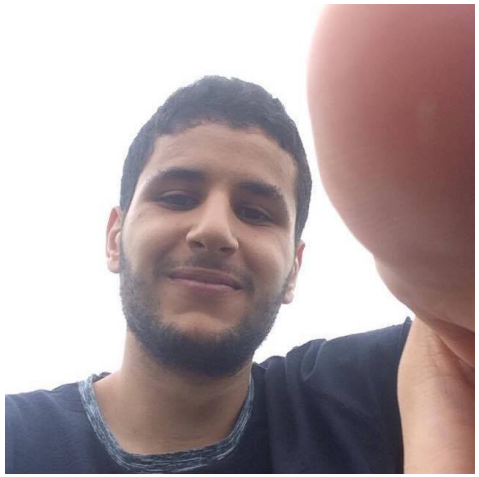
Regret bounds for online variational inference

Pierre Alquier

ACML – Nagoya, Nov. 18, 2019

Co-authors

Badr-Eddine Chérief-Abdellatif



ENSAE

IP PARIS

Emtiyaz Khan

RIKEN AIP

Center for Advanced Intelligence Project

Approximate Bayesian Inference team

https://emtiyaz.github.io/

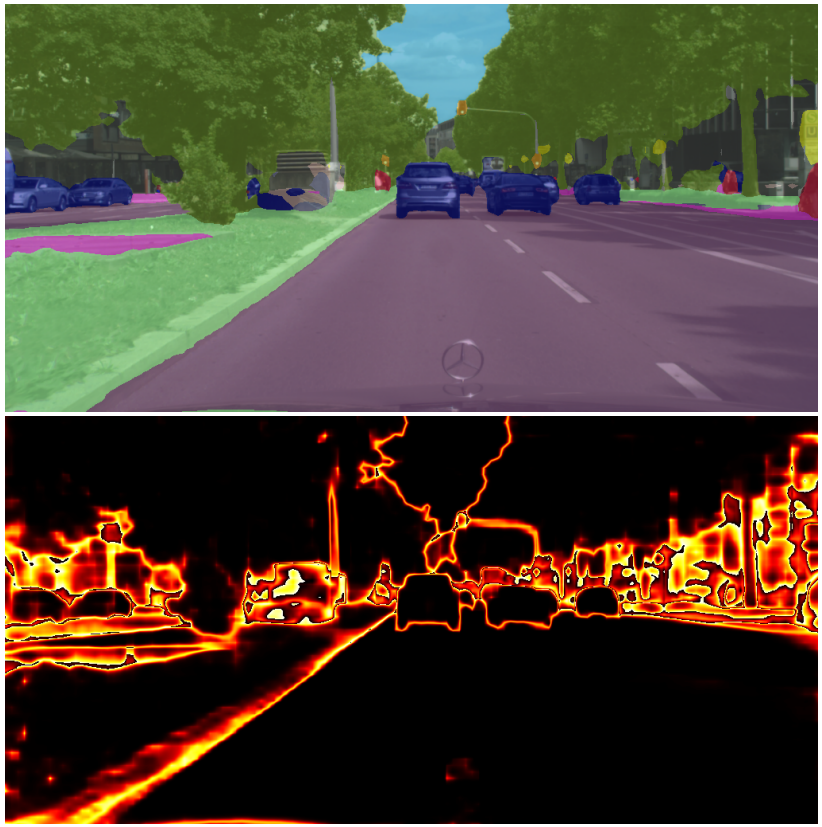
Motivation

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.

