Regret bounds for generalized Bayes updates

Pierre Alquier

RIKEN

AIP
Center for
Advanced Intelligence Project

AI seminar series – UCL AI Centre – May 19, 2021

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.

1st approach : "follow the regularized leader"

- set of predictors : $\{f_\theta, \theta \in \Theta\}$.
- $\ell_t(\theta) := \ell(f_\theta(x_t), y_t))$.

1st approach : "follow the regularized leader"

- set of predictors : $\{f_\theta, \theta \in \Theta\}$.
- $\ell_t(\theta) := \ell(f_\theta(x_t), y_t))$.

Follow The Regularized Leader – FTRL

$$\theta^t := \arg\min_{\theta} \left\{ \sum_{s=1}^{t-1} \ell_s(\theta) + \frac{\mathrm{pen}(\theta)}{\eta} \right\}.$$

1st approach : "follow the regularized leader"

- set of predictors : $\{f_\theta, \theta \in \Theta\}$.
- $\ell_t(\theta) := \ell(f_\theta(x_t), y_t))$.

Follow The Regularized Leader – FTRL

$$\theta^t := \arg\min_\theta \left\{ \sum_{s=1}^{t-1} \ell_s(\theta) + \frac{\mathrm{pen}(\theta)}{\eta} \right\}.$$

Quadratic penalty + linearization :

$$\theta^t := \arg\min_\theta \left\{ \sum_{s=1}^{t-1} \langle \theta, \nabla \ell_s(\theta^s) \rangle + \frac{\|\theta\|^2}{2\eta} \right\}.$$

1st approach : "follow the regularized leader"

- set of predictors : $\{f_\theta, \theta \in \Theta\}$.
- $\ell_t(\theta) := \ell(f_\theta(x_t), y_t))$.

Follow The Regularized Leader – FTRL

$$\theta^t := \arg\min_\theta \left\{ \sum_{s=1}^{t-1} \ell_s(\theta) + \frac{\mathrm{pen}(\theta)}{\eta} \right\}.$$

Differentiate :

$$0 = \sum_{s=1}^{t-1} \nabla \ell_s(\theta^s) + \frac{\theta^t}{\eta}.$$

1st approach : "follow the regularized leader"

- set of predictors : $\{f_\theta, \theta \in \Theta\}$.
- $\ell_t(\theta) := \ell(f_\theta(x_t), y_t))$.

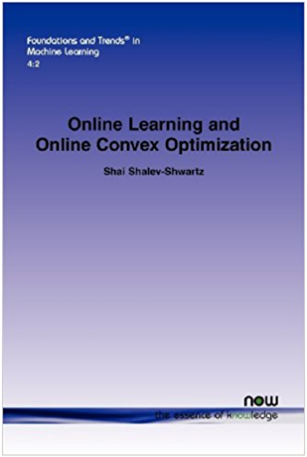
Follow The Regularized Leader – FTRL

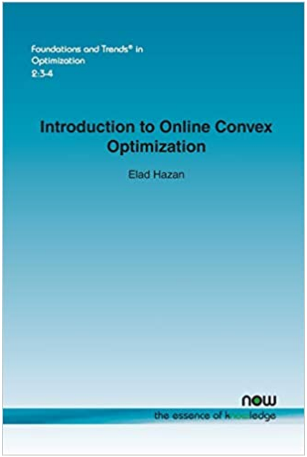
$$\theta^t := \arg\min_\theta \left\{ \sum_{s=1}^{t-1} \ell_s(\theta) + \frac{\mathrm{pen}(\theta)}{\eta} \right\}.$$

Online Gradient Algorithm – OGA

$$\theta^t := \theta^{t-1} - \eta \nabla \ell_{t-1}(\theta^{t-1}).$$

Theoretical properties of FTRL & OGA





2nd approach : (generalized) Bayes

Generalized Bayes, multiplicative weights, Exponential Weight Aggregation (EWA)...

EWA

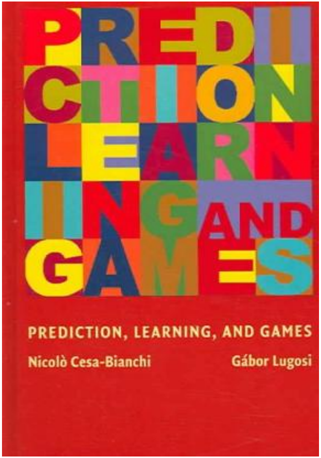
$$\rho^t(\theta) \propto \exp\left[-\eta \sum_{s=1}^{t-1} \ell_s(\theta)\right] \pi(\theta)$$

2nd approach : (generalized) Bayes

Generalized Bayes, multiplicative weights, Exponential Weight Aggregation (EWA)...

