Deep Learning with Bayesian Principles

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

With a significant help from

Roman Bachmann (RIKEN-AIP)

Xiangming Meng (RIKEN-AIP)

The Goal of My Research

"To understand the fundamental principles of learning from data and use them to develop algorithms that can learn like living beings."

Human Learning at the age of 6 months.

Converged at the age of 12 months

Transfer skills

at the age of 14 months

Bayesian
~~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!

Parisi, German I., et al. "Continual lifelong learning with neural networks: A review." *Neural Networks* (2019)
Friston, K. "The free-energy principle: a unified brain theory?." *Nature reviews neuroscience* (2010)
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 learning

Bayesian models

(GPs, BayesNets, PGMs,)

Bayesian inference

(Bayes rule)

Deep learning

Deep models

(MLP, CNN, RNN etc.)

Stochastic training

(SGD, RMSprop, Adam)

	Bayes	DL
Can handle large data and complex models?	✗	✓
Scalable training?	✗	✓
Can estimate uncertainty?	✓	✗
Can perform sequential / active /online / incremental learning?	✓	✗

Bringing the two together

To combine their complimentary strengths to solve challenging learning problems

Deep Learning with Bayesian Principles

- Bayesian principles as a general principle
  - To design/improve/generalize learning-algorithms
  - By computing “posterior approximations”
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Everything with one common principle.

Is this different from Bayesian Deep Learning?

Scope of the Tutorial

- Audience: Deep learners and Bayesians
- Goal: To bring the two together
- This tutorial is not about
  - Bayesian deep-learning methods
  - Classical Bayesian inference methods
  - Approximate Bayesian Inference
  - Uncertainty estimation
  - Generative Models, VAE, etc.
  - Gaussian processes and NN architectures

Disclaimer

- I might not have time to discuss many important/relevant works
  - If you think I should have included some of those, please send me email and I will try to include it the next time
- The content of the tutorial is based on my own biased opinion (and expertise)
  - A lot of it is based on my own work (28 slides out of 62)

Deep Learning
vs
Bayesian Learning

Deep Learning (DL)

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

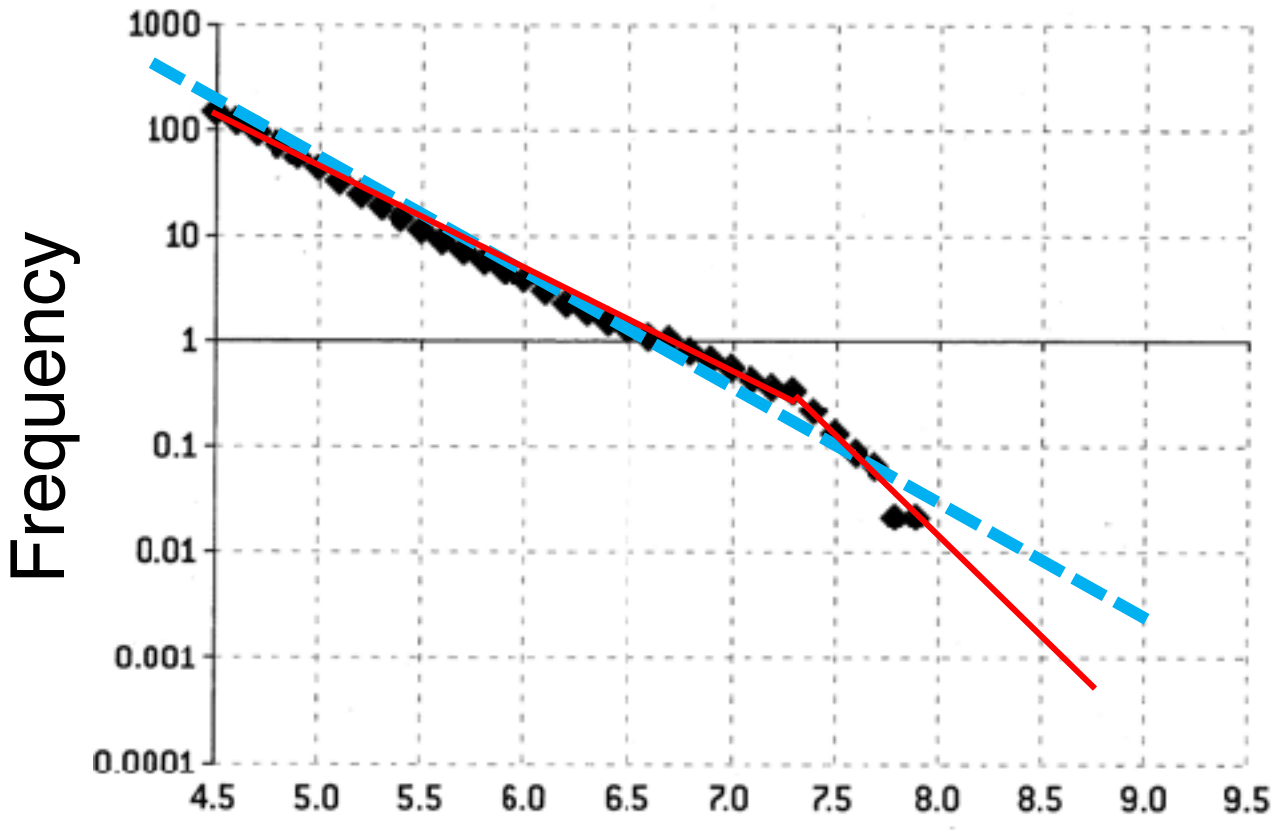
Model Params

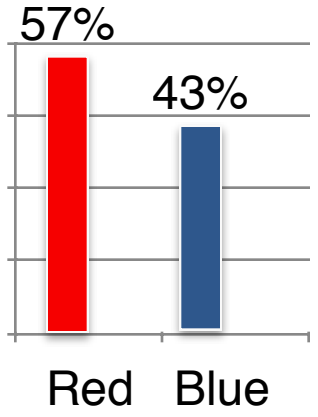
Deep Network

DL Algorithm: $\theta \leftarrow \theta - \rho H_\theta^{-1} \nabla_\theta \ell(\theta)$

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

Example: Which is a Better Fit?

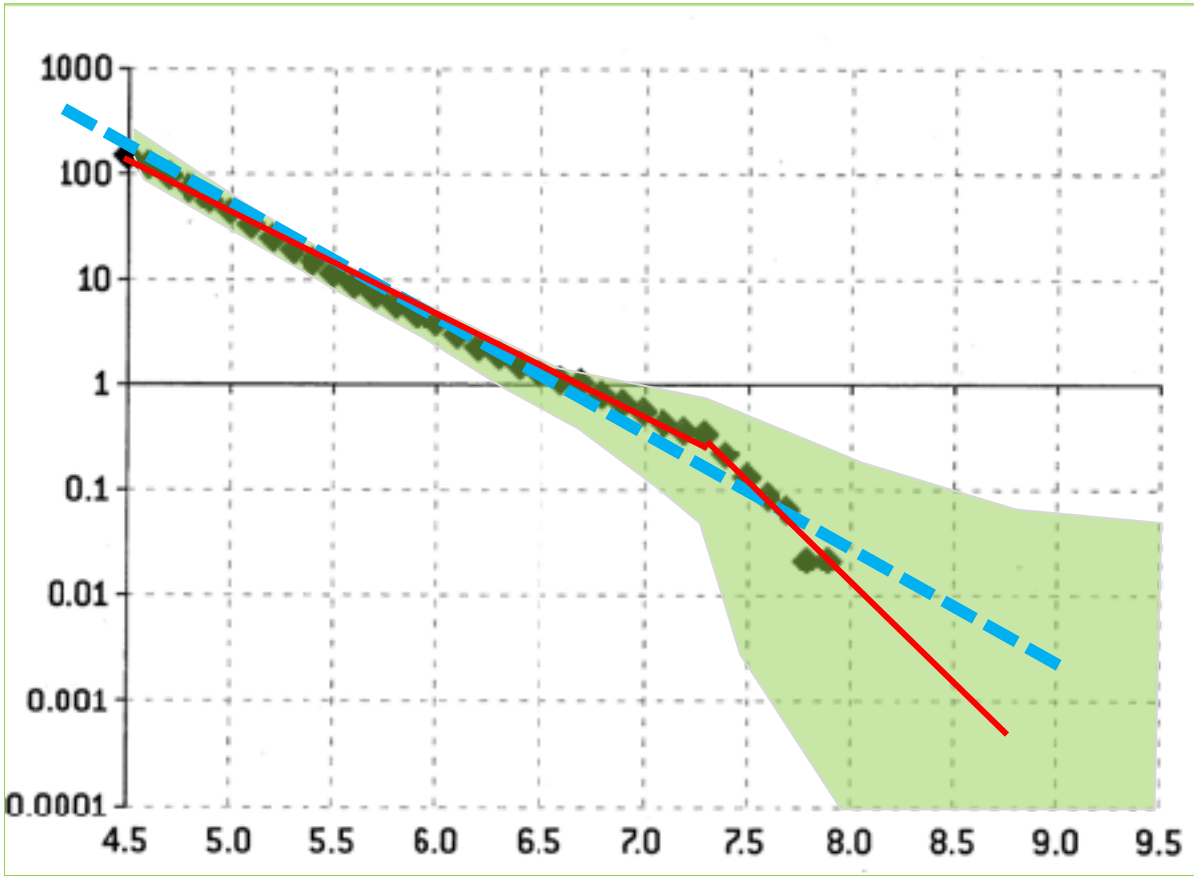




Red is more risky than the blue

Real data from Tohoku (Japan). Example taken from Nate Silver's book "The signal and noise"

Example: Which is a Better Fit?



Uncertainty: "What the model does not know"

Choose less risky options!

Avoid data bias with uncertainty!

Real data from Tohoku (Japan). Example taken from Nate Silver's book "The signal and noise"

