Presentation at the IIT Hyderabad (India), April 25, 2024

The Bayesian Learning Rule

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

Summary of recent research at https://emtiyaz.github.io/papers/symposium_2023.pdf
Slides available at https://emtiyaz.github.io/

Human Learning at the age of 6 months.

Converged at the
age of 12 months

Transfer skills

at the age of 14 months

Current state of ML

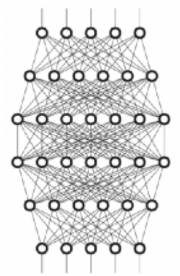
https://www.youtube.com/watch?v=TxobtWAFh8o The video is from 2017

Fixing Machine Learning

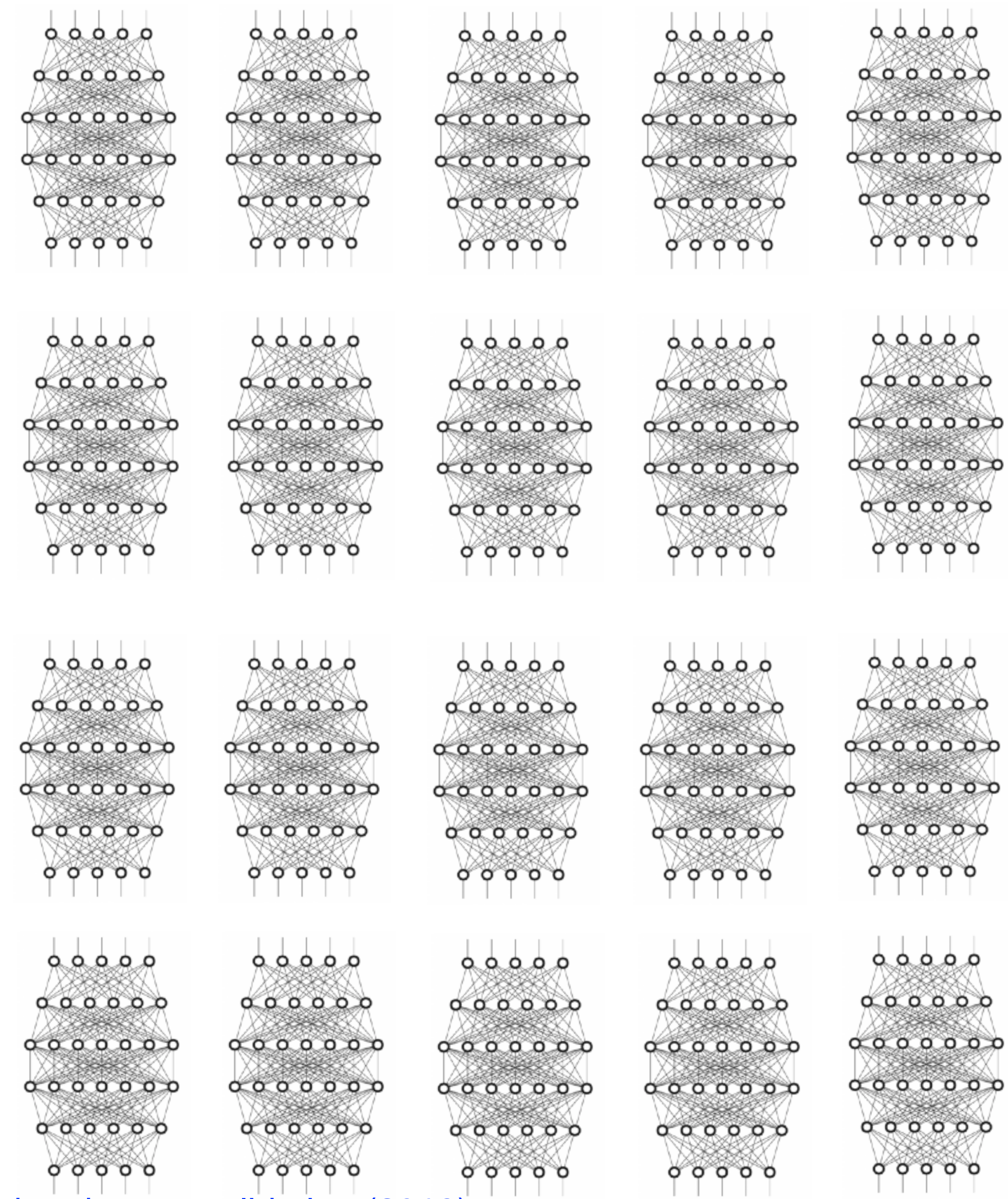
- Even a small change may need full retraining
  - Huge amount of resources only few can afford (costly & unsustainable) [1,2, 3]
  - Difficult to apply in “dynamic” settings (robotics, epidemiology, climate science etc)
- We need sustainable, transparent, trustworthy AI
  - Use reliable building blocks (data, model, metrics)
  - Switch to incremental, continual, lifelong learning
- The Bayesian Learning Rule as a solution to do so!

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

Standard



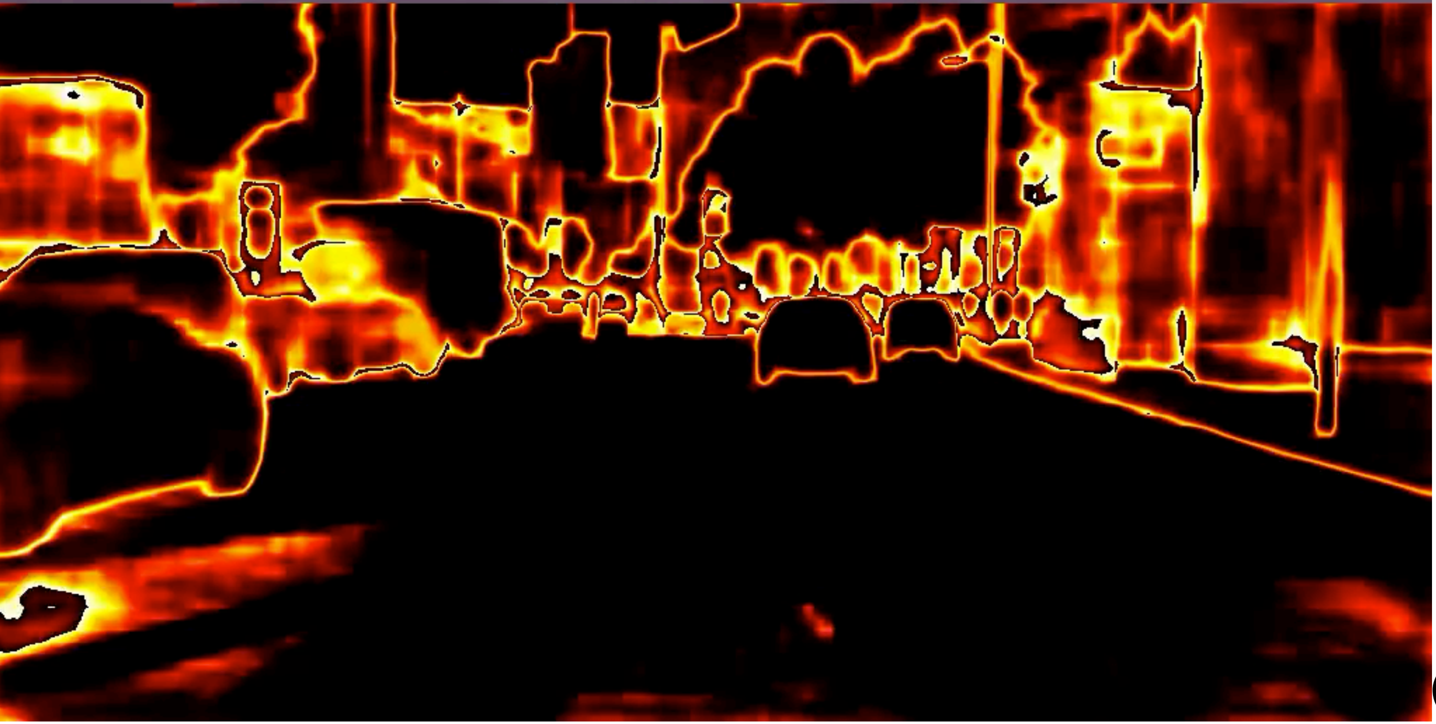
Bayes



$$\log \text{Partition} = \sum_{\text{all } s} \text{Leave-S-Out-CV}$$

1. Foong and Holmes, On the marginal likelihood and cross-validation (2019)

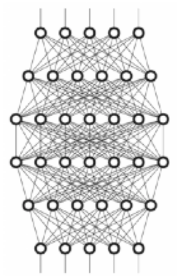
Image Segmentation



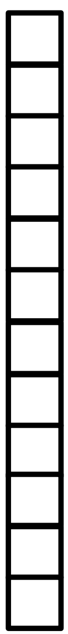
Uncertainty (how much the models differ from each other)

