Presentation at the 6th Workshop on Tractable Probabilistic Modeling, UAI 2023, Aug 4, 2023

The Bayesian Learning Rule

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

Summary of recent research at https://emtiyaz.github.io/papers/symposium_2022.pdf
Slides available at https://emtiyaz.github.io/

How to make AI that can adapt quickly?

Reasoning is crucial for this!

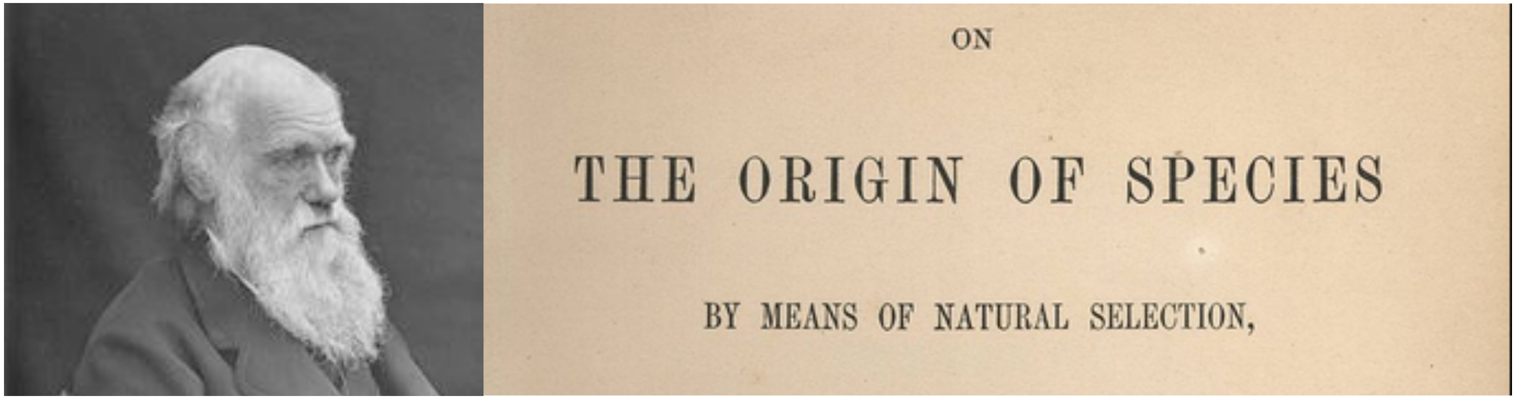
Human Learning at the age of 6 months.

Converged at the age of 12 months

Transfer skills

at the age of 14 months

See Section 6 (discussion) in Khan and Rue, 2021



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

The Origin of Algorithms

What are the common principles behind popular algorithms?

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021

Principles of “good” algorithms?

- Information Geometry of Bayes
  - To unify/generalize/improve learning-algorithms
  - Optimize for “posterior approximations”
- Bayesian Learning rule (BLR)
  - Derive many algorithms from optimization, deep learning, and Bayesian inference
- Natural Gradients are Everywhere!
  - Should also be there in Probabilistic programming and TPM etc., and I hope that this talks helps to build this bridge between TPM community and Approx Bayes.

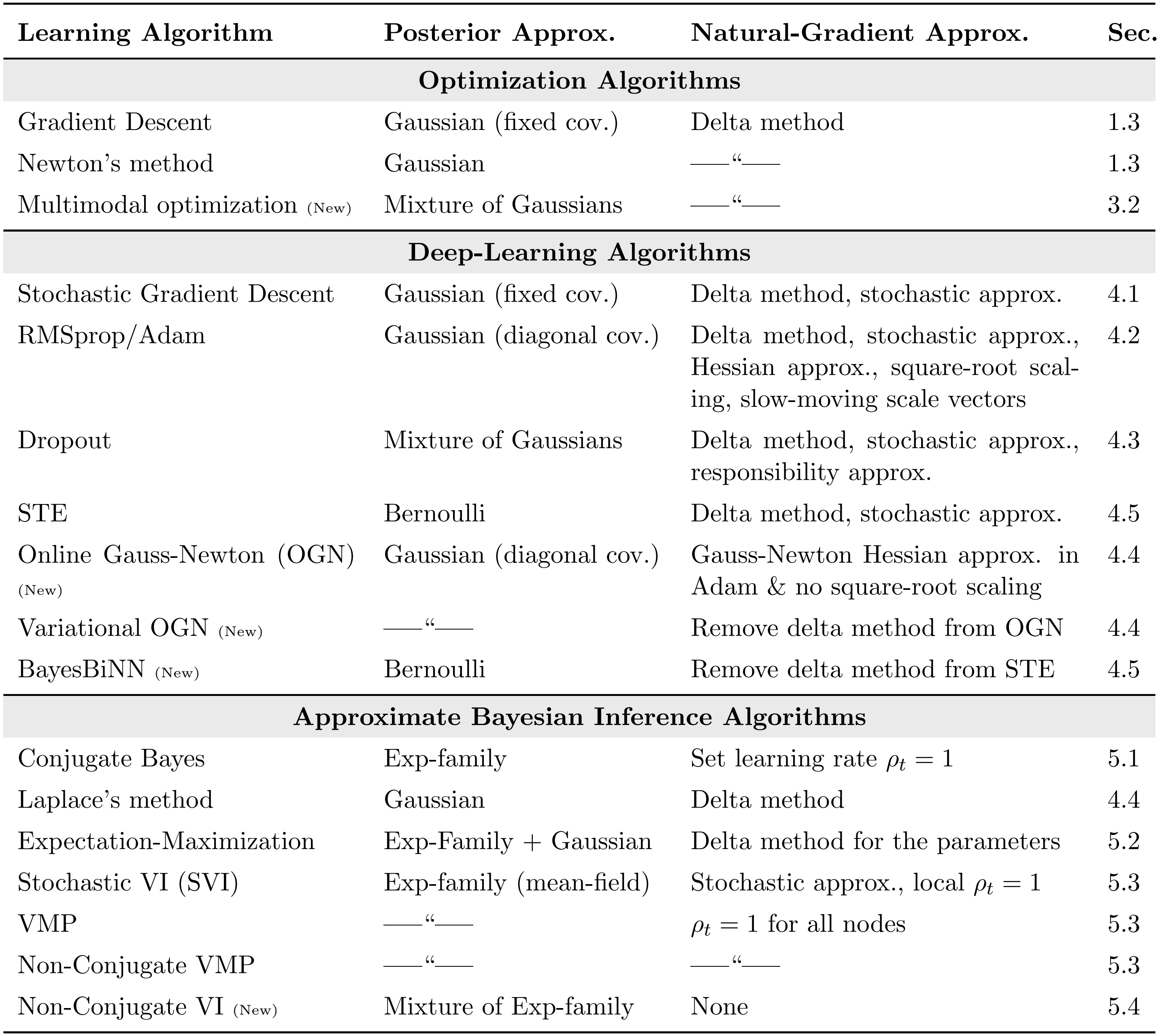
Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021

Bayesian Learning Rule

New information as natural gradients

Bayesian learning rule

See Table 1 in Khan and Rue, 2021



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

(Tractable) Bayesian Learning and Conjugate Computations

Multiplication of distribution = addition of (natural) params

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

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

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

No integrals needed! Tractability is often synonymous to "conjugate computations" [1] and this idea can be generalized through (natural) gradients.

