Presentation at the 3rd Bayes-Duality Workshop, Jun 26, 2025

Adaptive Bayesian Intelligence

(AGI meets ABI)

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo

https://emtiyaz.github.io

Summary of recent research at https://emtiyaz.github.io/papers/symposium_2024.pdf

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

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

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 220M + EUR 500K through the CREST-ANR grant! Thanks to the funding agencies!

Adaptive Bayesian Intelligence Team

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

Emtiyaz Khan
Team Leader

Thomas Möllenhoff
Research Scientist

Keigo Nishida
Special Postdoctoral Researcher
RIKEN BDR

Hugo Monzón Maldonado
Postdoctoral Researcher

Christopher Johannes Anders
Postdoctoral Researcher

Yohan Jung
Postdoctoral Researcher

Anita Yang
Part-Time Student
The University of Tokyo

Bai Cong
Part-Time Student
Tokyo Institute of Technology

Eiki Shimizu
Part-Time Student
Institute of Statistical Mathematics

Giulia Lanzillotta
Intern
ETH Zurich

Adrian R. Minut
Intern
Sapienza, University of Rome

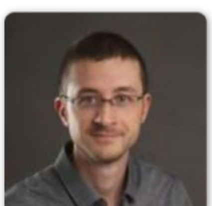
Florian Seligmann
Intern
Karlsruhe Institute of Technology

Guiomar Pescador Barrios
Intern
Imperial College London

Henrique Da Silva Gameiro
Intern
EPFL, Switzerland

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

Dharmesh Tailor
Remote Collaborator
University of Amsterdam

And many of our collaborators!

77. Rin Intachuen (Dec 2024-Apr 2025, Research Inter
76. Sin-Han Yang (Aug 2024-Mar 2025, Tech-Staff)
75. Alexander Timans (Jan-Mar 2025, Research Intern
74. Masaki Adachi (Jan-Mar 2025, Research Intern fron
73. Marco Miani (Oct 2024-Mar 2025, Research Intern f
72. Hyungi Lee (Nov 2024-Jan 2025, Research Intern fr
71. Zhedong Liu (Nov 2023-Dec 2024, Post-doc)
70. Wenlong Chen (Aug-Dec 2024, Research Intern fro
69. Avrajit Ghosh (Aug-Nov 2024, Research Intern fron
68. Clement Bazan (April-July 2024, Part-timer, Tokyo-T
67. Peter Nickl (May 2021-May 2024, Research Assistant,
66. Geoffrey Wolfer (Mar 2022-Mar 2024, Post-Doc throug
65. Lu Xu (Nov 2021-Dec 2023, Post-doc)
64. Erik Daxberger (Jun 2020-Oct 2023, remote collaborat
63. Gian Maria Marconi (Aug 2020-Sep 2023, Post-doc)
62. Etash Guha (May-Sep 2023, Research Intern from Gec
61. Naima Elosegui (June-Aug 2023, Research Intern from
60. Happy Buzaaba (Jun 2022-Aug 2023, Post-Doc for the
59. Ang Ming Liang (July 2022-Aug 2023, remote collabor
58. Yuesong Shen (May-June 2023, Research Intern from
57. Joe Austerweil (Sep 2022-May 2023, Visiting Professo
56. Wu Lin (Jan 2018-June 2023, PhD student at UBC, co-
55. Negar Safinianaini (Jan-July 2023, Post-Doc for the Ba
54. Erik Englesson (Feb-April 2023, Visiting PhD student f
53. Paul Chang (Mar 2021-May 2023, remote collaborator,
52. Ramansh Sharma (Aug 2022-Mar 2023, remote collab
51. Alexander Piche (Sep 2020-Mar 2023, remote collabor
50. Jooyeon Kim (Dec 2021-Feb 2022, Post-doc, started a
49. Pierre Alquier (Aug 2019-Dec 2022, Research Scientis
48. Ali Unlu (April-Sep 2022, intern at OIST, started as a F
47. Kenneth Chen (July-Sep 2022, intern at OIST)
46. David Tomàs Cuesta (Jan 2022-Apr 2022, Rotation Ph
45. Tojo Rakotoaritina (Jan 2022-Apr 2022, Rotation PhD
44. Happy Buzaaba (July 2020-Mar 2022, part-time PhD s
43. Ted Tinker (Sep 2021-Dec 2021, Rotation PhD student
42. Dharmesh Tailor (May 2019-Aug 2021, Research Assis
41. Siddharth Swaroop (Nov 2018-June 2021, remote coll
40. Evgenii Egorov (Jun 2020-May 2021, remote collabora
39. Peter Nickl (May 2020-Apr 2021, remote collaborator,
38. Fariz Ikhwantri (July 2020-March 2021, part-time MSc
37. Dimitri Meunier (May 2020-Nov 2020, remote collabor
36. Lucie Perrotta (Sep 2019-Mar2020, Intern from EPFL,
35. Xiangming Meng (July 2019-Mar 2020, Postdoc, starte
34. Farzaneh Mahdisoltani (Sep 2019-Feb 2020, Intern fro
33. Alexander Immer (March 2019- March 2020, Intern from EPFI
32. Roman Bachmann (July 2019-Feb 2020, Intern from EPFL, Sw
31. Kazuki Osawa (Nov 2019-Feb 2020, Trainee from Tokyo Tech)
30. Vincent Tan (May 2019-Jan 2020, Research assistant, went to
29. Hongyi Ding (July 2019-Jan 2020, Postdoc)
28. Anshuk Uppal (June-Dec 2019, Intern from IIIT Bangalore)
27. Michael Przystupa (July-Dec 2019, Intern from UBC Vancouve
26. Maciej Korzepa (Feb-Dec 2019, Intern from DTU, Copenhager
25. Matthias Bauer (Sep-Oct 2019, Intern from Max-Planck Instit
24. Pingbo Pan (May-Sep 2019, Intern from UT Sydney)
23. Pierre Orenstein (May-Sep 2019, Intern from France)
22. Benjamin Bray (May-August 2019, Intern from Georgia Tech)
21. Ehsan Abedi (March-August 2019, Intern from EPFL, Switzerla
20. Mark Goldstein (June-Sep 2019, Intern from NYU)
19. Anirudh Jain (Dec 2018-July 2019, Intern from ISM, India)
18. Runa Eschenhagen (Oct 2018-May 2019, Intern from Univers
17. Anand Subramanian (Feb-May 2019, Intern from JAIST, Japar
16. Dr. Parag Rastogi (Apr 2017-Mar 2019, Visiting Scientist, Univ
15. Ohiremen Dibua (Intern from Stanford University between Jul
14. Jiaxin Shi (Intern from Tsinghua University between July 2018
13. Hanna Tseran (Intern from University of Tokyo from Nov. 201
12. Si Kai Lee (Research Assistant from Dec 2017 to August 2018
11. Frederik Kunster (Intern from EPFL from Feb 2018 to August
10. Didrik Nielsen (Research assistant from March 2017 to August
9. Aaron Mishkin (Intern from UBC during Jan-Jun 2018, joined
8. Wu Lin (Research assistant from Jan-Dec 2017, joined UBC as
7. Nicolas Hubacher (Research Assistant from Jan-Dec 2017)
6. Zuozhu Liu (Intern from SUTD during June-Dec 2017).
5. Vaden Masrani (Intern from UBC during May-Oct 2017)
4. Salma El Aloui (Intern from École Polytechnique during Jun-S
3. Kimia Nadjahi (Intern from ENS Cachan during May-Sep 2017
2. Arnaud Robert (Intern from EPFL during Oct 2016 to April 201
1. Heiko Strathman (from UCL)

AI that can learn like us

Quickly adapt & continue to acquire new skills

Human Learning at the age of 6 months.

