Generalization bounds for variational inference

Pierre Alquier

École nationale de la statistique et de l'administration économique
ENSAE ParisTech
université PARIS-SACLAY

INRA SCIENCE & IMPACT

MaIAGE
Mathématiques et Informatique Appliquées du Génome à l'Environnement

Jouy-en-Josas
May 6, 2019

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

The likelihood

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

The likelihood

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

The posterior

$$\pi_n(\mathrm{d}\theta) \propto L_n(\theta)\pi(\mathrm{d}\theta).$$

Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by Q : $\frac{\mathrm{d}P_\theta}{\mathrm{d}Q} = p_\theta$. Prior π on Θ.

The likelihood

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

The posterior

$$\pi_n(\mathrm{d}\theta) \propto L_n(\theta)\pi(\mathrm{d}\theta).$$

The tempered posterior - $0 < \alpha < 1$

$$\pi_{n,\alpha}(\mathrm{d}\theta) \propto [L_n(\theta)]^\alpha \pi(\mathrm{d}\theta).$$

Computation of the posterior

- explicit form (conjugate models),

Computation of the posterior

- explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

Computation of the posterior

- explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

But...

- when the dimension is large, the convergence of MCMC can be extremely slow,

Computation of the posterior

- explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

But...

- when the dimension is large, the convergence of MCMC can be extremely slow,
- when the model is complex or when the sample size is large, each evaluation of $\pi_{n,\alpha}(\theta)$ can be expensive.

Computation of the posterior

- explicit form (conjugate models),
- MCMC algorithms : Metropolis-Hastings, Gibbs sampler, Langevin Monte Carlo...

But...

- when the dimension is large, the convergence of MCMC can be extremely slow,
- when the model is complex or when the sample size is large, each evaluation of $\pi_{n,\alpha}(\theta)$ can be expensive.

For these reasons, in the past 20 years, many methods targeting an approximation of $\pi_{n,\alpha}$ became popular : **ABC**, **EP algorithm**, variational inference, **approximate MCMC** ...

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

$$\tilde{\pi}_{n,\alpha} := \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

$$\tilde{\pi}_{n,\alpha} := \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Examples :

- parametric approximation

$$\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \right\}.$$

Variational approximations : definitions

Idea of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

$$\tilde{\pi}_{n,\alpha} := \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Examples :

- parametric approximation

$$\mathcal{F} = \left\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+\right\}.$$

- mean-field approximation, $\Theta = \Theta_1 \times \Theta_2$ and

$$\mathcal{F} : \{\rho : \rho(\mathrm{d}\theta) = \rho_1(\mathrm{d}\theta_1) \times \rho_2(\mathrm{d}\theta_2)\}.$$

Empirical lower bound (ELBO)

Note that :

$$
\begin{aligned}
\tilde{\pi}_{n,\alpha} &= \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\
&= \arg\min_{\rho \in \mathcal{F}} \underbrace{\left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}}_{-\mathrm{ELBO}(\rho)} .
\end{aligned}
$$

Empirical lower bound (ELBO)

Note that :

$$\begin{aligned}\tilde{\pi}_{n,\alpha} &= \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\ &= \arg\min_{\rho\in\mathcal{F}} \underbrace{\left\{-\alpha \int \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i)\rho(\mathrm{d}\theta) + \mathcal{K}(\rho,\pi)\right\}}_{-\mathrm{ELBO}(\rho)}.\end{aligned}$$

So we have the equivalent definition :

$$\tilde{\pi}_{n,\alpha} := \arg\max_{\rho\in\mathcal{F}} \mathrm{ELBO}(\rho).$$

Outline of the talk

After this introduction :

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

We will show general conditions, and many examples.

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

We will show general conditions, and many examples.

Section 3 will address the following question :

Are there efficient algorithms to (provably) compute $\tilde{\pi}_{n,\alpha}$?

Outline of the talk

After this introduction :

Section 2 will address the following question :

What are the conditions ensuring that $\tilde{\pi}_{n,\alpha}$ leads to good estimators ?

We will show general conditions, and many examples.

Section 3 will address the following question :

Are there efficient algorithms to (provably) compute $\tilde{\pi}_{n,\alpha}$?

We will see that fast algorithms from sequential optimization can be used in some cases. This also allows to do variational inference on a data stream that cannot be stored.

