Bayesian Principles for Learning Machines

Mohammad Emtiyaz Khan
RIKEN Center for AI Project, Tokyo
http://emtiyaz.github.io

AI that learn like humans

Learn and adapt quickly 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

Bayesian Principles

This talk

Human learning $\neq$ **Deep learning**

Life-long learning from small chunks of data in a non-stationary world

Bulk learning from a large amount of data in a stationary world

My current research focuses on reducing this gap!

1. Parisi, German I., et al. "Continual lifelong learning with neural networks: A review." *Neural Networks* (2019)

2. Geisler, W. S., and Randy L. D. "Bayesian natural selection and the evolution of perceptual systems." *Philosophical Transactions of the Royal Society of London. Biological Sciences* (2002)

Bayesian (Principles for) **Learning Machines**

- Bayes is essential for human-like learning, but infeasible
- Principle I: Bayes Learning Rule (estimation)
- Principle II: Bayes dual (explore-exploit)
- The way forward to human-like learning
- Disclaimer:Focus on the concepts rather than the details

Bayesian (Principles for) Learning Machines

- Bayes is essential for human-like learning, but infeasible
- Principle I: Bayes Learning Rule (estimation)
- Principle II: Bayes dual (explore-exploit)
- The way forward to human-like learning
- Disclaimer:Focus on the concepts rather than the details

Principle of Trial-and-Error

Frequentist: Empirical Risk Minimization (ERM) or Maximum Likelihood Principle, etc.

$$\min_{\theta} \ \ell(\mathcal{D}, \theta) \quad = \sum_{i=1}^{N} [y_i - f_\theta(x_i)]^2 + \gamma \theta^T \theta$$

Loss

Data

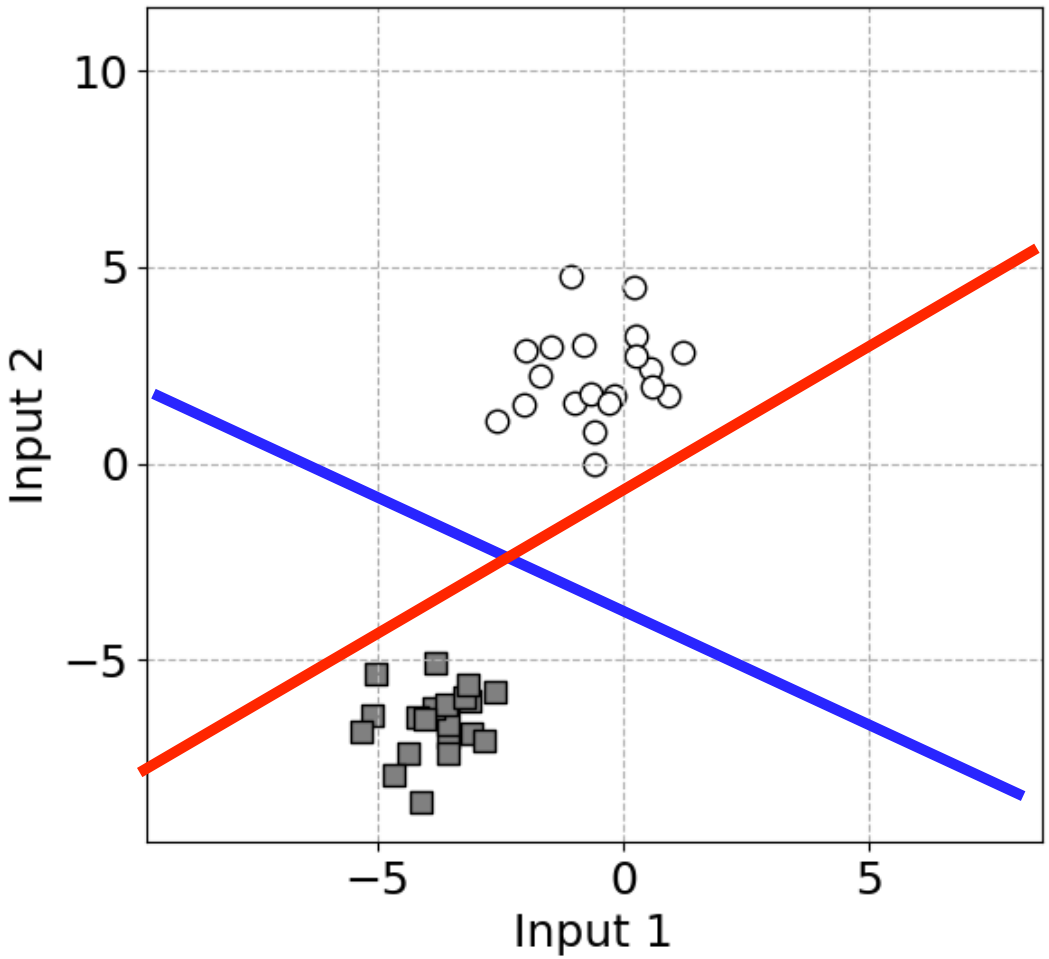
Model Params

Deep Network

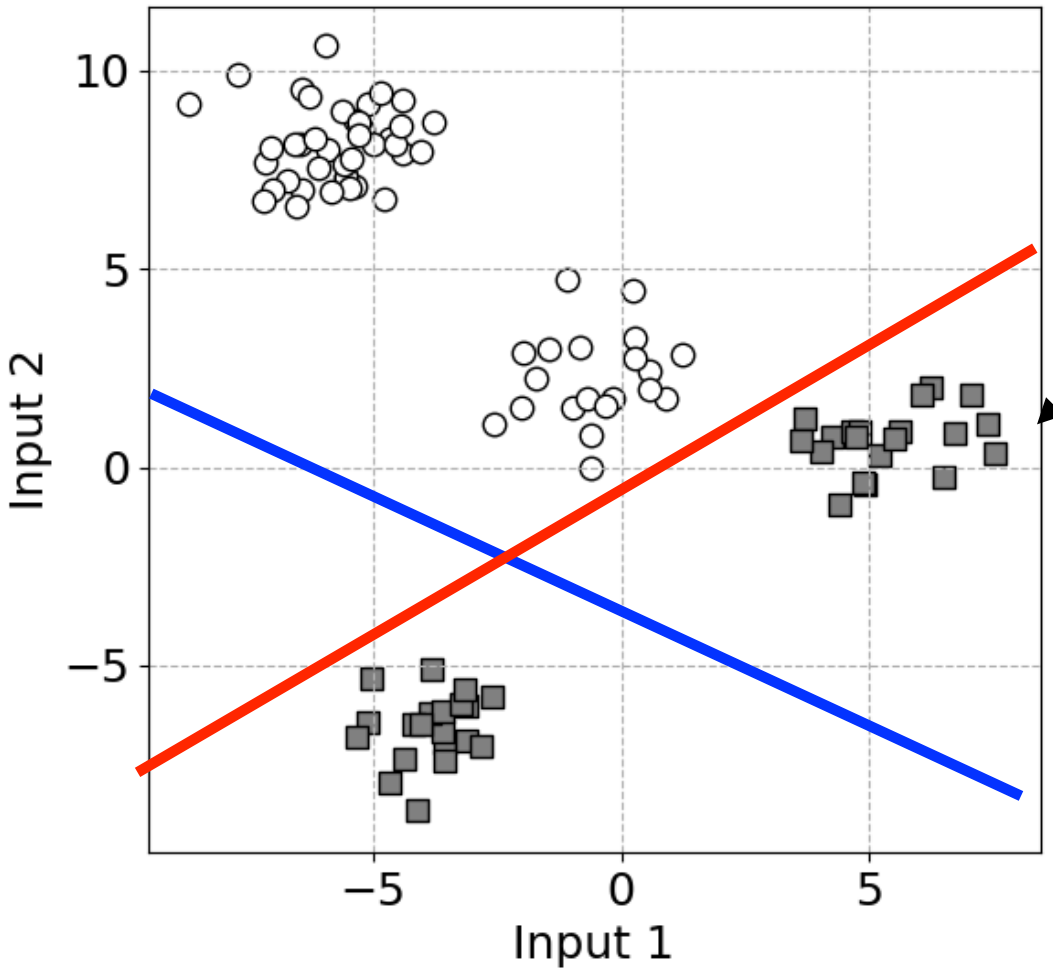
Deep Learning Algorithms: $\theta \leftarrow \theta - \rho H_\theta^{-1} \nabla_\theta \ell(\theta)$

Scales well to large data and complex model, and very good performance in practice.

Which is a good classifier?



Which is a good classifier?



