A Regret Bound for Online Variational Inference

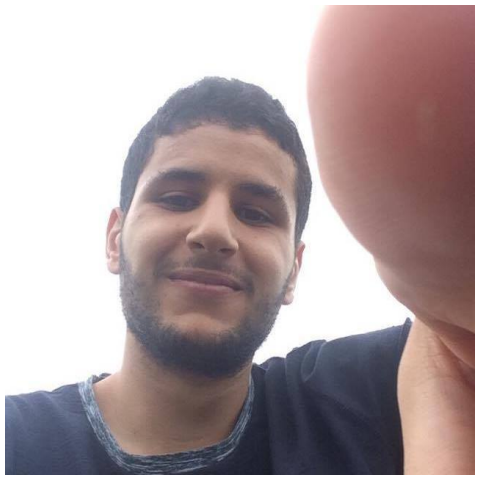
Pierre Alquier

RIKEN AIP
Center for
Advanced Intelligence Project

Heilbronn Institute – University of Bristol – Dec. 5, 2019

B.-E. Chérief-Abdellatif, P. Alquier, M. E. Khan (2019). *A regret bound for online variational inference*. ACML.

Badr-Eddine Chérief-Abdellatif



ENSAE

IP PARIS

Emtiyaz Khan

RIKEN

AIP Center for Advanced Intelligence Project

Approximate Bayesian Inference team

https://emtiyaz.github.io/

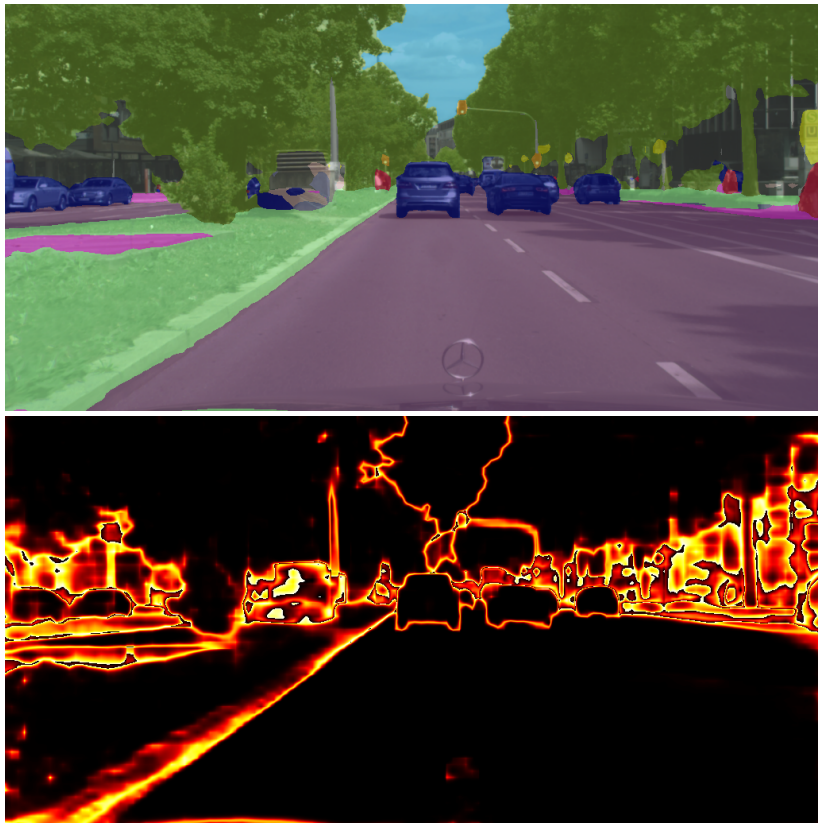
Motivation

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.