(By Roman Bachmann)

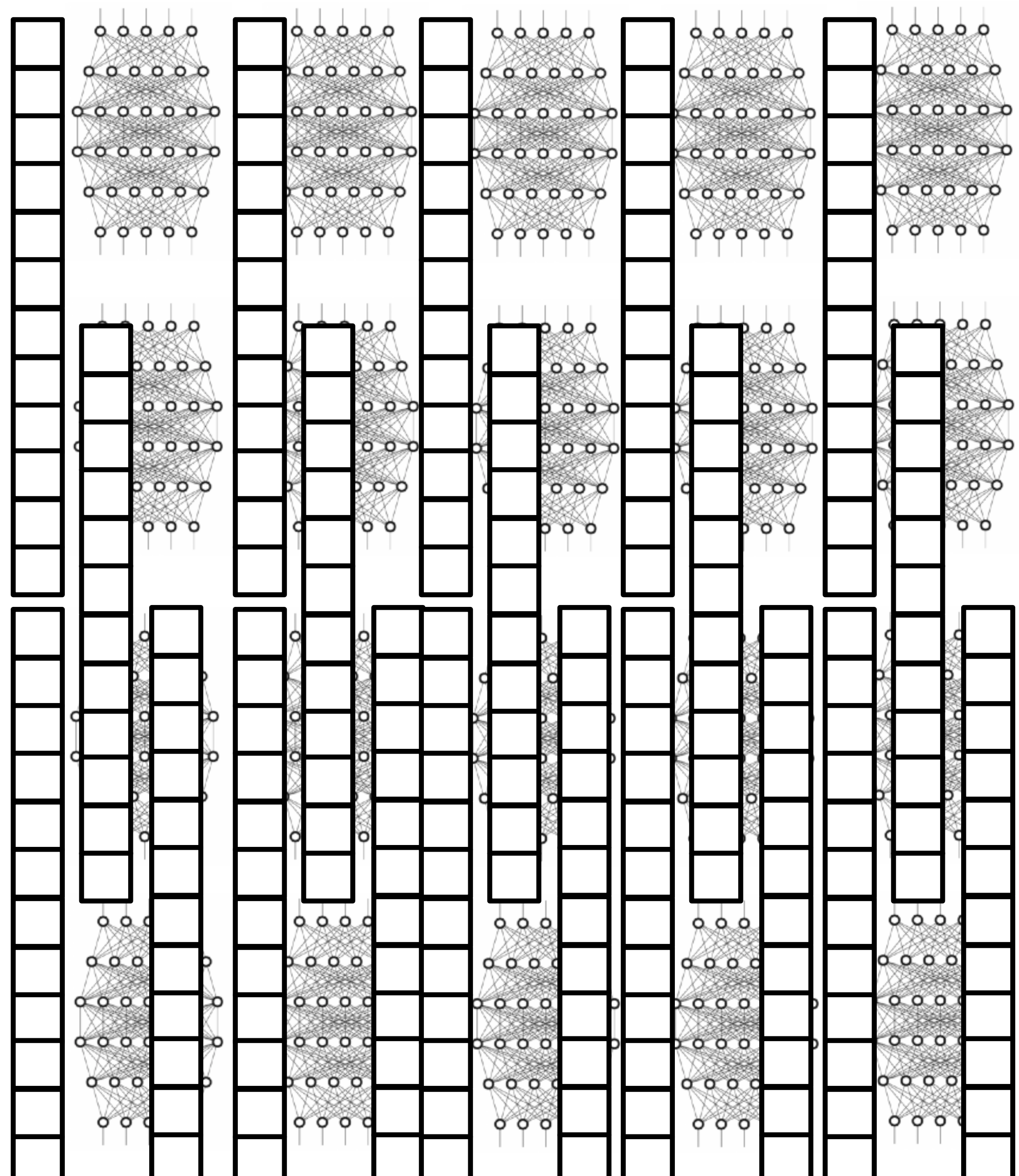
Standard



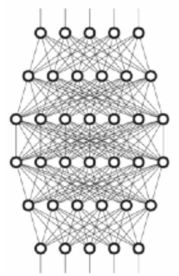
Weight vector



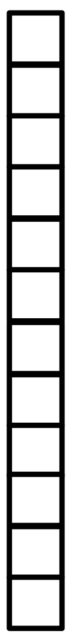
Bayes



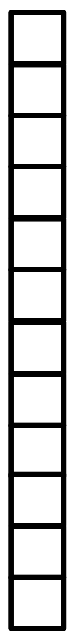
Standard



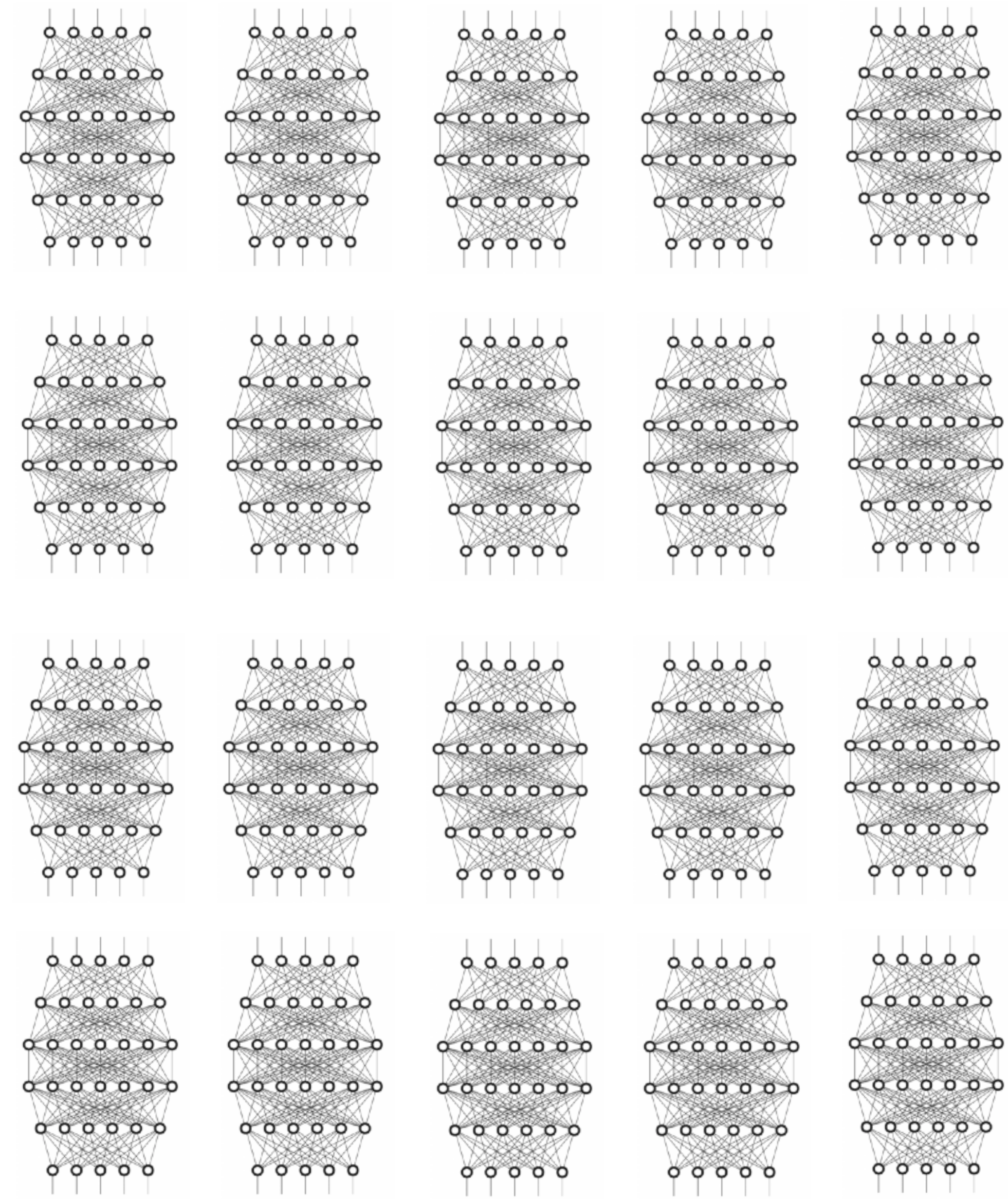
Weight vector



Variance vector



Approximate-Bayes



Learning Algorithms are Bayesian (Learning Rule, BLR)

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).

