Regret bounds for online variational inference

Pierre Alquier

RIKEN

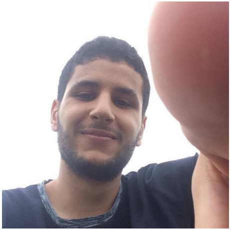
AIP
Center for
Advanced Intelligence Project

Mathematics of Online Decision Making
Simons Institute for the Theory of Computing, Berkeley
Oct. 29, 2020

Co-authors

B.-E. Chérief-Abdellatif, P. Alquier, M. E. Khan (2019). *A regret bound for online variational inference*. 11th Asian Conference on Machine Learning (ACML).

Badr-Eddine Chérief-Abdellatif



UNIVERSITY OF OXFORD DEPARTMENT OF STATISTICS

Emtiyaz Khan

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

RIKEN AIP Center for Advanced Intelligence Project

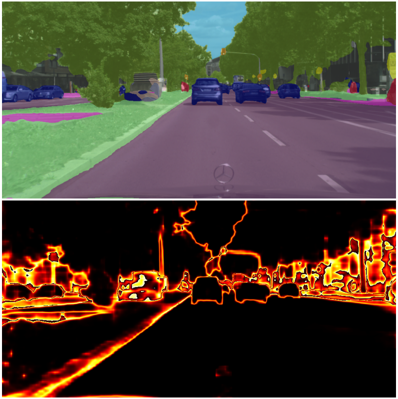
Motivation

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, M. E. Khan (2019). *Practical Deep Learning with Bayesian Principles*. NeurIPS.

Motivation

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, M. E. Khan (2019). *Practical Deep Learning with Bayesian Principles*. NeurIPS.

1. proposes a fast algorithm to approximate the posterior,
2. applies it to train Deep Neural Networks on CIFAR-10, ImageNet ...
3. observation : improved uncertainty quantification.

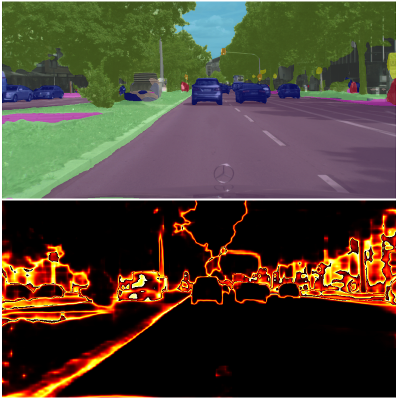


Picture : Roman Bachmann.

Motivation

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, M. E. Khan (2019). *Practical Deep Learning with Bayesian Principles*. NeurIPS.

1. proposes a fast algorithm to approximate the posterior,
2. applies it to train Deep Neural Networks on CIFAR-10, ImageNet ...
3. observation : improved uncertainty quantification.



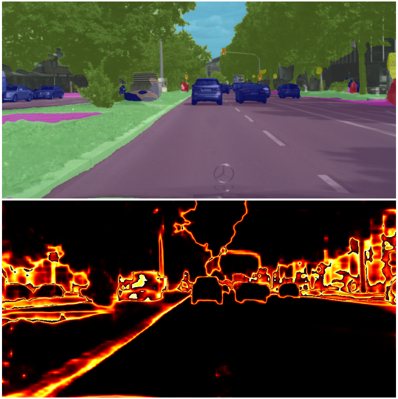
Picture : Roman Bachmann.

Objective : provide a theoretical analysis of this algorithm.

Motivation

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, M. E. Khan (2019). *Practical Deep Learning with Bayesian Principles*. NeurIPS.

1. proposes a fast algorithm to approximate the posterior,
2. applies it to train Deep Neural Networks on CIFAR-10, ImageNet ...
3. observation : improved uncertainty quantification.



Picture : Roman Bachmann.

Objective : provide a theoretical analysis of this algorithm.
First step : simplified versions.

