Machine Learning from a Bayesian Perspective

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

1. Summary at https://emtiyaz.github.io/papers/MLfromBayes.pdf
2. Slides at https://emtiyaz.github.io/

Information Processing == Bayes' Updating

Human Learning at the age of 6 months.

Converged at the age of 12 months

Transfer skills

at the age of 14 months

Failure of AI in “dynamic” setting

Robots need quick adaptation to be deployed
(for example, at homes for elderly care)

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

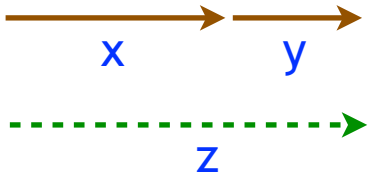
Fixing Machine Learning

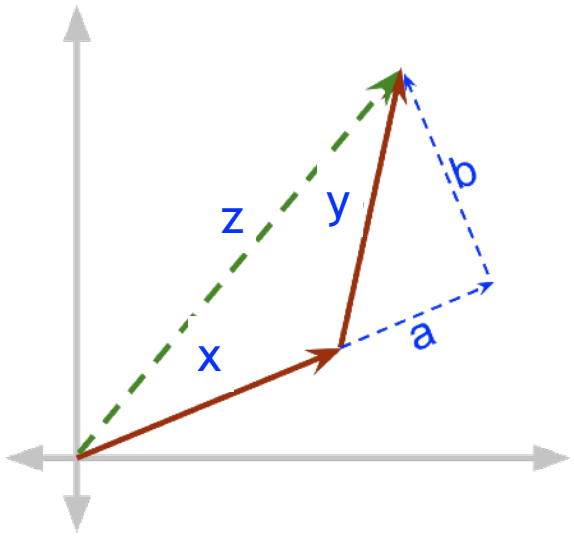
- Even a small change may need full retraining
  - Huge amount of resources only few can afford (costly & unsustainable) [1,2, 3]
  - Difficult to apply in “dynamic” settings (robotics, epidemiology, climate science etc)
- We need sustainable, transparent, trustworthy AI
  - Use reliable building blocks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

Information Procession 101

1. Think addition of numbers
2. Addition of Vectors [1]
3. Multiplication of Probabilities





$$p(x\,|\,y) \propto p(y\,|\,x)p(x)$$

$$\log p(x\,|\,y) = \log p(y\,|\,x) + \log p(x) + const\,.$$

1. https://firmfunda.com/maths/vector-algebra/vector-algebra-addition/vector-addition-first-principles

This Talk

- Value of information
  - Good or bad, old or new, here or there
  - Bayes' rule and Posterior uncertainty
- Multiplication through addition
  - Exp-family distribution
  - Conjugate Bayes
- Information Processing in general
  - Projection to exp-family
  - Bayesian Learning Rule and Deep learning