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

Geometry of Exponential Family

We will exploit the geometry of “minimal” exp-family

Natural parameters

Sufficient Statistics

Expectation parameters

$$q(\theta) \propto \exp\left[\lambda^\top T(\theta)\right] \qquad \mu := \mathbb{E}_q[T(\theta)]$$

$$\begin{aligned}\mathcal{N}(\theta|m, S^{-1}) &\propto \exp\left[-\frac{1}{2}(\theta - m)^\top S(\theta - m)\right] \\ &\propto \exp\left[(Sm)^\top\theta + \mathrm{Tr}\left(-\frac{S}{2}\theta\theta^\top\right)\right]\end{aligned}$$

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

Information Geometry of Bayes

Bayes' rule is 1-step of natural-gradient in the λ-space or equivalently a mirror-descent in the (dual) μ-space.

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

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

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

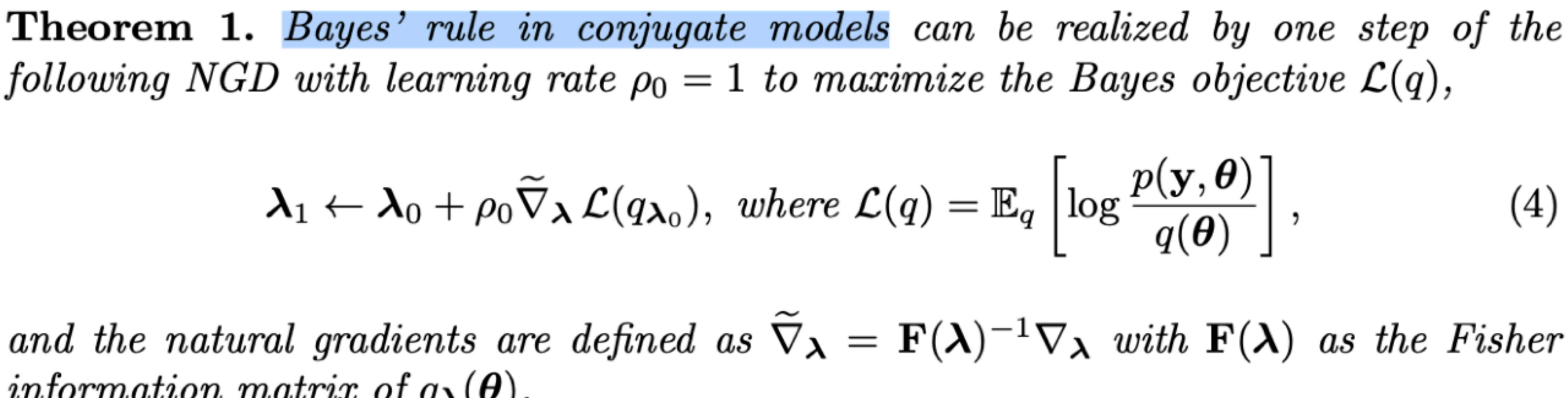
So Bayes' rule can be written as (for an arbitrary q)

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

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

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

Bayes' rule = Information-Geometric Optimization

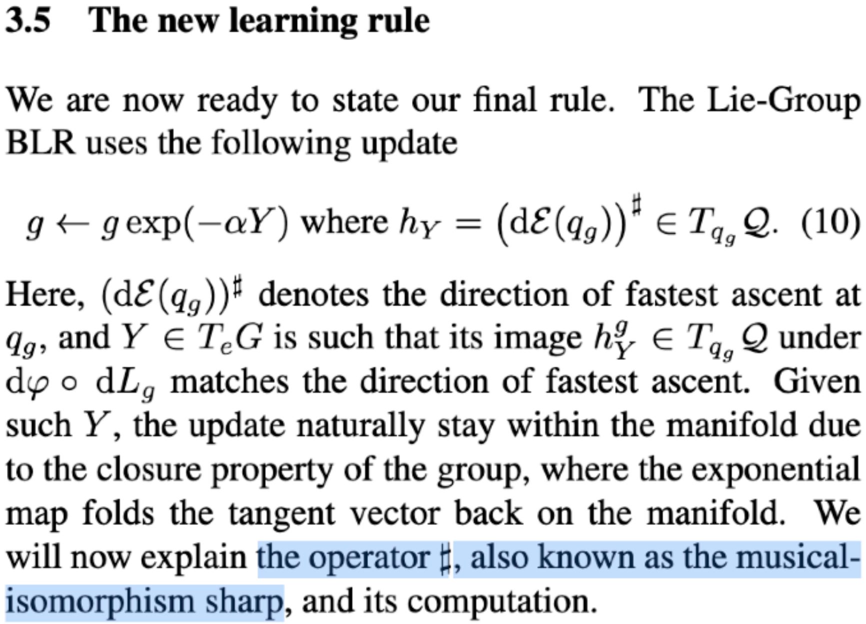


Theorem 1. *Bayes' rule in conjugate models can be realized by one step of the following NGD with learning rate $\rho_0 = 1$ to maximize the Bayes objective $\mathcal{L}(q)$,*

$$\boldsymbol{\lambda}_1 \leftarrow \boldsymbol{\lambda}_0 + \rho_0 \widetilde{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(q_{\boldsymbol{\lambda}_0}), \ \textit{where } \mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right], \tag{4}$$

and the natural gradients are defined as $\widetilde{\nabla}_{\boldsymbol{\lambda}} = \mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}}$ with $\mathbf{F}(\boldsymbol{\lambda})$ as the Fisher information matrix of $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$.

Such results can be written in more general forms (beyond conjugate models). We need to choose the class of q with an appropriate geometry)



3.5 The new learning rule

We are now ready to state our final rule. The Lie-Group BLR uses the following update

$$g \leftarrow g \exp(-\alpha Y) \text{ where } h_Y = \left(\mathrm{d}\mathcal{E}(q_g)\right)^\sharp \in T_{q_g}\mathcal{Q}. \quad (10)$$

Here, $(\mathrm{d}\mathcal{E}(q_g))^\sharp$ denotes the direction of fastest ascent at q_g, and $Y \in T_eG$ is such that its image $h_Y^g \in T_{q_g}\mathcal{Q}$ under $\mathrm{d}\varphi \circ \mathrm{d}L_g$ matches the direction of fastest ascent. Given such Y, the update naturally stay within the manifold due to the closure property of the group, where the exponential map folds the tangent vector back on the manifold. We will now explain the operator $\sharp$, also known as the musical-isomorphism sharp, and its computation.