Bayesian Principles

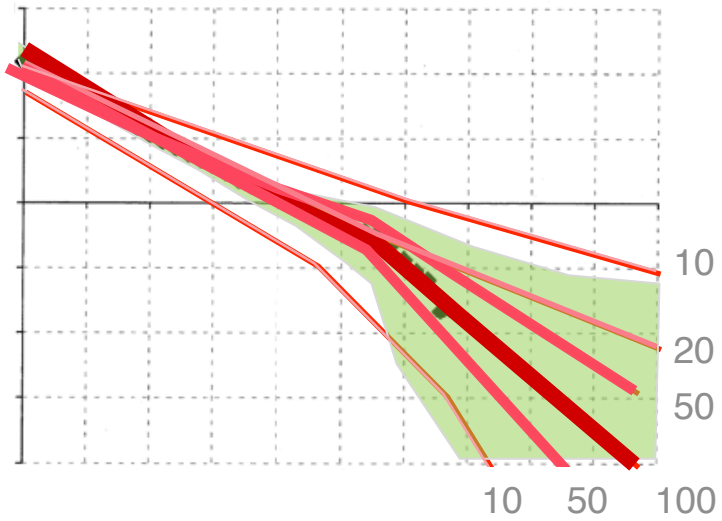
1. Sample $\theta \sim p(\theta)$ prior

2. Score $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y_i|f_\theta(x_i))$ Likelihood

3. Normalize

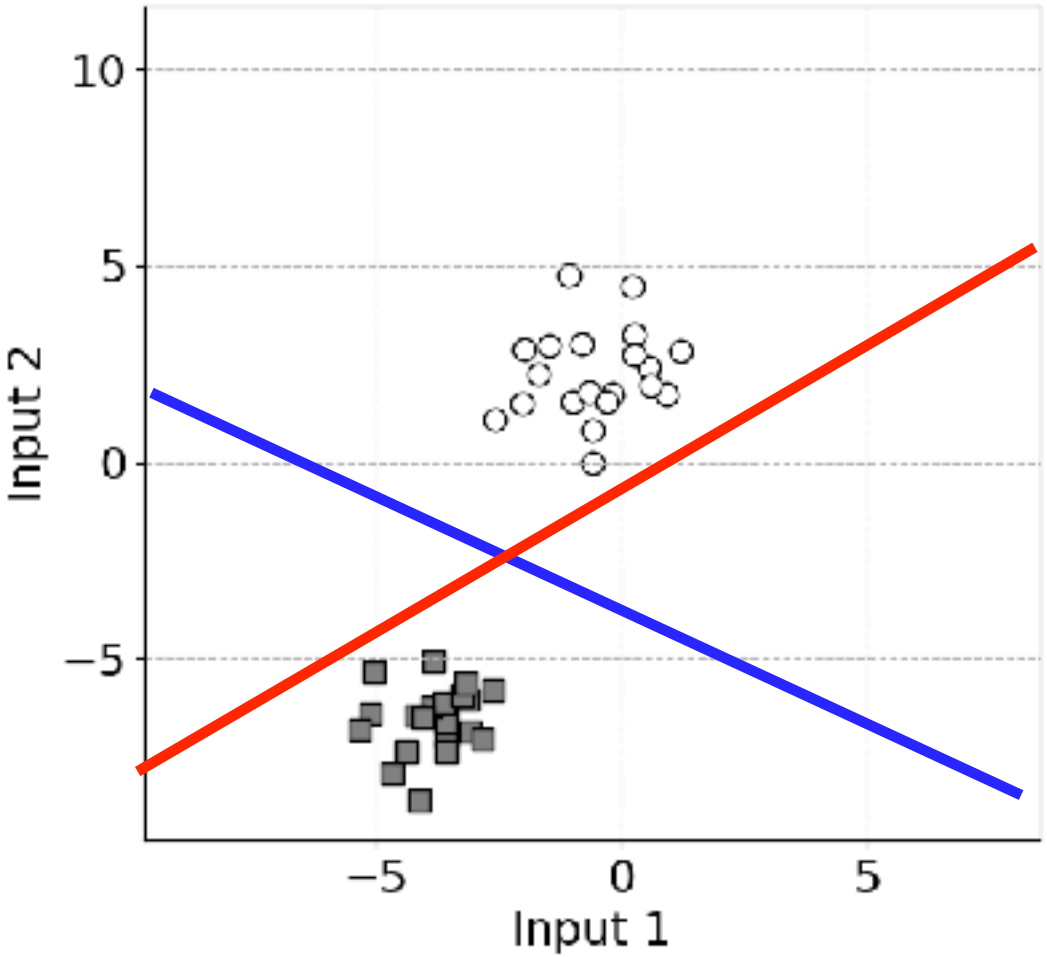
Posterior Likelihood x Prior

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

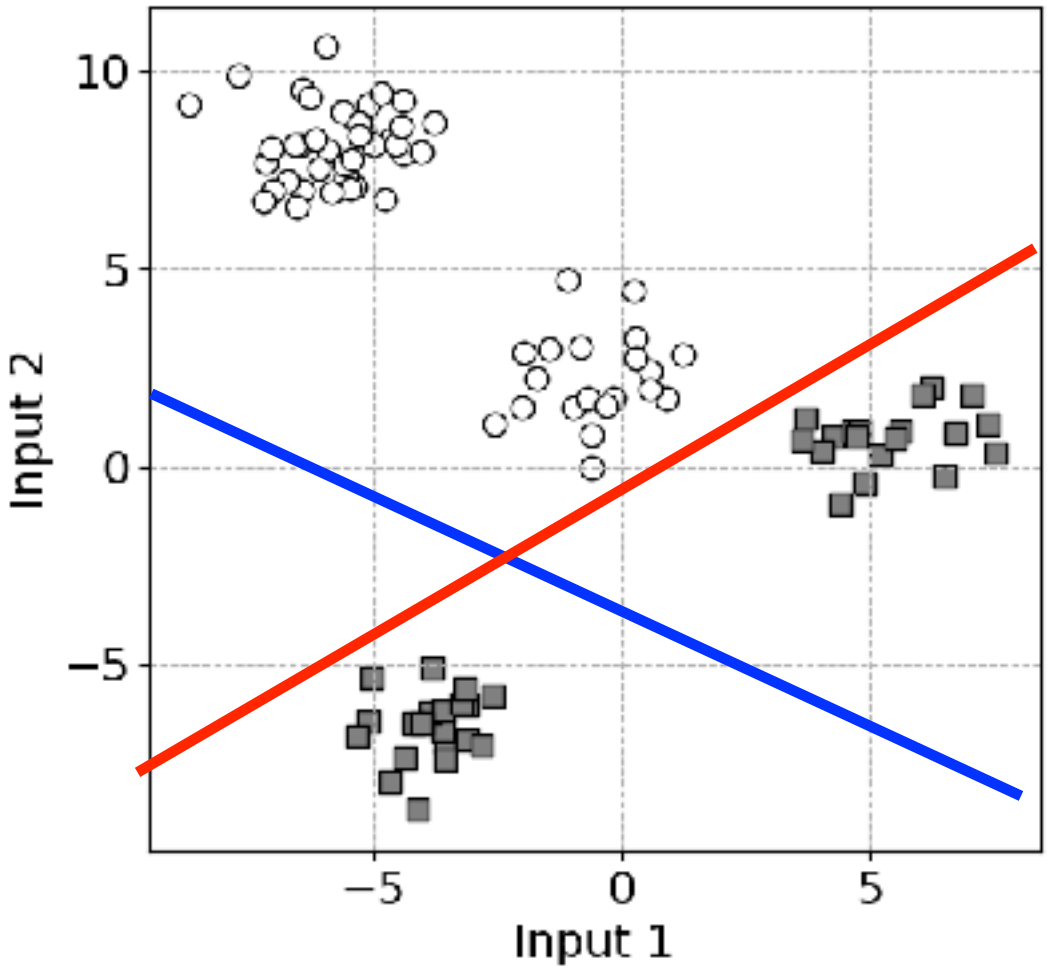


A global method: Integrates over all models
Does not scale to large problem

Which is a good classifier?

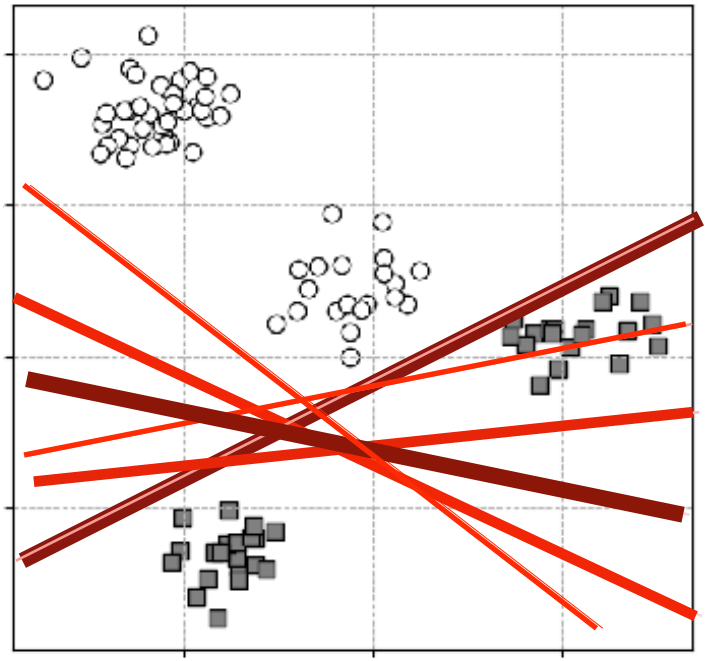


Which is a good classifier?



“What the model does not know”

Sequential Bayesian Inference



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

Set the prior to the previous posterior and recompute:

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

The global property enables sequential update

Bayesian learning

Integration (global)

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

Deep learning

Differentiation (local)

$$\theta \leftarrow \theta - \rho H_\theta^{-1} \nabla_\theta \ell(\theta)$$

	Bayes	DL
Can handle large data and complex models?	✗	✓
Scalable training?	✗	✓
Can estimate uncertainty?	✓	✗
Can perform sequential / active /online / incremental learning?	✓	✗

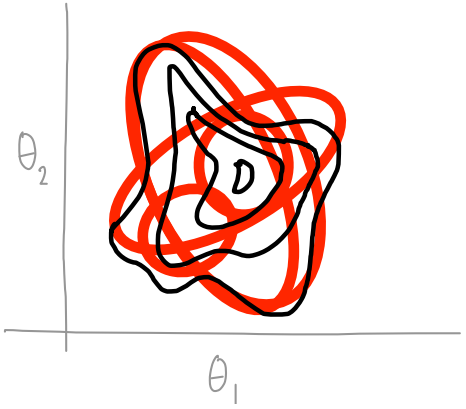
Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

Bayesian principles to derive Learning-Algorithms

Main ideas: Introduce “posterior approximations” and the “Bayesian learning rule” to estimate them

Complex ⟷ Simple



Bayes’ rule

Ensemble

Newton

Gradient Descent

Exponential Family Approximations

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	$\mu := \{\mathbb{E}_q(\theta), \mathbb{E}_q(\theta\theta^\top)\}$

Bayesian Learning Rule

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

Entropy

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 and Expectation parameters of an exponential family distribution q

Deep Learning algorithms can be obtained by

1. Choosing an appropriate approximation q,
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).

Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

Gradient Descent from Bayes

Gradient descent: $\theta \leftarrow \theta - \rho \nabla_\theta \ell(\theta)$

Bayes Learn Rule: $m \leftarrow m - \rho \nabla_m \ell(m)$

"Global" to "local"
$\mathbb{E}_q[\ell(\theta)] \approx \ell(m)$

$$m \leftarrow m - \rho \nabla_m \mathbb{E}_q[\ell(\theta)]$$

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

Derived by choosing Gaussian with fixed covariance

Gaussian distribution	$q(\theta) := \mathcal{N}(m, 1)$
Natural parameters	$\lambda := m$
Expectation parameters	$\mu := \mathbb{E}_q[\theta] = m$
Entropy	$\mathcal{H}(q) := \log(2\pi)/2$

Using stochastic gradients, we get SGD

1. Khan and Rue. "Learning-Algorithms from Bayesian Principles" (2019) (work in progress, an early draft available at https://emtiyaz.github.io/papers/learning_from_bayes.pdf)

Newton's Method from Bayes

Newton's method: $\theta \leftarrow \theta - H_\theta^{-1}\left[\nabla_\theta \ell(\theta)\right]$

$$Sm \leftarrow (1-\rho)Sm - \rho \nabla_{\mathbb{E}_q(\theta)} \mathbb{E}_q[\ell(\theta)]$$

$$-\frac{1}{2}S \leftarrow (1-\rho)\left(-\frac{1}{2}S\right) - \rho \nabla_{\mathbb{E}_q(\theta\theta^\top)} \mathbb{E}_q[\ell(\theta)]$$

$$\lambda \leftarrow (1-\rho)\lambda - \rho \nabla_\mu \mathbb{E}_q[\ell(\theta)] \qquad -\nabla_\mu \mathcal{H}(q) = \lambda$$

Derived by choosing a multivariate Gaussian

Gaussian distribution	$q(\theta) := \mathcal{N}(\theta \mid m, S^{-1})$
Natural parameters	$\lambda := \{Sm, -S/2\}$
Expectation parameters	$\mu := \{\mathbb{E}_q(\theta), \mathbb{E}_q(\theta\theta^\top)\}$

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).

Newton's Method from Bayes

Newton's method: $\theta \leftarrow \theta - H_\theta^{-1}\left[\nabla_\theta \ell(\theta)\right]$

Set ρ=1 to get $\quad m \leftarrow m - H_m^{-1}\left[\nabla_m \ell(m)\right]$

$$m \leftarrow m - \rho S^{-1} \nabla_m \ell(m)$$
$$S \leftarrow (1-\rho) S + \rho H_m$$

"Global" to "local"
$\mathbb{E}_q[\ell(\theta)] \approx \ell(m)$

Express in terms of gradient and Hessian of loss:

$$\nabla_{\mathbb{E}_q(\theta)} \mathbb{E}_q[\ell(\theta)] = \mathbb{E}_q[\nabla_\theta \ell(\theta)] - 2\mathbb{E}_q[H_\theta] m$$

$$\nabla_{\mathbb{E}_q(\theta\theta^\top)} \mathbb{E}_q[\ell(\theta)] = \mathbb{E}_q[H_\theta]$$

$$Sm \leftarrow (1-\rho) Sm - \rho \nabla_{\mathbb{E}_q(\theta)} \mathbb{E}_q[\ell(\theta)]$$
$$S \leftarrow (1-\rho) S - \rho 2 \nabla_{\mathbb{E}_q(\theta\theta^\top)} \mathbb{E}_q[\ell(\theta)]$$

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).

RMSprop/Adam from Bayes

RMSprop

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

Bayesian Learning rule for multivariate Gaussian

$$S \leftarrow (1-\rho)S + \rho(H_\theta)$$
$$m \leftarrow m - \alpha S^{-1}\nabla_\theta \ell(\theta)$$

To get RMSprop, make the following choices

- Choose Gaussian with diagonal covariance
- Replace Hessian by square of gradients
- Add square root for scaling vector

For Adam, use a Heavy-ball term with KL divergence as momentum (Appendix E in [1])

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).

Summary

- Gradient descent is derived using a Gaussian with fixed covariance, and estimating the mean
- Newton's method is derived using multivariate Gaussian
- RMSprop is derived using diagonal covariance
- Adam is derived by adding heavy-ball momentum term
- For "ensemble of Newton", use Mixture of Gaussians [1]
- To derive DL algorithms, we need to switch from a "global" to "local" approximation $\mathbb{E}_q[\ell(\theta)] \approx \ell(m)$
- Then, to improve DL algorithms, we just need to add some "global" touch to the DL algorithms

1. Lin, Wu, Mohammad Emtiyaz Khan, and Mark Schmidt. "Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations." *ICML* (2019).

Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

Bayes as Optimization

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

$$\ell(\theta) := -\log p(\mathcal{D}|\theta)p(\theta)$$

$$= \arg\min_{q \in \mathcal{P}} \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{H}(q)$$

All distribution

Distribution

Entropy

$$= \mathbb{E}_q[\ell(\theta)] + \mathbb{E}_q[\log q(\theta)] \quad = \mathbb{E}_q\left[\log \frac{q(\theta)}{e^{-\ell(\theta)}}\right]$$

$$\implies q_*(\theta) \propto e^{-\ell(\theta)} \propto p(\mathcal{D}|\theta)p(\theta) \propto p(\theta|\mathcal{D})$$

Good news: This holds for a generic loss function!

Zellner (1988), Bissiri, et al. (2016), Shawe-Taylor and Williamson (1997), Cesa-Bianchi and Lugosi (2006)

Bayes with Approximate Posterior

$$\arg\min_{q \in \mathcal{P}} \; \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{H}(q)$$

All distribution — Distribution — Entropy

Restrict the set of distribution from P to Q

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

This is known as Variational Inference, but along with the Bayesian learning rule, it enables us to derive many more algorithms (including Bayes' rule). So this is not just a method, but a principle.

Conjugate Bayesian Inference from Bayesian Principles

Ex: Linear model, Kalman filters, HMM, etc.

$$\ell(\theta) := -\log p(\mathcal{D}|\theta)p(\theta) = -\lambda_{\mathcal{D}}^\top T(\theta)$$

← Sufficient statistics of q

$$\begin{aligned}\ell(\theta) &:= (y - X\theta)^\top (y - X\theta) + \gamma\theta^\top\theta \\ &= -2\theta^\top (X^\top y) + \mathrm{Tr}\left[\theta\theta^\top \left(X^\top X + \gamma I\right)\right] + \mathrm{cnst}\end{aligned}$$

$$\implies \mathbb{E}_q[\ell(\theta)] = -\lambda_{\mathcal{D}}\mu \implies \nabla_\mu \mathbb{E}_q[\ell(\theta)] = -\lambda_{\mathcal{D}}$$

$$\lambda \leftarrow \lambda - \rho\left(\nabla_\mu \mathbb{E}_q[\ell(\theta)] + \lambda - \lambda_{\mathcal{D}}\right) \implies \lambda_* = \lambda_{\mathcal{D}}$$

Forward-backward, SVI, Variational message passing etc. are special cases of the same Bayesian principles

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

Laplace Approximation

Derived by choosing a multivariate Gaussian, then running the following Newton's update

$$m \leftarrow m - \rho S^{-1} \nabla_m \ell(m)$$
$$S \leftarrow (1-\rho) S + \rho H_m$$

Bayesian principles we discussed are general principles to derive learning algorithms

Calling them variational inference limits their scope!

Learning-Algorithms from Bayesian Principles

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

Given a loss, we can recover a variety of learning algorithms by choosing an appropriate q

- Classical algorithms: Least-squares, gradient descent, Newton's method, Kalman filters, Baum-Welch, Forward-backward, etc.
- Bayesian inference: EM, Laplace's method, SVI, VMP.
- Deep learning: SGD, RMSprop, Adam.
- Reinforcement learning: parameter-space exploration, natural policy-search.
- Continual learning: Elastic-weight consolidation.
- Online learning: Exponential-weight average.
- Global optimization: Natural evolutionary strategies, Gaussian homotopy, continuation method & smoothed optimization.