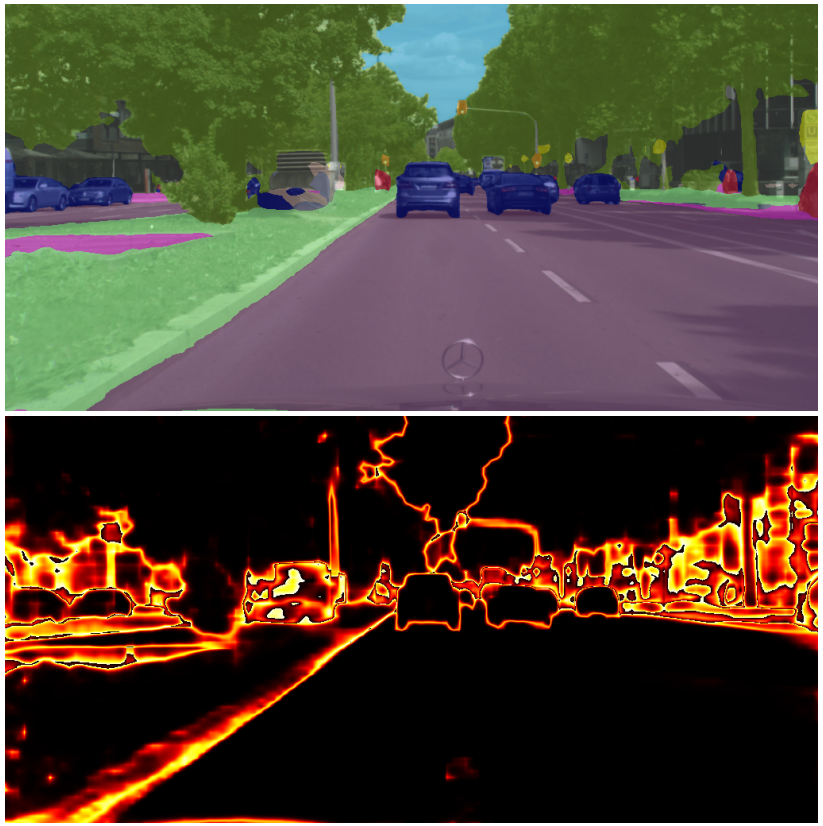
1. proposes a fast algorithm to approximate the posterior,
2. applies it to train Deep Neural Networks on CIFAR-10, ImageNet ...
3. observation : improved uncertainty quantification.



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.

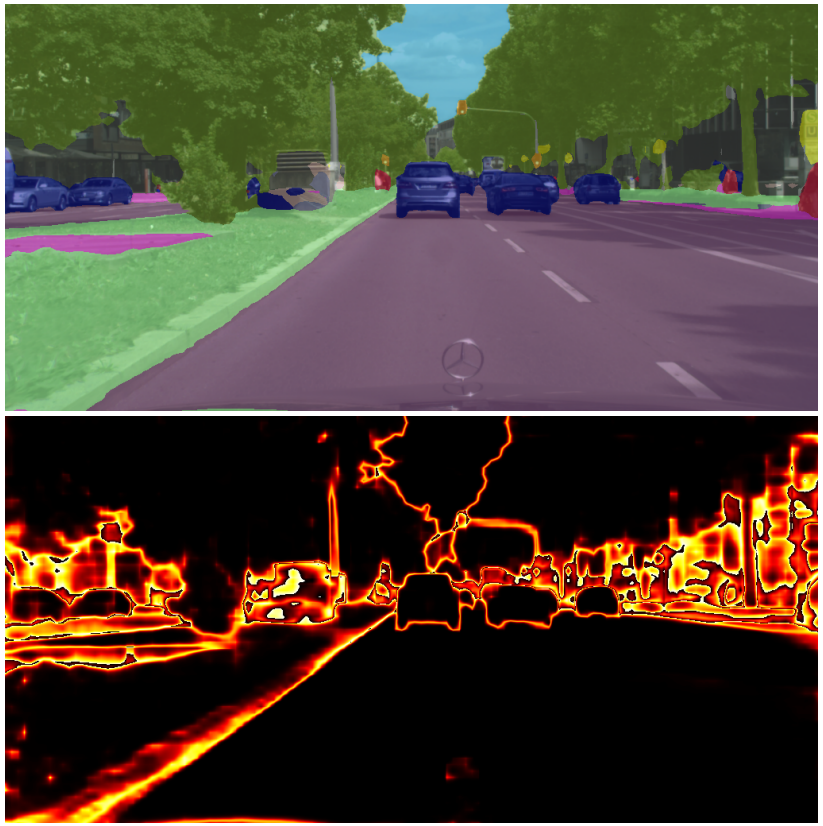


Objective : provide a theoretical analysis of this algorithm.