Bayes' Rule

The Value of Information and Posterior Uncertainty

Principle of Trial-and-Error

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

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

Loss

Data

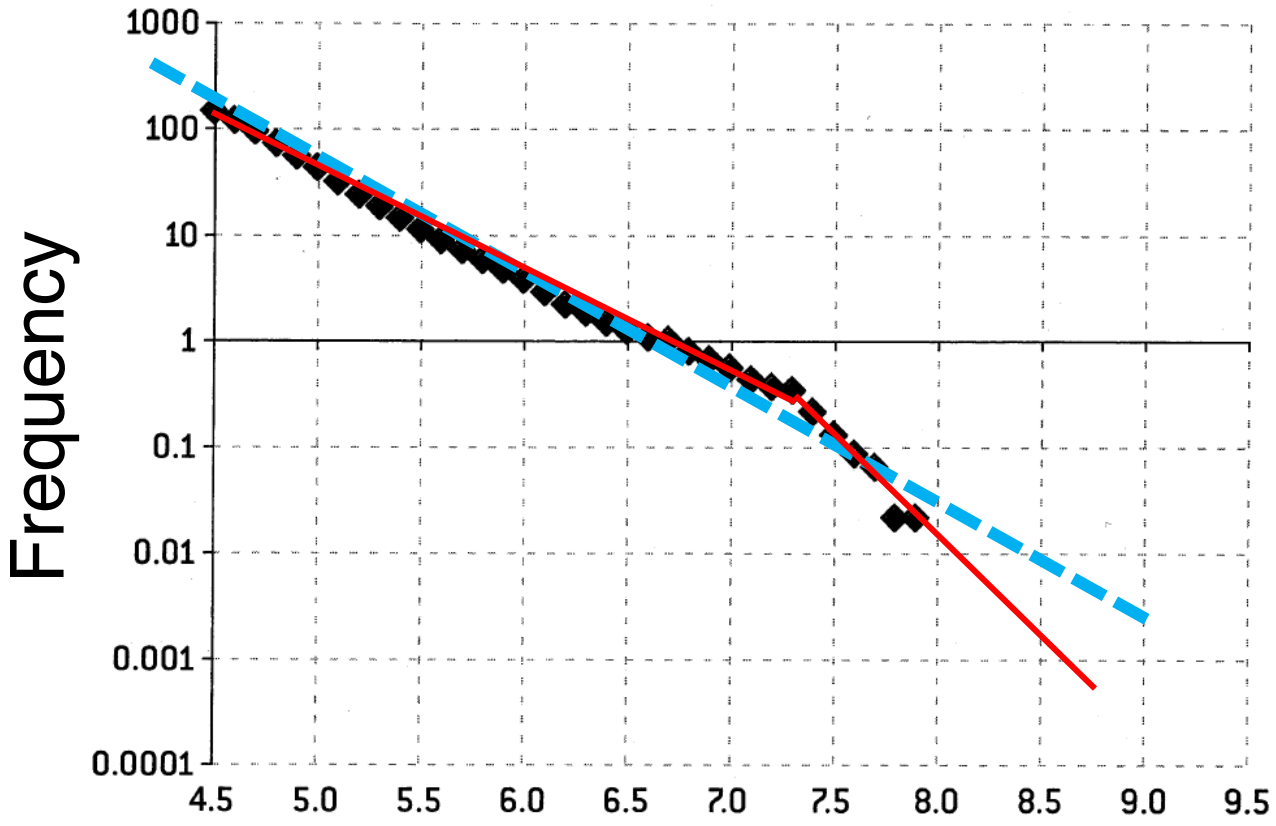
Model Params

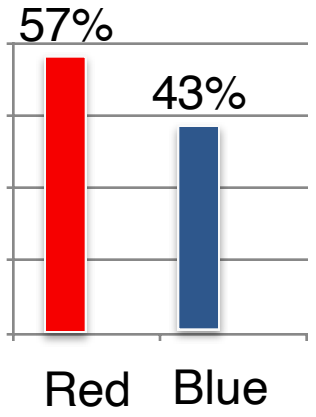
Deep Network

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

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

Example: Which is a Better Fit?

