Presentation at BIRS Workshop on Uncertainty Quantification for Neural Networks, Feb 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.

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

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

Quickly adapt & continue to acquire new skills.

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

Why haven't we solved it with Bayes?

- In theory, Bayes can solve these problems
  - By using the posterior uncertainty
- But, these are not Bayesian models
- And scale makes it infeasible
- Are there alternatives for Bayes?

Sensitivity and Uncertainty

- Sensitivity of (variational) posteriors to address uncertainty during knowledge transfer
  - Main point: the sensitivity is (essentially) freely available!
- Model sensitivity to data perturbation (addition/removal)
  - Beyond linear regression: conjugate-Bayes [1]
  - Beyond conjugacy [1,2]
  - For large models (VI for GPT-2, ImageNet) [3]
- Model perturbation: LLM model merging [4-5]
  - Federated learning [6] and connections to Bayes-duality

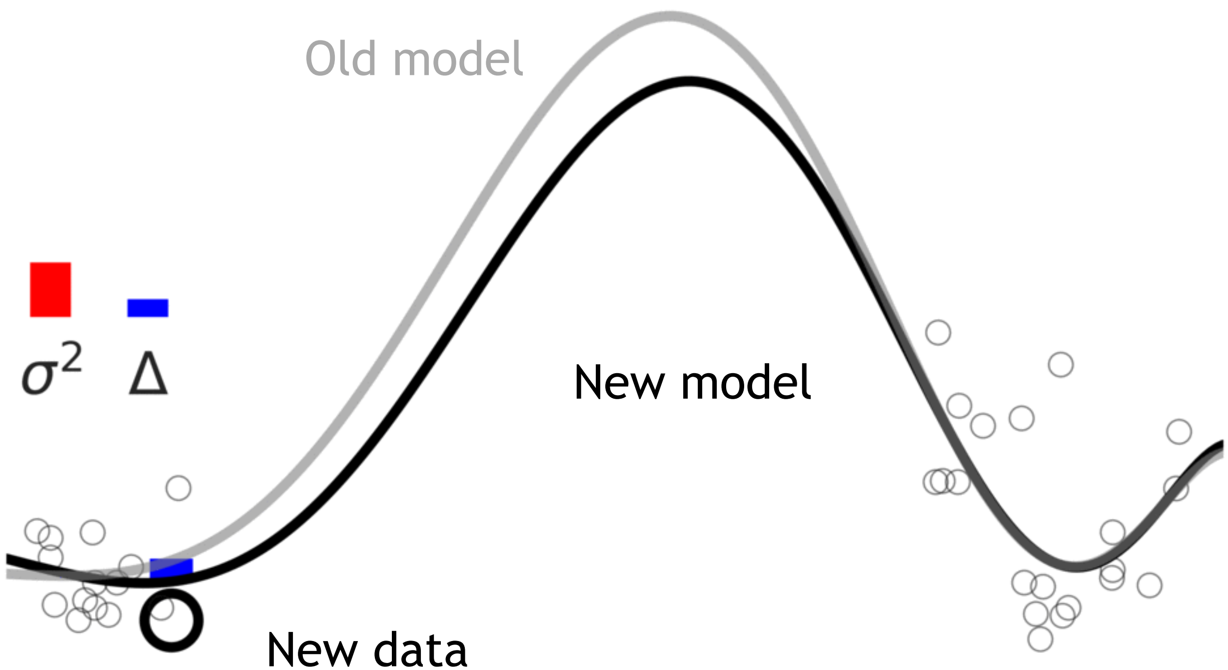
1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)
2. Khan and Rue, The Bayesian Learning Rule, JMLR (2023).
3. Shen et al. Variational Learning is Effective for Large Deep Networks, ICML (2024)
4. Daheim et al. Model merging by uncertainty-based gradient matching, 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

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

Bayes-Duality

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

Model's Sensitivity to Its Training Data



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

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

Closed-form Expression for Sensitivity

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

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

$$x_i^\top(\theta_t - \theta_t^{\backslash i}) = \underbrace{x_i^\top H_t^{-1} x_i}_{\text{PredictionVariance}} \underbrace{(y_i - x_i^\top \theta_t^{\backslash i})}_{\text{PredictionError}}$$

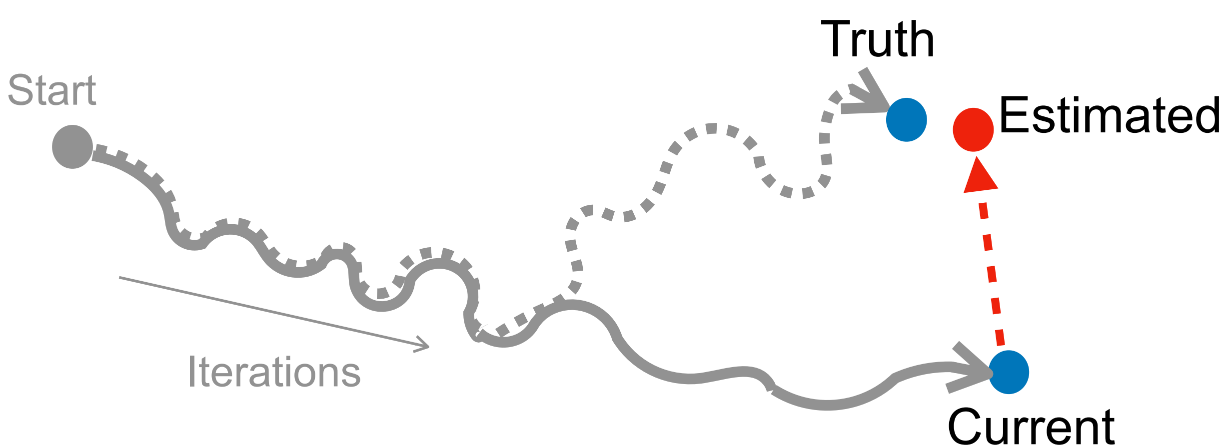
$$= - x_i^\top H_t^{-1} \nabla \ell_i(\theta_t^{\backslash i})$$

