Lectures on Variational Inference

2) Statistical Analysis of Variational Approximations

Pierre Alquier

RIKEN

AIP
Center for
Advanced Intelligence Project

Heilbronn Institute – University of Bristol – Nov. 27, 2019

Lecture 2

Reminder on the notations

Reminder on the notations

From now, and until Section 3, we will focus on parametric estimation in statistics.

Reminder on the notations

From now, and until Section 3, we will focus on parametric estimation in statistics.

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\{P_\theta, \theta \in \Theta\}$ with p.d.f p_θ. Prior π on Θ.

Reminder on the notations

From now, and until Section 3, we will focus on parametric estimation in statistics.
Assume that we observe $X_1, \dots, X_n$ i.i.d from P^0 in a model $\{P_\theta, \theta \in \Theta\}$ with p.d.f p_θ. Prior π on Θ. Until Section 3, we also assume a correct specification : $P^0 = P_{\theta_0}$.

Reminder on the notations

From now, and until Section 3, we will focus on parametric estimation in statistics.

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\{P_\theta, \theta \in \Theta\}$ with p.d.f p_θ. Prior π on Θ. Until Section 3, we also assume a correct specification : $P^0 = P_{\theta_0}$.

The likelihood

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

Reminder on the notations

From now, and until Section 3, we will focus on parametric estimation in statistics.
Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\{P_\theta, \theta \in \Theta\}$ with p.d.f p_θ. Prior π on Θ. Until Section 3, we also assume a correct specification : $P^0 = P_{\theta_0}$.

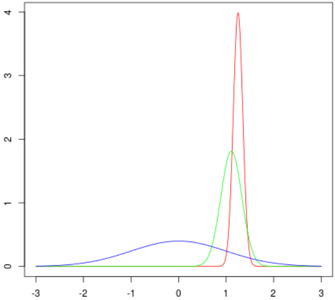
The likelihood

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

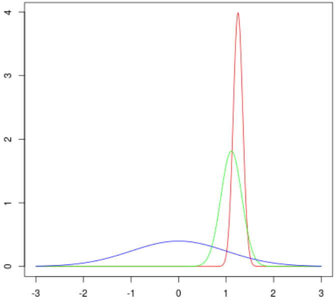
The tempered posterior - $0 < \alpha < 1$

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

Concentration of the posterior



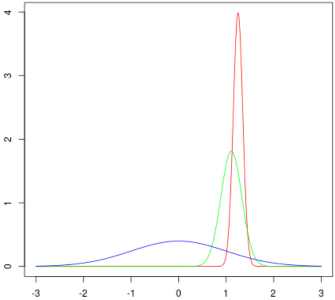
Concentration of the posterior



Do we have for some d, $\forall t > 0$

$$\mathbb{P}_{\theta \sim \pi_{n,\alpha}} \left[d(\theta, \theta_0) \geq t \right] \xrightarrow[n \to \infty]{} 0\,?$$

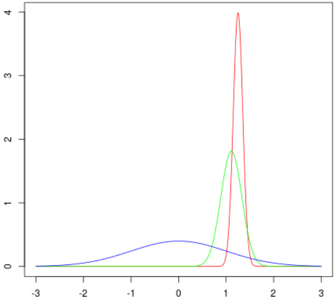
Concentration of the posterior

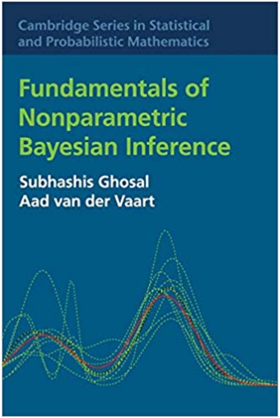


Do we have for some d, $\forall t > 0$

$$\mathbb{P}_{\theta \sim \pi_{n,\alpha}} \left[d(\theta, \theta_0) \geq t \right] \xrightarrow[n \to \infty]{?} 0\,?$$

Concentration of the posterior





Do we have for some d, $\forall t > 0$

$$\mathbb{P}_{\theta \sim \pi_{n,\alpha}}\left[d(\theta,\theta_0) \geq t\right] \xrightarrow[n\to\infty]{?} 0\,?$$

A simpler result

We will state a simpler results than the concentration theorems in this book, by following ideas from

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

A simpler result

We will state a simpler results than the concentration theorems in this book, by following ideas from

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

Definition - Rényi divergence

Assume that P and Q have respective densities p and q with respect to a measure μ, define, for $0 < \alpha < 1$,

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

A simpler result

We will state a simpler results than the concentration theorems in this book, by following ideas from

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

Definition - Rényi divergence

Assume that P and Q have respective densities p and q with respect to a measure μ, define, for $0 < \alpha < 1$,

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

We use $d(\theta, \theta_0) = D_\alpha(P_\theta, P_{\theta_0})$ to measure the concentration.