1. Khan and Rue. "Learning-Algorithms from Bayesian Principles" (2019) (work in progress, an early draft available at https://emtiyaz.github.io/papers/learning_from_bayes.pdf)

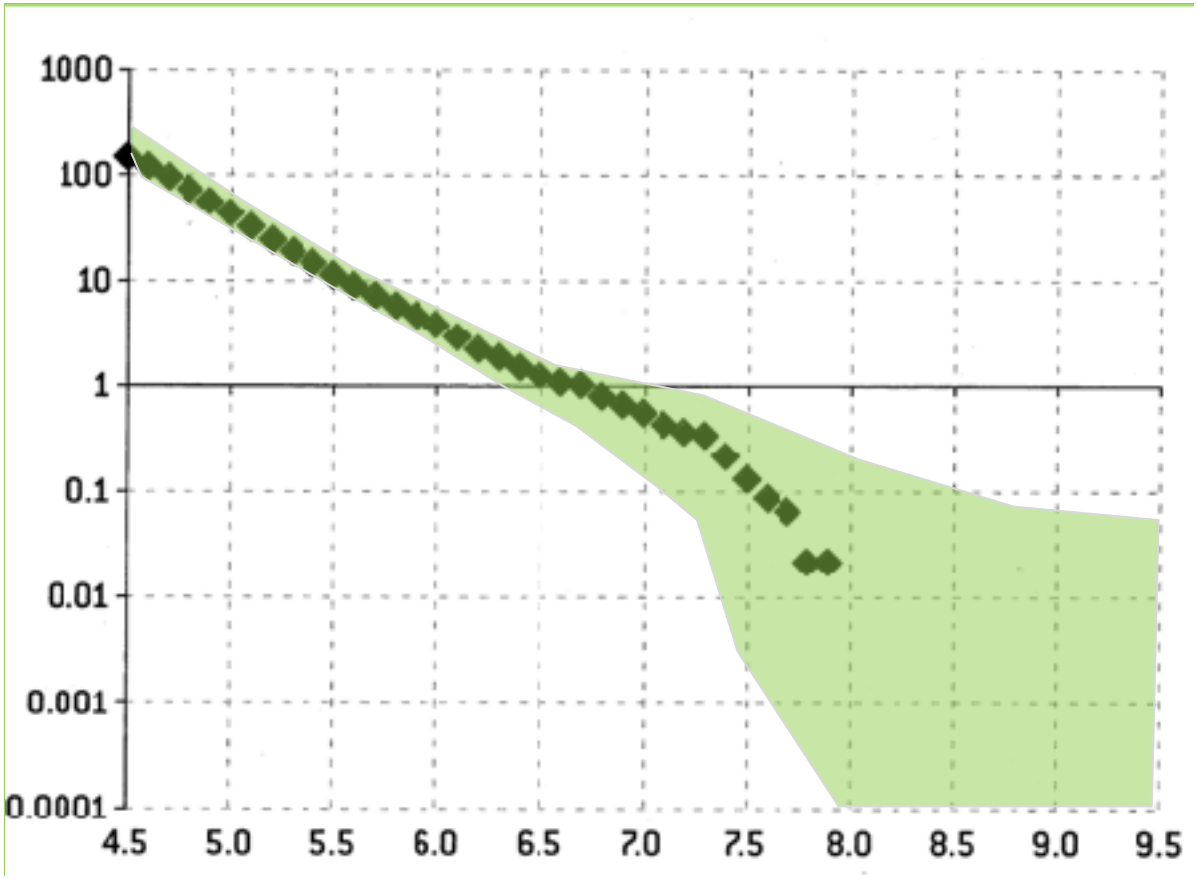
Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

Uncertainty Estimation for Deep Learning

New deep-learning algorithms

Uncertainty for Robust Decisions



Uncertainty: "What the model does not know"

Choose less risky options!

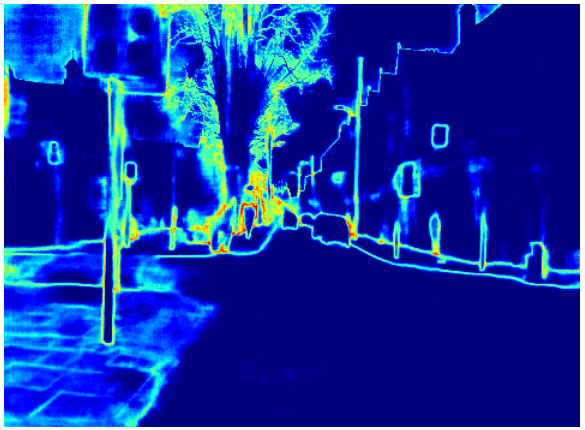
Avoid data bias with uncertainty!

Real data from Tohoku (Japan). Example taken from Nate Silver's book "The signal and noise"

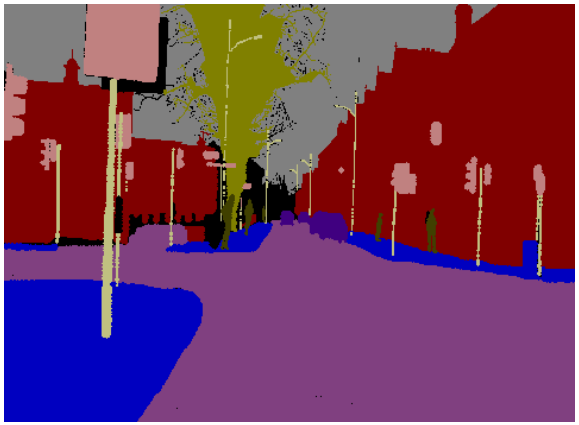
Uncertainty Estimation for Image segmentation

Image

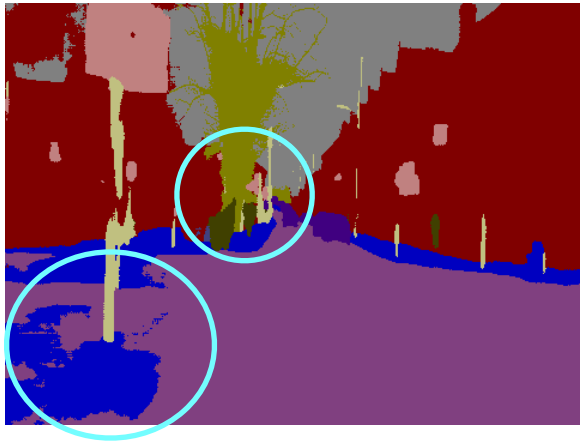
Uncertainty

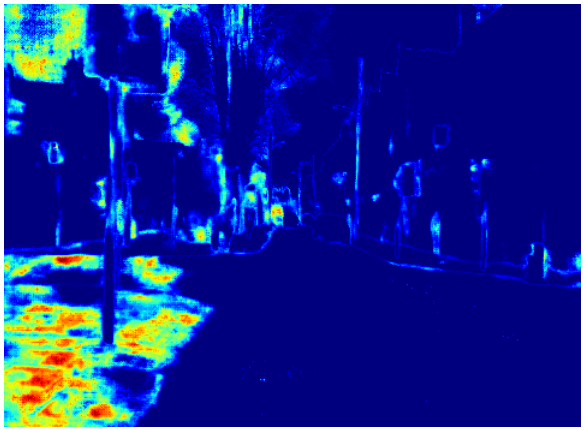


True Segments



Prediction





Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." *CVPR*. 2018.

(Some) Bayesian Deep Learning Methods

- SGD based (MC-dropout [1], SWAG [2], Laplace [3])
  - Pros: Scales well to large problems
  - Cons: Not flexible
- Variational inference methods [1, 2]

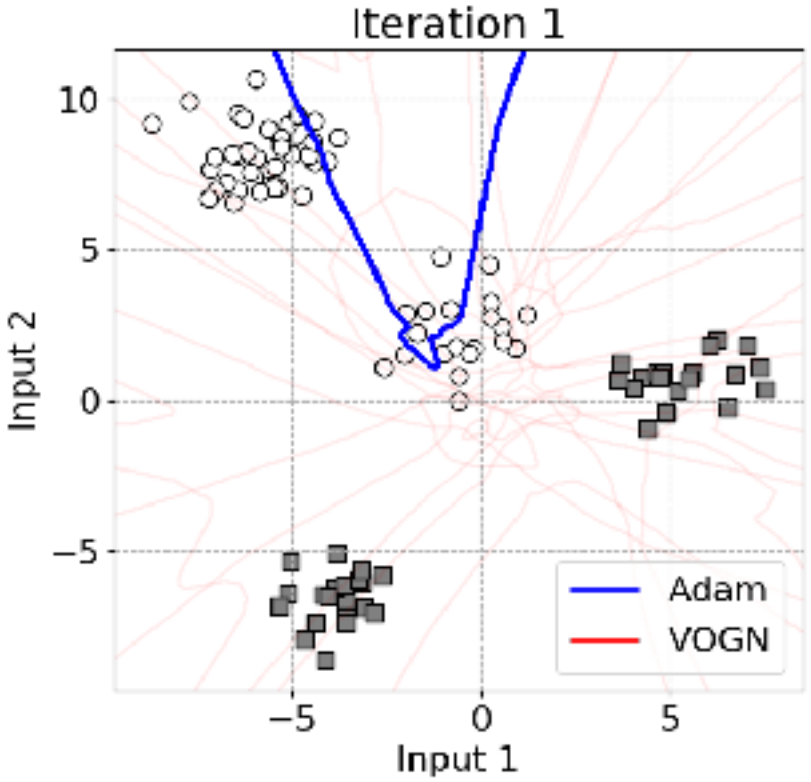
$$\lambda \leftarrow \lambda - \rho \nabla_{\lambda} \left(\mathbb{E}_q[\ell(\theta)] - \mathcal{H}(q) \right)$$

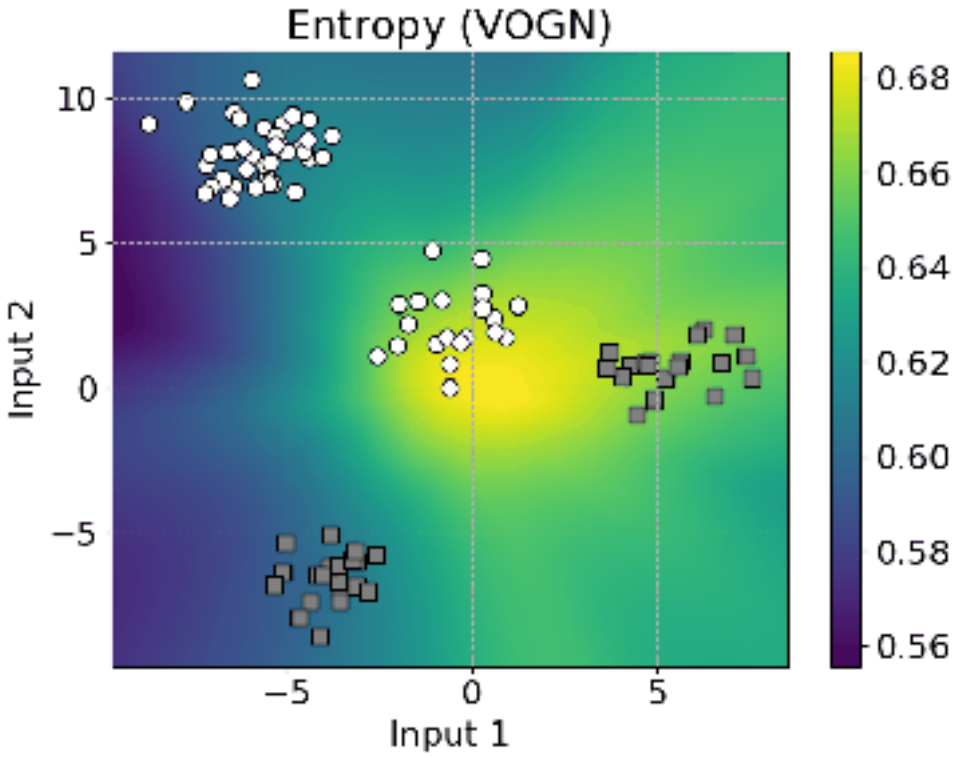
  - Pros: Enable flexible distributions
  - Cons: Do not scale to large problems (ImageNet)

1. Gal and Ghahramani. "Dropout as a bayesian approximation…" *ICML*. 2016.
2. Maddox, Wesley, et al. "A simple baseline for bayesian uncertainty in deep learning." *arXiv* (2019).
3. Ritter et al. "A scalable laplace approximation for neural networks." (2018).
4. Graves, Alex. "Practical variational inference for neural networks." *NeurIPS* (2011).
5. Blundell, Charles, et al. "Weight uncertainty in neural networks." *ICML* (2015).

Scaling up VI to ImageNet

VOGN, an Adam-like algorithm, for uncertainty





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).

Variational Online Gauss-Newton

- Improve RMSprop with the Bayesian "touch"
  - Remove the "local" approximation $\mathbb{E}_q[\ell(\theta)] \approx \ell(m)$
  - Use a second-order approximation
  - No square root of the scale
- Improve VOGN by using deep learning tricks
  - Momentum, batch norm, data augmentation etc

RMSprop

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

VOGN

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

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).

Adam to VOGN

"Adam" to "VOGN" in two lines of code change.

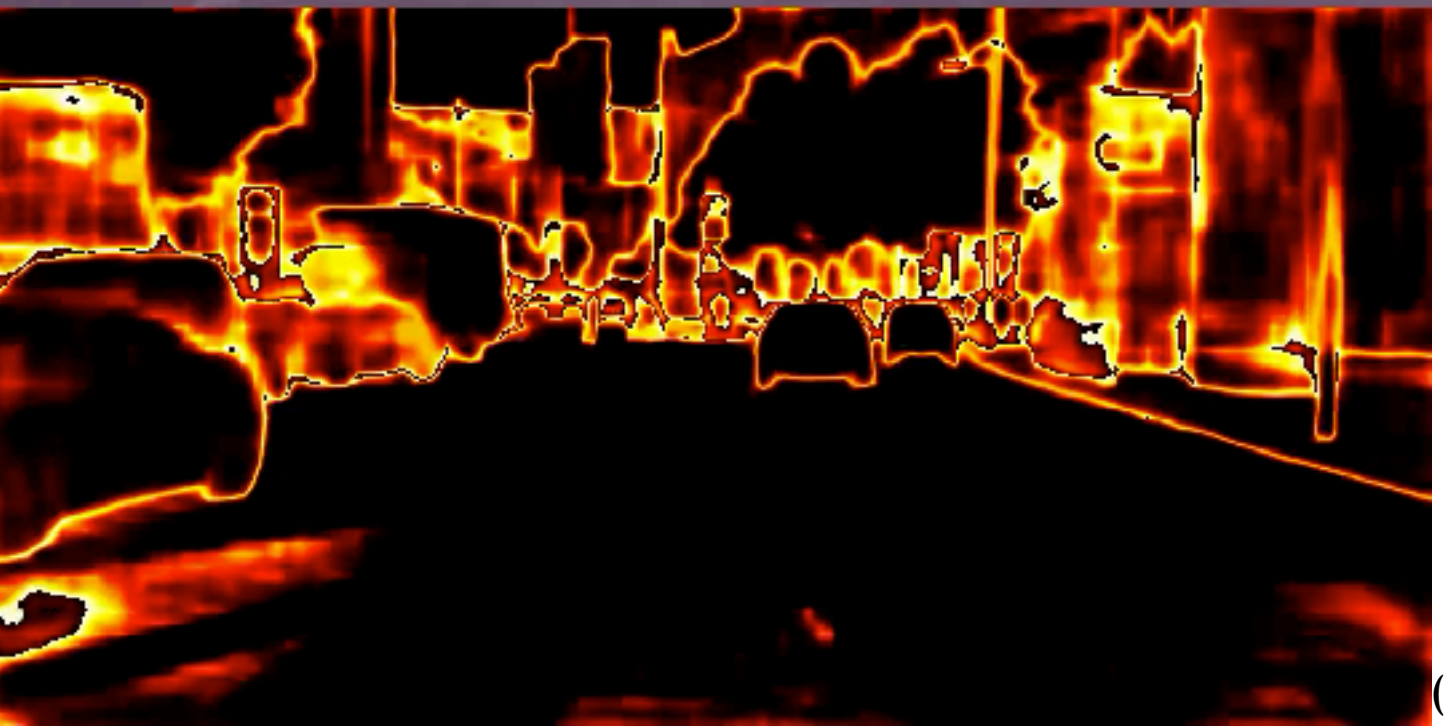
```
import torch
+import torchsso

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

-optimizer = torch.optim.Adam(model.parameters())
+optimizer = torchsso.optim.VOGN(model, dataset_size=len(train_loader.dataset))
```

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

Uses many practical tricks of DL to scale 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).

