Presentation at Workshop on Machine Learning and Applications at NUS+IMS, Oct 17, 2022

The Bayesian Learning Rule for Adaptive AI

Mohammad Emtiyaz Khan

RIKEN Center for AI Project, Tokyo

http://emtiyaz.github.io

Summary of research at https://emtiyaz.github.io/papers/symposium_2021.pdf

Slides available at https://emtiyaz.github.io/papers/Oct17_2022_NUS.pdf

AI that learn like humans

Quickly adapt to learn new skills, throughout their lives

Human Learning at the age of 6 months.

Converged at the age of 12 months

Transfer skills

at the age of 14 months

Fail because too slow to adapt

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

Adaptation in Machine Learning

- Machines are bad in quickly adapting to changes
  - Even small changes require a complete retraining-from-scratch
  - This is expensive, time consuming [1,2]
  - Example: Tesla AI Data-Engine for “self-driving cars” takes 70000 GPU hrs [3]
- Difficult to apply to domains with “dynamic” setting
  - Robotics, medicine, user interaction, epidemiology, climate science, etc.

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

July 14, 2021

Yann LeCun @ylecun · 7h

So many exciting new frontiers in ML, it's hard to give a short list, particularly in new application areas (e.g. in the physical and biological sciences).

But the Big Question is:
"How could machines learn as efficiently as humans and animals?"
This requires new paradigms.

“Solving” Adaptation

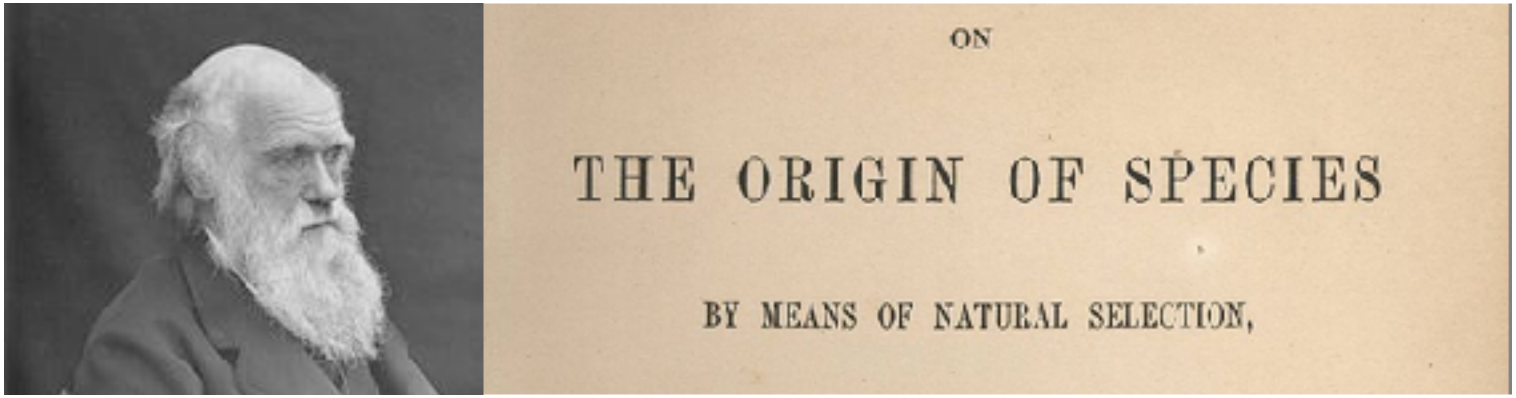
New learning principles to answer “When and how can a model quickly adapt?”

Today's talk

- New Learning Principles for Adaptive AI
- Unify algorithms with the Bayesian Learning rule (BLR) [1]
  – New work: SAM as Bayes [2]
- BLR's "dual" perspective to "solve" adaptation,
  – Bayesian Duality Principle [3, 8]
  – Continual learning with memory [4,5,6,7]
  – Reduce dependence on large data and compute

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
2. Moellenhoff and Khan, SAM as an optimal relaxation of Bayes, https://arxiv.org/abs/2210.01620, 2022
3. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
4. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020
5. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)
6. Daxberger et al., Improving CL by using the Principle of Gradient Reconstructions, Under review, 2022
7. Tailor, Chang, Swaroop, Solin, Khan. Memorable experiences of ML models (in preparation)
8. Khan, Bayesian duality principle (in preparation)

See Section 6 (discussion) in Khan and Rue, 2021



The Origin of Algorithms

What are the common principles behind “good” algorithms?

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021

The Bayesian Learning Rule

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

Håvard Rue
CEMSE Division, KAUST
Thuwal, Saudi Arabia
haavard.rue@kaust.edu.sa

Abstract

We show that many machine-learning algorithms are specific instances of a single algorithm called the *Bayesian learning rule*. The rule, derived from Bayesian principles, yields a wide-range of algorithms from fields such as optimization, deep learning, and graphical models. This includes classical algorithms such as ridge regression, Newton's method, and Kalman filter, as well as modern deep-learning algorithms such as stochastic-gradient descent, RMSprop, and Dropout. The key idea in deriving such algorithms is to approximate the posterior using candidate distributions estimated by using natural gradients. Different candidate distributions result in different algorithms and further approximations to natural gradients give rise to variants of those algorithms. Our work not only unifies, generalizes, and improves existing algorithms, but also helps us design new ones.

Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021

Bayesian learning rule

See Table 1 in Khan and Rue, 2021

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

A Bayesian Origin

E_q[log-lik] – $KL(q\,||\,$prior$)$

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

Entropy

Posterior approximation (expo-family)

Bayesian Learning Rule [1,2]

Natural and Expectation parameters of q

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

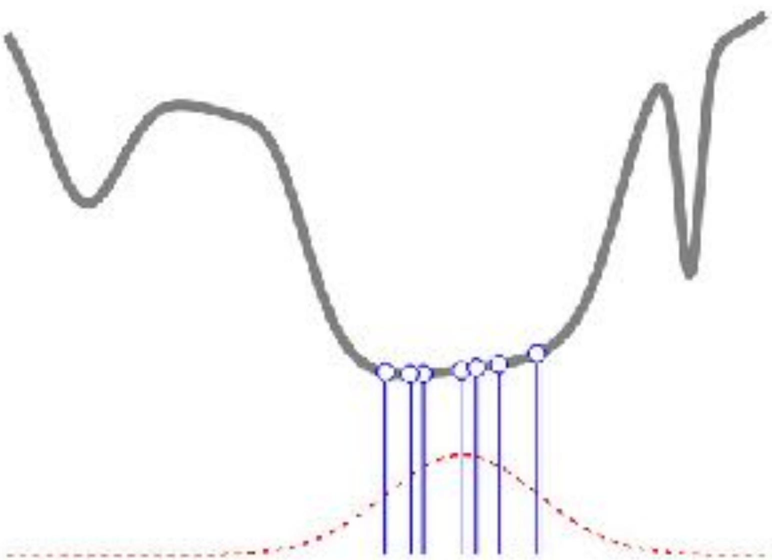
Natural gradients (information geometry)

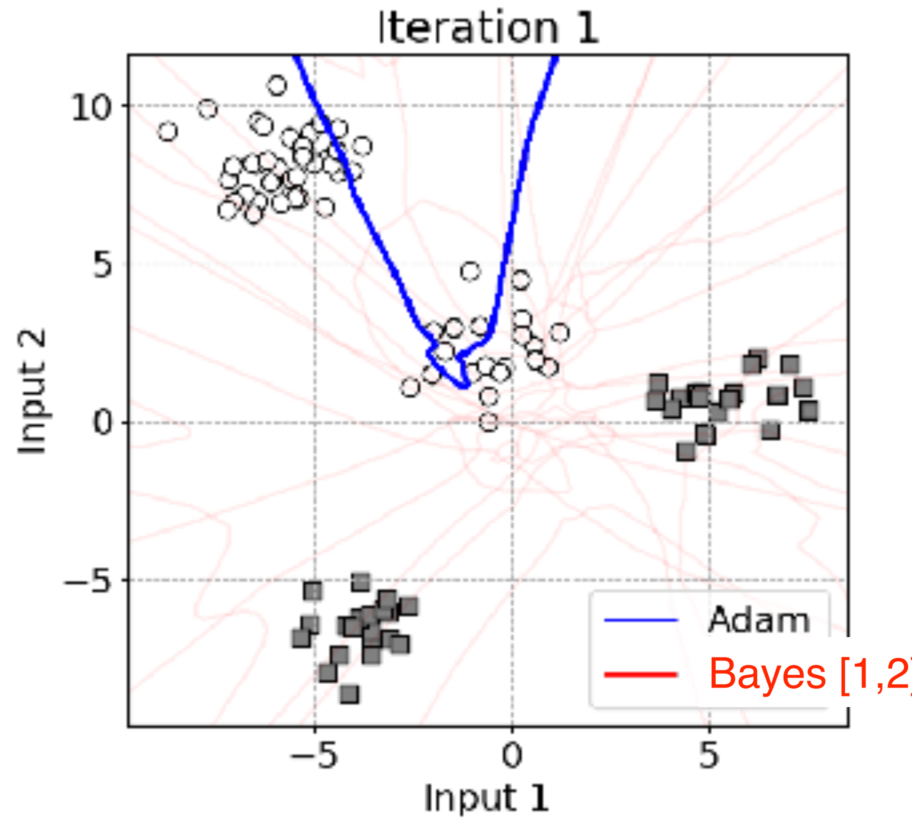
Many existing algorithms can be seen as special instances of the BLR, by using approximations to q and natural gradients.

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
2. Khan and Lin. "Conjugate-computation variational inference…." AIstats (2017).

Why use Bayesian averaging?

Choose an "ensemble" of almost equally good models (similar to sampling in SGD trajectories)





1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. "Practical Deep Learning with Bayesian Principles." NeurIPS (2019).