1. Kiral, Mollenhoff, Khan, The Lie-group Bayesian Learning Rule, AISTATS, 2023

Bayes as Optimization

Bayes rule:

$$\text{posterior} \propto \text{lik} \times \text{prior}$$

Bayes as optimization [1], aka variational inference:

$$\min_{q \in \mathcal{Q}} \; \mathbb{E}_q[\text{log-lik}] + \text{KL}(q \| \text{prior})$$

Generalized Approx Bayes:

log-lik + log-prior

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

Entropy

Posterior approximation (expo-family)

[1] Zellner, Optimal information processing and Bayes's theorem, The American Statistician, 1988.

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

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

$$\lambda \leftarrow \underbrace{(1-\rho)\lambda}_{\text{Old belief}} - \underbrace{\rho \nabla_{\mu} \mathbb{E}_q[\ell(\theta)]}_{\text{New information = natural gradients}}$$

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, arXiv, https://arxiv.org/abs/2107.04562, 2021

[2] Khan and Lin. "Conjugate-computation variational inference….” AIstats (2017).

Warning!

- This natural gradient is different from the one what we (often) encounter in machine learning for Maximum-Likelihood
  - In MLE, the loss is the negative log probability distribution

$$\min_{\theta} -\log q(\theta) \Rightarrow F(\theta)^{-1} \nabla \log q(\theta)$$

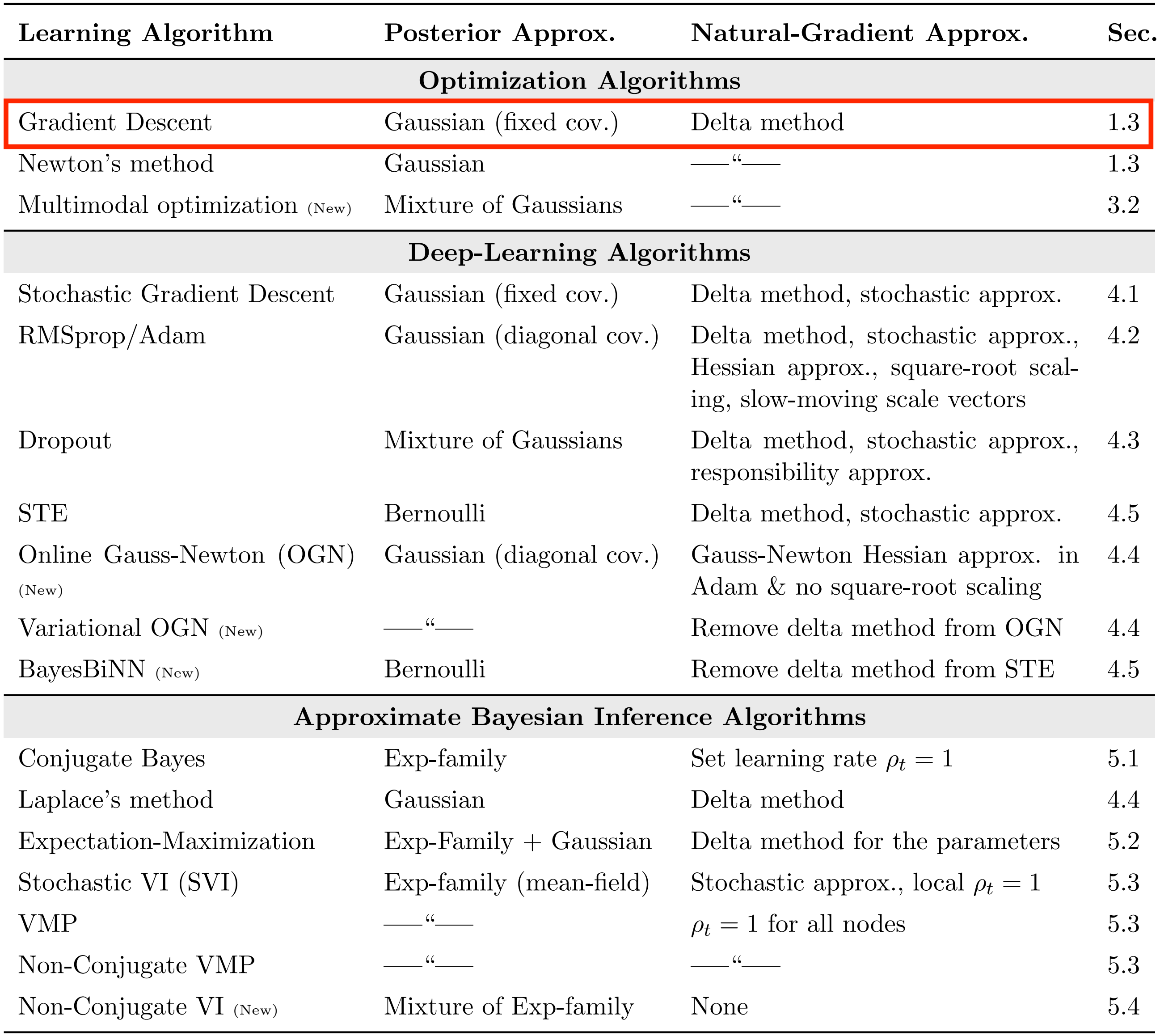
  - Here, loss and distribution are two different entities, even possible unrelated

$$\min_{q} \mathbb{E}_q[\ell(\theta)] - \mathscr{H}(q) \Rightarrow F(\lambda)^{-1} \nabla_\lambda \mathbb{E}_q[\ell(\theta)]$$

Gradient Descent from Bayesian Learning Rule

(Euclidean) gradients as natural gradients

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

See Section 1.3.1 in Khan and Rue, 2021

Gradient Descent from BLR

GD: $\theta \leftarrow \theta - \rho \nabla_\theta \ell(\theta)$

BLR: $m \leftarrow m - \rho \nabla_m \ell(m)$

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

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

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

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 $\mathcal{H}(q) := \log(2\pi)/2$

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

Put the expectation (Bayes) back in and use the Bayesian averaging.

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).

