0 \prod_{j=1}^{t+1} \exp(-\hat{\ell}_{j|t+1}) \qquad q_t \propto p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})$$

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Nguyen et al. Variational Continual Learning, ICLR (2018)

Eq. 4 in Khan (2025)

Posterior Correction for CL [1]

Quantify the error by writing q_{t+1} in terms of q_t

Batch: $$q_{t+1} = \arg\min \mathbb{E}_q[\ell_{t+1}] + \sum_{i=1}^{t} \mathbb{E}_q[\ell_i] + KL(q\|p_0)$$

$$q_{t+1} \propto p_0 \prod_{j=1}^{t+1} \exp(-\hat{\ell}_{j|t+1}) \qquad q_t \propto p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})$$

$$q_{t+1} \propto q_t \times \frac{p_0 \prod_{j=1}^{t+1} \exp(-\hat{\ell}_{j|t+1})}{p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})}$$

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Nguyen et al. Variational Continual Learning, ICLR (2018)

Eq. 4 in Khan (2025)

Posterior Correction for CL [1]

Quantify the error by writing q_{t+1} in terms of q_t

Batch: $$q_{t+1} = \arg\min \mathbb{E}_q[\ell_{t+1}] + \sum_{i=1}^{t} \mathbb{E}_q[\ell_i] + KL(q\|p_0)$$

$$q_{t+1} \propto p_0 \prod_{j=1}^{t+1} \exp(-\hat{\ell}_{j|t+1}) \qquad q_t \propto p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})$$

$$q_{t+1} \propto q_t \exp(-\hat{\ell}_{t+1|t+1}) \prod_{j=1}^{t} \exp\Big[-\Big(\hat{\ell}_{j|t+1} - \hat{\ell}_{j|t}\Big)\Big]$$

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Nguyen et al. Variational Continual Learning, ICLR (2018)

Eq. 4 in Khan (2025)

Posterior Correction for CL [1]

Quantify the error by writing q_{t+1} in terms of q_t

Batch: $$q_{t+1} = \arg\min \mathbb{E}_q[\ell_{t+1}] + \sum_{i=1}^{t} \mathbb{E}_q[\ell_i] + KL(q\|p_0)$$

$$q_{t+1} \propto p_0 \prod_{j=1}^{t+1} \exp(-\hat{\ell}_{j|t+1}) \qquad q_t \propto p_0 \prod_{j=1}^{t} \exp(-\hat{\ell}_{j|t})$$

$$q_{t+1} \propto q_t \exp(-\hat{\ell}_{t+1|t+1}) \prod_{j=1}^{t} \exp\Big[-\Big(\hat{\ell}_{j|t+1} - \hat{\ell}_{j|t}\Big)\Big]$$

$$q_{t+1} = \arg\min \underbrace{\mathbb{E}_q[\ell_{t+1}] + KL(q\|q_t)}_{VCL\,[2]} + \sum_{j=1}^{t} \underbrace{\mathbb{E}_q[\ell_j - \hat{\ell}_{j|t}]}_{correction}$$

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)
2. Nguyen et al. Variational Continual Learning, ICLR (2018)

Eq. 4 in Khan (2025)

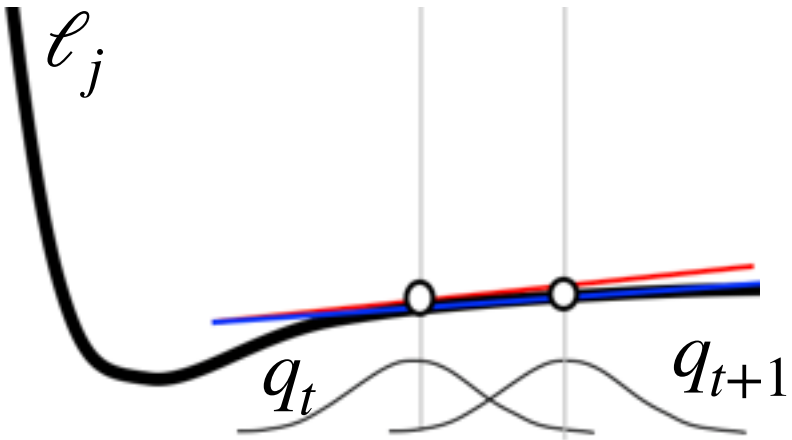
The Correction Term

Correction is a mathematical quantification of the "interference" between the past and the future.

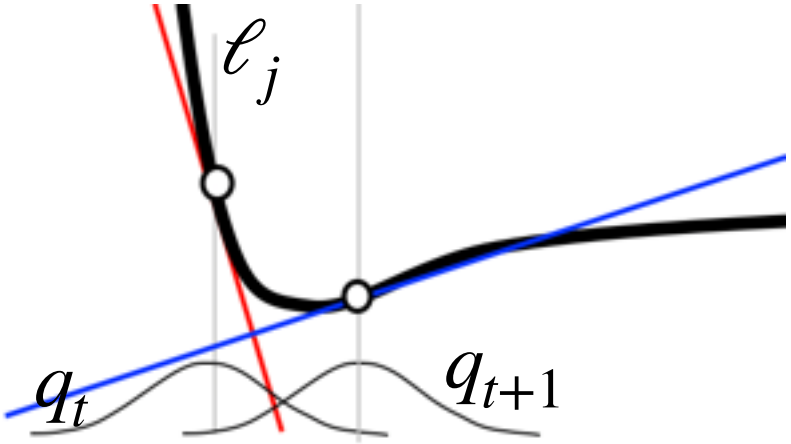
$$\hat{\ell}_{j|t+1} - \hat{\ell}_{j|t}$$

E.g., for isotropic Gaussian q, the sites are built with gradients and therefore correction = gradient mismatch

Small interference

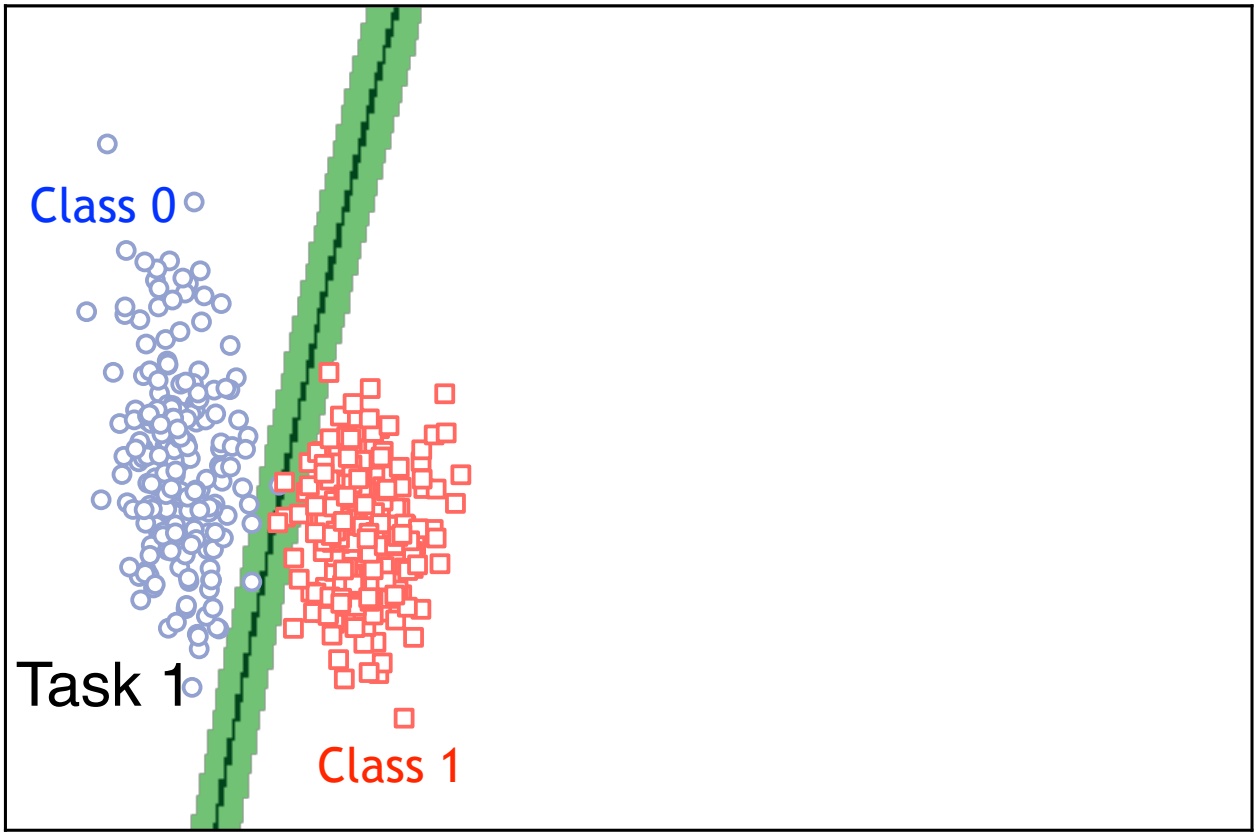


Large interference



CL with “Memorable” Past [1]

