Séminaire de Calcul Scientifique du CERMICS

École des Ponts
ParisTech

Variational Approximations in Machine Learning : Theory and Applications

Pierre Alquier (ENSAE)

25 juin 2018

Variational Approximations in Machine Learning: Theory and Applications

Pierre Alquier

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

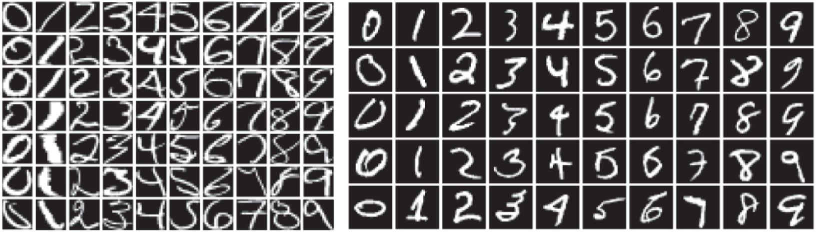
CERMICS - Monday, June 25, 2018

Learning vs. estimation

In many applications one would like to learn from a sample without being able to write the likelihood.

Learning vs. estimation

In many applications one would like to learn from a sample without being able to write the likelihood.



(a) USPS (b) MNIST

Typical machine learning problem

Main ingredients :

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:
  $\rightarrow$ for example $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$, $R(\theta) = \|\theta - \theta_0\|$ where θ_0 is a target parameter, ... we want $R(\theta)$ to be small. But note that it is unknown.

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:
  $\rightarrow$ for example $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$, $R(\theta) = \|\theta - \theta_0\|$ where θ_0 is a target parameter, ... we want $R(\theta)$ to be small. But note that it is unknown.
- an empirical proxy $r(\theta)$ for this criterion of success :

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:
  $\rightarrow$ for example $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$, $R(\theta) = \|\theta - \theta_0\|$ where θ_0 is a target parameter, ... we want $R(\theta)$ to be small. But note that it is unknown.
- an empirical proxy $r(\theta)$ for this criterion of success :
  $\rightarrow$ for example $r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(f_\theta(X_i) \neq Y_i)$.

PAC-Bayesian bounds

One more ingredient :

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

The PAC-Bayesian approach usually provides a "posterior distribution" $\hat{\rho}_\lambda$ and a theoretical guarantee :

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_{\rho}\left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right] + o(1).$$

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

The PAC-Bayesian approach usually provides a "posterior distribution" $\hat{\rho}_\lambda$ and a theoretical guarantee :

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right] + o(1).$$

Usually $o(1)$ is explicit, λ is some tuning-parameter to be calibrated (constrained to some range by theory), and $\hat{\rho}_\lambda$ is the "Gibbs posterior"

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Outline of the talk

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

1st example : general bound for batch learning

Context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.

1st example : general bound for batch learning

Context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.

1st example : general bound for batch learning

Context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.
- $R(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, f_\theta(X))]$ for any *bounded* loss function $|\ell(\cdot,\cdot)| \leq B$.

1st example : general bound for batch learning

Context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.
- $R(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, f_\theta(X))]$ for any *bounded* loss function $|\ell(\cdot,\cdot)| \leq B$.
- $r_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$.

1st example : general bound for batch learning

Context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.
- $R(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, f_\theta(X))]$ for any *bounded* loss function $|\ell(\cdot,\cdot)| \leq B$.
- $r_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$.
- any prior π.

Catoni's bound for batch learning

Theorem [Catoni 2007]

$$\forall \lambda > 0, \quad \mathbb{P}\Bigg\{ \int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$

$$\leq \inf_\rho \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right] \Bigg\}$$

$$\geq 1 - \varepsilon.$$

Catoni's bound for batch learning

Theorem [Catoni 2007]

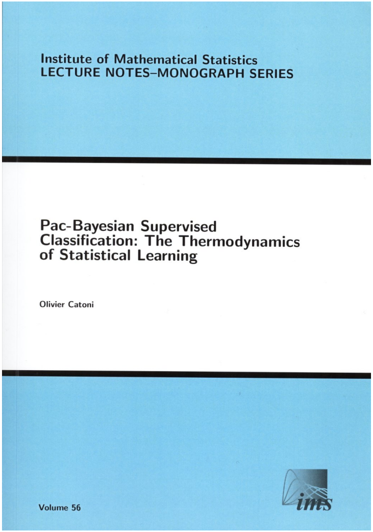
$$\forall\lambda > 0, \quad \mathbb{P}\Bigg\{ \int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$
$$\leq \inf_\rho \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right] \Bigg\}$$
$$\geq 1 - \varepsilon.$$

improving on seminal work :

Shawe-Taylor, J. & Williamson, R. C. (1997). A PAC Analysis of a Bayesian Estimator. *COLT'97*.