Practical DL with Bayes

RMSprop

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

BLR variant called VOGN

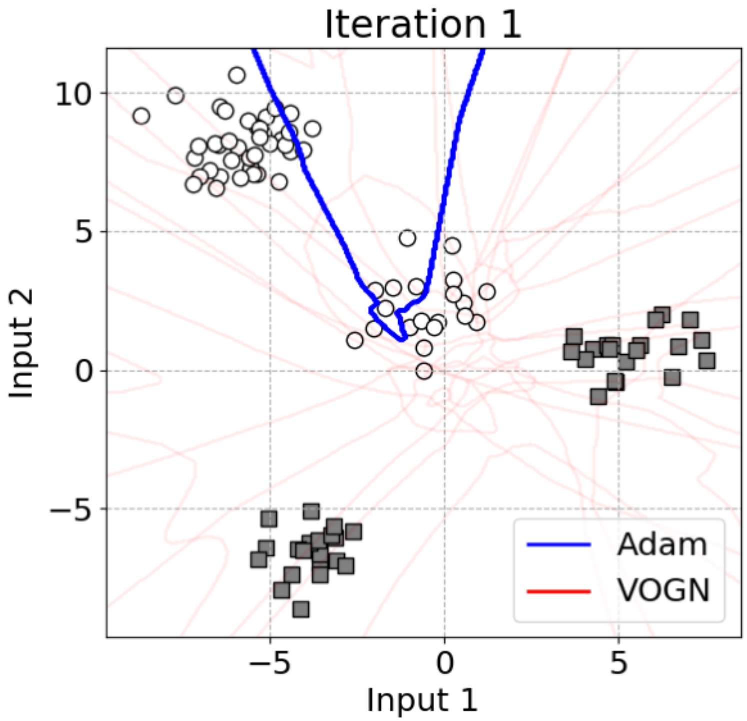
$$g \leftarrow \hat{\nabla}\ell(\theta), \text{ where } \theta \sim \mathcal{N}(m, \sigma^2)$$
$$s \leftarrow (1-\rho)s + \rho(\Sigma_i g_i^2)$$
$$m \leftarrow m - \alpha(s+\gamma)^{-1}\nabla_\theta \ell(\theta)$$
$$\sigma^2 \leftarrow (s+\gamma)^{-1}$$

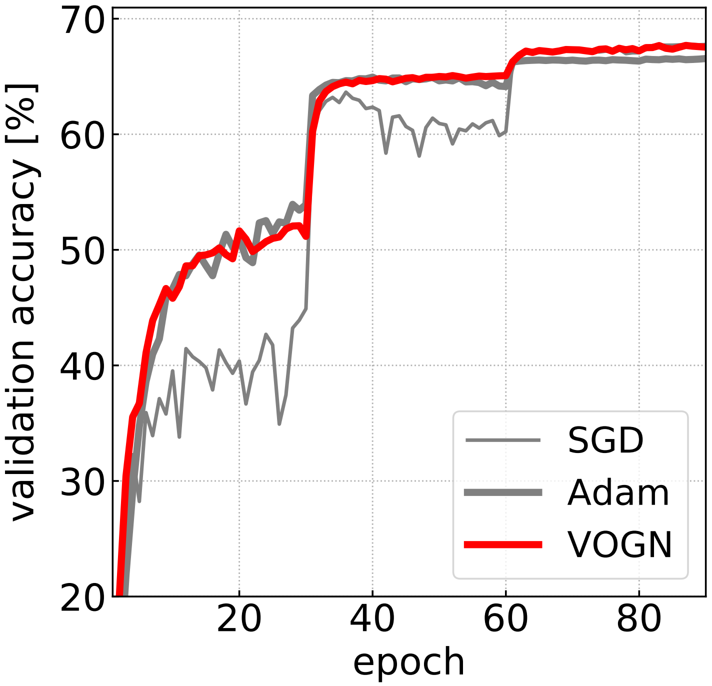
Available at https://github.com/team-approx-bayes/dl-with-bayes

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).
3. Lin et al. “Handling the positive-definite constraints in the BLR.” ICML (2020).

Uncertainty of Deep Nets

VOGN: A modification of Adam with similar performance on ImageNet, but better uncertainty





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).

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

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

Mixture-of-Gaussian Posteriors with an Improved Bayesian Learning Rule

Thomas Möllenhoff[1], Yuesong Shen[2], Gian Maria Marconi[1]
Peter Nickl[1], Mohammad Emtiyaz Khan[1]

[1] Approximate Bayesian Inference Team
RIKEN Center for AI Project, Tokyo, Japan

[2] Computer Vision Group
Technical University of Munich, Germany

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

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. "Practical Deep Learning with Bayesian Principles." NeurIPS (2019).
3. Lin et al. "Handling the positive-definite constraints in the BLR." ICML (2020).