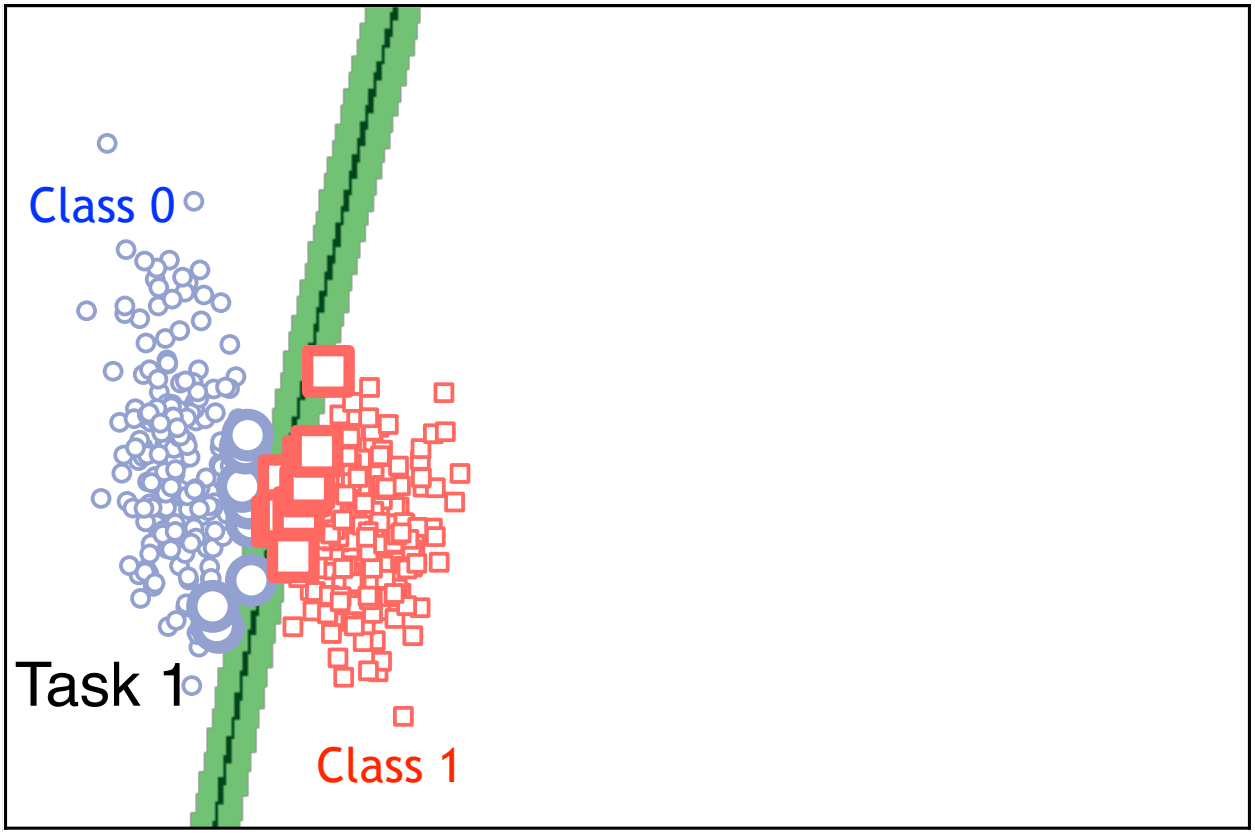
Choose memories where interference is more likely and add corrections to the variational objective.



1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

CL with “Memorable” Past [1]

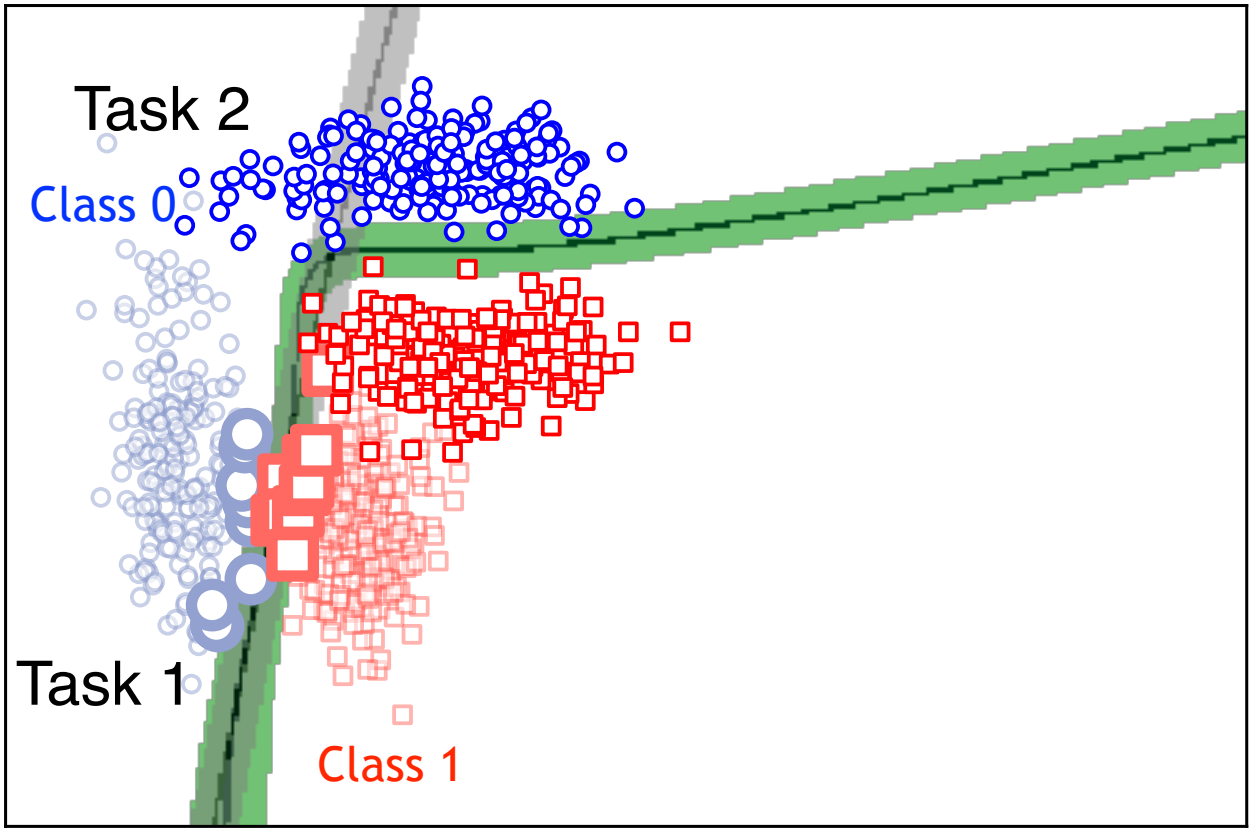
Choose memories where interference is more likely and add corrections to the variational objective.



1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

CL with “Memorable” Past [1]