McAllester, D. A. (1998). Some PAC-Bayesian Theorems. *COLT'98*.

Reference


Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES
Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning
Olivier Catoni
Volume 56
ims

Application : finite set of predictors $\theta_1, \dots, \theta_M$

With π the uniform distribution on $\{\theta_1, \dots, \theta_M\}$ we get

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_{\rho=\delta_{\theta_i}} \left[\int R \mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]$$

Application : finite set of predictors $\theta_1, \dots, \theta_M$

With π the uniform distribution on $\{\theta_1, \dots, \theta_M\}$ we get

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$

$$\leq \inf_{\rho=\delta_{\theta_i}} \left[\int R \mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right]$$

$$\leq \inf_{1 \leq i \leq M} \left[R(\theta_i) + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\log(M) + \log\left(\frac{2}{\varepsilon}\right) \right] \right]$$

Application : finite set of predictors $\theta_1, \dots, \theta_M$

With π the uniform distribution on $\{\theta_1, \dots, \theta_M\}$ we get

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$

$$\leq \inf_{\rho=\delta_{\theta_i}} \left[\int R\mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]$$

$$\leq \inf_{1\leq i\leq M} \left[R(\theta_i) + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\log(M) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]$$

$$= \inf_{1\leq i\leq M} R(\theta_i) + 2B\sqrt{\frac{2\log(M)}{n}} + \log\left(\frac{2}{\varepsilon}\right)\sqrt{\frac{1}{2n\log(M)}}$$

for $\lambda = \frac{\sqrt{2n\log(M)}}{B}$.

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

2nd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* assumption.

2nd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* assumption.
- any $(f_\theta, \theta \in \Theta)$.

2nd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* assumption.
- any $(f_\theta, \theta \in \Theta)$.
- given (X_1, Y_1), (X_2, Y_2), ..., (X_{t-1}, Y_{t-1}) and X_t we are asked to predict Y_t : by $\hat{Y}_t$. At some time T the game stops and we evaluate the *regret* :

$$\mathcal{R} = \sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t) - \inf_\theta \sum_{t=1}^{T} \ell(Y_t, f_\theta(X_t)),$$

ℓ is bounded by B and cvx. w.r.t its second argument.

2nd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* assumption.
- any $(f_\theta, \theta \in \Theta)$.
- given (X_1, Y_1), (X_2, Y_2), ..., (X_{t-1}, Y_{t-1}) and X_t we are asked to predict Y_t : by $\hat{Y}_t$. At some time T the game stops and we evaluate the *regret* :

$$\mathcal{R} = \sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t) - \inf_{\theta} \sum_{t=1}^{T} \ell(Y_t, f_\theta(X_t)),$$

  ℓ is bounded by B and cvx. w.r.t its second argument.
- at time t we can use as a proxy of the quality of θ : $r_{t-1}(\theta) = \sum_{h=1}^{t-1} \ell(Y_h, f_\theta(X_h))$.

2nd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* assumption.
- any $(f_\theta, \theta \in \Theta)$.
- given (X_1, Y_1), (X_2, Y_2), ..., (X_{t-1}, Y_{t-1}) and X_t we are asked to predict Y_t : by $\hat{Y}_t$. At some time T the game stops and we evaluate the *regret* :

$$\mathcal{R} = \sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t) - \inf_{\theta} \sum_{t=1}^{T} \ell(Y_t, f_\theta(X_t)),$$

  ℓ is bounded by B and cvx. w.r.t its second argument.
- at time t we can use as a proxy of the quality of θ : $r_{t-1}(\theta) = \sum_{h=1}^{t-1} \ell(Y_h, f_\theta(X_h))$.
- any prior π.