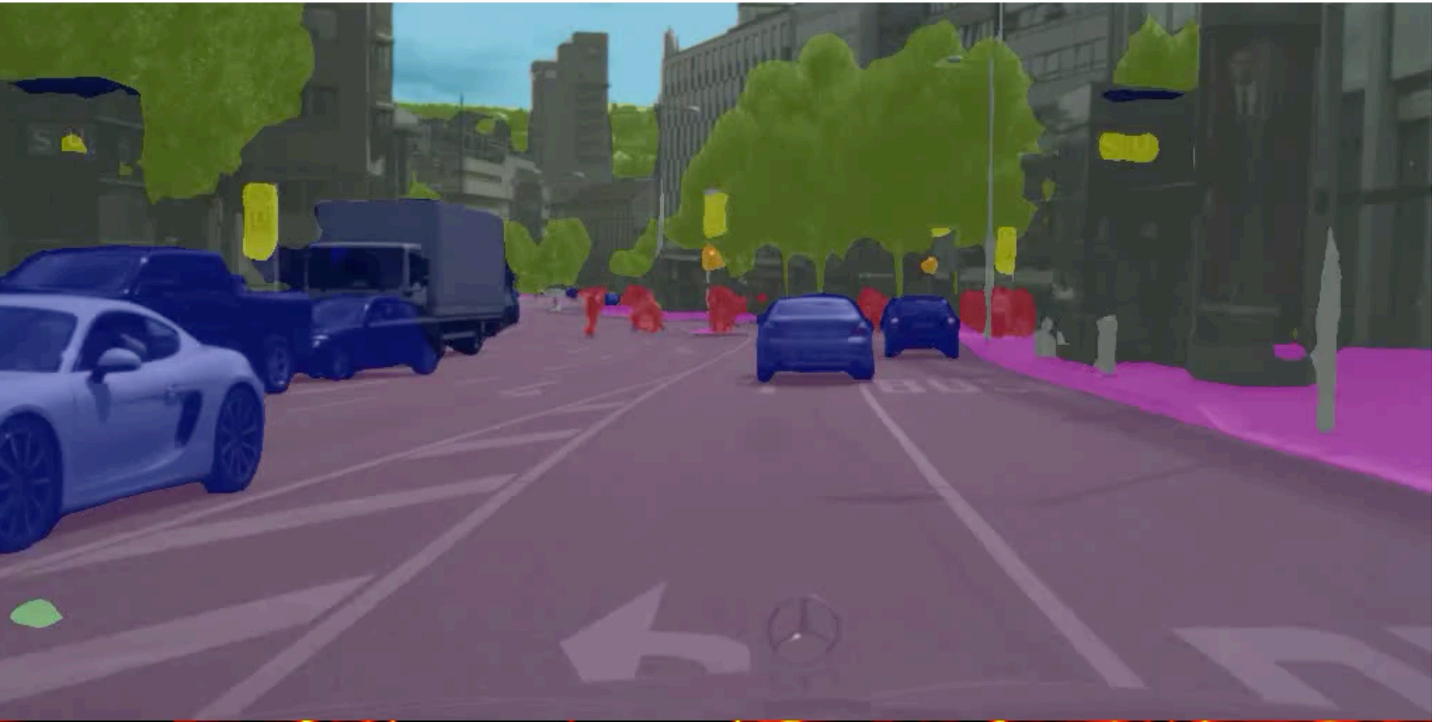
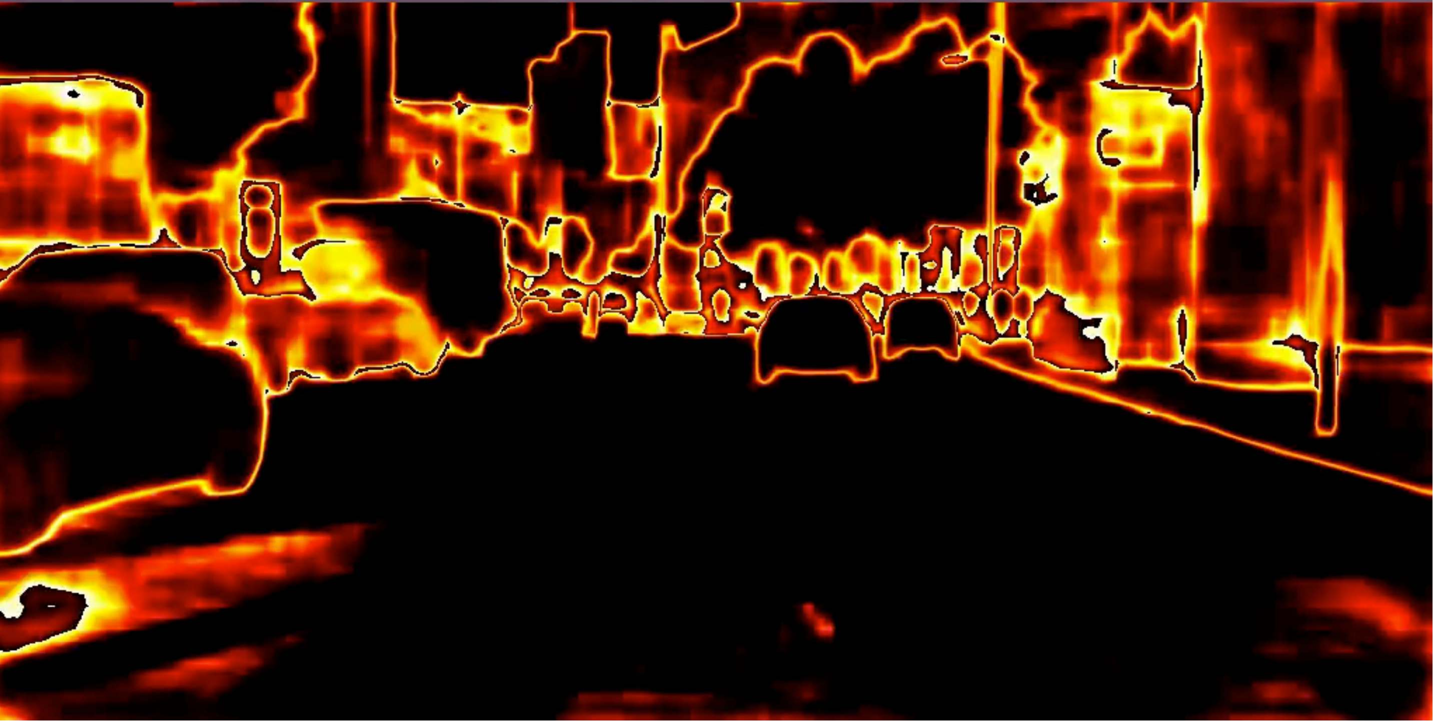


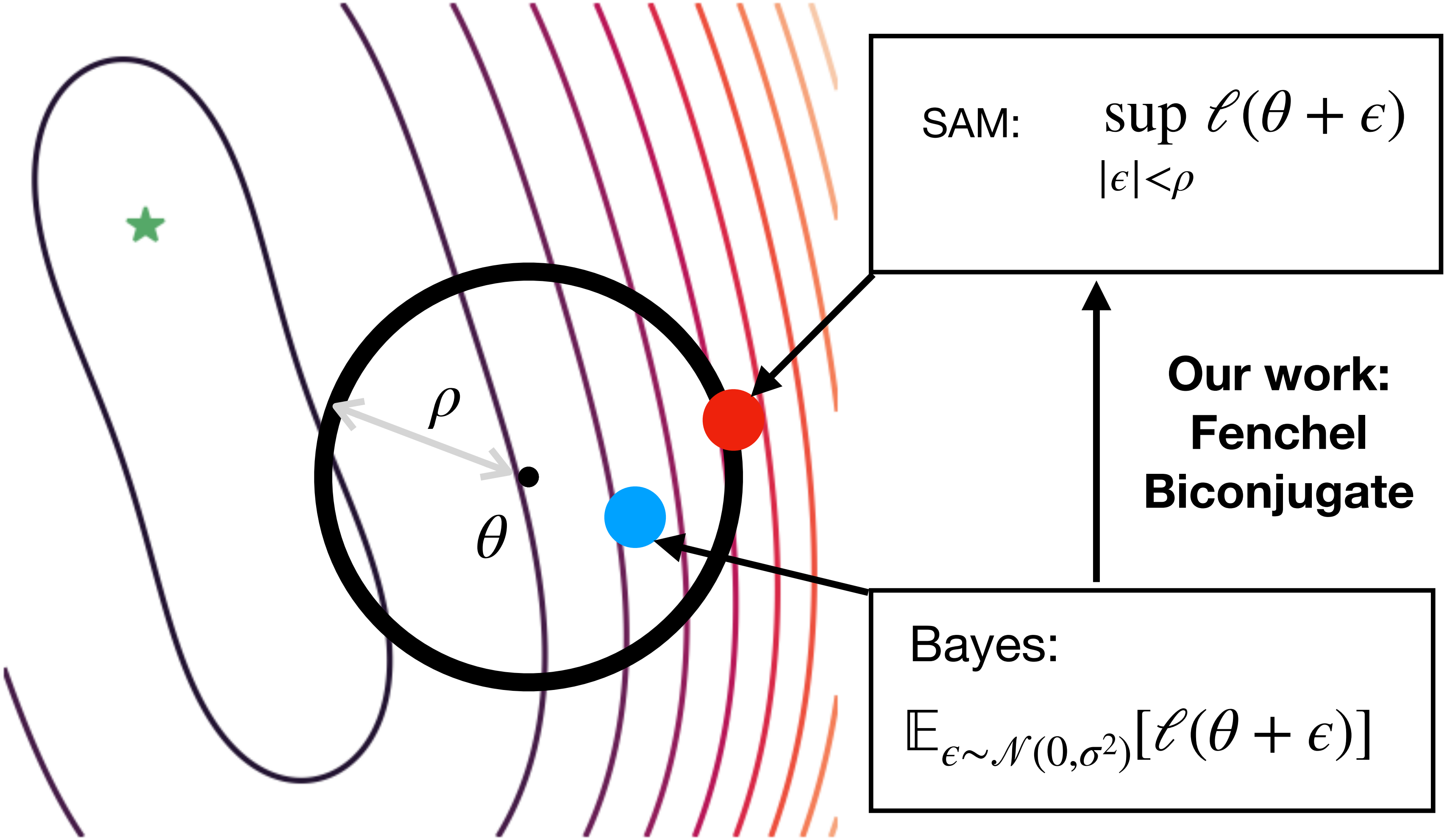
Image Segmentation



Uncertainty (entropy of class probs)