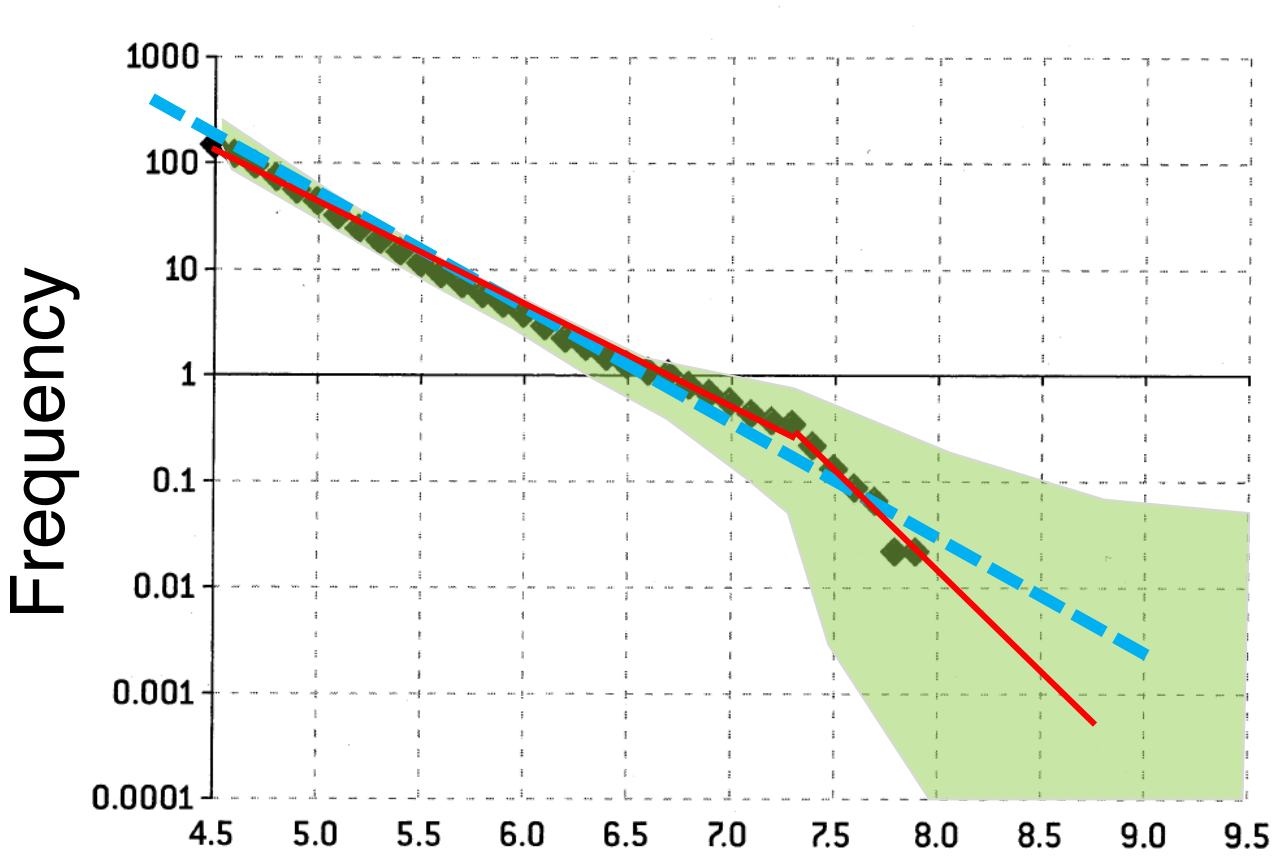




Red is more risky than the blue

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

Value of Information: Uncertainty



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

A Bayesian Principle

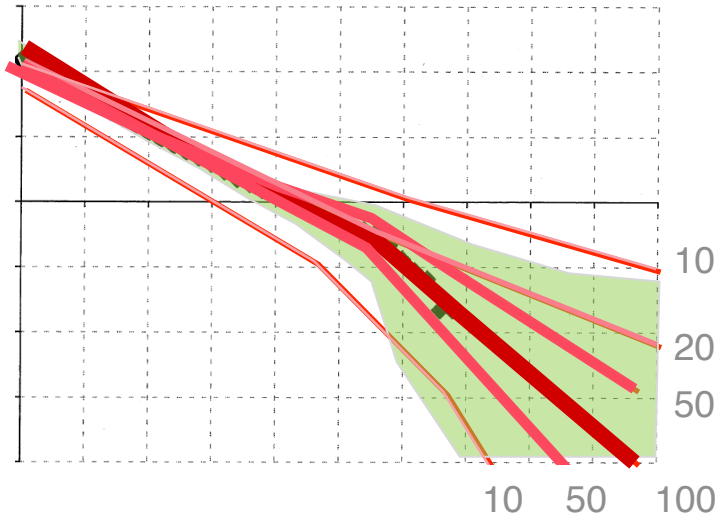
1. Sample $\theta \sim p(\theta)$ prior

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

3. Normalize

Posterior Likelihood x Prior

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



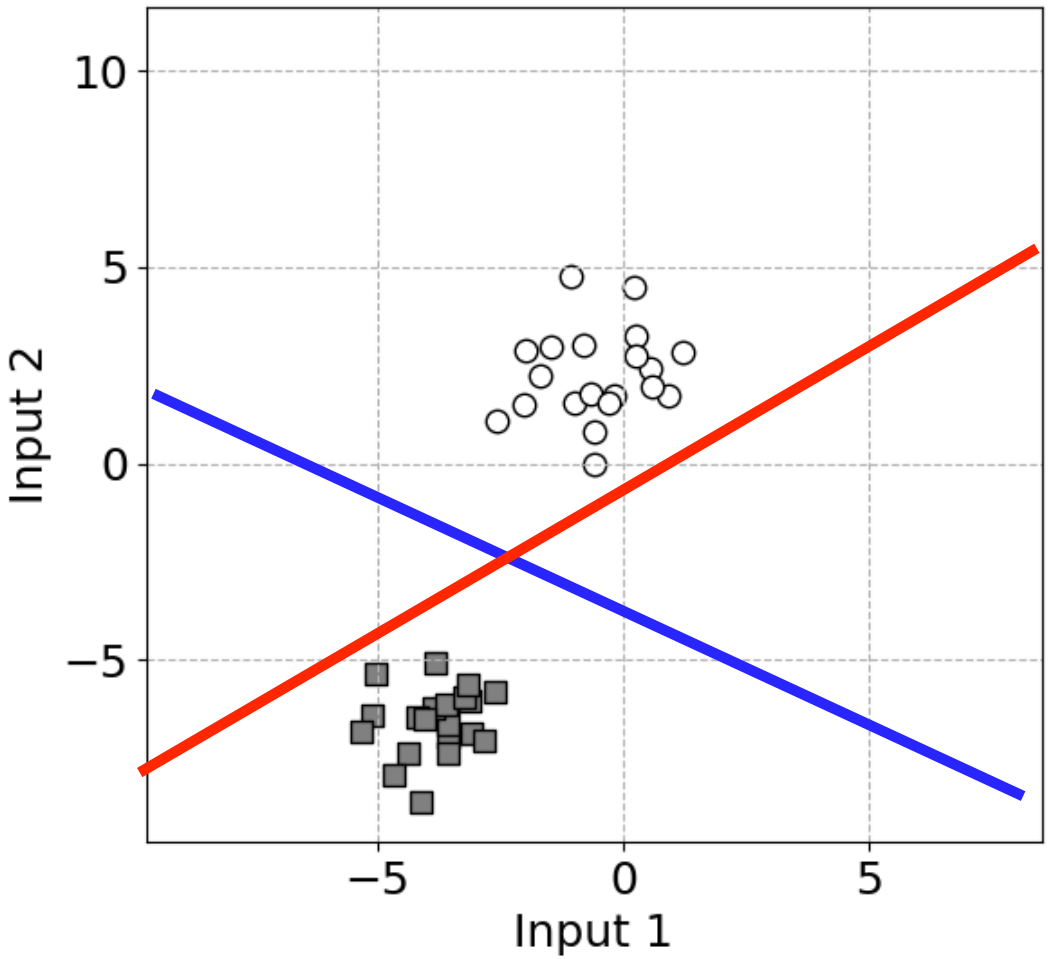
Now, think about the value of information!

