Approximate Bayesian Inference

Study of a few algorithms

Pierre Alquier

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

Università degli Studi di Padova, May 10, 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 $X_1, \ldots, X_n$ 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

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

The posterior

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

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

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

The posterior

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

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

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

Various reasons to use a tempered posterior

- more robust to model misspecification (at least empirically)

P. Grünwald. The Safe Bayesian : Learning the Learning Rate via the Mixability Gap *ALT* 2012.

Various reasons to use a tempered posterior

- more robust to model misspecification (at least empirically)

P. Grünwald. The Safe Bayesian : Learning the Learning Rate via the Mixability Gap *ALT* 2012.

- theoretical analysis easier

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

Computation of the posterior

- explicit form (conjugate models),

Computation of the posterior

- explicit form (conjugate models),
- MCMC algorithms.

Computation of the posterior

- explicit form (conjugate models),
- MCMC algorithms. Example : **Metropolis-Hastings**.

Metropolis-Hastings Algorithm (MH)

- arbitraty θ_0,
- given θ_n,
  1. draw $t_{n+1} \sim q(\cdot|\theta_n)$,
  2. $\theta_{n+1} = \begin{cases} t_{n+1} \text{ with probability } a(\theta_n, t_{n+1}) \\ \theta_n \text{ otherwise.} \end{cases}$

$$a(\theta, t) = \min\left[\frac{\pi_{n,\alpha}(t)q(\theta|t)}{\pi_{n,\alpha}(\theta)q(t|\theta)}, 1\right].$$

But...

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

But...

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

But...

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

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** ...

Outline of the talk

Outline of the talk

1. Introduction : algorithms for Bayesian inference

2. Noisy MCMC
   - Noisy MCMC : definition, and motivating example
   - Convergence study of noisy MCMC
   - Subsampling in MCMC

3. Variational approximations
   - Variational approximations : definition
   - Consistency of variational approximations
   - Applications

Co-authors on this project

Nial Friel

Florian Maire

Aidan Boland

Richard Everitt

Noisy Metropolis-Hastings

Metropolis-Hastings Algorithm

- arbitraty θ_0,
- given θ_n,
  1. draw $t_{n+1} \sim q(\cdot|\theta_n)$,
  2. $\theta_{n+1} = \begin{cases} t_{n+1} \text{ with probability } a(\theta_n, t_{n+1}) \\ \theta_n \text{ otherwise,} \end{cases}$

Noisy Metropolis-Hastings

Noisy Metropolis-Hastings Algorithm

- arbitraty θ_0,
- given θ_n,
  1. draw $t_{n+1} \sim q(\cdot|\theta_n)$,
  2. $\theta_{n+1} = \begin{cases} t_{n+1} \text{ with probability } \hat{a}(\theta_n, t_{n+1}, S_n) \\ \theta_n \text{ otherwise,} \end{cases}$

where $\hat{a}(\theta, t, S)$ is a numerical approximation of

$$a(\theta, t) = \min\left[\frac{\pi_{n,\alpha}(t)q(\theta|t)}{\pi_{n,\alpha}(\theta)q(t|\theta)}, 1\right]$$

that can be based (or not !) on some simulated r.v. S.

A motivating example

Example : Exponential Random Graph Model (ERGM)

Given a set of nodes $\{1, \dots, n\}$, and x a graph on these nodes represented by the adjacency matrix $x_{i,j} = 1 \Leftrightarrow$ "i and j are connected", and $s(x)$ be a vector of statistics. We define :

$$p_\theta(x) = \frac{\exp(\theta^T s(x))}{\sum_y \exp(\theta^T s(x))} = \frac{\exp(\theta^T s(x))}{Z(\theta)}.$$

A motivating example

Example : Exponential Random Graph Model (ERGM)

Given a set of nodes $\{1, \dots, n\}$, and x a graph on these nodes represented by the adjacency matrix $x_{i,j} = 1 \Leftrightarrow$ "i and j are connected", and $s(x)$ be a vector of statistics. We define :

$$p_\theta(x) = \frac{\exp(\theta^T s(x))}{\sum_y \exp(\theta^T s(x))} = \frac{\exp(\theta^T s(x))}{Z(\theta)}.$$