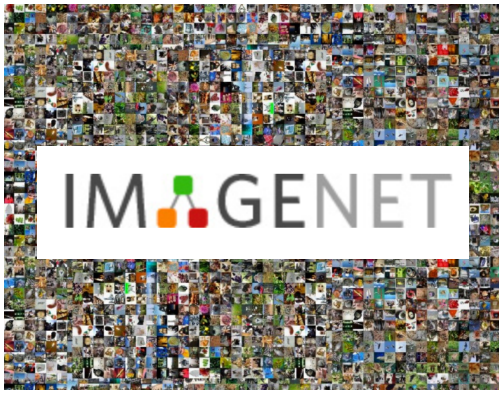
Misclassified by the red line, but not by the blue

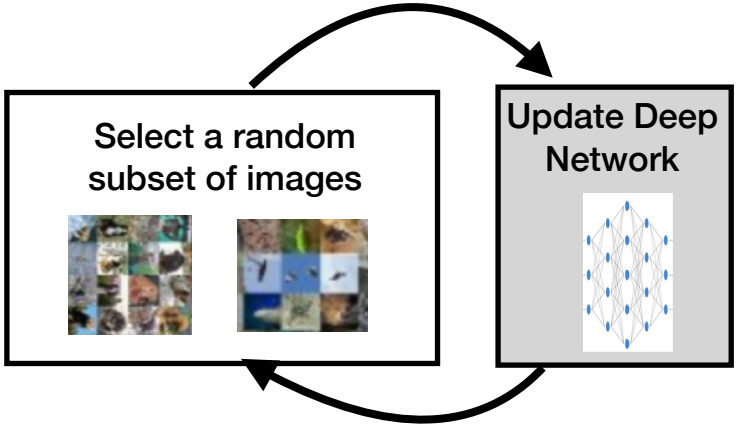
The main challenge of life-long learning:

"The Past can come back to hurt you"

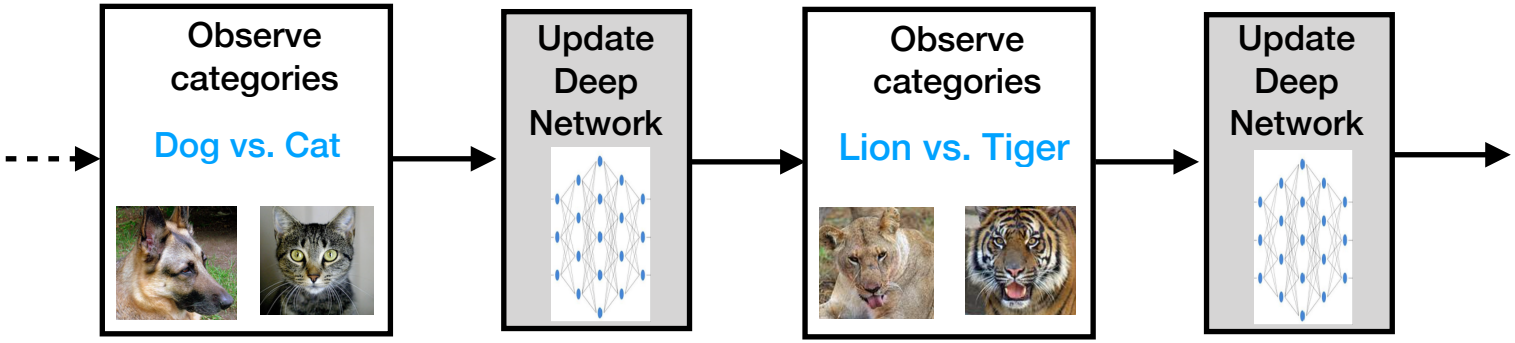
Life-Long Deep Learning?

Standard Deep Learning





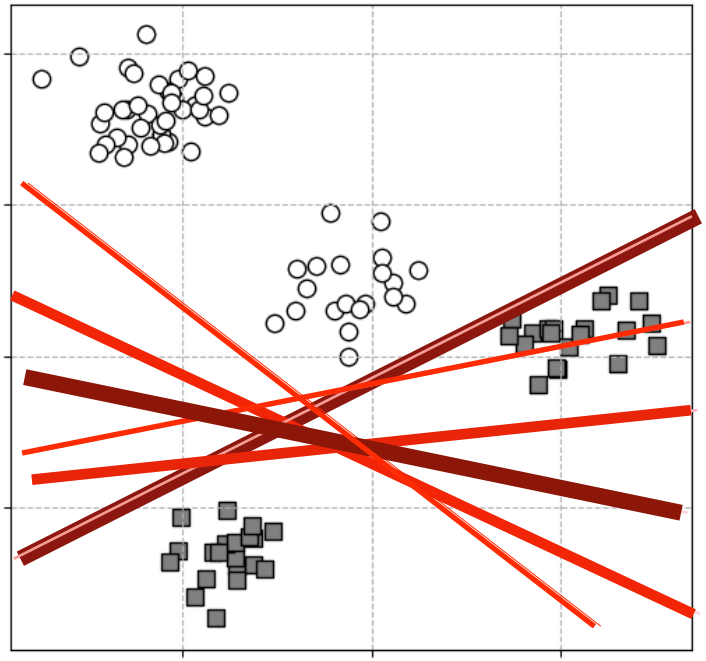
Continual Learning: past classes never revisited



Standard deep learning forgets the past data.

Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.

Bayesian Principles



(1) Keep your options open

$$p(\theta|\mathcal{D}_1) = \frac{p(\mathcal{D}_1|\theta)p(\theta)}{\int p(\mathcal{D}_1|\theta)p(\theta)d\theta}$$

(2) Revise with new evidence

$$p(\theta|\mathcal{D}_2,\mathcal{D}_1) = \frac{p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)}{\int p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)d\theta}$$

Similar ideas in sequential/online decision-making (uncertainty/randomization). Computation is infeasible.

Bayesian (Principles for) Learning Machines

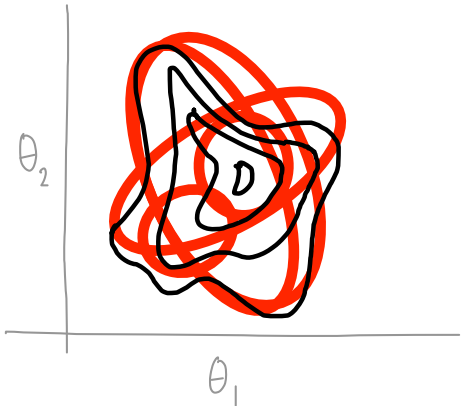
- Bayes is essential for human-like learning, but infeasible
- Principle I: Bayes Learning Rule (estimation)
- Principle II: Bayes dual (explore-exploit)
- The way forward to human-like learning
- Disclaimer:Focus on the concepts rather than the details

Learning-Algorithms from Bayes

Approximate Bayes with the Bayesian learning rule

Posterior Approximation <=> Learning-Algorithm

Complex ⟷ Simple



Bayes' rule

Mixture of Newton

Newton

Gradient Descent

Bayesian Learning Rule

Entropy

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

Exponential-family Approx.

Deep Learning algo: $\theta \leftarrow \theta - \rho H_{\theta}^{-1} \nabla_{\theta} \ell(\theta)$

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

Natural Gradient

Natural and Expectation parameters of an exponential family distribution q

Deep Learning algorithms can be obtained by

1. Choosing a Gaussian approximation,
2. Giving away the “global” property of the rule

1. Khan and Lin. "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models.” AIstats (2017).

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

Learning Algorithm	Posterior Approx.	Algorithmic Approx.	Sec.
Optimization Algorithms			
Gradient Descent	Gaussian (fixed cov.)	Delta approx.	1.4
Newton's method	Gaussian	—"—	1.4
Multimodel optimization (New)	Mixture of Gaussians	—"—	3.2
Deep-Learning Algorithms			
Stochastic Gradient Descent	Gaussian (fixed cov.)	Delta approx., Stochastic approx.	4.1
RMSprop/Adam	Gaussian (diagonal cov.)	Delta approx., Stochastic approx., Hessian approx., Square-root scaling, Slow-moving scale vectors	4.2, 4.3
Dropout	Mixture of Gaussians	Delta approx., Stochastic approx., Responsibility approx.	4.4
STE	Bernoulli	Delta approx., Stochastic approx.	4.6
Online Gauss-Newton (OGN) (New)	Gaussian (diagonal cov.)	Gauss-Newton Hessian approx. in Adam & no square-root scaling	4.5
Variational OGN (New)	—"—	Remove Delta approx. from OGN	4.5
Bayesian Binary NN (New)	—"—	Remove Delta approx. from STE	4.6
Approximate Bayesian Inference Algorithms			
Conjugate Bayes	Exp-family	Set learning rate $\rho_t = 1$	5.1
Laplace's method	Gaussian	Delta approx.	5.2
Expectation-Maximization	Exp-Family + Gaussian	Delta approx. for the parameters	5.3
Stochastic VI (SVI)	Exp-family (mean-field)	Stochastic approx., local rate $\rho_t = 1$	5.4
VMP	—"—	Set learning rate $\rho_t = 1$	5.4
Non-Conjugate VMP	—"—	—"—	5.4
Non-Conjugate VI (New)	Mixture of Exp-family	None	5.5