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.



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, without unrealistic assumptions on the data.

What can we expect ?

As an example, consider for one minute a classification problem, that is :

$$y_t \in \{0, 1\} \text{ and } \ell(\hat{y}_t, y_t) = \mathbf{1}_{\{y_t \neq \hat{y}_t\}}.$$

What can we expect ?

As an example, consider for one minute a classification problem, that is :

$$y_t \in \{0, 1\} \text{ and } \ell(\hat{y}_t, y_t) = \mathbf{1}_{\{y_t \neq \hat{y}_t\}}.$$

1. In the worst case scenario, y_t generated by an omniscient opponent by : $y_t = 1 - \hat{y}_t$. Then

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

What can we expect ?

As an example, consider for one minute a classification problem, that is :

$$y_t \in \{0, 1\} \text{ and } \ell(\hat{y}_t, y_t) = \mathbf{1}_{\{y_t \neq \hat{y}_t\}}.$$

1. In the worst case scenario, y_t generated by an omniscient opponent by : $y_t = 1 - \hat{y}_t$. Then

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

2. On the other hand, many real world phenomena can be "quite well" described by models. These models allow to do "sensible" predictions.

What can we expect ?

As an example, consider for one minute a classification problem, that is :

$$y_t \in \{0, 1\} \text{ and } \ell(\hat{y}_t, y_t) = \mathbf{1}_{\{y_t \neq \hat{y}_t\}}.$$

1. In the worst case scenario, y_t generated by an omniscient opponent by : $y_t = 1 - \hat{y}_t$. Then

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

2. On the other hand, many real world phenomena can be "quite well" described by models. These models allow to do "sensible" predictions.
   The extreme case would be the constraint $y_t = f(x_t)$, where $f \in \mathcal{F}$ for a known class $\mathcal{F}$ – the *realizable case*.
   Let's study it as a toy example when $\mathcal{F}$ is finite.

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t = f_{i(t)}(x_t)$, observe y_t,

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t = f_{i(t)}(x_t)$, observe y_t,
2. update $\begin{cases} C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}, \\ i(t+1) = \min C(t+1). \end{cases}$

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t = f_{i(t)}(x_t)$, observe y_t,
2. update $\begin{cases} C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}, \\ i(t+1) = \min C(t+1). \end{cases}$

Theorem

$$\forall T, \ \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq M - 1.$$

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t =$ "majority vote in $C(t)$", observe y_t,

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t =$ "majority vote in $C(t)$", observe y_t,
2. update $C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}$.

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t =$ "majority vote in $C(t)$", observe y_t,
2. update $C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}$.

Theorem

$$\forall T, \ \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \log_2(M).$$

A feasible objective

Two extremes :

- playing against the devil $y_t = 1 - \hat{y}_t$,
- assuming a true, exact model $\mathcal{F}$.

A feasible objective

Two extremes :

- playing against the devil $y_t = 1 - \hat{y}_t$,
- assuming a true, exact model $\mathcal{F}$.

Real-life is somewhere in between !

A feasible objective

Two extremes :

- playing against the devil $y_t = 1 - \hat{y}_t$,
- assuming a true, exact model $\mathcal{F}$.

Real-life is somewhere in between !