This result is the basis for most works in deep learning [2], but these extensions are too narrow (leave-one-out, at convergence, for data-attribution).

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)

A Broader Perspective

- Sensitivity is essential to answer “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.



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

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

Conjugate Exp-Fam 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

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

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

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

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

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

Linear Regression as a special case

$$q_t = \mathcal{N}(\theta_t, H_t^{-1})$$

$$\propto e^{-\frac{1}{2}\theta_t^\top H_t \theta + Tr\left(-\frac{1}{2}H_t \theta\theta^\top\right)}$$

$$T(\theta) = \left(\theta, \theta\theta^\top\right)$$

$$\lambda_t = \left(H_t\theta_t, -H_t/2\right)$$

$$q_t^{\backslash i} = \mathcal{N}(\theta_t^{\backslash i}, H_t^{\backslash i -1})$$

$$\lambda_t^{\backslash i} = \left(H_t^{\backslash i}\theta_t^{\backslash i}, -H_t^{\backslash i}/2\right)$$

$$e^{-\ell_i} \propto e^{-\frac{1}{2}(y_i - x_i^\top\theta)^2}$$

$$\propto e^{y_i x_i^\top \theta + Tr\left(-\frac{1}{2}x_i x_i^\top \theta\theta^\top\right)}$$

$$\widetilde{\lambda}_i = \left(y_i x_i, -x_i x_i^\top/2\right)$$

$$\lambda_t - \lambda_t^{\backslash i} = \widetilde{\lambda}_i \implies H_t\theta_t - H_t^{\backslash i}\theta_t^{\backslash i} = y_i x_i$$

$$H_t - H_t^{\backslash i} = x_i x_i^\top$$

$$\theta_t - \theta_t^{\backslash i} = H_t^{-1} x_i (y_i - x_i^\top \theta_t^{\backslash i})$$

This addresses all issues!

Group level sensitive (with just addition)

$$\lambda_t - \lambda_t^{\backslash i} = \sum_{i \in \mathcal{M}} \widetilde{\lambda}_i$$

Holds at every t during online updating

Can be generalized to neural-network training iterations too (but also to model perturbation and other types of perturbations).

We need "dual" coordinates: $\mu = \mathbb{E}_q[T(\theta)]$

Going beyond conjugacy

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

$$\lambda_t - \lambda_t^{\setminus i} = \widetilde{\lambda}_i \qquad\qquad = \nabla_{\mu_t} \mathbb{E}_{q_t}[-\ell_i]$$

$$e^{-\ell_i} \propto e^{\widetilde{\lambda}_i^\top T(\theta)} \quad \Longrightarrow -\ell_i = \widetilde{\lambda}_i^\top T(\theta) + const\,.$$

$$\Longrightarrow \; \mathbb{E}_{q_t}[-\ell_i] = \widetilde{\lambda}_i^\top \mu_t + const\,.$$

$$\Longrightarrow \; \nabla_{\mu_t} \mathbb{E}_{q_t}[-\ell_i] = \widetilde{\lambda}_i$$

Using this relation we can recover measures used in deep learning (Thm 2-4). Available for free!

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

Bayesian Learning Rule (BLR) [1]

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

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

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

BLR:

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

To estimate sensitivity, we take a step back

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

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

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).

Improved Variational Online Newton

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

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, 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)