Properties of the Rényi divergence

It is important to note that $D_\alpha(P,Q)$ does not depend on the choice of μ as can be "seen" from

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

Properties of the Rényi divergence

It is important to note that $D_\alpha(P, Q)$ does not depend on the choice of μ as can be "seen" from

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

Many properties derived in :

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

Properties of the Rényi divergence

It is important to note that $D_\alpha(P, Q)$ does not depend on the choice of μ as can be "seen" from

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

Many 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 :

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

A theorem in expectation

Define

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

A theorem in expectation

Define

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

Theorem (BPY 19, simplified version)

For any sequence (r_n) such that $r_n \geq 0$ and

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

we have

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]\right\} \leq \frac{1+\alpha}{1-\alpha} r_n.$$

From expectation to concentration

From the previous result, we can recover concentration results by a simple application of Markov inequality :

From expectation to concentration

From the previous result, we can recover concentration results by a simple application of Markov inequality :

$$\mathbb{P}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}) \geq t\right] \leq \frac{\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]}{t},$$

From expectation to concentration

From the previous result, we can recover concentration results by a simple application of Markov inequality :

$$\mathbb{P}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}) \geq t\right] \leq \frac{\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]}{t},$$

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{P}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}) \geq t\right]\right\} \leq \frac{(1+\alpha)}{(1-\alpha)t} r_n \xrightarrow[n\to\infty]{} 0,$$

From expectation to concentration

From the previous result, we can recover concentration results by a simple application of Markov inequality :

$$\mathbb{P}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}) \geq t\right] \leq \frac{\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]}{t},$$

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{P}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}) \geq t\right]\right\} \leq \frac{(1+\alpha)}{(1-\alpha)t} r_n \xrightarrow[n\to\infty]{} 0,$$

$$\mathbb{P}_{\theta\sim\pi_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}) \geq t\right] \xrightarrow[n\to\infty]{\text{proba.}} 0.$$

Generalization to variational approximations

Generalization to variational approximations

In the following paper, we extended BPY's approach to variational approximations.

P. Alquier & J. Ridgway (2017). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics (to appear), preprint arxiv :1706.09293.*

Generalization to variational approximations

In the following paper, we extended BPY's approach to variational approximations.

P. Alquier & J. Ridgway (2017). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics (to appear), preprint arxiv :1706.09293.*

Generalization to variational approximations

In the following paper, we extended BPY's approach to variational approximations.

P. Alquier & J. Ridgway (2017). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics (to appear), preprint arxiv :1706.09293.*

Reminder :

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

Concentration of VB

Theorem - AR 17

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

$$\mathbb{E}_{\theta\sim\rho_n}\left[\mathcal{K}(P_{\theta_0}, P_\theta)\right] \leq r_n$$

and

$$\mathcal{K}(\rho_n, \pi) \leq n r_n.$$

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

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]\right\} \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Proof (1/2)

Define the log-likelihood ratio : $r_n(\theta, \theta_0) = \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}$.

Proof (1/2)

Define the log-likelihood ratio : $r_n(\theta, \theta_0) = \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}$.

1. Its expectation :

$$\mathbb{E}_{X_1,\dots,X_n}[r_n(\theta, \theta_0)] = n\mathcal{K}(P_{\theta_0}, P_\theta).$$

Proof (1/2)

Define the log-likelihood ratio : $r_n(\theta,\theta_0) = \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}$.