Deep Learning with the BLR

RMSprop

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

BLR variant called VOGN

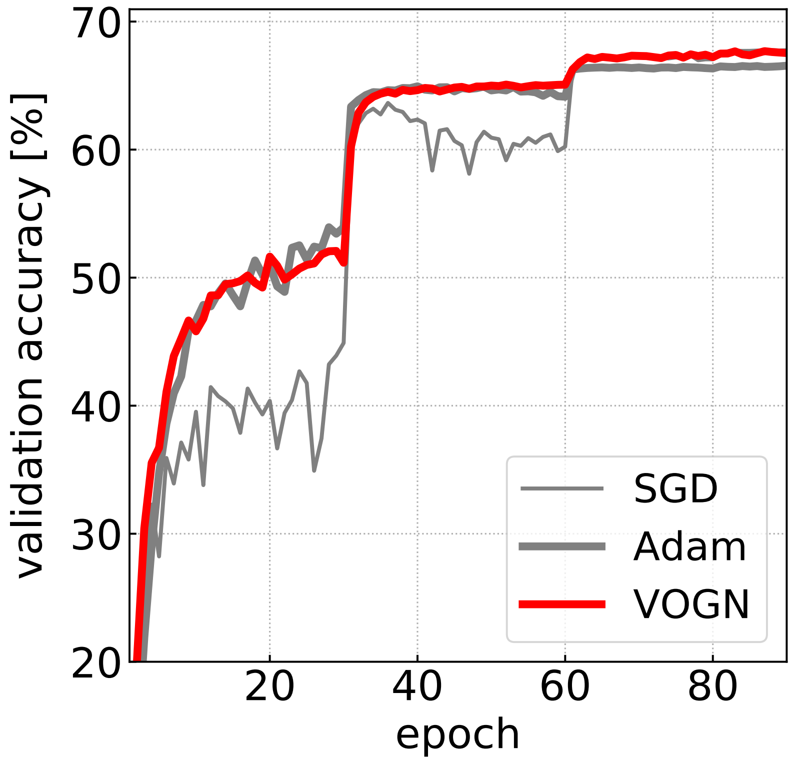
$$g \leftarrow \hat{\nabla}\ell(\theta), \text{ where } \theta \sim \mathcal{N}(m, \sigma^2)$$
$$s \leftarrow (1-\rho)s + \rho(\Sigma_i g_i^2)$$
$$m \leftarrow m - \alpha(s+\gamma)^{-1}\nabla_\theta \ell(\theta)$$
$$\sigma^2 \leftarrow (s+\gamma)^{-1}$$

Available at https://github.com/team-approx-bayes/dl-with-bayes

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

Uncertainty of Deep Nets

VOGN: A modification of Adam with similar performance on ImageNet, but better uncertainty



Code available at https://github.com/team-approx-bayes/dl-with-bayes

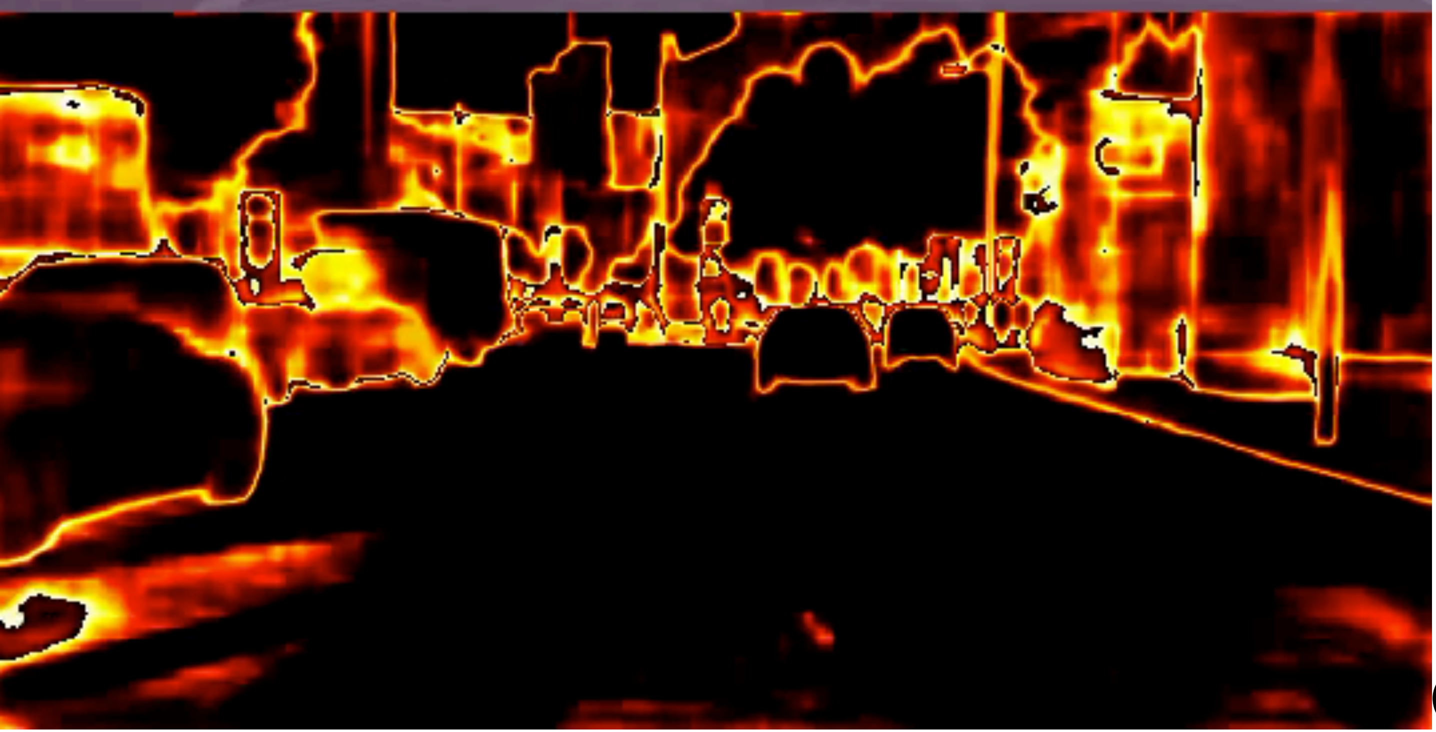
1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. "Practical Deep Learning with Bayesian Principles." NeurIPS (2019).
3. Lin et al. "Handling the positive-definite constraints in the BLR." ICML (2020).