Objective

Strategy such that

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \underbrace{\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)}_{\substack{= T \text{ in the worst case (devil),} \\ = 0 \text{ in the ideal case (true model),} \\ \text{almost always in between.}}} + \underbrace{B(T)}_{\text{as small as possible ! !}} .$$

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

Guarantees for OGA

Theorem - classical, see Shalev-Shwarz 2012

Assume ℓ is convex and L-Lipschitz. Then

$$\sum_{t=1}^{T} \ell_t(\theta_t) \leq \inf_{\theta \in \Theta} \left\{ \sum_{t=1}^{T} \ell_t(\theta) + \eta T L^2 + \frac{\|\theta\|^2}{2\eta} \right\}.$$

S. Shalev-Shwartz (2012). *Online learning and online convex optimization*. Foundations and Trends in Machine Learning.

Guarantees for OGA

Theorem - classical, see Shalev-Shwarz 2012

Assume ℓ is convex and L-Lipschitz. Then

$$\sum_{t=1}^{T} \ell_t(\theta_t) \leq \inf_{\theta \in \Theta} \left\{ \sum_{t=1}^{T} \ell_t(\theta) + \eta T L^2 + \frac{\|\theta\|^2}{2\eta} \right\}.$$

S. Shalev-Shwartz (2012). *Online learning and online convex optimization*. Foundations and Trends in Machine Learning.

$$\sum_{t=1}^{T} \ell_t(\theta_t) \leq \inf_{\|\theta\| \leq B} \sum_{t=1}^{T} \ell_t(\theta) + \eta T L^2 + \frac{B^2}{2\eta}.$$

Guarantees for OGA

Theorem - classical, see Shalev-Shwarz 2012

Assume ℓ is convex and L-Lipschitz. Then

$$\sum_{t=1}^{T} \ell_t(\theta_t) \leq \inf_{\theta \in \Theta} \left\{ \sum_{t=1}^{T} \ell_t(\theta) + \eta T L^2 + \frac{\|\theta\|^2}{2\eta} \right\}.$$

S. Shalev-Shwartz (2012). *Online learning and online convex optimization*. Foundations and Trends in Machine Learning.

$$\sum_{t=1}^{T} \ell_t(\theta_t) \leq \inf_{\|\theta\| \leq B} \sum_{t=1}^{T} \ell_t(\theta) + \eta T L^2 + \frac{B^2}{2\eta}.$$

The choice $\eta = B/(L\sqrt{2T})$ leads to

$$\sum_{t=1}^{T} \ell_t(\theta_t) \leq \inf_{\|\theta\| \leq B} \sum_{t=1}^{T} \ell_t(\theta) + BL\sqrt{2T}.$$

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 similarty define a sequential update for a variational approximation ?

Regret bounds for Bayesian inference

Theorem - classical, see Cesa-Bianchi & Lugosi 06

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

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)]$$

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

N. Cesa-Bianchi & G. Lugosi (2006). *Prediction, learning and games*. Cambridge.

Explicit regret bounds

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)]$$

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

Explicit regret bounds

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)]$$

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

Example 1 : finite set $\mathrm{card}(\Theta) = M$, prior π uniform.
Consider $q = \delta_\vartheta$. Then :

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_{\vartheta} \sum_{t=1}^{T} \ell_t(\vartheta) + \frac{\eta B^2 T}{8} + \frac{\log(M)}{\eta}.$$

Explicit regret bounds

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)]$$

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

Example 1 : finite set $\mathrm{card}(\Theta) = M$, prior π uniform. Consider $q = \delta_\vartheta$. Then :

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_\vartheta \sum_{t=1}^{T} \ell_t(\vartheta) + \frac{\eta B^2 T}{8} + \frac{\log(M)}{\eta}.$$

The choice $\eta = \sqrt{8\log(M)/(TB^2)}$ leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}[\ell_t(\theta)] \leq \inf_\vartheta \sum_{t=1}^{T} \ell_t(\vartheta) + B\sqrt{\frac{T\log(M)}{2}}.$$

Explicit regret bounds (2/2)

Example 2 : When $\Theta = \mathbb{R}^d$, using Gaussian priors and Gaussian q, and $\eta \sim \sqrt{T}$ "usually" leads to

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

Explicit regret bounds (2/2)

Example 2 : When $\Theta = \mathbb{R}^d$, using Gaussian priors and Gaussian q, and $\eta \sim \sqrt{T}$ "usually" leads to

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

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

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

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

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

Assume

1. $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex,
2. $\mu \mapsto KL(q_\mu, \pi)$ is α-strongly convex, then

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

Theoretical analysis of SVA

Theorem 1

Assume

1. $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex,
2. $\mu \mapsto KL(q_\mu, \pi)$ is α-strongly convex, then

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

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.*. Preprint arXiv.