PAC-Bayesian bound for online learning

Fix $\lambda > 0$ and define, at each time t,

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta) \text{ and } \hat{Y}_t = \int f_\theta(X_t)\hat{\rho}_{\lambda,t}(\mathrm{d}\theta).$$

PAC-Bayesian bound for online learning

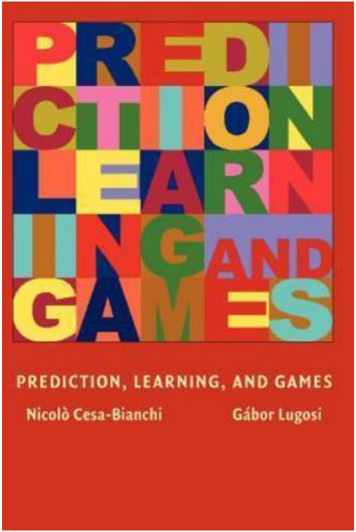
Fix $\lambda > 0$ and define, at each time t,

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta) \text{ and } \hat{Y}_t = \int f_\theta(X_t)\hat{\rho}_{\lambda,t}(\mathrm{d}\theta).$$

Theorem [Consequence of Audibert, 2006]

$$\sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t) \leq \inf_{\rho} \left\{ \int \sum_{t=1}^{T} \ell(Y_t, f_\theta(X_t))\rho(\mathrm{d}\theta) + \frac{\lambda T B^2}{2} + \frac{\mathcal{K}(\rho,\pi)}{\lambda} \right\}.$$

Reference


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

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

3rd example : Bayesian statistics

- $X_1, \ldots, X_n$ i.i.d from P_{θ_0}.

3rd example : Bayesian statistics

- $X_1, \dots, X_n$ i.i.d from P_{θ_0}.
- a statistical model $(P_\theta, \theta \in \Theta)$ dominated : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu} = p_\theta$.

3rd example : Bayesian statistics

- $X_1, \dots, X_n$ i.i.d from P_{θ_0}.
- a statistical model $(P_\theta, \theta \in \Theta)$ dominated : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu} = p_\theta$.
- criterion on θ : $\mathcal{K}(P_{\theta_0}, P_\theta)$ or $h(P_{\theta_0}, P_\theta)$.

3rd example : Bayesian statistics

- $X_1, \dots, X_n$ i.i.d from P_{θ_0}.
- a statistical model $(P_\theta, \theta \in \Theta)$ dominated : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu} = p_\theta$.
- criterion on θ : $\mathcal{K}(P_{\theta_0}, P_\theta)$ or $h(P_{\theta_0}, P_\theta)$.
- we measure the data-fit by the likelihood :

$$L(\theta|X_1^n) = \prod_{i=1}^{n} p_\theta(X_i).$$

3rd example : Bayesian statistics

- $X_1, \dots, X_n$ i.i.d from P_{θ_0}.
- a statistical model $(P_\theta, \theta \in \Theta)$ dominated : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu} = p_\theta$.
- criterion on θ : $\mathcal{K}(P_{\theta_0}, P_\theta)$ or $h(P_{\theta_0}, P_\theta)$.
- we measure the data-fit by the likelihood :

$$L(\theta | X_1^n) = \prod_{i=1}^{n} p_\theta(X_i).$$

- any prior π on Θ.

Posterior and variants

The posterior :

$$\begin{aligned}\pi(\theta|X_1^n) &\propto L(\theta|X_1^n)\pi(\theta) \\ &\propto \exp(-r_n(\theta))\pi(\theta)\end{aligned}$$

where $r_n(\theta) = -\sum_{i=1}^n \log p_\theta(X_i)$.

Posterior and variants

The posterior :

$$\begin{aligned}\pi(\theta|X_1^n) &\propto L(\theta|X_1^n)\pi(\theta)\\ &\propto \exp(-r_n(\theta))\pi(\theta)\end{aligned}$$

where $r_n(\theta) = -\sum_{i=1}^n \log p_\theta(X_i)$.

Tempered posterior (or fractional posterior), for $0 < \alpha \leq 1$:

$$\begin{aligned}\pi_\alpha(\theta|X_1^n) &\propto \exp(-\alpha r_n(\theta))\pi(\theta)\\ &\propto L(\theta|X_1^n)^\alpha \pi(\theta).\end{aligned}$$

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*.

- more robust to model misspecification (at least empirically)

P. Grünwald & 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*.

- more robust to model misspecification (at least empirically)

P. Grünwald & 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*.

PAC-Bayesian inequality for the tempered posterior

(Based on [Bhattacharya, D. Pati & Y. Yang, 2016]).