Outline of the talk

1. Introduction : variational Bayesian inference
   - Bayesian inference
   - Definition of variational approximations
   - Outline of the talk
2. Concentration of variational approximations of the posterior
   - Theoretical results
   - Applications
   - Extensions
3. Online variational inference
   - Sequential estimation problem
   - Online variational inference
   - Simulations

Tools for the consistency of VB

The α-Rényi divergence for $\alpha \in (0,1)$

$$D_\alpha(P,R) = \frac{1}{\alpha - 1} \log \int (\mathrm{d}P)^\alpha (\mathrm{d}R)^{1-\alpha}.$$

Tools for the consistency of VB

The α-Rényi divergence for $\alpha \in (0,1)$

$$D_\alpha(P,R) = \frac{1}{\alpha - 1} \log \int (\mathrm{d}P)^\alpha (\mathrm{d}R)^{1-\alpha}.$$

All the properties derived in :

T. Van Erven & P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 2014.

Among others, for $1/2 \leq \alpha$, link with Hellinger and Kullback :

$$\mathcal{H}^2(P,R) \leq D_\alpha(P,R) \xrightarrow[\alpha \nearrow 1]{} \mathcal{K}(P,R).$$

What do we know about $\pi_{n,\alpha}$?

$$\mathcal{B}(r) = \{\theta \in \Theta : \mathcal{K}(P_{\theta_0}, P_\theta) \leq r\}.$$

Theorem, variant of (Bhattacharya, Pati & Yang)

For any sequence (r_n) such that

$$-\log \pi[B(r_n)] \leq nr_n$$

we have

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\pi_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

A. Bhattacharya, D. Pati & Y. Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 2019.

Extension of previous result to VB

Theorem (A. & Ridgway)

If there is $\rho_n \in \mathcal{F}$ and (r_n) such that

$$\begin{cases} \int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n, \\ \text{and} \\ \mathcal{K}(\rho_n, \pi) \leq n r_n, \end{cases}$$

then, for any $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Extension of previous result to VB

Theorem (A. & Ridgway)

If there is $\rho_n \in \mathcal{F}$ and (r_n) such that

$$\begin{cases} \int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n, \\ \text{and} \\ \mathcal{K}(\rho_n, \pi) \leq n r_n, \end{cases}$$

then, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

P. Alquier & J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, to appear.

Misspecified case

Assume now that $X_1, \ldots, X_n$ i.i.d $\sim Q \notin \{P_\theta, \theta \in \Theta\}$. Put :

$$\theta^* := \arg\min_{\theta\in\Theta} \mathcal{K}(Q, P_\theta).$$

Theorem (A. and Ridgway)

Assume that there is $\rho_n \in \mathcal{F}$ such that

$$\int \mathbb{E}\left[\log \frac{\mathrm{d}P_{\theta^*}}{\mathrm{d}P_\theta}\right] \rho_n(\mathrm{d}\theta) \leq r_n \text{ and } \mathcal{K}(\rho_n, \pi) \leq n r_n,$$

then, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, Q)\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{\alpha}{1-\alpha}\mathcal{K}(Q, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} r_n.$$

First example : nonparametric regression

Nonparametric regression

- $Y_i = f(X_i) + \xi_i$,

First example : nonparametric regression

Nonparametric regression

- $Y_i = f(X_i) + \xi_i$,
- $\xi_i \sim \mathcal{N}(0, \sigma^2)$,

First example : nonparametric regression

Nonparametric regression

- $Y_i = f(X_i) + \xi_i$,
- $\xi_i \sim \mathcal{N}(0, \sigma^2)$,
- f is s-smooth with s unknown,

First example : nonparametric regression

Nonparametric regression

- $Y_i = f(X_i) + \xi_i$,
- $\xi_i \sim \mathcal{N}(0, \sigma^2)$,
- f is s-smooth with s unknown,
- prior : $f(\cdot) = \sum_{j=1}^{K} \beta_j \phi_j(\cdot)$, random K and β_j's, (φ_j) basis...

First example : nonparametric regression

Nonparametric regression

- $Y_i = f(X_i) + \xi_i$,
- $\xi_i \sim \mathcal{N}(0, \sigma^2)$,
- f is s-smooth with s unknown,
- prior : $f(\cdot) = \sum_{j=1}^{K} \beta_j \phi_j(\cdot)$, random K and β_j's, (φ_j) basis...
- variational approx : β_j mutually independent...

Under suitable assumptions, $r_n \sim \left(\frac{\log(n)}{n}\right)^{\frac{2s}{2s+1}}$.