Image Segmentation

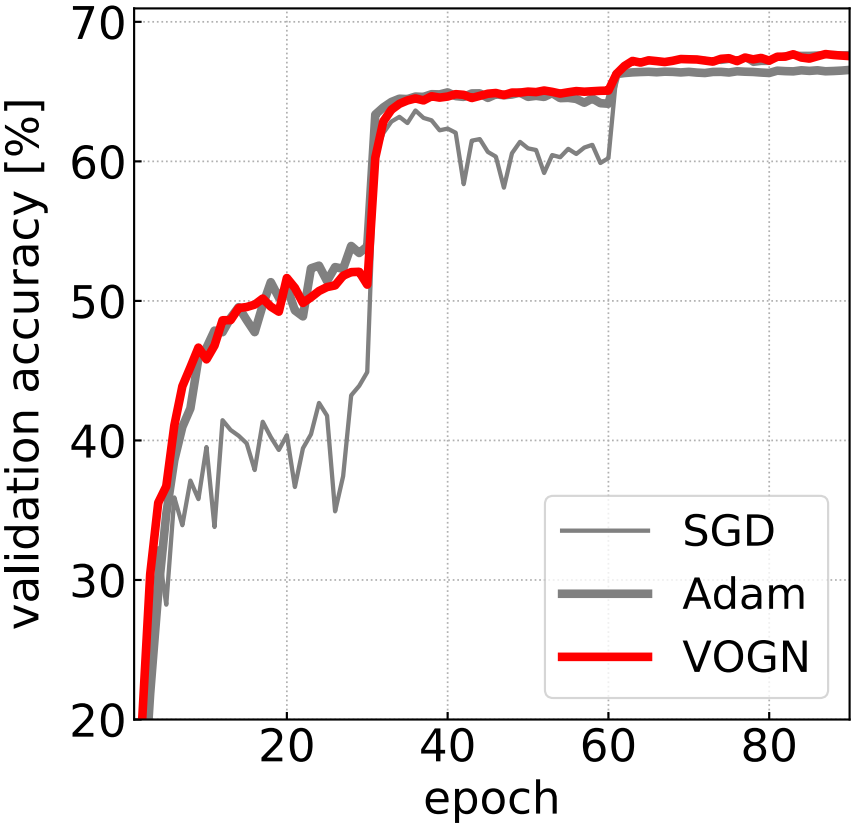


Uncertainty (entropy of class probs)

(By Roman Bachmann)

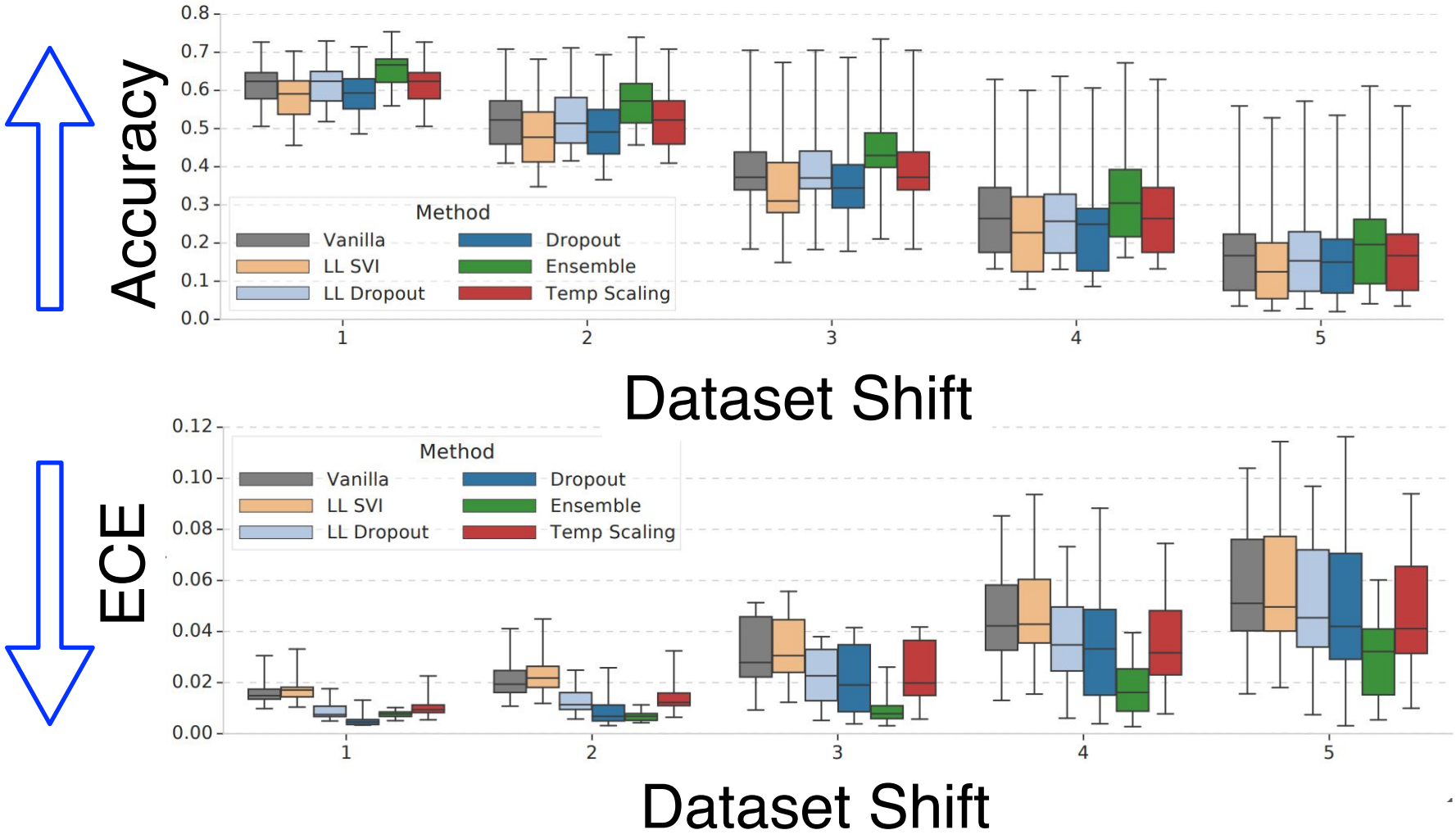
VOGN on ImageNet

State-of-the-art performance and convergence rate, while preserving benefits of Bayesian principles



1. Osawa et al. “Practical Deep Learning with Bayesian Principles.” NeurIPS (2019).

BDL methods do not really know that they are performing badly under dataset shit



1. Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." *NeurIPS* (2019).

Resources for Uncertainty in DL

- Yarin Gal's tutorial (http://bdl101.ml/)
- Benchmarks by OATML (http://bdlb.ml/)

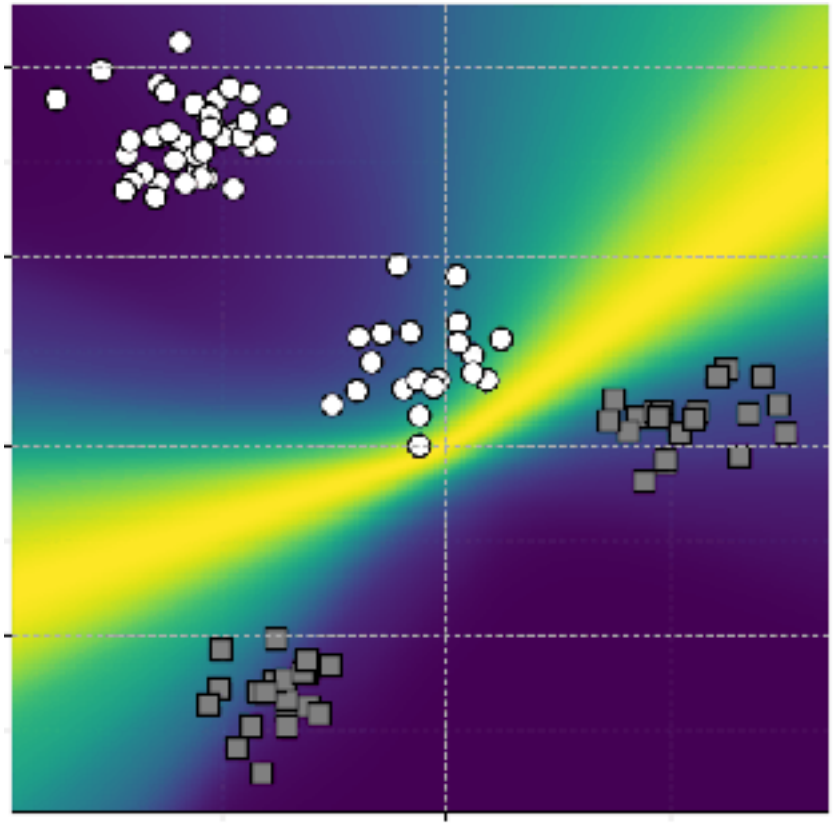
List of Benchmarks

Bayesian Deep Learning Benchmarks (BDL Benchmarks or `bdlb` for short), is an open-source framework that aims to bridge the gap between the design of deep probabilistic machine learning models and their application to real-world problems. Our currently supported benchmarks are:

- [x] Diabetic Retinopathy Diagnosis (in `alpha`, following Leibig et al.)
  - [x] Deterministic
  - [x] Monte Carlo Dropout (following Gal and Ghahramani, 2015)
  - [x] Mean-Field Variational Inference (following Peterson and Anderson, 1987, Wen et al., 2018)
  - [x] Deep Ensembles (following Lakshminarayanan et al., 2016)
  - [x] Ensemble MC Dropout (following Smith and Gal, 2018)
- [ ] Autonomous Vehicle's Scene Segmentation (in `pre-alpha`, following Mukhoti et al.)
- [ ] Galaxy Zoo (in `pre-alpha`, following Walmsley et al.)
- [ ] Fishyscapes (in `pre-alpha`, following Blum et al.)

Challenges in Uncertainty Estimation

- For non convex problem
  - Different local minima correspond to various solutions
  - Local approximations only capture “local uncertainty”
- Solutions
  - More flexible approximations?