IVON 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

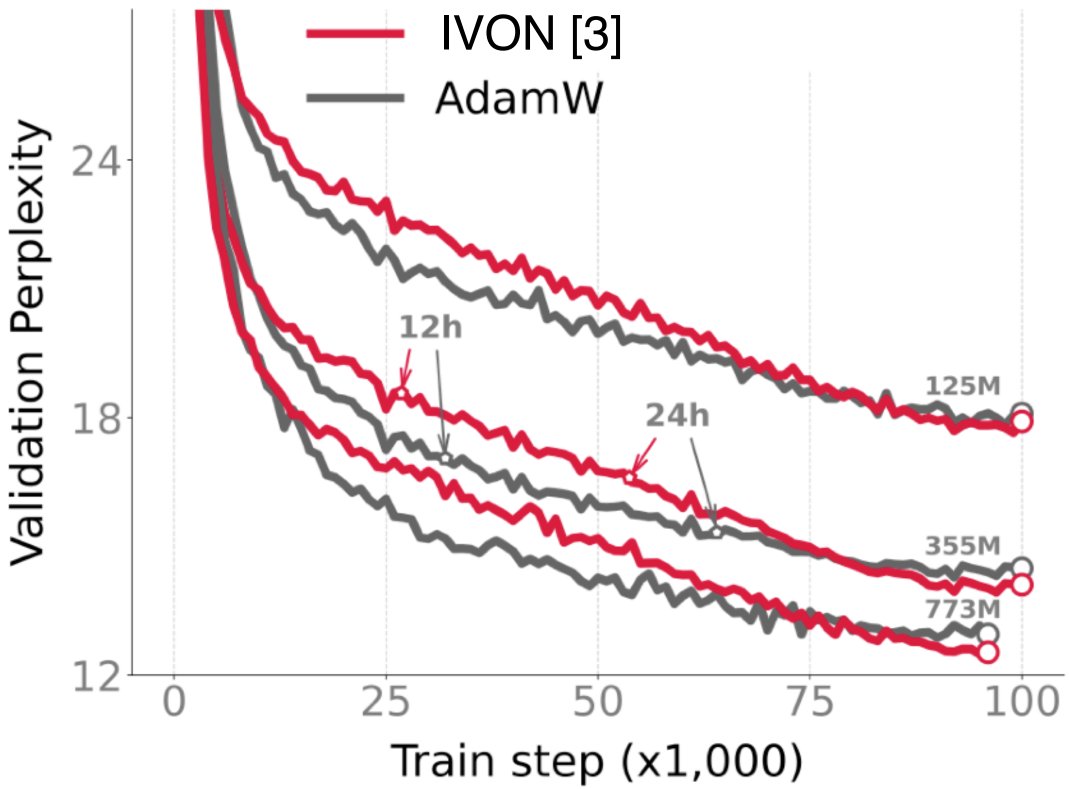
2 Computer Vision Group
Technical University of Munich, Germany

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 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)

Drop-in replacement of Adam

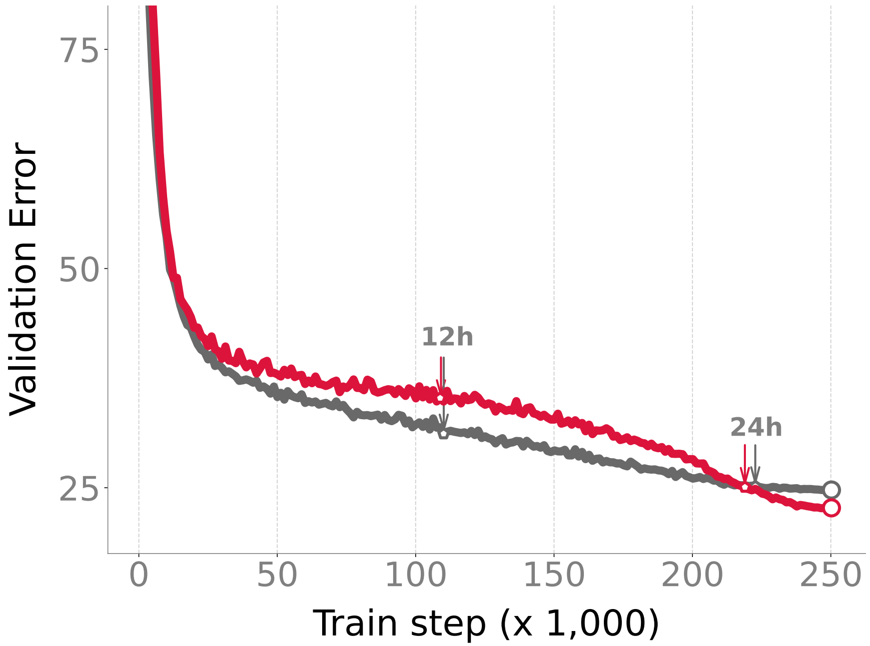
https://github.com/team-approx-bayes/ivon

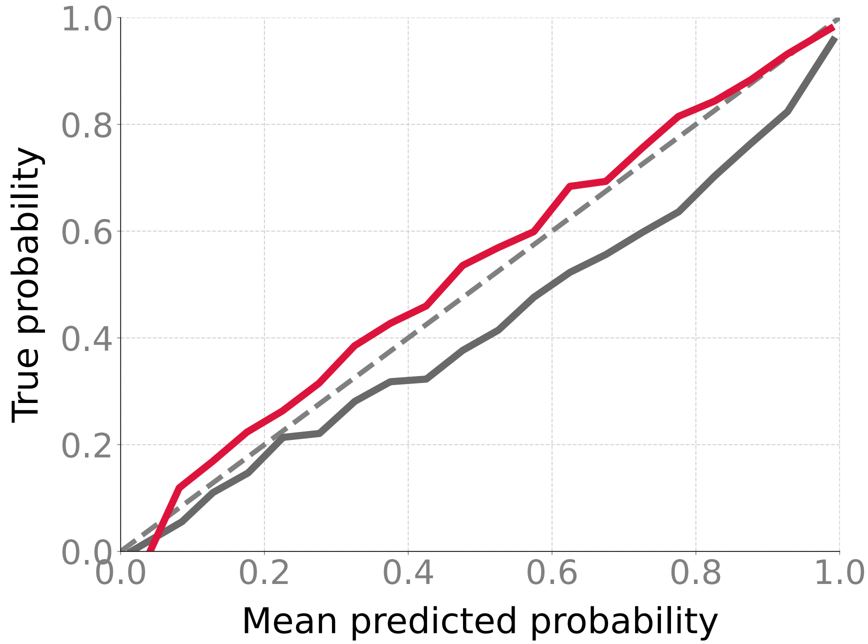
```
import torch
+import ivon

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()
```

Don’t use BBB
Use IVON!