The Bayesian Learning Rule

$$\min_{\theta} \ \ell(\theta) \qquad \text{vs} \qquad \min_{q \in \mathcal{Q}} \ \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{KL}(q||p_0)$$

Posterior approximation (expo-family)

Bayesian Learning Rule [1,2] (natural-gradient descent)

Natural parameters of q

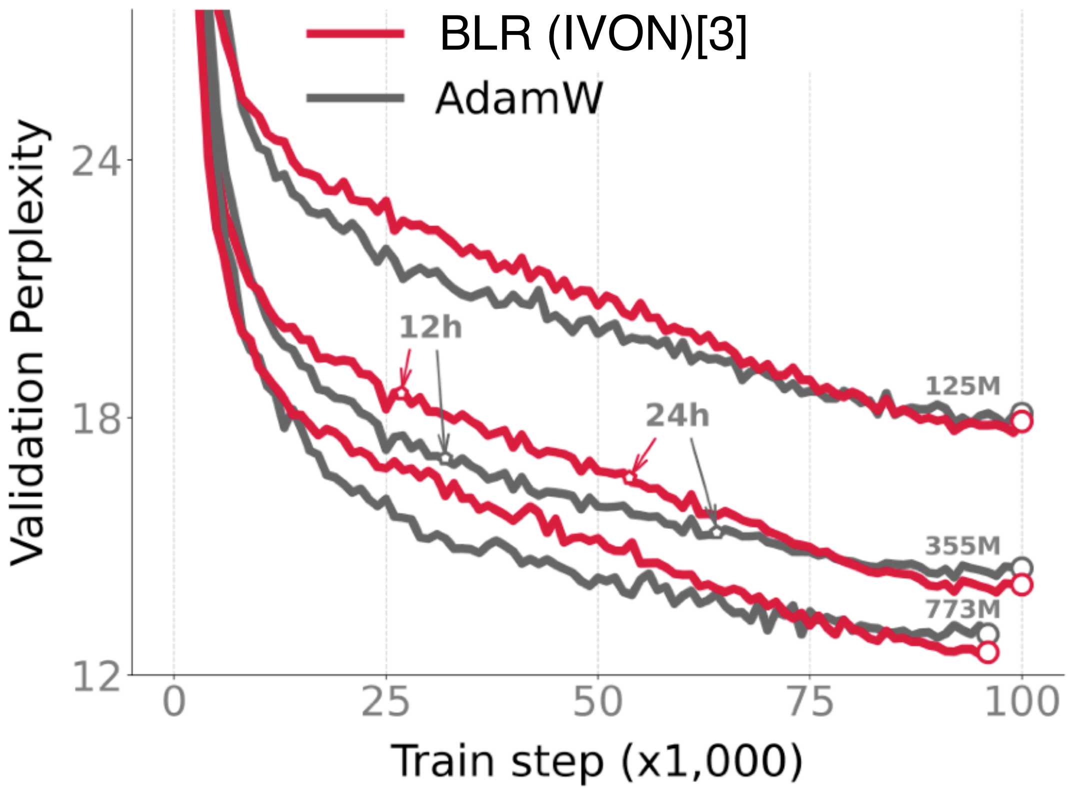
$$\lambda \leftarrow \lambda - \rho F(\lambda)^{-1} \nabla_{\lambda} \Big\{ \mathbb{E}_q[\ell(\theta)] - \mathcal{KL}(q||p_0) \Big\}$$

Posterior approximation q and the Bayesian learning rule open a new way to fix and improve many aspects of deep learning.

1. Khan and Rue, The Bayesian Learning Rule, JMLR, 2023
2. Khan and Lin. "Conjugate-computation variational inference…." AIstats, 2017

Better Performance (on GPT-2)

Better predictions & uncertainty at the same cost [2]



Trained on OpenWebText data (49.2B tokens).

On 773M, we get a gain of 0.5 in perplexity.

On 355M, we get a gain of 0.4 in perplexity.

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Shen et al. “Variational Learning is Effective for Large Deep Networks.” Under review (2024)

Comparison to Adam

RMSprop/Adam

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

BLR [1] variant called IVON [5]
(Improved Variational Online Newton)

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

Only tune initial value of h (a scalar)
Check out the blog: https://team-approx-bayes.github.io/blog/ivon/

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
3. Osawa et al. "Practical Deep Learning with Bayesian Principles." NeurIPS (2019).
4. Lin et al. "Handling the positive-definite constraints in the BLR." ICML (2020).
5. Shen et al. "Variational Learning is Effective for Large Deep Networks." Under review (2024)

Drop-in replacement of Adam

https://github.com/team-approx-bayes/ivon

```
import torch
+import ivon

train_loader = torch.utils.data.DataLoader(train_dataset)
test_loader = torch.utils.data.DataLoader(test_dataset)
model = MLP()

-optimizer = torch.optim.Adam(model.parameters())
+optimizer = ivon.IVON(model.parameters())

for X, y in train_loader:

+    for _ in range(train_samples):
+       with optimizer.sampled_params(train=True)
            optimizer.zero_grad()
            logit = model(X)
            loss = torch.nn.CrossEntropyLoss(logit, y)
            loss.backward()

    optimizer.step()
```

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

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

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

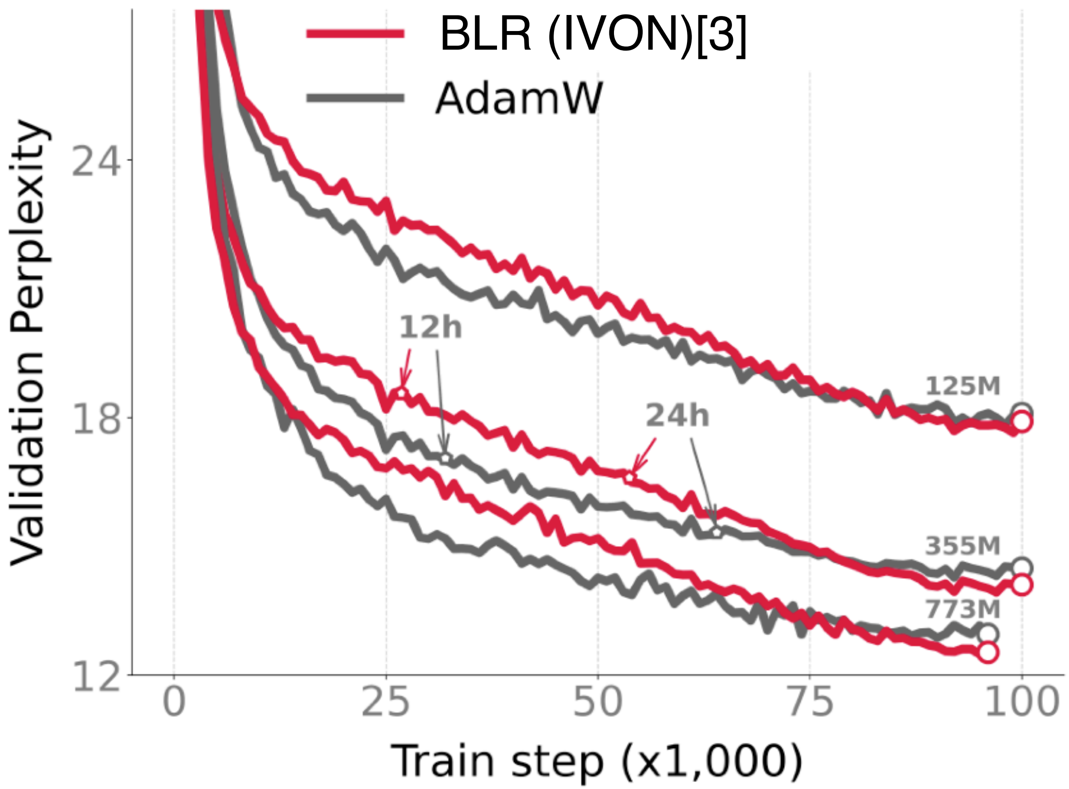
2 Computer Vision Group
Technical University of Munich, Germany

Dec 14th, 2021 — NeurIPS Workshop on Bayesian Deep Learning

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).