Image Segmentation

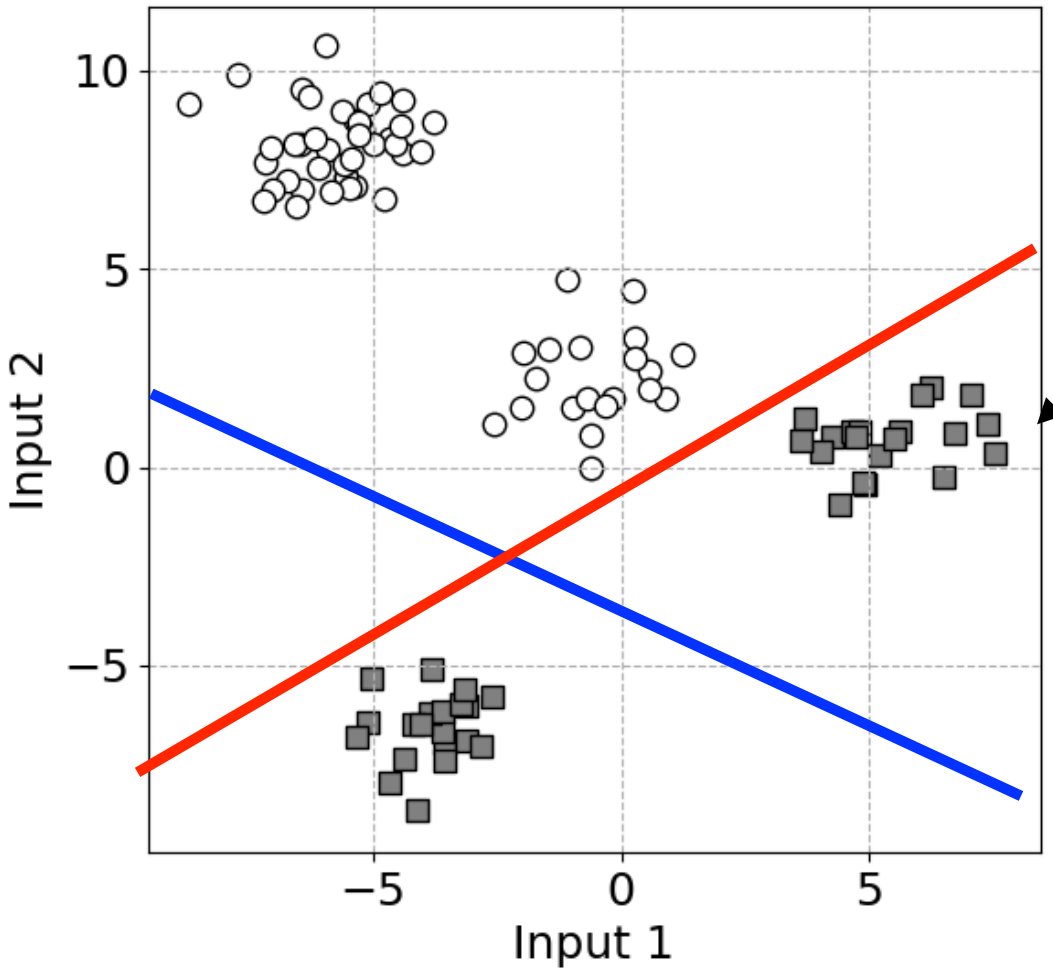
Uncertainty (entropy of class probs)

(By Roman Bachmann)

Which is a good classifier?



Which is a good classifier?

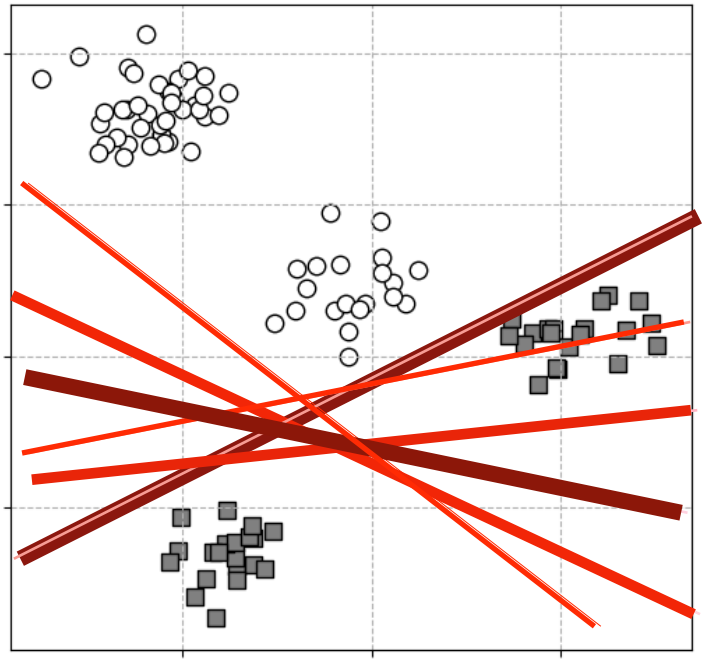


Misclassified by the red line, but not by the blue

What you don't know now, can hurt you later

"Uncertainty matters"