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

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

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

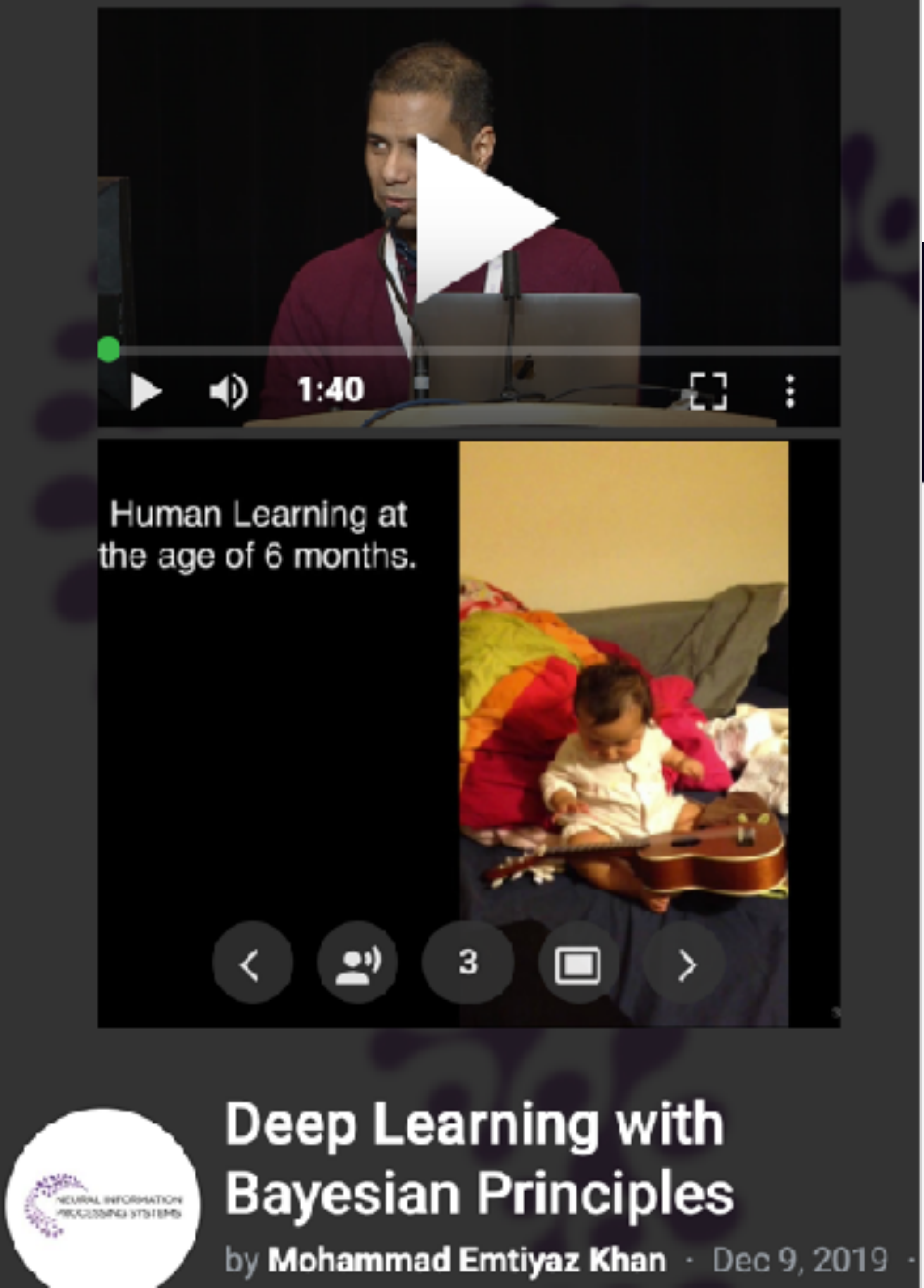
Image Segmentation

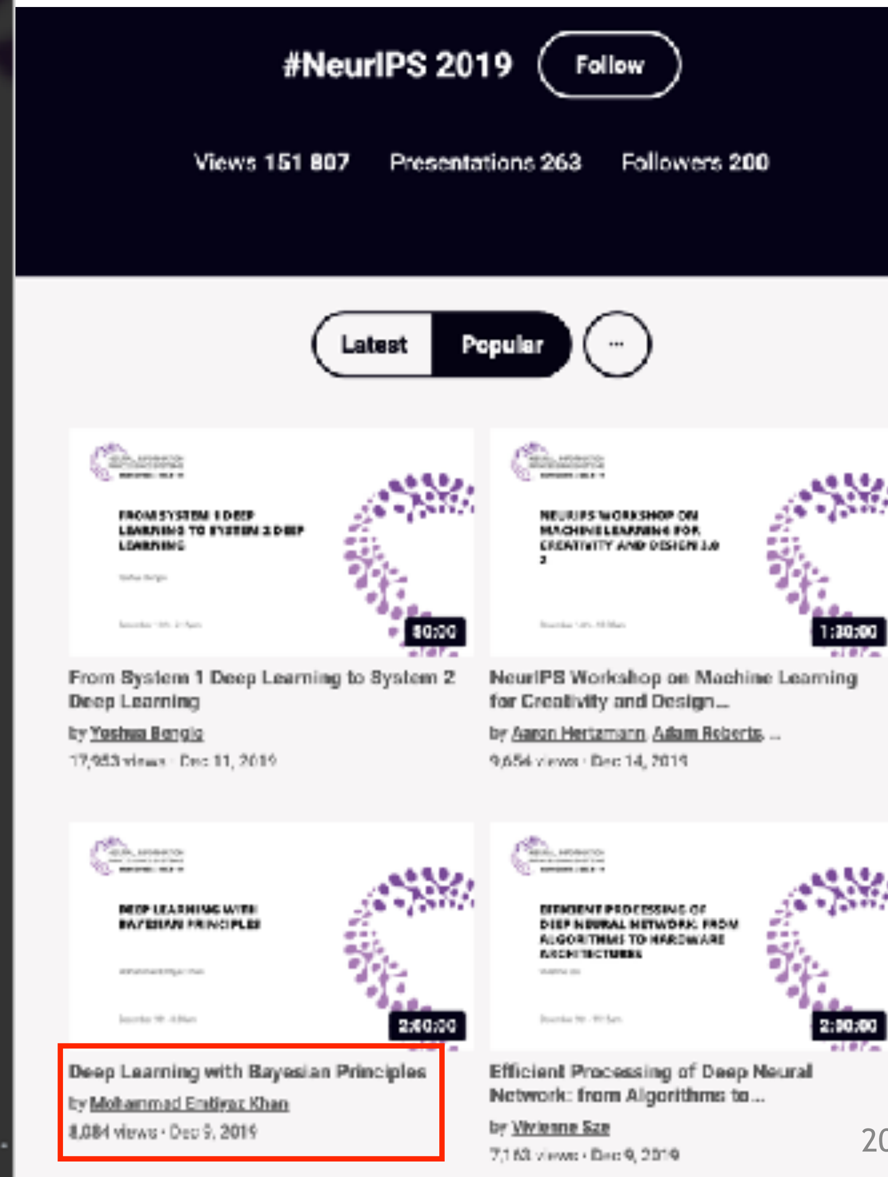


Uncertainty (entropy of class probs)

(By Roman Bachmann)

NeurIPS 2019 Tutorial





Robust DL with Bayes

Adding uncertainty to Adversarial Weight-perturbation methods

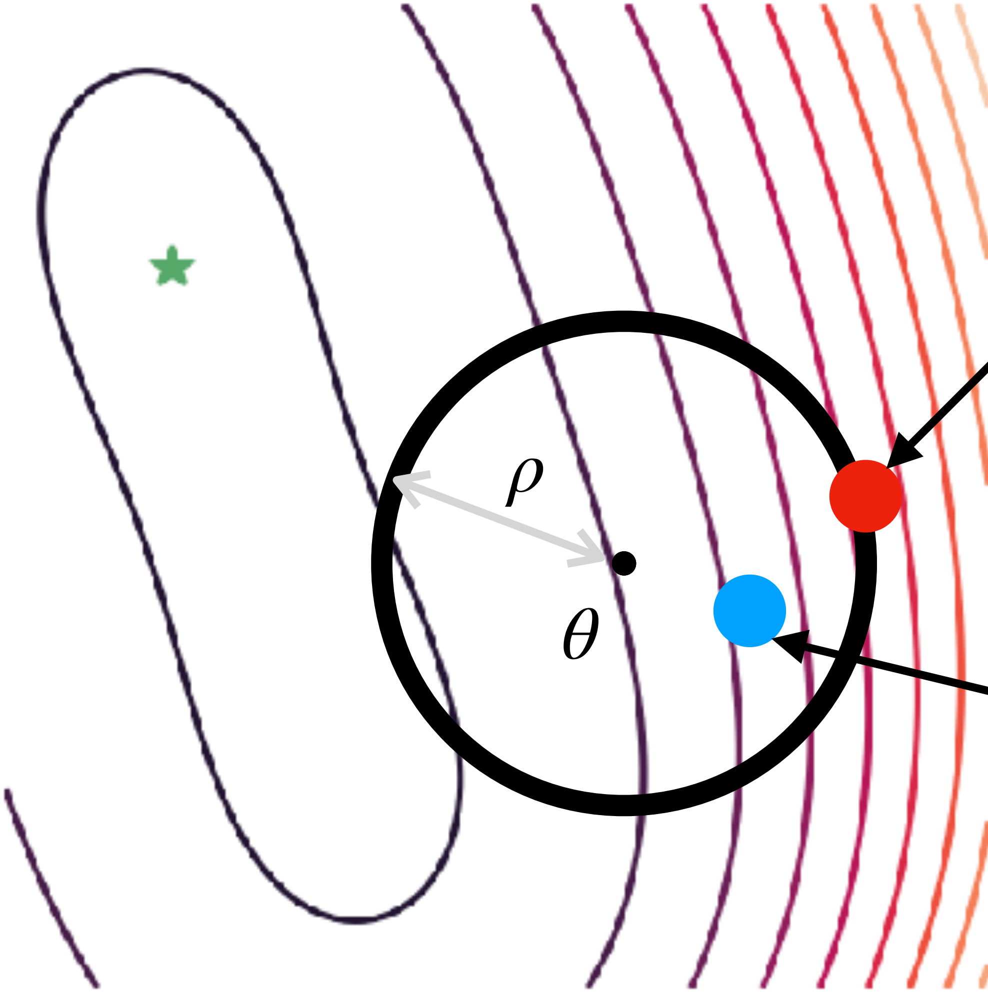
Thomas Moellenhoff

Robust Deep-Learning

- Sharpness-Aware Minimization (SAM)[1]
  - Huge improvements over SGD/Adam
  - Now used to train all sorts of models
  - Why does it work, and how to improve it?
- SAM as an “optimal” relaxation of Bayes [2]

1. Foret et al. Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR, 2021
2. Moellenhoff and Khan, SAM as an optimal relaxation of Bayes, https://arxiv.org/abs/2210.01620, 2022

SAM as an Optimal relaxation of Bayes



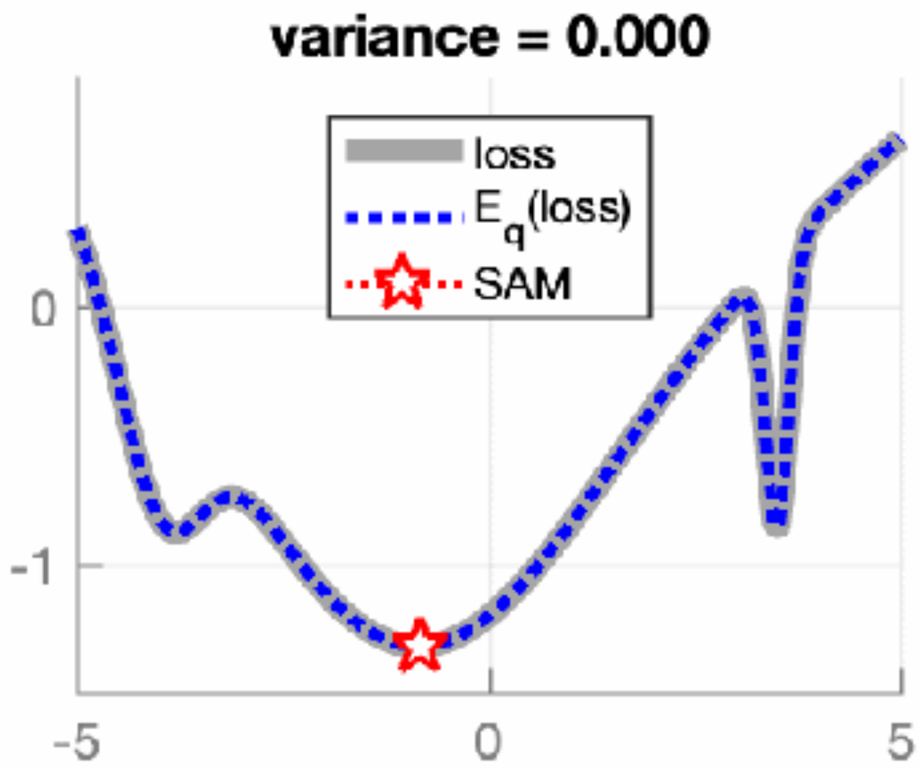
SAM: $\sup_{|\epsilon|<\rho} \ell(\theta + \epsilon)$