GPT-2 with Bayes

Better performance and uncertainty at the same cost



Trained on OpenWebText data (49.2B tokens).

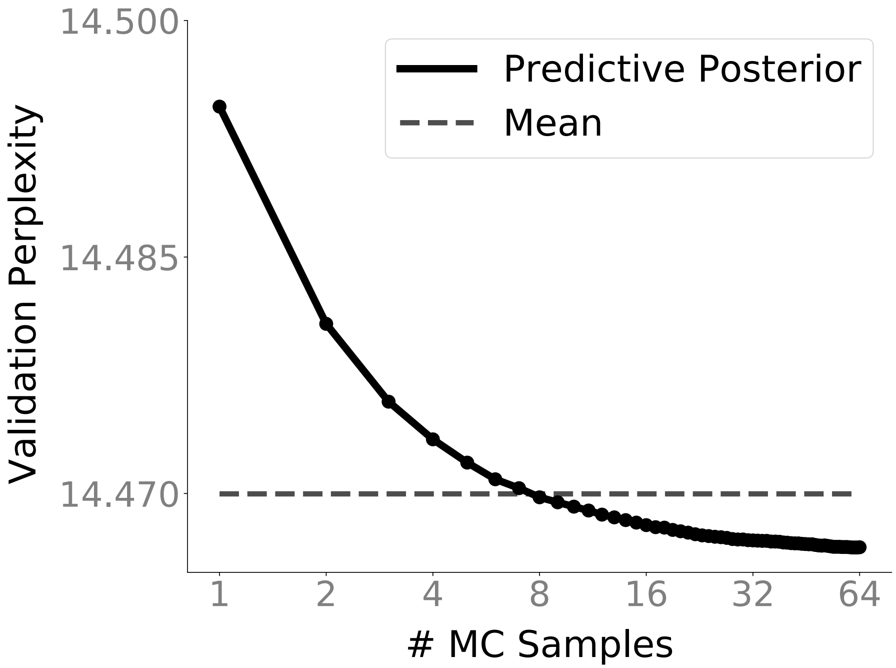
On 773M, we get a gain of 0.5 in perplexity.

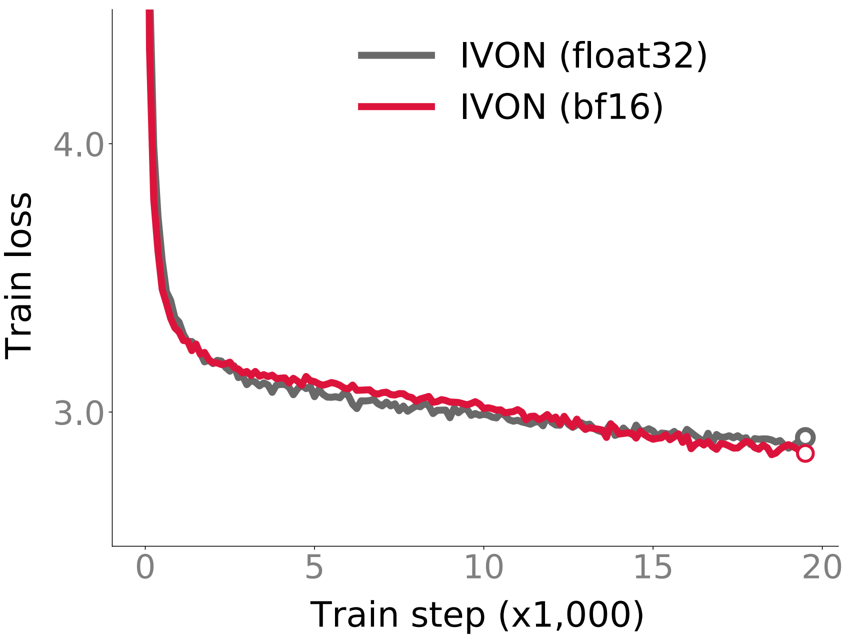
On 355M, we get a gain of 0.4 in perplexity.

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. Shen et al. “Variational Learning is effective for large neural networks.” (Under review)

GPT-2 with Bayes

Posterior averaging improve the result. Can also train on low-precision (a stable optimizer)

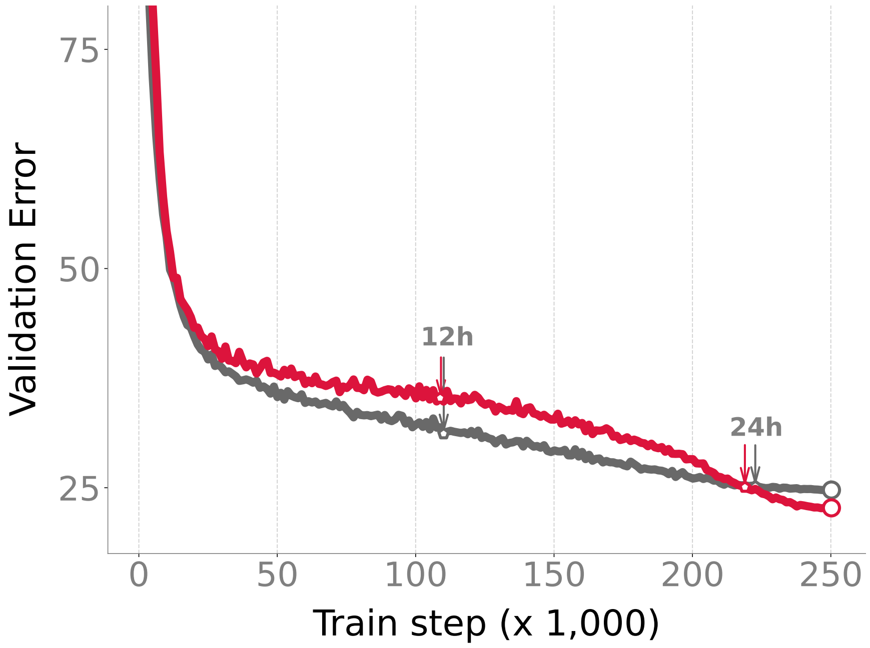


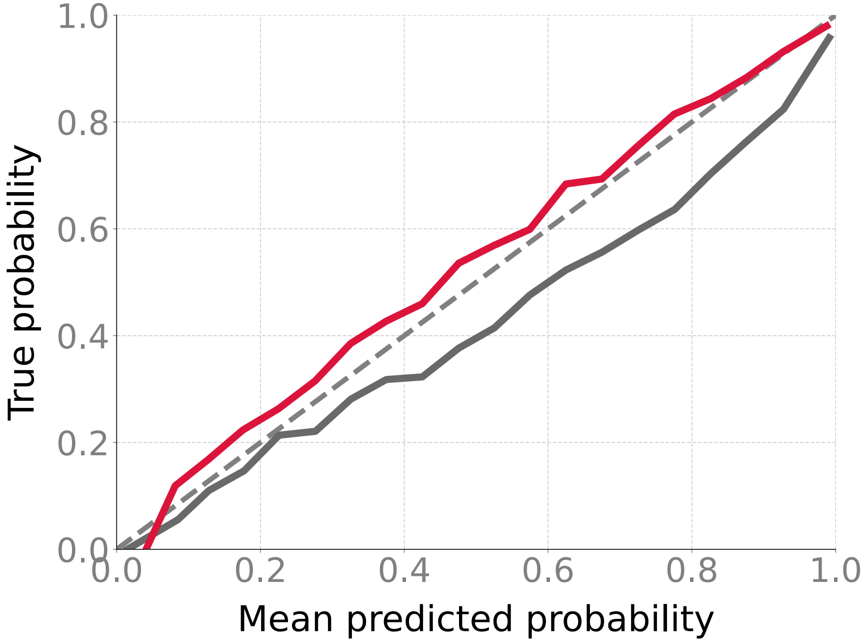


1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. “Practical Deep Learning with Bayesian Principles.” NeurIPS (2019).
3. Shen et al. “Variational Learning is effective for large neural networks.” (Under review)