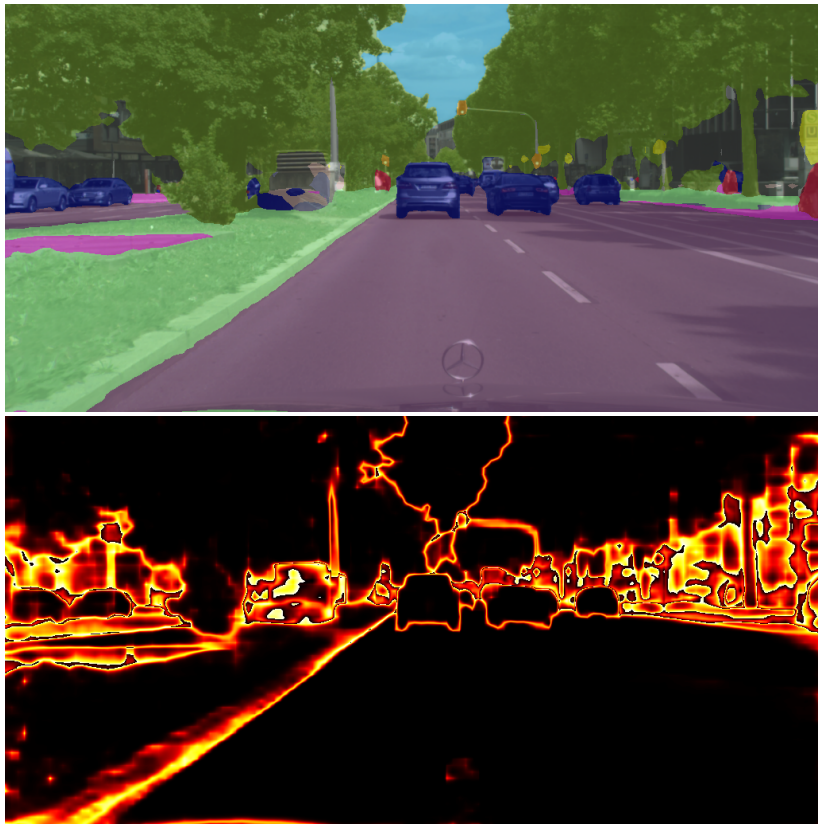


Picture : Roman Bachmann.

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.



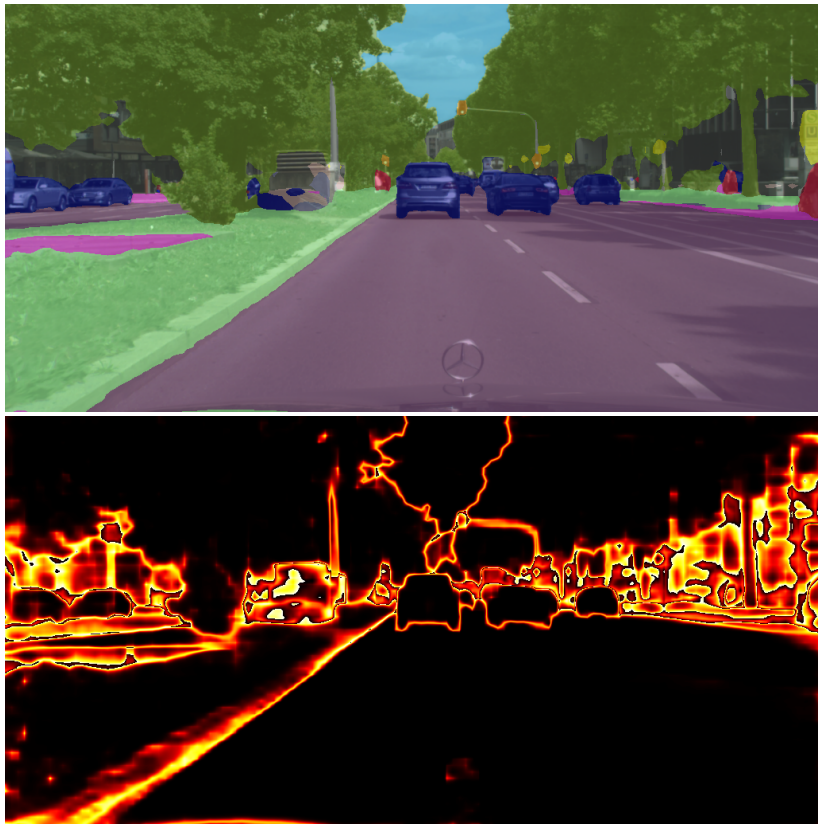
Picture : Roman Bachmann.