Our work: Fenchel Biconjugate

Bayes: $\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)}[\ell(\theta + \epsilon)]$

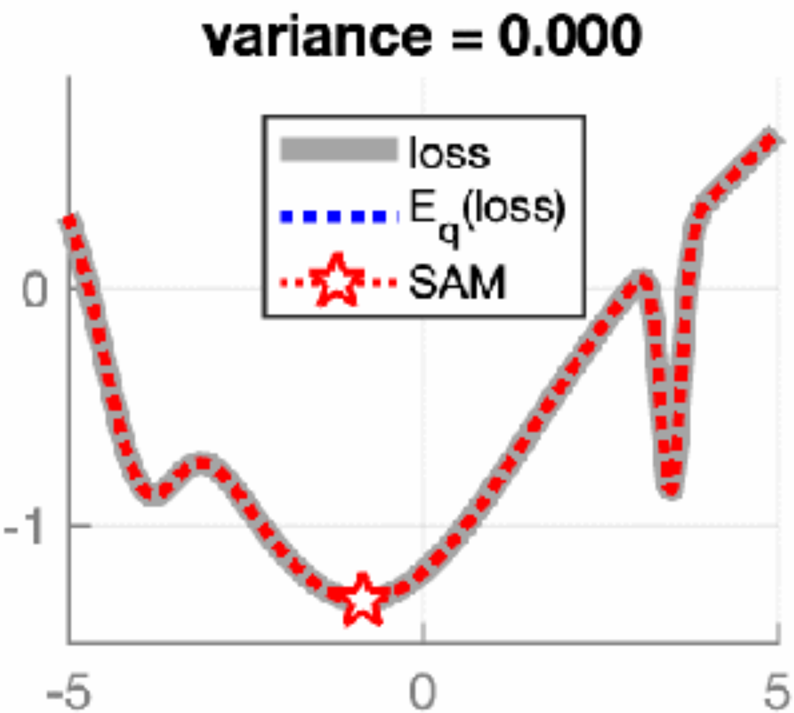
SAM = an Optimal Relaxation of Bayes

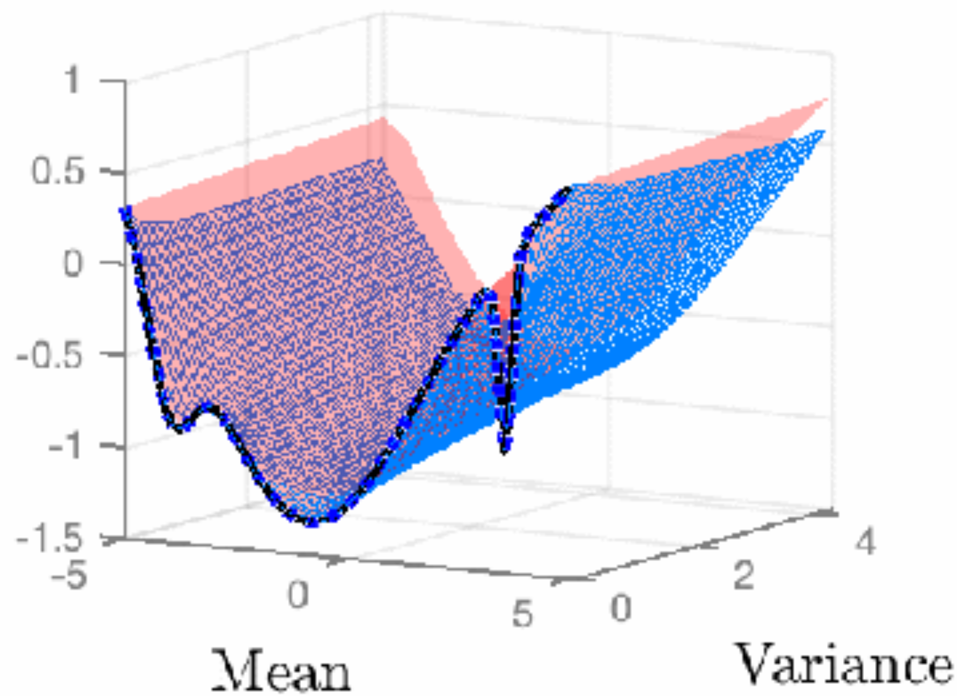
SAM (red star) upper bounds the Bayesian $\mathbb{E}_q[\ell]$



SAM = an Optimal Relaxation of Bayes

SAM minimizes the best-Concave upper bound to $\mathbb{E}_q[\ell]$ wrt the mean, while keeping variance fixed.





Bayesian-SAM

An Adam-style algorithm, derived using the BLR, where "perturbation-size" is automatically found using σ^2 (or s)

SAM with RMSprop

$$
\begin{aligned}
g_1 &\leftarrow \hat{\nabla}\ell(\theta) \\
\epsilon &\leftarrow \rho \frac{g_1}{\|g_1\|} \\
g &\leftarrow \hat{\nabla}\ell(\theta+\epsilon) \\
s &\leftarrow (1-\rho)s + \rho g^2 \\
\theta &\leftarrow \theta - \alpha(\sqrt{s}+\delta)^{-1} g
\end{aligned}
$$

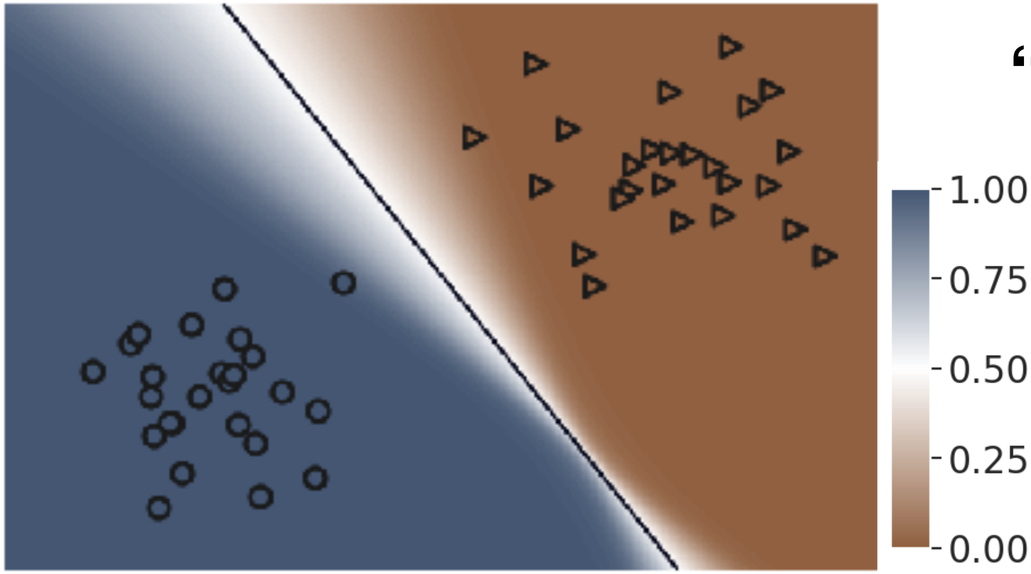
SAM with BLR

$$
\begin{aligned}
g_1 &\leftarrow \hat{\nabla}\ell(\theta), \text{ where } \theta \sim \mathcal{N}(m, \sigma^2) \\
\epsilon &\leftarrow \rho \frac{g_1}{s} \\
g &\leftarrow \hat{\nabla}\ell(\theta+\epsilon) \\
s &\leftarrow (1-\rho)s + \rho \sqrt{s}|g_1| \\
m &\leftarrow m - \alpha(s+\delta)^{-1}\nabla_\theta \ell(\theta) \\
\sigma^2 &\leftarrow (s+\delta)^{-1}
\end{aligned}
$$

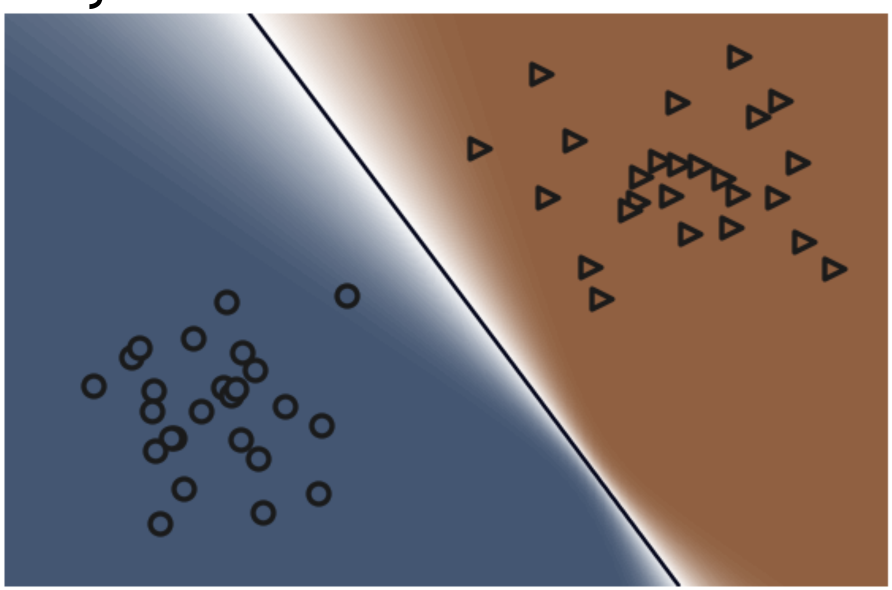
1. Foret et al. Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR, 2021
2. Moellenhoff and Khan, SAM as an optimal relaxation of Bayes, https://arxiv.org/abs/2210.01620, 2022

