Presentation at UW-RIKEN AIP Workshop, March 20, 2025

Posterior's Sensitivity to Address AI's Uncertainty

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

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

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

Current state of ML

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

AI that can learn like us

AI that is low-cost, sustainable, transparent, trustworthy, reliable, composable, modular….

How to represent and adapt the knowledge?
Sensitivity to Perturbation (Duality)

Bayes-Duality

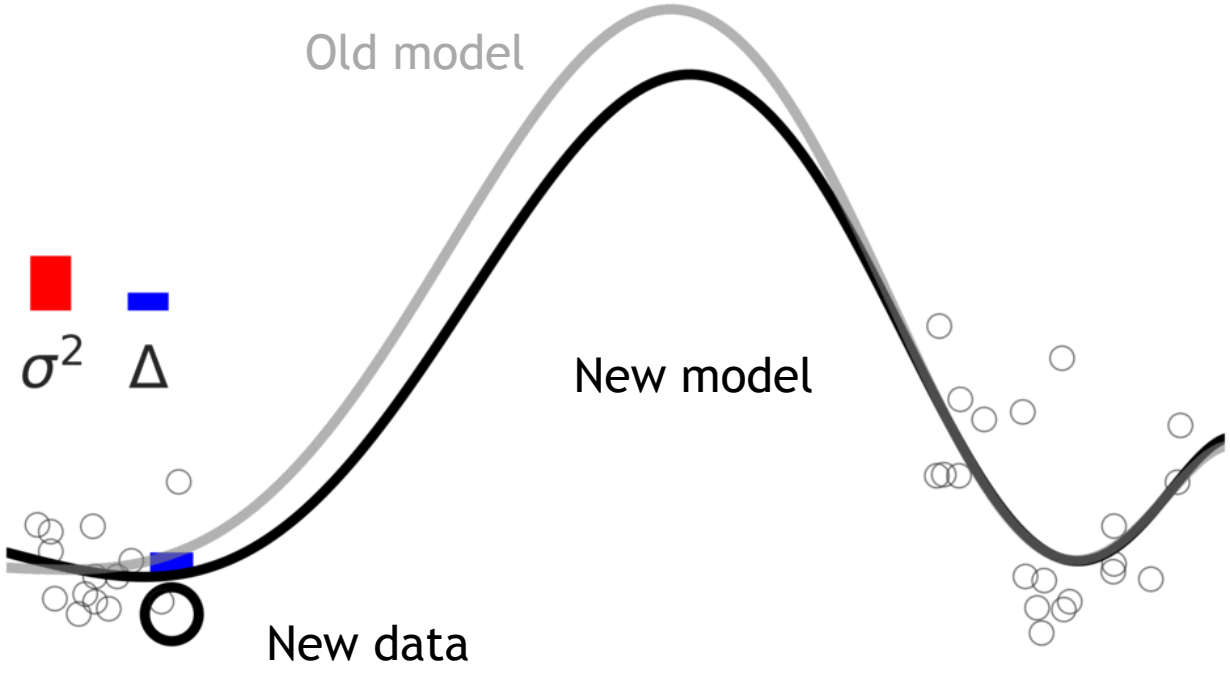
https://tenor.com/view/clockwork-gears-brain-gif-16784329

Sensitivity and Uncertainty

- Sensitivity of variational-posteriors for free
- Model sensitivity to data perturbation [1-3]
- Model perturbation: LLM model merging [4-5] and Federated learning [6]

1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)
2. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
3. Shen et al. Variational Learning is Effective for Large Deep Networks, ICML (2024)
4. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
5. Moldanado et al. How to Weight Multitask Finetuning? Fast Previews via Bayesian Model-Merging, (2024)
6. Swaroop et al. Connecting Federated ADMM to Bayes, ICLR, 2024

Model's Sensitivity to Data



Model is more sensitive to examples that are "far enough" (in the uncertain territories)

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

Sensitivity and Uncertainty

Linear regression $\ell_i = (y_i - x_i^\top \theta)^2/2$

$$\theta_t = \underbrace{H_t^{-1}}_{\text{Hessian}} \sum_{j=1}^{t} x_j y_j \implies \theta_t - \theta_t^{\setminus i} = H_t^{-1} x_i (y_i - x_i^\top \theta_t^{\setminus i})$$

$$x_i^\top (\theta_t - \theta_t^{\setminus i}) = \underbrace{x_i^\top H_t^{-1} x_i}_{\substack{\text{PredictionVariance} \\ \text{(epistemic)}}} \underbrace{(y_i - x_i^\top \theta_t^{\setminus i})}_{\substack{\text{Prediction Error} \\ \text{(aleatoric)}}}$$