Better Calibration

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





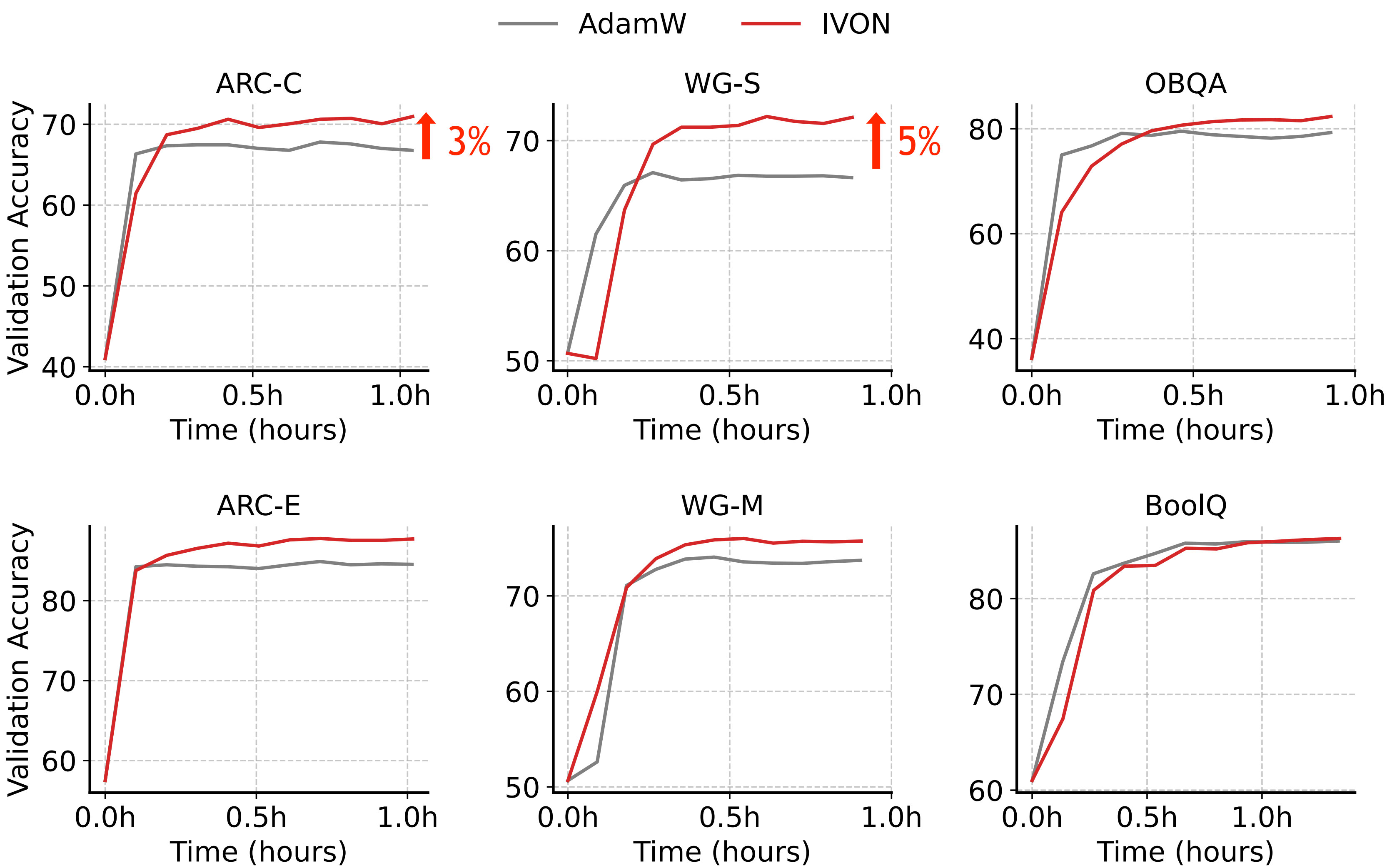
No Severe Overfitting

….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		$\mathbf{75.14}_{\pm 0.34}$	$\mathbf{93.30}_{\pm 0.19}$	$\mathbf{0.912}_{\pm 0.009}$	$\mathbf{0.021}_{\pm 0.003}$	$\mathbf{0.344}_{\pm 0.003}$