Better Calibration

2% better accuracy over AdamW and 1% over SGD. Better calibration (ECE of 0.022 vs 0.066)





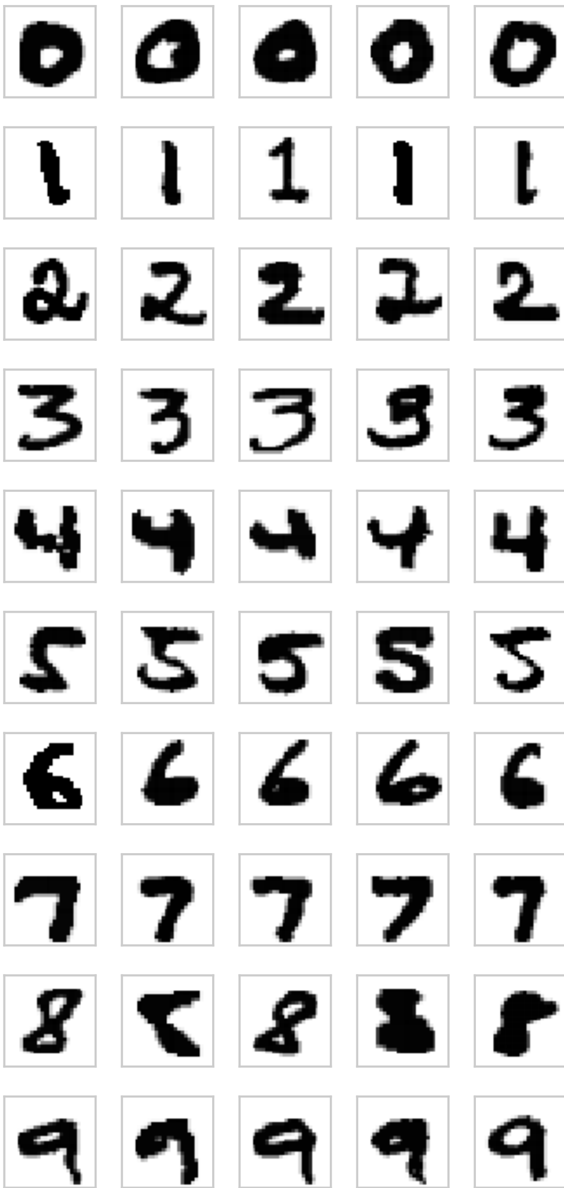
No Severe Overfitting

….like AdamW while improving accuracy over SGD consistently & better uncertainty

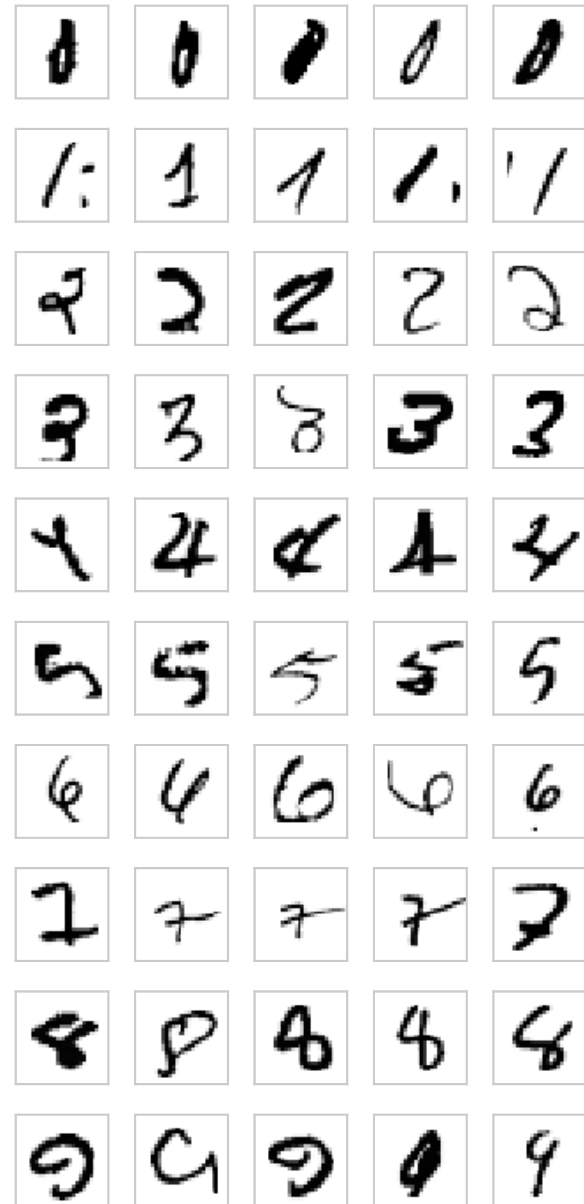
Dataset & Model	Epochs	Method	Top-1 Acc. ↑	Top-5 Acc. ↑	NLL ↓	ECE ↓	Brier ↓
ImageNet-1k **ResNet-50** (25.6M params)	100	AdamW	$74.56_{\pm 0.24}$	$92.05_{\pm 0.17}$	$1.018_{\pm 0.012}$	$0.043_{\pm 0.001}$	$0.352_{\pm 0.003}$
		SGD	$\mathbf{76.18}_{\pm 0.09}$	$\mathbf{92.94}_{\pm 0.05}$	$\mathbf{0.928}_{\pm 0.003}$	$0.019_{\pm 0.001}$	$\mathbf{0.330}_{\pm 0.001}$
		IVON@mean	$\mathbf{76.14}_{\pm 0.11}$	$92.83_{\pm 0.04}$	$0.934_{\pm 0.002}$	$0.025_{\pm 0.001}$	$\mathbf{0.330}_{\pm 0.001}$
		IVON	$\mathbf{76.24}_{\pm 0.09}$	$\mathbf{92.90}_{\pm 0.04}$	$\mathbf{0.925}_{\pm 0.002}$	$\mathbf{0.015}_{\pm 0.001}$	$\mathbf{0.330}_{\pm 0.001}$
	200	AdamW	+2% $75.16_{\pm 0.14}$	$92.37_{\pm 0.03}$	$1.018_{\pm 0.003}$	$0.066_{\pm 0.002}$	$0.349_{\pm 0.002}$
		SGD	+1% $76.63_{\pm 0.45}$	$93.21_{\pm 0.25}$	$0.917_{\pm 0.026}$	$0.038_{\pm 0.009}$	$0.326_{\pm 0.006}$
		IVON@mean	$77.30_{\pm 0.08}$	$93.58_{\pm 0.05}$	$0.884_{\pm 0.002}$	$0.035_{\pm 0.002}$	$\mathbf{0.316}_{\pm 0.001}$
		IVON	$\mathbf{77.46}_{\pm 0.07}$	$\mathbf{93.68}_{\pm 0.04}$	$\mathbf{0.869}_{\pm 0.002}$	$\mathbf{0.022}_{\pm 0.002}$	$\mathbf{0.315}_{\pm 0.001}$
TinyImageNet **ResNet-18** (11M params, wide)	200	AdamW	+15% $47.33_{\pm 0.90}$	$71.54_{\pm 0.95}$	$6.823_{\pm 0.235}$	$0.421_{\pm 0.008}$	$0.913_{\pm 0.018}$
		SGD	+1% $61.39_{\pm 0.18}$	$82.30_{\pm 0.22}$	$1.811_{\pm 0.010}$	$0.138_{\pm 0.002}$	$0.536_{\pm 0.002}$
		IVON@mean	$\mathbf{62.41}_{\pm 0.15}$	$\mathbf{83.77}_{\pm 0.18}$	$1.776_{\pm 0.018}$	$0.150_{\pm 0.005}$	$0.532_{\pm 0.002}$
		IVON	$\mathbf{62.68}_{\pm 0.16}$	$\mathbf{84.12}_{\pm 0.24}$	$\mathbf{1.528}_{\pm 0.010}$	$\mathbf{0.019}_{\pm 0.004}$	$\mathbf{0.491}_{\pm 0.001}$
TinyImageNet **PreResNet-110** (4M params, deep)	200	AdamW	+10% $50.65_{\pm 0.0*}$	$74.94_{\pm 0.0*}$	$4.487_{\pm 0.0*}$	$0.357_{\pm 0.0*}$	$0.812_{\pm 0.0*}$
		AdaHessian	$55.03_{\pm 0.53}$	$78.49_{\pm 0.34}$	$2.971_{\pm 0.064}$	$0.272_{\pm 0.005}$	$0.690_{\pm 0.008}$
		SGD	+2% $59.39_{\pm 0.50}$	$81.34_{\pm 0.30}$	$2.040_{\pm 0.040}$	$0.176_{\pm 0.006}$	$0.577_{\pm 0.007}$
		IVON @mean	$\mathbf{60.85}_{\pm 0.39}$	$\mathbf{83.89}_{\pm 0.14}$	$1.584_{\pm 0.009}$	$0.053_{\pm 0.002}$	$\mathbf{0.514}_{\pm 0.003}$
		IVON	$\mathbf{61.25}_{\pm 0.48}$	$\mathbf{84.13}_{\pm 0.17}$	$\mathbf{1.550}_{\pm 0.009}$	$\mathbf{0.049}_{\pm 0.002}$	$\mathbf{0.511}_{\pm 0.003}$
CIFAR-100 **ResNet-18** (11M params, wide)	200	AdamW	+11% $64.12_{\pm 0.43}$	$86.85_{\pm 0.51}$	$3.357_{\pm 0.071}$	$0.278_{\pm 0.005}$	$0.615_{\pm 0.008}$
		SGD	+.7% $74.46_{\pm 0.17}$	$92.66_{\pm 0.06}$	$1.083_{\pm 0.007}$	$0.113_{\pm 0.001}$	$0.376_{\pm 0.001}$
		IVON@mean	$74.51_{\pm 0.24}$	$92.74_{\pm 0.19}$	$1.284_{\pm 0.013}$	$0.152_{\pm 0.003}$	$0.399_{\pm 0.002}$
		IVON	$\mathbf{75.14}_{\pm 0.34}$	$\mathbf{93.30}_{\pm 0.19}$	$\mathbf{0.912}_{\pm 0.009}$	$\mathbf{0.021}_{\pm 0.003}$	$\mathbf{0.344}_{\pm 0.003}$

