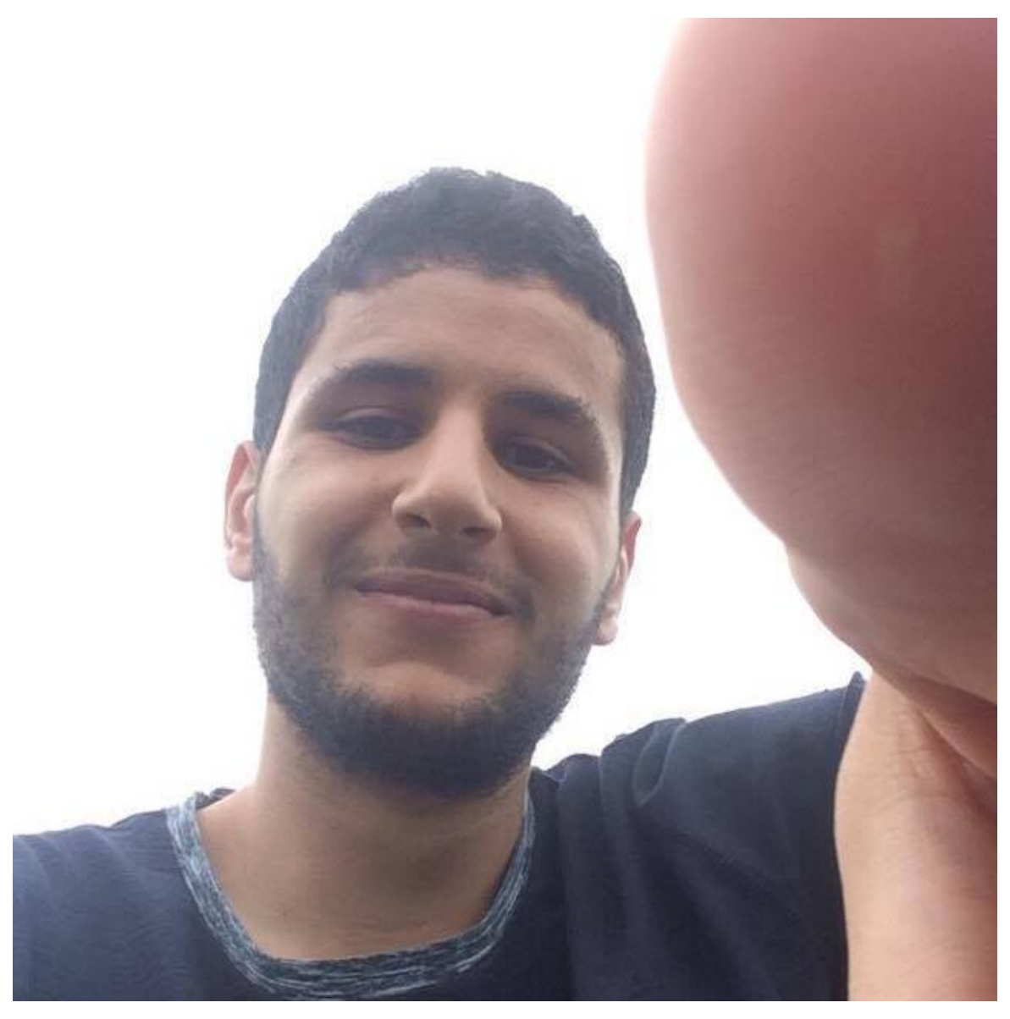
A Regret Bound for Online Variational Inference

Badr-Eddine Chérief-Abdellatif, Pierre Alquier, Emtiyaz Khan

ENSAE Paris & RIKEN AIP

B.-E. Chérief-Abdellatif and E. Khan



Approximate Bayesian Inference team

https : //emtiyaz.github.io/

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,

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

Note that θ_{t+1} can be obtained by:

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

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

Bayesian learning and variational inference (VI)

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

Not tractable in general, leading to variational approximations:

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

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

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

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

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

Q2: can we derive similar results for online VI?