Converged at the age of 12 months

Transfer skills

at the age of 14 months

Teacher-Student Learning?

Current state of Machine Learning

https://www.youtube.com/watch?v=TxobtWAFh8o The video is from 2017

Retraining from Scratch

Even when changes are tiny.

It is costly, undemocratic and unsustainable.

Adaptive Intelligence

How do brains adapt quickly?
What do they optimize and how?

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. Sternberg. What is intelligence really? the futile search for a holy grail. Learning & Individual Differences (2024)

Adaptive Bayesian Intelligence

- Adaptive Intelligence = Bayesian Computation
- Part 1: Bayesian Learning Rule [1]
  - (Emti) Foundational way to derive learning-algorithms
  - (Thomas and Nico) Application to DL: IVON [2]
- Part 2: Posterior Correction [3]
  - (Emti) Foundational way to derive adaptation-algorithms
  - (Emti) Application to continual learning [4-5], model merging [6]
  - (Siddharth and Thomas) Federated Learning
- More application to DL:
  - (Kenichi, Keigo) Low-Precision training, (Cong-Bai) IVON-LoRA

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. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020
6. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).

“The fact that many different approaches point to the same actual algorithm is a major strength of Bayesianity”

—E. T. Jaynes, discussion of [1]

1. Zellner, Optimal Information Processing and Bayes’ Theorem. The American Statistician (1988)

Optimization

Gradient Descent
Newton's Method
Multimodal Optimization

Deep-Learning

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

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).

Variational Formulation of Bayes' Rule

Bayes' Rule: $p_t(\theta) \propto p_0(\theta) \prod_{j=1}^{t} \text{lik}_j(\theta)$

Variational Inference to find an approximation $q_t(\theta)$

$$q_t = \arg\min_{q \in \mathcal{Q}} \sum_{j=1}^{t} \mathbb{E}_q[\underbrace{-\log \text{lik}_j}_{= \ell_j}] + KL(q \| \underbrace{p_0}_{\propto e^{-\ell_0}})$$

$$= \arg\min_{q \in \mathcal{Q}} \sum_{j=0}^{t} \mathbb{E}_q[\ell_j] - \mathscr{H}(q)$$

We will use this variational formulation to discover the inherent Bayesian nature of (non-Bayesian) algorithms.

Exponential Family

Natural parameters

Sufficient Statistics

Expectation parameters

$$q(\theta) \propto \exp\left[\lambda^\top T(\theta)\right] \qquad \mu := \mathbb{E}_q[T(\theta)]$$

$$\mathcal{N}(\theta|m, S^{-1}) \propto \exp\left[-\frac{1}{2}(\theta - m)^\top S(\theta - m)\right]$$

$$\propto \exp\left[(Sm)^\top \theta + \mathrm{Tr}\left(-\frac{S}{2}\theta\theta^\top\right)\right]$$

Gaussian distribution	$q(\theta) := \mathcal{N}(\theta\|m, S^{-1})$
Natural parameters	$\lambda := \{Sm, -S/2\}$
Expectation parameters	$\mu := \{\mathbb{E}_q(\theta), \mathbb{E}_q(\theta\theta^\top)\}$

1. Wainwright and Jordan, Graphical Models, Exp Fams, and Variational Inference Graphical models 2008
2. Malago et al., Towards the Geometry of Estimation of Distribution Algos based on Exp-Fam, FOGA, 2011

Bayesian Learning Rule (BLR) [1]

Deep Learning to find θ

$$\min_{\theta} \bar{\ell}(\theta) = \sum_{j=0}^{t} \ell_j(\theta)$$

SGD or Adam

$$\theta \leftarrow \theta - \rho \; P^{-1} \underbrace{\nabla \bar{\ell}(\theta)}_{\text{Gradient}}$$

Variational Learning to find $q_\lambda(\theta)$

$$\min_{q_\lambda \in \mathcal{Q}} \mathcal{L}(q_\lambda) = \sum_{j=1}^{t} \mathbb{E}_{q_\lambda}[\ell_j] + KL(q_\lambda \| \underbrace{p_0}_{\propto e^{-\ell_0}})$$

Bayesian Learning Rule

$$\lambda \leftarrow \lambda - \rho \; \underbrace{F(q_\lambda)^{-1} \nabla \mathcal{L}(q_\lambda)}_{\text{Natural Gradient}}$$

$$\lambda \leftarrow \lambda - \rho \; \nabla_\mu \mathcal{L}(\lambda)$$

Algorithms (such as SGD/Adam) are special cases of BLR obtained by choosing specific exp-family q_λ with natural parameter λ and expectation parameter μ.

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).

Deriving Gradient Descent from BLR

Derived by choosing Gaussian with fixed covariance

Gaussian distribution	$q(\theta) := \mathcal{N}(m, 1)$
Natural parameters	$\lambda := m$
Expectation parameters	$\mu := \mathbb{E}_q[\theta] = m$
Entropy	$\mathcal{H}(q) := \log(2\pi)/2$

BLR: $$\lambda \leftarrow \lambda - \rho\, \nabla_\mu \left(\mathbb{E}_q[\bar{\ell}] - \mathscr{H}(q) \right)$$

$$m \leftarrow m - \rho\, \nabla_m \mathbb{E}_q[\bar{\ell}]$$

$$m \leftarrow m - \rho\, \mathbb{E}_q[\nabla_\theta \bar{\ell}]$$ Bonnet's theorem

$$m \leftarrow m - \rho \nabla \bar{\ell}(m)$$ First-order delta method

GD: $$\theta \leftarrow \theta - \rho \nabla \bar{\ell}(\theta)$$

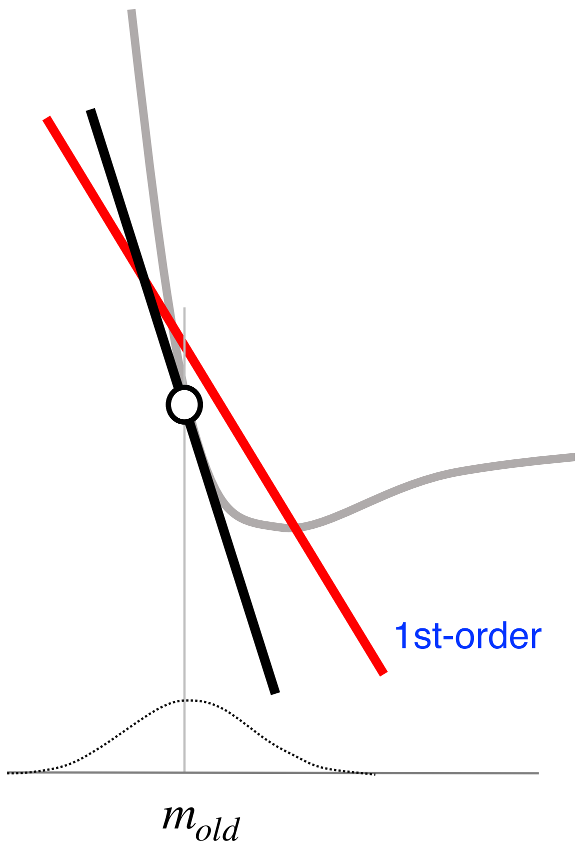
Bayesian learning rule:

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

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).