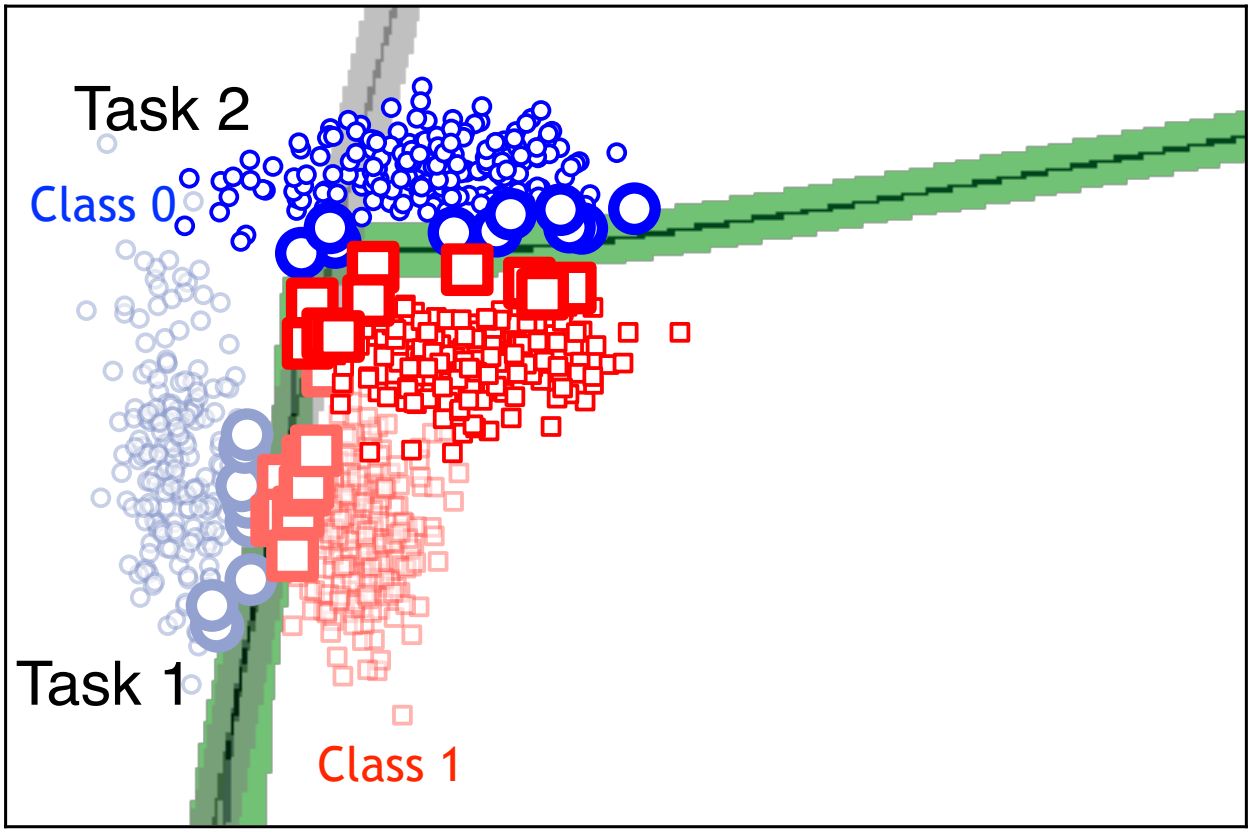
Choose memories where interference is more likely and add corrections to the variational objective.



1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

CL with “Memorable” Past [1]

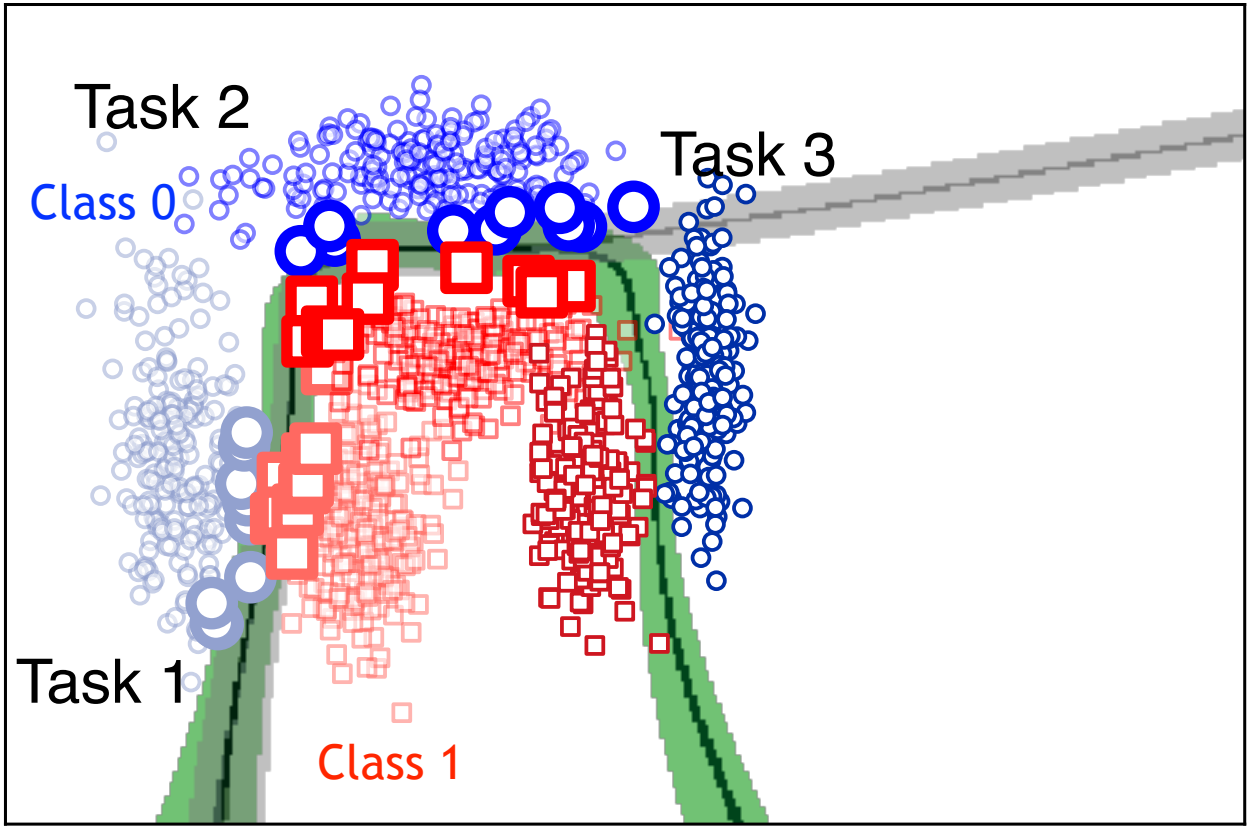
Choose memories where interference is more likely and add corrections to the variational objective.



1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

CL with “Memorable” Past [1]

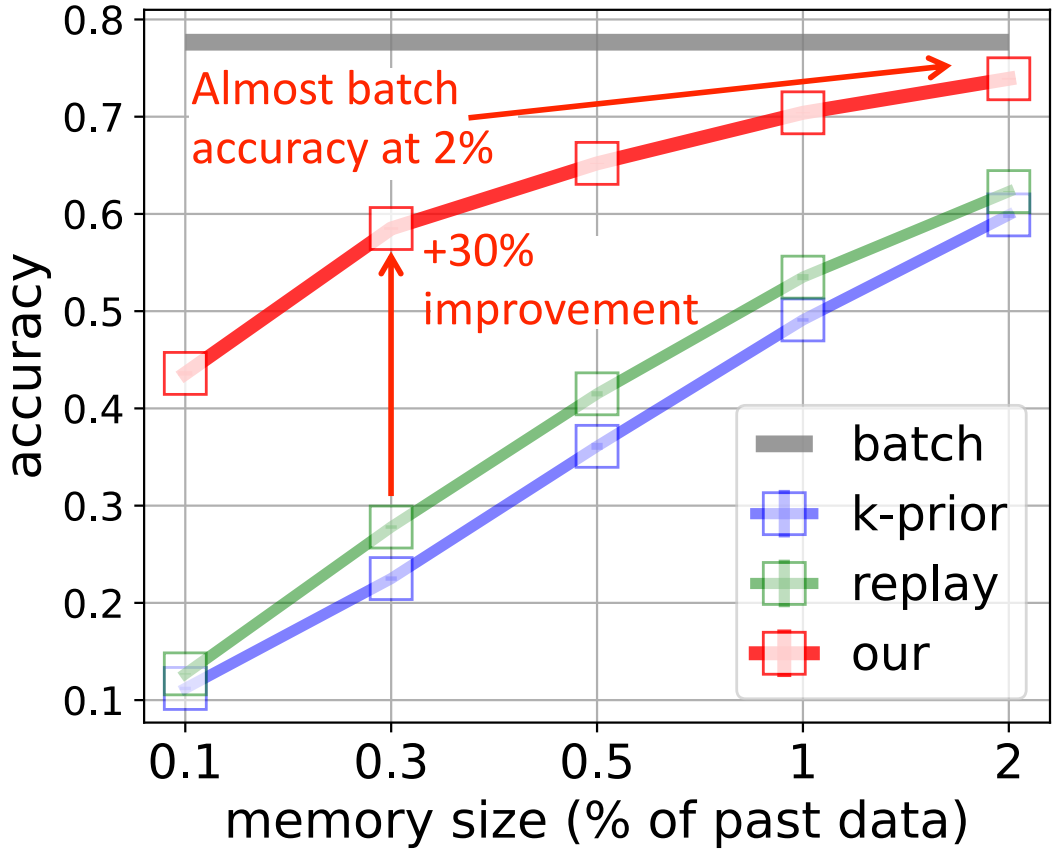
Choose memories where interference is more likely and add corrections to the variational objective.