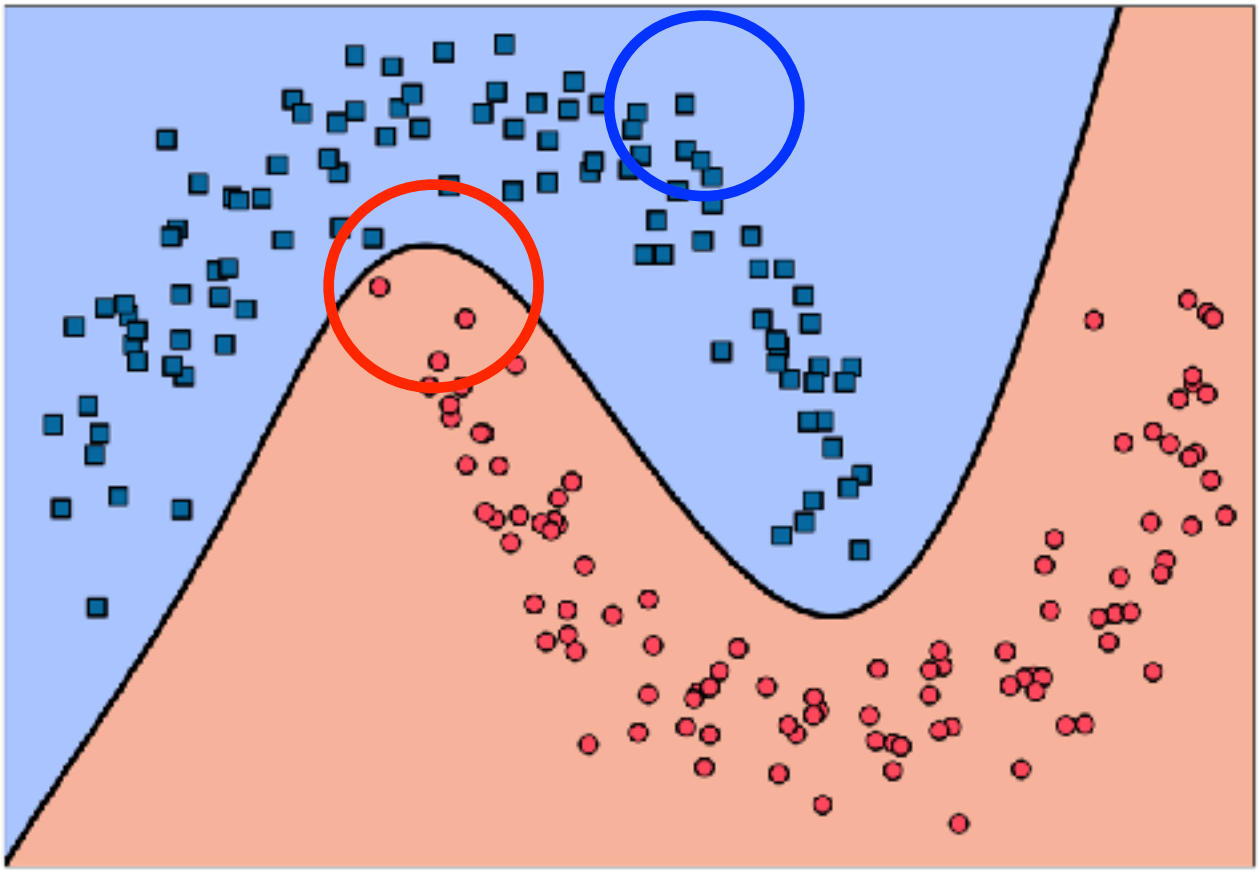


Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

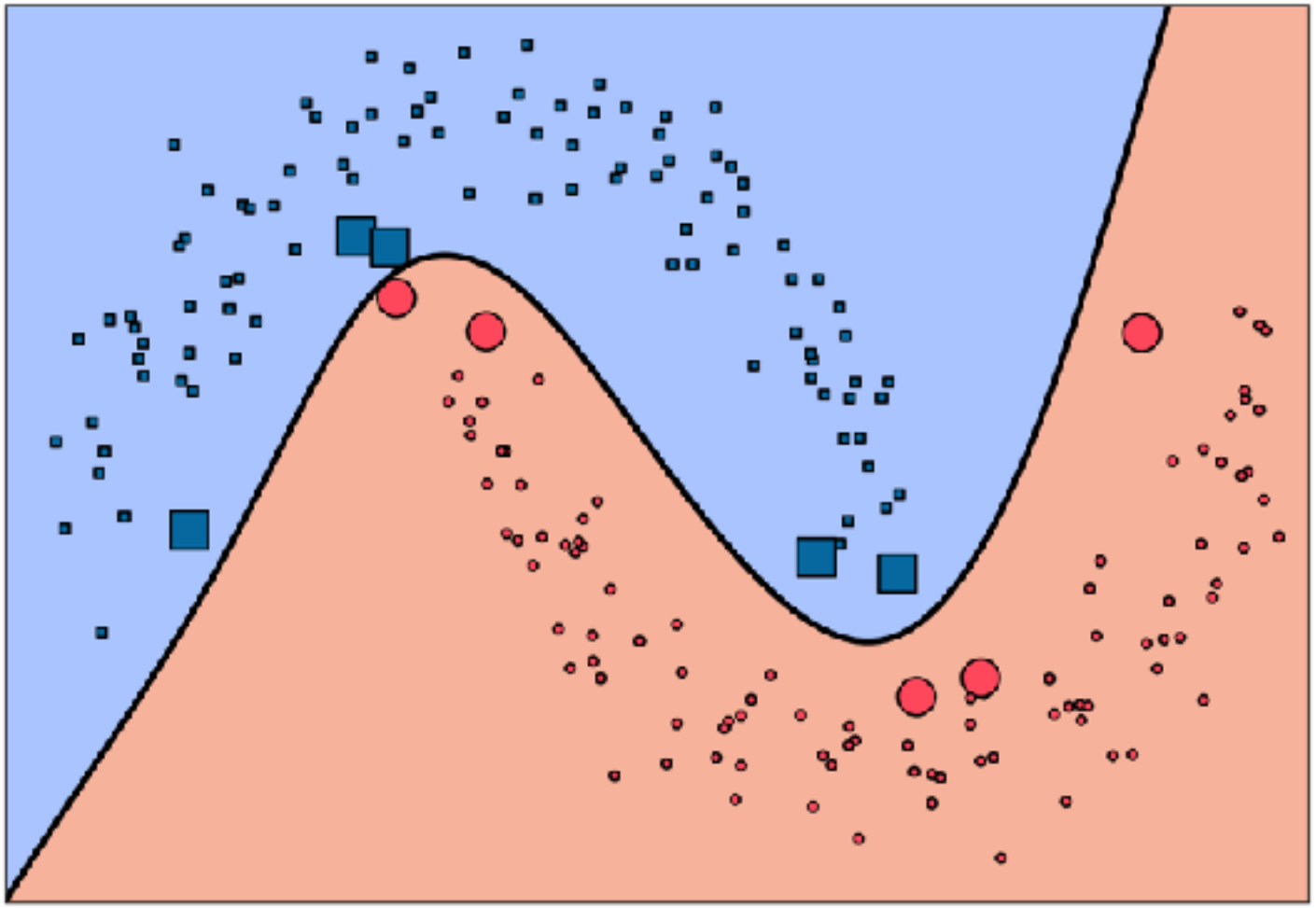
Importance of Data Examples

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



Model view vs Data view

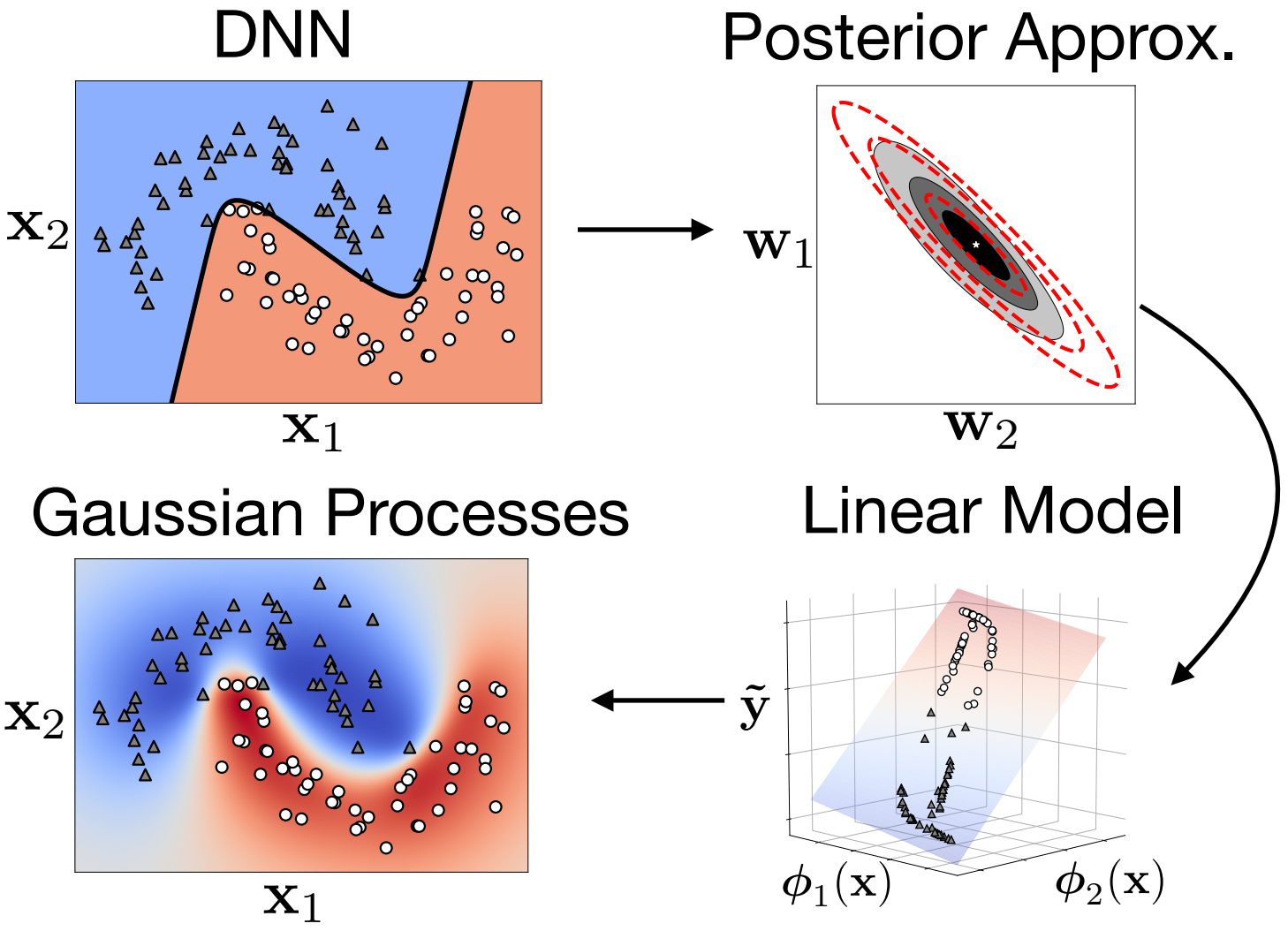
Bayes “automatically” defines data-Importance



Data view

(By Roman Bachmann)

DNN to GP



1. Khan et al., Approximate Inference Turns Deep Networks into Gaussian Processes, NeurUPS, 2019

“Global” to “Local”

Posterior approximations connect “global” parameters (e.g. DNN weights) to “local” parameters (e.g. data examples)

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

neural network

“Dual” variables

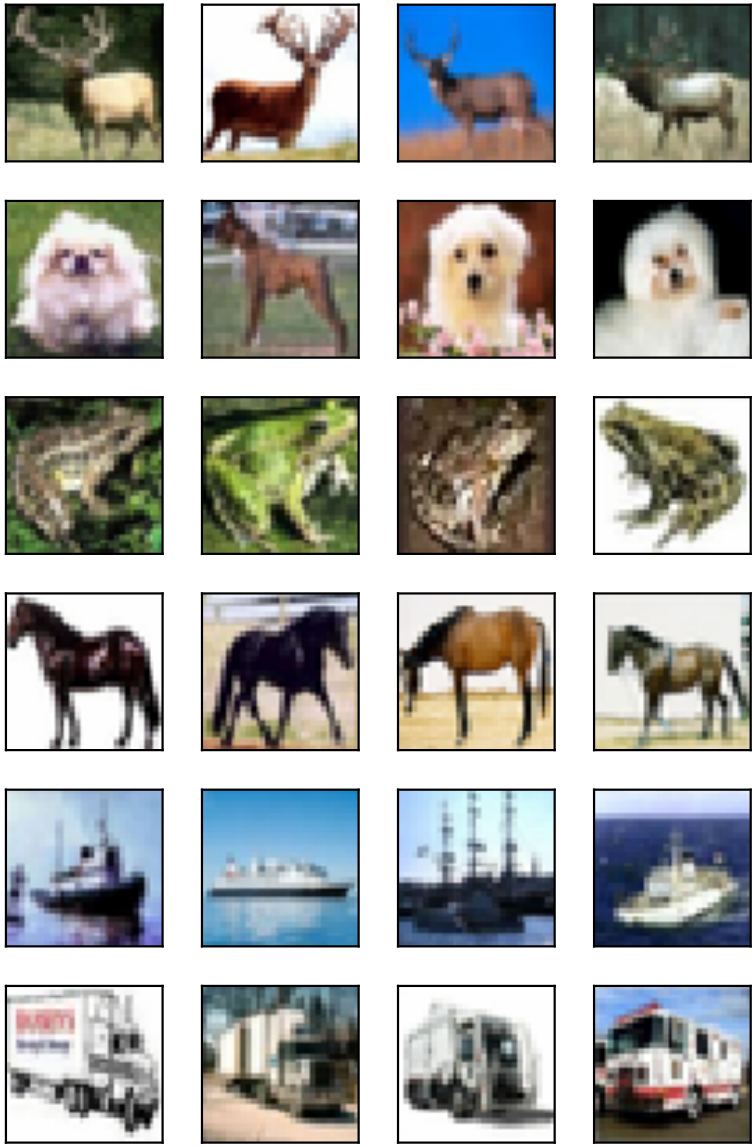
The local parameters can be seen as “dual” variables that define the “importance” of the data

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).