(By Roman Bachmann)

SAM as an Optimal relaxation of Bayes



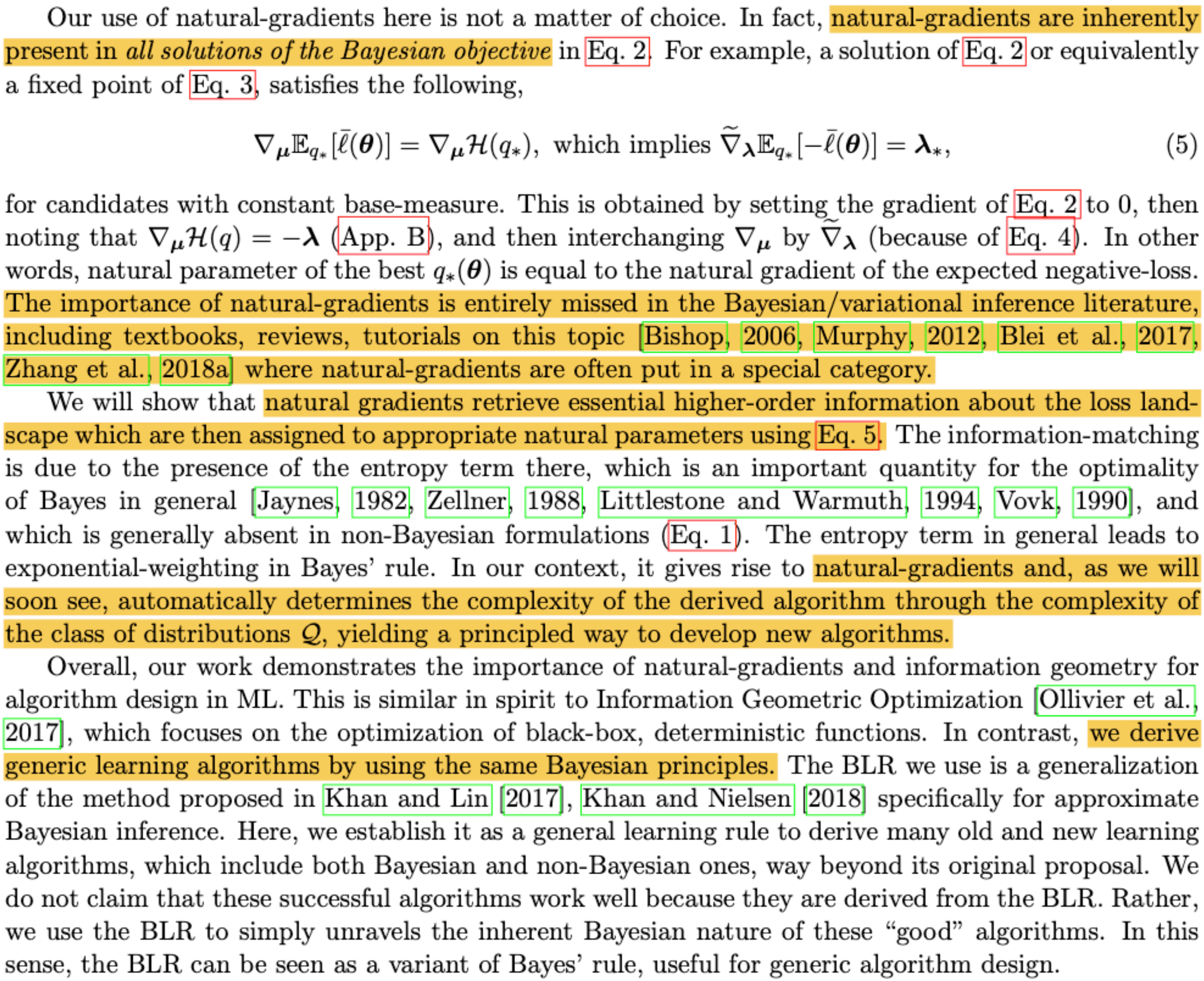
SAM: $\sup_{|\epsilon|<\rho} \ell(\theta + \epsilon)$

Our work: Fenchel Biconjugate

Bayes: $\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)}[\ell(\theta + \epsilon)]$

1. Foret et al. Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR, 2021
2. Moellenhoff and Khan, SAM as an Optimal Relaxation of Bayes, Under review, 2022

See Section 1.2 in Khan and Rue, 2021



Our use of natural-gradients here is not a matter of choice. In fact, natural-gradients are inherently present in *all solutions of the Bayesian objective* in Eq. 2. For example, a solution of Eq. 2 or equivalently a fixed point of Eq. 3, satisfies the following,

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q_*}[\bar{\ell}(\boldsymbol{\theta})] = \nabla_{\boldsymbol{\mu}} \mathcal{H}(q_*), \text{ which implies } \widetilde{\nabla}_{\boldsymbol{\lambda}} \mathbb{E}_{q_*}[-\bar{\ell}(\boldsymbol{\theta})] = \boldsymbol{\lambda}_*, \tag{5}$$

for candidates with constant base-measure. This is obtained by setting the gradient of Eq. 2 to 0, then noting that $\nabla_{\boldsymbol{\mu}} \mathcal{H}(q) = -\boldsymbol{\lambda}$ (App. B), and then interchanging $\nabla_{\boldsymbol{\mu}}$ by $\widetilde{\nabla}_{\boldsymbol{\lambda}}$ (because of Eq. 4). In other words, natural parameter of the best $q_*(\boldsymbol{\theta})$ is equal to the natural gradient of the expected negative-loss. The importance of natural-gradients is entirely missed in the Bayesian/variational inference literature, including textbooks, reviews, tutorials on this topic [Bishop, 2006, Murphy, 2012, Blei et al., 2017, Zhang et al., 2018a] where natural-gradients are often put in a special category.