1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

K-Priors with Compact Memory

We can Hessian corrections to build a memory



Batch accuracy with just < 2% of past data

The results are on split ImageNet-1000 classification with ViT-L/14 model

1. Jung et al. Compact Memory for Continual Logistic Regression, NeurIPS (2025)

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)

Variance Reduction for SGD

Speed-up SGD by reducing the gradient variance [1]. Essentially, we compute full-batch gradient at an older θ_{old} and then take a few steps of SGD, and iterate,

$$\theta \leftarrow \theta - \rho \left[\nabla \ell_j(\theta) \right.$$

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Variance Reduction for SGD

Speed-up SGD by reducing the gradient variance [1]. Essentially, we compute full-batch gradient at an older θ_{old} and then take a few steps of SGD, and iterate,

$$\theta \leftarrow \theta - \rho \left[\nabla \ell_j(\theta) \qquad \sum_{i=1}^{t} \nabla \ell_j(\theta_{old}) \right]$$

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Variance Reduction for SGD

Speed-up SGD by reducing the gradient variance [1]. Essentially, we compute full-batch gradient at an older θ_{old} and then take a few steps of SGD, and iterate,

$$\theta \leftarrow \theta - \rho \left[\nabla \ell_j(\theta) - \nabla \ell_j(\theta_{old}) + \sum_{i=1}^{t} \nabla \ell_j(\theta_{old}) \right]$$

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Variance Reduction for SGD

Speed-up SGD by reducing the gradient variance [1]. Essentially, we compute full-batch gradient at an older θ_{old} and then take a few steps of SGD, and iterate,

$$\theta \leftarrow \theta - \rho \left[\nabla \ell_j(\theta) - \nabla \ell_j(\theta_{old}) + \sum_{i=1}^{t} \nabla \ell_j(\theta_{old}) \right]$$

An earlier variant won the Lagrange prize in 2017! However, this approach cannot be easily generalized to other algorithms like Newton's method or Adam.

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Variance Reduction for SGD

Speed-up SGD by reducing the gradient variance [1]. Essentially, we compute full-batch gradient at an older θ_{old} and then take a few steps of SGD, and iterate,

$$\theta \leftarrow \theta - \rho \left[\nabla \ell_j(\theta) - \nabla \ell_j(\theta_{old}) + \sum_{i=1}^{t} \nabla \ell_j(\theta_{old}) \right]$$

An earlier variant won the Lagrange prize in 2017! However, this approach cannot be easily generalized to other algorithms like Newton's method or Adam.

Posterior Correction generalize this approach [2] and allows us to derive new variants.

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing SVRG with Posterior Correction [1]

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

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

1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing SVRG with Posterior Correction [1]

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

$$q \leftarrow q^{1-\rho} \prod_{j=1}^{t} \exp(-\rho\, \hat{\ell}_j)$$

1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing SVRG with Posterior Correction [1]

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

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

1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing SVRG with Posterior Correction [1]

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

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

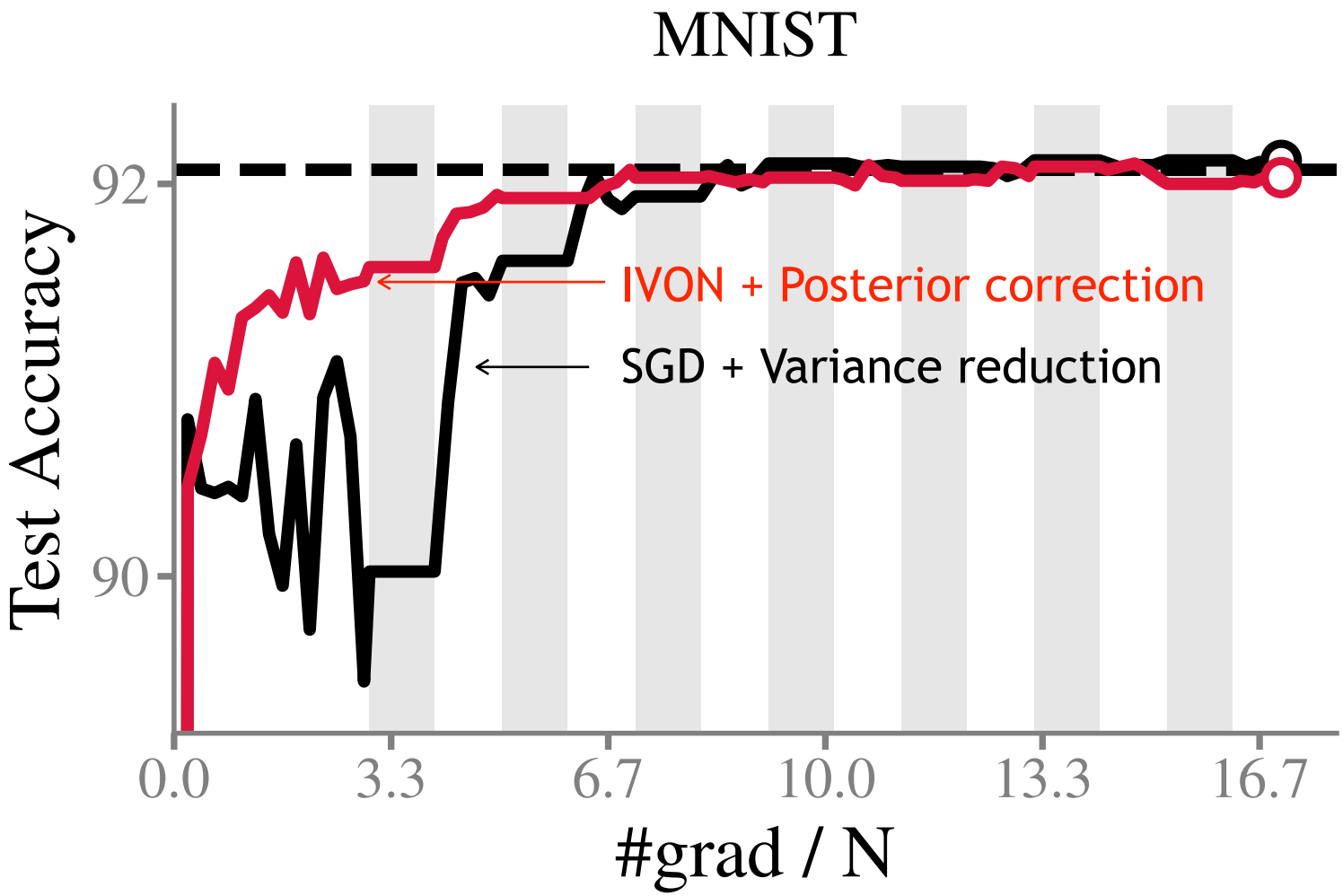
In the future, we aim to "correct" this as new data arrives

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

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

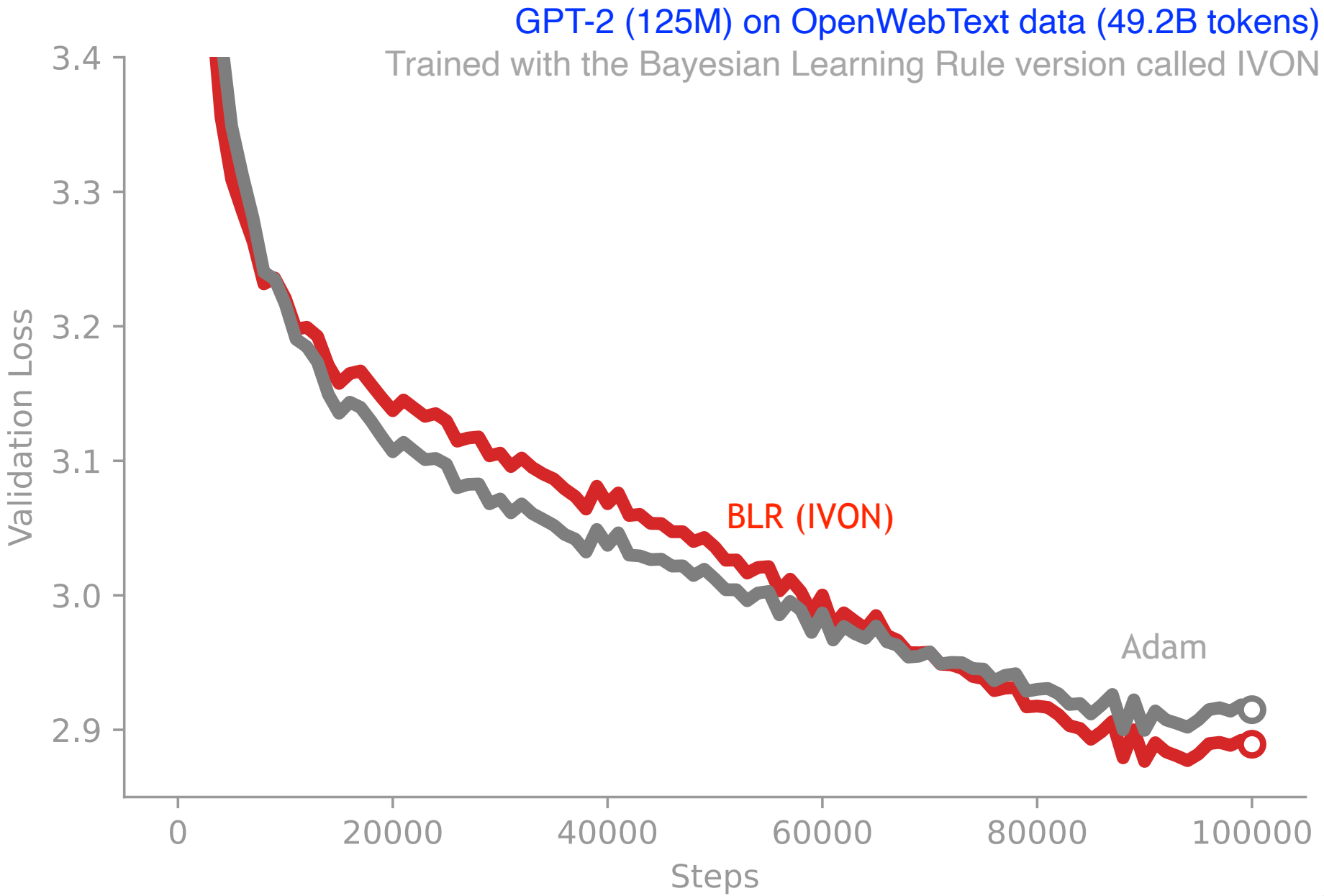
1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Logistic Regression