Two options for online VI

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

Sequential Variational Approximation (SVA):

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

Streaming Variational Bayes (SVB):

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

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

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

Theoretical analysis of SVA

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

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

Application to Gaussian approximation leads to

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

Theoretical analysis of SVB

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

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

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

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

Test on a simulated dataset

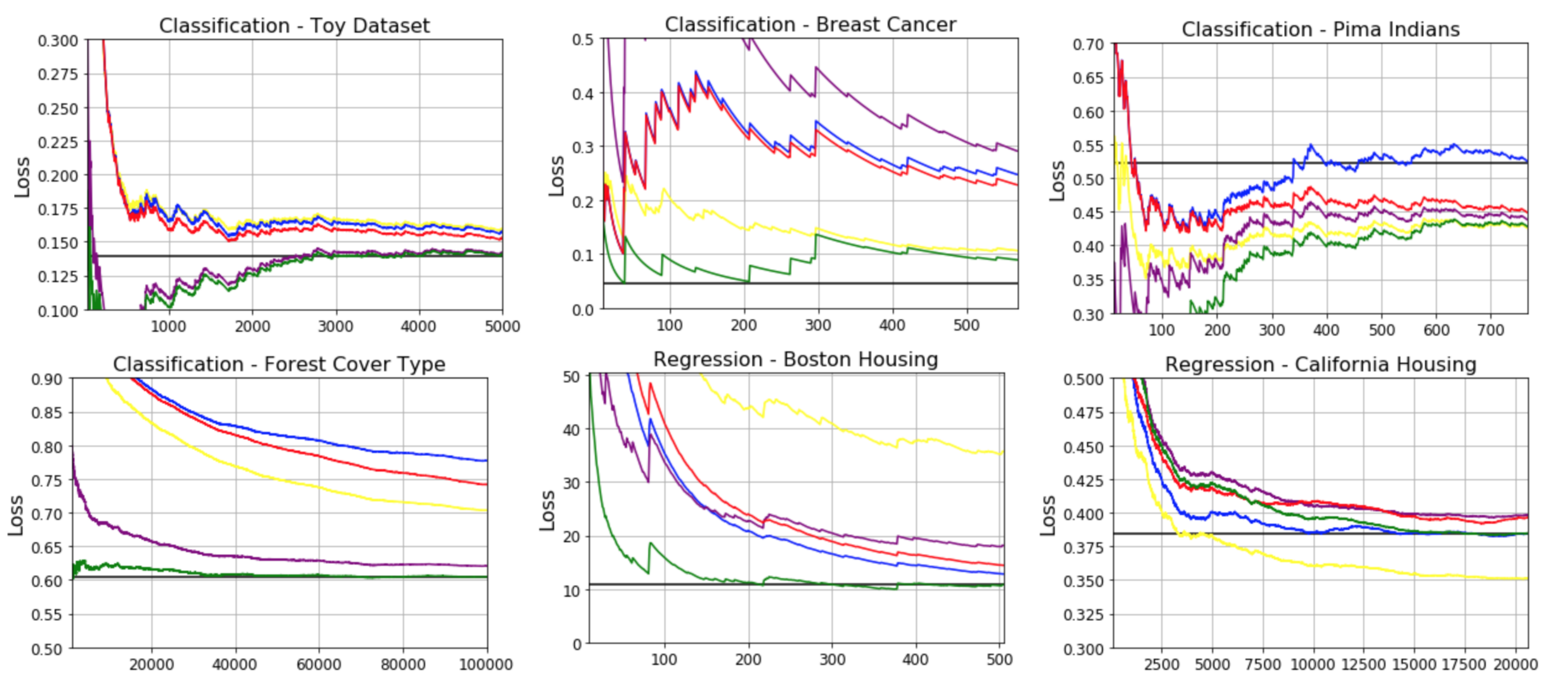


Figure: Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green) for the convex hinge loss and the squared loss functions. The black line shows the average total cumulative loss in hindsight. We see that in most cases NGVI outperforms the other algorithms. The last plot (California Housing dataset) shows the consistency of our algorithms for a nonconvex loss $\bar{L}_t$.

Open questions

Analysis of SVB in the general case.

Analysis of the uncertainty quantification.

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

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

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

http://alquier.ensae.net/ pierre.alquier.stat@gmail.com