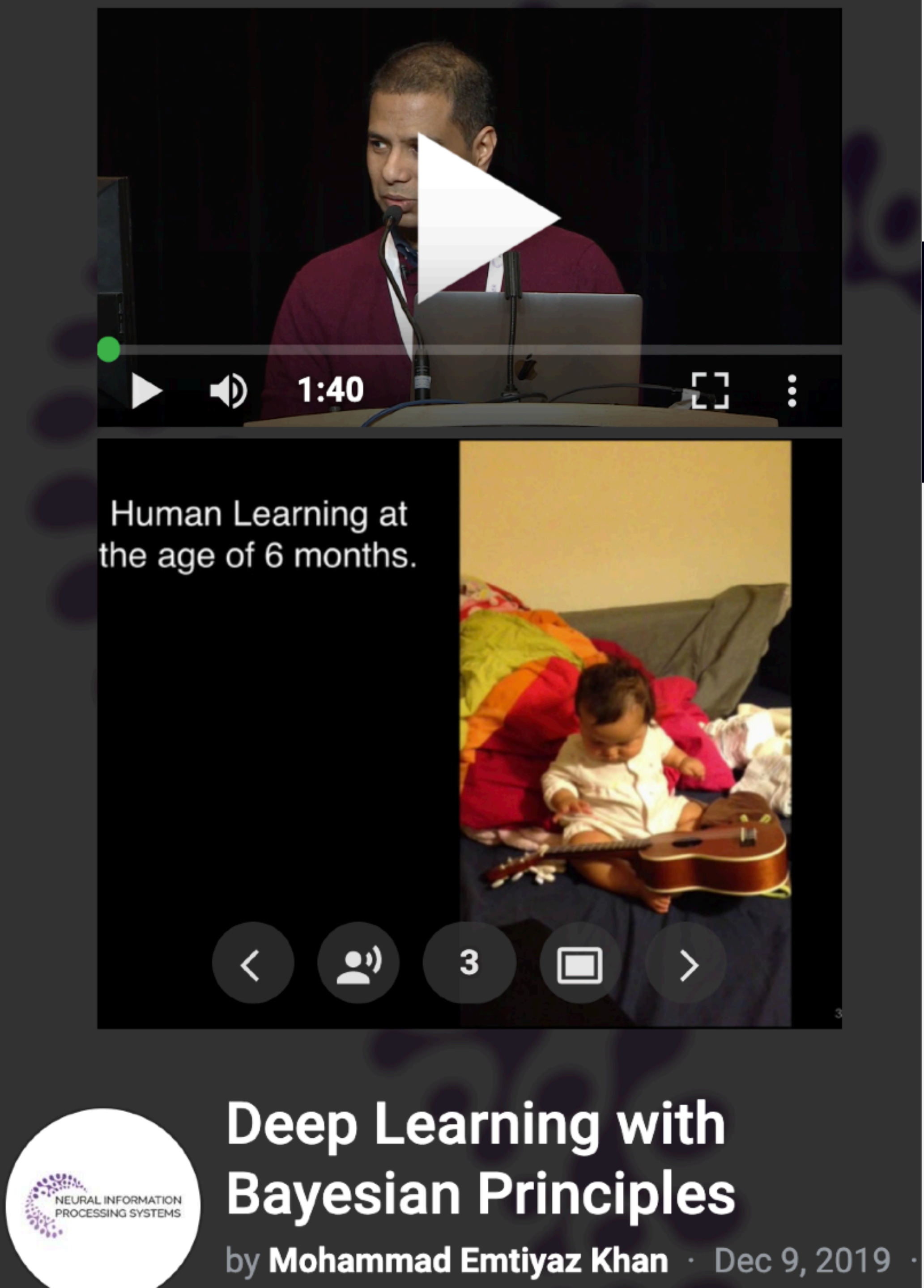
We will show that natural gradients retrieve essential higher-order information about the loss landscape which are then assigned to appropriate natural parameters using Eq. 5. The information-matching is due to the presence of the entropy term there, which is an important quantity for the optimality of Bayes in general [Jaynes, 1982, Zellner, 1988, Littlestone and Warmuth, 1994, Vovk, 1990], and which is generally absent in non-Bayesian formulations (Eq. 1). The entropy term in general leads to exponential-weighting in Bayes' rule. In our context, it gives rise to natural-gradients and, as we will soon see, automatically determines the complexity of the derived algorithm through the complexity of the class of distributions $\mathcal{Q}$, yielding a principled way to develop new algorithms.

Overall, our work demonstrates the importance of natural-gradients and information geometry for algorithm design in ML. This is similar in spirit to Information Geometric Optimization [Ollivier et al., 2017], which focuses on the optimization of black-box, deterministic functions. In contrast, we derive generic learning algorithms by using the same Bayesian principles. The BLR we use is a generalization of the method proposed in Khan and Lin [2017], Khan and Nielsen [2018] specifically for approximate Bayesian inference. Here, we establish it as a general learning rule to derive many old and new learning algorithms, which include both Bayesian and non-Bayesian ones, way beyond its original proposal. We do not claim that these successful algorithms work well because they are derived from the BLR. Rather, we use the BLR to simply unravels the inherent Bayesian nature of these "good" algorithms. In this sense, the BLR can be seen as a variant of Bayes' rule, useful for generic algorithm design.

Principles of “good” algorithms?

- Information Geometry of Bayes
  - To unify/generalize/improve learning-algorithms
  - Optimize for “posterior approximations”
- Bayesian Learning rule (BLR)
  - Derive many algorithms from optimization, deep learning, and Bayesian inference
- Natural Gradients are Everywhere!

Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
Lin et al., Tractable structured natural gradient descent using local parameterizations, ICML 2021