Theorem [Alquier & Ridgway, 2017]

For any $\alpha \in (1/2, 1)$,

$$\mathbb{E}\left[\int h^2(P_\theta, P_{\theta_0})\pi_\alpha(\mathrm{d}\theta|X_1^n)\right] \leq \inf_\rho \left\{\frac{\alpha}{1-\alpha}\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)}\right\}.$$

Concentration of the tempered posterior

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

Corollary

For any sequence (ε_n) such that

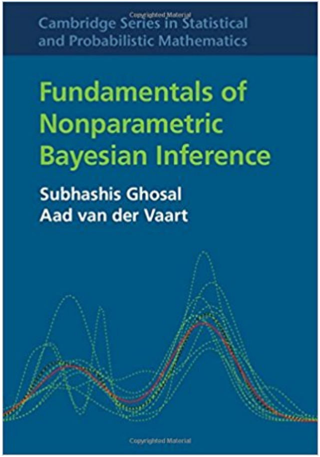
$$-\log \pi[B(r_n)] \leq n\varepsilon_n$$

we have

$$\mathbb{E}\left[\int h^2(P_\theta, P_{\theta_0})\pi_\alpha(\mathrm{d}\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

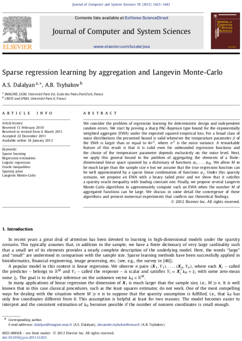
Reference

The (more classical) case $\alpha = 1$ is covered in depth in :



Computations ? A natural idea : MCMC methods

For the Gibbs posterior :



Journal of Computer and System Sciences 78 (2012) 1423–1443

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Sparse regression learning by aggregation and Langevin Monte-Carlo

A.S. Dalalyan[a,*], A.B. Tsybakov[b]

[a] IMAGINE, LIGM, Université Paris Est, Ecole des Ponts ParisTech, France
[b] CREST and LPMA, Université Paris 6, France

ARTICLE INFO

Article history:
Received 13 February 2010
Received in revised form 6 March 2011
Accepted 22 December 2011
Available online 18 January 2012

Keywords:
Sparse learning
Regression estimation
Logistic regression
Oracle inequalities
Sparsity prior
Langevin Monte-Carlo

ABSTRACT

We consider the problem of regression learning for deterministic design and independent random errors. We start by proving a sharp PAC-Bayesian type bound for the exponentially weighted aggregate (EWA) under the expected squared empirical loss. For a broad class of noise distributions the presented bound is valid whenever the temperature parameter β of the EWA is larger than or equal to $4\sigma^2$, where σ^2 is the noise variance. A remarkable feature of this result is that it is valid even for unbounded regression functions and the choice of the temperature parameter depends exclusively on the noise level. Next, we apply this general bound to the problem of aggregating the elements of a finite-dimensional linear space spanned by a dictionary of functions $\phi_1, \dots, \phi_M$. We allow M to be much larger than the sample size n but we assume that the true regression function can be well approximated by a sparse linear combination of functions ϕ_j. Under this sparsity scenario, we propose an EWA with a heavy tailed prior and we show that it satisfies a sparsity oracle inequality with leading constant one. Finally, we propose several Langevin Monte-Carlo algorithms to approximately compute such an EWA when the number M of aggregated functions can be large. We discuss in some detail the convergence of these algorithms and present numerical experiments that confirm our theoretical findings.

1. Introduction

In recent years a great deal of attention has been devoted to learning in high-dimensional models under the sparsity scenario. This typically assumes that, in addition to the sample, we have a finite dictionary of very large cardinality such that a small set of its elements provides a nearly complete description of the underlying model. Here, the words "large" and "small" are understood in comparison with the sample size. Sparse learning methods have been successfully applied in bioinformatics, financial engineering, image processing, etc. (see, e.g., the survey in [46]).