Understand Model Behavior

Small influence



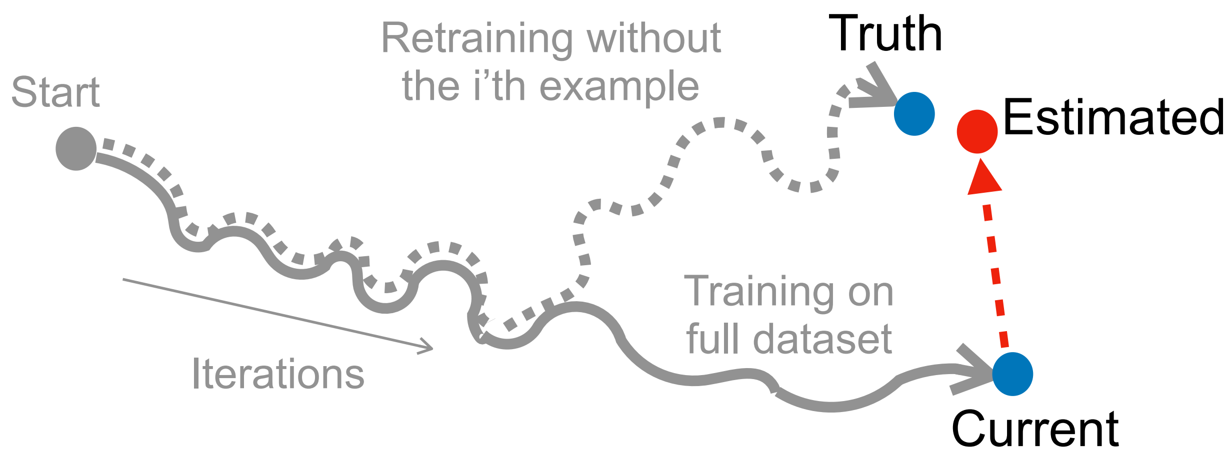
Large influence



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

Memory Perturbation Equation

Past that has the most influence on the present



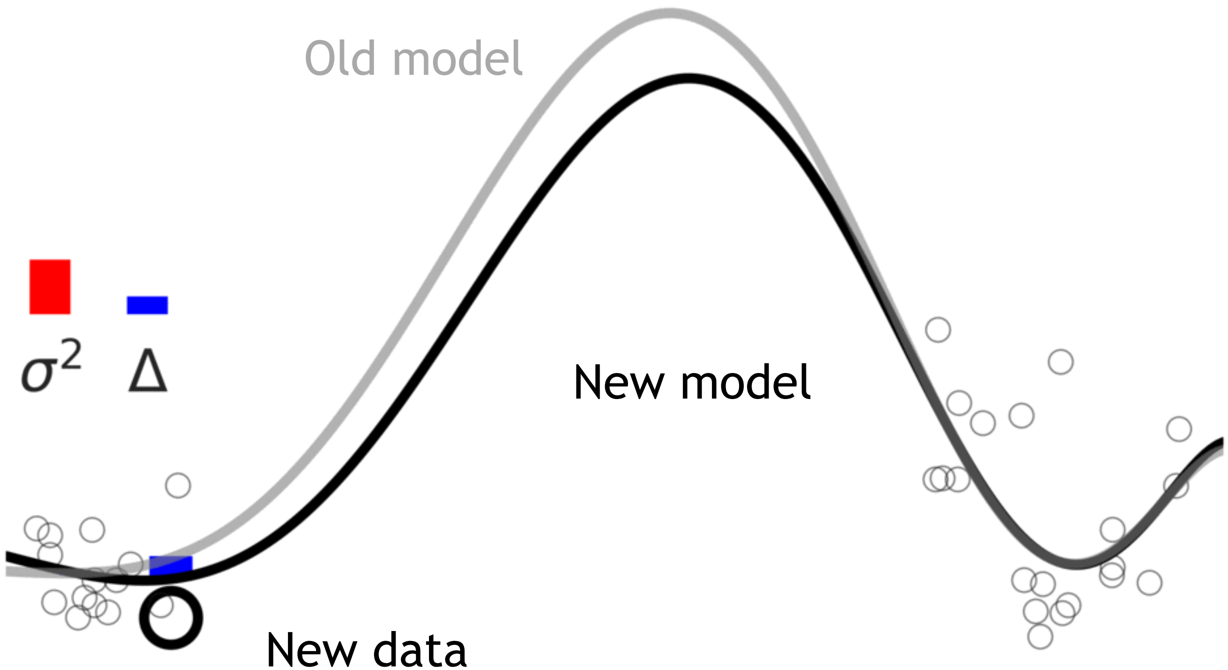
Estimating it without retraining: Using the BLR, we can recover all sorts of influence criteria used in literature.