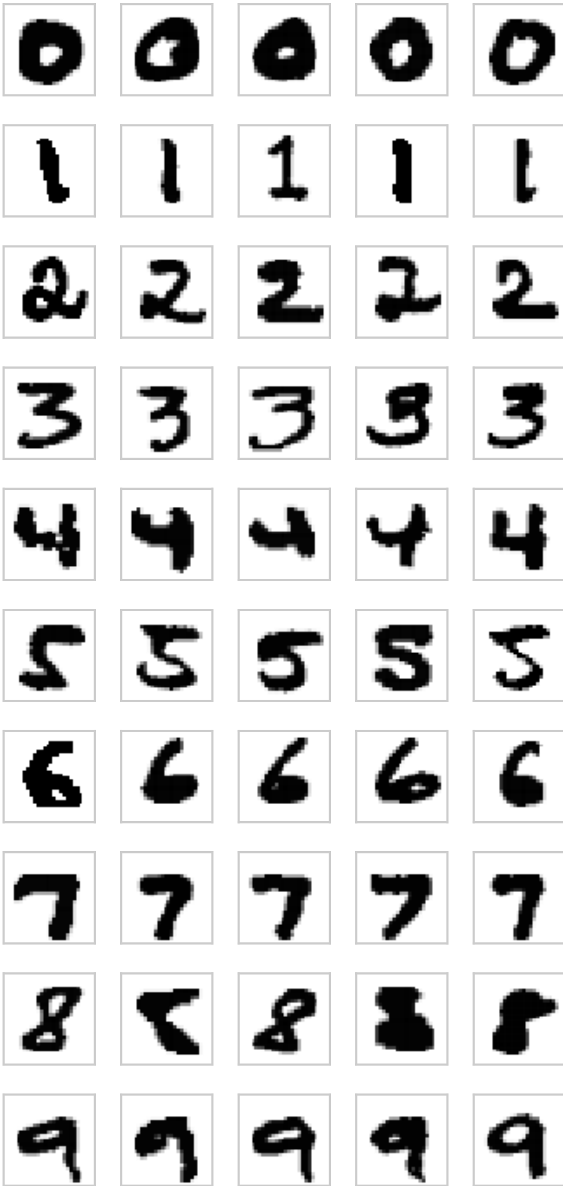
LoRA Finetuning

Llama 2
(7 billion)



1. Bai et al. “Variational Low-Rank Adaptation Using IVON”, FITML workshop at NeurIPS 2024

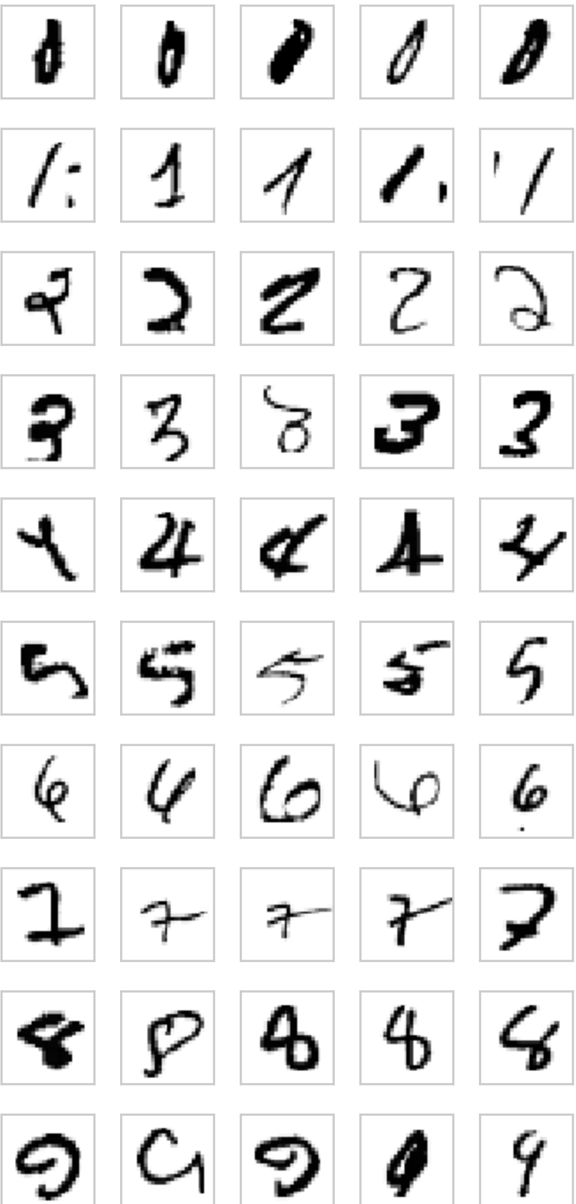
Low Sensitivity



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

High Sensitivity

Traffic light (ImageNet)

What class is this?

