Concentration of tempered posteriors and their variational approximations

James Ridgway
Joint work with Pierre Alquier

workshop Statistics/Learning at Paris-Saclay
January 2018

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 - $0 < \alpha < 1$

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

Classic way to deal with posteriors: Monte Carlo

- Monte Carlo algorithms are widely used to deal with posteriors or tempered posteriors (e.g. MCMC, SMC)

Classic way to deal with posteriors: Monte Carlo

- Monte Carlo algorithms are widely used to deal with posteriors or tempered posteriors (e.g. MCMC, SMC)
- Issues:
  - Computational complexity
  - Lack of non asymptotic theory, under investigation for behaviour in high dimension etc.
    Recent research filling the gap in this direction for log-concave problems:

    Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017

Variational Bayes

Variational Bayes is a deterministic approximation of some probability measure.

Let $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$. We define the VB approximation $\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)$ by

$$\tilde{\pi}_{n,\alpha}(\cdot|X_1^n) = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}(\cdot|X_1^n)).$$

where the Kullback-Leibler divergence is

$$\mathcal{K}(P,R) = \begin{cases} \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)\mathrm{d}P \text{ if } P \ll R \\ +\infty \text{ otherwise.} \end{cases}$$

Variational Bayes

Variational Bayes is a deterministic approximation of some probability measure.

Let $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$. We define the VB approximation $\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)$ by

$$\tilde{\pi}_{n,\alpha}(\cdot|X_1^n) = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}(\cdot|X_1^n)).$$

where the Kullback-Leibler divergence is

$$\mathcal{K}(P, R) = \begin{cases} \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)\mathrm{d}P \text{ if } P \ll R \\ +\infty \text{ otherwise.} \end{cases}$$

What family of distribtion $\mathcal{F}$?

Two common choices:

- Parametric family:

$$\mathcal{F} = \{q_\vartheta(d\theta), \vartheta \in \Theta'\}$$

What family of distribtion $\mathcal{F}$?

Two common choices:

- Parametric family:

$$\mathcal{F} = \{q_\vartheta(d\theta), \vartheta \in \Theta'\}$$

- Mean field:

$$\mathcal{F}^{\mathrm{mf}} := \Big\{\rho(\mathrm{d}\theta) = \bigotimes_{i=1}^{p} \rho_i(\mathrm{d}\theta_i) \in \mathcal{M}_1^+(\Theta), \quad \forall i = 1, \cdots, p \quad \rho_i \in \mathcal{M}_1^+(\Theta_i), \quad \Theta = \Theta_1 \times \cdots \times \Theta_p\Big\},$$

Previous results

In a previous paper

P. Alquier, J. R., and N. Chopin. On the properties of variational approximations of Gibbs posterior.
Journal of Machine Learning Research, 17(239):1–41, 2016

- We studied variational approximations of Gibbs posteriors with bounded risk. Fractional posteriors do not fall in this category.
- pseudo-posterior of interest are defined for a risk $r_n(\theta)$

$$\pi_\gamma(\mathrm{d}\theta) \propto \exp\left(-\gamma r_n(\theta)\right)\pi(\theta)$$

Previous results

In a previous paper

P. Alquier, J. R., and N. Chopin. On the properties of variational approximations of Gibbs posterior.
Journal of Machine Learning Research, 17(239):1–41, 2016

- We studied variational approximations of Gibbs posteriors with bounded risk. Fractional posteriors do not fall in this category.
- pseudo-posterior of interest are defined for a risk $r_n(\theta)$

$$\pi_\gamma(\mathrm{d}\theta) \propto \exp\left(-\gamma r_n(\theta)\right)\pi(\theta)$$

Definition

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

$$D_\alpha(P,R) = \begin{cases} \frac{1}{\alpha-1} \log \int \left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)^{\alpha-1} \mathrm{d}P \text{ if } P \ll R \\ +\infty \text{ otherwise.} \end{cases}$$

Definition

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

$$D_\alpha(P,R) = \begin{cases} \frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)^{\alpha-1}\mathrm{d}P \text{ if } P \ll R \\ +\infty \text{ otherwise.} \end{cases}$$

In particular, for $1/2 \leq \alpha$, link with Hellinger and Kullback:

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

Concentration of tempered posterior

$$\mathcal{B}(r) = \left\{ \theta \in \Theta : \mathcal{K}(P_{\theta_0}, P_\theta) \leq r \text{ and } \mathrm{Var}\left[\log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right] \leq r. \right\}$$