Influence = predictError x predictVariance

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

Memory Perturbation

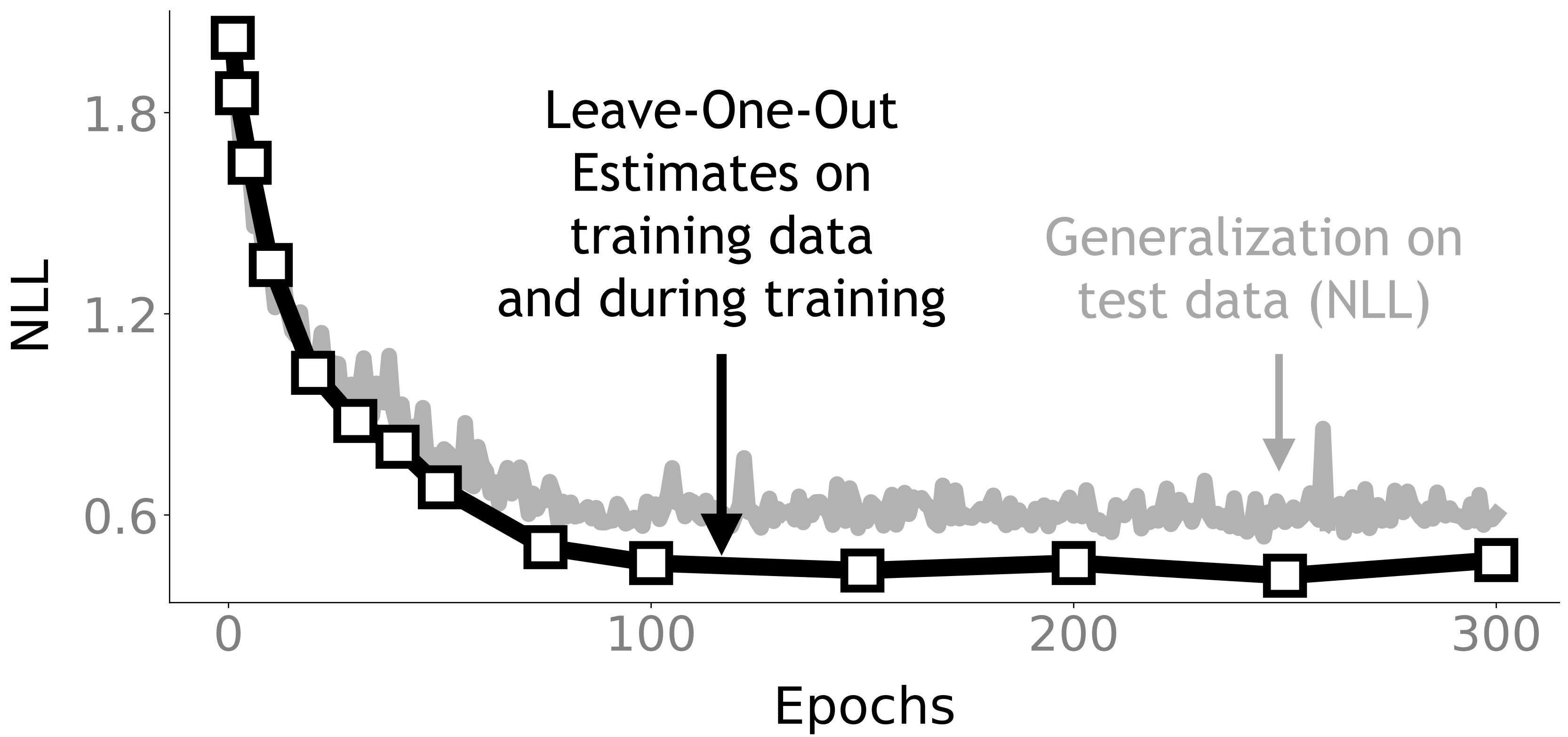
Influence (Δ) = predictionError *predictionVariance



1. Cook. Detection of Influential Observations in Linear Regression. Technometrics. ASA 1977
2. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS, 2023

Predict Future Performance

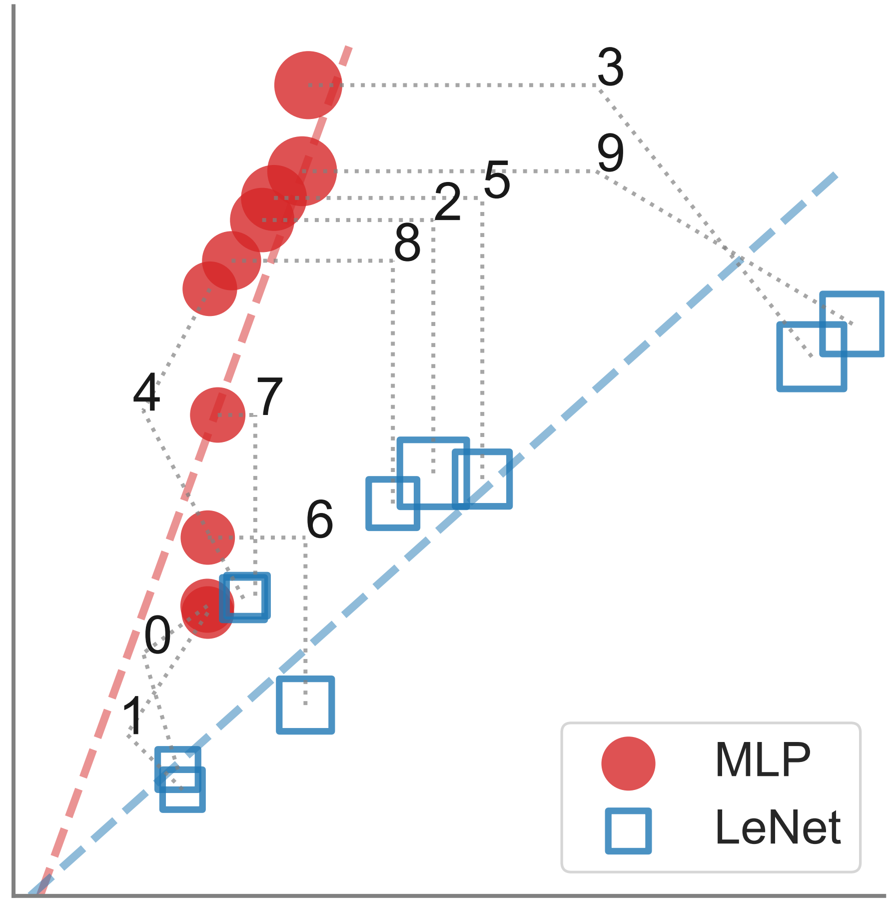
CIFAR10 on ResNet-20 using IVON



Answering “What-If” Questions

What if we removed a class from MNIST?

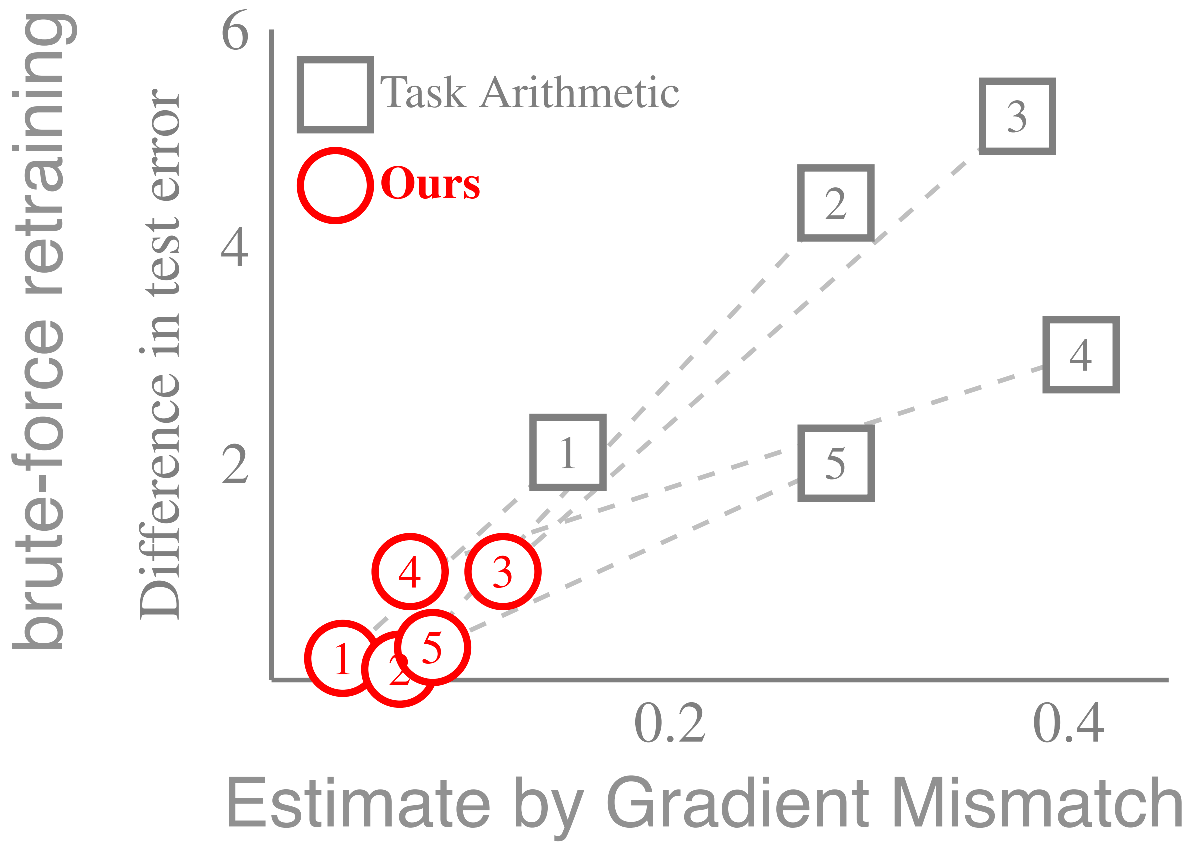
Estimates on training data (no retraining)