Improving “overconfident” SAM

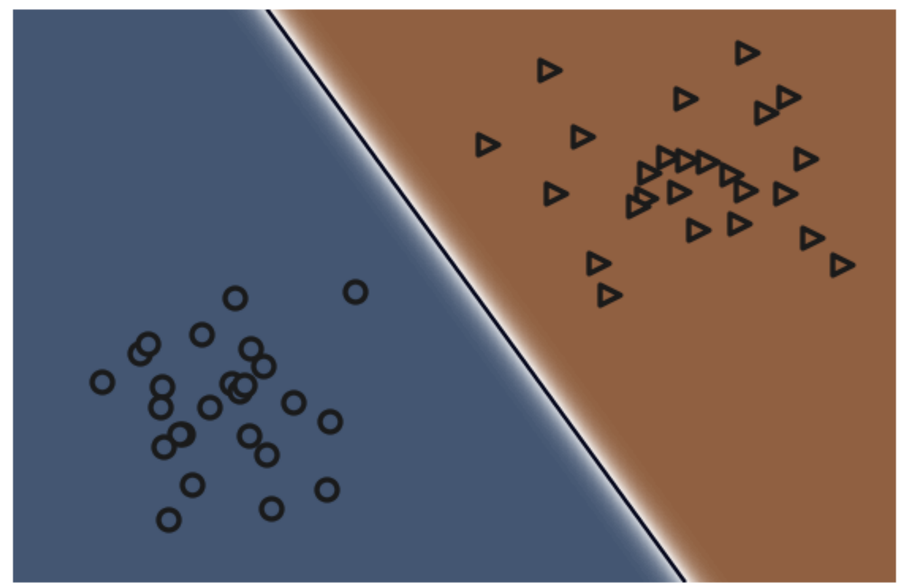
Bayes



Bayesian SAM



SAM



Towards Solving Adaptation

By using a dual perspective of the BLR to solve continual learning

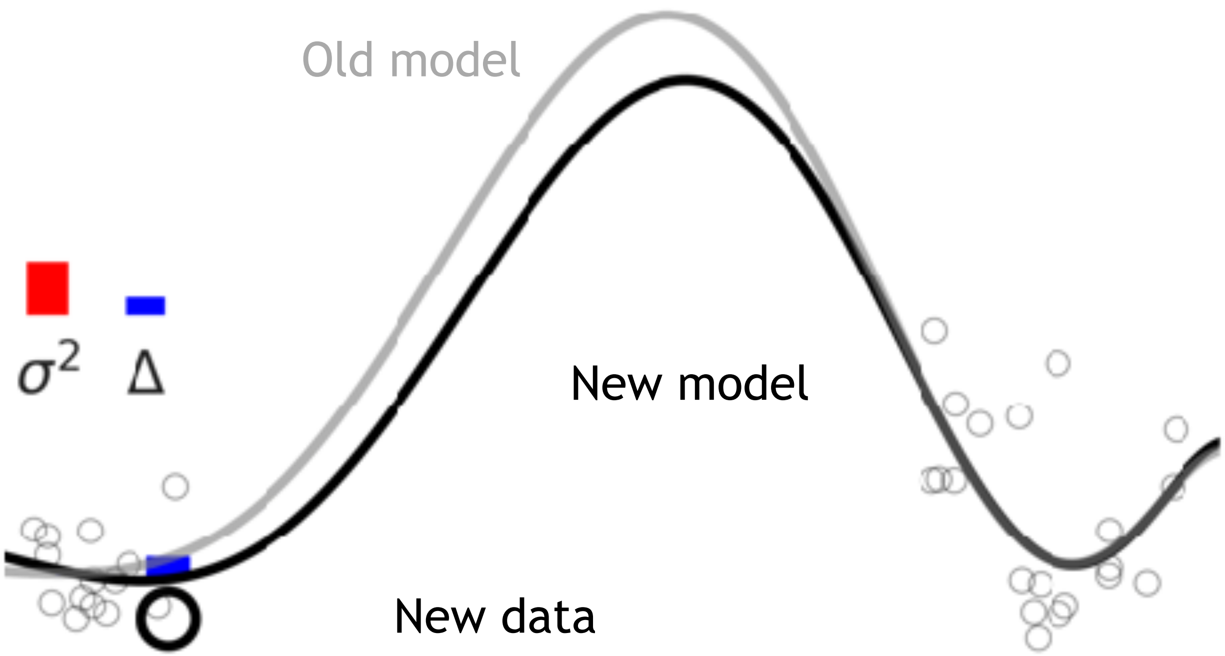
How to adapt the knowledge?
Perturbation, Sensitivity, and Duality

https://tenor.com/view/clockwork-gears-brain-gif-16784329

“Model Change” and Uncertainty

We can “predict” how much a model is going to change by using the uncertainty

“Model-change” (Δ) $\propto$ “Uncertainty (σ^2)”



1. Cook. Detection of Influential Observations in Linear Regression. Technometrics. ASA 1977

See Section 5.4 in Khan and Rue, 2021 for local parameterization
See Section 3 in ADAM et al. 2021 for dual parameterization

BLR Solutions & Their Duality

$$\ell(\theta) = \sum_{i=0}^{N} \ell_i(\theta) \qquad \lambda \leftarrow (1-\rho)\lambda - \sum_{i=0}^{N} \rho \nabla_{\mu} \mathbb{E}_q[\ell_i(\theta)]$$

$$\lambda^* = \sum_{i=0}^{N} \underbrace{\nabla_{\mu^*} \mathbb{E}_{q^*}[-\ell_i(\theta)]}_{\widetilde{\lambda}_i^*}$$

Global and local natural parameter

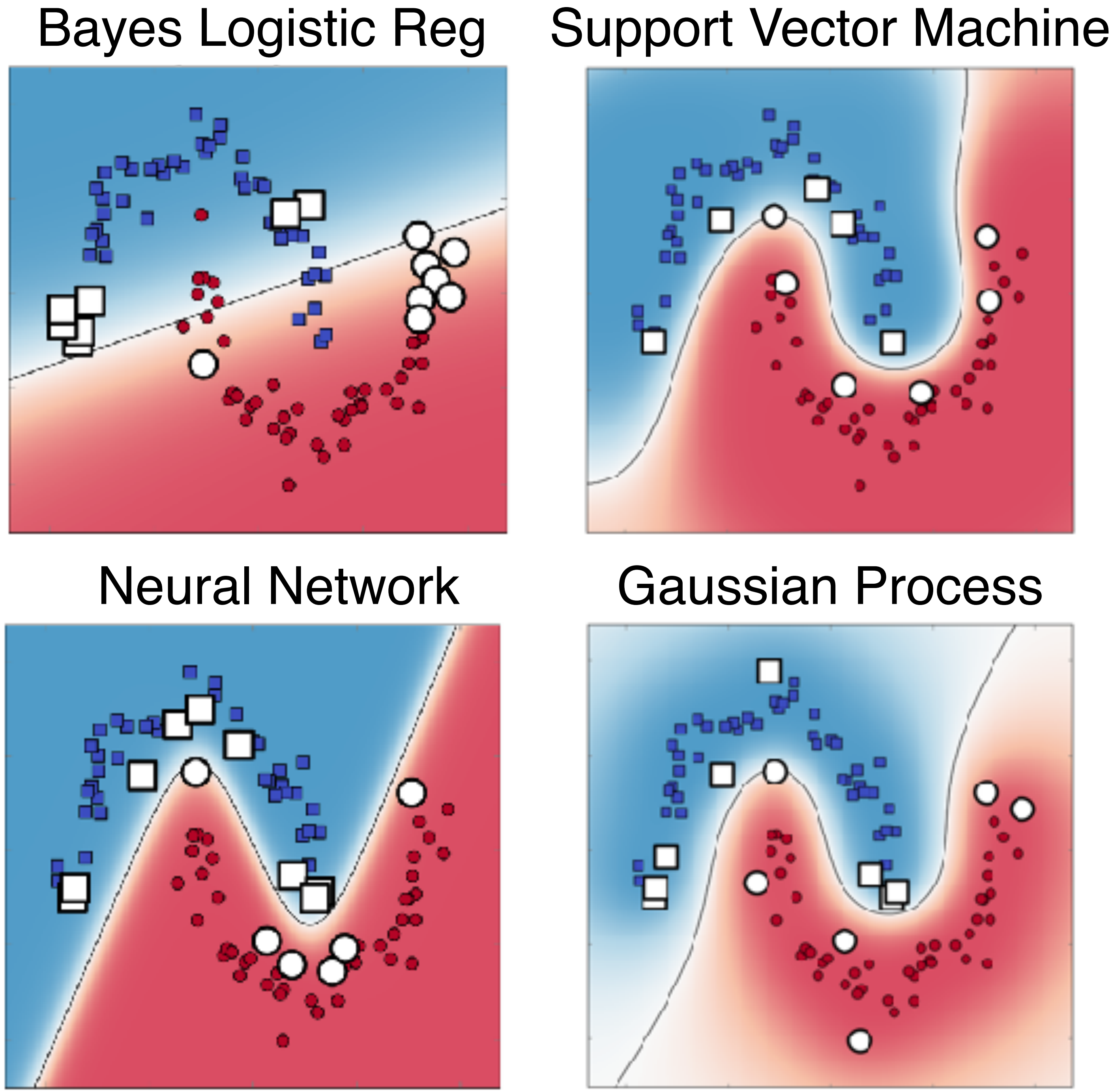
Local parameters are Lagrange Multipliers, measuring the sensitivity of BLR solutions to local perturbation [1,2]. They can be used to tell apart relevant vs irrelevant data.

