Consistency of Variational Inference

Badr-Eddine Chérief-Abdellatif
Under the supervision of Pierre Alquier

Second-year PhD students team days
Ecole Doctorale Mathématiques Hadamard
20 May 2019

Outline of the talk

1. Tempered Variational Bayes
   - Tempered posteriors
   - Variational Bayes
2. ELBO maximization
   - Mixture models
   - Model selection
3. Consistency of VB
   - Theoretical results
   - Efficient algorithms

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$ associated with a likelihood L_n. We define a prior π on Θ.

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$ associated with a likelihood L_n. We define a prior π on Θ.

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^0 in a model $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$ associated with a likelihood L_n. We define a prior π on Θ.

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

- Easier to sample from

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

Various reasons to use a tempered posterior

- Easier to sample from

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- Robust to model misspecification

P. Grünwald and T. Van Ommen (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*.

Various reasons to use a tempered posterior

- Easier to sample from

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- Robust to model misspecification

P. Grünwald and T. Van Ommen (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*.

- Theoretical analysis easier

A. Bhattacharya, D. Pati & Y. Yang (2016). Bayesian fractional posteriors. *Preprint arxiv :1611.01125*.

1. Tempered Variational Bayes
   - Tempered posteriors
   - Variational Bayes
2. ELBO maximization
   - Mixture models
   - Model selection
3. Consistency of VB
   - Theoretical results
   - Efficient algorithms

Variational Bayes

$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\ = \arg\max_{\rho\in\mathcal{F}} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho,\pi) \right\}.$$

Variational Bayes

$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha})$$
$$= \arg\max_{\rho\in\mathcal{F}} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \pi) \right\}.$$

Examples :

Variational Bayes

$$
\begin{aligned}
\tilde{\pi}_{n,\alpha} &= \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\
&= \arg\max_{\rho \in \mathcal{F}} \left\{ \alpha \int \ell_n(\theta) \rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \pi) \right\}.
\end{aligned}
$$

Examples :

- parametric approximation

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

Variational Bayes

$$\tilde{\pi}_{n,\alpha} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha})$$
$$= \arg\max_{\rho \in \mathcal{F}} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \pi) \right\}.$$

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

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

Mixture models

Mixture models

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

Mixture models

Mixture models

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

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 :

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

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

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

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

Coordinate Descent algorithm

Algorithm 1 Coordinate Descent Variational Bayes for mixtures

Input: a dataset $(X_1, ..., X_n)$, priors π_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 $(w_j^i)_{1 \le j \le K}$

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

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

Numerical example on Gaussian mixtures

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

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)

1. Tempered Variational Bayes
   - Tempered posteriors
   - Variational Bayes
2. ELBO maximization
   - Mixture models
   - Model selection
3. Consistency of VB
   - Theoretical results
   - Efficient algorithms

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 a countable number of models, define $\tilde{\pi}^K_{n,\alpha}$ a variational approximation of the tempered posterior in model K :

Model selection

Assume that we have a countable number of models, define $\tilde{\pi}_{n,\alpha}^K$ a variational approximation of the tempered posterior in model K :

ELBO maximization program

$$\tilde{\pi}_{n,\alpha}^K = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K) \rho_K(d\theta_K) - \mathcal{K}(\rho_K, \pi_K) \right\}$$

Model selection

Assume that we have a countable number of models, define $\tilde{\pi}^K_{n,\alpha}$ a variational approximation of the tempered posterior in model K :

ELBO maximization program

$$\tilde{\pi}^K_{n,\alpha} = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K)\rho_K(d\theta_K) - \mathcal{K}(\rho_K, \pi_K) \right\}$$

ELBO

$$\mathrm{ELBO}(K) = \alpha \int \ell_n(\theta_K)\tilde{\pi}^K_{n,\alpha}(d\theta_K) - \mathcal{K}(\tilde{\pi}^K_{n,\alpha}, \pi_K)$$

ELBO criterion

Model selection criterion

$$\hat{K} = \arg\max_{K \geq 1} \left\{ \mathrm{ELBO}(K) - \log\left(\frac{1}{b_K}\right) \right\}$$

ELBO criterion

Model selection criterion

$$\hat{K} = \arg\max_{K \geq 1} \left\{ \mathrm{ELBO}(K) - \log\left(\frac{1}{b_K}\right) \right\}$$

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

Technical condition for posterior concentration

If the model is well-specified ($\exists \theta^0 \in \Theta$, $P_{\theta^0} = P^0$) :

Technical condition for posterior concentration

If the model is well-specified ($\exists \theta^0 \in \Theta$, $P_{\theta^0} = P^0$) :

Prior mass condition for concentration of tempered posteriors

The rate (r_n) is such that

$$\pi[\mathcal{B}(r_n)] \geq e^{-nr_n}$$

where $\mathcal{B}(r) = \{\theta \in \Theta : \mathcal{K}(P_{\theta^0}, P_\theta) \leq r\}$.