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.*. Preprint arXiv.

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.*. Preprint arXiv.

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 lower and upper-bounded variance).

Theoretical analysis of SVB

Theorem 2

1. We use Gaussian approximation q_μ with μ in a bounded set (diameter $= D$),
2. $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex,
3. $\mu \mapsto KL(q_\mu, \pi)$ is α-strongly convex, then

SVB with adequate η leads to

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

Theoretical analysis of SVB

Theorem 2

1. We use Gaussian approximation q_μ with μ in a bounded set (diameter $= D$),
2. $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$ is L-Lipschitz and convex,
3. $\mu \mapsto KL(q_\mu, \pi)$ is α-strongly convex, then

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

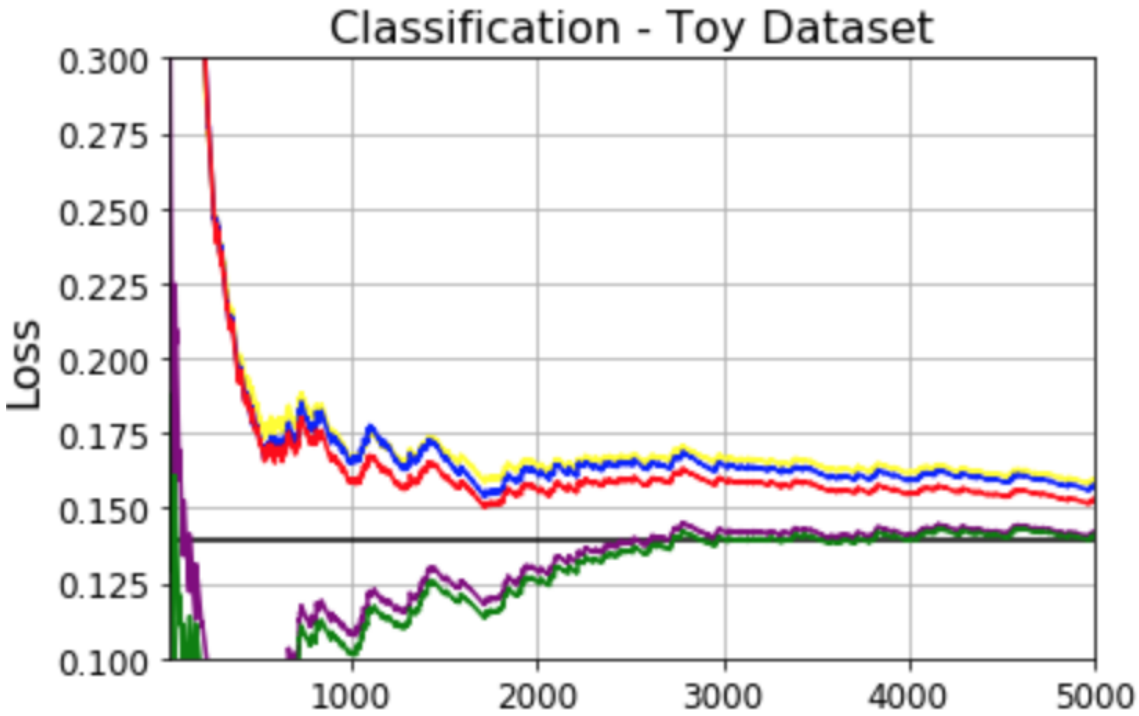


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Breast dataset

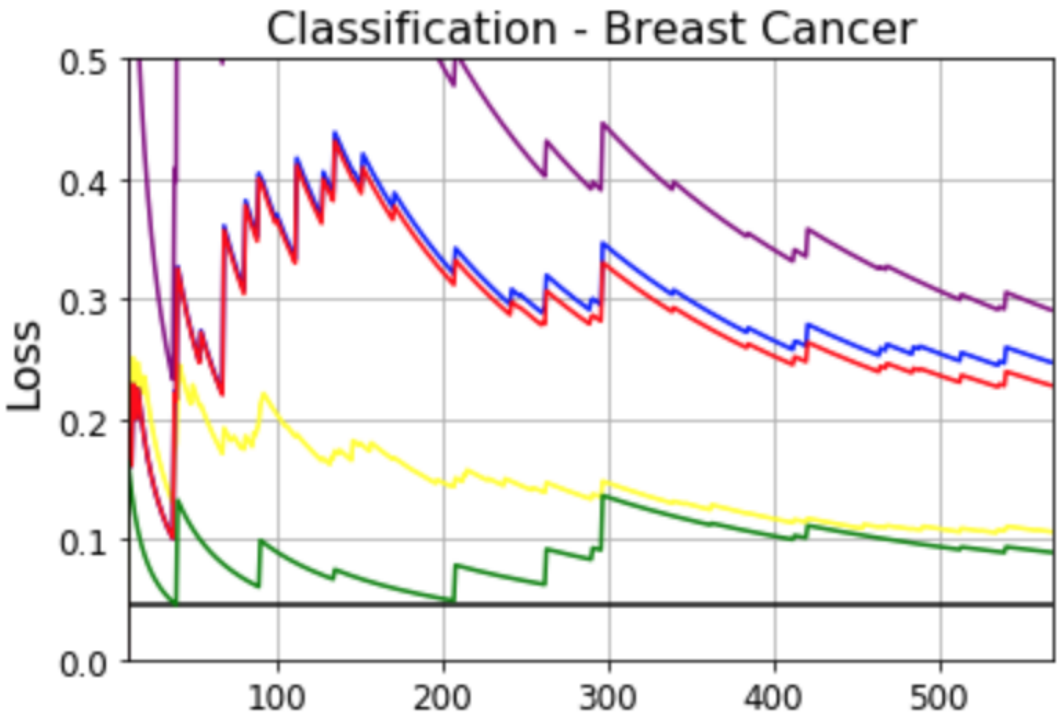


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

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 it does not satisfy our assumptions...

Assume we use approximations in the exponential family