Low Sensitivity

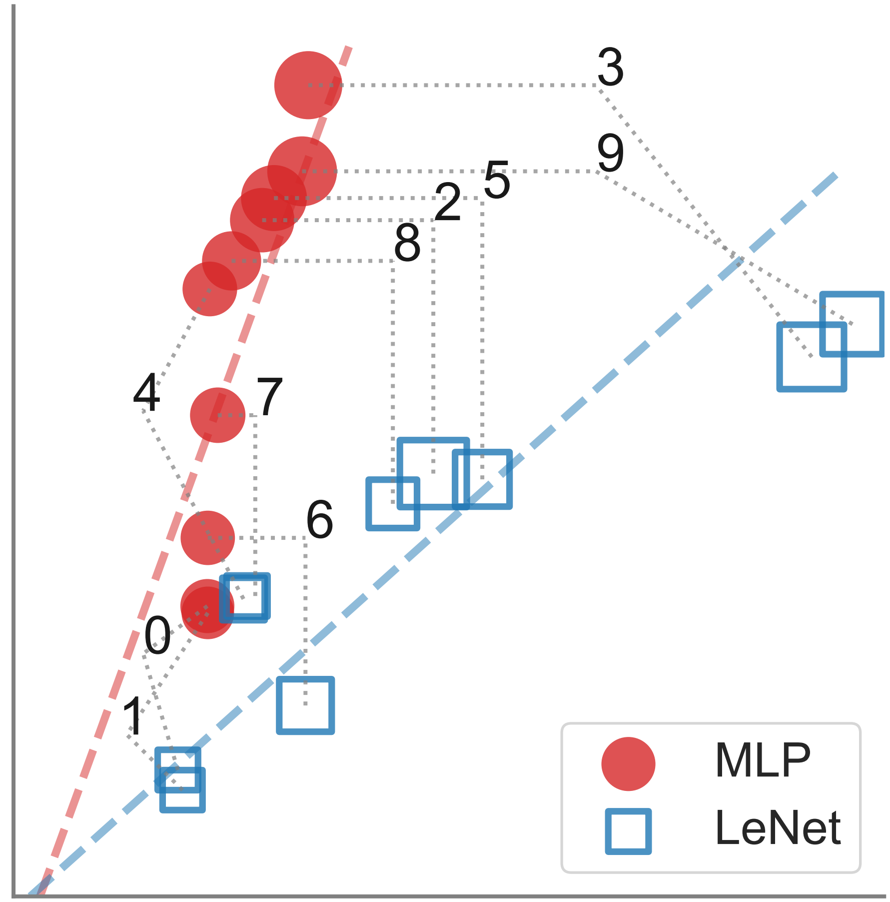
High Sensitivity

Chihuahua class (ImageNet)

Low Sensitivity

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

$$\begin{tikzcd} & \theta_{1+2} & \\ \theta_1 \arrow[ur] & & \theta_2 \arrow[ul] \\ \mathscr{D}_1 \arrow[u] & & \mathscr{D}_2 \arrow[u] \end{tikzcd}$$

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].

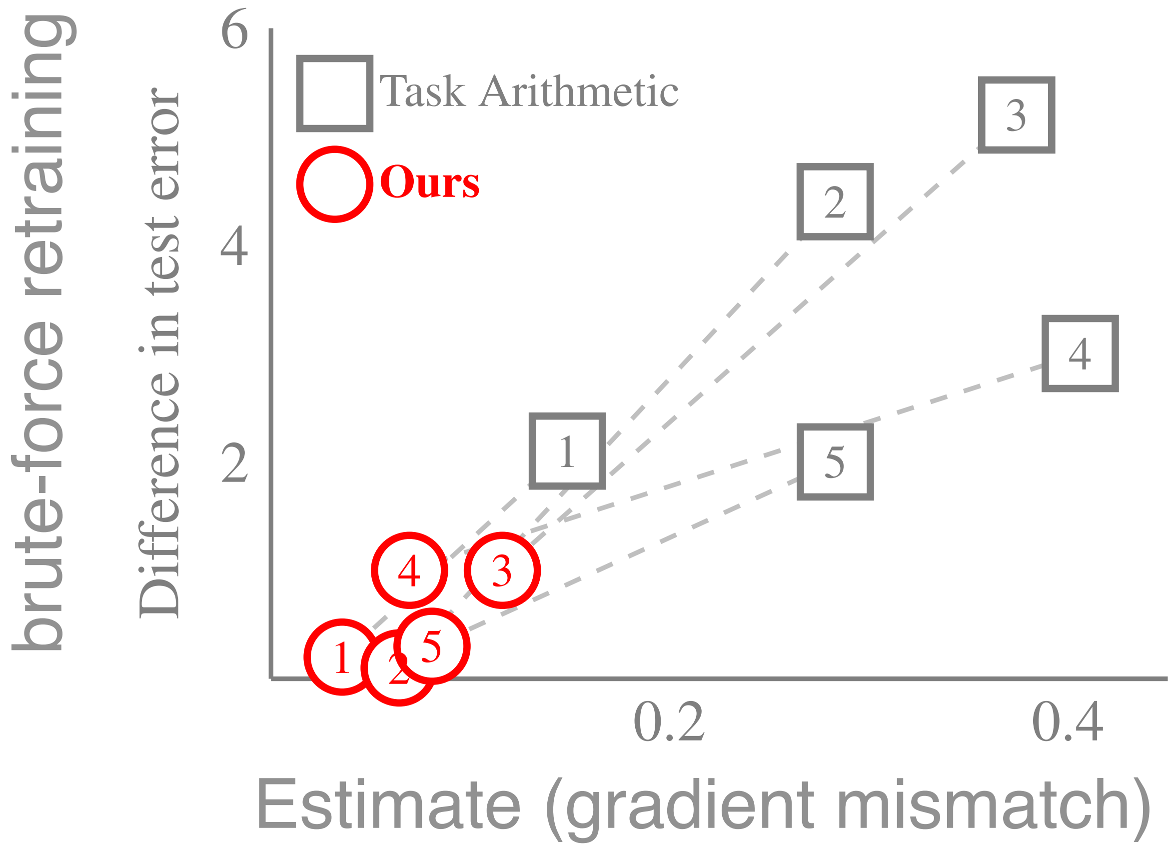
A generalization is to use $\alpha_1\lambda_1 + \alpha_2\lambda_2$ [3], eg, use Hessian which is necessarily better [2]

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

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

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



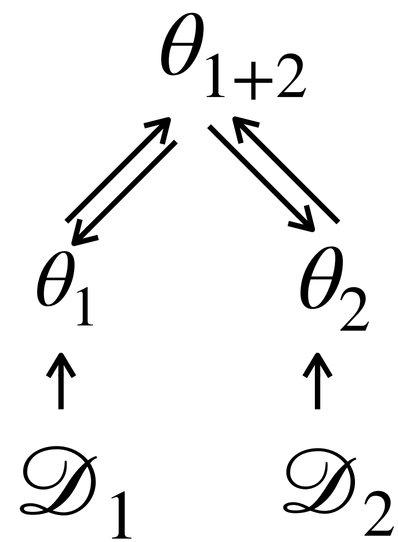
RoBERTa
on IMDB

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

Bayesian Duality

The variables $\widetilde{\lambda}_i$ are dual variables (Lagrange multipliers). In fact, variational posteriors have an equivalent dual representation in terms of $\widetilde{\lambda}_i$ [1-4]