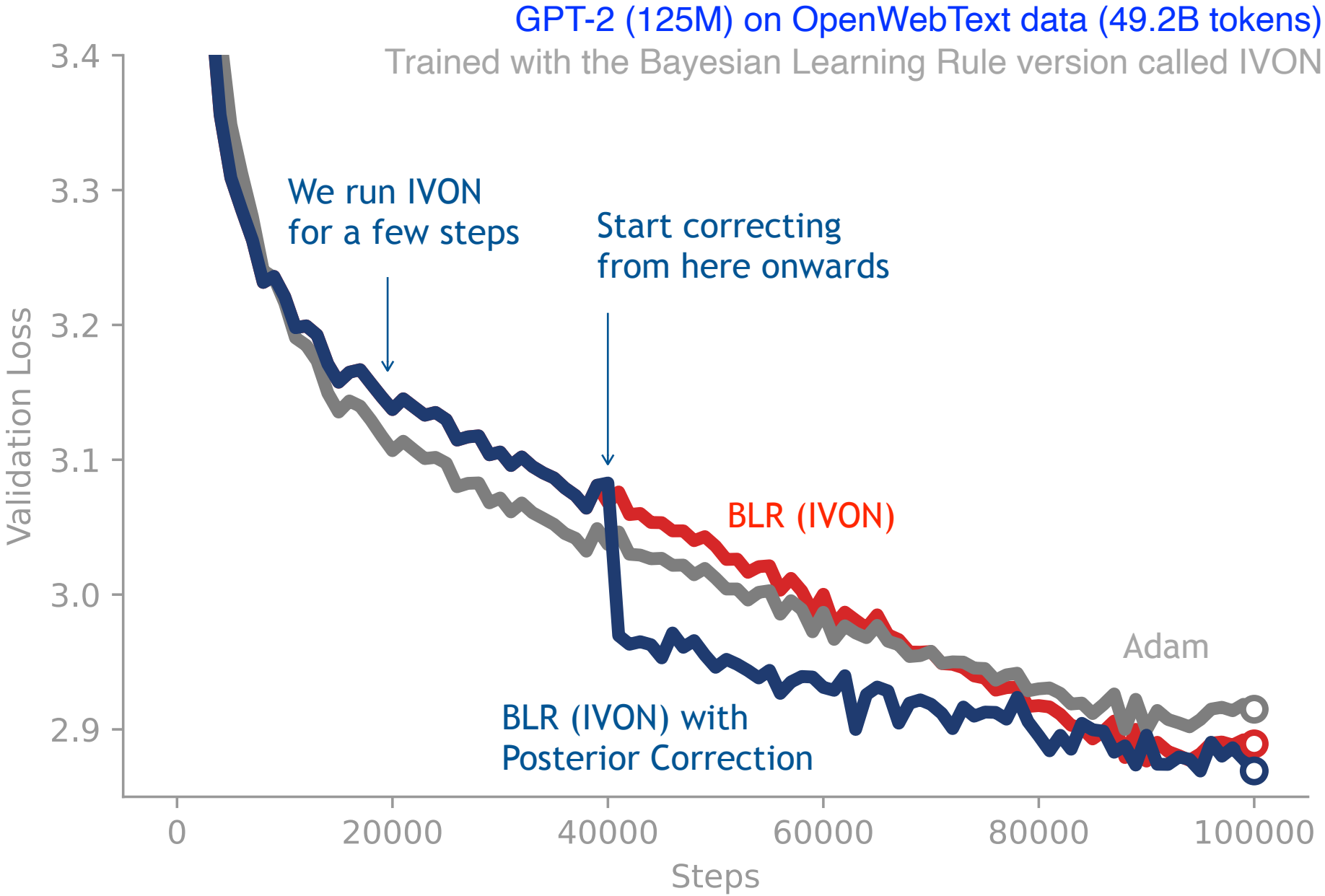
1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Posterior Correction can boost LLM training



1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Posterior Correction can boost LLM training



1. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing ADMM

Alternating Direction Method of Multiplier is a classical method from 70s for distributed optimization, but there are no good variants for Newton and Adam optimizers

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing ADMM

Alternating Direction Method of Multiplier is a classical method from 70s for distributed optimization, but there are no good variants for Newton and Adam optimizers

A simplified version for 2 clients with losses ℓ_1, ℓ_2

$$\theta_1 \leftarrow \theta_1 - \rho \left[\nabla \ell_1(\theta_1) \right.$$

$$\theta_2 \leftarrow \theta_2 - \rho \left[\nabla \ell_2(\theta_2) \right.$$

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing ADMM

Alternating Direction Method of Multiplier is a classical method from 70s for distributed optimization, but there are no good variants for Newton and Adam optimizers

A simplified version for 2 clients with losses ℓ_1, ℓ_2

$$\theta_1 \leftarrow \theta_1 - \rho \left[\nabla \ell_1(\theta_1) \sum_{i=1}^{2} \nabla \ell_j(\theta_j^{old}) \right]$$

$$\theta_2 \leftarrow \theta_2 - \rho \left[\nabla \ell_2(\theta_2) \sum_{i=1}^{2} \nabla \ell_j(\theta_j^{old}) \right]$$

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing ADMM

Alternating Direction Method of Multiplier is a classical method from 70s for distributed optimization, but there are no good variants for Newton and Adam optimizers

A simplified version for 2 clients with losses ℓ_1, ℓ_2

$$\theta_1 \leftarrow \theta_1 - \rho \left[\nabla \ell_1(\theta_1) - \nabla \ell_1(\theta_1^{old}) + \sum_{i=1}^{2} \nabla \ell_j(\theta_j^{old}) \right]$$

$$\theta_2 \leftarrow \theta_2 - \rho \left[\nabla \ell_2(\theta_2) - \nabla \ell_2(\theta_2^{old}) + \sum_{i=1}^{2} \nabla \ell_j(\theta_j^{old}) \right]$$

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

Generalizing ADMM

Alternating Direction Method of Multiplier is a classical method from 70s for distributed optimization, but there are no good variants for Newton and Adam optimizers

A simplified version for 2 clients with losses ℓ_1, ℓ_2

$$\theta_1 \leftarrow \theta_1 - \rho \left[\nabla \ell_1(\theta_1) - \nabla \ell_1(\theta_1^{old}) + \sum_{i=1}^{2} \nabla \ell_j(\theta_j^{old}) \right]$$

$$\theta_2 \leftarrow \theta_2 - \rho \left[\nabla \ell_2(\theta_2) - \nabla \ell_2(\theta_2^{old}) + \sum_{i=1}^{2} \nabla \ell_j(\theta_j^{old}) \right]$$

We can easily generalize ADMM by replacing gradient corrections by posterior corrections [2]

1. Johnson and Zhang, Accelerating SGD using predictive variance reduction, NeurIPS (2013)
2. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)