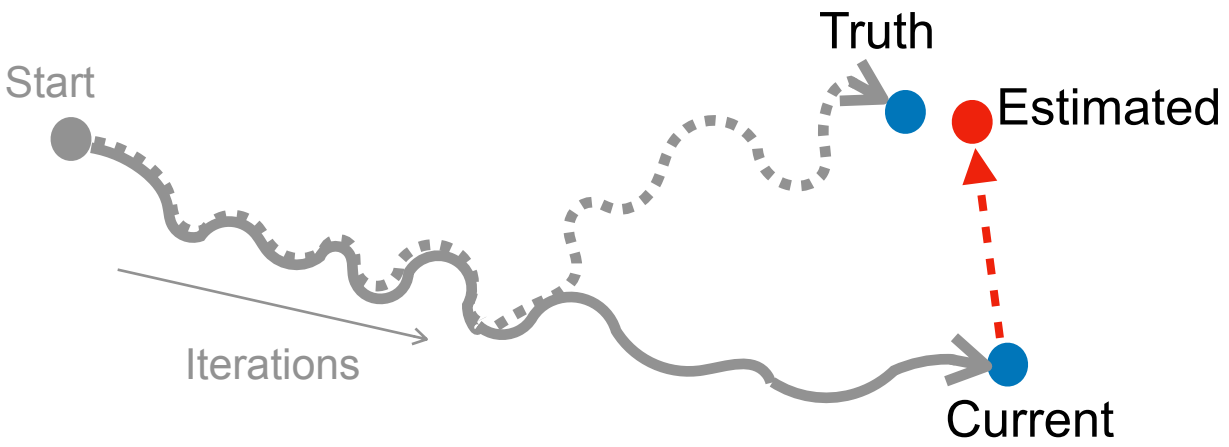
$$= - x_i^\top H_t^{-1} \nabla \ell_i(\theta_t^{\setminus i})$$

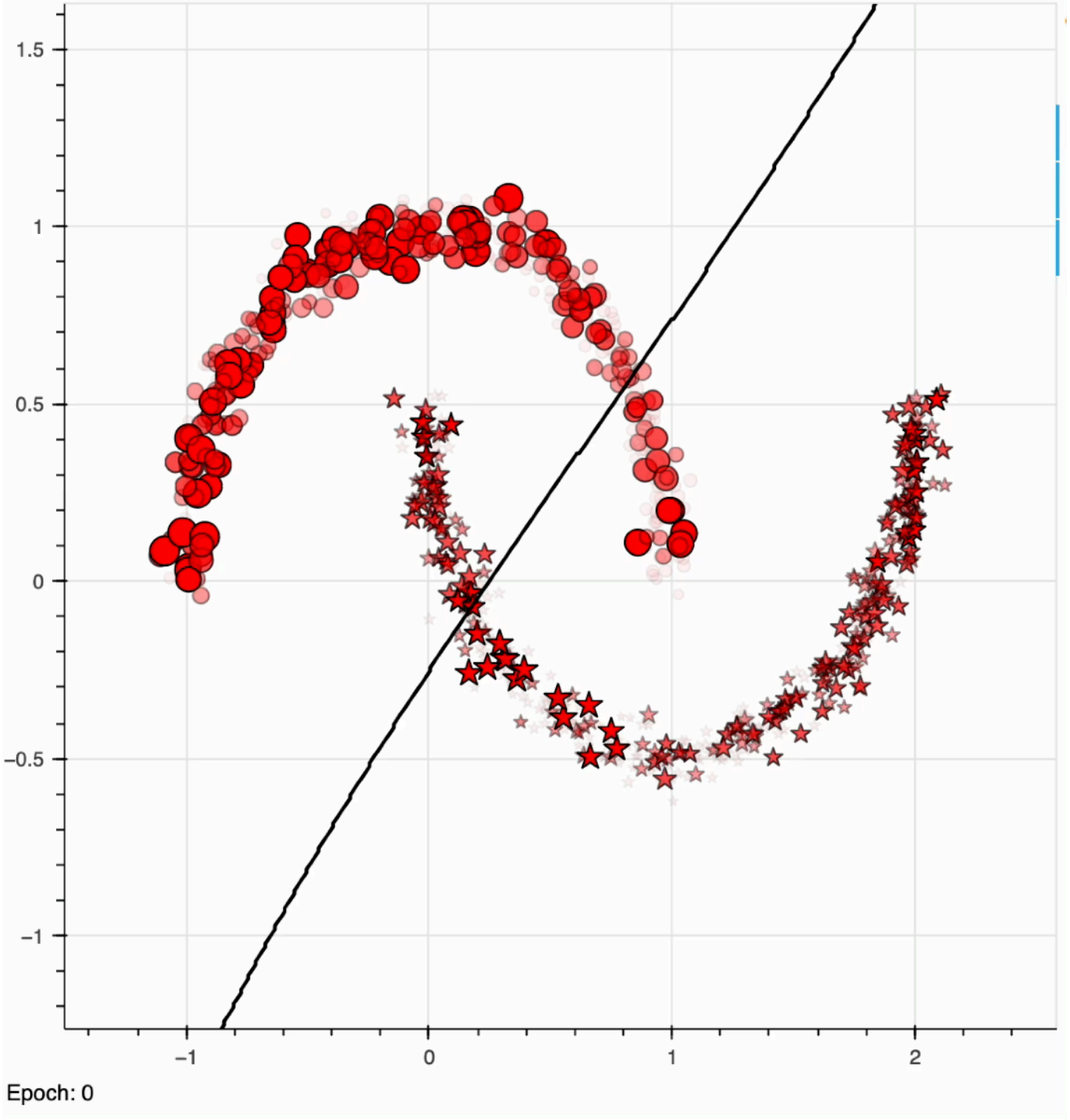
Forms the basis for most works in deep learning [2], with most works analyzing leave-one-out sensitivity to data-perturbation at convergence.