Technical condition for posterior concentration

If the model is well-specified ($\exists \theta^0 \in \Theta$, $P_{\theta^0} = P^0$) :

Prior mass condition for concentration of tempered posteriors

The rate (r_n) is such that

$$\pi[\mathcal{B}(r_n)] \geq e^{-nr_n}$$

where $\mathcal{B}(r) = \{\theta \in \Theta : \mathcal{K}(P_{\theta^0}, P_\theta) \leq r\}$.

Prior mass condition for concentration of Variational Bayes

The rate (r_n) is such that there exists $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta^0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n, \ \text{ and } \ \mathcal{K}(\rho_n, \pi) \leq nr_n.$$

Consistency of the variational approximation

P. Alquier & J. Ridgway (2017). Concentration of Tempered Posteriors and of their Variational Approximations. *Preprint arxiv :1706.09293.*

Theorem (Alquier, Ridgway)

Under the prior mass condition, for any $\alpha \in (0,1)$,

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

Consistency for mixture models

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

Theorem (C.-A., Alquier)

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

$$\mathbb{E}\left[\int D_\alpha(P_{p,\theta_1,\dots,\theta_K}, P_{p^0,\theta_1^0,\dots,\theta_K^0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta)\right] \leq \frac{1+\alpha}{1-\alpha} 2Kr_n.$$

Consistency of the true approximation

P. Alquier & J. Ridgway (2017). Concentration of Tempered Posteriors and of their Variational Approximations. *Preprint arxiv :1706.09293.*

Theorem (Alquier, Ridgway)

If there is a true model ($\exists K_0$, $\exists \theta^0 \in \Theta_{K_0}$, $P_{\theta^0} = P^0$), then under the prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{K_0}(d\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Consistency of the selected approximation

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

Theorem (C.-A.)

If there is a true model ($\exists K_0$, $\exists \theta^0 \in \Theta_{K_0}$, $P_{\theta^0} = P^0$), then under the prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{\log(\frac{1}{b_{K_0}})}{n(1-\alpha)}.$$

Robustness to misspecification (Gaussian mixtures)

The true distribution P^0 is such that $\mathbb{E}|X| < +\infty$.
Let $L \geq 1$, $b_K = 2^{-K}$, $\pi_K = \mathcal{D}_K(\alpha_1, \dots, \alpha_K) \bigotimes \mathcal{N}(0, \mathcal{V}^2)^{\otimes n}$ and

$$r_{n,K} = \left[\frac{8K \log(nK)}{n} \bigvee \left(\frac{8K \log(n\mathcal{V})}{n} + \frac{8KL^2}{n\mathcal{V}^2} \right) \right] + \frac{K \log(2)}{n(1-\alpha)}.$$

Theorem

For any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta) \right] \leq \inf_{K \geq 0} \left\{ \frac{\alpha}{1-\alpha} \inf_{\theta^* \in \mathcal{S}_K \times [-L,L]^K} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} r_{n,K} \right\}.$$

Applications

If there is a true model $(\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0)$:

Applications

If there is a true model $(\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0)$:

Gaussian mixtures : $\theta = (p, (m_1, \sigma_1^2), ..., (m_K, \sigma_K^2))$

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta) \right] = \mathcal{O}\left(\frac{K_0 \log(nK_0)}{n} \right)$$

Applications

If there is a true model ($\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0$) :

Gaussian mixtures : $\theta = (p, (m_1, \sigma_1^2), ..., (m_K, \sigma_K^2))$

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta)\right] = \mathcal{O}\left(\frac{K_0 \log(nK_0)}{n}\right)$$

Probabilistic PCA : $\theta \in \mathbb{R}^{d\times K}$

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta)\right] = \mathcal{O}\left(\frac{dK_0 \log(dn)}{n}\right)$$

Efficient algorithms

Question

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

Efficient algorithms

Question

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

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

Efficient algorithms

Question

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

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 the tempered posterior...

Some online 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**

Some online 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

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

A regret bound for SVA

Theorem (C.-A., Alquier & 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 (C.-A., Alquier & 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 (C.-A., Alquier & 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 :

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

Thank you !