EWA

$$\rho^t(\theta) \propto \exp\left[-\eta \sum_{s=1}^{t-1} \ell_s(\theta)\right] \pi(\theta)$$



EWA as FTRL

It is known that

$$\rho^t = \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\Theta)} \left\{ \sum_{s=1}^{t-1} \underbrace{\mathbb{E}_{\theta\sim\rho}[\ell_s(\theta)]}_{=: L_s(\rho)} + \underbrace{\frac{\mathrm{KL}(\rho\|\pi)}{\eta}}_{=: \frac{\mathrm{pen}(\rho)}{\eta}} \right\}.$$

That is, EWA is a special case of FTRL.

$$\mathrm{KL}(\rho\|\pi) = \begin{cases} \mathbb{E}_{\theta\sim\rho}\left[\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right)\right] & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

1st objective

We will study a more general version of FTRL on ρ :

$$\rho^t = \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim \rho}[\ell_s(\theta)] + \frac{D(\rho \| \pi)}{\eta} \right\},$$

for more general divergences D.

P. Alquier. *Non-exponentially Weighted Aggregation : Regret Bounds for Unbounded Loss Functions*. Accepted for ICML 2021.

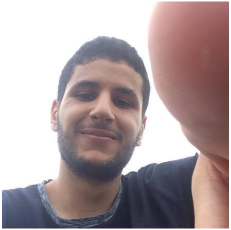
2nd objective

EWA is often non feasible in practice. We will thus modify it : we will constrain ρ^t to belong to a feasible set of probability distributions (e.g. : Gaussian).

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

Co-authors

Badr-Eddine Chérief-Abdellatif



UNIVERSITY OF OXFORD

Emtiyaz Khan

RIKEN

AIP Center for Advanced Intelligence Project

Approximate Bayesian Inference team

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

1. Generalized Bayes update
   - Formula for the posterior : non-exponential weights
   - Regret bound
2. Online variational inference
   - The algorithms : SVA and SVB
   - Regret bounds

1. Generalized Bayes update
   - Formula for the posterior : non-exponential weights
   - Regret bound

2. Online variational inference
   - The algorithms : SVA and SVB
   - Regret bounds

Reminder

$$\rho^t = \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta\sim\rho}[\ell_s(\theta)] + \frac{D_\phi(\rho\|\pi)}{\eta} \right\},$$

Reminder

$$\rho^t = \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta\sim\rho}[\ell_s(\theta)] + \frac{D_\phi(\rho\|\pi)}{\eta} \right\},$$

where

$$D_\phi(\rho\|\pi) = \begin{cases} \mathbb{E}_{\theta\sim\pi}\left[\phi\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right)\right] & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise,} \end{cases}$$

and $\phi : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ with :

- ϕ convex,
- $\phi(1) = 0$,
- $\inf_{x\geq 0} \phi(x) > -\infty$.

Differential of the convex conjugate

Assume that ϕ is differentiable, strictly convex. Put

$$\tilde{\phi}(x) = \begin{cases} \phi(x) \text{ if } x \geq 0, \\ +\infty \text{ if } x < 0. \end{cases}$$

Differential of the convex conjugate

Assume that ϕ is differentiable, strictly convex. Put

$$\tilde{\phi}(x) = \begin{cases} \phi(x) \text{ if } x \geq 0, \\ +\infty \text{ if } x < 0. \end{cases}$$

Then

$$\tilde{\phi}^* = \sup_{x \in \mathbb{R}} [xy - \tilde{\phi}(x)] = \sup_{x \geq 0} [xy - \phi(x)]$$

is differentiable and for any $y \in \mathbb{R}$,

$$\nabla \tilde{\phi}^*(y) = \arg\max_{x \geq 0} \left\{ xy - \phi(x) \right\}.$$

Formula for ρ^t

Assume moreover that $\tilde{\phi}^*(\lambda - a) - \lambda \to \infty$ when $\lambda \to \infty$, for any $a \geq 0$. Then :

$$\lambda_t = \arg\min_{\lambda \in \mathbb{R}} \left\{ \int \tilde{\phi}^* \left(\lambda - \eta \sum_{s=1}^{t-1} \ell_s(\theta) \right) \pi(\mathrm{d}\theta) - \lambda \right\}$$

exists, and

$$\rho^t(\mathrm{d}\theta) = \nabla \tilde{\phi}^* \left(\lambda_t - \eta \sum_{s=1}^{t-1} \ell_s(\theta) \right) \pi(\mathrm{d}\theta).$$