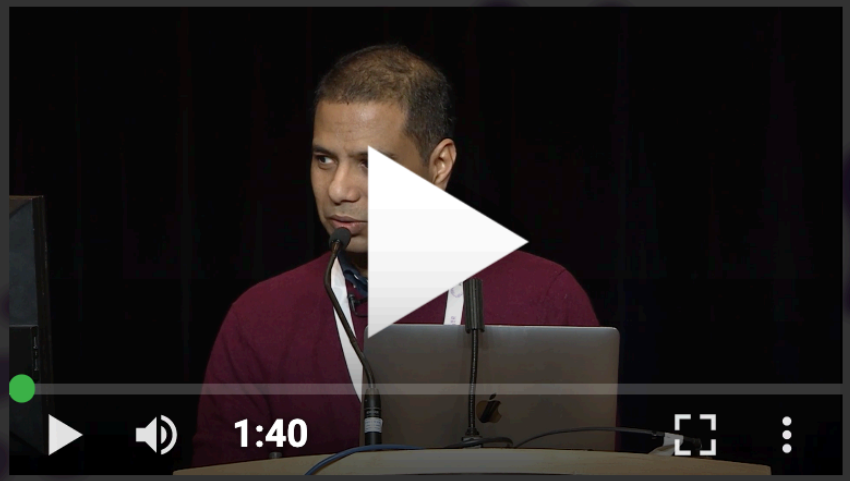
Khan and Rue. "Learning-Algorithms from Bayesian Principles" (2020)

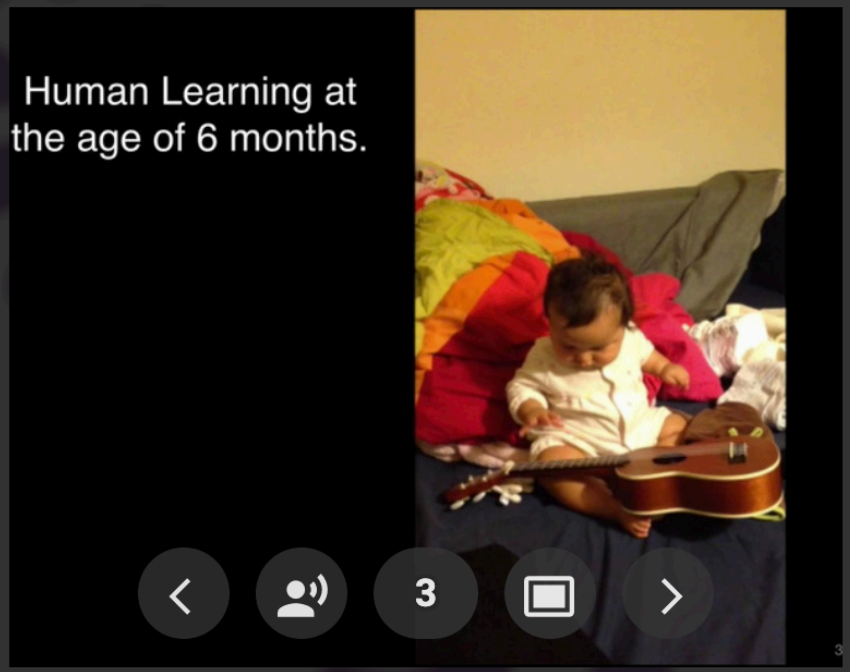
Work in progress (draft available at https://emtiyaz.github.io/papers/learning_from_bayes.pdf)

We can design new algorithms by relaxing the approximations.

For example, to estimate uncertainty via Adam, we can put back the expectation wrt q. This gives us the VOGN algorithm.

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 Bayesian Principles

by **Mohammad Emtiyaz Khan** · Dec 9, 2019 ·

NeurIPS 2019 Tutorial

#NeurIPS 2019 Follow

Views **151 807** Presentations **263** Followers **200**

Latest Popular ···

From System 1 Deep Learning to System 2 Deep Learning

by Yoshua Bengio

17,953 views · Dec 11, 2019

NeurIPS Workshop on Machine Learning for Creativity and Design...

by Aaron Hertzmann, Adam Roberts, ...

9,654 views · Dec 14, 2019

Deep Learning with Bayesian Principles

by Mohammad Emtiyaz Khan

8,084 views · Dec 9, 2019

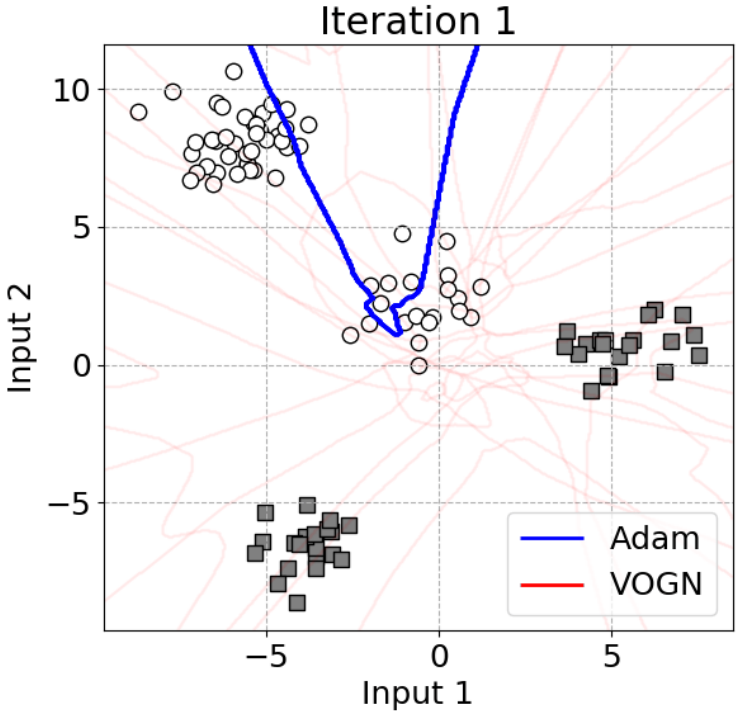
Efficient Processing of Deep Neural Network: from Algorithms to...