1. Its expectation :

$$\mathbb{E}_{X_1,\dots,X_n}\left[r_n(\theta,\theta_0)\right] = n\mathcal{K}(P_{\theta_0},P_\theta).$$

2. Its exponential moment :

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\exp\left[-\alpha r_n(\theta,\theta_0)\right]\right\} = \prod_{i=1}^{n}\int\left(\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)}\right)^{\alpha} p_{\theta_0}(x_i)\mathrm{d}x_i$$

$$= \prod_{i=1}^{n}\exp\left[\log\int p_\theta(x_i)^{\alpha}p_{\theta_0}(x_i)^{1-\alpha}\mathrm{d}x_i\right]$$

$$= \exp\left[-n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right].$$

Proof (2/2)

Start from the exponential moment :

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\exp\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right]\right\}=1.$$

Proof (2/2)

Start from the exponential moment :

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\exp\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right]\right\}=1.$$

Expectation w.r.t $\theta \sim \pi$ and Fubini :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\pi}\left\{\exp\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right]\right\}=1.$$

Proof (2/2)

Start from the exponential moment :

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\exp\left[-\alpha r_n(\theta,\theta_0) + n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right]\right\} = 1.$$

Expectation w.r.t $\theta \sim \pi$ and Fubini :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\pi}\left\{\exp\left[-\alpha r_n(\theta,\theta_0) + n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right]\right\} = 1.$$

Then :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left\{\exp\left[-\alpha r_n(\theta,\theta_0) + n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0}) + \log\left(\frac{\pi(\theta)}{\tilde{\pi}_{n,\alpha}(\theta)}\right)\right]\right\} = 1.$$

Proof (2/2)

Then :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left\{\exp\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right.\right.$$
$$\left.\left.+\log\left(\frac{\pi(\theta)}{\tilde{\pi}_{n,\alpha}(\theta)}\right)\right]\right\}=1.$$

Jensen's inequality :

$$\exp\left\{\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right.\right.$$
$$\left.\left.+\log\left(\frac{\pi(\theta)}{\tilde{\pi}_{n,\alpha}(\theta)}\right)\right]\right\}\leq 1.$$

Proof (2/2)

Jensen's inequality :

$$\exp\left\{\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0})\right.\right.$$
$$\left.\left.+\log\left(\frac{\pi(\theta)}{\tilde{\pi}_{n,\alpha}(\theta)}\right)\right]\right\}\leq 1.$$

Rearranging :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta,P_{\theta_0}\right]$$
$$\leq\frac{1}{n(1-\alpha)}\mathbb{E}_{X_1,\dots,X_n}\left\{\underbrace{\alpha\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[r_n(\theta,\theta_0)\right]+\mathcal{K}(\tilde{\pi}_{n,\alpha},\pi)}\right\}$$

Proof (2/2)

Jensen's inequality :

$$\exp\left\{\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[-\alpha r_n(\theta,\theta_0)+n(1-\alpha)D_\alpha(P_\theta,P_{\theta_0}) + \log\left(\frac{\pi(\theta)}{\tilde{\pi}_{n,\alpha}(\theta)}\right)\right]\right\}\leq 1.$$

Rearranging :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta,P_{\theta_0}\right] \leq \frac{1}{n(1-\alpha)}\mathbb{E}_{X_1,\dots,X_n}\left\{\underbrace{\alpha\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[r_n(\theta,\theta_0)\right]+\mathcal{K}(\tilde{\pi}_{n,\alpha},\pi)}_{=-\alpha\mathrm{ELBO}(\tilde{\pi}_{n,\alpha})+\text{ constant}}\right\}$$

Proof (2/2)

Rearranging :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}\right]$$
$$\leq \frac{1}{n(1-\alpha)}\mathbb{E}_{X_1,\dots,X_n}\left\{\underbrace{\alpha\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[r_n(\theta,\theta_0)\right] + \mathcal{K}(\tilde{\pi}_{n,\alpha},\pi)}_{=-\alpha\mathrm{ELBO}(\tilde{\pi}_{n,\alpha})+\text{ constant}}\right\}$$

As $\pi_{n,\alpha}$ minimizes the ELBO :

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0}\right]$$
$$\leq \frac{1}{n(1-\alpha)}\mathbb{E}_{X_1,\dots,X_n}\left\{\inf_{\rho\in\mathcal{F}}\left[\alpha\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta,\theta_0)\right] + \mathcal{K}(\rho,\pi)\right]\right\}$$
$$\leq \frac{1}{n(1-\alpha)}\inf_{\rho\in\mathcal{F}}\left\{n\alpha\mathbb{E}_{\theta\sim\rho}\left[\mathcal{K}(P_{\theta_0},P_\theta)\right] + \mathcal{K}(\rho,\pi)\right\}.$$

Proof (2/2)