The classical example : KL and exponential weights

- $\phi(x) = x \log(x)$,

The classical example : KL and exponential weights

- $\phi(x) = x \log(x)$,
- $\tilde{\phi}^*(y) = \exp(y - 1)$,

The classical example : KL and exponential weights

- $\phi(x) = x\log(x)$,
- $\tilde{\phi}^*(y) = \exp(y-1)$,
- $\nabla\tilde{\phi}^*(y) = \exp(y-1)$,

The classical example : KL and exponential weights

- $\phi(x) = x\log(x)$,
- $\tilde{\phi}^*(y) = \exp(y-1)$,
- $\nabla\tilde{\phi}^*(y) = \exp(y-1)$,

$$\rho^t(\mathrm{d}\theta) = \exp\left[\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta) - 1\right]\pi(\mathrm{d}\theta).$$

The classical example : KL and exponential weights

- $\phi(x) = x\log(x)$,
- $\tilde{\phi}^*(y) = \exp(y-1)$,
- $\nabla\tilde{\phi}^*(y) = \exp(y-1)$,

$$\rho^t(\mathrm{d}\theta) = \exp\left[\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta) - 1\right]\pi(\mathrm{d}\theta).$$

$$\rho^t(\mathrm{d}\theta) = \frac{\exp\left[-\eta\sum_{s=1}^{t-1}\ell_s(\theta)\right]\pi(\mathrm{d}\theta)}{\int\exp\left[-\eta\sum_{s=1}^{t-1}\ell_s(\vartheta)\right]\pi(\mathrm{d}\vartheta)}.$$

The χ^2 divergence

- $\phi(x) = x^2 - 1$,

The χ^2 divergence

- $\phi(x) = x^2 - 1$,
- $\tilde{\phi}^*(y) = (y^2/4)1_{\{y \geq 0\}}$,

The χ^2 divergence

- $\phi(x) = x^2 - 1$,
- $\tilde{\phi}^*(y) = (y^2/4)1_{\{y \geq 0\}}$,
- $\nabla \tilde{\phi}^*(y) = (y/2)_+$,

The χ^2 divergence

- $\phi(x) = x^2 - 1$,
- $\tilde{\phi}^*(y) = (y^2/4)1_{\{y\geq 0\}}$,
- $\nabla\tilde{\phi}^*(y) = (y/2)_+$,

$$\rho^t(\mathrm{d}\theta) = \left[\frac{\lambda_t - \eta\sum_{s=1}^{t-1}\ell_s(\theta)}{2}\right]_+ \pi(\mathrm{d}\theta).$$

Some references

- the formula was known for a finite Θ :

M. D. Reid, R. M. Frongillo, R. C. Williamson, N. Mehta (2015). *Generalized mixability via entropic duality*. COLT.

- the proof for the general case relies on :

R. Agrawal, T. Horel (2020). *Optimal bounds between f-divergences and integral probability metrics*. ICML.

- comparable PAC-Bayes bounds (no online update) :

P. Alquier and B. Guedj (2018). *Simpler PAC-Bayesian bounds for hostile data*. Machine Learning.

- defense of the generalized Bayes update :

J. Knoblauch, J. Jewson, T. Damoulas (2019). *Generalized variational inference : Three arguments for deriving new posteriors.*. Preprint arXiv.

- more : see the paper.

General regret bound

Assume there is a norm $\|\cdot\|$ such that

General regret bound

Assume there is a norm $\|\cdot\|$ such that

- $\rho \mapsto \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)]$ is L-Lipschitz w.r.t $\|\cdot\|$,

General regret bound

Assume there is a norm $\|\cdot\|$ such that

- $\rho \mapsto \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)]$ is L-Lipschitz w.r.t $\|\cdot\|$,
- $\rho \mapsto D_\phi(\rho\|\pi)$ is α-strongly convex w.r.t $\|\cdot\|$.

General regret bound

Assume there is a norm $\|\cdot\|$ such that

- $\rho \mapsto \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)]$ is L-Lipschitz w.r.t $\|\cdot\|$,
- $\rho \mapsto D_\phi(\rho\|\pi)$ is α-strongly convex w.r.t $\|\cdot\|$.