Taylor vs Bayes

Why do we recover optimization algorithm from BLR?



GD: $\theta \leftarrow \theta - \rho \nabla_\theta \bar{\ell}(\theta_{old})$

Taylor's surrogate: $\sum_i \theta^\top \nabla \ell_i(\theta_{old})$

BLR with isotropic Gaussian

$$m \leftarrow m - \rho \, \mathbb{E}_{q_{old}}[\nabla \bar{\ell}(\theta)]$$

Bayes's surrogate: $\sum_i \theta^\top \mathbb{E}_{q_{old}}[\nabla \ell_i]$

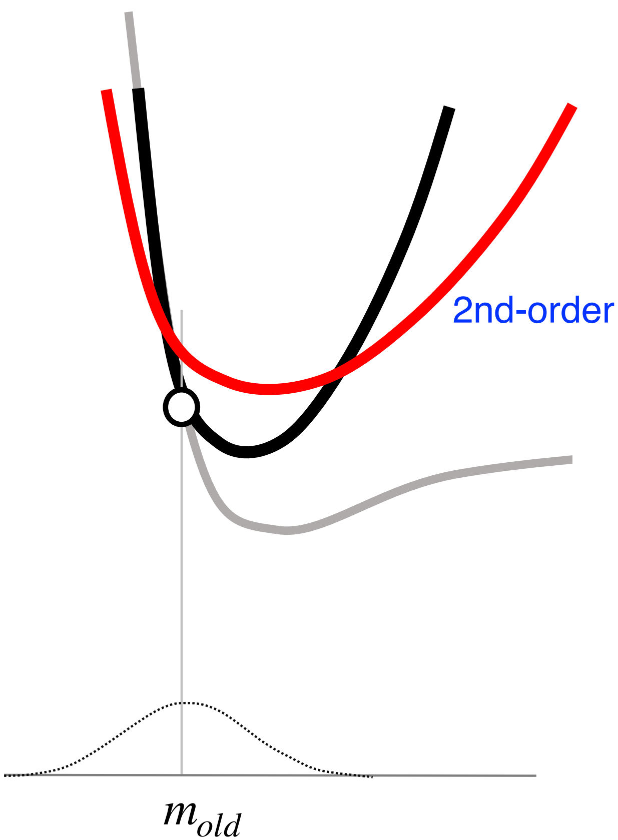
BLR generalizes Taylor!

Eq. 18 in Khan and Nielsen (2018), Eq. 59 Khan and Rue (2023), Eq. 3 in Section 2 in Khan (2025)

Bayes Generalizes Taylor

BLR with full cov Gaussian:

$$\sum_i \theta^\top \mathbb{E}_{q_{old}}[\nabla \ell_i] + \frac{1}{2}(\theta - m_{old})^\top \mathbb{E}_{q_{old}}[\nabla^2 \ell_i](\theta - m_{old})$$



BLR with exponential-family:

Suff stats

$$q_{old} \propto \exp(T(\theta)^\top \lambda_{old})$$

Natural gradients

$$= \exp\Big(- \sum_{i=0}^{t} \boxed{T(\theta)^\top \nabla_\mu \mathbb{E}_{q_{old}}[\ell_i]} \Big)$$

Site $\hat{\ell}_{i|old}(\theta)$

Sites are important for adaptation!

Dual-Representation of the BLR

$$q_t \propto \exp(T(\theta)^\top \lambda_t) \;\; = \exp\Big(- \sum_{i=0}^{t} \underbrace{T(\theta)^\top \nabla_\mu \mathbb{E}_{q_t}[\ell_i]}_{\text{Site } \hat{\ell}_{i|t}(\theta)} \Big)$$

$$q_t \propto \prod_{i=0}^{t} \exp(-\hat{\ell}_{i|t}) \quad \Longleftrightarrow \quad \lambda_t = \sum_{i=0}^{t} \nabla_\mu \mathbb{E}_{q_t}[\ell_i]$$

Posterior Sites Natural parameters Natural gradients

Natural Gradients are additive (representation theorem). Largest ones are the most influential.

1. Khan et al. Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models. ICML (2013)
2. Khan and Nielsen. Fast yet Simple Natural-Gradient Descent for Variational Inference … ISITA (2018)
3. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Processes. NeurIPS (2019)
4. Adam et al. Dual Parameterization of Sparse Variational Gaussian Processes. NearIPS (2021)
5. Chang et al. Memory-Based Dual Gaussian Processes for Sequential Learning. ICML (2023)
6. Moellenhoff et al. Federated ADMM from Bayesian Duality. arXiv (2025)

Continual Learning

Elastic Weight Consolidation
Variational Continual Learning
Memory Replay Methods
Functional Regularization

Model Merging

Task Arithmetic
Fisher/Hessian-Based Merging
Ensembles Methods

Posterior Correction [1]

Unlearning and Influence

Student-Teacher Learning

Knowledge Distillation
Learning with Privileged information
Incremental SVMs

Federated Learning

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

1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)

Adaptive Intelligence

How do brains adapt quickly?
What do they optimize and how?

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. Sternberg. What is intelligence really? the futile search for a holy grail. Learning & Individual Differences (2024)

Variational Formulation of Online Bayesian Inference

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

Variational formulation:

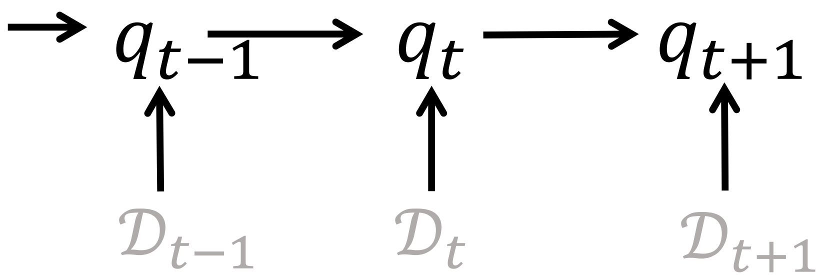
Batch: $$q_{t+1} = \arg\min_q \sum_{j=1}^{t+1} \mathbb{E}_q[\ell_j] + KL(q\|p_0)$$

Online [1]: $$\hat{q}_{t+1} = \arg\min_q \ \mathbb{E}_q[\ell_{t+1}] + KL(q\|q_t)$$

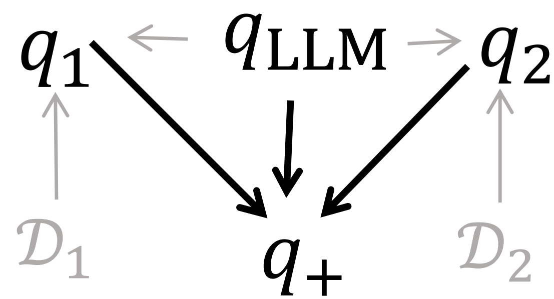
How inaccurate is $\hat{q}_{t+1}$? Can we correct it to exactly recover q_{t+1}? This is the goal of posterior correction.