Test Performance (NLL) by brute-force retraining

Answering “What-If” Questions

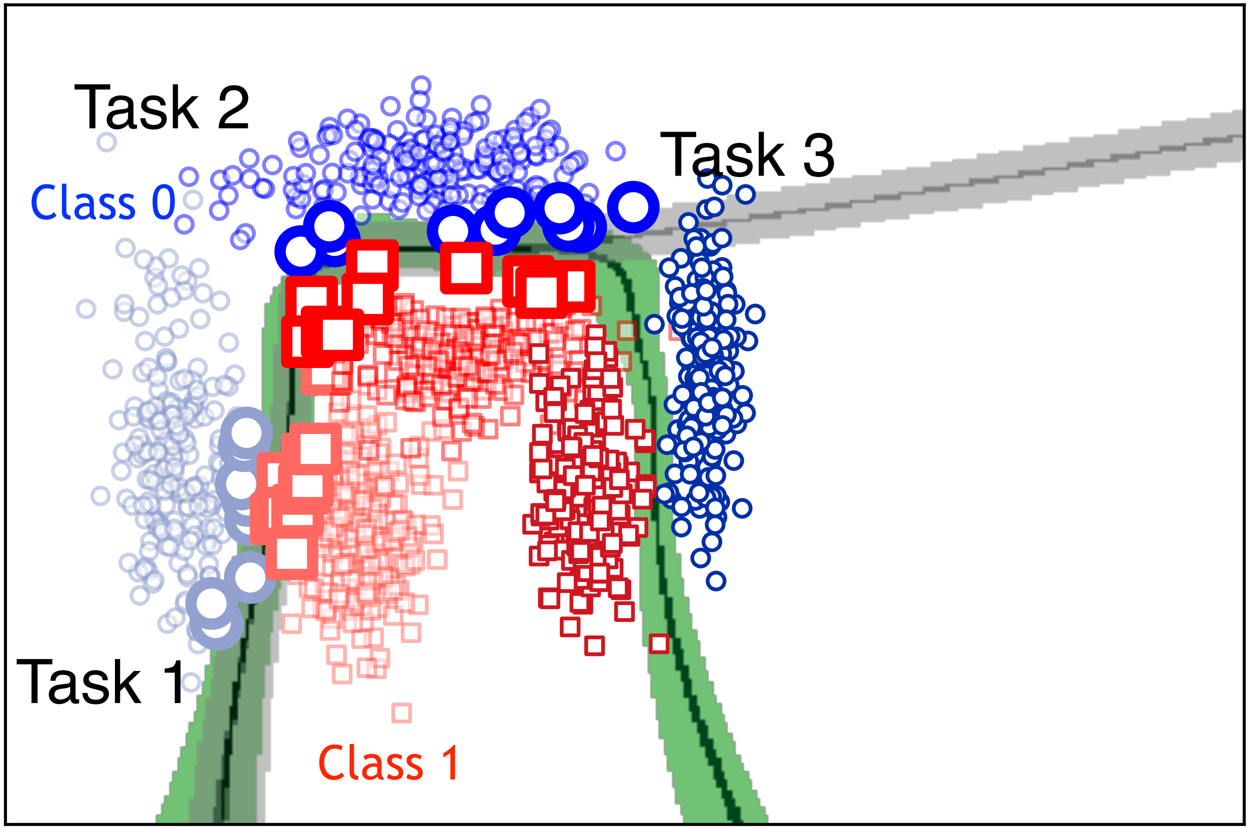
What if we merge fine-tuned large-language models?



RoBERTa
on IMDB

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

Learn Continually



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

Bayesian Learning Rule [1]

- Bridge DL & Bayesian learning [2-5]
  - SOTA on GPT-2 and ImageNet [5]
- Improve DL [5-7]
  - Calibration, uncertainty, memory etc.
  - Understand and fix model behavior
- Towards human-like quick adaptation

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
2. Khan, et al. Fast and scalable Bayesian deep learning by weight-perturbation in Adam, ICML (2018).
3. Osawa et al. Practical Deep Learning with Bayesian Principles, NeurIPS (2019).
4. Lin et al. Handling the positive-definite constraints in the BLR, ICML (2020).
5. Shen et al. Variational Learning is Effective for Large Deep Networks, Under review.
6. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
7. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)

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 (2021-2027) and Kakenhi Grants (2019-2021).

Bayes-Duality Workshop 2024

June 12-21, 2024, featuring around 20 speakers
https://bayesduality.github.io/workshop_2024.html


Adam White
University of Alberta, Canada


Alexander Immer
ETH, Switzerland


Arindam Banerjee
University of Illinois Urbana-Champaign, US


Daiki Chijiwa
NTT Corporation, Japan


Ehsan Amid
Google DeepMind, US


Eugene Ndiaye
Apple, France


Frank Nielsen
Sony Computer Science Laboratories, Japan


Jonghyun Choi
Seoul National University, South Korea


Juho Lee
KAIST, South Korea


Haavard Rue
KAUST, Saudi Arabia


Hossein Mobahi
Google Research, US

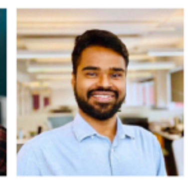

Martin Mundt
TU Darmstadt, Germany

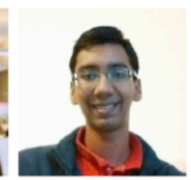

Matt Jones
University of Colorado, US

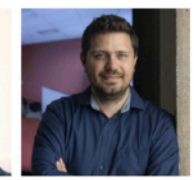

Nico Daheim
TU Darmstadt, Germany


Razvan Pascanu
Google DeepMind, US


Rupam Mahmood
University of Alberta, Canada

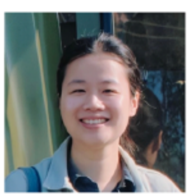

Sarath Chandar
École Polytechnique de Montréal, Canada

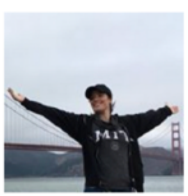

Siddharth Swaroop
Harvard University, US


Stephan Mandt
University of California, US


Tom Rainforth
University of Oxford, UK


Vincent Fortuin
Helmholtz AI, Germany


Yingzhen Li
Imperial College London, UK


Zelda Mariet
Bioptimus, US

Team Approx-Bayes

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


Emtiyaz Khan
Team Leader


Thomas Möllenhoff
Research Scientist


Geoffrey Wolfer
Special Postdoctoral Resesarcher


Hugo Monzón Maldonado
Postdoctoral Researcher

Many thanks to our group members and collaborators (many not on this slide).

We are always looking for new collaborations.


Keigo Nishida
Postdoctoral Researcher
RIKEN BDR


Zhedong Liu
Postdoctoral Researcher


Peter Nickl
Research Assistant


Joseph Austerweil
Visiting Scientist
University of Winsconsin-Madison


Pierre Alquier
Visiting Scientist
ESSEC Business School


Dharmesh Tailor
Remote Collaborator
University of Amsterdam