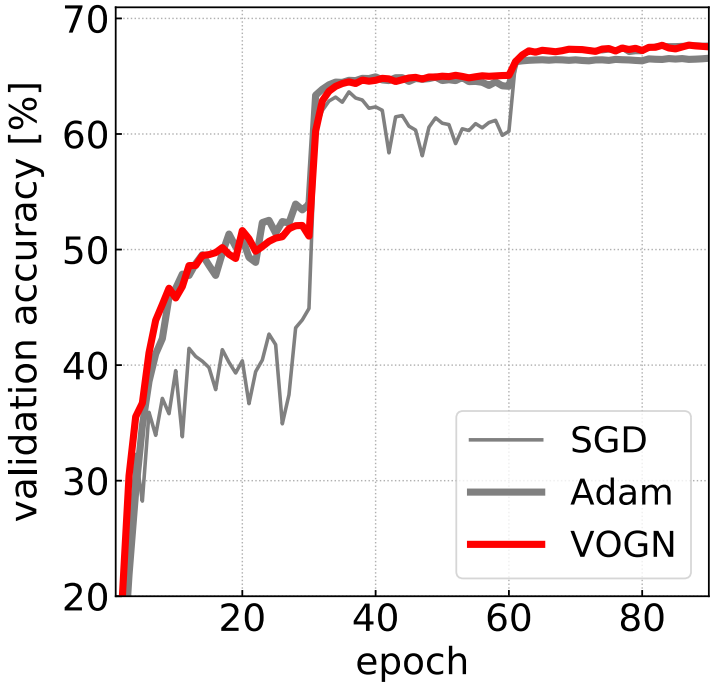
by Vivienne Sze

7,163 views · Dec 9, 2019

Uncertainty Estimation in DL

VOGN: A modification of Adam but match the performance on ImageNet



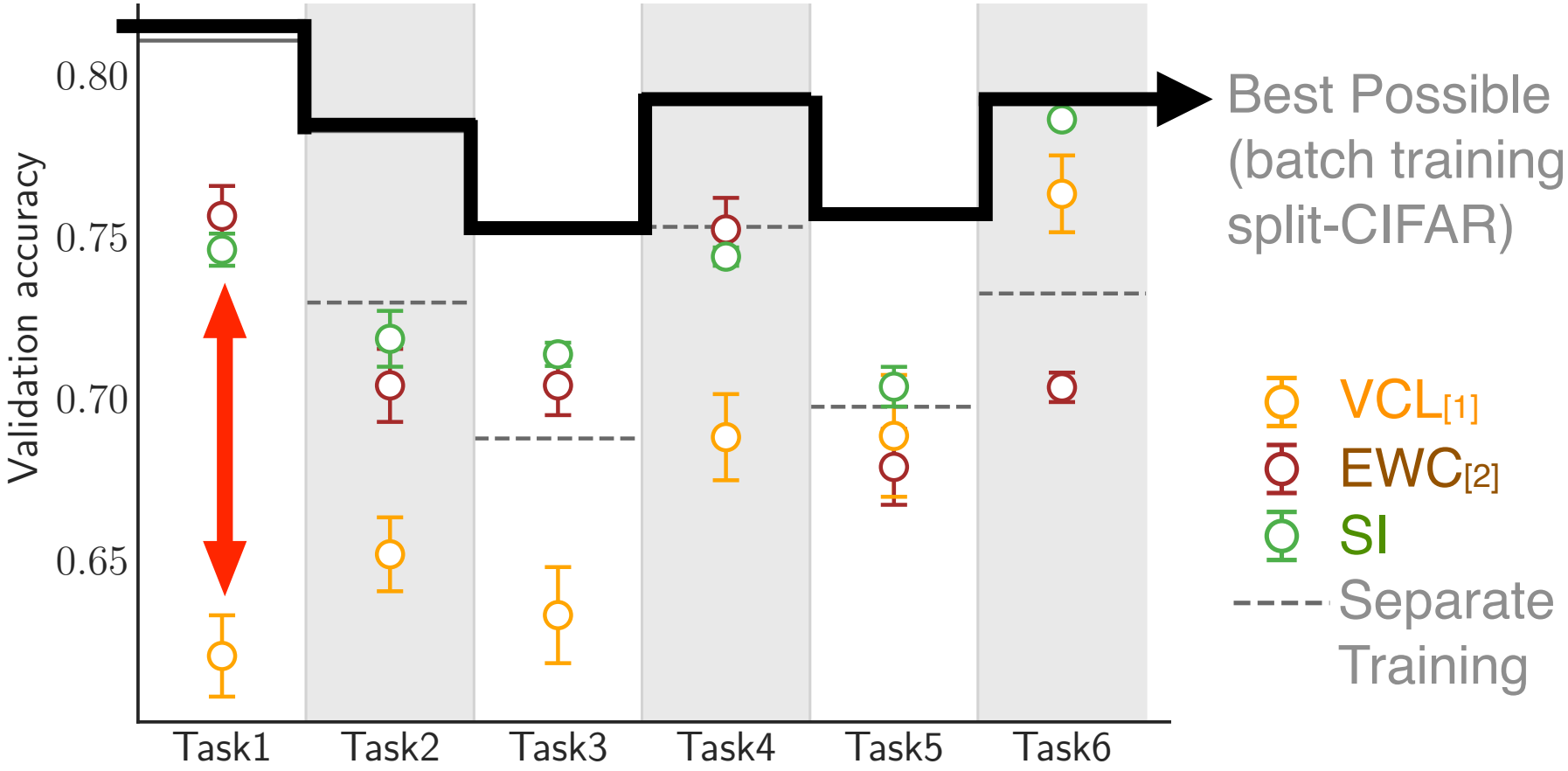


Code available at https://github.com/team-approx-bayes/dl-with-bayes

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).

Continual Learning: Fixing Bayes

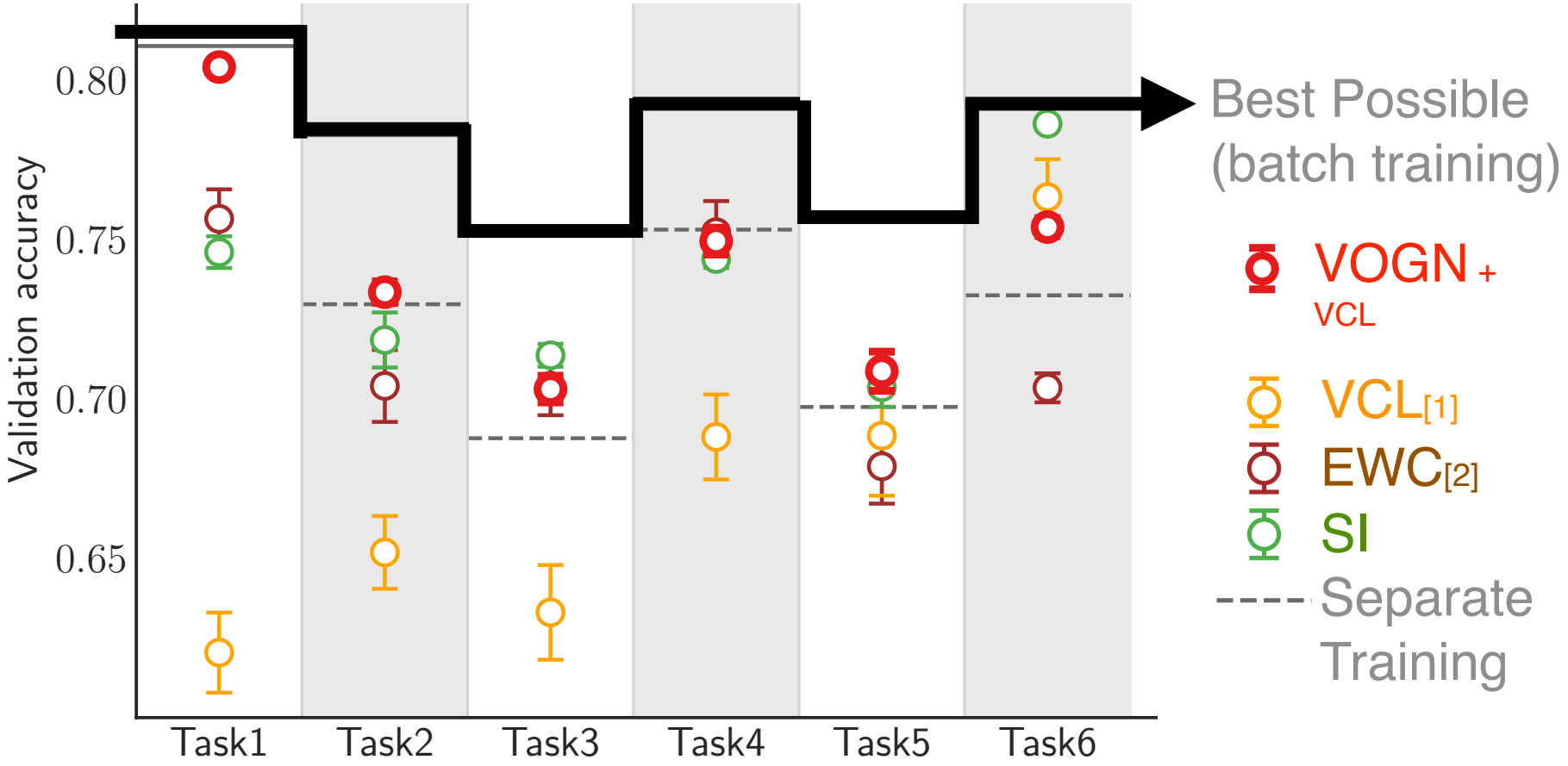
VCL is Bayesian method, trained using SGD, & performs poorly



1. Nguyen et al. “Variational Continual Learning.” ICLR (2018).
2. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS 2017*

Continual Learning: Fixing Bayes

Bayesian Learning Rule (VOGN) fixes the gap



1. Nguyen et al. “Variational Continual Learning.” ICLR (2018).
2. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS 2017*

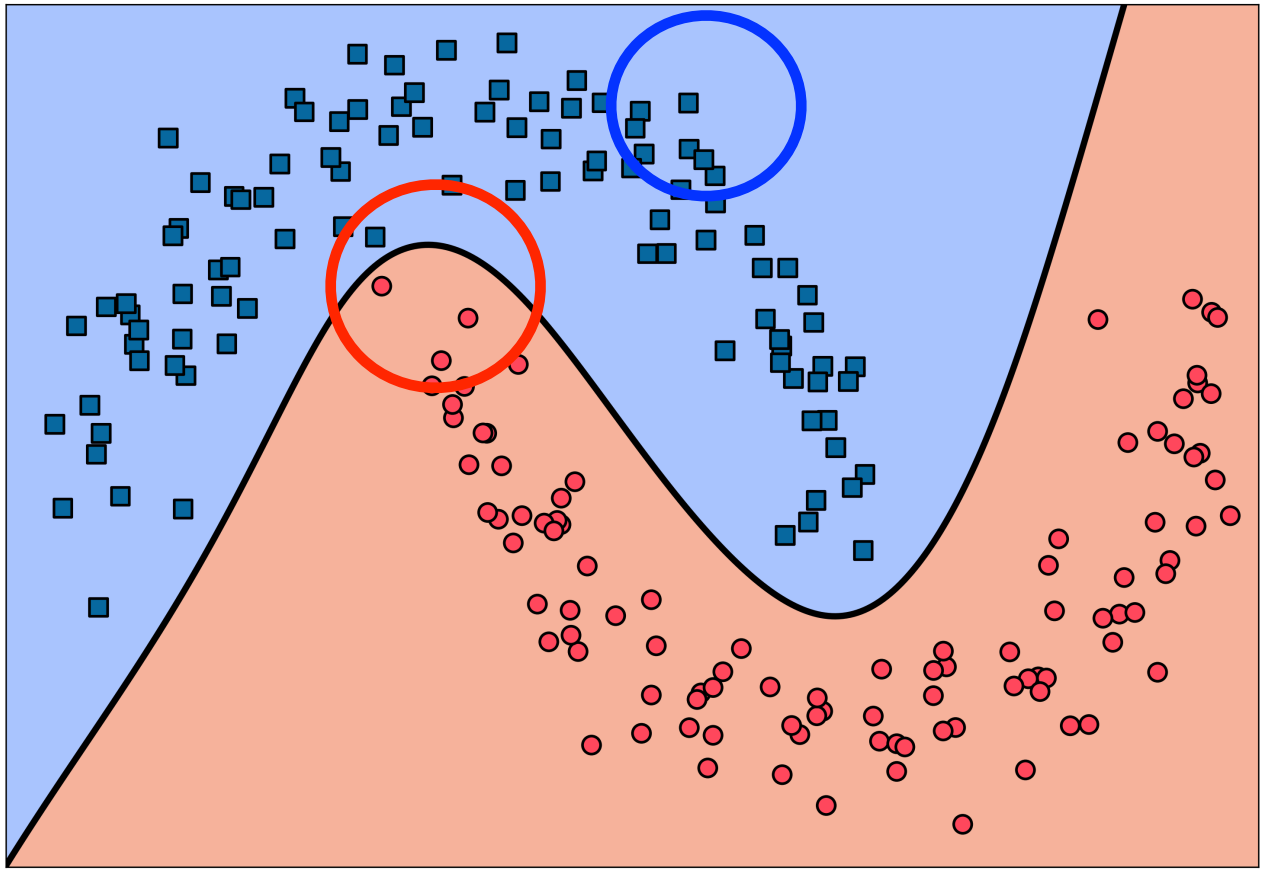
Bayesian (Principles for) Learning Machines