1. ADAM, Chang, Khan, Solin, Dual parameterization of SVGP, NeurIPS, 2021
2. Khan, Bayesian duality principle, in preparation

"Memorable" Experiences

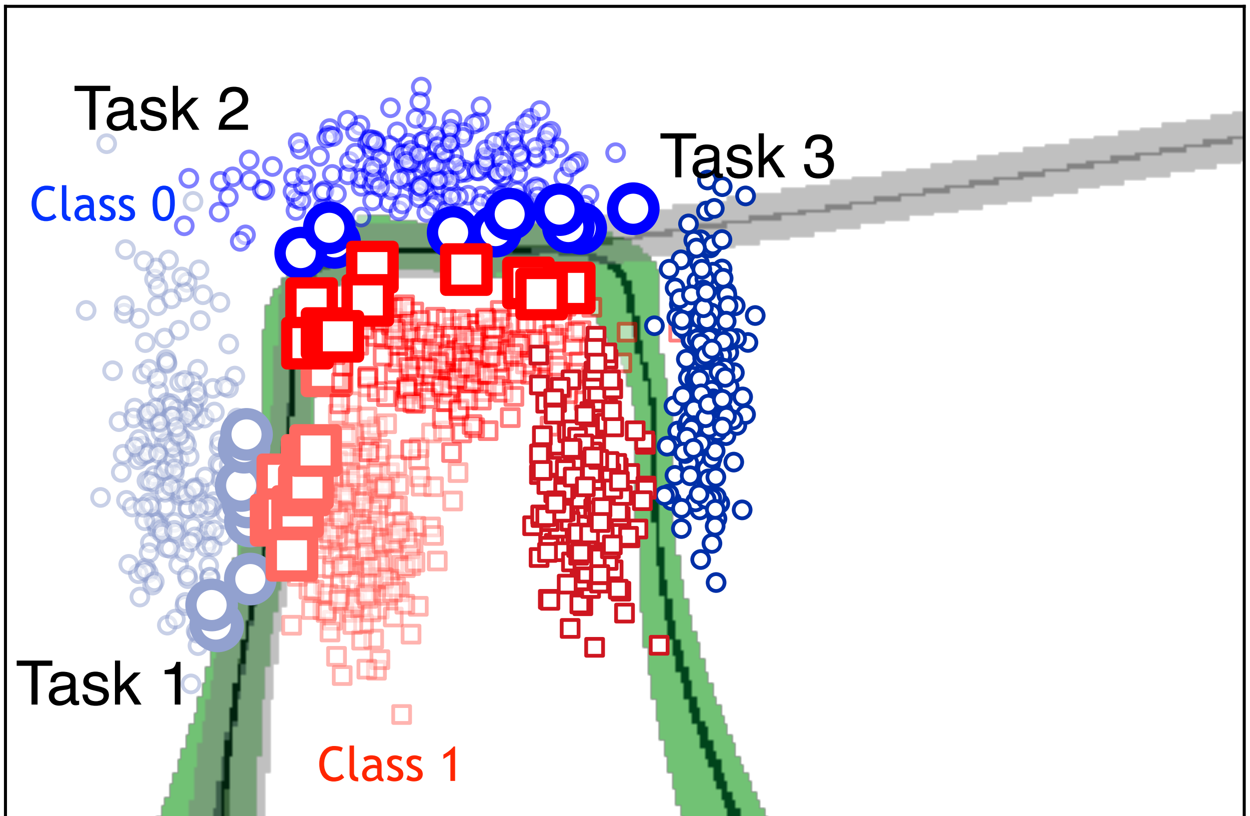
1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020
2. Tailor, Chang, Swaroop, Solin, Khan. Memorable experiences of ML models (in preparation)

The tool applies to a wide variety of ML models, ranging from linear models, SVMs, and neural networks



Continual Learning

Avoid forgetting by using memorable examples [1,2]



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

Functional Regularization of Memorable Past (FROMP) [4]

Previous approaches used weight-regularization [1,2]

$$q_{new}(\theta) = \min_{q \in \mathcal{Q}} \mathbb{E}_{q(\theta)}[\ell_{new}(\theta)] - \mathcal{H}(q) - \mathbb{E}_{q(\theta)}[\log q_{old}(\theta)]$$

New data

Weight-regularizer

Replace it by a functional regularizer using a Dual GP-view of DNNs [2]

$$\mathbb{E}_{\tilde{q}_\theta(\mathbf{f})}[\log \tilde{q}_{\theta_{old}}(\mathbf{f})]$$

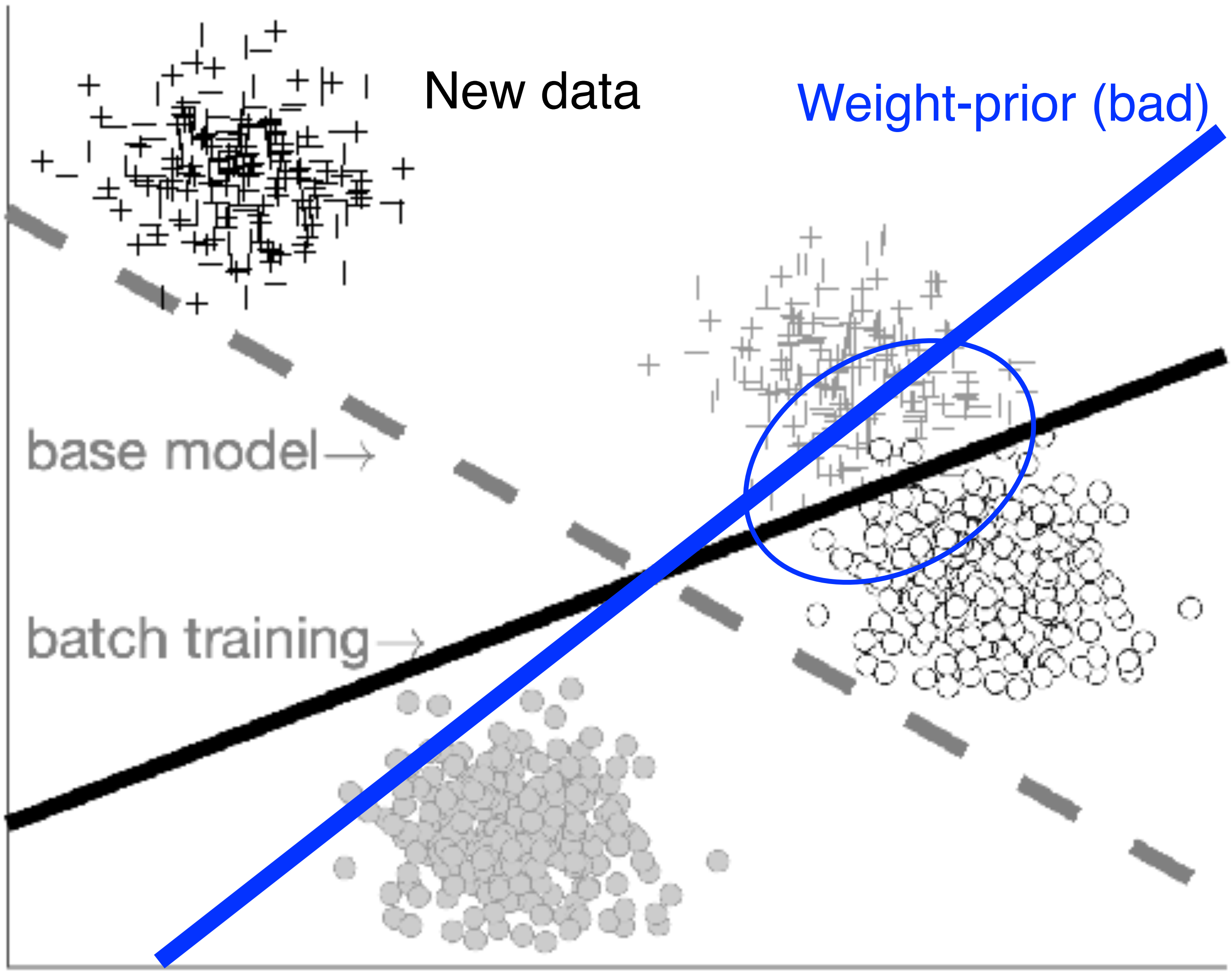
$$[\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]^\top K_{old}^{-1}[\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]$$

Kernels weighs examples according to their memorability

Forces network-outputs to be similar

1. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS* 2017
2. Nguyen et al., Variational Continual Learning, ICLR, 2018
3. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
4. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

How to improve over weight priors?



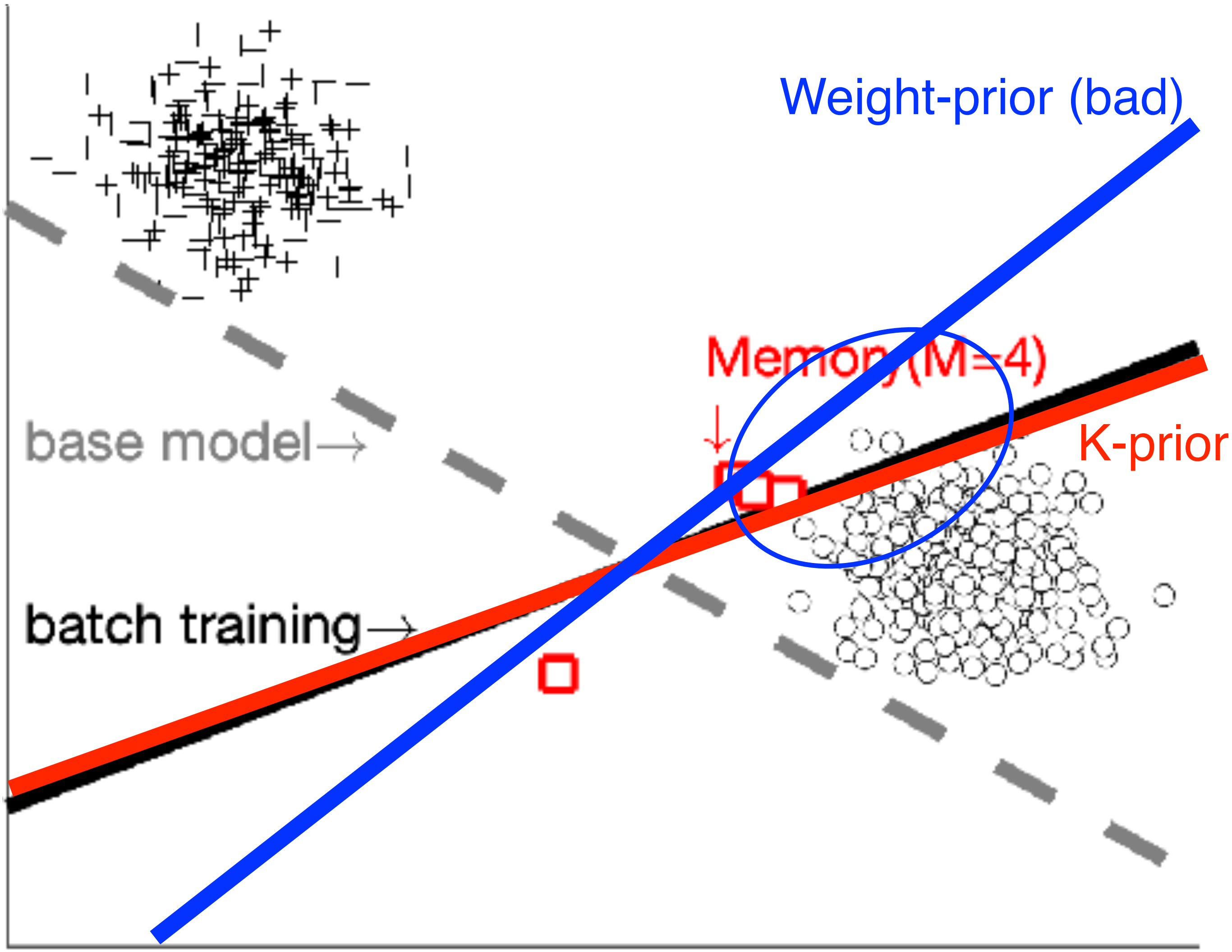
'Add Data' task.