Then

$$a(\theta, t) = \min\left[\frac{\pi(t)\left[\exp(t^T s(x)) Z(\theta)\right]^\alpha q(\theta|t)}{\pi(\theta)\left[\exp(\theta^T s(x)) Z(t)\right]^\alpha q(t|\theta)}, 1\right].$$

Approximation of $a(\cdot,\cdot)$ in ERGM

$$a(\theta,t) = \min\left[\frac{\pi(t)\left[\exp(t^T s(x)) Z(\theta)\right]^{\alpha} q(\theta|t)}{\pi(\theta)\left[\exp(\theta^T s(x)) Z(t)\right]^{\alpha} q(t|\theta)}, 1\right].$$

and we cannot compute Z.

Approximation of $a(\cdot,\cdot)$ in ERGM

$$a(\theta,t) = \min\left[\frac{\pi(t)\left[\exp(t^T s(x)) Z(\theta)\right]^\alpha q(\theta|t)}{\pi(\theta)\left[\exp(\theta^T s(x)) Z(t)\right]^\alpha q(t|\theta)}, 1\right].$$

and we cannot compute Z. However,

$$\mathbb{E}_{x\sim P_t}\left(\frac{\exp(\theta^T s(x))}{\exp(t^T s(x))}\right) = \sum_x \frac{\exp(\theta^T s(x))}{\exp(t^T s(x))}\frac{\exp(t^T s(x))}{Z(t)} = \frac{Z(\theta)}{Z(t)}$$

Approximation of $a(\cdot,\cdot)$ in ERGM

$$a(\theta,t) = \min\left[\frac{\pi(t)\left[\exp(t^T s(x)) Z(\theta)\right]^{\alpha} q(\theta|t)}{\pi(\theta)\left[\exp(\theta^T s(x)) Z(t)\right]^{\alpha} q(t|\theta)}, 1\right].$$

and we cannot compute Z. However,

$$\mathbb{E}_{x\sim P_t}\left(\frac{\exp(\theta^T s(x))}{\exp(t^T s(x))}\right) = \sum_x \frac{\exp(\theta^T s(x))}{\exp(t^T s(x))}\frac{\exp(t^T s(x))}{Z(t)} = \frac{Z(\theta)}{Z(t)}$$

so we can draw $S_N = (x_1,\ldots,x_N)$ iid from P_t (feasible) and

$$\hat{a}(\theta,t,S_N) = \min\left[\frac{\pi(t)\exp(\alpha t^T s(x)) q(\theta)}{\pi(\theta)\exp(\alpha\theta^T s(x)) q(t)}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\theta^T s(x_i))}{\exp(t^T s(x_i))}\right)^{\alpha}, 1\right].$$

Theoretical study of noisy MCMC

Note that noisy MCMC produces a Markov chain, but there is no reason for $\pi_{n,\alpha}$ to be invariant for this chain.

Theoretical study of noisy MCMC

Note that noisy MCMC produces a Markov chain, but there is no reason for $\pi_{n,\alpha}$ to be invariant for this chain. However :

Theorem

Assume :

- $\mathbb{E}_S \left|a(\theta,t) - \hat{a}(\theta,t,S)\right| \leq \delta(\theta,t)$.
- The kernel P associated with $a(\theta,t)$ is uniformly ergodic :

$$\forall \theta_0, \quad \|\delta_{\theta_0} P^M - \pi_{n,\alpha}\|_{\mathrm{TV}} \leq C\rho^M .$$

Then $\|\delta_{\theta_0} P^M - \delta_{\theta_0} \hat{P}^M\|_{\mathrm{TV}} \leq 2K(C,\rho) \sup_{\theta} \int q(\mathrm{d}t|\theta)\delta(\theta,t)$

where $\hat{P}$ is the kernel of noisy MCMC, $K(C,\rho)$ is known.