As $\pi_{n,\alpha}$ minimizes the ELBO :

$$\begin{aligned}
&\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta,P_{\theta_0}\right] \\
&\quad\leq \frac{1}{n(1-\alpha)}\mathbb{E}_{X_1,\dots,X_n}\left\{\inf_{\rho\in\mathcal{F}}\left[\alpha\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta,\theta_0)\right]+\mathcal{K}(\rho,\pi)\right]\right\} \\
&\quad\leq \frac{1}{n(1-\alpha)}\inf_{\rho\in\mathcal{F}}\left\{n\alpha\mathbb{E}_{\theta\sim\rho}\left[\mathcal{K}(P_{\theta_0},P_\theta)\right]+\mathcal{K}(\rho,\pi)\right\}.
\end{aligned}$$

We end the proof by using the assumption that there is a $\rho\in\mathcal{F}$ such that $\mathbb{E}_{\theta\sim\rho}[\mathcal{K}(P_{\theta_0},P_\theta)]\leq r_n$ and $\mathcal{K}(\rho,\pi)\leq nr_n$:

$$\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta,P_{\theta_0})\right]\leq\frac{1}{n(1-\alpha)}\left[\alpha nr_n+nr_n\right]=\frac{1+\alpha}{1-\alpha}r_n.$$

A toy example : Gaussian variables

Toy example : assume that $X_1, \dots, X_n$ are i.i.d from $P_{\theta_0} = \mathcal{N}(\theta_0, 1)$.

A toy example : Gaussian variables

Toy example : assume that $X_1, \dots, X_n$ are i.i.d from $P_{\theta_0} = \mathcal{N}(\theta_0, 1)$. Cauchy prior $\theta \sim \pi = \mathcal{C}(0,1)$.

A toy example : Gaussian variables

Toy example : assume that $X_1, \dots, X_n$ are i.i.d from $P_{\theta_0} = \mathcal{N}(\theta_0, 1)$. Cauchy prior $\theta \sim \pi = \mathcal{C}(0, 1)$.
Gaussian approximation of the posterior :

$$\mathcal{F} = \left\{ \mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

A toy example : Gaussian variables

Toy example : assume that $X_1, \dots, X_n$ are i.i.d from $P_{\theta_0} = \mathcal{N}(\theta_0, 1)$. Cauchy prior $\theta \sim \pi = \mathcal{C}(0,1)$. Gaussian approximation of the posterior :

$$\mathcal{F} = \left\{ \mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

Note that $\mathcal{K}(P_{\theta_0}, P_\theta) = |\theta - \theta_0|^2/2$ and so

$$\mathbb{E}_{\theta \sim \mathcal{N}(m,\sigma^2)} \left[\mathcal{K}(P_{\theta_0}, P_\theta) \right] = \frac{|m - \theta_0|^2 + \sigma^2}{2}$$

A toy example : Gaussian variables

Toy example : assume that $X_1, \dots, X_n$ are i.i.d from $P_{\theta_0} = \mathcal{N}(\theta_0, 1)$. Cauchy prior $\theta \sim \pi = \mathcal{C}(0,1)$.
Gaussian approximation of the posterior :

$$\mathcal{F} = \left\{ \mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

Note that $\mathcal{K}(P_{\theta_0}, P_\theta) = |\theta - \theta_0|^2/2$ and so

$$\mathbb{E}_{\theta \sim \mathcal{N}(m,\sigma^2)} \left[\mathcal{K}(P_{\theta_0}, P_\theta) \right] = \frac{|m - \theta_0|^2 + \sigma^2}{2}$$

Moreover, rough upper bounds lead to

$$\mathcal{K}(\mathcal{N}(m,\sigma^2), \pi) \leq \log \left(\sqrt{\frac{\pi}{2\sigma^2}} \right) + \log(1+2m^2) + \sqrt{\frac{2\sigma^2}{\pi}}$$

A toy example : Gaussian variables

Toy example : assume that $X_1, \dots, X_n$ are i.i.d from $P_{\theta_0} = \mathcal{N}(\theta_0, 1)$. Cauchy prior $\theta \sim \pi = \mathcal{C}(0, 1)$.
Gaussian approximation of the posterior :

$$\mathcal{F} = \left\{ \mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

Note that $\mathcal{K}(P_{\theta_0}, P_\theta) = |\theta - \theta_0|^2/2$ and so

$$\mathbb{E}_{\theta \sim \mathcal{N}(m,\sigma^2)} \left[\mathcal{K}(P_{\theta_0}, P_\theta) \right] = \frac{|m - \theta_0|^2 + \sigma^2}{2} \leq r_n?$$