Theorem

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{D_\phi(\rho\|\pi)}{\eta} \right\}.$$

Bound for EWA : the conditions

- known result : $KL(\rho\|\pi)$ is 1-strongly convex with respect to $\|\cdot\|_{\mathrm{TV}}$;

Bound for EWA : the conditions

- known result : $KL(\rho\|\pi)$ is 1-strongly convex with respect to $\|\cdot\|_{\mathrm{TV}}$;
- we have :

$$\left|\int \ell_t(\theta)\rho(\mathrm{d}\theta) - \int \ell_t \rho'(\mathrm{d}\theta)\right| \leq \int \ell_t(\theta) \left|\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta)\right| \pi(\mathrm{d}\theta)$$

$$\leq L \underbrace{\int \left|\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta)\right| \pi(\mathrm{d}\theta)}_{=2\|\rho-\rho'\|_{\mathrm{TV}}}$$

on the condition that $0 \leq \ell_t(\theta) \leq L$ for any θ.

Bound for EWA

Assume $0 \leq \ell_t(\theta) \leq L$ for any θ, t, then

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \eta L^2 T + \frac{\mathrm{KL}(\rho\|\pi)}{\eta} \right\}.$$

Bound for EWA

Assume $0 \leq \ell_t(\theta) \leq L$ for any θ, t, then

$$\sum_{t=1}^{T} \mathbb{E}_{\theta \sim \rho^t}[\ell_t(\theta)] \leq \inf_{\rho \in \mathcal{P}(\Theta)} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta \sim \rho}[\ell_t(\theta)] + \eta L^2 T + \frac{\mathrm{KL}(\rho \| \pi)}{\eta} \right\}.$$

(This is a well-known result).

Bound with χ^2 : the conditions

- $\phi(x) = x^2 - 1$ is 2-strongly convex so D_ϕ is 2-strongly convex with respect to the $L_2(\pi)$ norm.

Bound with χ^2 : the conditions

- $\phi(x) = x^2 - 1$ is 2-strongly convex so D_ϕ is 2-strongly convex with respect to the $L_2(\pi)$ norm.
- we have

$$\left| \int \ell_t(\theta)\rho(\mathrm{d}\theta) - \int \ell_t \rho'(\mathrm{d}\theta) \right| \leq \int \ell_t(\theta) \left| \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right| \pi(\mathrm{d}\theta)$$

$$\leq L \left(\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right)^2 \pi(\mathrm{d}\theta) \right)^{1/2}$$

on the condition that $\left(\int \ell_t(\theta)^2 \pi(\mathrm{d}\theta)\right)^{1/2} \leq L$.

Bound with χ^2

Assume $\int \ell_t(\theta)^2 \pi(\mathrm{d}\theta) \leq L^2$ for any t, then

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho^t}[\ell_t(\theta)] \leq \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim\rho}[\ell_t(\theta)] + \frac{\eta L^2 T}{2} + \frac{\chi^2(\rho\|\pi)}{\eta} \right\}.$$

1. Generalized Bayes update
   - Formula for the posterior : non-exponential weights
   - Regret bound

2. Online variational inference
   - The algorithms : SVA and SVB
   - Regret bounds

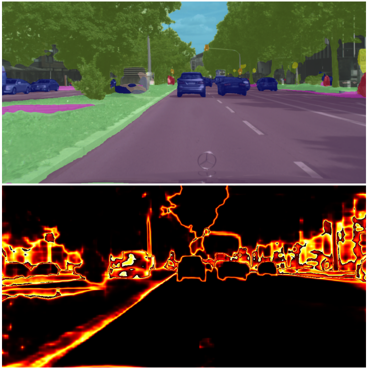
Motivation

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.

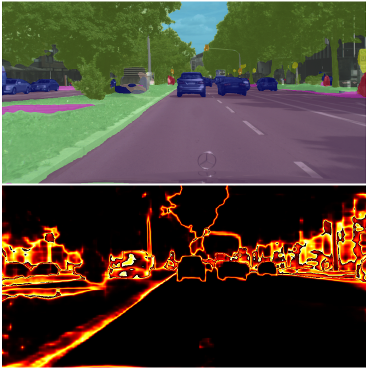


Picture : Roman Bachmann.

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.

Sequential Variational Approximation (SVA)

We restrict ρ to belong to $\mathcal{F} = \{q_\mu, \mu \in M\}$ a parametric family. Example : Gaussian distributions.

Sequential Variational Approximation (SVA)

We restrict ρ to belong to $\mathcal{F} = \{q_\mu, \mu \in M\}$ a parametric family. Example : Gaussian distributions.