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.

The sequential prediction problem

Sequential prediction problem

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given

The sequential prediction problem

Sequential prediction problem

1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$

The sequential prediction problem

Sequential prediction problem

1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$

The sequential prediction problem

Sequential prediction problem

1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2.
   1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3.
   1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

Objective :

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

Objective : make sure that we learn to predict well **as soon as possible**.

The sequential prediction problem

Sequential prediction problem

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

Objective : make sure that we learn to predict well **as soon as possible**. Keep

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$$

as small as possible.

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. $\min_\theta \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\}.$

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)

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

Not tractable in general, leading to variational approximations :

$$\begin{aligned}\tilde{\pi}_{t+1}(\theta) &= \arg\min_{q \in \mathcal{F}} KL(q, \pi_{t+1}) \\ &= \arg\min_{q \in \mathcal{F}} \left\{ \mathbb{E}_{\theta \sim q} \left[\sum_{s=1}^{t} \ell_s(\theta) \right] + \frac{KL(q, \pi)}{\eta} \right\}.\end{aligned}$$

Bayesian learning and variational inference (VI)

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

Not tractable in general, leading to variational approximations :

$$\begin{aligned}\tilde{\pi}_{t+1}(\theta) &= \arg\min_{q \in \mathcal{F}} KL(q, \pi_{t+1}) \\ &= \arg\min_{q \in \mathcal{F}} \left\{ \mathbb{E}_{\theta \sim q} \left[\sum_{s=1}^{t} \ell_s(\theta) \right] + \frac{KL(q, \pi)}{\eta} \right\}.\end{aligned}$$

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)

Under the assumption that the loss is bounded by B, the Bayesian 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)

Under the assumption that the loss is bounded by B, the Bayesian update leads to

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

Derivation of the infimum and $\eta \sim \sqrt{T}$ "usually" leads to

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

Regret bounds for Bayesian inference

Theorem (classical result)

Under the assumption that the loss is bounded by B, the Bayesian update leads to

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

Derivation of the infimum and $\eta \sim \sqrt{T}$ "usually" leads to

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

Q2 : can we derive similar results for online VI ?

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, $\mu = (m, \sigma)$.