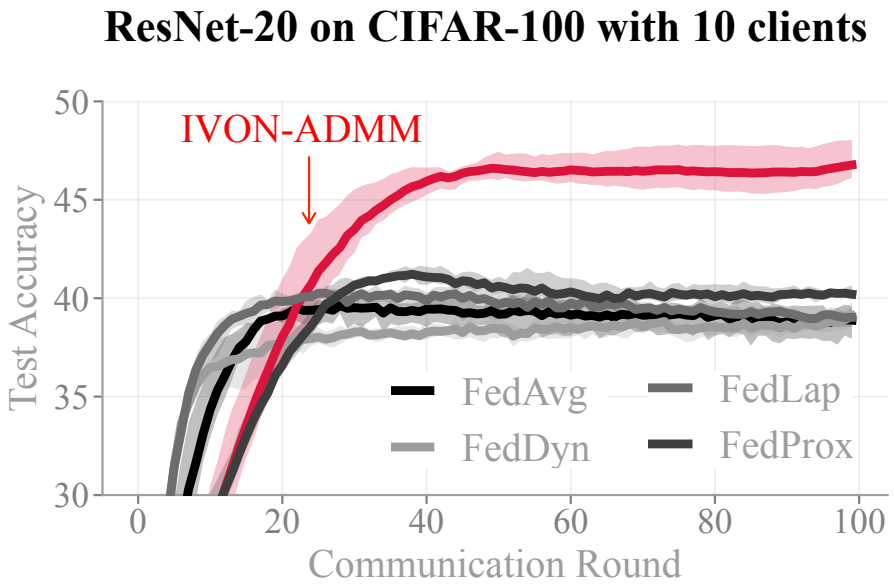
Distributed Deep Learning

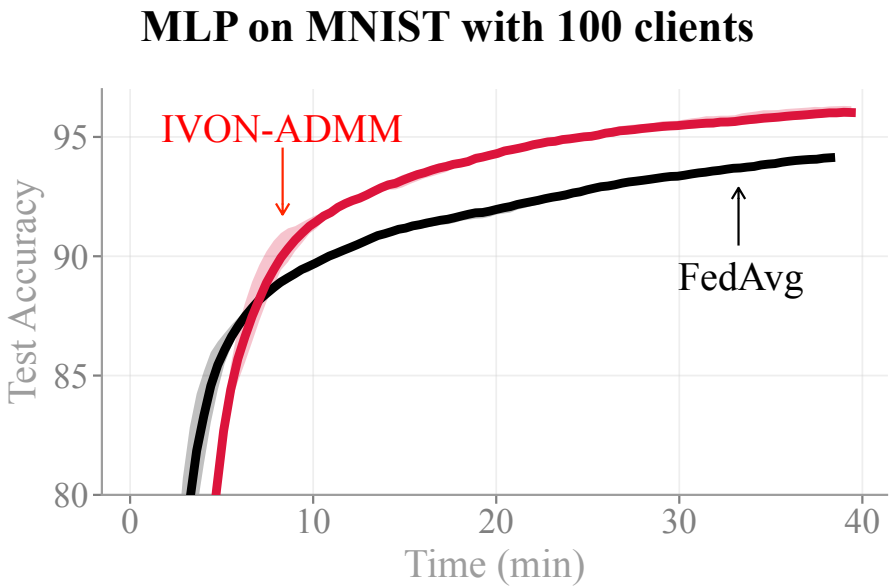
Bayesian generalizations with posterior corrections converge faster with no increase in costs [1]

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

Distributed Deep Learning

Bayesian generalizations with posterior corrections converge faster with no increase in costs [1]

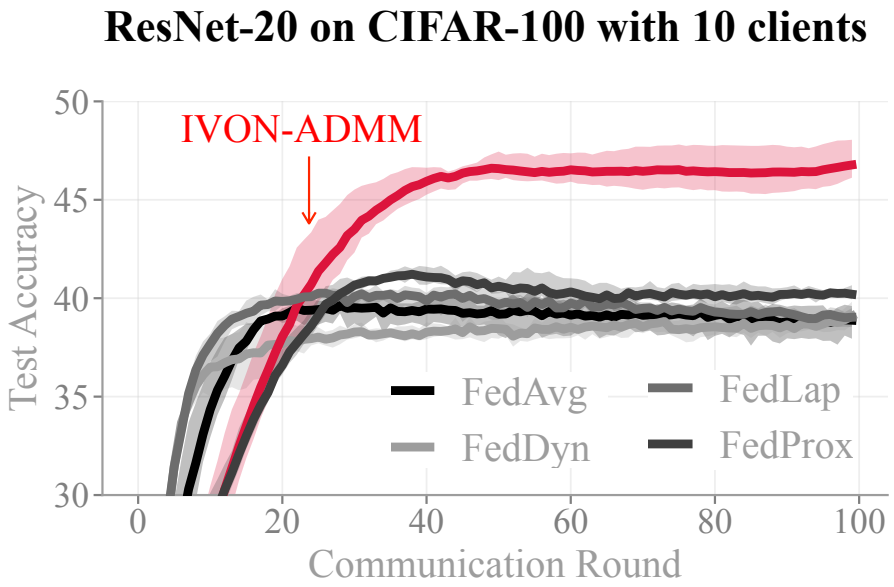


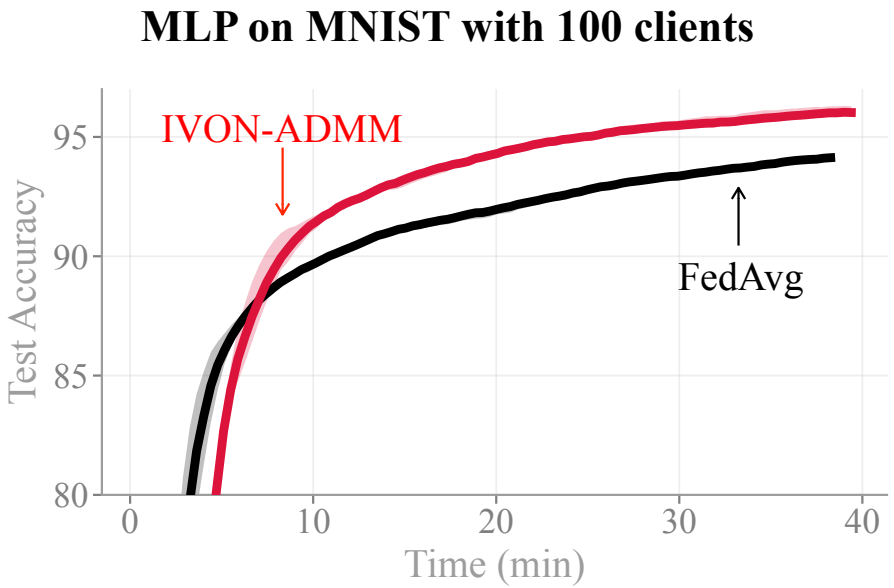


1. Moellenhoff et al. Federated ADMM from Bayesian Duality. ICLR (2026)

Distributed Deep Learning

Bayesian generalizations with posterior corrections converge faster with no increase in costs [1]



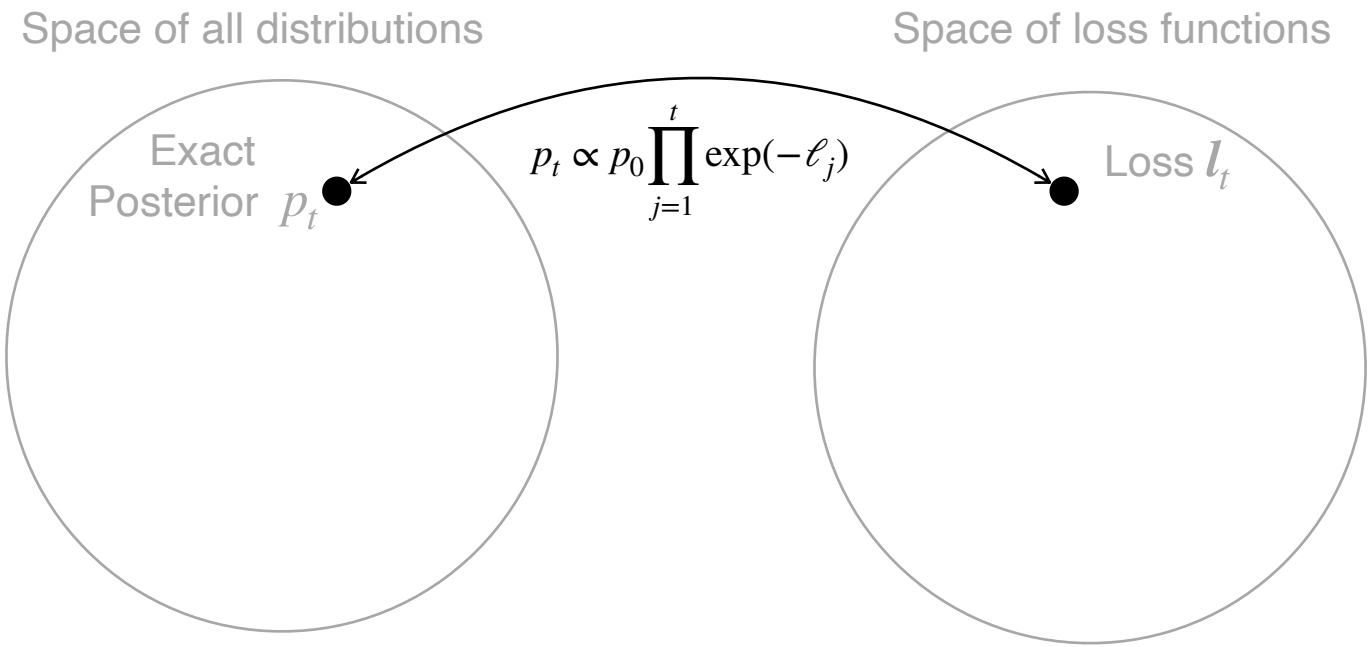


We hope to scale this to LLMs in the near future.