1. Nguyen et al. Variational continual learning. ICLR (2018)

Continual Learning

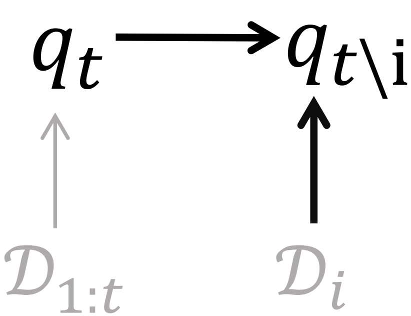


Model Merging

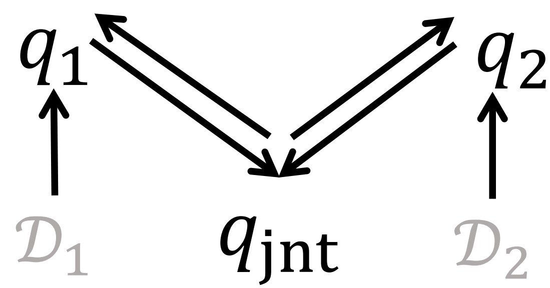


Posterior Correction [1]

Unlearning and Influence

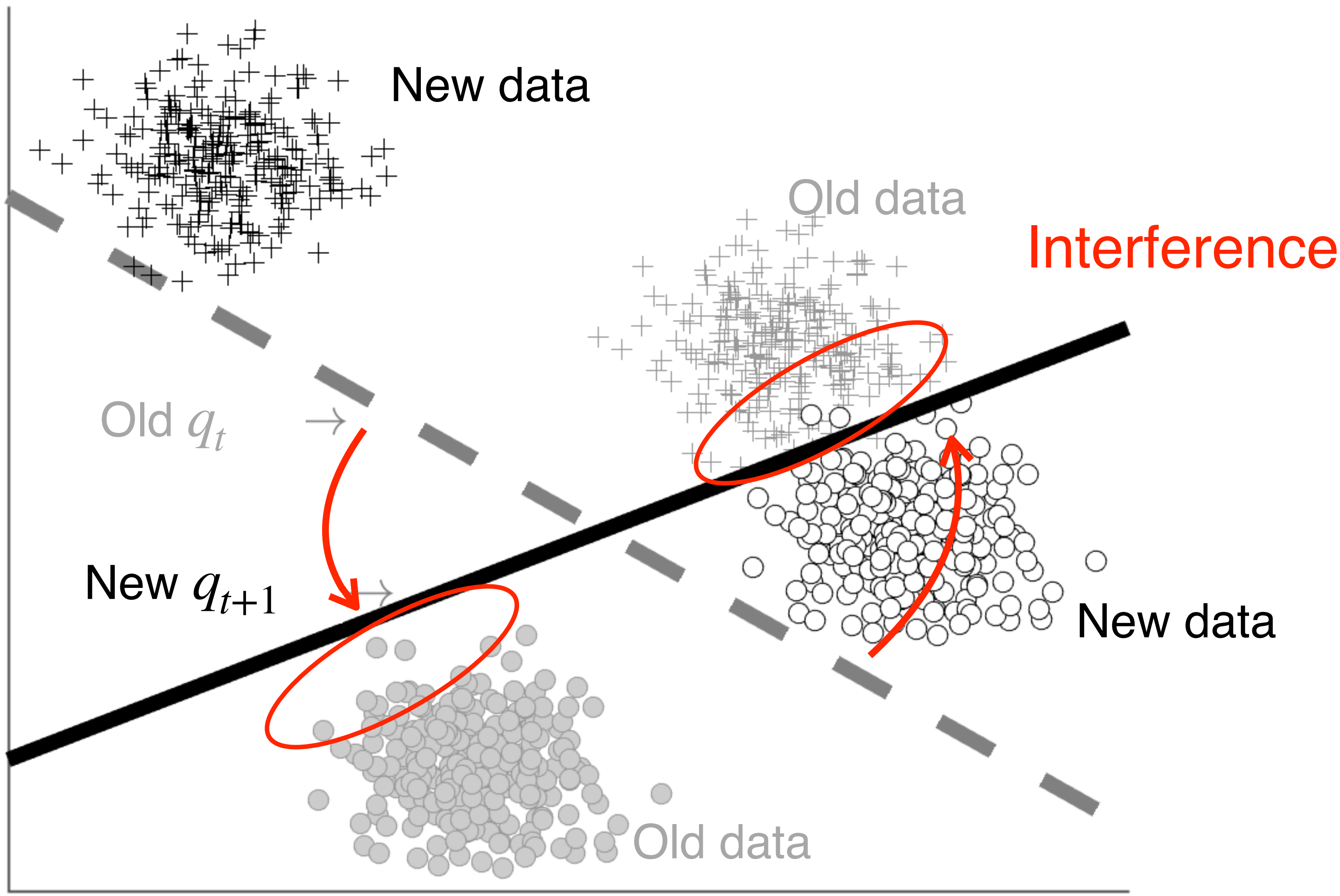


Federated Learning



1. Khan, Knowledge Adaptation as Posterior Correction, arXiv (2025)

Correct the Past due to the Interference Created by the Future



Posterior Correction

We will use the site functions to correct the posterior!

$$\frac{q_t}{\prod_{i=0}^{t} \exp(-\hat{\ell}_{j|t})}$$

Batch: $$q_{t+1} = \arg\min_q \sum_{j=1}^{t+1} \mathbb{E}_q[\ell_j] + KL(q\|p_0)$$

$$= \arg\min_q \; \mathbb{E}_q[\ell_{t+1}] + KL(q\|q_t) + \sum_{j=0}^{t} \mathbb{E}_q[\ell_j - \hat{\ell}_{j|t}]$$

Correction

Online: $$\hat{q}_{t+1} = \arg\min_q \; \mathbb{E}_q[\ell_{t+1}] + KL(q\|q_t)$$

Very simple proof (3 lines). Exact recovery in general!

1. Khan, Knowledge Adaptation as Posterior Approximation, arXiv (2025)

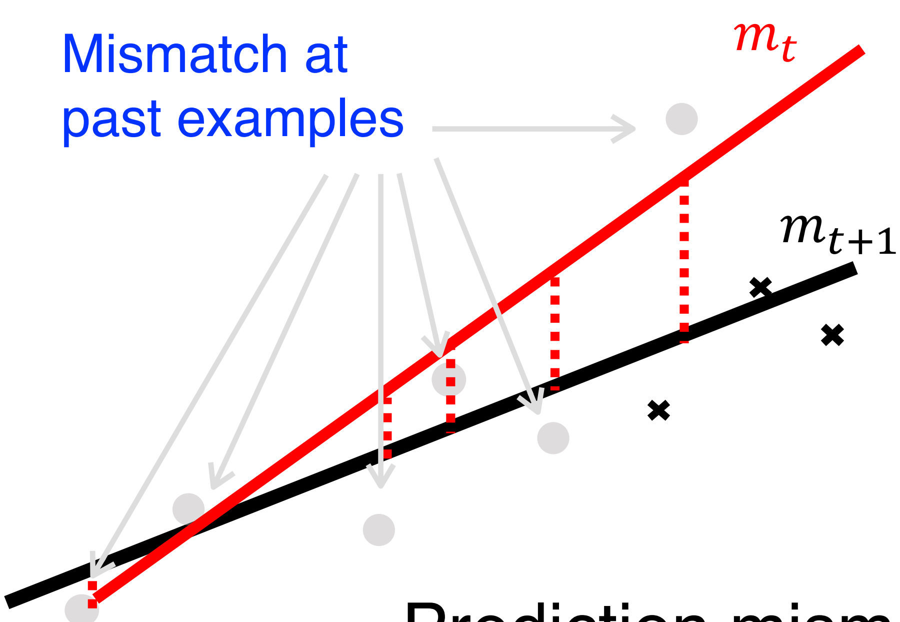
Eq. 7 in Khan (2025)

Correction as Prediction Mismatch

Linear regression with isotropic Gaussian posterior

$$m_{t+1} = \arg\min_{m} \; \mathbb{E}_q[\frac{1}{2}(y_{t+1} - x_{t+1}^\top \theta)^2] + KL\left[\mathcal{N}(m, I) \| \mathcal{N}(m_t, I)\right]$$

$$+ \sum_{j=1}^{t} \frac{1}{2}(x_j^\top m_t - x_j^\top m)^2 \; + \dots$$



Error due to mean-field is fixed by the correction!

$$\frac{1}{2}(m - m_t)^\top \Big(\sum_{j=1}^{t} x_j x_j^\top \Big)(m - m_t)$$

Prediction mismatch is simpler to implement!