Bayesian inference and variational approximations

(Generalized) Bayesian inference

$$\pi(\theta|x_1, y_1, \ldots, x_n, y_n) \propto \exp\left[-\eta \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\right] \pi(\theta)$$

Bayesian inference and variational approximations

(Generalized) Bayesian inference

$$\pi(\theta|x_1, y_1, \dots, x_n, y_n) \propto \exp\left[-\eta \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\right] \pi(\theta)$$

It is well known that

$$\pi(\cdot|x_1, y_1, \dots, x_n, y_n) = \arg\min_{p} \left\{ \mathbb{E}_{\theta\sim p}\left[\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\right] + \frac{KL(p,\pi)}{\eta} \right\}.$$

Bayesian inference and variational approximations

(Generalized) Bayesian inference

$$\pi(\theta|x_1, y_1, \dots, x_n, y_n) \propto \exp\left[-\eta \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\right] \pi(\theta)$$

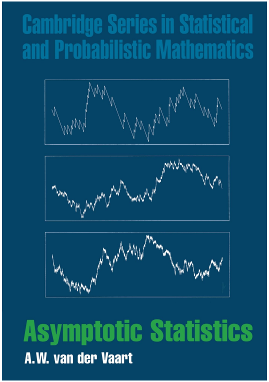
It is well known that

$$\pi(\cdot|x_1, y_1, \dots, x_n, y_n) = \arg\min_{p} \left\{ \mathbb{E}_{\theta\sim p}\left[\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\right] + \frac{KL(p,\pi)}{\eta} \right\}.$$

Variational approximation

$$\pi_n^{\mathrm{approx}}(\theta) := \arg\min_{p\in\mathcal{F}} \left\{ \mathbb{E}_{\theta\sim p}\left[\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\right] + \frac{KL(p,\pi)}{\eta} \right\}.$$

Consistency of Bayesian estimators



- in order to ensure consistency at rate r_n, many conditions including the prior mass condition :

$$\log \pi(B_{r_n}) \geq n r_n, \text{ where}$$

$$B_r = \left\{ \theta : \frac{\sum_{i=1}^n [\ell(f_\theta(x_i), y_i) - \ell(f_{\theta*}(x_i), y_i)]}{n} \leq r \right\}$$

- note that this condition implies, for $p = \pi$ restricted to B_{r_n},

$$\mathbb{E}_{\theta \sim p} \left[\frac{\sum_{i=1}^n \ell(f_\theta(x_i), y_i)}{n} \right] + \frac{KL(p, \pi)}{n} \leq \frac{\sum_{i=1}^n \ell(f_\theta^*(x_i), y_i)}{n} + 2 r_n .$$

Consistency of variational approximations

P. Alquier, J. Ridgway , N. Chopin (2016). On the Properties of Variational Approximations of Gibbs Posteriors. *JMLR*.

P. Alquier & J. Ridgway (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*.

Y. Yang, D. Pati & A. Bhattacharya (2020). α-Variational Inference with Statistical Guarantees. *The Annals of Statistics*.

F. Zhang & C. Gao (2020). Convergence Rates of Variational Posterior Distributions. *The Annals of Statistics*.

These papers show that the variational approximation of the posterior in $\mathcal{F}$ concentrates at the rate r_n if there is $\rho \in \mathcal{F}$ such that

$$\mathbb{E}_{\theta\sim p}\left[\frac{\sum_{i=1}^{n}\ell(f_\theta(x_i),y_i)}{n}\right]+\frac{KL(p,\pi)}{n}\leq\frac{\sum_{i=1}^{n}\ell(f_\theta^*(x_i),y_i)}{n}+2r_n.$$

Question : can this be extended to the online setting ?