Theorem A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125*, 2016

For any sequence (r_n) such that

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

we have

$$\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \pi_{n,\alpha}(\mathrm{d}\theta) \leq \frac{2(1+\alpha)}{1-\alpha} r_n \right] \geq 1 - \frac{2}{n r_n}.$$

General result for VB approximation

Theorem (P. Alquier and J R. Concentration of tempered posterior and their variational approximations. *arXiv:1706.09293*, pages 1–24, 2017)

Fix $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$. Assume that $r_n > 0$ is such that there is distribution $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \le r_n, \quad \int \mathbb{E}\left[\log^2\left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right)\right]\rho_n(\mathrm{d}\theta) \le r_n \tag{1}$$

and

$$\mathcal{K}(\rho_n, \pi) \le n r_n. \tag{2}$$

Then, for any $\alpha \in (0,1)$, for any $(\varepsilon, \eta) \in (0,1)^2$,

$$\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \le \frac{(\alpha+1)r_n + \alpha\sqrt{\frac{r_n}{n\eta}} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}}{1-\alpha}\right] \ge 1-\varepsilon-\eta.$$

Remark and connection to Bayesian statistics

Put $\mathcal{F} = \mathcal{M}_1^+$,

- Define $B(r)$, for $r > 0$, as

$$B(r) = \left\{ \theta \in \Theta : \mathcal{K}(P_{\theta_0}, P_\theta) \leq r, \mathrm{Var}\left[\log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right] \leq r \right\}.$$

- Taking ρ_n as π restricted to $B(r_n)$, $\rho_n = \pi_{|B(r_n)}$: (1) is satisfied and (2) can be written

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

A simpler result in expectation

Theorem

If we only require that there is $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n$$

and

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

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

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

Misspecified case

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

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

Theorem

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

$$\int \mathcal{K}(P_{\theta^*}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n \text{ and } \mathcal{K}(\rho_n, \pi) \leq n r_n,$$

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

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

1 Introduction

2 Variational aproach

3 Main results

4 Examples
- Gaussian vb
- Matrix completion

Gaussian VB

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

Gaussian VB

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

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

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 $M(\cdot)$ and $p \in \mathbb{N}^\star \cup \{\frac{1}{2}\}$

$$|\log p_\theta(X_1) - \log p_{\theta'}(X_1)| \leq M(X_1) \|\theta - \theta'\|_2^{2p}$$

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 $\alpha \in (0,1)$, for any η, ϵ

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

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

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

1. Introduction
2. Variational aproach
3. Main results
4. Examples
   - Gaussian vb
   - Matrix completion

Matrix completion: notations

- The parameter θ is a matrix $M \in \mathbb{R}^{m\times p}$, with $m, p \geq 1$.
- Under P_M,

$$Y_k = M_{i_k, j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1,\dots,m\}\times\{1,\dots,p\})$. The noise ε_k is i.i.d $\mathcal{N}(0,\sigma^2)$, σ^2 known.

Matrix completion: notations

- The parameter θ is a matrix $M \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$.
- Under P_M,

$$Y_k = M_{i_k, j_k} + \varepsilon_k$$

  where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1, \dots, m\} \times \{1, \dots, p\})$. The noise ε_k is i.i.d $\mathcal{N}(0, \sigma^2)$, σ^2 known.
- Usual assumption: M is low-rank.

Prior specification - main idea

Define:

$$\underbrace{M}_{p\times m} = \underbrace{U}_{p\times k}\underbrace{V^T}_{k\times m}.$$

Prior specification - main idea

Define:

$$\underbrace{M}_{p\times m} = \underbrace{U}_{p\times k}\underbrace{V^T}_{k\times m}.$$

$$M = \sum_{\ell=1}^{k} U_{\cdot,\ell}(V_{\cdot,\ell})^T$$

with k large - e.g. $k = \min(p, m)$.

Prior specification - main idea

Define:

$$\underbrace{M}_{p\times m} = \underbrace{U}_{p\times k}\underbrace{V^T}_{k\times m}.$$

$$M = \sum_{\ell=1}^{k} U_{\cdot,\ell}(V_{\cdot,\ell})^T$$

with k large - e.g. $k = \min(p, m)$.
Definition of π:

- $U_{\cdot,\ell}, V_{\cdot,\ell} \sim \mathcal{N}(0, \gamma_\ell I)$,
- γ_ℓ is itself random, such that most of the $\gamma_\ell \simeq 0$

$$\frac{1}{\gamma_\ell} \sim \mathrm{Gamma}(a, b).$$

Variational approximation

Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction.
In *Proceedings of KDD Cup and Workshop*, 2007 Mean-field

approximation, $\mathcal{F}$ given by:

$$\rho(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma) = \bigotimes_{i=1}^{m} \rho_{U_i}(\mathrm{d}U_{i,\cdot}) \bigotimes_{j=1}^{p} \rho_{V_j}(\mathrm{d}V_{j,\cdot}) \bigotimes_{k=1}^{K} \rho_{\gamma_k}(\gamma_k).$$

Variational approximation

Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction.
In *Proceedings of KDD Cup and Workshop*, 2007 Mean-field approximation, $\mathcal{F}$ given by:

$$\rho(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma) = \bigotimes_{i=1}^{m} \rho_{U_i}(\mathrm{d}U_{i,\cdot}) \bigotimes_{j=1}^{p} \rho_{V_j}(\mathrm{d}V_{j,\cdot}) \bigotimes_{k=1}^{K} \rho_{\gamma_k}(\gamma_k).$$

It can be shown that

1. ρ_{U_i} is $\mathcal{N}(\mathbf{m}_{i,\cdot}^T, \mathcal{V}_i)$,
2. ρ_{V_j} is $\mathcal{N}(\mathbf{n}_{j,\cdot}^T, \mathcal{W}_j)$,
3. ρ_{γ_k} is $\Gamma(a + (m_1 + m_2)/2, \beta_k)$,

for some $m \times K$ matrix $\mathbf{m}$ whose rows are denoted by $\mathbf{m}_{i,\cdot}$, some $p \times K$ matrix $\mathbf{n}$ and some vector $\beta = (\beta_1, \dots, \beta_K)$.

The VB algorithm

The parameters are updated iteratively through the formulae

1. **moments of** U:

$$\mathbf{m}_{i,\cdot}^{T} := \frac{2\alpha}{n}\mathcal{V}_i \sum_{k:i_k=i} Y_{i_k,j_k}\mathbf{n}_{j_k,\cdot}^{T}$$

$$\mathcal{V}_i^{-1} := \frac{2\alpha}{n}\sum_{k:i_k=i}\left[\mathcal{W}_{j_k} + \mathbf{n}_{j_k,\cdot}\mathbf{n}_{j_k,\cdot}^{T}\right] + \left(a + \frac{m_1+m_2}{2}\right)\mathbf{diag}(\beta)^{-1}$$

2. **moments of** V:

$$\mathbf{n}_{j,\cdot}^{T} := \frac{2\alpha}{n}\mathcal{W}_j \sum_{k:j_k=j} Y_{i_k,j_k}\mathbf{m}_{i_k,\cdot}^{T}$$

$$\mathcal{W}_j^{-1} := \frac{2\alpha}{n}\sum_{k:j_k=j}\left[\mathcal{V}_{i_k} + \mathbf{m}_{i_k,\cdot}\mathbf{m}_{i_k,\cdot}^{T}\right] + \left(a + \frac{m_1+m_2}{2}\right)\mathbf{diag}(\beta)^{-1}$$

3. **moments of** γ:

$$\beta_k := \frac{1}{2}\left[\sum_{i=1}^{m_1}\left(\mathbf{m}_{i,k}^2 + (\mathcal{V}_i)_{k,k}\right) + \sum_{j=1}^{m_2}\left(\mathbf{n}_{j,k}^2 + (\mathcal{V}_j)_{k,k}\right)\right].$$

Application of our theorem

Theorem

Assume $M = \bar{U}\bar{V}^T$ *where*

$$\bar{U} = (\bar{U}_{1,\cdot}|\dots|\bar{U}_{r,\cdot}|0|\dots|0) \text{ and } \bar{V} = (\bar{V}_{1,\cdot}|\dots|\bar{V}_{r,\cdot}|0|\dots|0)$$

and $\sup_{i,k}|U_{i,k}|, \sup_{j,k}|V_{j,k}| \leq B$. *Take* $a > 0$ *as any constant and* $b = \frac{B^2}{512(nmp)^4[(m \vee p)K]^2}$. *Then*

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

$$\text{where } r_n = \frac{\mathcal{C}(a, \sigma^2, B) r \max(m,p)\log(nmp)}{n}.$$

Thank you!

P. Alquier and J R. Concentration of tempered posterior and their variational approximations. *arXiv:1706.09293*, pages 1–24, 2017.

P. Alquier, J. R., and N. Chopin. On the properties of variational approximations of Gibbs posterior. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125*, 2016.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, 2007.