Bayesian Principles



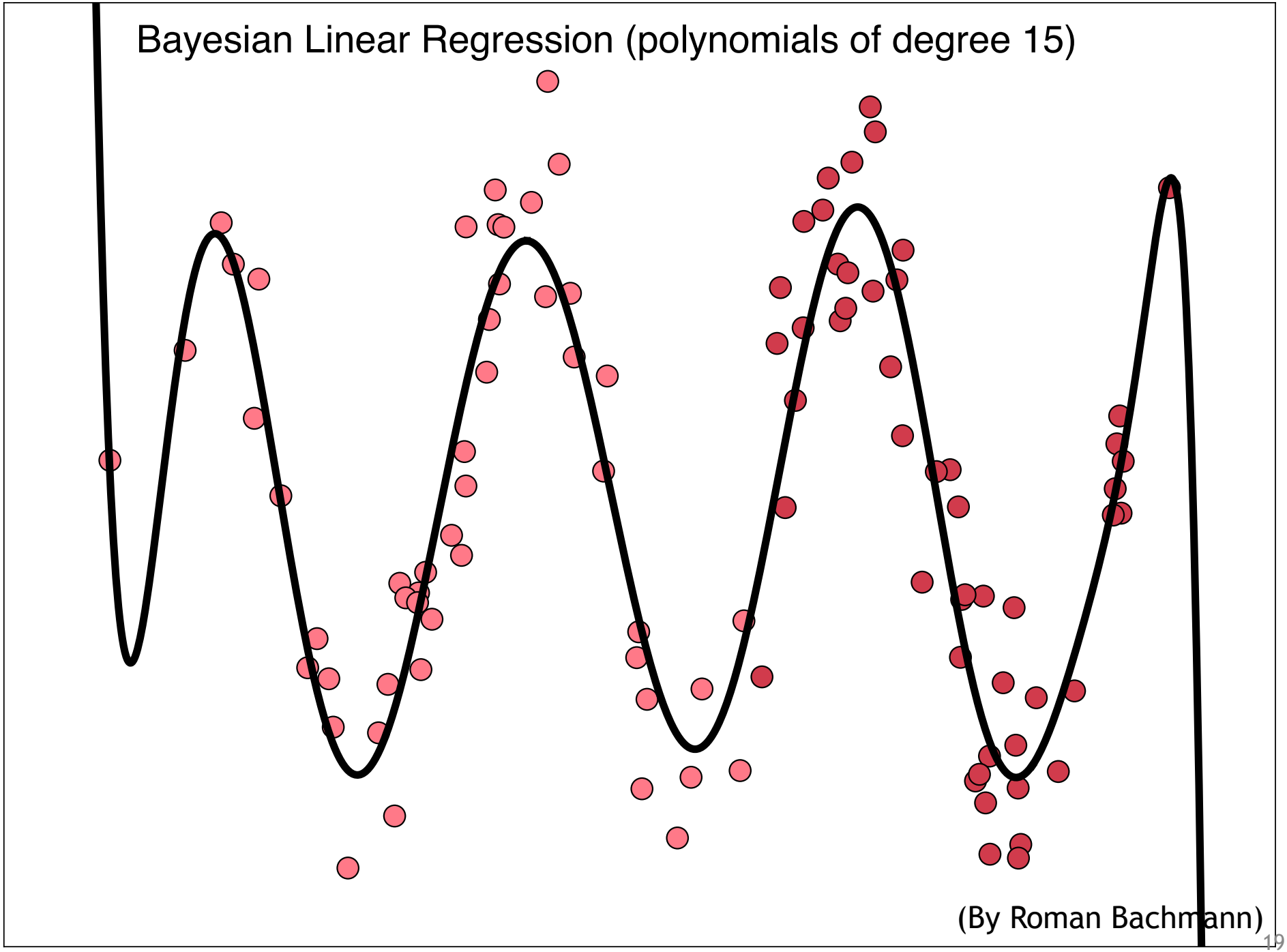
(1) Keep your options open

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

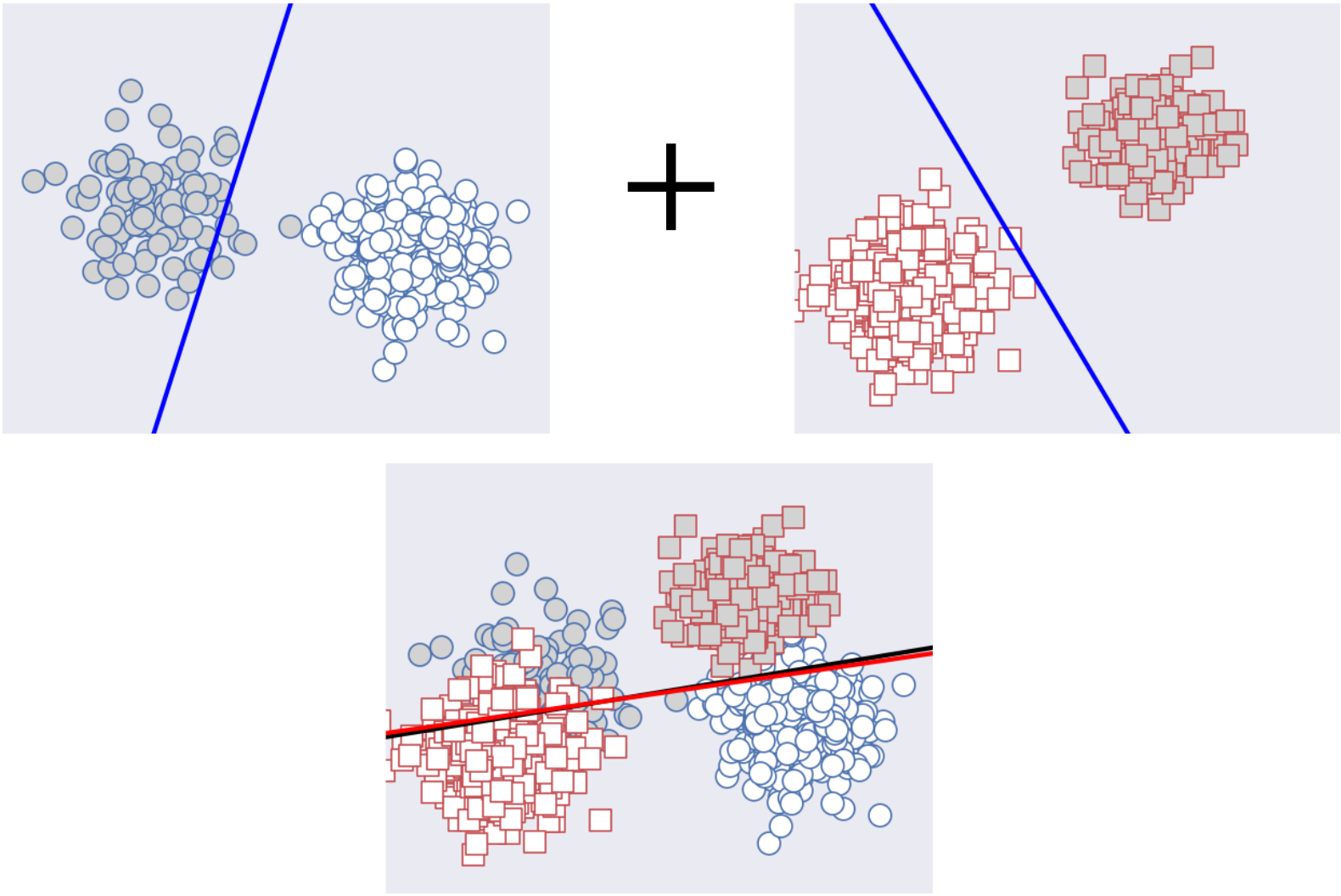
(2) Revise with new evidence

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

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


Bayesian Linear Regression (polynomials of degree 15)
(By Roman Bachmann)

Model Merging



Conjugate Bayes

Multiplication by addition

Exponential-Family distribution

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

Bayes and Conjugate Computations [1]

Multiplication of distribution = addition of (natural) params

Bayes rule: $\text{posterior} \propto \text{lik} \times \text{prior}$

$$e^{\lambda_{\text{post}}^\top T(\theta)} \propto e^{\lambda_{\text{lik}}^\top T(\theta)} \times e^{\lambda_{\text{prior}}^\top T(\theta)}$$

$$\text{log-posterior} = \text{log-lik} + \text{log-prior}$$

$$\lambda_{\text{post}} = \lambda_{\text{lik}} + \lambda_{\text{prior}}$$

1. Khan and Lin, Conjugate computation variational inference, AISTATS, 2017.

General Information Processing

Projection to Exp-Family

Bayesian Learning Rule

For deep learning

Principle of Trial-and-Error

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

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

Loss

Data

Model Params

Deep Network

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

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

Bayes and Conjugate Computations [1]

Multiplication of distribution = addition of (natural) params

Bayes rule: $\text{posterior} \propto \text{lik} \times \text{prior}$

$$e^{\lambda_{\text{post}}^\top T(\theta)} \propto e^{\lambda_{\text{lik}}^\top T(\theta)} \times e^{\lambda_{\text{prior}}^\top T(\theta)}$$

$$\text{log-posterior} = \text{log-lik} + \text{log-prior}$$