Bayesian learning and variational inference (VI)

$$\pi_{t+1}(\theta) := \pi(\theta|x_1, y_1, \ldots, x_t, y_t) \propto \exp\left(-\eta \sum_{s=1}^{t} \ell_s(\theta)\right) \pi(\theta).$$

Bayesian learning and variational inference (VI)

$$\pi_{t+1}(\theta) := \pi(\theta|x_1, y_1, \dots, x_t, y_t) \propto \exp\left(-\eta \sum_{s=1}^{t} \ell_s(\theta)\right) \pi(\theta).$$

Formula for the online update of π_{t+1} :

$$\pi_{t+1}(\theta) \propto \exp\left(-\eta \ell_t(\theta)\right) \pi_t(\theta).$$

Bayesian learning and variational inference (VI)

$$\pi_{t+1}(\theta) := \pi(\theta|x_1, y_1, \dots, x_t, y_t) \propto \exp\left(-\eta \sum_{s=1}^{t} \ell_s(\theta)\right) \pi(\theta).$$

Formula for the online update of π_{t+1} :

$$\pi_{t+1}(\theta) \propto \exp\left(-\eta \ell_t(\theta)\right) \pi_t(\theta).$$

Q1 : can we similarly define a sequential update for a variational approximation ?

Regret bounds for Bayesian inference

Theorem (classical result) for bounded loss $\ell \leq B$

Bayes update leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_q \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q}[\ell_t(\theta)] + \frac{\eta B^2 T}{8} + \frac{KL(q,\pi)}{\eta} \right\}.$$

Regret bounds for Bayesian inference

Theorem (classical result) for bounded loss $\ell \leq B$

Bayes update leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_q \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q}[\ell_t(\theta)] + \frac{\eta B^2 T}{8} + \frac{KL(q,\pi)}{\eta} \right\}.$$

Under the prior mass condition with $r_T = \sqrt{\frac{\log T}{T}}$ and $\eta \sim r_T$,

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_\theta \sum_{t=1}^{T} \ell_t(\theta) + \mathcal{O}(\sqrt{T\log(T)}).$$

Regret bounds for Bayesian inference

Theorem (classical result) for bounded loss $\ell \leq B$

Bayes update leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_q \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q}[\ell_t(\theta)] + \frac{\eta B^2 T}{8} + \frac{KL(q,\pi)}{\eta} \right\}.$$

Under the prior mass condition with $r_T = \sqrt{\frac{\log T}{T}}$ and $\eta \sim r_T$,

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_\theta \sum_{t=1}^{T} \ell_t(\theta) + \mathcal{O}(\sqrt{T\log(T)}).$$

Q2 : can we derive similar results for online VI ?

Online gradient algorithm (OGA)

Given

- a set of predictors $\{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, e.g $f_\theta(x) = \langle \theta, x \rangle$,
- an initial guess θ_1,

$$\hat{y}_t = f_{\theta_t}(x_t) \quad \text{and} \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \ell(f_{\theta_t}(x_t), y_t).$$

Online gradient algorithm (OGA)

Given

- a set of predictors $\{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, e.g $f_\theta(x) = \langle \theta, x \rangle$,
- an initial guess θ_1,

$$\hat{y}_t = f_{\theta_t}(x_t) \quad \text{and} \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \ell_t(\theta_t).$$

Online gradient algorithm (OGA)

Given

- a set of predictors $\{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, e.g $f_\theta(x) = \langle \theta, x \rangle$,
- an initial guess θ_1,

$$\hat{y}_t = f_{\theta_t}(x_t) \quad \text{and} \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \ell_t(\theta_t).$$

Note that θ_{t+1} can be obtained by :

1. $$\min_\theta \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$
2. $$\min_\theta \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\}.$$

Two options for online VI

Parametric VI : $\mathcal{F} = \{q_\mu, \mu \in M\}$.

Two options for online VI

Parametric VI : $\mathcal{F} = \{q_\mu, \mu \in M\}$.

1. Sequential Variational Approximation (SVA) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$

2. Streaming Variational Bayes (SVB) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\},$$

Two options for online VI

Parametric VI : $\mathcal{F} = \{q_\mu, \mu \in M\}$.

1. Sequential Variational Approximation (SVA) :

$$\theta_{t+1} = \arg\min_\theta \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$

$$\mu_{t+1} = \arg\min_\mu \left\{ \left\langle \mu, \sum_{s=1}^{t} \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_s}} [\ell_s(\theta)] \right\rangle + \frac{KL(q_\mu, \pi)}{\eta} \right\}.$$

2. Streaming Variational Bayes (SVB) :

$$\theta_{t+1} = \arg\min_\theta \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\},$$

Two options for online VI

Parametric VI : $\mathcal{F} = \{q_\mu, \mu \in M\}$.