Knowledge-Adaptation Prior

Posterior correction with isotropic Gaussian reduces to "prediction or gradient mismatch" (K-priors) [1]

$$m_{t+1} = \arg\min_{m} \ell_{t+1} + \frac{\rho}{2}\|m - m_t\|^2 + \sum_{j=1}^{t} \ell_j\left(\hat{y}_j(m_t),\ \hat{y}_j(m)\right)$$

Many adaptation methods (assuming linearity) reduce this mismatch [2-8] & Posterior Correction generalizes it!

1. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS (2021).
2. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.
3. Benjamin et al. Measuring and regularizing networks in function space. ICLR 2019.
4. Buzzega et al. Dark experience for general continual learning: a strong, simple baseline. NeurIPS 2020.
5. Cauwenberghs and Poggio. Incremental and decremental SVM learning. NeurIPS, 2001.
6. Vapnik and Izmailov. Learning using privileged information: similarity control and …. JMLR, 2015.
7. Lopez-Paz and Ranzato. Gradient episodic memory for continual learning, NIPS'17
8. Csató and Opper. Sparse on-line Gaussian processes. Neural computation, 2002.

Generalization to Non-Linear Cases

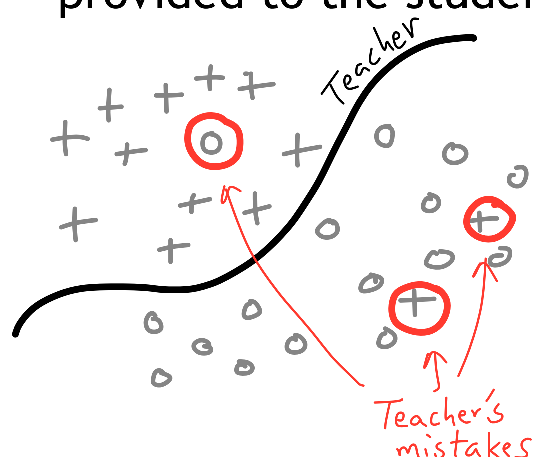
Requires an additional effort to "avoid past mistakes"

$$m_{t+1} = \arg\min_m \ell_{t+1} + \frac{\rho}{2}\|m - m_t\|^2 + \sum_{j=1}^{t} \ell_j\left(\hat{y}_j(m_t),\ \hat{y}_j(m)\right)$$

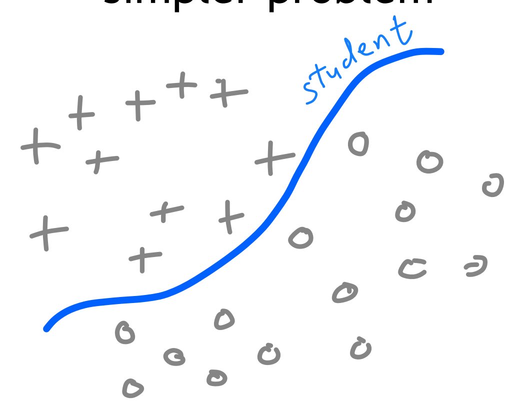
$$+ \sum_{i=1}^{t} r_{i|t}\left[f_i(m) - f_i^{lin}(m)\right]$$

Similar to student-teacher learning [1,2]

Teacher's mistakes provided to the student



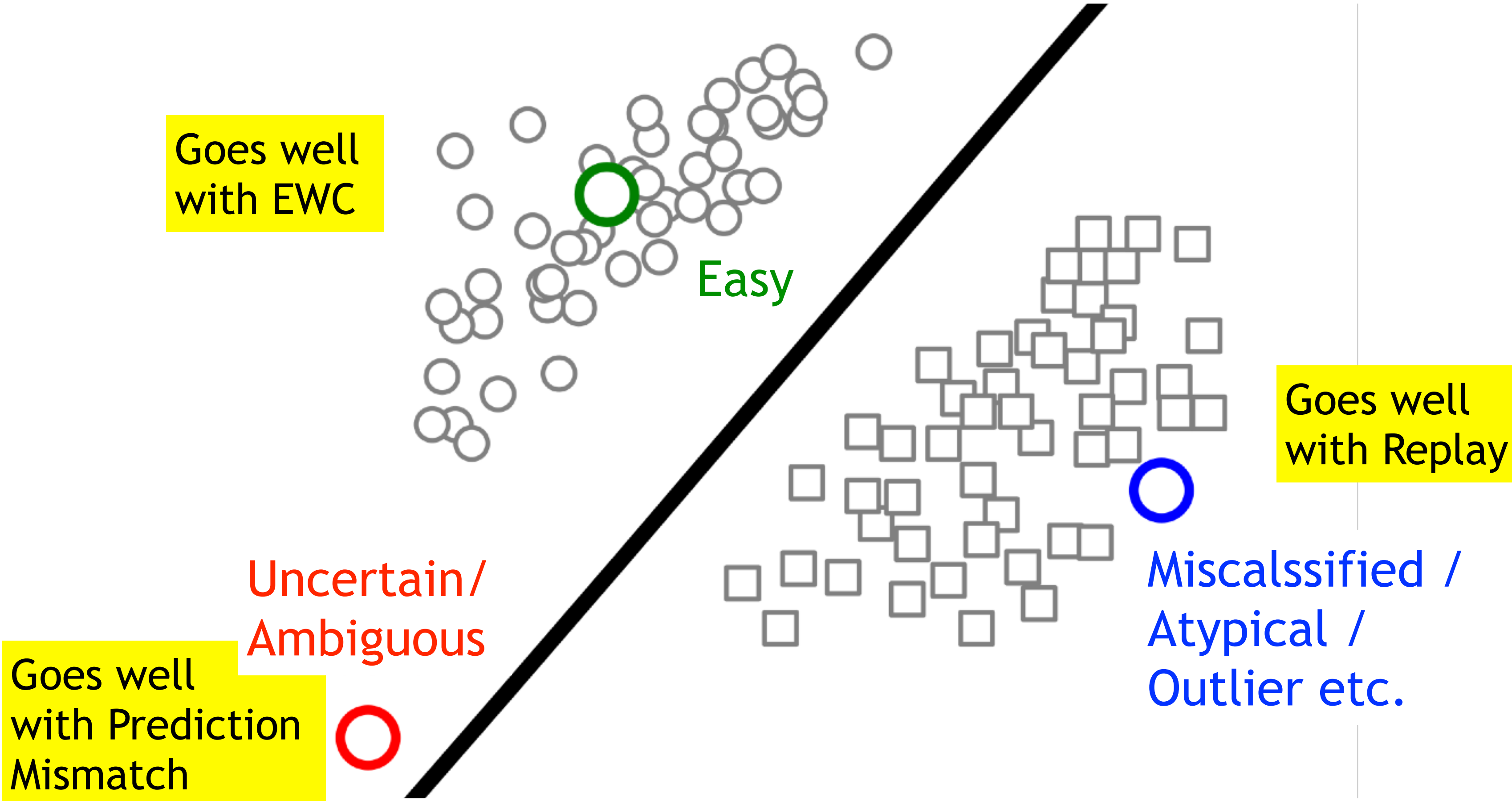
Student solves a simpler problem



1. Hinton et al. Distilling the knowledge in a neural network, arXiv, 2015.
2. Vapnik and Izmailov. Learning using privileged information: similarity control and …. JMLR, 2015.