$$\lambda_{\text{post}} = \lambda_{\text{lik}} + \lambda_{\text{prior}}$$

This idea can be generalized through natural-gradients.

$$\lambda_{\text{post}} = \nabla_\mu \mathbb{E}_q[\text{log-lik} + \text{log-prior}]$$

Natural gradient

Posterior "approximation"

1. Khan and Lin, Conjugate computation variational inference, AISTATS, 2017.

Bayes Rule as (Natural) Gradient Descent

$$\lambda_{\text{post}} \leftarrow \lambda_{\text{lik}} + \lambda_{\text{prior}}$$

Expected log-lik and log-prior are linear in μ [1]

$$\mathbb{E}_q[\text{log-lik}] = \lambda_{\text{lik}}^\top \mathbb{E}_q[T(\theta)] = \lambda_{\text{lik}}^\top \mu$$

Gradient wrt μ is simply the natural parameter

$$\nabla_\mu \mathbb{E}_q[\text{log-lik}] = \lambda_{\text{lik}}$$

So Bayes' rule can be written as (for an arbitrary q)

$$\lambda_{\text{post}} \leftarrow \nabla_\mu \mathbb{E}_q[\text{log-lik} + \text{log-prior}]$$

As an analogy, think of least-square = 1-step of Newton

1. Khan, Variational-Bayes Made Easy, AABI 2023.

See Appendix A in Khan and Rue, 2021

Bayes Rule 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

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

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

Holds for any loss function (generalized-posterior)

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

The Bayesian Learning Rule

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

Entropy

Posterior approximation (expo-family)

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

Natural and Expectation parameters of q

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

Old belief

New information = natural gradients

Exploiting posterior's information geometry to derive existing algorithms as special instances by approximating q and natural gradients.