Noisy MCMC for ERGM

Corollary for ERGM

Assume that

- the parameter space is bounded : $\sup_{\theta\in\Theta} \|\theta\| = \mathcal{T} < \infty$,
- there is a $c > 0$ such that $c \leq \pi(\theta), q(\theta|t) \leq 1/c$.

$$\text{Then :} \quad \|\delta_{\theta_0} P^M - \delta_{\theta_0} \hat{P}^M\|_{\mathrm{TV}} \leq \frac{\mathcal{C}(\mathcal{T}, c, s)}{\sqrt{N}}.$$

P. Alquier, N. Friel, R. G. Everitt & A. Boland. Noisy Monte-Carlo : Convergence of Markov Chains with Approximate Transition Kernels. *Statistics and Computing*, 2016.

Noisy MCMC for ERGM

Corollary for ERGM

Assume that

- the parameter space is bounded : $\sup_{\theta \in \Theta} \|\theta\| = \mathcal{T} < \infty$,
- there is a $c > 0$ such that $c \leq \pi(\theta), q(\theta|t) \leq 1/c$.

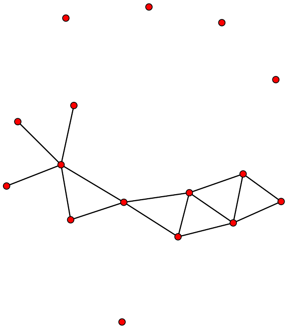
$$\text{Then :} \quad \|\delta_{\theta_0} P^M - \delta_{\theta_0} \hat{P}^M\|_{\mathrm{TV}} \leq \frac{\mathcal{C}(\mathcal{T}, c, s)}{\sqrt{N}}.$$

P. Alquier, N. Friel, R. G. Everitt & A. Boland. Noisy Monte-Carlo : Convergence of Markov Chains with Approximate Transition Kernels. *Statistics and Computing*, 2016.

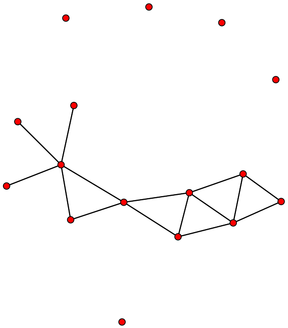
Important generalization to the geometrically ergodic P, using the Wasserstein distance rather than total variation :

D. Rudolf & N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 2018.

Simulations : Florentine Family Business Dataset



Simulations : Florentine Family Business Dataset



$$s(x) = (s_1(x), s_2(x))$$

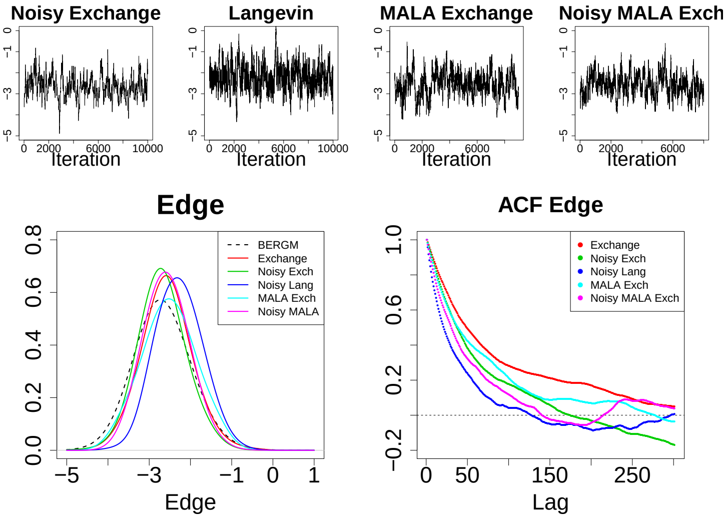
- $s_1(x)$ number of edges,
- $s_2(x)$ number of 2-stars.