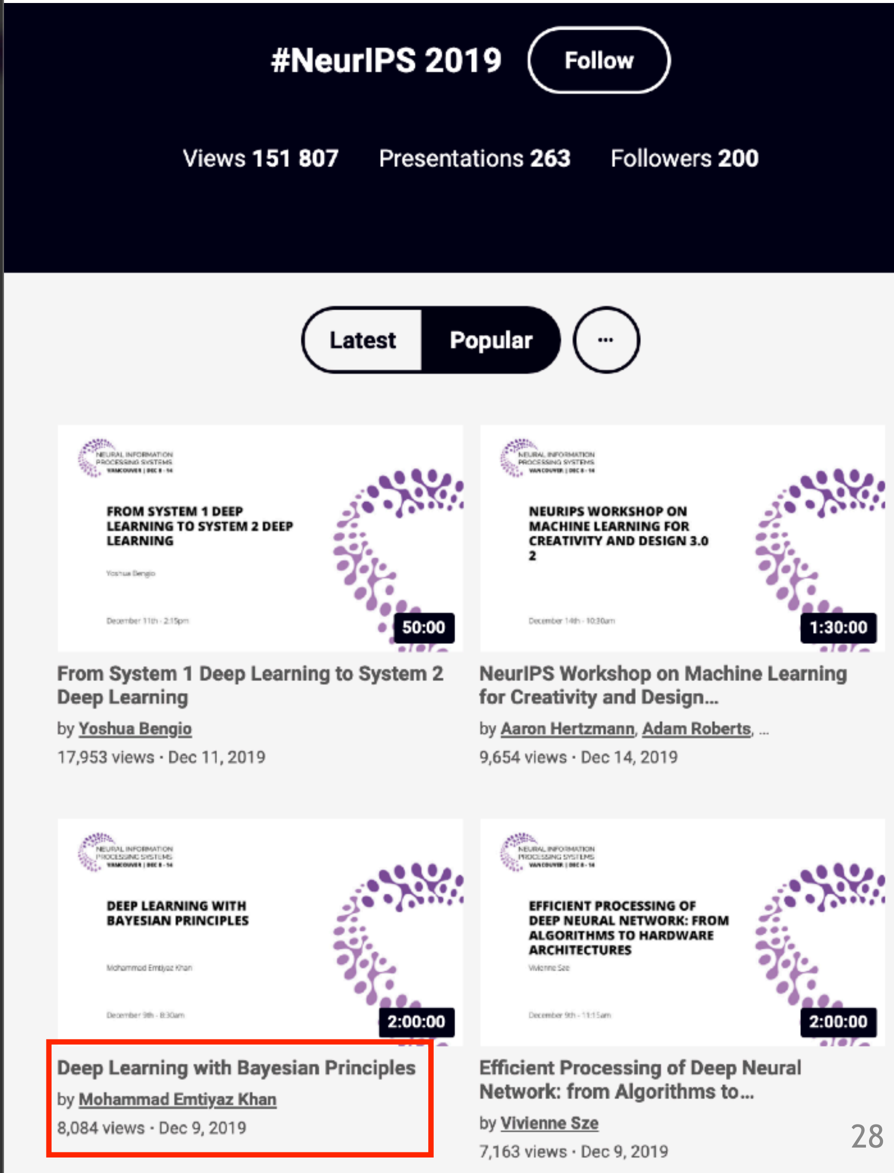


Human Learning at the age of 6 months.

Deep Learning with Bayesian Principles

by Mohammad Emtiyaz Khan · Dec 9, 2019

NeurIPS 2019 Tutorial



#NeurIPS 2019

Views 151 807 Presentations 263 Followers 200

FROM SYSTEM 1 DEEP LEARNING TO SYSTEM 2 DEEP LEARNING

From System 1 Deep Learning to System 2 Deep Learning

by Yoshua Bengio

17,953 views · Dec 11, 2019

NEURIPS WORKSHOP ON MACHINE LEARNING FOR CREATIVITY AND DESIGN 3.0 2

NeurIPS Workshop on Machine Learning for Creativity and Design...

by Aaron Hertzmann, Adam Roberts, ...

9,654 views · Dec 14, 2019

DEEP LEARNING WITH BAYESIAN PRINCIPLES

Deep Learning with Bayesian Principles

by Mohammad Emtiyaz Khan

8,084 views · Dec 9, 2019

EFFICIENT PROCESSING OF DEEP NEURAL NETWORK: FROM ALGORITHMS TO HARDWARE ARCHITECTURES

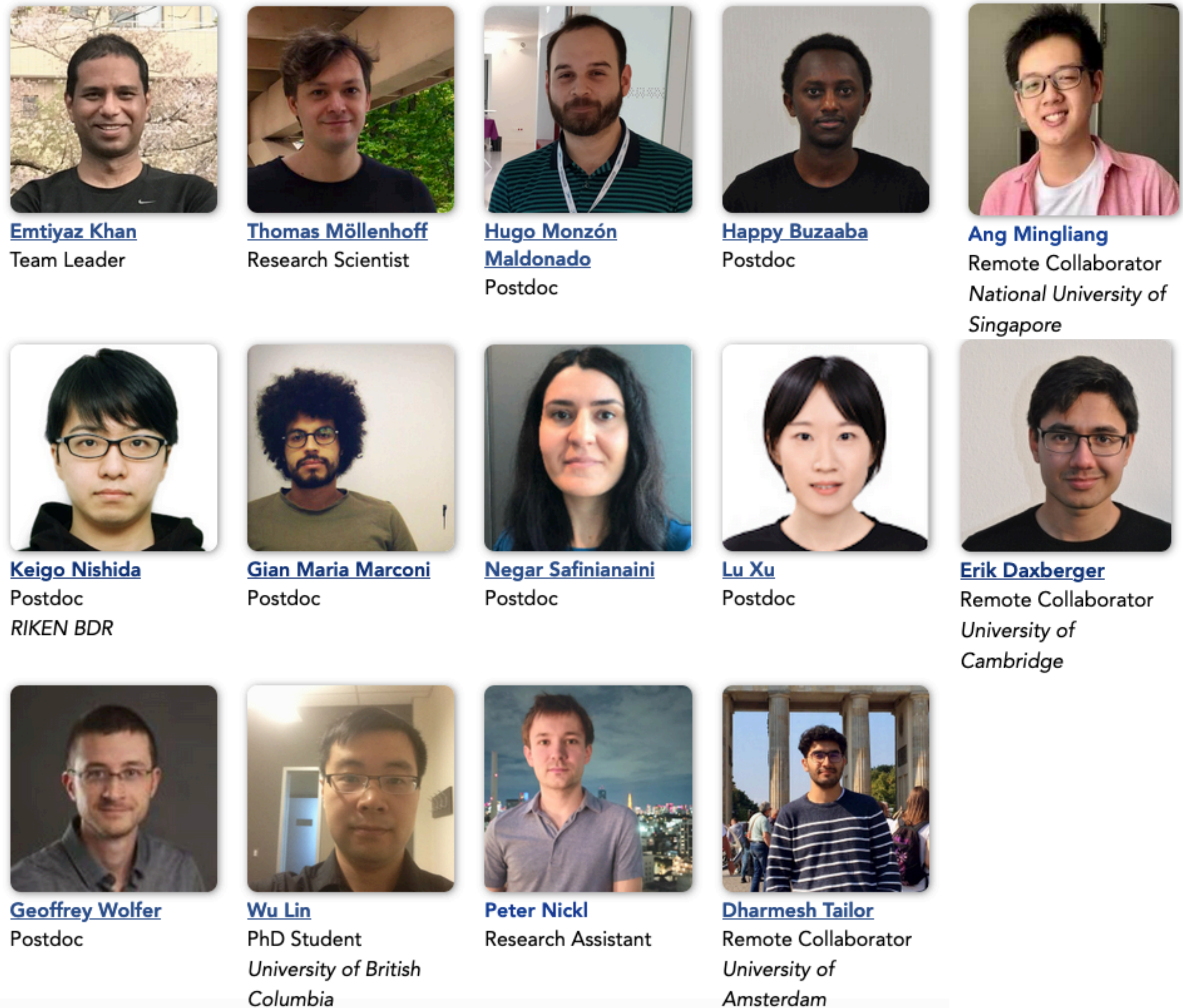
Efficient Processing of Deep Neural Network: from Algorithms to...

by Vivienne Sze

7,163 views · Dec 9, 2019

Approximate Bayesian Inference Team

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



Emtiyaz Khan
Team Leader

Thomas Möllenhoff
Research Scientist

Hugo Monzón Maldonado
Postdoc

Happy Buzaaba
Postdoc

Ang Mingliang
Remote Collaborator
National University of Singapore

Keigo Nishida
Postdoc
RIKEN BDR

Gian Maria Marconi
Postdoc

Negar Safinianaini
Postdoc

Lu Xu
Postdoc

Erik Daxberger
Remote Collaborator
University of Cambridge

Geoffrey Wolfer
Postdoc

Wu Lin
PhD Student
University of British Columbia

Peter Nickl
Research Assistant

Dharmesh Tailor
Remote Collaborator
University of Amsterdam