$$q_\lambda(\theta) = a(\theta)b(\lambda)\exp\left[\langle \lambda, S(\theta)\rangle\right].$$

NGVI

Assume we use approximations in the exponential family

$$q_\lambda(\theta) = a(\theta)b(\lambda)\exp\left[\langle \lambda, S(\theta)\rangle\right].$$

We can re-parametrize this set with $\mu = \mathbb{E}_{\theta\sim q_\lambda}[T(\theta)] = F(\lambda)$.

NGVI

Assume we use approximations in the exponential family

$$q_\lambda(\theta) = a(\theta)b(\lambda)\exp\left[\langle \lambda, S(\theta)\rangle\right].$$

We can re-parametrize this set with $\mu = \mathbb{E}_{\theta\sim q_\lambda}[T(\theta)] = F(\lambda)$. The NGVI algorithm is given by

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

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

NGVI

Assume we use approximations in the exponential family

$$q_\lambda(\theta) = a(\theta)b(\lambda)\exp\left[\langle \lambda, S(\theta)\rangle\right].$$

We can re-parametrize this set with $\mu = \mathbb{E}_{\theta\sim q_\lambda}[T(\theta)] = F(\lambda)$. The NGVI algorithm is given by

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

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

Problem : under this parametrisation, $\mu \mapsto \mathbb{E}_{\theta\sim q_\lambda}[\ell_t(\theta)]$ is generally not convex...

Thank you !