1. Cook. Detection of Influential Observations in Linear Regression. Technometrics. ASA (1977)
2. Koh and Liang. Understanding Black-Box Predictions via Influence Functions. ICML (2017)

We hope to address broader scenarios

- Sensitivity & uncertainty essential for “what-if” questions
  - Data Perturbation: What if we add/remove a class? All NY times articles? Continual/active learning
  - Model Perturbation: What if we merge separately fine-tuned LLMs? Federated/distributed learning
  - Algorithm perturbation, etc. etc.





Binary classification on the Two Moons dataset.

Big markers with red indicate influential examples for an iterations of an Adam-like algorithm (IVON).

This is a simpler version only using gradient wrt mean.


Peter Nickl


Lu Xu


Dharmesh Tailor


Thomas Moellenhoff

Memory-Perturbation

Broadening data-attribution by using posterior-sensitivity

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

Conjugate Exponential-Family Models

$$\theta_t - \theta_t^{\setminus i} = H_t^{-1} x_i (y_i - x_i^\top \theta_t^{\setminus i}) \;\; = - H_t^{-1} \nabla \ell_i(\theta_t^{\setminus i})$$

We will extend this to posterior's sensitivity

$$q_t \propto \prod_{j=0}^{t} e^{-\ell_j} \qquad q_t^{\setminus i} \propto \prod_{j=0, j\neq i}^{t} e^{-\ell_j} \qquad \frac{q_t}{q_t^{\setminus i}} \propto e^{-\ell_i}$$

$$e^{\lambda_t^\top T(\theta)} \qquad e^{(\lambda_t^{\setminus i})^\top T(\theta)} \qquad e^{\tilde{\lambda}_i^\top T(\theta)}$$

Natural parameter

$$\lambda_t - \lambda_t^{\setminus i} = \tilde{\lambda}_i$$

Lin-reg is a special case [1, Thm. 1]

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

Generalization using Natural Gradients

$$\lambda_t - \lambda_t^{\setminus i} = \widetilde{\lambda}_i$$

This can be generalized to cover all sorts perturbation, e.g., during training, to handle model merging, continual learning, federated learning etc.

How? The $\widetilde{\lambda}_i$ can be written as gradient wrt "dual" coordinates expectation parameters $\mu = \mathbb{E}_q[T(\theta)]$

$$\widetilde{\lambda}_{i|t} = \nabla_{\mu_t} \mathbb{E}_{q_t}[-\ell_i]$$

A type of natural Gradients (see Sec 2 in [1])

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

Bayesian learning rule

Learning Algorithm	Posterior Approx.	Natural-Gradient Approx.
Optimization Algorithms		
Gradient Descent	Gaussian (fixed cov.)	Delta method
Newton's method	Gaussian	——"——
Multimodal optimization (New)	Mixture of Gaussians	——"——
Deep-Learning Algorithms		
Stochastic Gradient Descent	Gaussian (fixed cov.)	Delta method, stochastic approx.
RMSprop/Adam	Gaussian (diagonal cov.)	Delta method, stochastic approx., Hessian approx., square-root scaling, slow-moving scale vectors
Dropout	Mixture of Gaussians	Delta method, stochastic approx., responsibility approx.
STE	Bernoulli	Delta method, stochastic approx.
Online Gauss-Newton (OGN) (New)	Gaussian (diagonal cov.)	Gauss-Newton Hessian approx. in Adam & no square-root scaling
Variational OGN (New)	——"——	Remove delta method from OGN
BayesBiNN (New)	Bernoulli	Remove delta method from STE
Approximate Bayesian Inference Algorithms		
Conjugate Bayes	Exp-family	Set learning rate $\rho_t = 1$
Laplace's method	Gaussian	Delta method
Expectation-Maximization	Exp-Family + Gaussian	Delta method for the parameters
Stochastic VI (SVI)	Exp-family (mean-field)	Stochastic approx., local $\rho_t = 1$
VMP	——"——	$\rho_t = 1$ for all nodes
Non-Conjugate VMP	——"——	——"——
Non-Conjugate VI (New)	Mixture of Exp-family	None