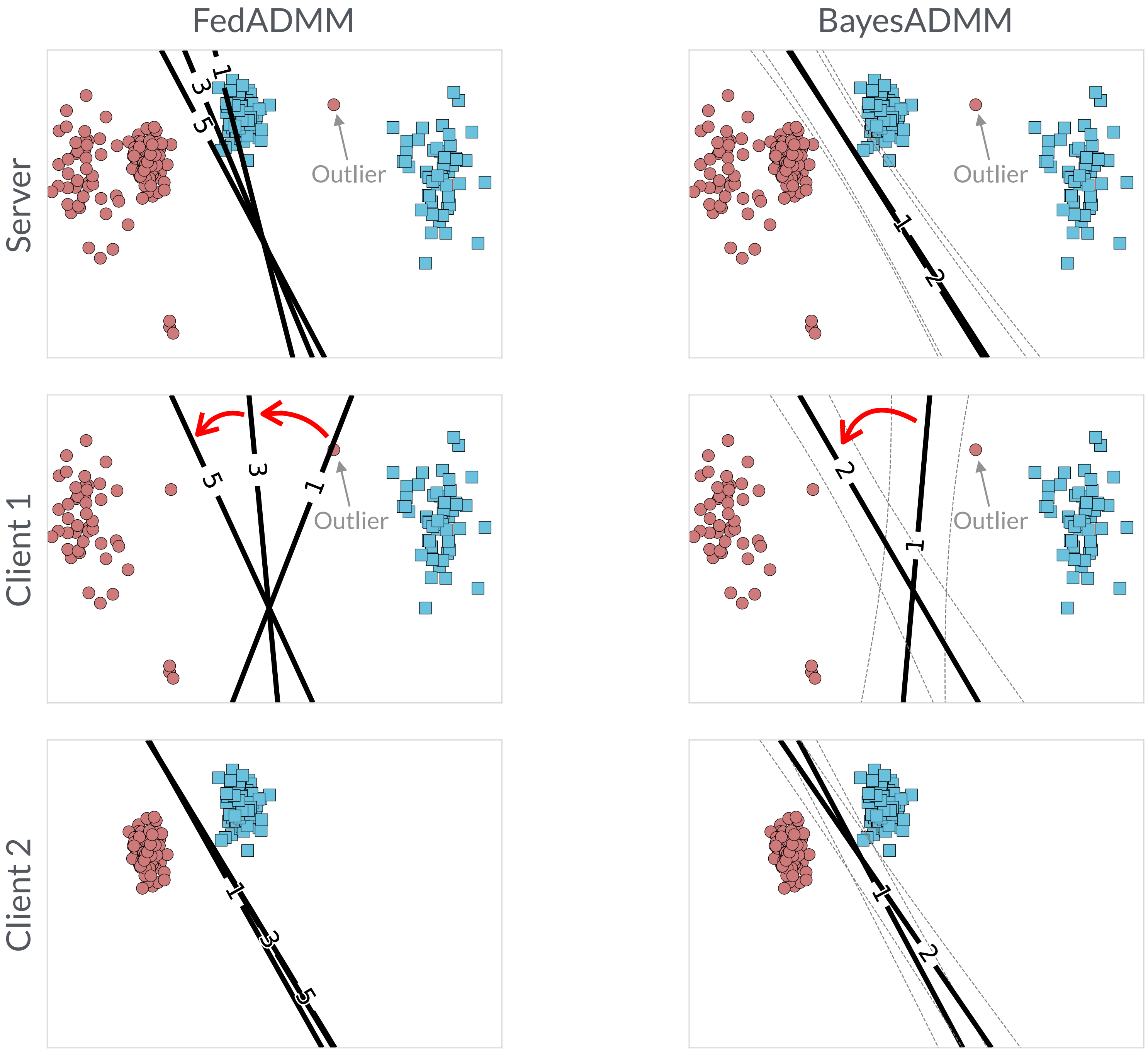
Federated Learning



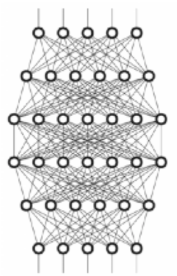
Eg, dual variables in federated ADMM automatically emerges through $\widetilde{\lambda}_i$ in variational Bayes [4]