More examples covered in the paper

1. logistic regression,

More examples covered in the paper

1. logistic regression,
2. matrix completion : we prove that the approx. in

Y. J. Lim & Y. W. Teh. Variational Bayesian approach to movie rating prediction. *Proceedings of KDD cup and workshop*, 2007.

leads to minimax-optimal estimation.

				
Claire	4	?	3	...
Nial	?	4	?	...
Brendon	?	5	4	...
Andrew	?	4	?	...
Adrian	1	?	?	...
Damien	?	1	?	...
⋮	⋮	⋮	⋮	⋱

An important example : mixture models

Mixture models

- $P_\theta = P_{p,\theta_1,\dots,\theta_K} = \sum_{j=1}^{K} p_j q_{\theta_j}$,

An important example : mixture models

Mixture models

- $P_\theta = P_{p,\theta_1,\dots,\theta_K} = \sum_{j=1}^K p_j q_{\theta_j}$,
- prior π : $p = (p_1, \dots, p_K) \sim \pi_p = \mathcal{D}(\alpha_1, \dots, \alpha_K)$ and the θ_j's are independent from π_θ.

An important example : mixture models

Mixture models

- $P_\theta = P_{p,\theta_1,\dots,\theta_K} = \sum_{j=1}^K p_j q_{\theta_j}$,
- prior π : $p = (p_1, \dots, p_K) \sim \pi_p = \mathcal{D}(\alpha_1, \dots, \alpha_K)$ and the θ_j's are independent from π_θ.

Tempered posterior :

$$L_n(\theta)^\alpha \pi(\theta) \propto \left(\prod_{i=1}^n \sum_{j=1}^K p_j q_{\theta_j}(X_i) \right)^\alpha \pi_p(p) \prod_{j=1}^K \pi_\theta(\theta_j).$$

An important example : mixture models

Mixture models

- $P_\theta = P_{p,\theta_1,\dots,\theta_K} = \sum_{j=1}^K p_j q_{\theta_j}$,
- prior π : $p = (p_1, \dots, p_K) \sim \pi_p = \mathcal{D}(\alpha_1, \dots, \alpha_K)$ and the θ_j's are independent from π_θ.

Tempered posterior :

$$L_n(\theta)^\alpha \pi(\theta) \propto \left(\prod_{i=1}^n \sum_{j=1}^K p_j q_{\theta_j}(X_i) \right)^\alpha \pi_p(p) \prod_{j=1}^K \pi_\theta(\theta_j).$$

Variational approximation :

$$\tilde{\pi}_{n,\alpha}(p,\theta) = \rho_p(p) \prod_{j=1}^K \rho_j(\theta_j).$$

ELBO maximization for mixtures

Optimization program

$$\min_{\rho=(\rho_p,\rho_1,\dots,\rho_K)} \left\{ -\alpha \sum_{i=1}^{n} \int \log\left(\sum_{j=1}^{K} p_j q_{\theta_j}(X_i) \right) \rho(d\theta) + \mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j) \right\}$$

$$-\log\left(\sum_{j=1}^{K} p_j q_{\theta_j}(X_i) \right) = \min_{\omega^i \in \mathcal{S}_K} \left\{ -\sum_{j=1}^{K} \omega_j^i \log(p_j q_{\theta_j}(X_i)) + \sum_{j=1}^{K} \omega_j^i \log(\omega_j^i) \right\}$$

Coordinate Descent algorithm

Algorithm 1 Coordinate Descent Variational Bayes for mixtures

Input: a dataset $(X_1, ..., X_n)$, priors $\pi_p, \{\pi_j\}_{j=1}^K$ and a family $\{q_\theta / \theta \in \Theta\}$
Output: a variational approximation $\rho_p(p) \prod_{j=1}^K \rho_j(\theta_j)$
Initialize variational factors ρ_p, $\{\rho_j\}_{j=1}^K$
until convergence of the objective function **do**
for $i = 1, ..., n$ **do**
  for $j = 1, ..., K$ **do**
    set $w_j^i = \exp\left(\int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) \right)$
  end for
  normalize $(w_j^i)_{1 \leq j \leq K}$
end for
set $\rho_p(dp) \propto \exp\left(\alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \log(p_j) \right) \pi_p(dp)$
for $j = 1, ..., K$ **do**
  set $\rho_j(d\theta_j) \propto \exp\left(\alpha \sum_{i=1}^n \omega_j^i \log(q_{\theta_j}(X_i)) \right) \pi_j(d\theta_j)$
end for

Numerical example on Gaussian mixtures

Gaussian mixture $\sum_{j=1}^{3} p_j \mathcal{N}(\theta_j, 1)$ and Gaussian prior on θ_j.
Sample size $n = 1000$, we report the MAE over 10 replications.