1. Sequential Variational Approximation (SVA) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$

$$\mu_{t+1} = \arg\min_{\mu} \left\{ \left\langle \mu, \sum_{s=1}^{t} \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_s}}[\ell_s(\theta)] \right\rangle + \frac{KL(q_\mu, \pi)}{\eta} \right\}.$$

2. Streaming Variational Bayes (SVB) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\},$$

$$\mu_{t+1} = \arg\min_{\mu} \left\{ \left\langle \mu, \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \right\rangle + \frac{KL(q_\mu, q_{\mu_t})}{\eta} \right\}.$$

SVA & SVB are tractable, and not equivalent

Example : Gaussian prior $\theta \sim \pi = \mathcal{N}(0, s^2 I)$ and mean-field Gaussian approximation, $\mu = (m, \sigma)$.

$$
\begin{aligned}
\mathrm{SVA} : m_{t+1} &\leftarrow m_t - \eta s^2 \bar{g}_{m_t}, \qquad g_{t+1} \leftarrow g_t + \bar{g}_{\sigma_t}, \\
\sigma_{t+1} &\leftarrow h\left(\eta s g_{t+1}\right) s, \\
\mathrm{SVB} : m_{t+1} &\leftarrow m_t - \eta \sigma_t^2 \bar{g}_{m_t}, \\
\sigma_{t+1} &\leftarrow \sigma_t h\left(\eta \sigma_t \bar{g}_{\sigma_t}\right)
\end{aligned}
$$

where $h(x) := \sqrt{1 + x^2} - x$ is applied componentwise, as well as the multiplication of two vectors, and

$$
\begin{aligned}
\bar{g}_{m_t} &= \frac{\partial}{\partial m} \mathbb{E}_{\theta \sim \pi_{m_t, \sigma_t}}[\ell_t(\theta)], \\
\bar{g}_{\sigma_t} &= \frac{\partial}{\partial \sigma} \mathbb{E}_{\theta \sim \pi_{m_t, \sigma_t}}[\ell_t(\theta)].
\end{aligned}
$$

Theoretical analysis of SVA

Theorem 1

Under convexity and L-Lipschitz assumption on the loss, under α-strong convexity assumption on the KL term, SVA leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\mu \in M} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu}}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{KL(q_\mu, \pi)}{\eta} \right\}.$$

Theoretical analysis of SVA

Theorem 1

Under convexity and L-Lipschitz assumption on the loss, under α-strong convexity assumption on the KL term, SVA leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\mu \in M} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu}}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{KL(q_\mu, \pi)}{\eta} \right\}.$$

Application to Gaussian approximation leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + (1 + o(1)) \frac{2L}{\alpha} \sqrt{dT \log(T)}.$$

Comments on the assumptions

The assumptions :

1. $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex ?

Comments on the assumptions

The assumptions :

1. $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex ?

Proposition

Assume $\theta \mapsto \ell_t(\theta)$ is $L/2$-Lipschitz and convex, and $\mu = (m, \Sigma)$ is a location scale parameter, then : satisfied.

Comments on the assumptions

The assumptions :

1. $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex ?

Proposition

Assume $\theta \mapsto \ell_t(\theta)$ is $L/2$-Lipschitz and convex, and $\mu = (m, \Sigma)$ is a location scale parameter, then : satisfied.

Proof : Lipschitz : in our paper ; convex :

J. Domke (2019). *Provable smoothness guarantees for black-box variational inference.* NeurIPS 2019.

Comments on the assumptions

The assumptions :

1. $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex ?

Proposition

Assume $\theta \mapsto \ell_t(\theta)$ is $L/2$-Lipschitz and convex, and $\mu = (m, \Sigma)$ is a location scale parameter, then : satisfied.