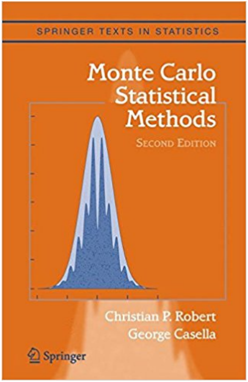
A popular model in this context is linear regression. We observe n pairs $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$, where each $\boldsymbol{X}_i$ – called the predictor – belongs to $\mathbb{R}^M$ and Y_i – called the response – is scalar and satisfies $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\lambda}_0 + \xi_i$ with some zero-mean noise ξ_i. The goal is to develop inference on the unknown vector $\boldsymbol{\lambda}_0 \in \mathbb{R}^M$.

In many applications of linear regression the dimension of $\boldsymbol{X}_i$ is much larger than the sample size, i.e., $M \gg n$. It is well known that in this case classical procedures, such as the least squares estimator, do not work. One of the most compelling ways for dealing with the situation where $M \gg n$ is to suppose that the sparsity assumption is fulfilled, i.e., that $\boldsymbol{\lambda}_0$ has only few coordinates different from 0. This assumption is helpful at least for two reasons: The model becomes easier to interpret and the consistent estimation of $\boldsymbol{\lambda}_0$ becomes possible if the number of nonzero coordinates is small enough.

* Corresponding author.
E-mail addresses: dalalyan@imagine.enpc.fr (A.S. Dalalyan), alexandre.tsybakov@ensae.fr (A.B. Tsybakov).

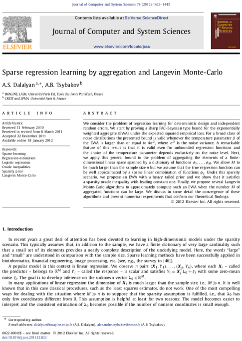
0022-0000/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2011.12.023

In Bayesian statistics :



Computations ? A natural idea : MCMC methods

For the Gibbs posterior :



Journal of Computer and System Sciences 78 (2012) 1423–1443
Contents lists available at SciVerse ScienceDirect
Journal of Computer and System Sciences
www.elsevier.com/locate/jcss

Sparse regression learning by aggregation and Langevin Monte-Carlo

A.S. Dalalyan[a,*], A.B. Tsybakov[b]
[a] IMAGINE, LIGM, Université Paris Est, Ecole des Ponts ParisTech, France
[b] CREST and LPMA, Université Paris 6, France

ARTICLE INFO
Article history:
Received 13 February 2010
Received in revised form 6 March 2011
Accepted 22 December 2011
Available online 18 January 2012
Keywords:
Sparse learning
Regression estimation
Logistic regression
Oracle inequalities
Sparsity prior
Langevin Monte-Carlo

ABSTRACT
We consider the problem of regression learning for deterministic design and independent random errors. We start by proving a sharp PAC-Bayesian type bound for the exponentially weighted aggregate (EWA) under the expected squared empirical loss. For a broad class of noise distributions the presented bound is valid whenever the temperature parameter β of the EWA is larger than or equal to $4\sigma^2$, where σ^2 is the noise variance. A remarkable feature of this result is that it is valid even for unbounded regression functions and the choice of the temperature parameter depends exclusively on the noise level. Next, we apply this general bound to the problem of aggregating the elements of a finite-dimensional linear space spanned by a dictionary of functions $\phi_1, \dots, \phi_M$. We allow M to be much larger than the sample size n but we assume that the true regression function can be well approximated by a sparse linear combination of functions ϕ_j. Under this sparsity scenario, we propose an EWA with a heavy tailed prior and we show that it satisfies a sparsity oracle inequality with leading constant one. Finally, we propose several Langevin Monte-Carlo algorithms to approximately compute such an EWA when the number M of aggregated functions can be large. We discuss in some detail the convergence of these algorithms and present numerical experiments that confirm our theoretical findings.

1. Introduction

In recent years a great deal of attention has been devoted to learning in high-dimensional models under the sparsity scenario. This typically assumes that, in addition to the sample, we have a finite dictionary of very large cardinality such that a small set of its elements provides a nearly complete description of the underlying model. Here, the words "large" and "small" are understood in comparison with the sample size. Sparse learning methods have been successfully applied in bioinformatics, financial engineering, image processing, etc. (see, e.g., the survey in [46]).