Three types of Examples

Very similar to Support Vectors!

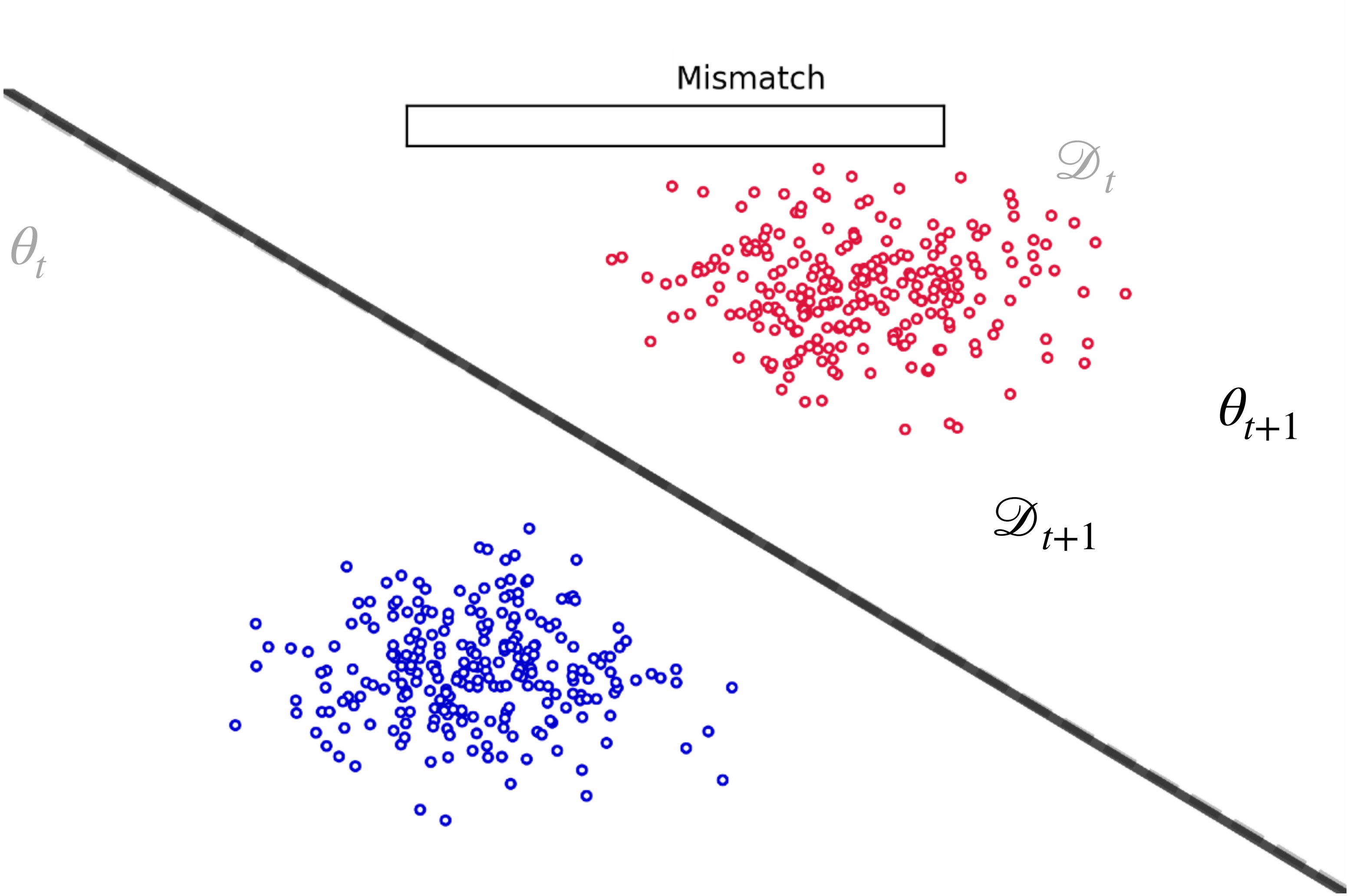


How to Solve Adaptation!

- Three kinds of regularizations required for three different kinds of examples
    1. Weight regularization for examples where both feature and predictions do not change
    2. Prediction matching handles examples where features are static but predictions need adjustments
    3. Memory replay handles examples with large prediction errors and dynamic features
- Any adaptive learning require a balance these three
- Memory requirements increase as we move from 1 to 3.
- These sets characterize the difficulty of adaptations.

From Quick to Slow Adaptation

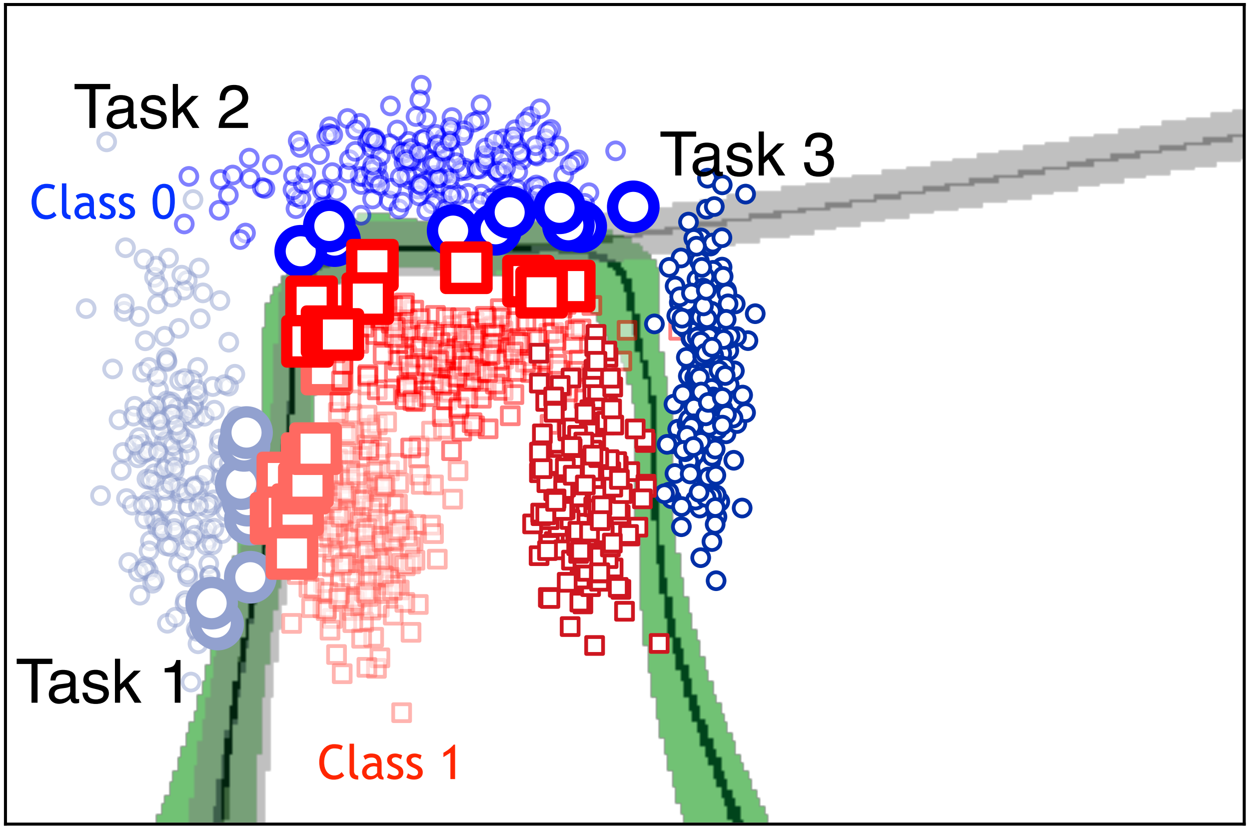
Correction as Information Gain



Quick Adaptation with Compact Memory

Choose memories where interference is more likely.