Proof : Lipschitz : in our paper ; convex :

J. Domke (2019). *Provable smoothness guarantees for black-box variational inference.* NeurIPS 2019.

2. $\mu \mapsto KL(q_\mu, \pi)$ is α-strongly convex ?

Comments on the assumptions

The assumptions :

1. $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex ?

Proposition

Assume $\theta \mapsto \ell_t(\theta)$ is $L/2$-Lipschitz and convex, and $\mu = (m, \Sigma)$ is a location scale parameter, then : satisfied.

Proof : Lipschitz : in our paper ; convex :

J. Domke (2019). *Provable smoothness guarantees for black-box variational inference.* NeurIPS 2019.

2. $\mu \mapsto KL(q_\mu, \pi)$ is α-strongly convex ?

$\rightarrow$ True for many examples, for example when q_μ and π are Gaussian (with upper-bounded variance).

Theoretical analysis of SVB

Theorem 2

Using Gaussian approximations, assuming the loss is convex, L-Lipschitz and the parameter space bounded (diameter $= D$), SVB with adequate η leads to

$$\sum_{t=1}^{T} \ell_t\left(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\right) \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + DL\sqrt{2T}.$$

Theoretical analysis of SVB

Theorem 2

Using Gaussian approximations, assuming the loss is convex, L-Lipschitz and the parameter space bounded (diameter $= D$), SVB with adequate η leads to

$$\sum_{t=1}^{T} \ell_t\left(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\right) \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + DL\sqrt{2T}.$$

If, moreover, the loss is H-strongly convex,

$$\sum_{t=1}^{T} \ell_t\left(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\right) \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + \frac{L^2(1+\log(T))}{H}.$$

Test on a simulated dataset

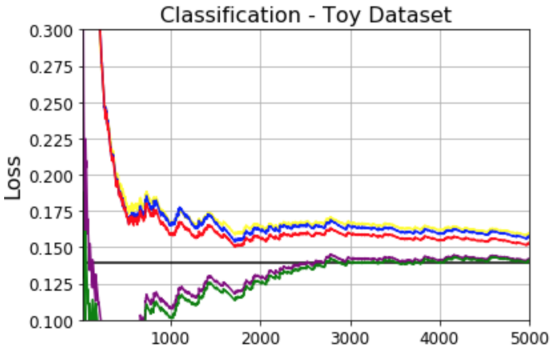


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Breast dataset

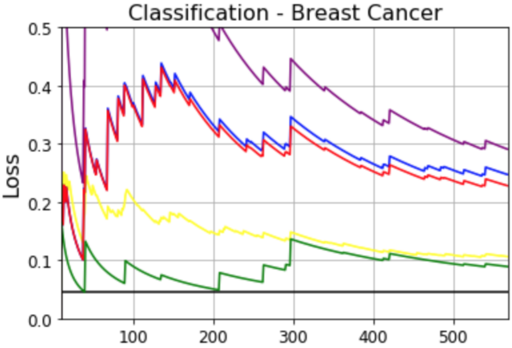


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Open questions

Open questions

1. Analysis of SVB in the general case.

Open questions

1. Analysis of SVB in the general case.
2. Analysis of the uncertainty quantification.

Open questions

1. Analysis of SVB in the general case.
2. Analysis of the uncertainty quantification.
3. NGVI is the next step in going closer to algorithms used to train Neural Networks with Bayesian principles. But being based on a different parametrization, it does not satisfy our convexity assumption...

Open questions

1. Analysis of SVB in the general case.
2. Analysis of the uncertainty quantification.
3. NGVI is the next step in going closer to algorithms used to train Neural Networks with Bayesian principles. But being based on a different parametrization, it does not satisfy our convexity assumption...

   Uses exponential family approximations $\{q_\mu, \mu \in M\}$ where m is the mean parameter. Denoting λ the natural parameter (with $\lambda = F(\mu)$),

$$\lambda_{t+1} = (1-\rho)\lambda_t + \rho \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_t}} \left[\ell_t(\theta)\right],$$

M. E. Khan, D. Nielsen (2018). *Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models*. ISITA.

Thank you !