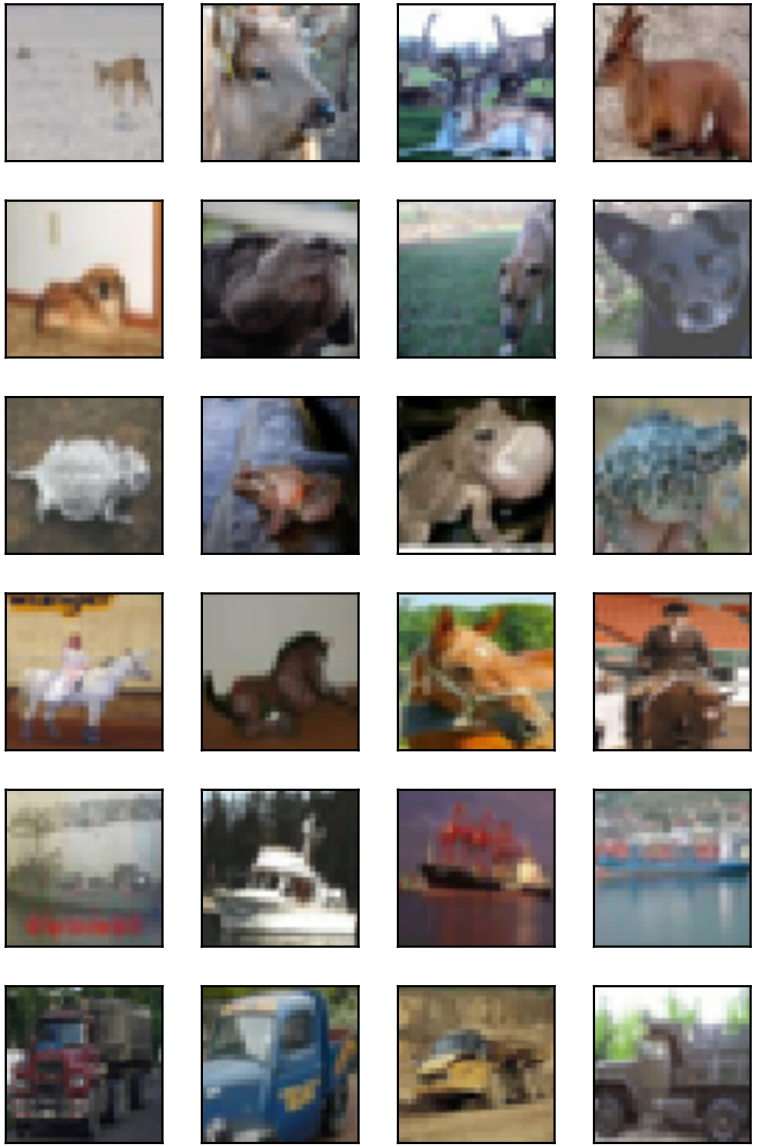
Least Important

Most Important

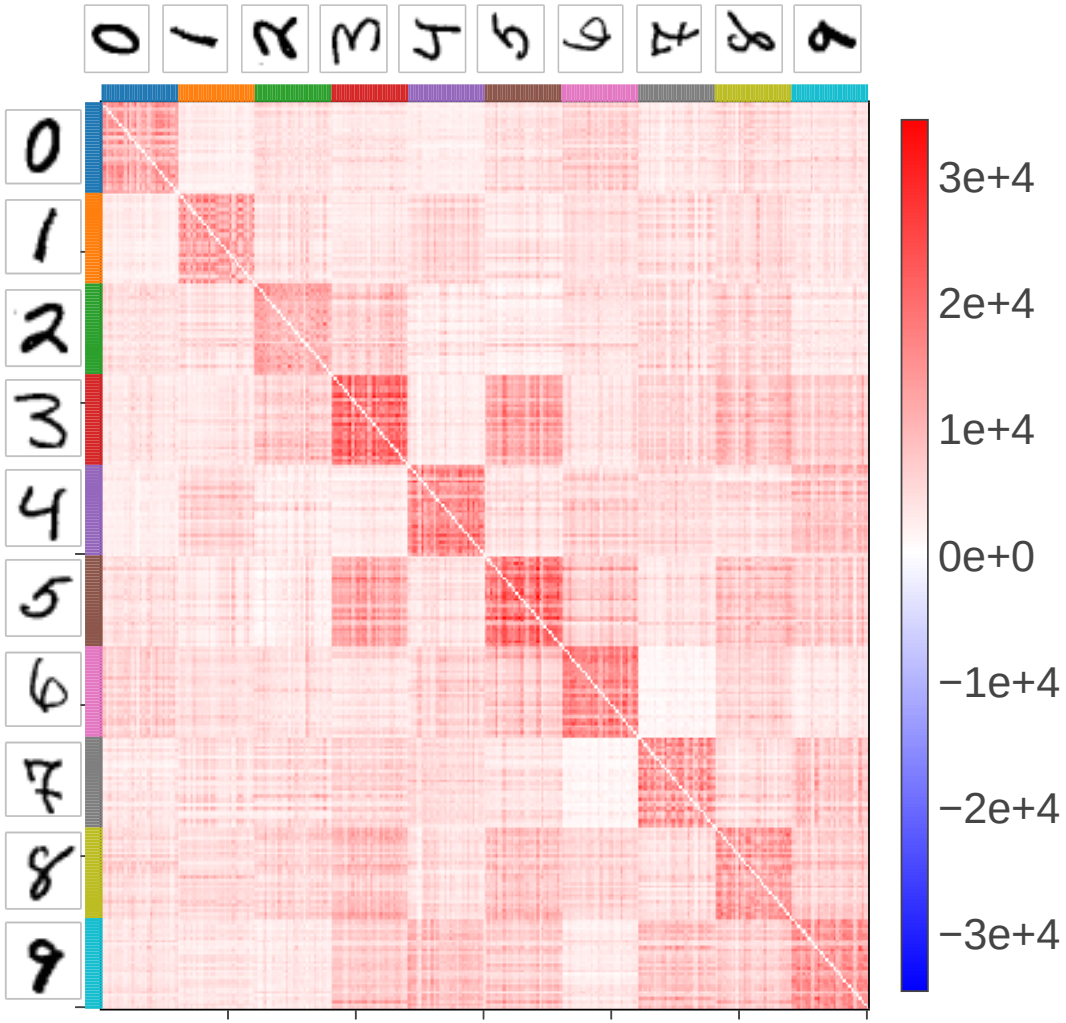
Least Important



Most Important



Similarity (Kernel) Matrix



$$K_{ij} := \phi_i^\top \phi_j$$

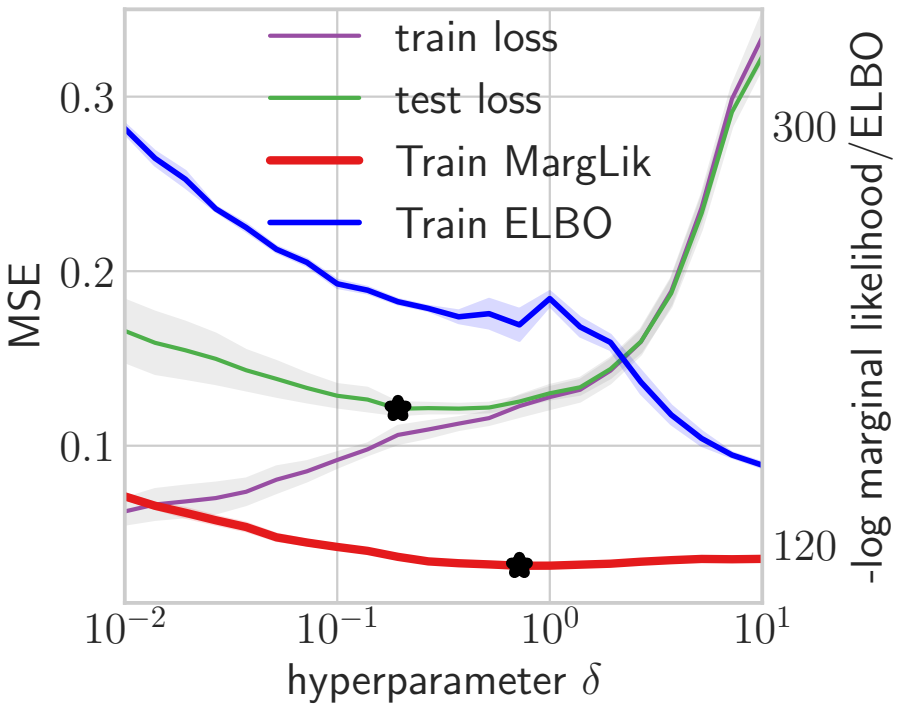
For DNN, with a specific Gaussian approximation, we obtain Neural Tangent Kernel

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

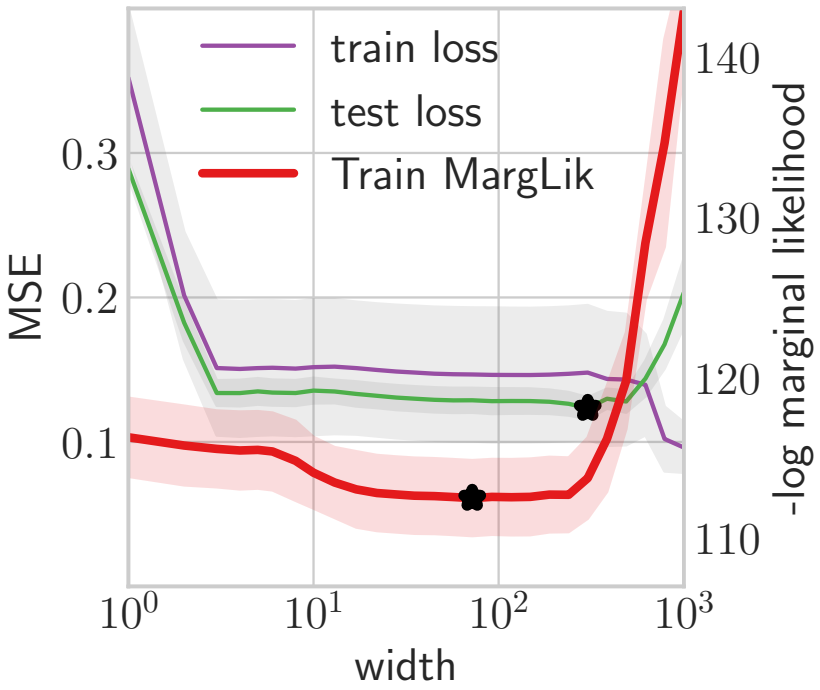
Model Selection

Tune hyper parameters with GP Marginal likelihood

Regularization parameter



NN width



Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

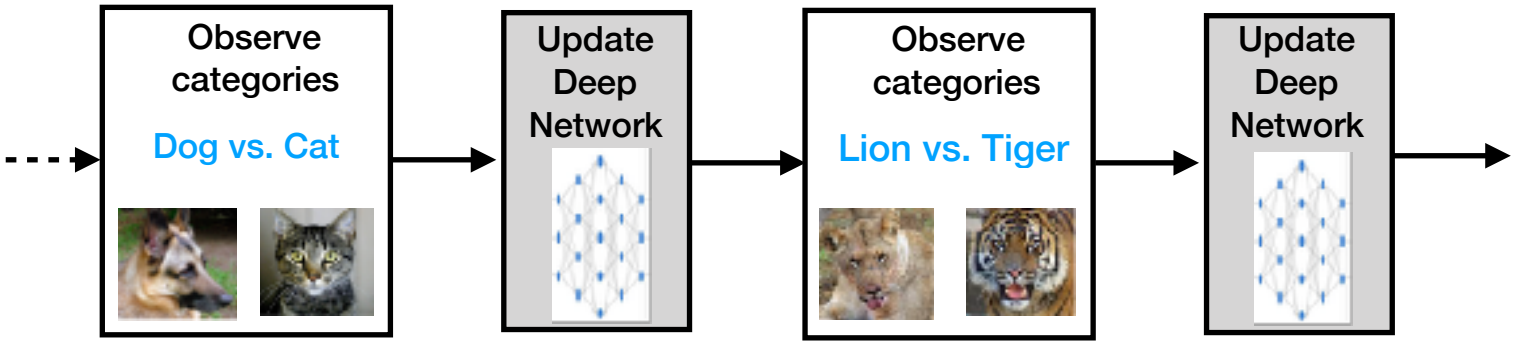
Towards Life-Long Learning

Continual and active learning
(unpublished)

Continual Learning

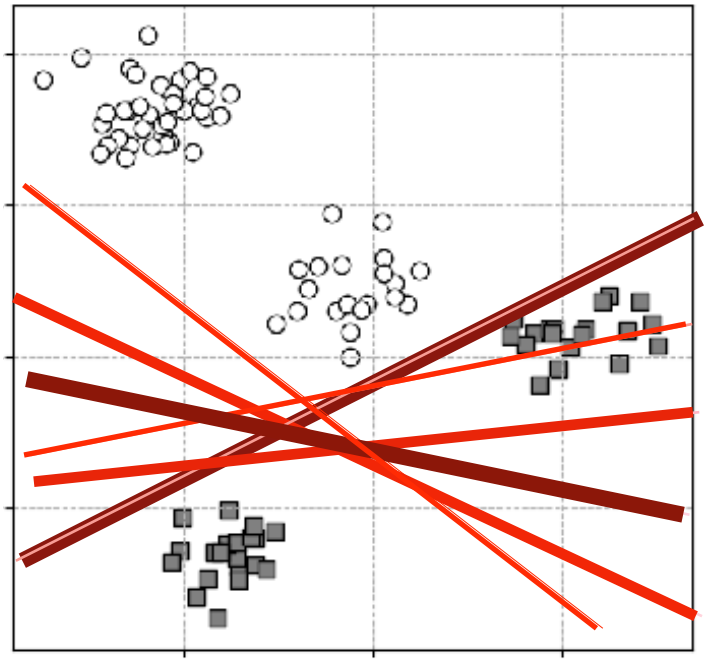
Standard Deep Learning

Continual Learning: past classes never revisited



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

Continual Learning with Bayes



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

Set the prior to the previous posterior and recompute:

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

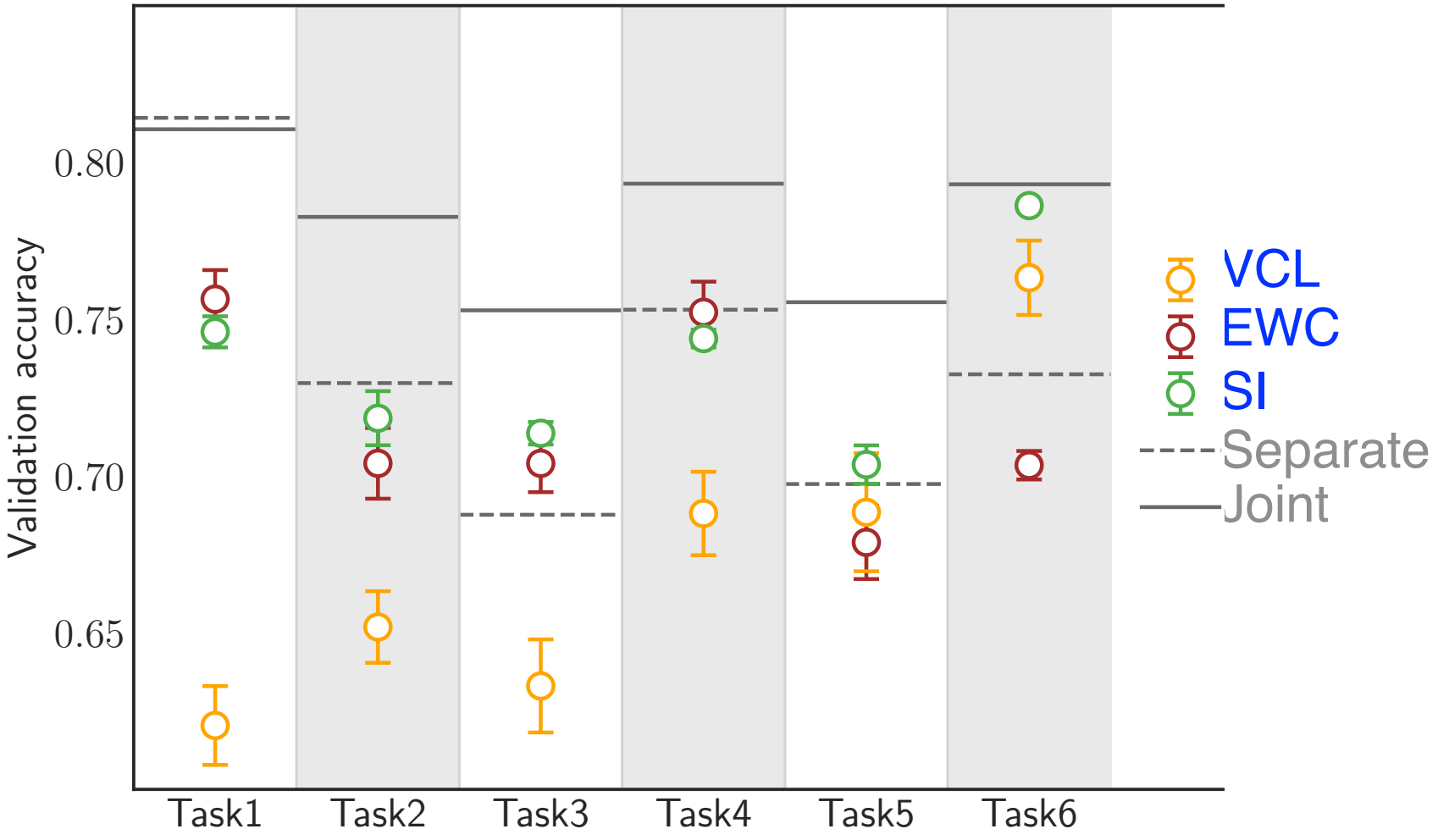
Computing posterior is challenging, so we can use posterior approximations