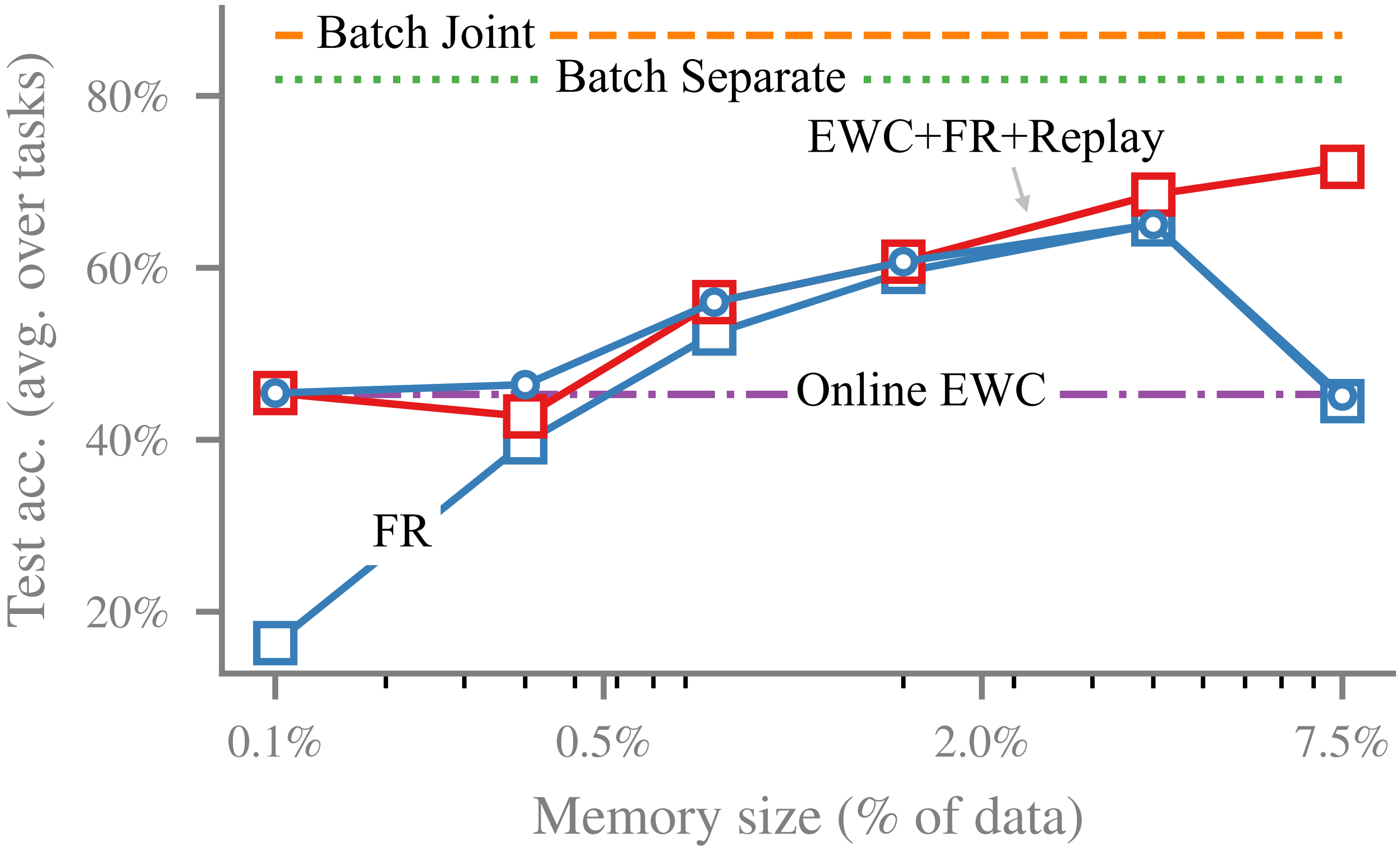
Small correction $\Longrightarrow$ Small memory $\Longrightarrow$ Quick adaptation



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

Combine Methods to Reduce Correction

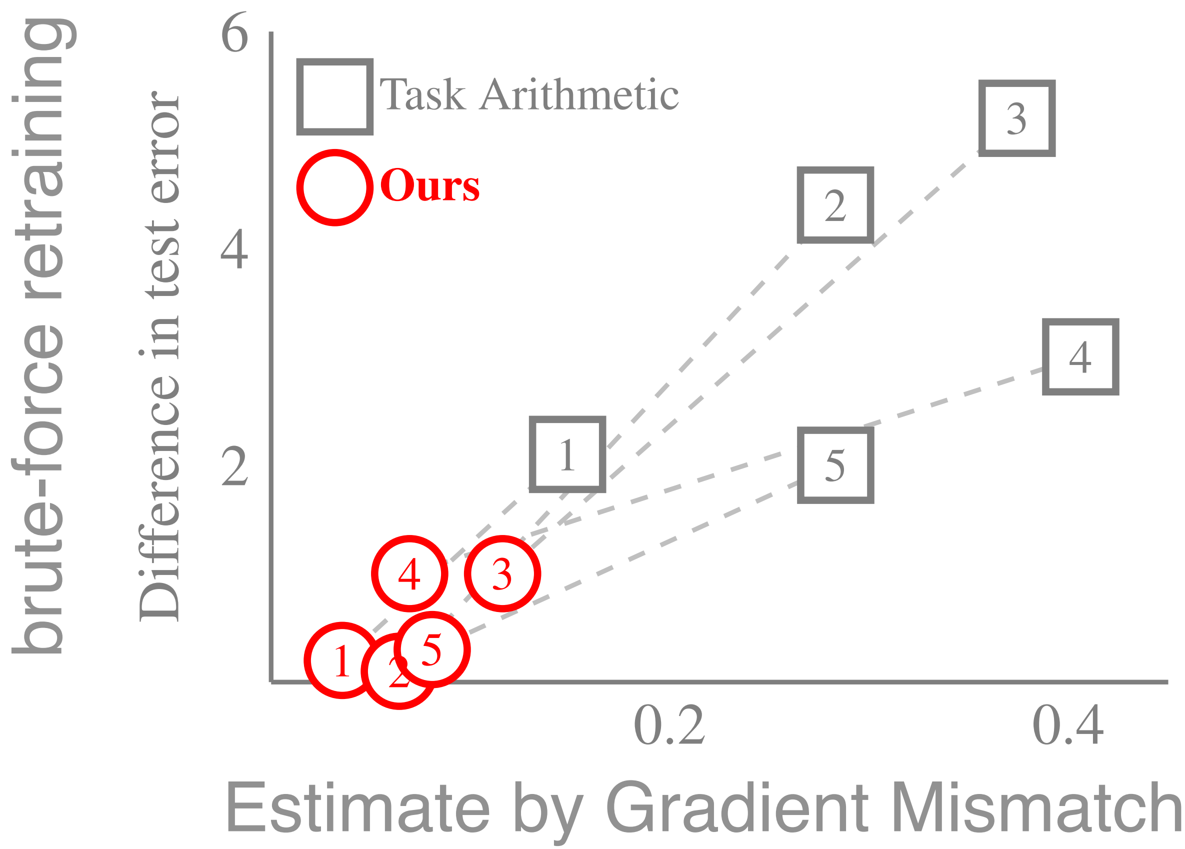
Get 78% accuracy with 7.5% (random) memory