Algo.	p	θ_1	θ_2	θ_3
$\mathrm{VB}_{\alpha=0.5}$	0.03 (0.02)	0.14 (0.30)	0.38 (1.11)	0.05 (0.05)
$\mathrm{VB}_{\alpha=1}$	0.03 (0.02)	0.14 (0.21)	0.36 (0.97)	0.06 (0.04)
EM	0.03 (0.02)	0.14 (0.22)	0.36 (0.97)	0.06 (0.05)

Mixture models : convergence rates

Theorem (Chérief-Abdellatif, A.)

Chose $\frac{2}{K} \leq \alpha_j \leq 1$ and assume that estimation in (q_θ) (without mixture) at rate r_n.

Mixture models : convergence rates

Theorem (Chérief-Abdellatif, A.)

Chose $\frac{2}{K} \leq \alpha_j \leq 1$ and assume that estimation in (q_θ) (without mixture) at rate r_n. Then

$$\mathbb{E}\left[\int D_\alpha(P_{p,\theta_1,\dots,\theta_K}, P_{p^0,\theta_1^0,\dots,\theta_K^0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha}\mathrm{cst}.Kr_n.$$

Mixture models : convergence rates

Theorem (Chérief-Abdellatif, A.)

Chose $\frac{2}{K} \leq \alpha_j \leq 1$ and assume that estimation in (q_θ) (without mixture) at rate r_n. Then

$$\mathbb{E}\left[\int D_\alpha(P_{p,\theta_1,\dots,\theta_K}, P_{p^0,\theta_1^0,\dots,\theta_K^0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \\ \leq \frac{1+\alpha}{1-\alpha}\mathrm{cst}.Kr_n.$$

B.-E. Chérief-Abdellatif, P. Alquier. Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures. *Electronic Journal of Statistics*, 2018.

Model selection

D. Blei, A. Kucukelbir & J. McAuliffe. Variational inference : A review for statisticians. *JASA*, 2017.

Model selection

D. Blei, A. Kucukelbir & J. McAuliffe. Variational inference : A review for statisticians. *JASA*, 2017.

The relationship between the ELBO and log $p(\mathbf{x})$ has led to using the variational bound as a model selection criterion. This has been explored for mixture models (Ueda and Ghahramani 2002; McGrory and Titterington 2007) and more generally (Beal and Ghahramani 2003). The premise is that the bound is a good approximation of the marginal likelihood, which provides a basis for selecting a model. Though this sometimes works in practice, selecting based on a bound is not justified in theory. Other research has used variational approximations in the log predictive density to use VI in cross-validation-based model selection (Nott et al. 2012).

Model selection

Assume that we have K models, define $\tilde{\pi}^k_{n,\alpha}$ a variational approximation of the tempered posterior in model k, and r^k_n its convergence rate if model k is correct. Put :

$$\hat{k} = \arg\max_k \mathrm{ELBO}(\tilde{\pi}^k_{n,\alpha}).$$

Model selection

Assume that we have K models, define $\tilde{\pi}^k_{n,\alpha}$ a variational approximation of the tempered posterior in model k, and r^k_n its convergence rate if model k is correct. Put :

$$\hat{k} = \arg\max_k \mathrm{ELBO}(\tilde{\pi}^k_{n,\alpha}).$$

Theorem (Chérief-Abdellatif)

If the true model is actually k_0,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}^{\hat{k}}_{n,\alpha}(d\theta | X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} r_n^{k_0} + \frac{\log(K)}{n(1-\alpha)}.$$

Model selection

Assume that we have K models, define $\tilde{\pi}^k_{n,\alpha}$ a variational approximation of the tempered posterior in model k, and r_n^k its convergence rate if model k is correct. Put :

$$\hat{k} = \arg\max_k \mathrm{ELBO}(\tilde{\pi}^k_{n,\alpha}).$$

Theorem (Chérief-Abdellatif)

If the true model is actually k_0,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}^{\hat{k}}_{n,\alpha}(d\theta | X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} r_n^{k_0} + \frac{\log(K)}{n(1-\alpha)}.$$

B.-E. Chérief-Abdellatif. Consistency of ELBO maximization for model selection. *Proceedings of AABI* 2018.