They all compute natural gradients.

$$\widetilde{\lambda}_{j|t} = \nabla_{\mu_t} \mathbb{E}_{q_t}[-\ell_j]$$

To estimate sensitivity, we take a step back

$$\lambda_t - \lambda_t^{\setminus i} \approx \widetilde{\lambda}_{i|t}$$

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

Bayesian Learning Rule (BLR) [1]

Many ML algorithms compute the quantity (approx.).
IOW, they are approximately Bayesian!

$$q_t \propto \prod_{j=0}^{t} e^{-\ell_j} \quad = \arg\min_{q \in \mathbb{Q}} \sum_{j=1}^{t} \mathbb{E}_q[\ell_j] + KL(q \| p_0)$$

$$\lambda_t = \sum_{j=0}^{t} \underbrace{\nabla_{\mu_t} \mathbb{E}_{q_t}[-\ell_j]}_{\widetilde{\lambda}_{j|t}} \quad \Longrightarrow \quad \lambda_t = \sum_{j=0}^{t} \widetilde{\lambda}_{j|t}$$

BLR:

$$\lambda_t \leftarrow (1-\rho)\lambda_t + \rho \sum_{j=0}^{t} \widetilde{\lambda}_{j|t}$$

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

Sensitivity Estimates: Adam and IVON

RMSprop/Adam

1 $\hat{g} \leftarrow \hat{\nabla}\ell(\theta)$

2 $\hat{h} \leftarrow \hat{g}^2$

3 $h \leftarrow (1-\rho)h + \rho\hat{h}$

4 $\theta \leftarrow \theta - \alpha(\hat{g} + \delta m)/(\sqrt{h} + \delta)$

5

Sensitivity

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

1 $\hat{g} \leftarrow \hat{\nabla}\ell(\theta)$ where $\theta \sim \mathcal{N}(m, \sigma^2)$

2 $\hat{h} \leftarrow \hat{g} \cdot (\theta - m)/\sigma^2$

3 $h \leftarrow (1-\rho)h + \rho\hat{h} \; + \rho^2(h-\hat{h})^2/(2(h+\delta))$

4 $m \leftarrow m - \alpha(\hat{g} + \delta m)/(h + \delta)$

5 $\sigma^2 \leftarrow 1/(N(h+\delta))$

Sensitivity

Sensitivity is cheaply obtained by using 1 step of the algorithms.
Adam's sensitivity (uncertainty) is poorer compared to IVON.
Check out the blog: https://team-approx-bayes.github.io/blog/ivon/

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).
4. Shen, et al. "Variational Learning is Effective for Large Deep Networks." ICML (2024)