Moreover, rough upper bounds lead to

$$\mathcal{K}(\mathcal{N}(m, \sigma^2), \pi) \leq \log\left(\sqrt{\frac{\pi}{2\sigma^2}} \right) + \log(1 + 2m^2) + \sqrt{\frac{2\sigma^2}{\pi}} \leq nr_n?$$

Gaussian example (continued)

$$\frac{|m-\theta_0|^2+\sigma^2}{2} \leq r_n$$

$$\log\left(\sqrt{\frac{\pi}{2\sigma^2}}\right) + \log(1+2m^2) + \sqrt{\frac{2\sigma^2}{\pi}} \leq n r_n$$

Gaussian example (continued)

$$\frac{|m-\theta_0|^2+\sigma^2}{2} \leq r_n$$

$$\log\left(\sqrt{\frac{\pi}{2\sigma^2}}\right) + \log(1+2m^2) + \sqrt{\frac{2\sigma^2}{\pi}} \leq n r_n$$

For example satisfied by $m=\theta_0$, $\sigma^2 = 1/n$ and

$$r_n = \frac{\frac{1}{2}\log\left(\frac{n\pi}{2}\right) + \log(1+2\theta_0^2) + \sqrt{\frac{\pi}{2}}}{n}.$$

Gaussian example (continued)

$$\frac{|m-\theta_0|^2+\sigma^2}{2} \leq r_n$$

$$\log\left(\sqrt{\frac{\pi}{2\sigma^2}}\right) + \log(1+2m^2) + \sqrt{\frac{2\sigma^2}{\pi}} \leq n r_n$$

For example satisfied by $m = \theta_0$, $\sigma^2 = 1/n$ and

$$r_n = \frac{\frac{1}{2}\log\left(\frac{n\pi}{2}\right) + \log(1+2\theta_0^2) + \sqrt{\frac{\pi}{2}}}{n}.$$

We can apply our theorem :

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]\right\} \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Gaussian example (continued)

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]\right\} \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Gaussian example (continued)

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]\right\} \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Here, we have actually

$$D_\alpha(P_\theta, P_{\theta_0}) = \frac{|\theta-\theta_0|^2}{2\alpha}.$$

Gaussian example (continued)

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P_{\theta_0})\right]\right\} \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Here, we have actually

$$D_\alpha(P_\theta, P_{\theta_0}) = \frac{|\theta-\theta_0|^2}{2\alpha}.$$

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[|\theta-\theta_0|^2\right]\right\}$$
$$\leq \left(\frac{\alpha(1+\alpha)}{1-\alpha}\right)\frac{\frac{1}{2}\log\left(\frac{n\pi}{2}\right)+\log(1+2\theta_0^2)+\sqrt{\frac{\pi}{2}}}{n}.$$

Second example : matrix completion

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

Matrix completion (continued)

Reminder on matrix completion :

$$M = \sum_{\ell=1}^{k} U_{\cdot,\ell}(V_{\cdot,\ell})^T \text{ is } p \times m$$

with prior π given by

- $U_{\cdot,\ell}, V_{\cdot,\ell} \sim \mathcal{N}(0, \gamma_\ell I)$,
- $\frac{1}{\gamma_\ell} \sim \mathrm{Gamma}(a, b)$.

Mean-field variational approximation with Gaussian and inverse gamma disitributions on U, V and γ_ℓ respectively.

Matrix completion : rate of convergence

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_M, P_{M_0})\right]\right\} \\ = \mathcal{O}\left(\frac{\mathrm{rank}(M_0)(m+p)+\log(nmp)}{n}\right).$$

Generalization to variational approximations

Misspecified case

Misspecified case

Assume we observe $X_1, \dots, X_n$ i.i.d from P^0 and use a model $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$, but it is possible that $P^0 \notin \mathcal{M}$.

Misspecified case

Assume we observe $X_1, \dots, X_n$ i.i.d from P^0 and use a model $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$, but it is possible that $P^0 \notin \mathcal{M}$.