- Bayes is essential for human-like learning, but infeasible
- Principle I: Bayes Learning Rule (estimation)
  - Unify, generalize, and improve learning-algorithms
  - Application: Uncertainty estimation in deep learning
- Principle II: Bayes dual (explore-exploit)
  - Knowledge transfer using a dual representation
  - Application: Continual learning of deep networks
- The way forward to human-like learning

- Disclaimer:Focus on the concepts rather than the details

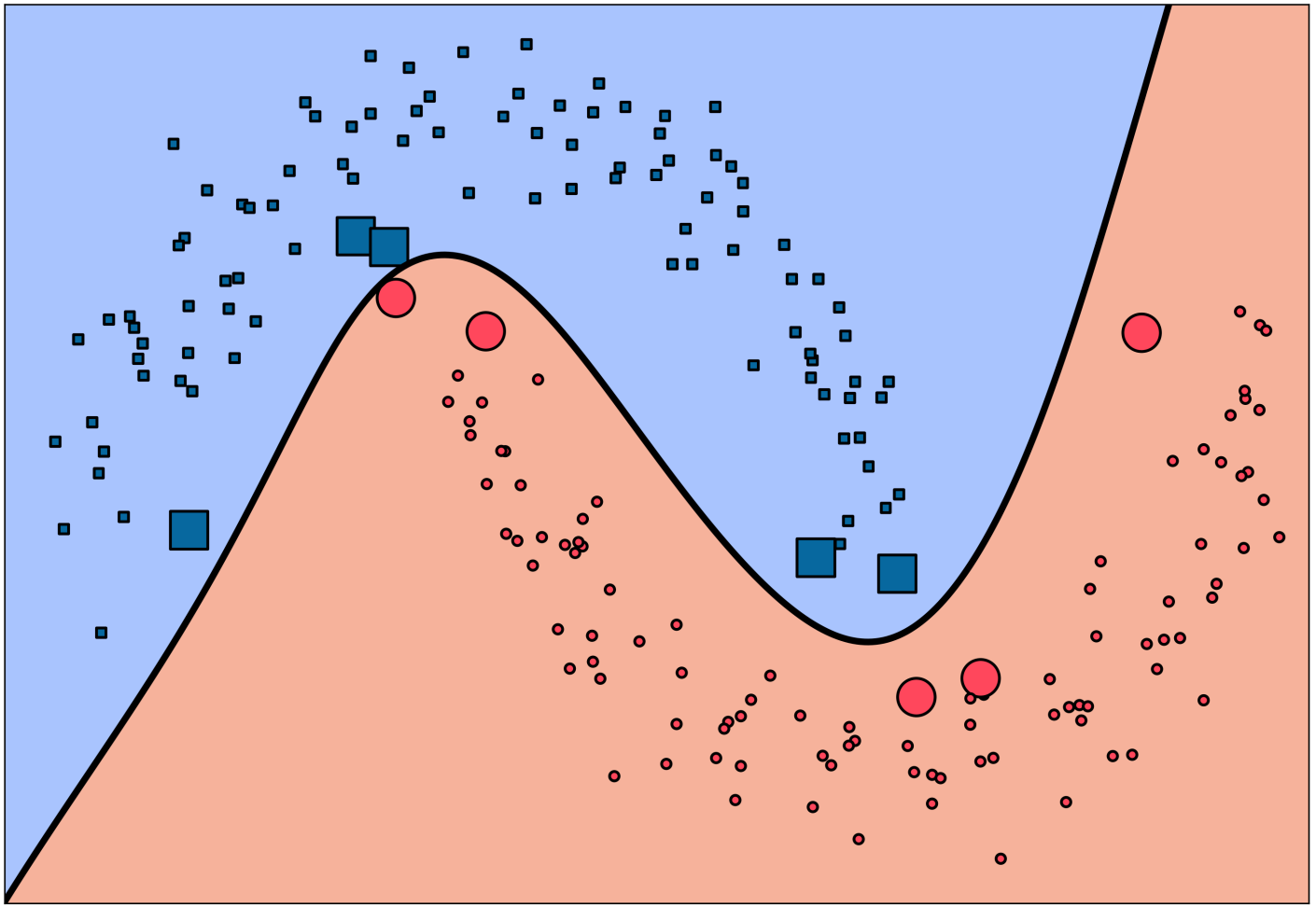
Relevance of Data Examples

Which examples are most relevant for the classifier? Red circle vs Blue circle.



Model view vs Data view

Bayes “automatically” defines data-relevance



Data view (Very much like SVMs)

(By Roman Bachmann)

Bayes Duality

- DNN2GP: Gaussian approx fom Bayes learning rule connect NN to Linear models & Gaussian Process (GPs) [1].

$$\sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) \approx \sum_{i=1}^{N} \frac{1}{\sigma_i^2} [\tilde{y}_i - \phi_i(x_i)^\top \theta]^2$$

neural network

"Dual" variables obtained from $\nabla_\mu \mathbb{E}_q[\ell_i(\theta)]$

(For Gaussian approx, obtained from Jacobian, residual etc.)

- sigma_i^2 defines the "relevance" of the data examples. We call more relevant ones the "memorable examples".
- Natural-gradients give "dual variables" (Bayes Duality)

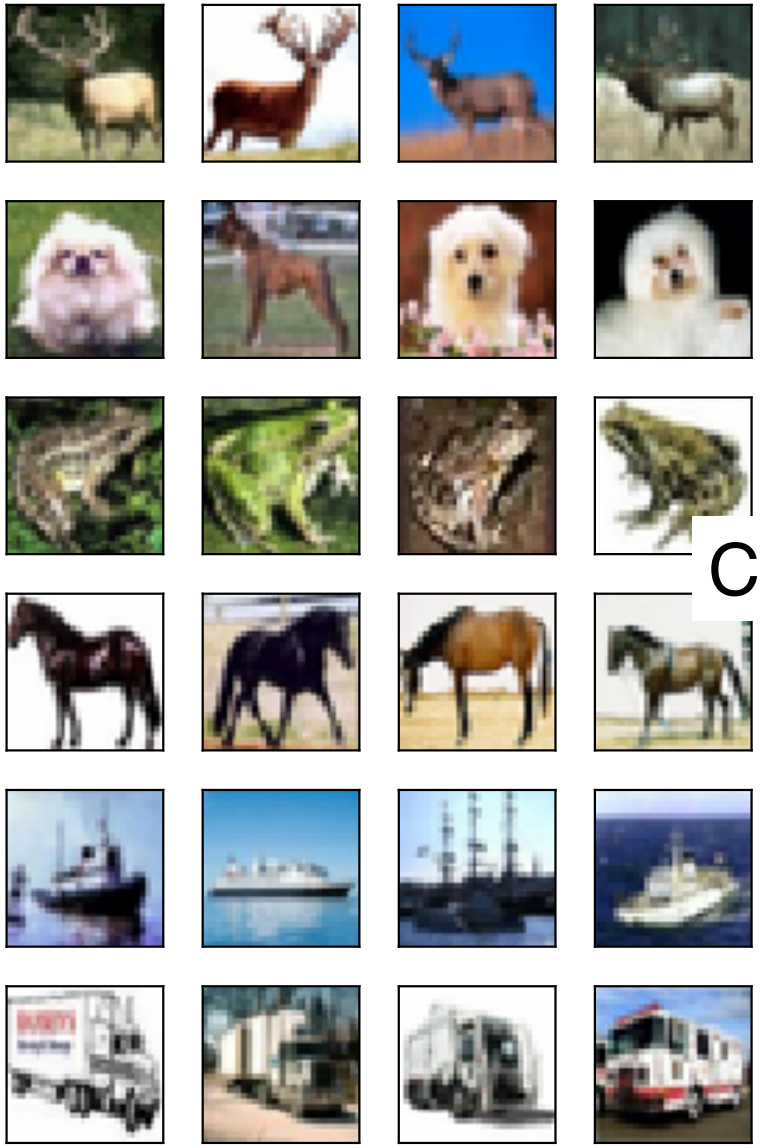
1. Khan et al. "Approximate Inference Turns Deep Networks into Gaussian Processes." *NeurIPS* (2019).