Binary classification with Logistic regression (Zero offset, ie, decision boundary pass through the origin).

Each task N=500, each class 250 examples.

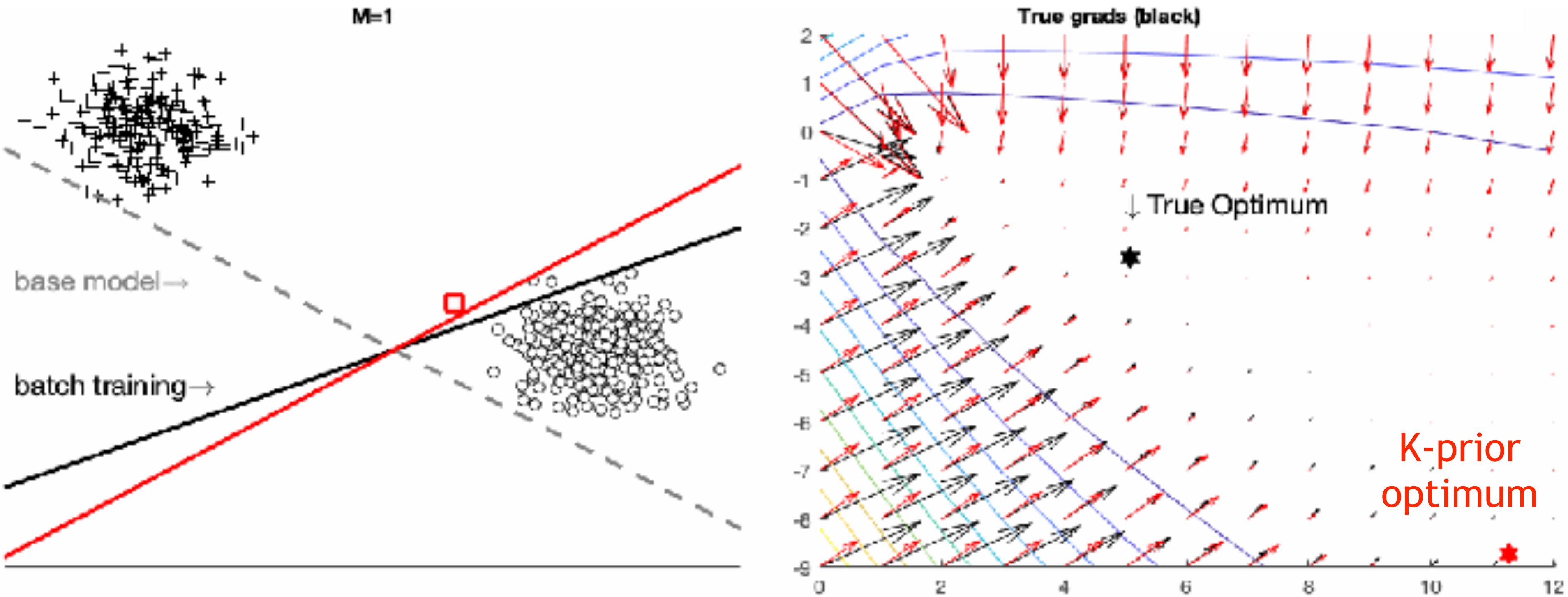
1. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.

Knowledge-Adaptation Priors



K-priors use past-memory $\mathscr{M}$ (size M) in addition to the base model.

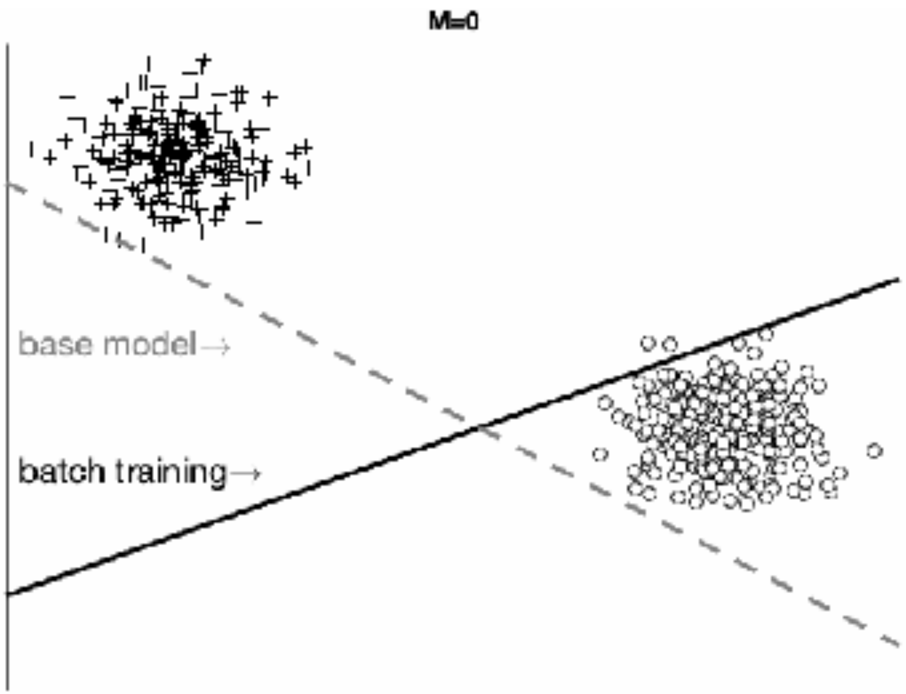
Knowledge-Adaptation Prior (K-priors)

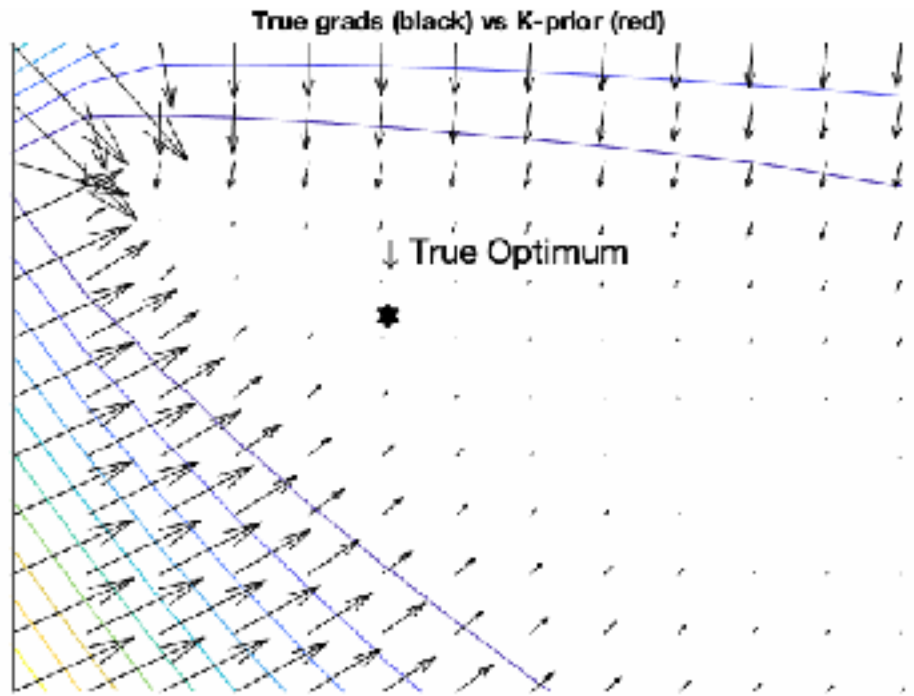


K-priors reconstruct the "gradients of past" by combining weight and functional regularizers

1. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)

Knowledge-Adaptation Prior (K-priors)