Numerical Results

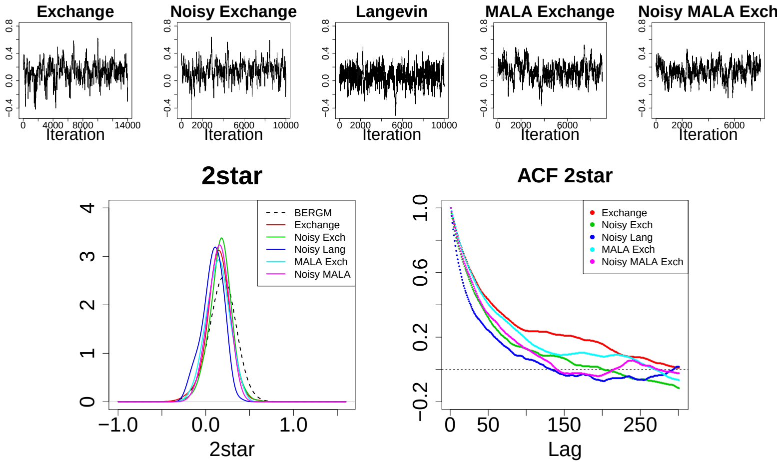
	Edge		2-star	
Method	Mean	SD	Mean	SD
BERGM	-2.675	0.647	0.188	0.155
Exchange	-2.573	0.568	0.146	0.133
Noisy Exchange	-2.686	0.526	0.167	0.122
Noisy Langevin	-2.281	0.513	0.081	0.119
MALA Exchange	-2.518	0.62	0.136	0.128
Noisy MALA	-2.584	0.498	0.144	0.113

Table – Posterior means and standard deviations.

Chains, density and ACF plot for the edge statistic.

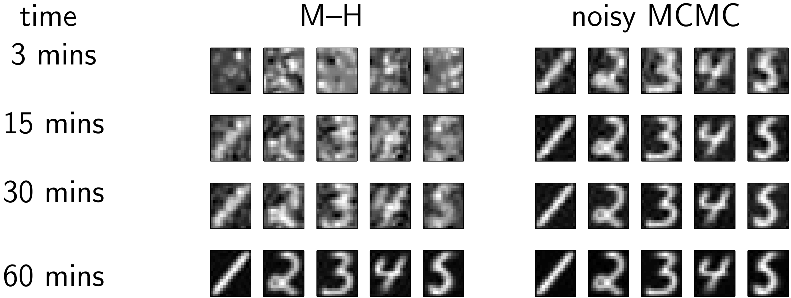


Chains, density and ACF plot for the 2-star stat.



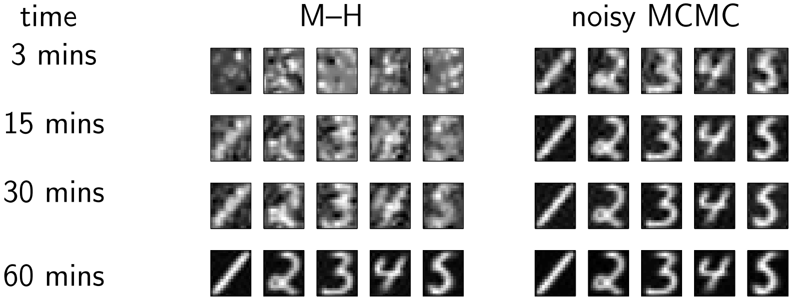
Subsampling in MCMC

Idea to approximate $\hat{a}$ when the sample size n is too large : evaluate $\hat{a}$ on a subsample of the data.



Subsampling in MCMC

Idea to approximate $\hat{a}$ when the sample size n is too large : evaluate $\hat{a}$ on a subsample of the data.



F. Maire, N. Friel, P. Alquier, Informed Sub-Sampling MCMC : Approximate Bayesian Inference for Large Datasets. *Statistics and Computing*, 2019.

Outline of the talk

1. Introduction : algorithms for Bayesian inference
2. Noisy MCMC
   - Noisy MCMC : definition, and motivating example
   - Convergence study of noisy MCMC
   - Subsampling in MCMC
3. Variational approximations
   - Variational approximations : definition
   - Consistency of variational approximations
   - Applications

Co-authors on this project

James Ridgway

Nicolas Chopin

Co-authors on this project

James Ridgway

Nicolas Chopin

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

Co-authors on this project

James Ridgway

Nicolas Chopin

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

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

Examples :

Variational approximations

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

Examples :

- parametric approximation

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

Variational approximations

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

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

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

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 $X_1, \dots, X_n$ 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 $\alpha \in (0,1)$,

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

Example 1 : Gaussian VB

- Let $\Theta = \mathbb{R}^p$.