(Some) Regularization-based Continual Learning Methods

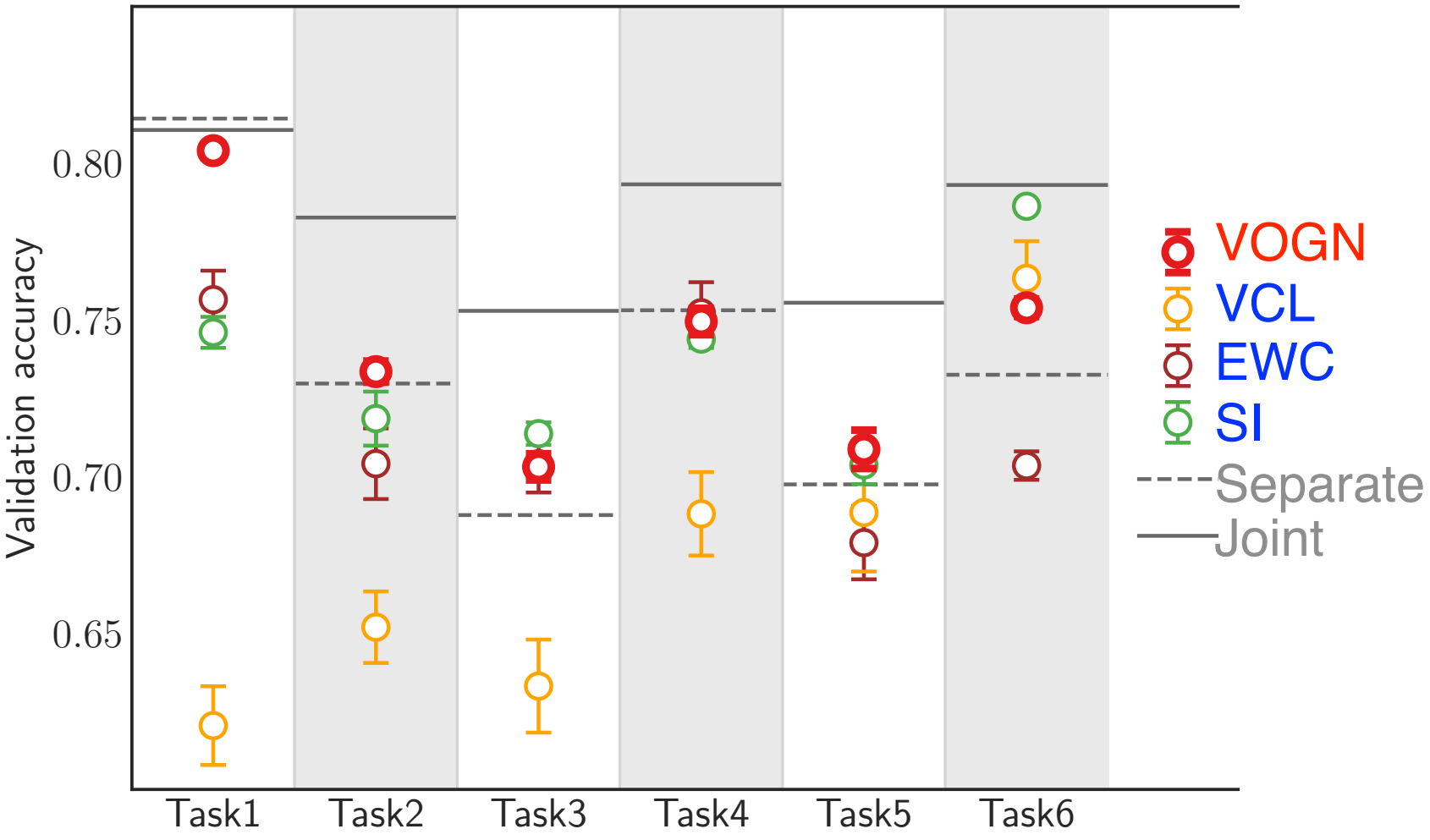
- Elastic-weight consolidation (EWC) [1]
  - Based on a diagonal Laplace approximation
  - [2] considers structured Laplace
- Synaptic Intelligence (SI) [3]
- Variational Continual learning (VCL) [4]
  - Based on variational inference
- With better approximations, we expect accuracy to improve, but unfortunately we don't see this!

1. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS* (2017).
2. Ritter et al. "Online structured laplace … for overcoming catastrophic forgetting." *NeurIPs*. 2018.
3. Zenke et al. "Continual learning through synaptic intelligence." *ICML,* 2017.
4. Nguyen, Cuong V., et al. "Variational continual learning." *arXiv preprint arXiv:1710.10628* (2017).

Principle is Broken: Better Approximation don't give better results!

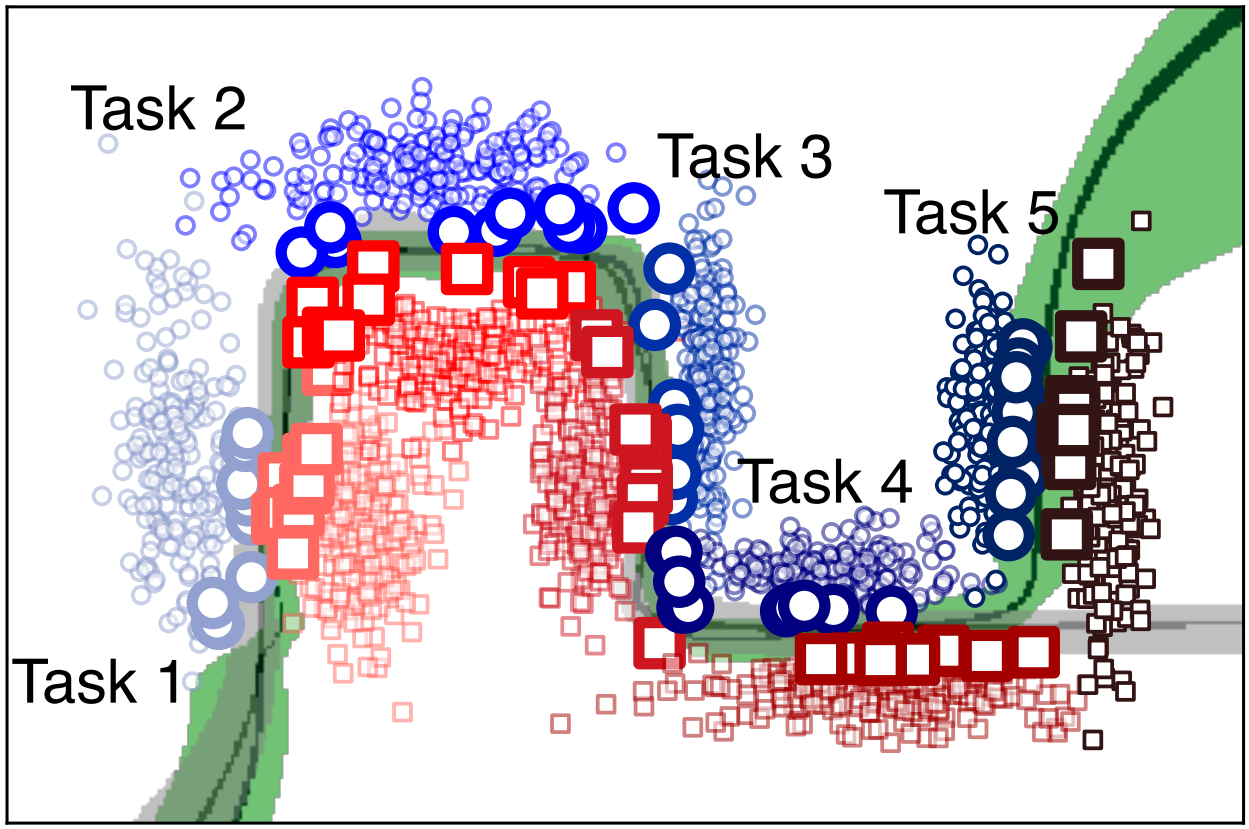


VOGN improves the gap

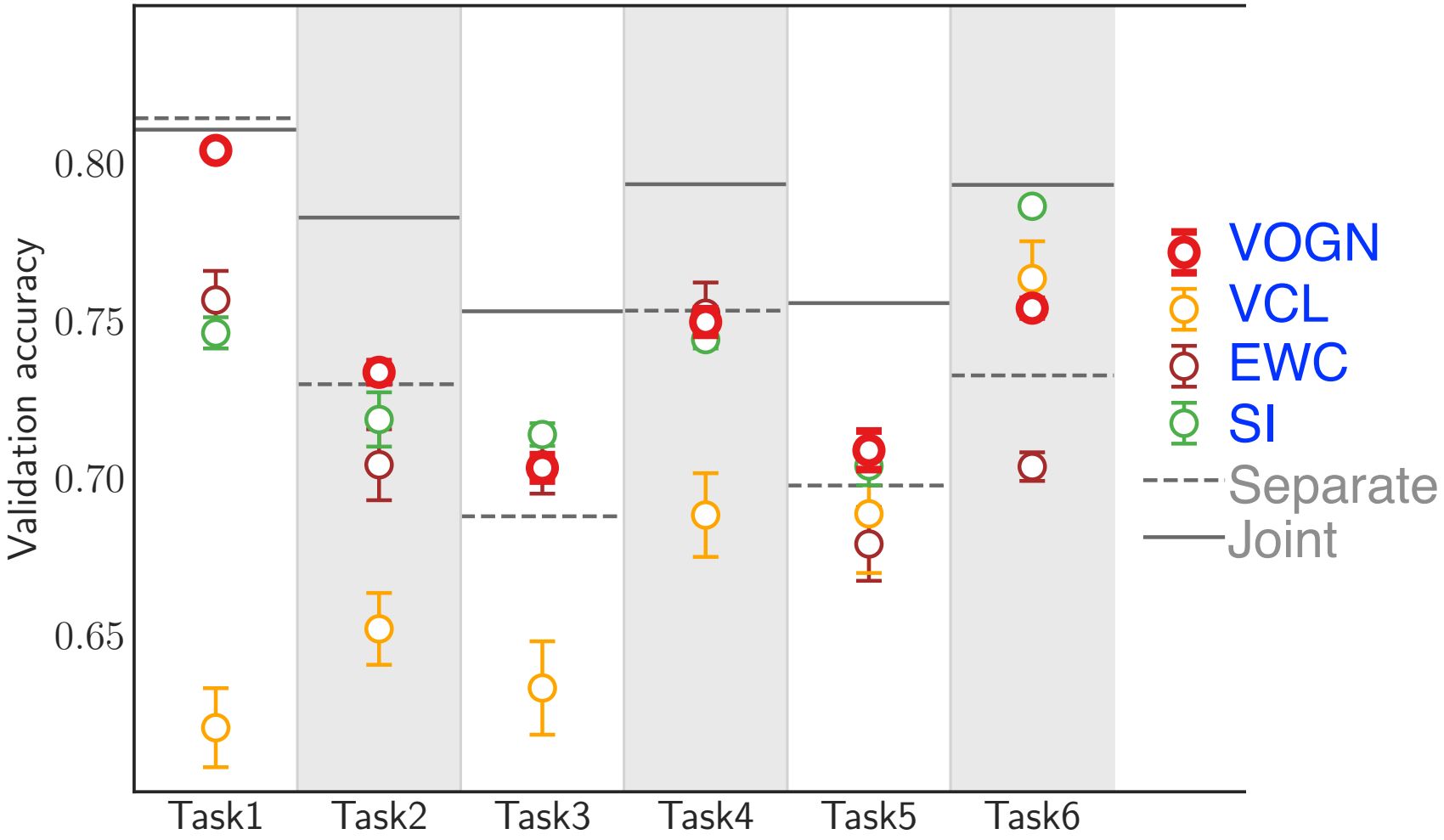


Functional Regularization of Memorable Past (FROMP)

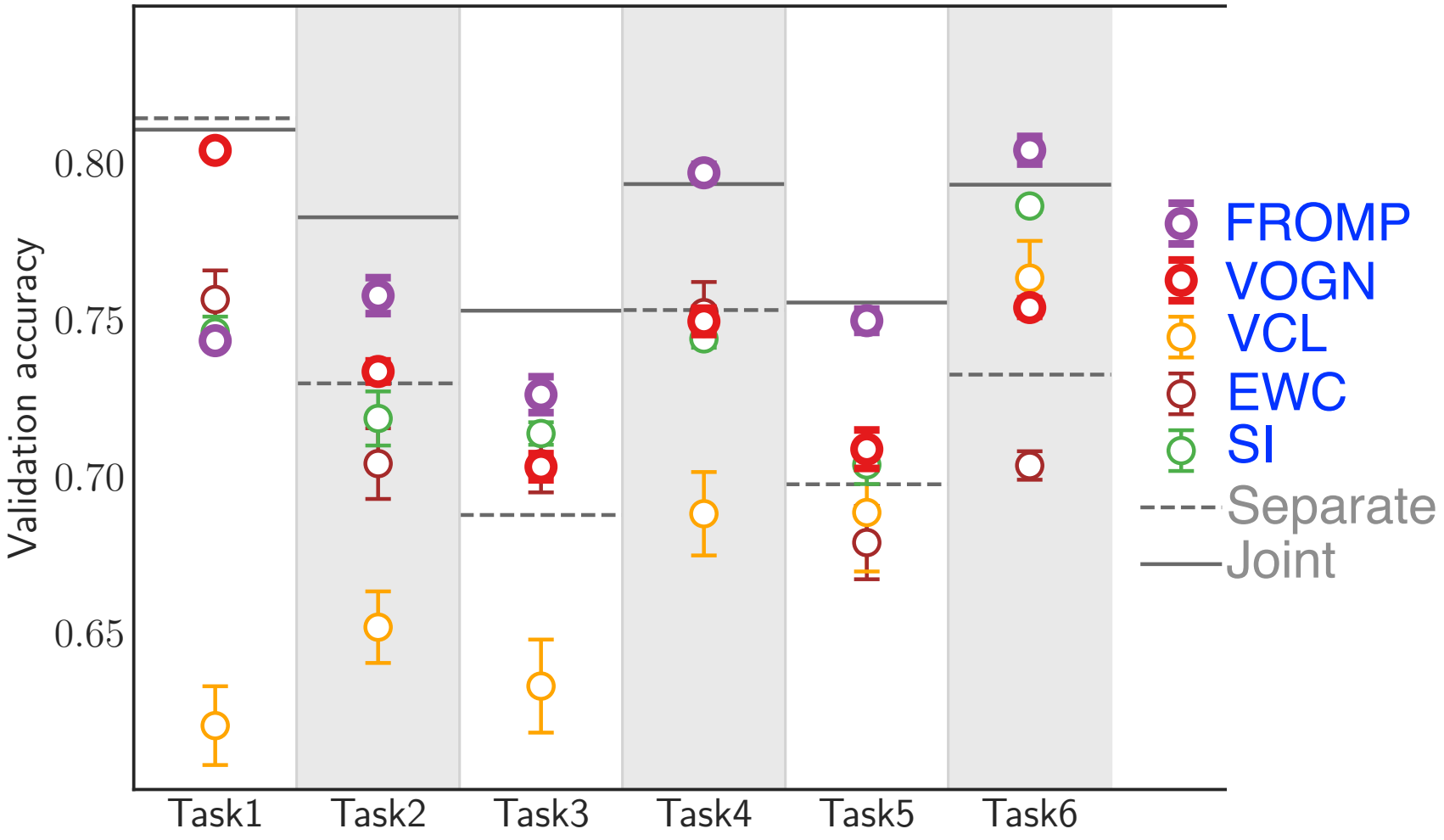
Identify, memorize, and regularize the past using Laplace Approximation (similar to EWC)



FROMP improves over EWC!



FROMP improves over EWC!