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

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

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

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

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

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

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

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

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

Test on a simulated dataset

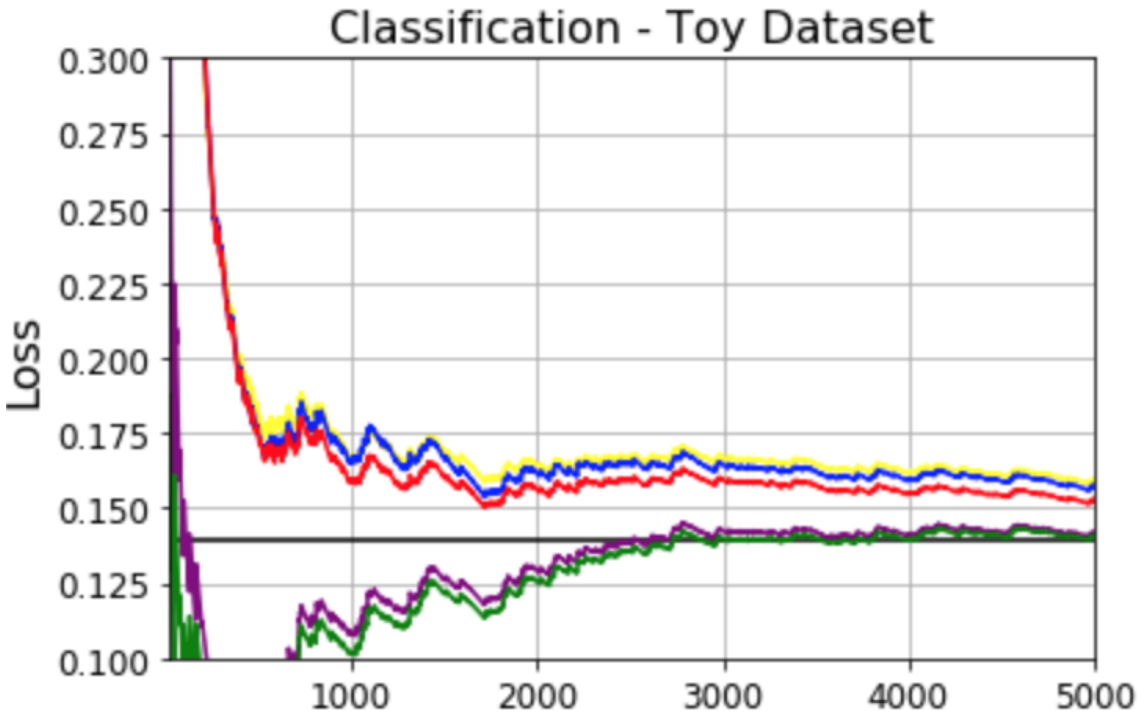


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Breast dataset

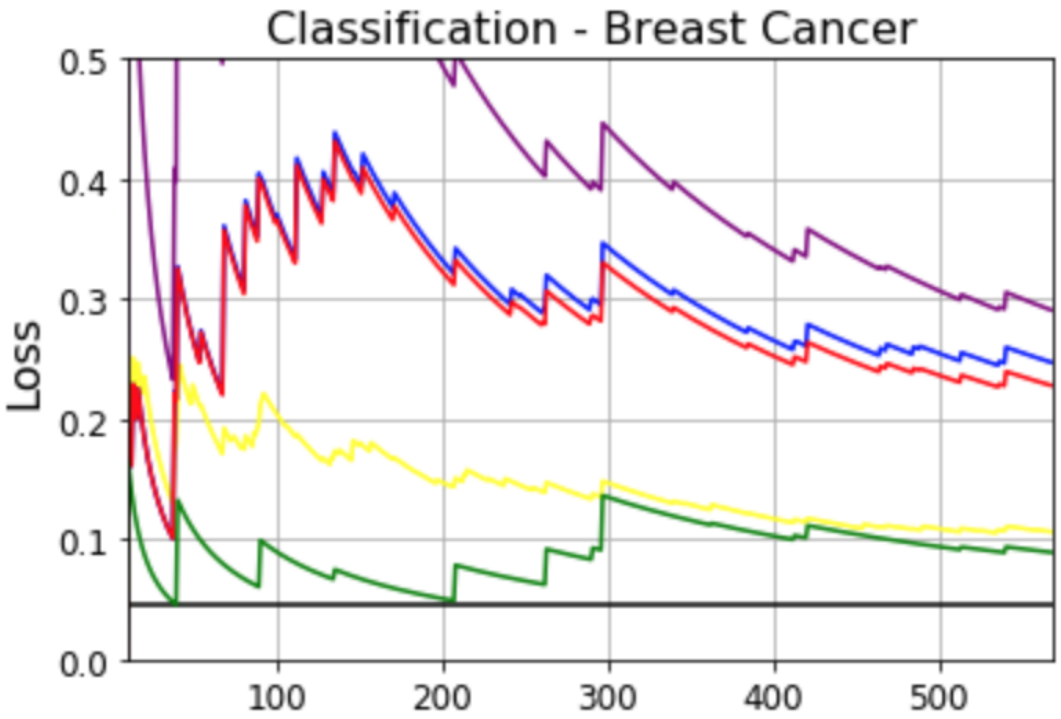


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

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 !