Least Relevant

MNIST

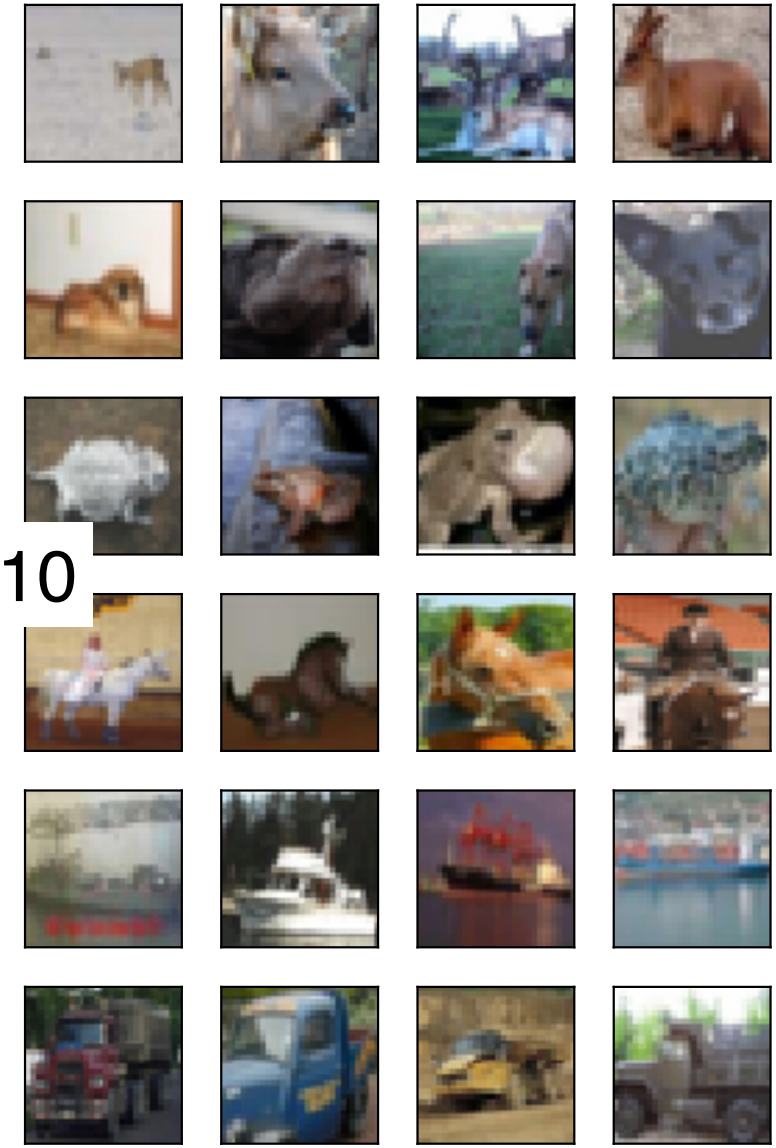
Most Relevant

Least Relevant

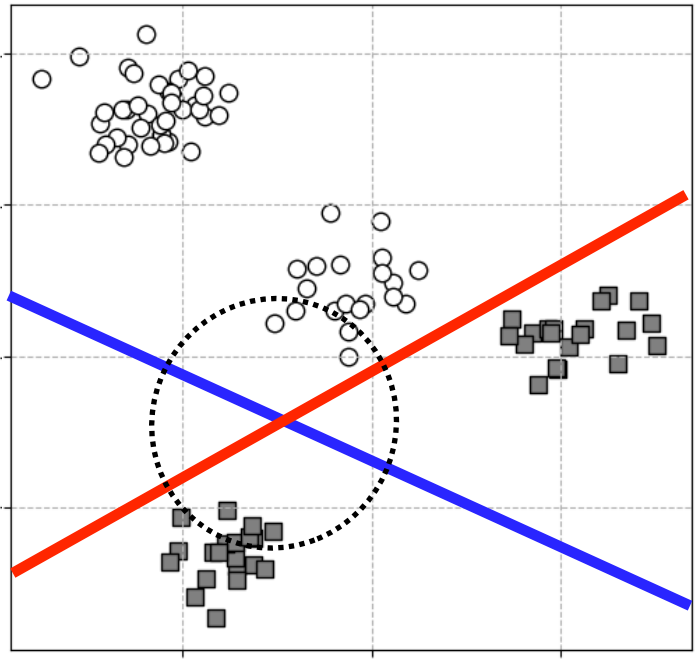


CIFAR-10

Most Relevant



Continual Learning with Bayes Dual



Before: Use past posterior as prior for the next task

$$p(\theta|\mathcal{D}_2,\mathcal{D}_1) = \frac{p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)}{\int p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)d\theta}$$

Gauss Approx (Weight regularization [1])

$$(\theta - \theta_{old})^{\top}\Sigma_{old}^{-1}(\theta - \theta_{old})$$

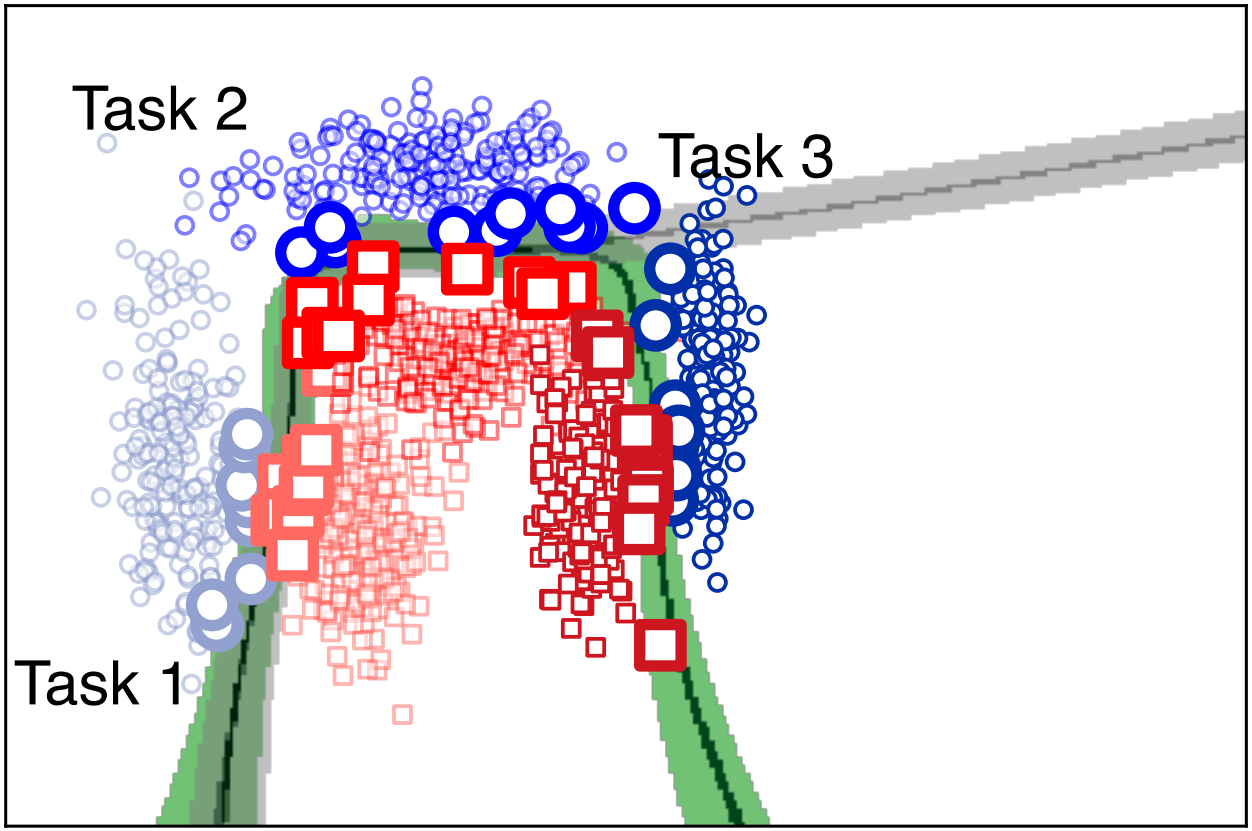
New Idea: Don't let the predictions of memorable examples change (functional regularization with GPs)

$$\left[f(X_m) - f_{old}(X_m)\right]^{\top} K_{old}(X_m, X_m)^{-1}\left[f(X_m) - f_{old}(X_m)\right]$$

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

Functional Regularization of Memorable Past (FROMP)