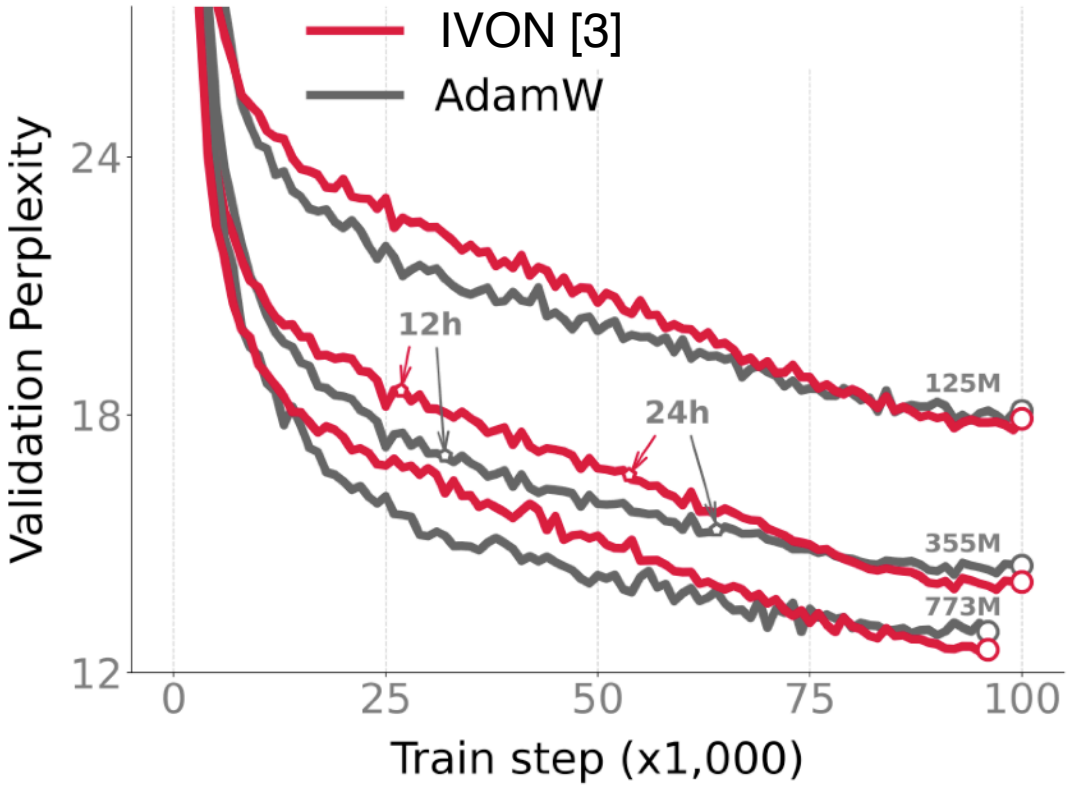
Drop-in replacement of Adam

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()
```

GPT-2 with IVON

Better performance & uncertainty at the same cost



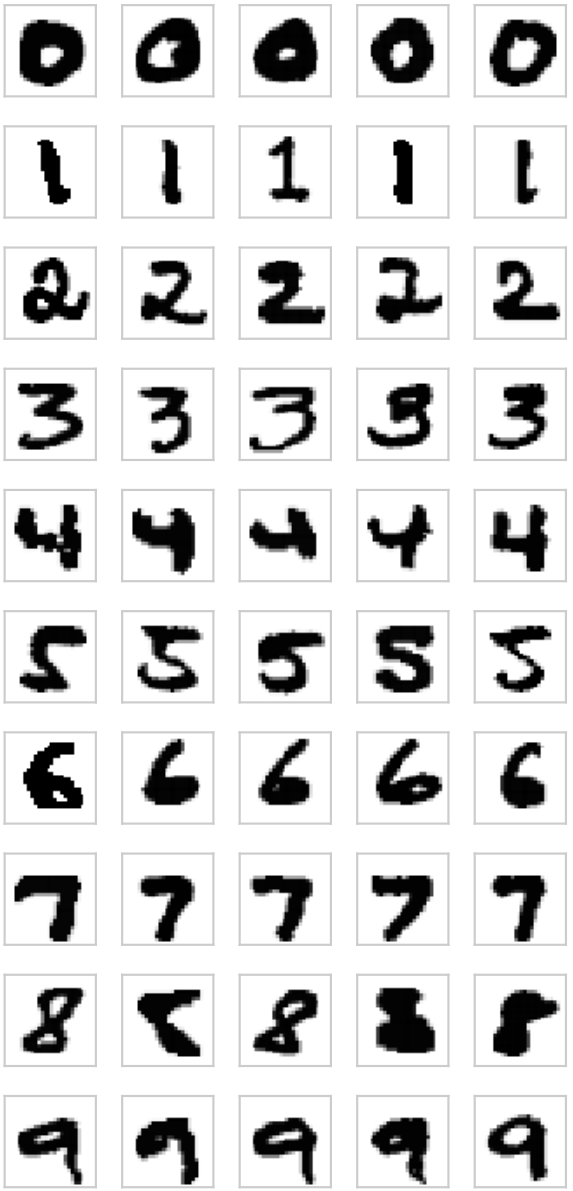
Trained on OpenWebText data (49.2B tokens).

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 Deep Networks, ICML (2024)

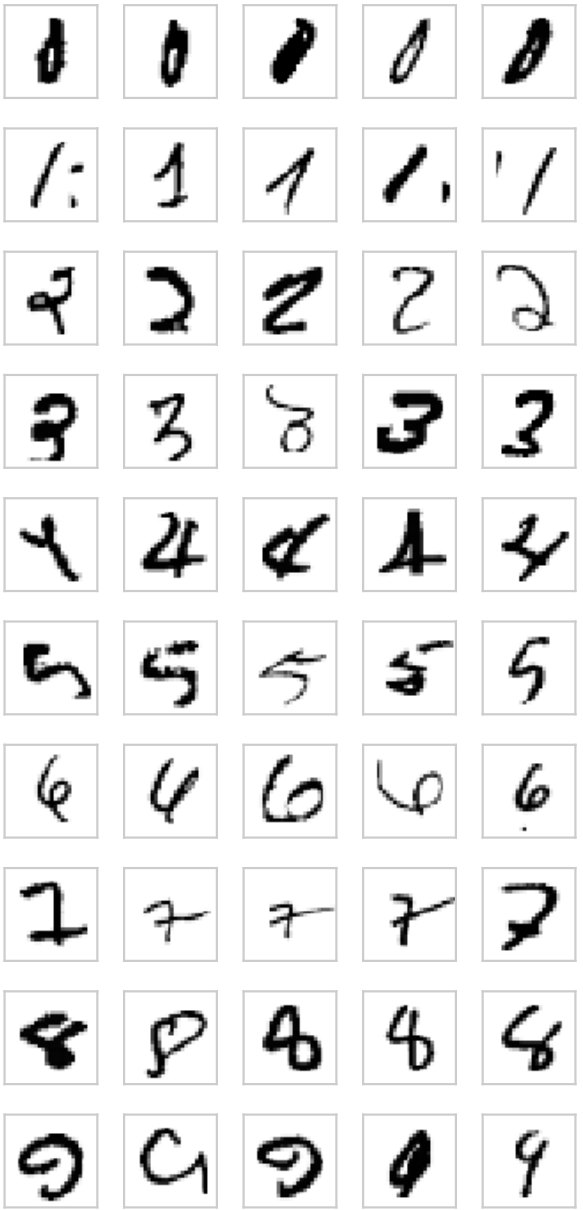
Low Sensitivity



To estimate sensitivity, just take a step back

$$\lambda_t^{\setminus i} - \lambda_t \approx -\widetilde{\lambda}_{i|t}$$

High Sensitivity



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

Guess the ImageNet class [1]

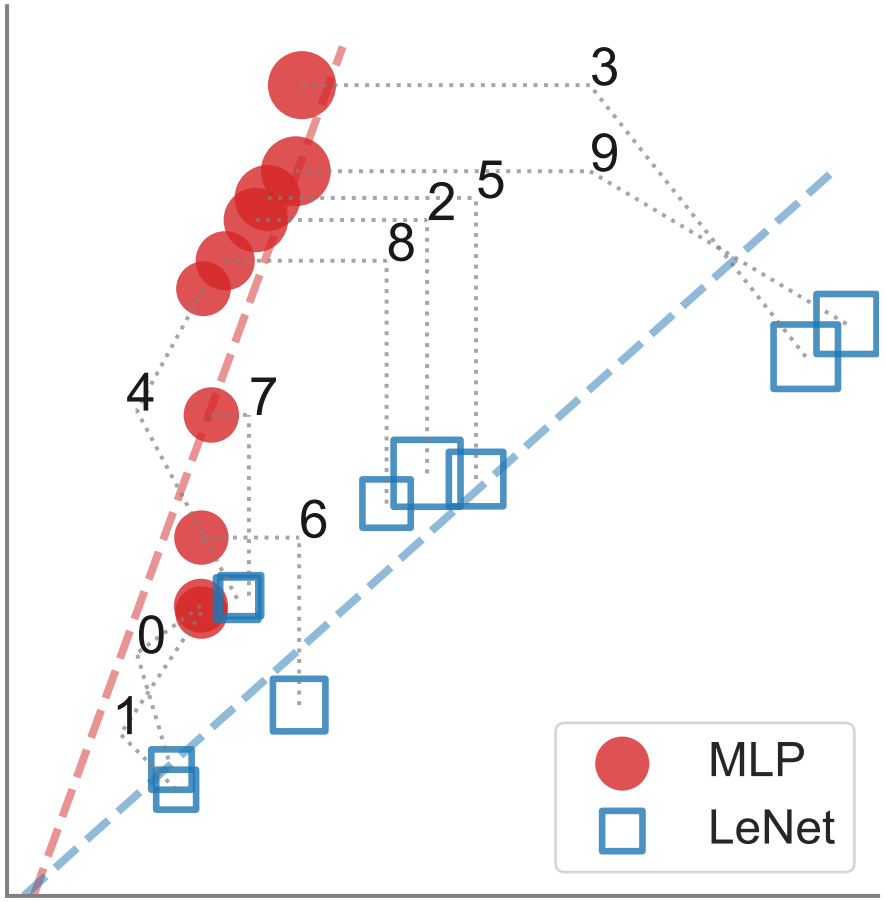





1. Shen et al. Variational Learning is Effective for Large Deep Networks, ICML, 2024

Answering “What-If” Questions

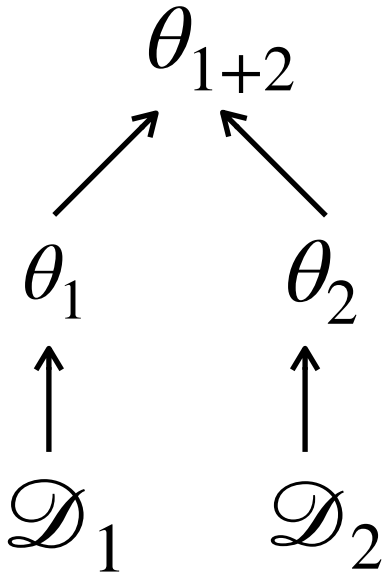
What if we removed a class from MNIST?



Estimates on training data (no retraining)

Test Performance (NLL) by brute-force retraining

Model Merging



Given θ_1 fine-tuned on $\mathscr{D}_1$ and θ_2 fine-tuned on $\mathscr{D}_2$, merge them (to estimate θ_{1+2}).

Simplest strategy: $\alpha_1\theta_1 + \alpha_2\theta_2$ [1].