Challenges in Continual Learning with Bayesian Approaches

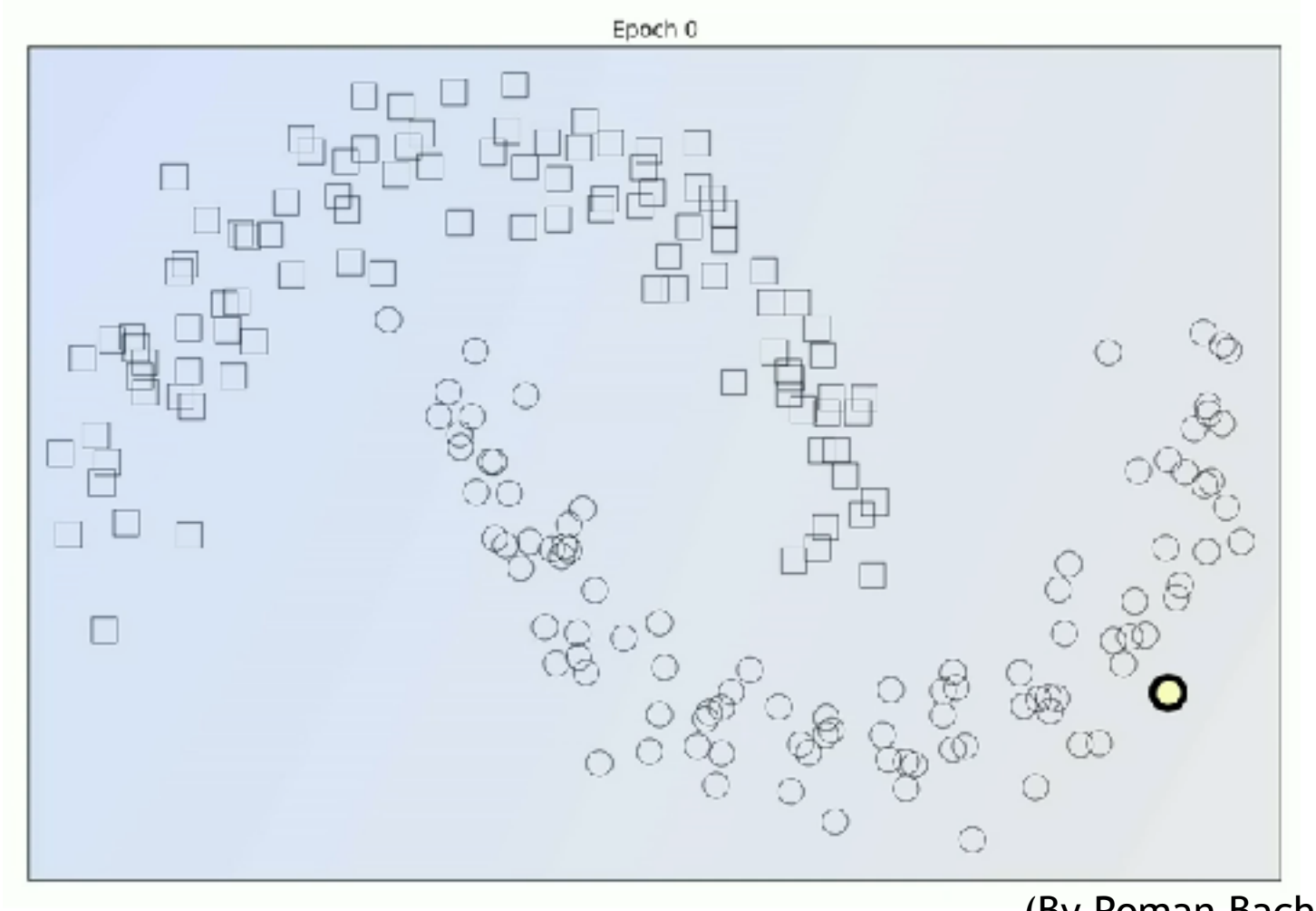
- Computing exact posterior is not tractable
- Approximations do not always behave the way we want them to
  - They can miss important information from the past and lead to forgetting
- Working with the data space could be one solution.
- There are plenty of non-Bayesian solutions and many promising
  - Links to Bayesian principles?

Summary of Continual Learning

Better approximations should give better performance (or at least we should aim for that)

Active Deep Learning

Select “Important” examples while training with Adam



(By Roman Bachmann)

Bayesian Principles: Theory, Derivation, and Related Works

Step A: Express Bayes rule as optimization

Step B: Introduce posterior approximation

Step C: Estimate approximations using the Bayesian learning rule

References for Bayes as Optimization

$$\arg\min_{q \in \mathcal{P}} \; \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{H}(q)$$

- Bayesian statistics
  1. Jaynes, Edwin T. "Information theory and statistical mechanics." *Physical review* (1957)
  2. Zellner, A. "Optimal information processing and Bayes's theorem." *The American Statistician* (1988)
  3. Bissiri, Pier Giovanni, Chris C. Holmes, and Stephen G. Walker. "A general framework for updating belief distributions." *RSS: Series B (Statistical Methodology)* (2016)
- PAC-Bayes
  4. Shawe-Taylor, John, and Robert C. Williamson. "A PAC analysis of a Bayesian estimator." COLT 1997.
  5. Alquier, Pierre. "PAC-Bayesian bounds for randomized empirical risk minimizers." *Mathematical Methods of Statistics* 17.4 (2008): 279-304.
- Online-learning (Exponential Weight Aggregates)
  6. Cesa-Bianchi, Nicolo, and Gabor Lugosi. *Prediction, learning, and games*. 2006.
- Free-energy principle
  7. Friston, K. "The free-energy principle: a unified brain theory?." *Nature neuroscience* (2010)

References for Posterior Approximations

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

- Variational inference

  1. Hinton, Geoffrey, and Drew Van Camp. "Keeping neural networks simple by minimizing the description length of the weights." *COLT* 1993.
  2. Jordan, Michael I., et al. "An introduction to variational methods for graphical models." *Machine learning* 37.2 (1999): 183-233.

- Entropy-regularized / Maximum-entropy RL

  3. Williams, Ronald J., and Jing Peng. "Function optimization using connectionist reinforcement learning algorithms." *Connection Science* 3.3 (1991): 241-268.
  4. Ziebart, Brian D. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. Diss. figshare, 2010. (see chapter 5)

- Parameter-Space Exploration in RL

  5. Rückstiess, Thomas, et al. "Exploring parameter space in reinforcement learning." *Paladyn, Journal of Behavioral Robotics* 1.1 (2010): 14-24.
  6. Plappert, Matthias, et al. "Parameter space noise for exploration." *arXiv preprint arXiv: 1706.01905* (2017)
  7. .Fortunato, Meire, et al. "Noisy networks for exploration." *arXiv preprint arXiv: 1706.10295* (2017).

More References for Posterior Approximations

- Evolution strategy $\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\theta)}[\ell(\theta)]$
  1. Wierstra, Daan, et al. "Natural evolution strategies." 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence). IEEE, 2008.
- Gaussian Homotopy
  2. Mobahi, Hossein, and John W. Fisher III. "A theoretical analysis of optimization by Gaussian continuation." *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.
- Smoothing-based Optimization
  3. Leordeanu, Marius, and Martial Hebert. "Smoothing-based optimization." *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.
- Graduated Optimization
  4. Hazan, Elad, Kfir Yehuda Levy, and Shai Shalev-Shwartz. "On graduated optimization for stochastic non-convex problems." *International conference on machine learning*. 2016.
- Stochastic Search
  5. Zhou, Enlu, and Jiaqiao Hu. "Gradient-based adaptive stochastic search for non-differentiable optimization." *IEEE Transactions on Automatic Control* 59.7 (2014): 1818-1832.

Bayesian Learning Rule and Related Works

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

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

Natural-Gradient VI: $\lambda \leftarrow \lambda - \rho F_q^{-1} \nabla_{\lambda} \left(\mathbb{E}_q[\ell(\theta)] - \mathcal{H}(q) \right)$

Fisher Information Matrix

Also equivalent to a mirror-descent algorithm.The Geometry of the mirror-descent is defined by the log partition function of the posterior approximation.

1. Khan and Lin. "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models." AIstats (2017).
2. Raskutti, Garvesh, and Sayan Mukherjee. "The information geometry of mirror descent." *IEEE Transactions on Information Theory* 61.3 (2015): 1451-1457.

References for Step C: Natural-Gradient VI

1. Sato, Masa-aki. "Fast learning of on-line EM algorithm." Technical Report, ATR Human Information Processing Research Laboratories (1999).
2. Sato, Masa-Aki. "Online model selection based on the variational Bayes." *Neural computation* 13.7 (2001): 1649-1681.
3. Winn, John, and Christopher M. Bishop. "Variational message passing." *Journal of Machine Learning Research* 6.Apr (2005): 661-694.
4. Honkela, Antti, et al. "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes." *Journal of Machine Learning Research* 11.Nov (2010): 3235-3268.
5. Knowles, David A., and Tom Minka. "Non-conjugate variational message passing for multinomial and binary regression." *NeurIPS*. (2011).
6. Hoffman, Matthew D., et al. "Stochastic variational inference." *JMLR* (2013).
7. Salimans, Tim, and David A. Knowles. "Fixed-form variational posterior approximation through stochastic linear regression." *Bayesian Analysis* 8.4 (2013): 837-882.
8. Sheth, Rishit, and Roni Khardon. "Monte Carlo Structured SVI for Two-Level Non-Conjugate Models." *arXiv preprint arXiv:1612.03957* (2016).
9. Khan and Lin. "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models." AIstats (2017).
10. Khan and Nielsen. "Fast yet simple natural-gradient descent for variational inference in complex models." (*2018) ISITA*.
11. Zhang, Guodong, et al. "Noisy natural gradient as variational inference." *ICML* (2018).

Black-Box VI & Bayesian Learning rule

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

Black-Box VI [1]: $\lambda \leftarrow \lambda - \rho \nabla_{\lambda} \left(\mathbb{E}_q[\ell(\theta)] - \mathcal{H}(q) \right)$

Black-box VI is more generally applicable (beyond exponential-family), but we cannot derive learning-algorithms from it (even for conjugate Bayesian models)

1. Ranganath, Rajesh, Sean Gerrish, and David Blei. "Black box variational inference." *Artificial Intelligence and Statistics*. 2014.

Deep Learning with Bayesian Principles

- Bayesian principles as common principles
  - By computing "posterior approximations"
- Derive many existing algorithms,
  - Deep Learning (SGD, RMSprop, Adam)
  - Exact Bayes, Laplace, Variational Inference, etc
- Design new deep-learning algorithms
  - Uncertainty, data importance, life-long learning
- Impact: Many learning-algorithms with a common set of principles.

Open Challenges

- Deep Learning + Bayes Learning
  - Principles of “trial and error” and “bayes” together
- How to achieve Life-long deep learning?
- How to compute better posterior approx?
- How to compute higher-order gradients?

Towards Life-long learning

- For life-long learning, we need
  - Perception: how you want to see the world?
  - Action: what you want to see in the world?
- Posterior approximation connects the two
  - Models are representation of the world
  - Approximations are representation of the model
  - They help us learn the model through actions
  - Act to appropriately “fill” the data space

1. Friston, K. "The free-energy principle: a unified brain theory?." *Nature neuroscience* (2010)

Learning-Algorithms from Bayesian Principles

Coming soon!

A preliminary version is at

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

Havard Rue (KAUST)

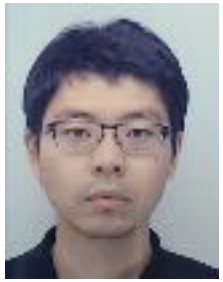
Acknowledgements

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

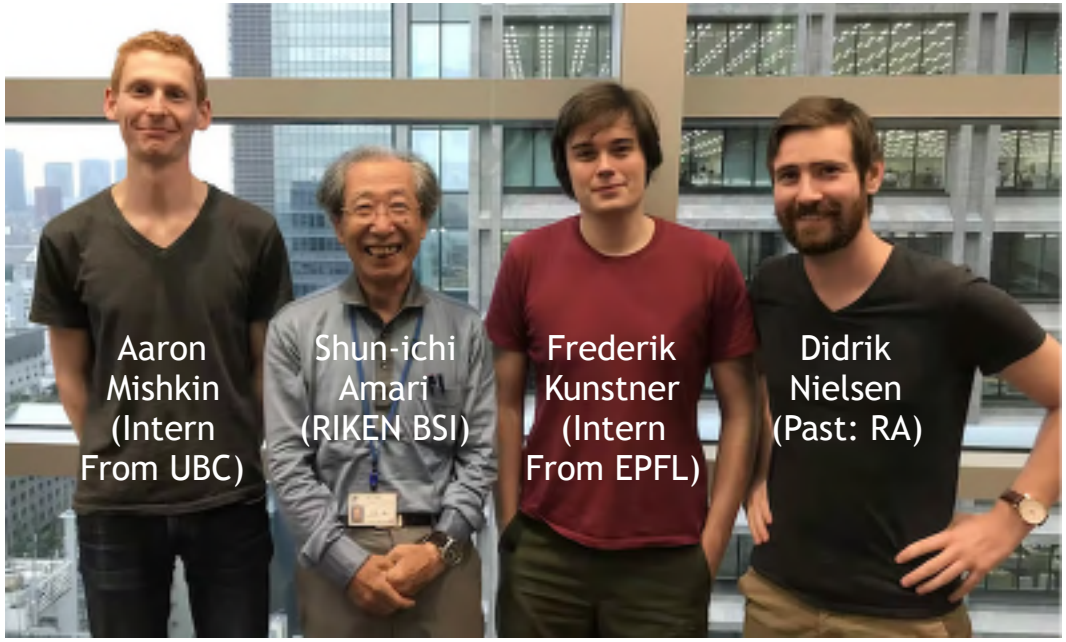
Wu Lin
(Past: RA)

Nicolas Hubacher
(Past: RA)

Masashi Sugiyama
(Director RIKEN-AIP)



Voot Tangkaratt
(Postdoc, RIKEN-AIP)



External Collaborators

Zuozhu Liu
(Intern from SUTD)

RAIDEN

Mark Schmidt
(UBC)

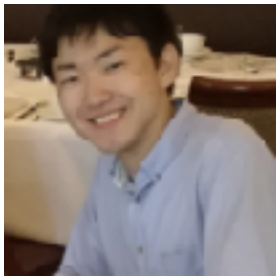
Reza Babanezhad
(UBC)

Yarin Gal
(UOxford)

Akash Srivastava
(UEdinburgh)

Acknowledgements

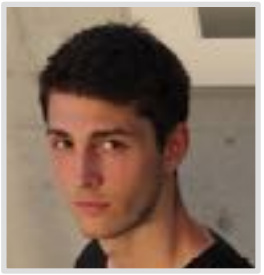
Slides, papers, & code are at emtiyaz.github.io



Kazuki Osawa (Tokyo Tech)

Rio Yokota (Tokyo Tech)

Anirudh Jain (Intern from IIT-ISM, India)



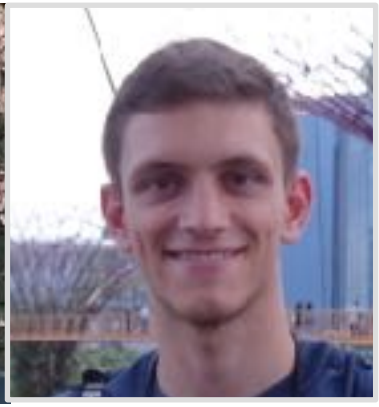
Runa Eschenhagen (Intern from University of Osnabruck)

Siddharth Swaroop (University of Cambridge)

Rich Turner (University of Cambridge)

Alexander Immer (Intern from EPFL)

Ehsan Abedi (Intern from EPFL)



Maciej Korzepa (Intern from DTU)

Pierre Alquier (RIKEN AIP)

Havard Rue (KAUST)

Approximate Bayesian Inference Team