Example 1 : Gaussian VB

- Let $\Theta = \mathbb{R}^p$.
- We start with the family of approximations

$$\mathcal{F}_{\mathcal{G}}^{\Phi} := \left\{ \Phi(d\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathcal{G} \subset \mathcal{S}_+^d(\mathbb{R}) \right\},$$

Example 1 : Gaussian VB

- Let $\Theta = \mathbb{R}^p$.
- We start with the family of approximations

$$\mathcal{F}_{\mathcal{G}}^{\Phi} := \left\{ \Phi(d\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathcal{G} \subset \mathcal{S}_+^d(\mathbb{R}) \right\},$$

- We assume that for a model $\{p_\theta, \theta \in \Theta\}$ there exists a measurable real valued function $M(\cdot)$ such that

$$|\log p_\theta(X_1) - \log p_{\theta'}(X_1)| \leq M(X_1) \|\theta - \theta'\|_2$$

Furthermore we assume that
$\mathbb{E}M(X_1) =: B_1, \quad \mathbb{E}M^2(X_1) =: B_2 < \infty.$

Application of the result

Theorem

Let the family of approximation be $\mathcal{F}$ with $\mathcal{F}^{\Phi}_{\sigma^2 I} \subset \mathcal{F}$ as defined above. We put

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee C\frac{d}{n}\log n$$

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

$$\mathbb{E}\left[\int D_{\alpha}(P_{\theta}, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Stochastic Variational Bayes

- To implement the idea we write

$$\mathcal{F}_B^{\Phi} = \left\{ \Phi(d\theta; m, CC^t), \quad (m, C) \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d \right\}.$$

$$F : x = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} \mapsto \mathbb{E}\left[f(x, \xi) \right] = \mathcal{K}(\rho_{m,C}, \pi_n)$$

where $\xi \sim \mathcal{N}(0, I_d)$

Stochastic Variational Bayes

- To implement the idea we write

$$\mathcal{F}_B^{\Phi} = \left\{ \Phi(d\theta; m, CC^t), \quad (m, C) \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d \right\}.$$

$$F : x = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} \mapsto \mathbb{E}\left[f(x, \xi) \right] = \mathcal{K}(\rho_{m,C}, \pi_n)$$

where $\xi \sim \mathcal{N}(0, I_d)$

- The optimization problem can be written

$$\min_{x \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d} \mathbb{E}\left[f(x, \xi) \right],$$

where

$$f((m, C), \xi) := \log p_{m + C\xi}(Y_1^n) + \log \frac{\mathrm{d}\Phi_{m, CC^t}}{\mathrm{d}\pi}(m + C\xi)$$

We can use stochastic gradient descent

Algorithm 1 Stochastic VB

Input : x_0, X_1^n, γ_T

For $i \in \{1, \cdots, T\}$,

a. Sample $\xi_t \sim \mathcal{N}(0, I_d)$

b. Update

$$x_t \leftarrow \mathcal{P}_{\mathbb{B}}\left(x_{t-1} - \gamma_T \nabla f(x_{t-1}, \xi_t)\right)$$

End For .

Output : $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$

where ∇f is the gradient of the integrand in the objective function

- Assume that f is convex in its first component x and that it has L-Lipschitz gradients.
- Define $\tilde{\pi}^{k}_{n,\alpha}(\mathrm{d}\theta|X_1^n)$ to be the k-th iterate of the algorithm

Theorem

For some C,

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee \left\{ \frac{d}{n} \left[\frac{1}{2} \log\left(\vartheta^2 n^2 C\right) + \frac{1}{n\vartheta^2} \right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n} \right\}$$

with $\gamma_k = \frac{B}{L\sqrt{2k}}$, we get

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}^{k}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{1}{1-\alpha}\sqrt{\frac{2BL}{k}}.$$

Example 2 : nonparametric regression

Nonparametric regression

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

Example 2 : nonparametric regression

Nonparametric regression

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

Example 2 : 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,

Example 2 : 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...