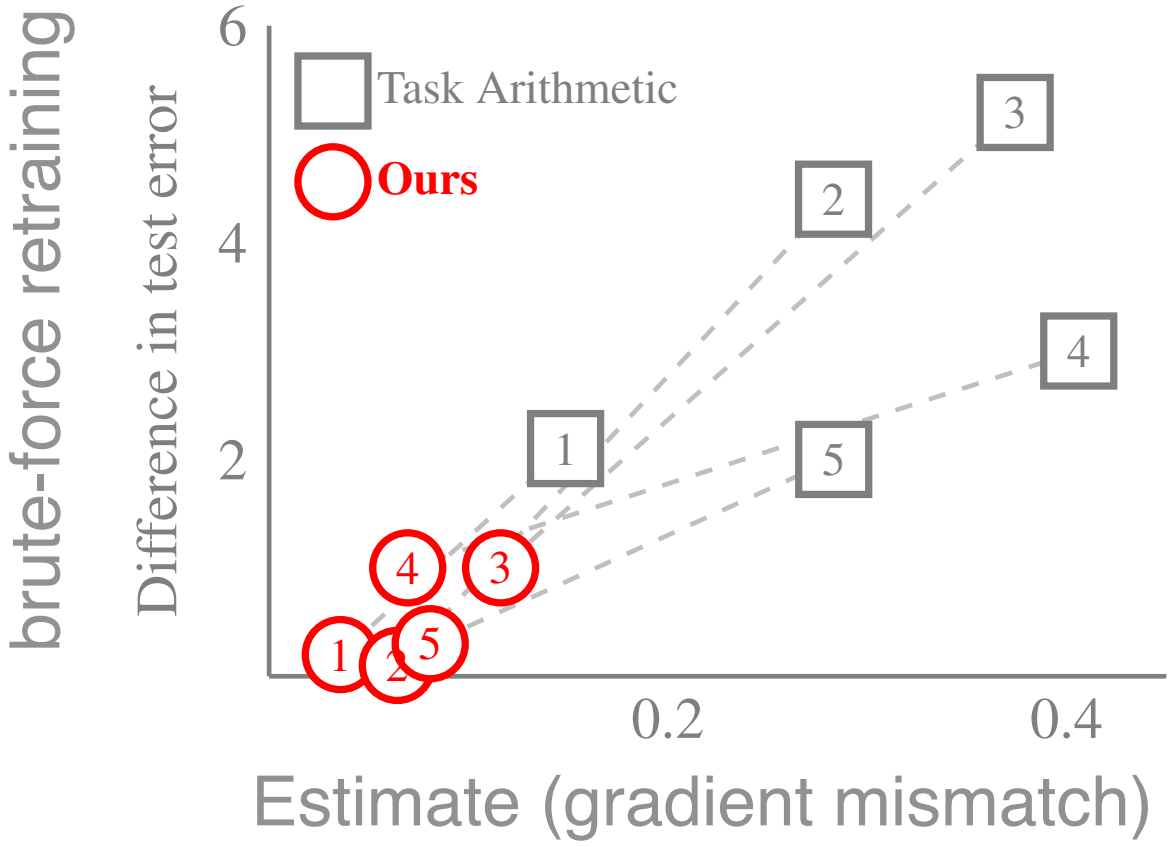
A generalization is to use $\alpha_1\lambda_1 + \alpha_2\lambda_2$ [3], eg, use Hessian which is necessarily better [2]

$$H_{1+2}\theta_{1+2} \approx \alpha_1 H_1\theta_1 + \alpha_2 H_2\theta_2$$

$$\implies \theta_{1+2} - \theta_1 \approx H_{1+2}^{-1}\nabla\ell_1(\theta_1) \text{ (Thm 1, [2])}$$

1. Wortsman et al. Robust fine-tuning of zero-shot models, CVPR 2022
2. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
3. Maldonado et al. ….. Fast Previews via Bayesian Model-Merging (under review, 2024)

“What-if” we merged models



RoBERTa
on IMDB

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

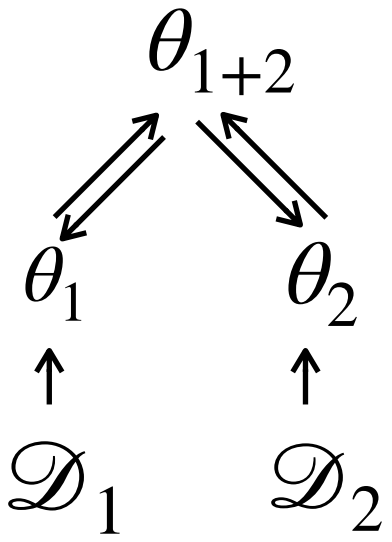
Federated Learning

The $\widetilde{\lambda}_i$ are dual variables (Lagrange multiplier) [1-4]

Eg, dual variables in federated ADMM automatically emerges through $\widetilde{\lambda}_i$ in variational Bayes [4]

$$\lambda_{1+2} \leftarrow \widetilde{\lambda}_1 + \widetilde{\lambda}_2$$

Federated Learning



1. Khan et al. Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models, ICML, 2013
2. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Processes, NeurIPS, 2019