More extensions

1. more general models with latent variables :

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

More extensions

1. more general models with latent variables :

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

2. case $\alpha = 1$, *i.e* approximation of the "usual" posterior :

F. Zhang & C. Gao. Convergence Rates of Variational Posterior Distributions. *Preprint arXiv*, 2017.

More extensions

1. more general models with latent variables :

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

2. case $\alpha = 1$, *i.e* approximation of the "usual" posterior :

F. Zhang & C. Gao. Convergence Rates of Variational Posterior Distributions. *Preprint arXiv*, 2017.

3. approximation based on another distance, for example :

$$\tilde{\pi}_{n,\alpha} := \arg\min_{\rho\in\mathcal{F}} \mathcal{W}(\rho, \pi_{n,\alpha}) \text{ (Wasserstein distance)},$$

J. Huggins, T. Campbell, M. Kasprzak & T. Broderick. Practical bounds on the error of Bayesian posterior approximations : a nonasymptotic approach. *Preprint arXiv*, 2018.

Outline of the talk

Sequential estimation problem

Sequential estimation problem

1

1. initialize θ_1,

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$
3.
   1. update $\theta_2 \to \theta_3$,

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$
3.
   1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,

Sequential estimation problem

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$
3. 1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $-\log p_{\theta_3}(x_3)$
4. ...

Sequential estimation problem

1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$
3.
   1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $-\log p_{\theta_3}(x_3)$
4. ...

Objective :

Sequential estimation problem

1. 1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2. 1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$
3. 1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $-\log p_{\theta_3}(x_3)$
4. ...

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

Sequential estimation problem

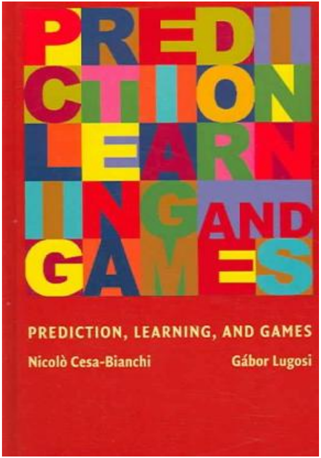
1.
   1. initialize θ_1,
   2. x_1 revealed,
   3. incur loss $-\log p_{\theta_1}(x_1)$
2.
   1. update $\theta_1 \to \theta_2$,
   2. x_2 revealed,
   3. incur loss $-\log p_{\theta_2}(x_2)$
3.
   1. update $\theta_2 \to \theta_3$,
   2. x_3 revealed,
   3. incur loss $-\log p_{\theta_3}(x_3)$
4. ...

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

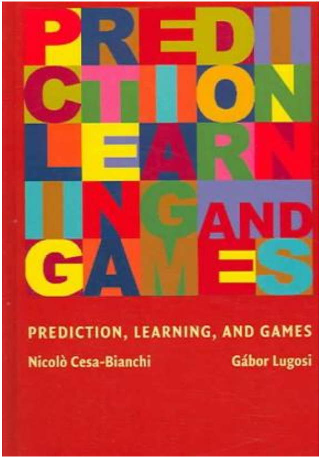
$$\sum_{t=1}^{T} [-\log p_{\theta_t}(x_t)]$$

as small as possible for any T, **without stochastic assumptions on the data.**

Reference


PREDICTION LEARNING AND GAMES
PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi

Reference



The regret :

$$\mathrm{R}(T) = \sum_{t=1}^{T} [-\log p_{\theta_t}(x_t)] - \inf_{\theta \in \Theta} \sum_{t=1}^{T} [-\log p_{\theta}(x_t)].$$

EWA strategy / multipicative update...

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).

Algorithm 2 Exponentially Weighted Aggregation

1: **for** $t = 1, 2, \dots$ **do**
2: $\theta_t = \mathbb{E}_{\theta \sim p_t}[\theta]$,
3: x_t revealed, update $p_{t+1}(\mathrm{d}\theta) = \frac{[p_\theta(x_t)]^\alpha p_t(\mathrm{d}\theta)}{\int [p_\vartheta(x_t)]^\alpha p_t(\mathrm{d}\vartheta)}$.
4: **end for**

EWA strategy / multipicative update...

- learning rate $\alpha > 0$.
- initialize $p_1 = \pi$ (the prior).

Algorithm 2 Exponentially Weighted Aggregation

1: **for** $t = 1, 2, \dots$ **do**
2: $\quad \theta_t = \mathbb{E}_{\theta \sim p_t}[\theta]$,
3: $\quad x_t$ revealed, update $p_{t+1}(\mathrm{d}\theta) = \frac{[p_\theta(x_t)]^\alpha p_t(\mathrm{d}\theta)}{\int [p_\vartheta(x_t)]^\alpha p_t(\mathrm{d}\vartheta)}$.
4: **end for**

Note that $p_t = \pi_{n,\alpha}$ the tempered posterior, so problem : how can we compute θ_t ?