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

Example 3 : matrix completion

In

P. Alquier, J. Ridgway , N. Chopin. On the Properties of Variational Approximations of Gibbs Posteriors. *JMLR*, 2016.

we proved that the variational approximations used in the matrix completion problem do not change the rate of convergence.

Example 4 : model selection

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

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

Example 4 : model selection

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

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

Theorem

If the true model is actually k_0,

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

Example 4 : model selection

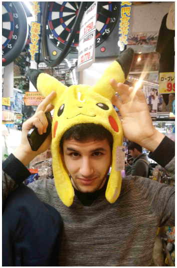
Theorem

If the true model is actually k_0,

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

This result is actually due to my PhD student Badr-Eddine Chérief-Abdellatif.

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



Example 5 : mixture models

VB for mixtures

- $P_{p,\theta} = \sum_{i=1}^{K} p_i q_{\theta_i}$,

Example 5 : mixture models

VB for mixtures

- $P_{p,\theta} = \sum_{i=1}^{K} p_i q_{\theta_i}$,
- VB approximation : the θ_i's are mutually independent and independent from $(p_1, \dots, p_K)$.

Example 5 : mixture models

VB for mixtures

- $P_{p,\theta} = \sum_{i=1}^{K} p_i q_{\theta_i}$,
- VB approximation : the θ_i's are mutually independent and independent from $(p_1, \dots, p_K)$.

Under suitable assumptions, $r_n \sim \frac{K \log(n)}{n}$.

Example 5 : mixture models

VB for mixtures

- $P_{p,\theta} = \sum_{i=1}^{K} p_i q_{\theta_i}$,
- VB approximation : the θ_i's are mutually independent and independent from $(p_1, \dots, p_K)$.

Under suitable assumptions, $r_n \sim \frac{K \log(n)}{n}$.

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

What's next ?

Case $\alpha = 1$

$$[L(\theta)]^{\alpha}\,\pi(\mathrm{d}\theta) = L(\theta)\pi(\mathrm{d}\theta)$$

F. Zhang & C. Gao (2017). Convergence Rates of Variational Posterior Distributions. *Preprint arxiv :1712.02519.*

What's next ?

Case $\alpha = 1$

$$[L(\theta)]^{\alpha} \, \pi(\mathrm{d}\theta) = L(\theta)\pi(\mathrm{d}\theta)$$

F. Zhang & C. Gao (2017). Convergence Rates of Variational Posterior Distributions. *Preprint arxiv :1712.02519.*

- we only proved pointwise convergence. What would be conditions ensuring that credible intervals given by the variational approximation are correct ?

What's next ?

Case $\alpha = 1$

$$[L(\theta)]^{\alpha}\,\pi(\mathrm{d}\theta) = L(\theta)\pi(\mathrm{d}\theta)$$

F. Zhang & C. Gao (2017). Convergence Rates of Variational Posterior Distributions. *Preprint arxiv :1712.02519.*

- we only proved pointwise convergence. What would be conditions ensuring that credible intervals given by the variational approximation are correct ?
- many recent papers study approximations based on other divergences or distances than $\mathcal{K}$: Rényi, Wasserstein, ...

What's next ?

Case $\alpha = 1$

$$[L(\theta)]^{\alpha} \pi(\mathrm{d}\theta) = L(\theta)\pi(\mathrm{d}\theta)$$

F. Zhang & C. Gao (2017). Convergence Rates of Variational Posterior Distributions. *Preprint arxiv :1712.02519.*

- we only proved pointwise convergence. What would be conditions ensuring that credible intervals given by the variational approximation are correct ?
- many recent papers study approximations based on other divergences or distances than $\mathcal{K}$: Rényi, Wasserstein, ...
- analysis of online variational inference (work in progress with Emti Khan and Badr-Eddine Chérief-Abdellatif)...

Thank you !