3. Adam et al. Dual Parameterization of Sparse Variational Gaussian Processes, NeurIPS, 2021
4. Swaroop et al. Connecting Federated ADMM to Bayes, ICLR, 2024

Sensitivity and Uncertainty

- Sensitivity of (variational) posteriors to address uncertainty during knowledge transfer
  - without increasing the cost
- Model sensitivity to data perturbation [1-3]
- Model perturbation: LLM model merging [4-5] and Federated learning [6]

1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)
2. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
3. Shen et al. Variational Learning is Effective for Large Deep Networks, ICML (2024)
4. Daheim et al. Model merging by uncertainty-based gradient matching, ICLR (2024).
5. Moldanado et al. How to Weight Multitask Finetuning? Fast Previews via Bayesian Model-Merging, (2024)
6. Swaroop et al. Connecting Federated ADMM to Bayes, ICLR, 2024

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

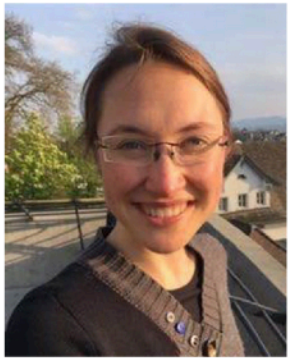
Received total funding of JPY 220M + EUR 500K through the CREST-ANR grant! Thanks to JST for their generous funding!

Bayes-Duality Workshop (June 25-27, 2025)