A regret bound for EWA

From now, $\theta \mapsto [-\log p_\theta(x_t)]$ is convex + bounded : $|\cdot| \leq C$.

A regret bound for EWA

From now, $\theta \mapsto [-\log p_\theta(x_t)]$ is convex + bounded : $|\cdot| \leq C$.

Theorem

$$\sum_{t=1}^{T} [-\log p_{\theta_t}(x_t)] \leq \inf_p \left[\sum_{t=1}^{T} \mathbb{E}_{\theta\sim p}[-\log p_\theta(x_t)] \right.$$
$$\left. + \frac{\alpha C^2 T}{2} + \frac{\mathcal{K}(p,\pi)}{\alpha} \right].$$

A regret bound for EWA

From now, $\theta \mapsto [-\log p_{\theta}(x_t)]$ is convex + bounded : $|\cdot| \leq C$.

Theorem

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{p}\left[\sum_{t=1}^{T}\mathbb{E}_{\theta\sim p}[-\log p_{\theta}(x_t)] + \frac{\alpha C^2 T}{2} + \frac{\mathcal{K}(p,\pi)}{\alpha}\right].$$

Under similar assumptions than in the batch case, that is, the prior gives enough mass to relevant θ, and $\alpha \sim 1/\sqrt{T}$,

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta}\sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst}.\sqrt{T}$$

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst.}\sqrt{T}$$

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst}.\sqrt{T}$$

$$\frac{1}{T}\sum_{t=1}^{T} \log \frac{q(x_t)}{p_{\theta_t}(x_t)} \leq \inf_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \log \frac{q(x_t)}{p_{\theta}(x_t)} + \frac{\mathrm{cst}}{\sqrt{T}}.$$

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \leq \inf_{\theta\in\Theta} \sum_{t=1}^{T}[-\log p_{\theta}(x_t)] + \mathrm{cst}.\sqrt{T}$$

$$\frac{1}{T}\sum_{t=1}^{T} \log \frac{q(x_t)}{p_{\theta_t}(x_t)} \leq \inf_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \log \frac{q(x_t)}{p_{\theta}(x_t)} + \frac{\mathrm{cst}}{\sqrt{T}}.$$

Assuming that $x_1, \dots, x_T$ are actually i.i.d from Q, with density q, define

$$\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta_T,$$

we have ("online-to-batch" conversion) :

$$\mathbb{E}\left[\mathcal{K}\left(Q, P_{\hat{\theta}_T}\right)\right] \leq \inf_{\theta\in\Theta} \mathcal{K}\left(Q, P_{\theta}\right) + \frac{\mathrm{cst}}{\sqrt{T}}.$$

Variational approximations of EWA

B.-E. Chérief-Abdellatif, P. Alquier & M. E. Khan. A Generalization Bound for Online Variational Inference. *Preprint arXiv*, 2018.

Variational approximations of EWA

B.-E. Chérief-Abdellatif, P. Alquier & M. E. Khan. A Generalization Bound for Online Variational Inference. *Preprint arXiv*, 2018.

Parametric variational approximation :

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

Objective : propose a way to update $\mu_t \to \mu_{t+1}$ so that q_{μ_t} leads to similar performances as p_t in EWA...

SVA and SVB strategies

Algorithm 3 SVA (Sequential Variational Approximation)

1: **for** $t = 1, 2, \ldots$ **do**

2: $\theta_t = \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,

3: x_t revealed, update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left[\mu^T \nabla_\mu \sum_{i=1}^{t} \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_i)] + \frac{\mathcal{K}(q_\mu, \pi)}{\alpha} \right].$$

4: **end for**

SVA and SVB strategies

Algorithm 3 SVA (Sequential Variational Approximation)

1: **for** $t = 1, 2, \dots$ **do**

2: $\theta_t = \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,

3: x_t revealed, update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left[\mu^T \nabla_\mu \sum_{i=1}^{t} \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_i)] + \frac{\mathcal{K}(q_\mu, \pi)}{\alpha} \right].$$

4: **end for**

SVB (Streaming Variational Bayes) has update

$$\mu_{t+1} = \arg\min_{\mu \in M} \left[\mu^T \nabla_\mu \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)] + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\alpha} \right].$$