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

The Bayesian-Duality Principle

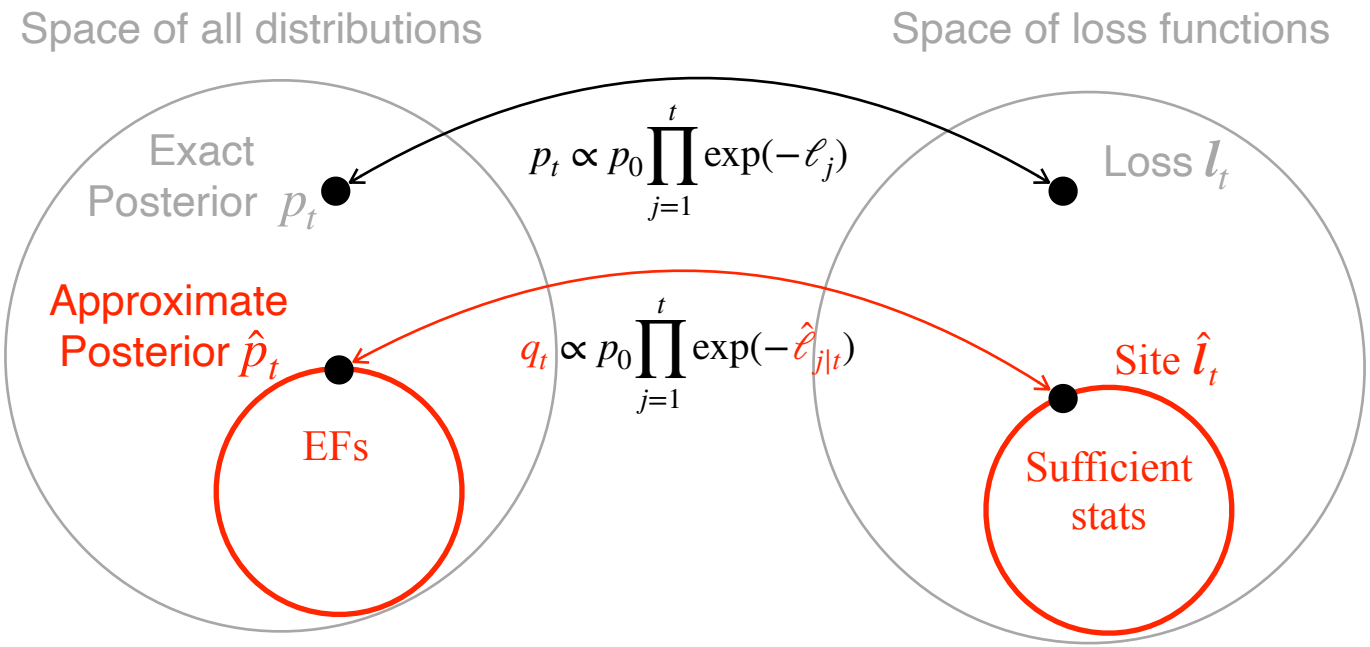
Why are we able to unify all sorts of knowledge adaptation tasks? This is due to duality of Bayes. Every (variational) posterior has a dual pairing



1. Khan et al. Fast Dual Variational Inference for Non-Conjugate LGMs. ICML (2013)

The Bayesian-Duality Principle

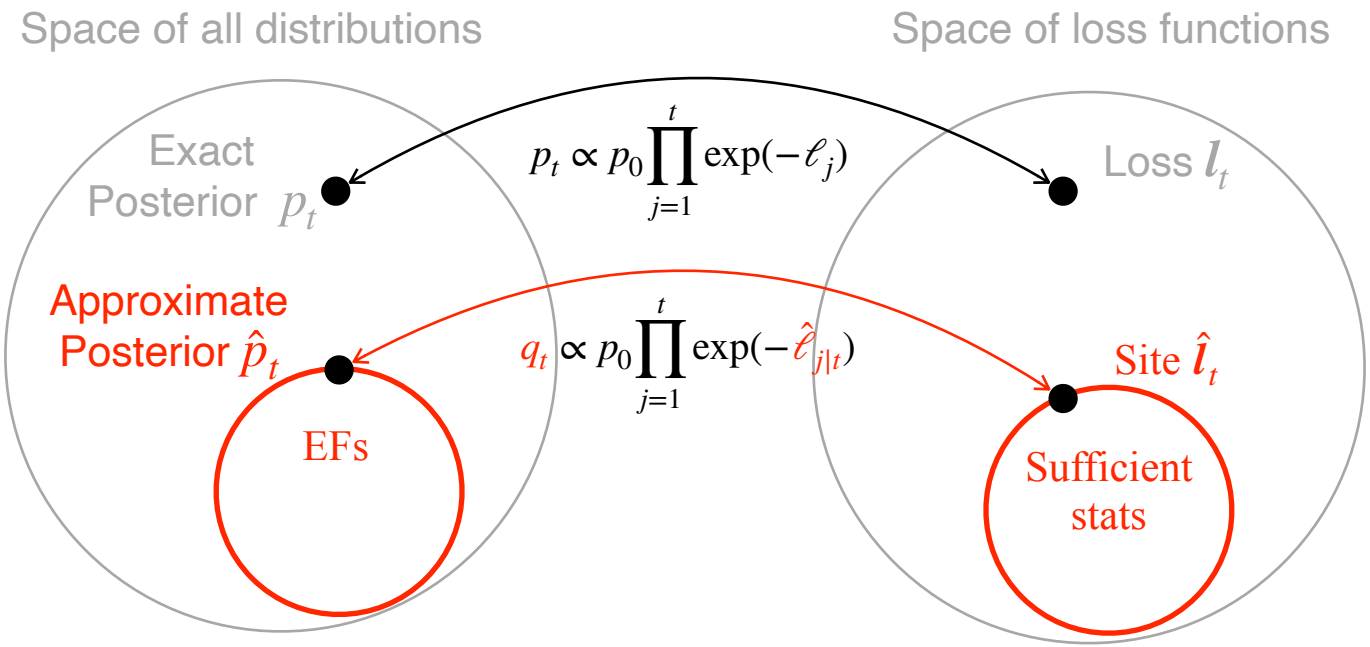
Why are we able to unify all sorts of knowledge adaptation tasks? This is due to duality of Bayes. Every (variational) posterior has a dual pairing



1. Khan et al. Fast Dual Variational Inference for Non-Conjugate LGMs. ICML (2013)

The Bayesian-Duality Principle

Why are we able to unify all sorts of knowledge adaptation tasks? This is due to duality of Bayes. Every (variational) posterior has a dual pairing



Information processing is additive in the dual space.

1. Khan et al. Fast Dual Variational Inference for Non-Conjugate LGMs. ICML (2013)

Duality of (Exact) Bayes

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

Log Marginal Likelihood

KL

$$\log \int dp_0 \prod_{j=1}^{t} \exp(-\ell_j) = \min_{q \in \mathcal{P}} \sum_{j=1}^{t} \mathbb{E}_q[\ell_j] + KL(q \| p_0)$$

KL

Log Marginal Likelihood

$$KL(p_t \| p_0) = \min_{f \in \mathcal{P}^*} \sum_{j=1}^{t} \mathbb{E}_{p_t}[f_j] + \log \int dp_0 \prod_{j=1}^{t} \exp(-f_j)$$

The duality of VB posterior extends this result to show that $(q_t, \hat{\boldsymbol{l}}_t)$ is also a dual pair.