Theorem - AR 17

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

$$\mathbb{E}_{\theta\sim\rho}\left\{\mathbb{E}_{X\sim P^0}\left[\log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right)\right]\right\} \leq r_n$$

and $\mathcal{K}(\rho_n, \pi) \leq n r_n$. Then

$$\mathbb{E}_{X_1,\dots,X_n}\left\{\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[D_\alpha(P_\theta, P^0)\right]\right\} \leq \frac{\alpha}{1-\alpha}\mathcal{K}(P_{\theta^*}, P^0) + \frac{1+\alpha}{1-\alpha} r_n.$$

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 :

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

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

If the true model is actually k_0,

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

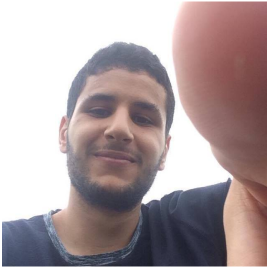
Model selection

Theorem

If the true model is actually k_0,

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

This result is actually due to my PhD student :



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

Models with hidden variables (1/2)

The results presented so far do not include approximations in models with hidden variables.

Models with hidden variables (1/2)

The results presented so far do not include approximations in models with hidden variables.

VB for mixtures

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

Models with hidden variables (1/2)

The results presented so far do not include approximations in models with hidden variables.

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 $(p_1, \dots, p_K)$.

Models with hidden variables (1/2)

The results presented so far do not include approximations in models with hidden variables.

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

Models with hidden variables (1/2)

The results presented so far do not include approximations in models with hidden variables.

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 $(p_1, \ldots, p_K)$.

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

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

Models with hidden variables (2/2)

For a general approach for models with hidden variables, including

1. mixture models,
2. hidden Markov chains,
3. ...

see :

Y. Yang, D. Pati & A. Bhattacharya (2017). α-Variational Inference with Statistical Guarantees. *The Annals of Statistics (to appear)*, preprint arXiv :1712.08983.

The case $\alpha = 1$

Our paper does not cover the case $\alpha = 1$:

Case $\alpha = 1$

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

The case $\alpha = 1$

Our paper does not cover the case $\alpha = 1$:

Case $\alpha = 1$

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

Covered in the following paper – note that this case requires **much stronger** assumptions :

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

More general machine learning problem

Reminder of the context of machine learning :

1. $X_1, \dots, X_n$ i.i.d from P^0,
2. $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.
3. $R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i)$.

More general machine learning problem

Reminder of the context of machine learning :

1. $X_1, \dots, X_n$ i.i.d from P^0,
2. $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.
3. $R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i)$.

Gibbs posterior, EWA, ...

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)] \pi(\theta).$$

More general machine learning problem

Reminder of the context of machine learning :

1. $X_1, \dots, X_n$ i.i.d from P^0,
2. $R(\theta) = \mathbb{E}_{X \sim P^0}[\ell(\theta, X)]$.
3. $R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i)$.

Gibbs posterior, EWA, ...

$$\pi_{n,\alpha}(\theta) \propto \exp[-\alpha n R_n(\theta)] \pi(\theta).$$

We define a variational approximation of the Gibbs posterior :

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

Variational approximation of Gibbs posteriors

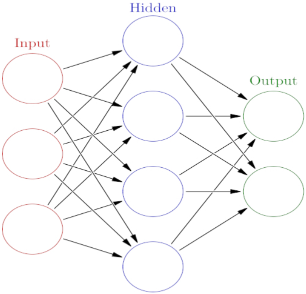
Bounds on the generalization error of the variational approximation of the Gibbs posterior

$$\mathbb{E}_{\theta\sim\tilde{\pi}_{n,\alpha}}\left[R(\theta)\right]$$

provided in the paper :

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

Neural networks



Source : Wikipedia.

Prior π : independent

$$\theta_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_\ell^2).$$

Variational approximation : independent

$$\theta_{i,j}^{(\ell)} \sim \mathcal{N}(m_{i,j}^{(\ell)}, (\sigma_{i,j}^{(\ell)})^2).$$

Neural networks for non parametric regression

Badr-Eddine Chérief-Abdellatif proved that, in regression with quadratic loss, suitable neural networks estimate β-Hölder functions at rate :

$$\mathcal{O}\left(\frac{\log(n)^2}{n^{\frac{2\beta}{2\beta+1}}}\right).$$

Neural networks for non parametric regression

Badr-Eddine Chérief-Abdellatif proved that, in regression with quadratic loss, suitable neural networks estimate β-Hölder functions at rate :

$$\mathcal{O}\left(\frac{\log(n)^2}{n^{\frac{2\beta}{2\beta+1}}}\right).$$

B.-E. Chérief-Abdellatif (2018). *Convergence Rates of Variational Inference in Sparse Deep Learning*. Preprint arXiv :1908 :04847.

Thank you !