ImageNet with ResNet-18 [1]

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

Reducing Correction Improves Performance in LLM fine-tuning

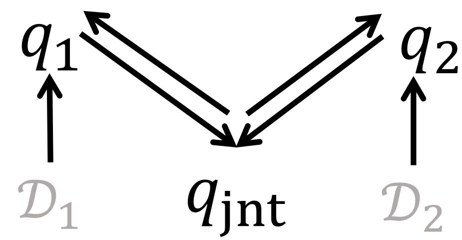


RoBERTa on IMDB

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

Summary of Federated Learning, Model Merging, and Memories etc.

Recover q_{jnt} from q_1 and q_2

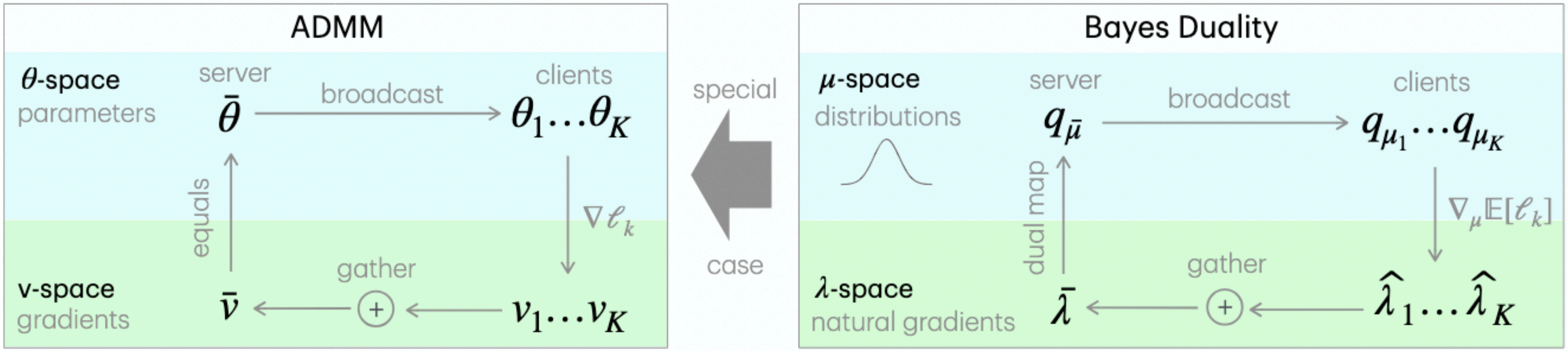


$$q_{jnt} = \arg\min_{q} KL(q\|q_1 q_2) + \sum_{j=1}^{2} \mathbb{E}_q[\ell_j - \hat{\ell}_{j|j}]$$

By choosing different q, we get different strategies (better q gives better merging) [1,2]. Same is true for federated learning [3,4]. All of them will benefit from compact memories designed to reduce corrections [5].

1. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
2. Monzon et al. How to Weight Multitask Finetuning? Fast Previews via Bayesian Model-Merging, 2024
3. Swaroop, Khan, Doshi, Connecting Federated ADMM to Bayes, ICLR 2025
4. Moellenhoff et al. Federated ADMM from Bayes Duality, arXiv, 2025
5. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)

ADMM as a special case of Bayes (Dual)



Algorithm 1 BayesADMM (Fig. 2b) for Gaussians with diagonal covariance. Additional steps when compared to FederatedADMM are highlighted in red. Implementation details are in App. D.

Hyperparameters: Prior precision $\delta > 0$, step-sizes $\rho > 0$ and $\gamma > 0$.
Initialize: $\mathbf{v}_k \leftarrow 0$, $\mathbf{u}_k \leftarrow 0$, $\bar{\mathbf{m}} \leftarrow 0$, $\bar{\mathbf{s}} \leftarrow \delta$, $\alpha \leftarrow 1/(1+\rho K)$.

while not converged **do**
  Broadcast $\bar{\mathbf{m}}$ and $\bar{\mathbf{s}}$ to all clients.
  for each client $1, \ldots, K$ in parallel **do**
    Local training on $\ell_k(\boldsymbol{\theta}) + \boldsymbol{\theta}^\top \mathbf{v}_k - \frac{1}{2}\boldsymbol{\theta}^\top(\mathbf{u}_k \boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta} - \bar{\mathbf{m}}\|^2_{\bar{\mathbf{s}}}$ ▷ Using IVON [53]
    $\mathbf{v}_k \leftarrow \mathbf{v}_k + \gamma(\mathbf{s}_k \mathbf{m}_k - \bar{\mathbf{s}}\bar{\mathbf{m}})$
    $\mathbf{u}_k \leftarrow \mathbf{u}_k + \gamma(\mathbf{s}_k - \bar{\mathbf{s}})$ ▷ An additional dual variable.
  end for
  Gather $\mathbf{m}_k$, $\mathbf{v}_k$ and $\mathbf{s}_k$, $\mathbf{u}_k$ from all clients.
  $\bar{\mathbf{m}} \leftarrow (1-\alpha)\,\texttt{Mean}(\mathbf{s}_{1:K}\mathbf{m}_{1:K}) + \alpha\,\texttt{Sum}(\mathbf{v}_{1:K})$
  $\bar{\mathbf{s}} \leftarrow (1-\alpha)\,\texttt{Mean}(\mathbf{s}_{1:K}) + \alpha\,[\delta\mathbf{1} + \texttt{Sum}(\mathbf{u}_{1:K})]$ ▷ Two additional steps for precision $\bar{\mathbf{s}}$
  $\bar{\mathbf{m}} \leftarrow \bar{\mathbf{m}}/\bar{\mathbf{s}}$
end while

Adaptive Bayesian Intelligence

- Adaptive Intelligence = Bayesian Computation
- Part 1: Bayesian Learning Rule [1]
  - Foundational way to derive learning-algorithms
  - Application to Deep Learning [2]
- Part 2: Posterior Correction [3]
  - Foundational way to derive adaptation-algorithms
  - Application to continual learning [4-5]
  - But also for LLM merging, Federated Learning etc.
- Adaptive Bayesian Intelligence: A roadmap.

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

Questions for the future

- What should the algorithm remember?
- And what new experiences should it seek?
- Memory should be chosen to minimize the corrections that may arise in the future.
- New experiences should be chosen to enable easy-enough corrections (not too daunting for the learner)
- Future is unknown but the algorithm has the freedom to explore by “fixing the past & choosing the future”

Fixing
Choosing