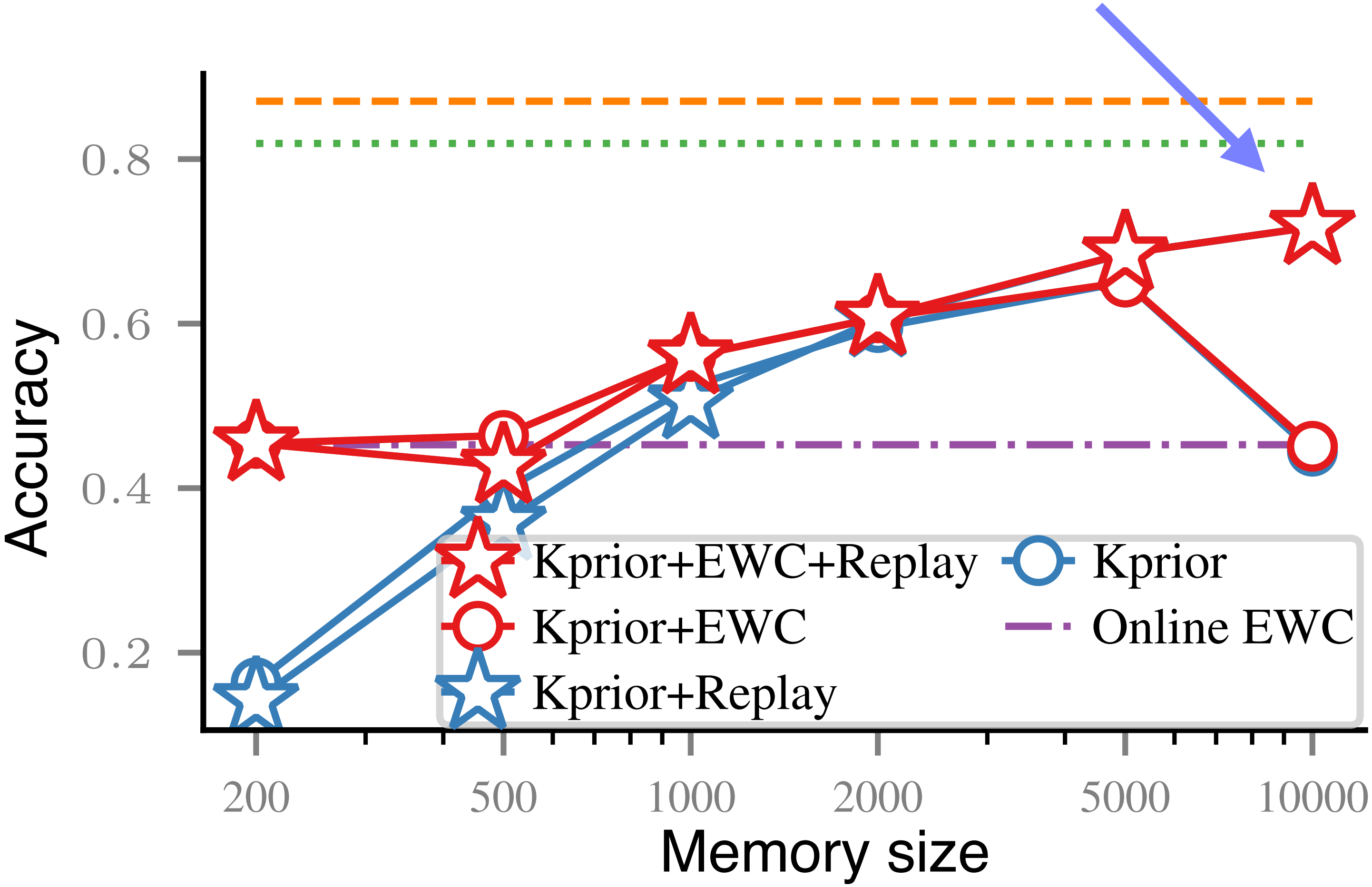


K-priors reconstruct the "gradients of past" by combining weight and functional regularizers

1. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)

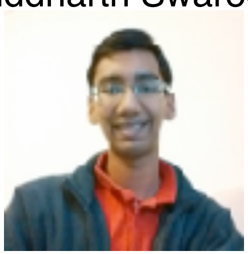
Continual Learning on ImageNet

80% of the batch performance with 10% memory



Erik Daxberger

Siddharth Swaroop

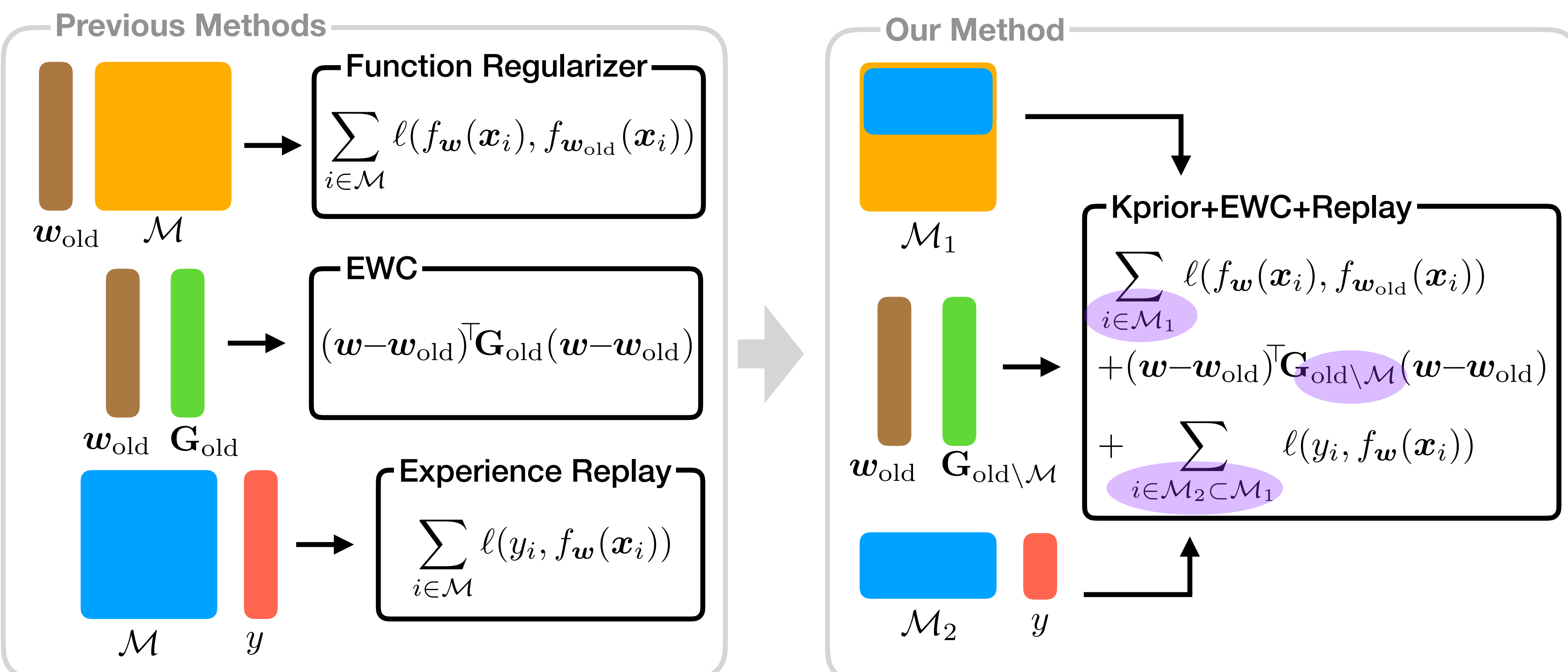


Kazuki Osawa

1. Daxberg, Swaroop, Osawa, Yokota, Turner, Hernandez, Khan, Improving Continual Learning by Accurate Gradient Reconstruction of the Past, under review.

Improving Continual Learning by using the principle of gradient reconstruction

Combine previous approaches to “minimize error in gradient of the past” (use memory)



1. Daxberg, Swaroop, Osawa, Yokota, Turner, Hernandez, Khan, Improving Continual Learning by Accurate Gradient Reconstruction of the Past, under review.

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 around USD 3 million through JST's CREST-ANR and Kakenhi Grants.

Summary of the talk

- New Learning Principles for Adaptive AI
- Unify algorithms with the Bayesian Learning rule (BLR) [1]
  - New work: SAM as Bayes [2]
- BLR's "dual" perspective to "solve" adaptation,
  - Continual learning with memory [4,5,6,7]
  - Bayesian Duality Principle [3, 8]
- "Solve" adaptation
  - When and how can a model quickly adapt?

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
2. Moellenhoff and Khan, SAM as an optimal relaxation of Bayes, https://arxiv.org/abs/2210.01620, 2022
3. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
4. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020
5. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)
6. Daxberger et al., Improving CL by using the Principle of Gradient Reconstructions, Under review, 2022
7. Tailor, Chang, Swaroop, Solin, Khan. Memorable experiences of ML models (in preparation)
8. Khan, Bayesian duality principle (in preparation)

Approximate Bayesian Inference Team

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

Emtiyaz Khan
Team Leader

Pierre Alquier
Research Scientist

Hugo Monzón Maldonado
Postdoc

Happy Buzaaba
Postdoc

Erik Daxberger
Remote Collaborator
University of Cambridge

Paul Chang
Remote Collaborator
Aalto University

Gian Maria Marconi
Postdoc

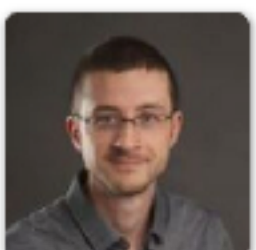
Thomas Möllenhoff
Postdoc

Lu Xu
Postdoc

Jooyeon Kim
Postdoc

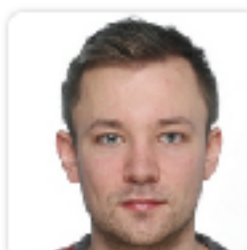
Alexandre Piché
Remote Collaborator
MILA

Ali Unlu
Intern, *Okinawa Institute of Science*



Geoffrey Wolfer
Postdoc

Wu Lin
PhD Student
University of British Columbia



Peter Nickl
Research Assistant

Dharmesh Tailor
Remote Collaborator
University of Amsterdam

Ang Mingliang
Remote Collaborator
National University of Singapore

Kenneth Chen
Intern, *Okinawa Institute of Science and Technology*