FTRL on this set :

$$\mu^t = \arg\min_{\mu \in M} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] + \frac{D_\phi(q_\mu, \pi)}{\eta} \right\}.$$

Sequential Variational Approximation (SVA)

We restrict ρ to belong to $\mathcal{F} = \{q_\mu, \mu \in M\}$ a parametric family. Example : Gaussian distributions.

FTRL on this set :

$$\mu^t = \arg\min_{\mu \in M} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta \sim q_\mu}[\ell_s(\theta)] + \frac{D_\phi(q_\mu, \pi)}{\eta} \right\}.$$

Linearization gives :

SVA

$$\mu^t = \arg\min_{\mu \in M} \left\{ \sum_{s=1}^{t-1} \left\langle \mu, \nabla \mathbb{E}_{\theta \sim q_{\mu^s}}[\ell_s(\theta)] \right\rangle + \frac{D_\phi(q_\mu, \pi)}{\eta} \right\}.$$

Streaming Variational Bayes (SVB)

(OGA) can actually be obtained via :

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

OR

$$\theta^t := \arg\min_{\theta} \left\{ \langle \theta, \nabla \ell_{t-1}(\theta^{t-1}) \rangle + \frac{\|\theta - \theta^{t-1}\|^2}{2\eta} \right\}$$

Streaming Variational Bayes (SVB)

(OGA) can actually be obtained via :

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

OR

$$\theta^t := \arg\min_{\theta} \left\{ \left\langle \theta, \nabla \ell_{t-1}(\theta^{t-1}) \right\rangle + \frac{\|\theta - \theta^{t-1}\|^2}{2\eta} \right\}$$

SVB

$$\mu^t = \arg\min_{\mu \in M} \left\{ \left\langle \mu, \nabla \mathbb{E}_{\theta \sim q_{\mu^{t-1}}} [\ell_{t-1}(\theta)] \right\rangle + \frac{D_\phi(q_\mu, q_{\mu^{t-1}})}{\eta} \right\}.$$

SVA & SVB are tractable, and not equivalent

Example : Gaussian prior $\theta \sim \pi = \mathcal{N}(0, s^2 I)$, $D_\phi = \mathrm{KL}$ 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

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

Theoretical analysis of SVA

Two assumptions :

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

Theoretical analysis of SVA

Two 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.

Theoretical analysis of SVA

Two 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 :

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

Theoretical analysis of SVA

Two 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 :

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

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

Theoretical analysis of SVA

Two 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 :

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

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

   For example true when q_μ is Gaussian with $\mu = (m, \Sigma)$ and $D_\phi = \mathrm{KL}$.

Theoretical analysis of SVA

Theorem

Under the previous assumptions SVA leads to

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

Theoretical analysis of SVA

Theorem

Under the previous assumptions SVA leads to

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

Application to Gaussian approximation with KL :

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

Theoretical analysis of SVB

Theorem 2

Using Gaussian approximations and $D_\phi = \mathrm{KL}$, 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 and $D_\phi = \mathrm{KL}$, 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

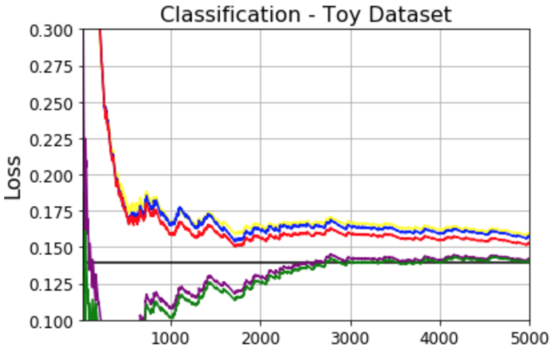


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Breast dataset

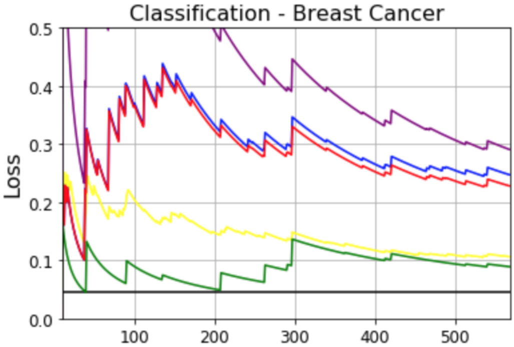


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

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 $\lambda = F(\mu)$),

$$\lambda^t = (1-\rho)\lambda^{t-1} + \rho \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu^{t-1}}} [\ell_t(\theta)] ,$$

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

Thank you !