https://bayesduality.github.io/workshop_2025.html

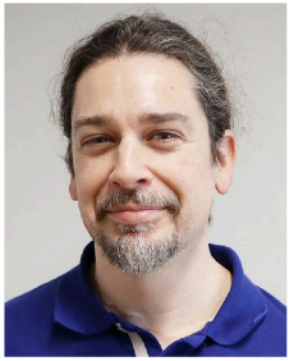
Abeba Birhane

Trinity College
Dublin, Ireland



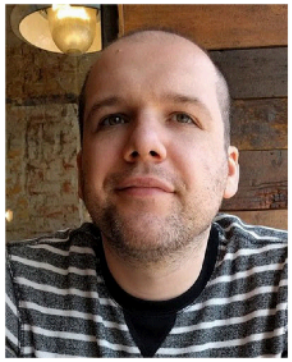
Rebekka Burkholz

Helmholtz Center
CISPA, Germany



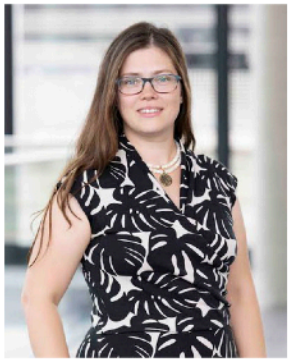
André Martins

Instituto Superior
Tecnico, Portugal



Razvan Pascanu

Deepmind, UK



Anna Rohrbach

TU Darmstadt,
Germany

Marcus Rohrbach

TU Darmstadt,
Germany

Daniel Roy

University of
Toronto, Canada

Diverse topics: Bayes, Optimization, Information Geometry, Continual Learning, Federated Learning, Active learning, RL, Model understanding, Data Attributions, LLMs, etc.

Team Approx-Bayes

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

Sin-Han Yang
Technical Staff

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

Marco Miani
Intern
Technical University of Denmark

Rin Intachuen
Intern
Mahidol University

Alexander Timans
Intern
University of Amsterdam

Masaki Adachi
Intern
University of Oxford

Adrian R. Minut
Intern
Sapienza, University of Rome

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

Visit us! Let's collaborate!
Also see open (post-doc) positions on the webpage