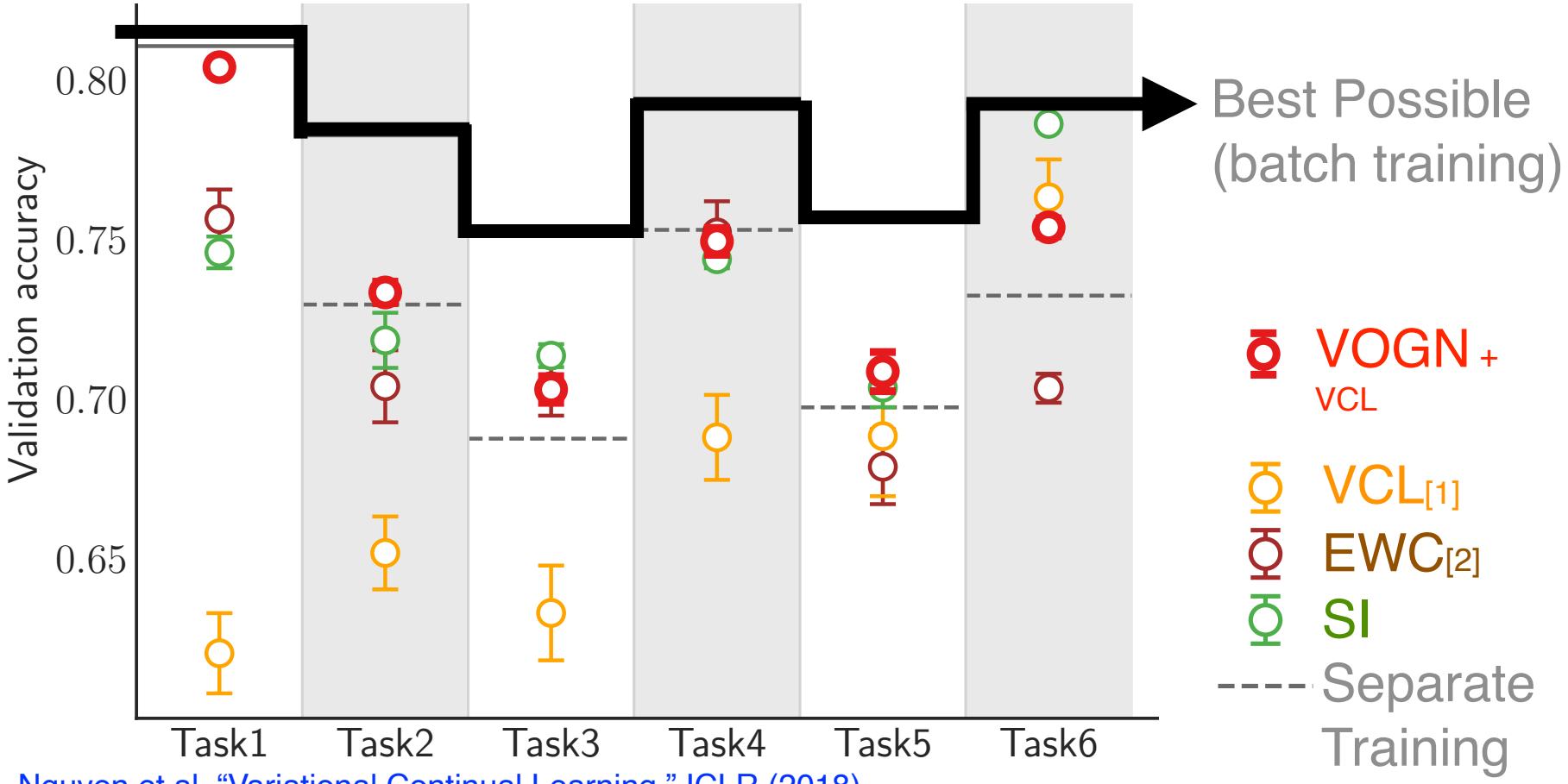
Regularize the function outputs.
Simply adds an additional term in Adam.



1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, Openreview, 2019

Continual Learning: Improving Bayes

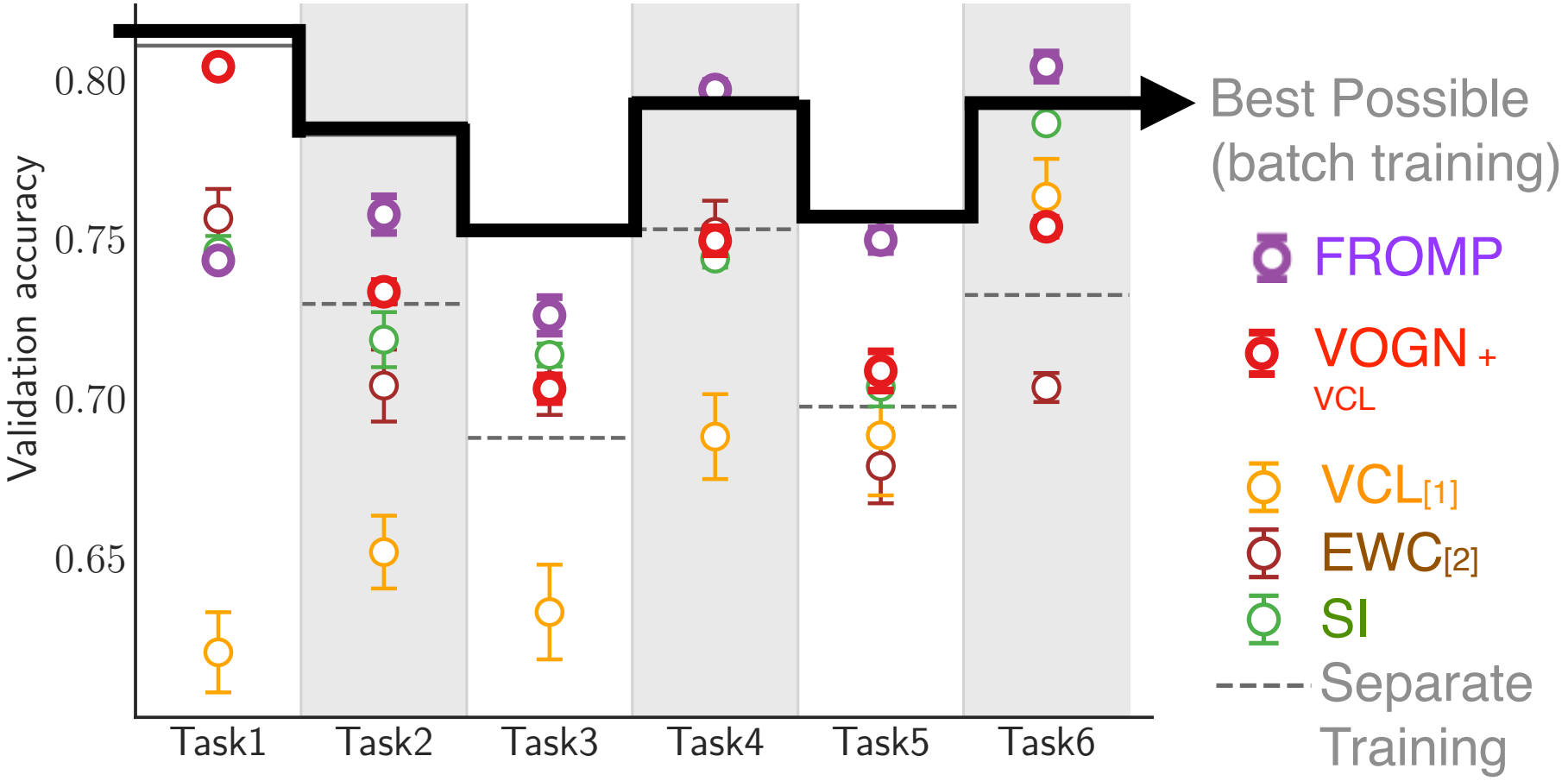
VOGN uses Gaussian posterior as prior in “weight-space”, does not perform significantly better than other methods



1. Nguyen et al. “Variational Continual Learning.” ICLR (2018).
2. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS 2017*

Continual Learning: Improving Bayes

FROMP uses a GP prior in “function-space” over the “memorable pasts” and improves the performance.