NGVI strategy

NGVI (Natural Gradient Variational Inference) : fix some $\beta > 0$,

$$\mu_{t+1} = \arg\min_{\mu\in M}\left[\mu^T \nabla_\mu \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)] + \frac{\mathcal{K}(q_\mu, \pi)}{\alpha} + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\beta}\right].$$

NGVI strategy

NGVI (Natural Gradient Variational Inference) : fix some $\beta > 0$,

$$\mu_{t+1} = \arg\min_{\mu \in M} \left[\mu^T \nabla_\mu \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)] + \frac{\mathcal{K}(q_\mu, \pi)}{\alpha} + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\beta} \right].$$

M. E. Khan & W. Lin. Conjugate-computation variational inference : Converting variational inference in non-conjugate models to inferences in conjugate models. *AISTAT*, 2017.

An example : SVB with Gaussian approximations

As an example, assume that $\theta \in \mathbb{R}^d$, the prior is $\pi = \mathcal{N}(0, s^2 I)$ and that we use the variational approximation family : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix} \right)$.

An example : SVB with Gaussian approximations

As an example, assume that $\theta \in \mathbb{R}^d$, the prior is $\pi = \mathcal{N}(0, s^2 I)$ and that we use the variational approximation family : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix} \right)$.

In this case, the update in SVB is :

$$
\begin{aligned}
m_{t+1} &= m_t - \alpha \sigma_t^2 \odot \nabla_{m=m_t} \mathbb{E}_{\theta \sim q_{m,\sigma_t}} [-\log p_\theta(x_t)] \\
\sigma_{t+1} &= \sigma_t \odot h \left(\frac{\alpha \sigma_t \nabla_{\sigma=\sigma_t} \mathbb{E}_{\theta \sim q_{m_t,\sigma}} [-\log p_\theta(x_t)]}{2} \right)
\end{aligned}
$$

where $\odot$ means "componentwise multiplication" and $h(x) = \sqrt{1+x^2} - x$ is also applied componentwise.

An example : SVB with Gaussian approximations

As an example, assume that $\theta \in \mathbb{R}^d$, the prior is $\pi = \mathcal{N}(0, s^2 I)$ and that we use the variational approximation family : $q_\mu = q_{m,\sigma} = \mathcal{N}\left(m, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix}\right)$.

In this case, the update in SVB is :

$$
\begin{aligned}
m_{t+1} &= m_t - \alpha \sigma_t^2 \odot \nabla_{m=m_t} \mathbb{E}_{\theta \sim q_{m,\sigma_t}} [-\log p_\theta(x_t)] \\
\sigma_{t+1} &= \sigma_t \odot h\left(\frac{\alpha \sigma_t \nabla_{\sigma=\sigma_t} \mathbb{E}_{\theta \sim q_{m_t,\sigma}} [-\log p_\theta(x_t)]}{2} \right)
\end{aligned}
$$

where $\odot$ means "componentwise multiplication" and $h(x) = \sqrt{1+x^2} - x$ is also applied componentwise. We also have explicit formulas for SVA and NGVI (see the paper).

A regret bound for SVA

Theorem (Chérief-Abdellatif, A. & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex.

A regret bound for SVA

Theorem (Chérief-Abdellatif, A. & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta \sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. (this is for example the case as soon as the log-likelihood is concave in θ and L-Lipschitz, and μ is a location-scale parameter).

A regret bound for SVA

Theorem (Chérief-Abdellatif, A. & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. Assume that $\mu \mapsto \mathcal{K}(p_\mu,\pi)$ is γ-strongly convex. Then SVA satisfies :

$$
\begin{aligned}
&\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)] \\
&\leq \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}[-\log p_\theta(x_t)]\right] + \frac{\alpha L^2 T}{\gamma} + \frac{\mathcal{K}(q_\mu,\pi)}{\alpha}\right\}.
\end{aligned}
$$

A regret bound for SVA

Theorem (Chérief-Abdellatif, A. & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. Assume that $\mu \mapsto \mathcal{K}(p_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)]$$

$$\leq \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}[-\log p_\theta(x_t)]\right] + \frac{\alpha L^2 T}{\gamma} + \frac{\mathcal{K}(q_\mu,\pi)}{\alpha}\right\}.$$

For SVB : some results in the Gaussian case.