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

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

BLR for large deep networks

RMSprop/Adam

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

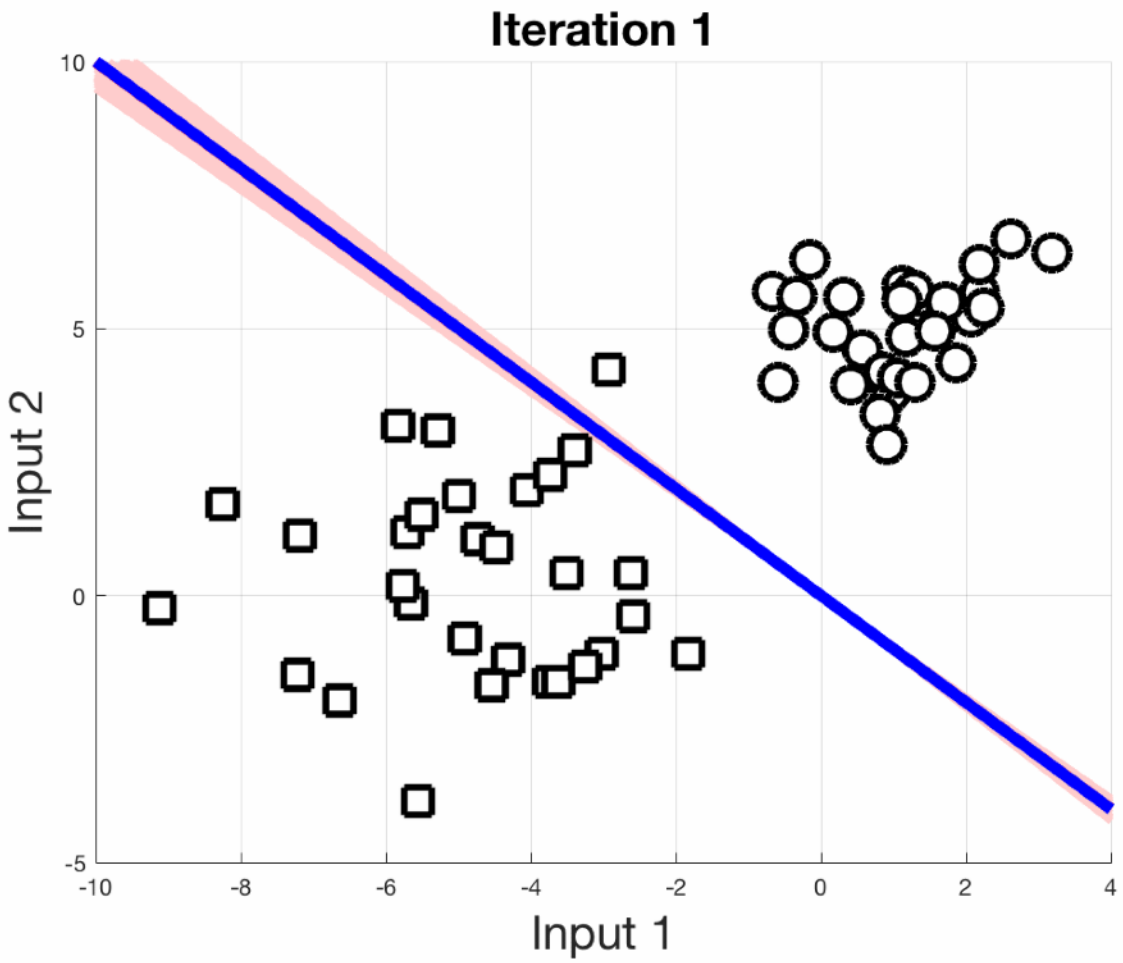
BLR variant
Improved Variational Online Newton (IVON)

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

Only tune initial value of h (a scalar)

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.” Under review (2024)

Logistic Regression



Variational Online Newton method

Frequentist (Adam)