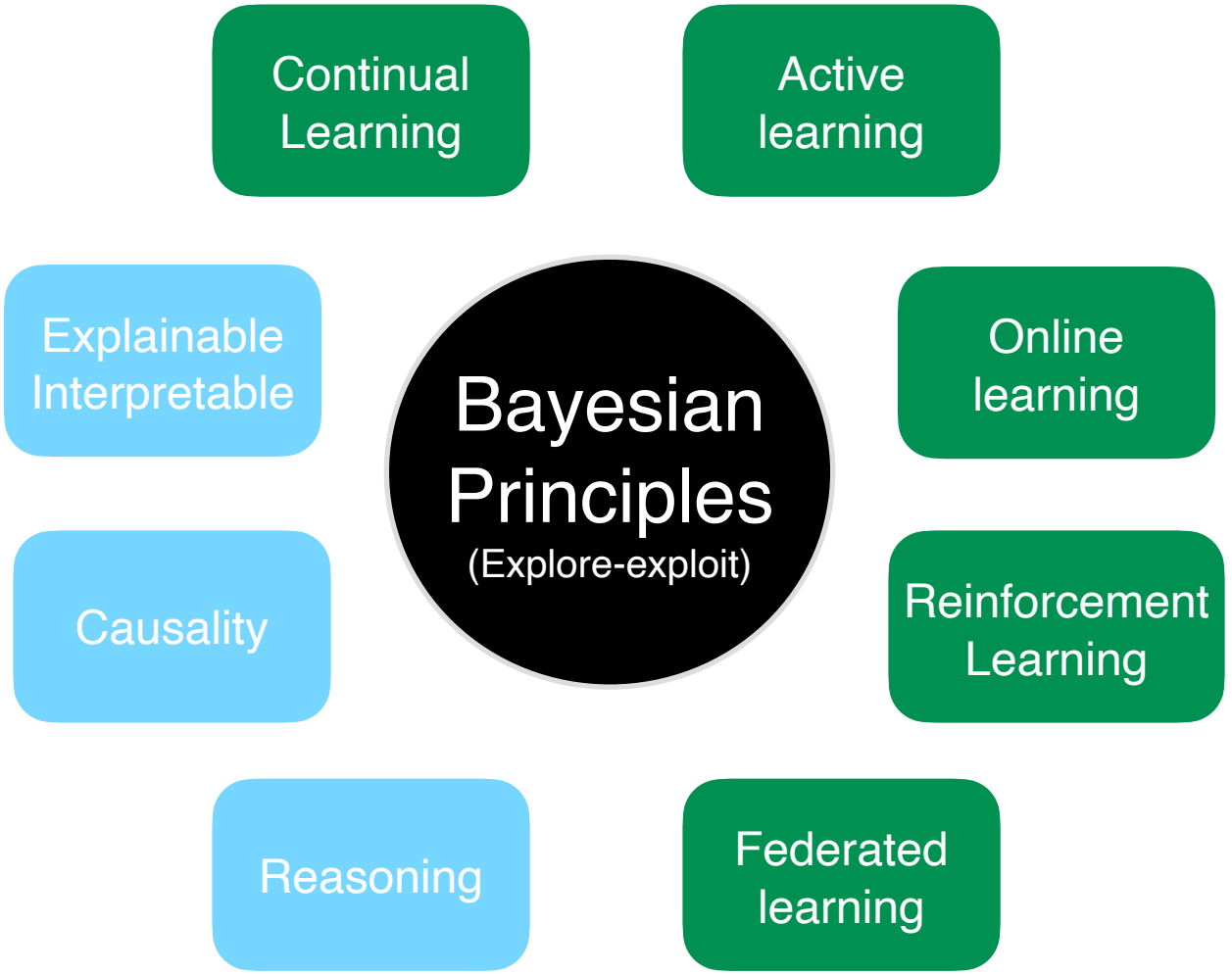
1. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, Openreview, 2019

Bayesian (Principles for) Learning Machines

- Bayes is essential for human-like learning, but infeasible
- Principle I: Bayes Learning Rule (estimation)
- Principle II: Bayes dual (explore-exploit)
- The way forward to human-like learning
- Disclaimer:Focus on the concepts rather than the details

Human Learning at the age of 6 months.

Bayes is indispensable for an AI that learns as efficiently as we do



How to design AI that learn like us?

- Three questions
  - Q1: What do we know? (model)
  - Q2: What do we not know? (uncertainty)
  - Q3: What do we need to know? (action & exploration)
- Posterior approximation is the key
  - (Q1) Models == representation of the world
  - (Q2) Posterior approximations == representation of the model
  - (Q3) The Bayes-dual representation will enable
    - represent learned knowledge,
    - reuse them in novel situations,
    - interact with the environment to collect new knowledge

Learning-Algorithms from Bayesian Principles

Mohammad Emtiyaz Khan
RIKEN center for Advanced Intelligence Project
Tokyo, Japan

Håvard Rue
CEMSE Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Version of November 3, 2020
DRAFT ONLY

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, Adam, and Dropout. The key idea is to estimate posterior approximations using the Bayesian learning rule. Different approximations then result in different algorithms and further algorithmic approximations give rise to variants of those algorithms. Our work shows that Bayesian principles not only unify, generalize, and improve existing learning-algorithms, but also help us design new ones.

Available at

https://emtiyaz.github.io/papers/learning_from_bayes.pdf

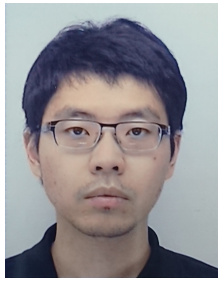
Acknowledgements

Slides, papers, & code are at emtiyaz.github.io

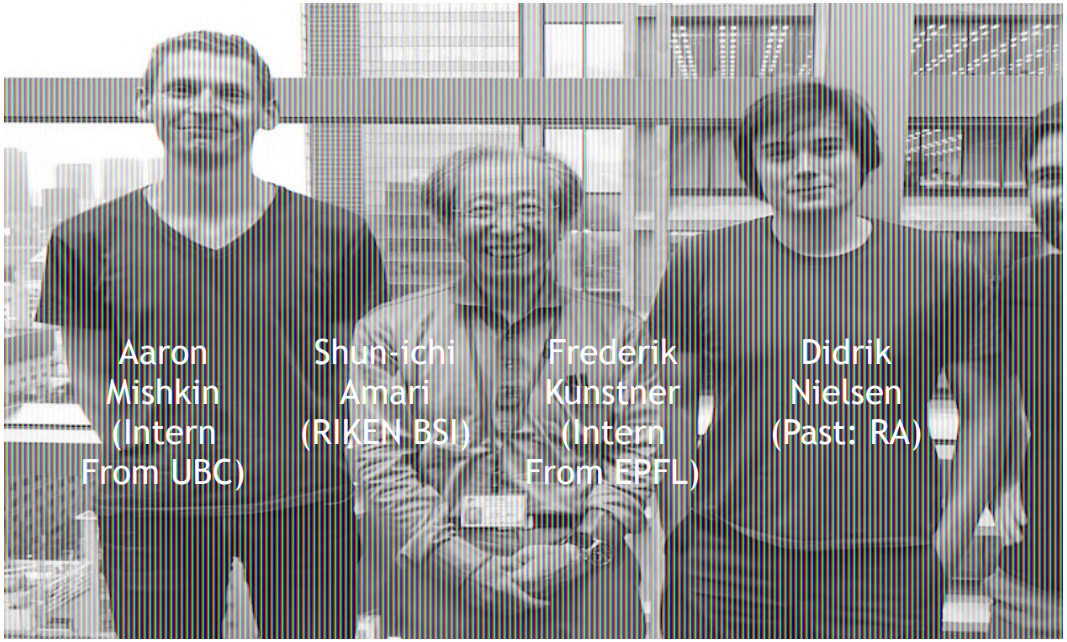
Wu Lin
(Past: RA)

Nicolas Hubacher
(Past: RA)

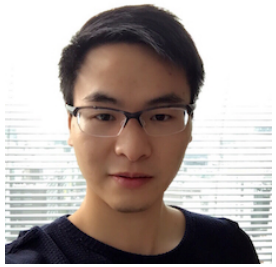
Masashi Sugiyama
(Director RIKEN-AIP)



Voot Tangkaratt
(Postdoc, RIKEN-AIP)



External Collaborators



Zuozhu Liu
(Intern from SUTD)

RAIDEN &
Tsubame

Mark Schmidt
(UBC)

Reza Babanezhad
(UBC)

Yarin Gal
(UOxford)

Akash Srivastava
(UEdinburgh)

Acknowledgements

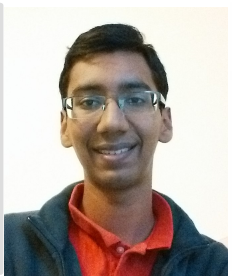
Slides, papers, & code are at emtiyaz.github.io

Kazuki Osawa (Tokyo Tech)

Rio Yokota (Tokyo Tech)

Anirudh Jain (Intern from IIT-ISM, India)

Runa Eschenhagen (Intern from University of Osnabruck)



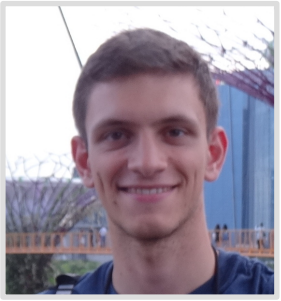
Siddharth Swaroop (University of Cambridge)

Rich Turner (University of Cambridge)

PingBo Pan (Intern from UT Sydney)

Alexander Immer (Intern from EPFL)

Ehsan Abedi (Intern from EPFL)

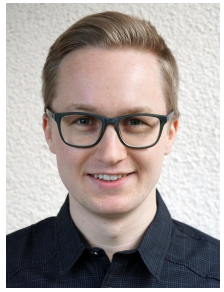


Maciej Korzepa (Intern from DTU)

Pierre Alquier (RIKEN AIP)

Havard Rue (KAUST)

Xiangming Meng Former Post-Doc at RIKEN



Roman Bachmann (Intern from EPFL)

Approximate Bayesian Inference Team

Emtiyaz Khan
Team Leader

Pierre Alquier
Research Scientist

Gian Maria Marconi
Postdoc

Thomas Möllenhoff
Postdoc

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

Open post-doc/RA position:
https://www.riken.jp/en/careers/researchers/20201022_2/index.html

Wu Lin
PhD Student
University of British Columbia

Dharmesh Tailor
Research Assistant

Fariz Ikhwantri
Part-time Student
Tokyo Institute of Technology

Happy Buzaaba
Part-time Student
University of Tsukuba

Evgenii Egorov
Remote Collaborator
Skoltech

Siddharth Swaroop
Remote Collaborator
University of Cambridge

Dimitri Meunier
Remote Collaborator
ENSAE Paris

Peter Nickl
Remote Collaborator
TU Darmstadt

Erik Daxberger
Remote Collaborator
University of Cambridge

Alexandre Piché
Remote Collaborator
MILA