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

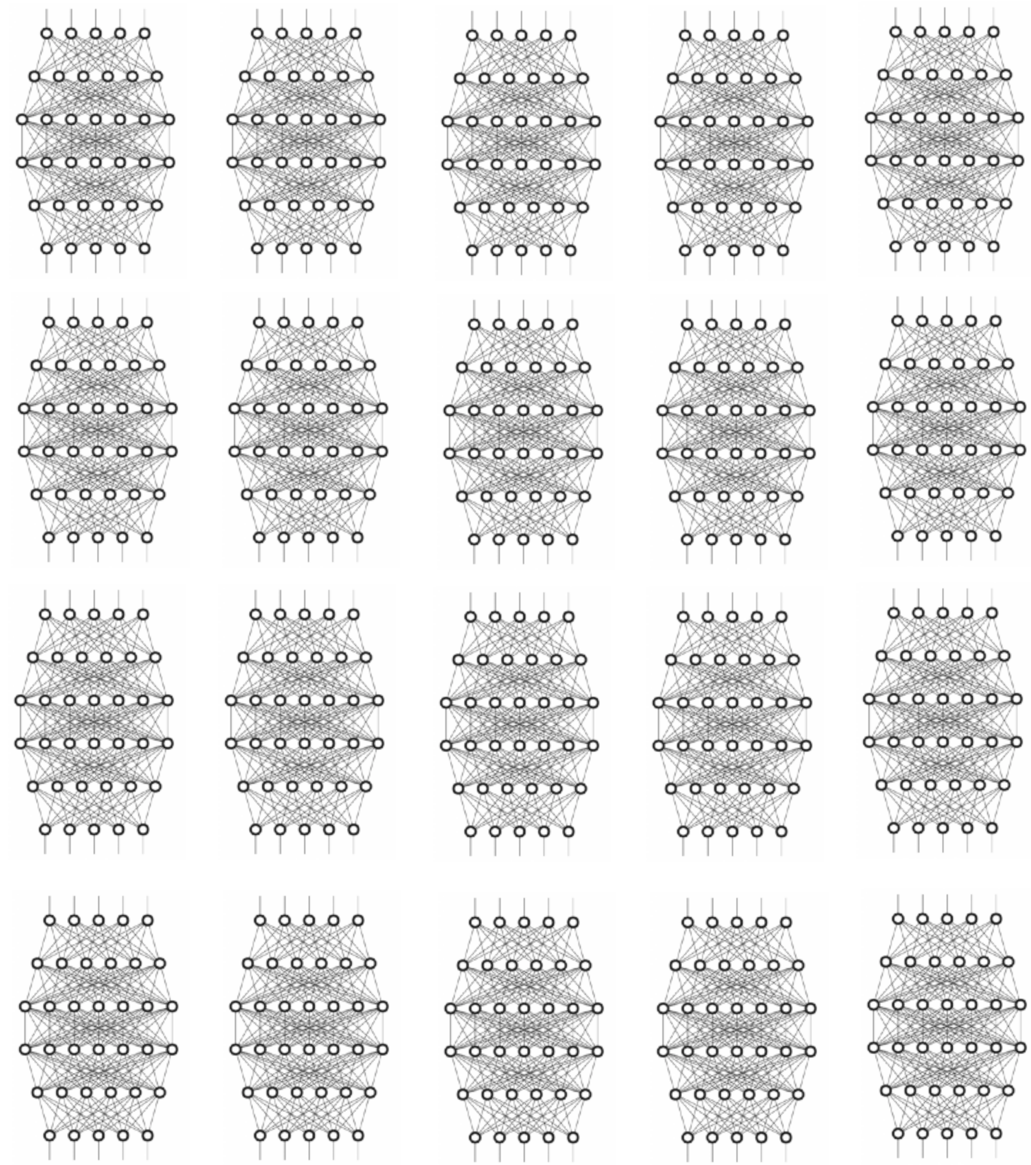
1. Khan et al. Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models, ICML, 2013
2. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Processes, NeurIPS, 2019
3. Adam et al. Dual Parameterization of Sparse Variational Gaussian Processes, NeurIPS, 2021
4. Swaroop et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


FedADMM
BayesADMM
Server
Client 1
Client 2
Outlier
Outlier
Outlier
Outlier
1
3
5
1
2
5
3
1
2
1
1
3
5
1
2

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)

Sensitivity and Uncertainty

- Sensitivity of (variational) posteriors to address uncertainty during knowledge transfer
  - Main point: it is essentially available for free!
- Model sensitivity to training-data perturbation
  - Beyond linear regression: conjugate-Bayes [1]
  - Beyond conjugacy [2]
  - For large models (GPT-2, ImageNet) [3]
- Model perturbation: LLM model merging [4-5]
  - Federated learning [6] and connections to duality

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

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

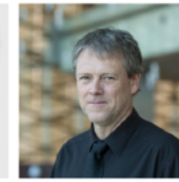

Ehsan Amid
Google DeepMind, US


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

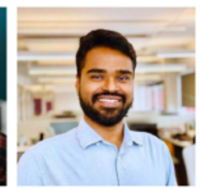

Martin Mundt
TU Darmstadt, Germany

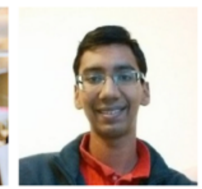

Matt Jones
University of Colorado, US


Nico Daheim
TU Darmstadt, Germany


Razvan Pascanu
Google DeepMind, US

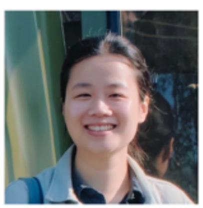

Rupam Mahmood
University of Alberta, Canada


Sarath Chandar
École Polytechnique de Montréal, Canada


Siddharth Swaroop
Harvard University, US


Tom Rainforth
University of Oxford, UK


Vincent Fortuin
Helmholtz AI, Germany


Yingzhen Li
Imperial College London, UK


Zelda Mariet
Bioptimus, US

Every June in Tokyo (June 25-27, 2025)

Attendees are from a diverse research interests: Bayes, Duality, Continual/Federated/Active learning, RL, Experiment Design 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