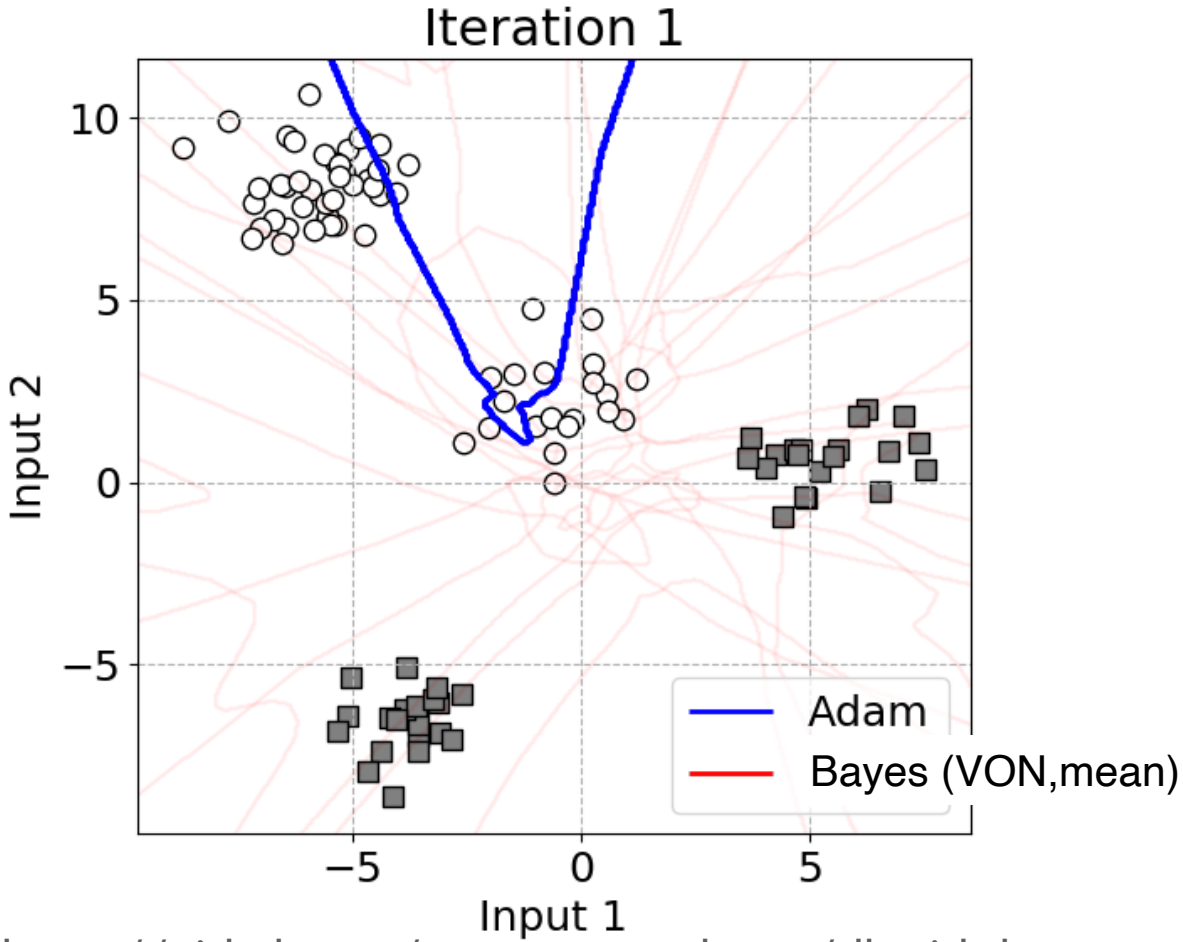
Bayes (VON,mean)

Bayes (samples)

Logistic Regression
Minibatch = 5,
Learning rates = (0.01, 0.01)

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

Deep Learning

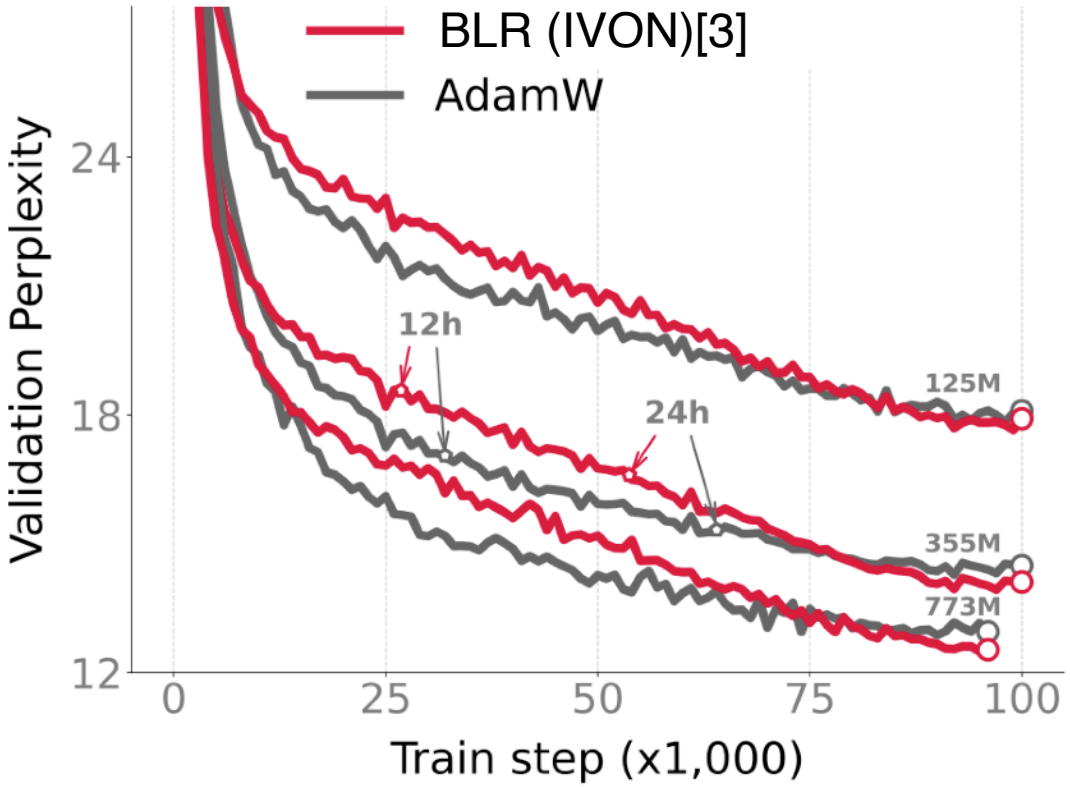


Code available at https://github.com/team-approx-bayes/dl-with-bayes

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. “Practical Deep Learning with Bayesian Principles.” NeurIPS (2019).

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)

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 Aggregate)
  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)

Related Frormulations

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

1. Ingo Rechenberg, Evolutionsstrategie – Optimierung technischer Systeme nach Prinzipien der biologischen Evolution (PhD thesis) 1971.

- 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.

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).

References for 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).