A regret bound for SVA

Theorem (Chérief-Abdellatif, A. & Khan)

Assume that $\mu \mapsto \mathbb{E}_{\theta\sim q_\mu}[-\log p_\theta(x_t)]$ is L-Lipschitz and convex. Assume that $\mu \mapsto \mathcal{K}(p_\mu, \pi)$ is γ-strongly convex. Then SVA satisfies :

$$\sum_{t=1}^{T}[-\log p_{\theta_t}(x_t)]$$
$$\leq \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}[-\log p_\theta(x_t)]\right] + \frac{\alpha L^2 T}{\gamma} + \frac{\mathcal{K}(q_\mu,\pi)}{\alpha}\right\}.$$

For SVB : some results in the Gaussian case. For NGVI : we were not able to derive regret bounds until now.

Test on a simulated dataset

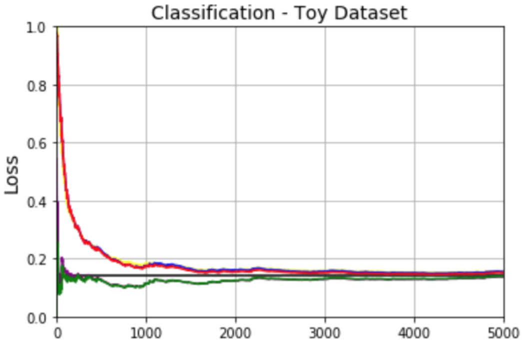


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Breast dataset

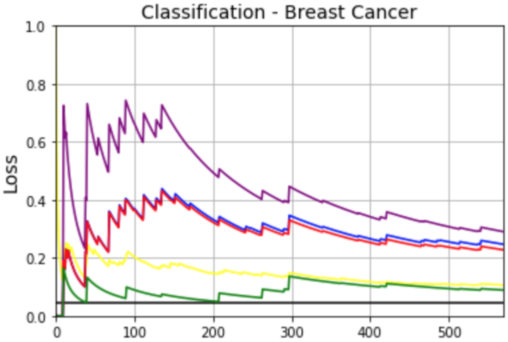


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Pima Indians dataset

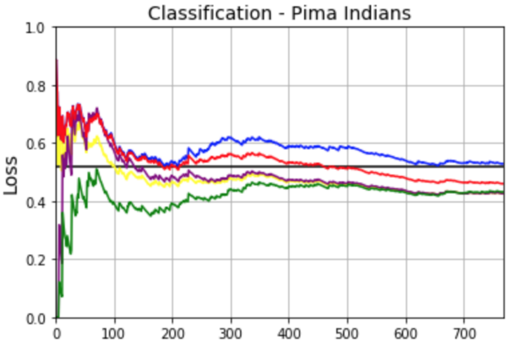


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Boston Housing dataset

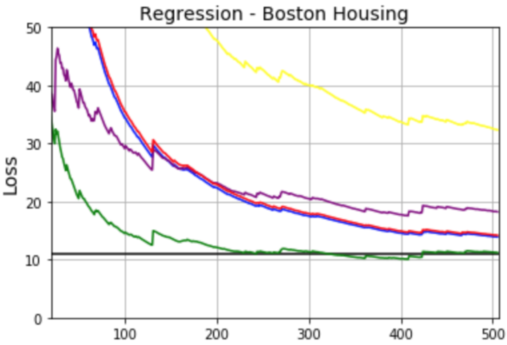


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Forest Cover Type dataset

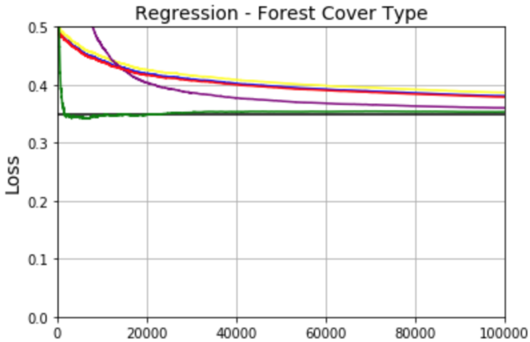


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Conclusions

1. Using online-to-batch conversion, we now have algorithms for variational inference with provable statistical properties after a finite number of steps.

Conclusions

1. Using online-to-batch conversion, we now have algorithms for variational inference with provable statistical properties after a finite number of steps.
2. SVA, SVB competitive with OGA (online gradient algorithm, "non-Bayesian").

Conclusions

1. Using online-to-batch conversion, we now have algorithms for variational inference with provable statistical properties after a finite number of steps.
2. SVA, SVB competitive with OGA (online gradient algorithm, "non-Bayesian").
3. NGVI is the best method on all datasets. Its theoretical analysis is thus an important open problem. Cannot be done with our current techniques (using natural parameters in exponential models lead to non-convex objectives).

Thank you !