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)

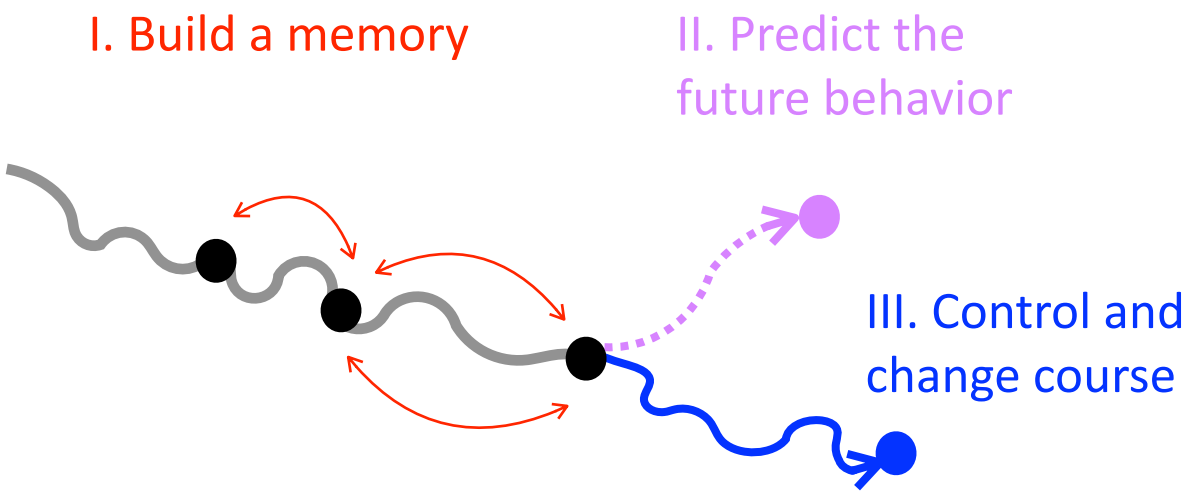
Adaptive Bayesian Intelligence

- Developing Adaptive Intelligence via Bayesian Principles
- Part 1: Bayesian Learning Rule [1]
  - Unifies many machine-learning algorithms
  - We use it to improve Deep Learning [2]
- Part 2: Posterior Correction [3]
  - Unifies many knowledge-adaptation methods
  - We use it to improve Continual learning [4], Variance reduction [5], and Distributed optimization [6]
- Bayes for the next-generation adaptive intelligence

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023)
2. Shen et al. Variational Learning is Effective for Large Deep Networks, ICML (2024)
3. Khan. Knowledge Adaptation as Posterior Correction, arXiv (2025)
4. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS (2021)
5. Daheim et al. SVRG and Beyond with Posterior Correction, ICML (2026)
6. Moellenhoff et al. Federated ADMM from Bayesian Duality. ICLR (2026)

Towards Sustainable AI Training

A sustainable AI marketplace to “reduce, reuse, recycle” models instead of training new models from scratch.



Adaptive methods enable us to give more control to the developer and make better use of existing resources.

See the discussion in Sec 5 in Khan (2025)

Bayesian Principles to Build the Next-Generation Adaptive Intelligence [2]

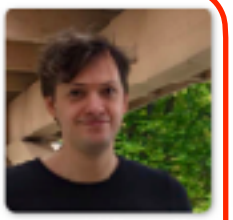
- How can we reduce the cost of training AI?
  - What should the algorithm remember and what new experiences it should seek?
  - Build a memory of the past, inject prior knowledge, and design a curriculum to slowly explore the future
  - Note: Correction is (variational) information gain
- To truly reduce the cost, we also need
  1. Encourage parsimony in data and parameters
  2. Use local learning (brain-like learning)
  3. Perform active self-guidance
- Adaptive Bayesian Intellignce is a path to sustainability

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

Adaptive Bayesian Intelligence Team

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

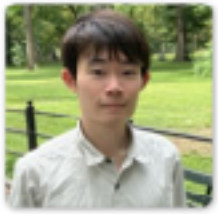
Emtiyaz Khan
Team Director



Thomas Möllenhoff
Deputy Team Director

Keigo Nishida
Special Postdoctoral Researcher
RIKEN BDR

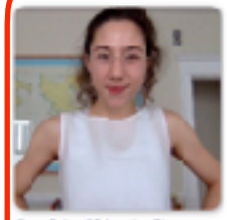
Christopher J. Anders
Postdoctoral Researcher



Hajime Ueda
Part-Time Student
The University of Tokyo

Bai Cong
Part-Time Student
Tokyo Institute of Technology

Roshni Kamath
Intern
TU Darmstadt



Sophia Sklaviadis
Intern
Instituto Superior Tecnico

Elia Mounier-Poulat
Intern
EPFL, Switzerland

Adrian R. Ninut
Intern
Sapienza, University of Rome

Sungjun Lim
Intern
Yonsei University

Joseph Austerweil
Visiting Scientist
University of Winsconsin-Madison

Pierre Alquier
Visiting Scientist
ESSEC Business School

Geoffrey Wolfer
Visiting Scientist
Waseda University

Rio Yokota
Visiting Scientist
Tokyo Institute of Technology

Kosuke Segiyama
Visiting Scientist
Waseda University

Nico Daheim joined as a post-doc at TU Darmstadt today!

And many of our interns & collaborators over the past 10 years (2016-2026)

86. Eiki Shimizu (Oct 20
85. Dharmesh Tailor (Oc
84. Geoffrey Wolfer (Aug
83. Henrique Gameiro (A
82. Giulia Lanzillotta (Ap
81. Guiomar Pescador Ba
80. Anita Yang (May 202
79. Adrian R. Minut (Jan
78. Florian Sligmann (Ap
77. Rin Intachuen (Dec 2
76. Joe Austerweil (Aug.
75. Sin-Han Yang (Aug 2
74. Alexander Timans (J.
73. Masaki Adachi (Jan-M
72. Marco Miani (Oct 202
71. Hyungi Lee (Nov 202
70. Zhedong Liu (Nov 20
69. Wenlong Chen (Aug-
68. Avrajit Ghosh (Aug-M
67. Clement Bazan (Apri
66. Peter Nickl (May 202
65. Geoffrey Wolfer (Mar
64. Lu Xu (Nov 2021-De
63. Erik Daxberger (Jun

62. Gian Maria Marconi (
61. Etash Guha (May-Sep
60. Naima Elosegui (June
59. Happy Buzaaba (Jun
58. Ang Ming Liang (July
57. Yuesong Shen (May-
56. Joe Austerweil (Sep
55. Wu Lin (Jan 2018-Jun
54. Negar Safinianaini (J
53. Erik Englesson (Feb-
52. Paul Chang (Mar 202
51. Ramansh Sharma (A
50. Alexander Piche (Sep
49. Jooyeon Kim (Dec 20
48. Pierre Alquier (Aug 2
47. Ali Unlu (April-Sep 20
46. Kenneth Chen (July-S
45. David Tomàs Cuesta
44. Tojo Rakotoaritina (J
43. Happy Buzaaba (July
42. Ted Tinker (Sep 2021

41. Dharmesh Tailor (Ma
40. Siddharth Swaroop (
39. Evgenii Egorov (Jun
38. Peter Nickl (May 202
37. Fariz Ikhwantri (July
36. Dimitri Meunier (May
35. Lucie Perrotta (Sep 2
34. Xiangming Meng (Jul
33. Farzaneh Mahdisoltan
32. Alexander Immer (M
31. Roman Bachmann (J
30. Kazuki Osawa (Nov 2
29. Vincent Tan (May 20
28. Hongyi Ding (July 20
27. Anshuk Uppal (June-
26. Michael Przystupa (Ju
25. Maciej Korzepa (Feb-
24. Matthias Bauer (Sep-
23. Pingbo Pan (May-Sep
22. Pierre Orenstein (Ma
21. Benjamin Bray (May-

20. Ehsan Abedi (March-
19. Mark Goldstein (June
18. Anirudh Jain (Dec 20
17. Runa Eschenhagen (
16. Anand Subramanian
15. Dr. Parag Rastogi (Ap
14. Ohiremen Dibua (Int
13. Jiaxin Shi (Intern fro
12. Hanna Tseran (Interr
11. Sl Kal Lee (Research
10. Frederik Kunster (Int
9. Didrik Nielsen (Resea
8. Aaron Mishkin (Inter
7. Wu Lin (Research ass
6. Nicolas Hubacher (Re
5. Zuozhu Liu (Intern fr
4. Vaden Masrani (Inter
3. Salma El Aloui (Inter
2. Kimia Nadjahi (Inter
1. Arnaud Robert (Inter

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!

Thank you!