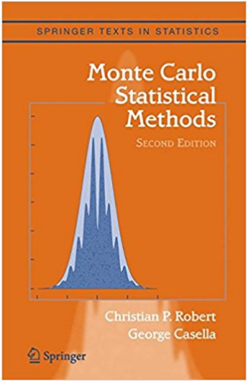
A popular model in this context is linear regression. We observe n pairs $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$, where each $\boldsymbol{X}_i$ – called the predictor – belongs to $\mathbb{R}^M$ and Y_i – called the response – is scalar and satisfies $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\lambda}_0 + \xi_i$ with some zero-mean noise ξ_i. The goal is to develop inference on the unknown vector $\boldsymbol{\lambda}_0 \in \mathbb{R}^M$.

In many applications of linear regression the dimension of $\boldsymbol{X}_i$ is much larger than the sample size, i.e., $M \gg n$. It is well known that in this case classical procedures, such as the least squares estimator, do not work. One of the most compelling ways for dealing with the situation where $M \gg n$ is to suppose that the sparsity assumption is fulfilled, i.e., that $\boldsymbol{\lambda}_0$ has only few coordinates different from 0. This assumption is helpful at least for two reasons: The model becomes easier to interpret and the consistent estimation of $\boldsymbol{\lambda}_0$ becomes possible if the number of nonzero coordinates is small enough.

* Corresponding author.
E-mail addresses: dalalyan@imagine.enpc.fr (A.S. Dalalyan), alexandre.tsybakov@ensae.fr (A.B. Tsybakov).

0022-0000/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2011.12.023

In Bayesian statistics :



Problems : often slow, no guarantees on the quality...

Variational Bayes methods

Idea : approximate the posterior distribution $\pi(\theta|X_1^n)$. We fix a convenient family of probability distributions $\mathcal{F}$ and approximate the posterior by $\tilde{\pi}(\theta)$:

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

Jordan, M. *et al* (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*.

Variational Bayes methods

Idea : approximate the posterior distribution $\pi(\theta|X_1^n)$. We fix a convenient family of probability distributions $\mathcal{F}$ and approximate the posterior by $\tilde{\pi}(\theta)$:

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

Jordan, M. *et al* (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*.

$\mathcal{F}$ is either parametric or non-parametric. In the parametric case, the problem boils down to an optimization problem :

$$\mathcal{F} = \{\rho_a, a \in \mathcal{A} \subset \mathbb{R}^d\} \dashrightarrow \min_{a\in\mathcal{A}} \mathcal{K}(\rho_a, \pi(\cdot|X_1^n)).$$

Variational Bayes methods

Idea : approximate the posterior distribution $\pi(\theta|X_1^n)$. We fix a convenient family of probability distributions $\mathcal{F}$ and approximate the posterior by $\tilde{\pi}(\theta)$:

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

Jordan, M. *et al* (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*.

$\mathcal{F}$ is either parametric or non-parametric. In the parametric case, the problem boils down to an optimization problem :

$$\mathcal{F} = \{\rho_a, a \in \mathcal{A} \subset \mathbb{R}^d\} \dashrightarrow \min_{a \in \mathcal{A}} \mathcal{K}(\rho_a, \pi(\cdot|X_1^n)).$$

Theoretical guarantees on the approximation ?

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

VB in the machine learning framework

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Then :

$$\begin{aligned}\mathcal{K}(\rho_a, \hat{\rho}_\lambda) &= \int \log\left[\frac{\mathrm{d}\rho_a}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\hat{\rho}_\lambda}\right]\mathrm{d}\rho_a \\ &= \lambda\int r(\theta)\rho_a(\mathrm{d}\theta) + \mathcal{K}(\rho_a, \pi) + \log\int \exp[-\lambda r]\mathrm{d}\pi.\end{aligned}$$

VB in the machine learning framework

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Then :

$$\mathcal{K}(\rho_a, \hat{\rho}_\lambda) = \int \log\left[\frac{\mathrm{d}\rho_a}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\hat{\rho}_\lambda}\right]\mathrm{d}\rho_a$$
$$= \lambda \int r(\theta)\rho_a(\mathrm{d}\theta) + \mathcal{K}(\rho_a, \pi) + \log\int \exp[-\lambda r]\mathrm{d}\pi.$$

We put

$$\tilde{a}_\lambda = \arg\min_{a\in\mathcal{A}}\left[\lambda\int r(\theta)\rho_a(\mathrm{d}\theta) + \mathcal{K}(\rho_a, \pi)\right] \text{ and } \tilde{\rho}_\lambda = \rho_{\hat{a}_\lambda}.$$

A PAC-Bound for VB Approximation

Theorem

Alquier, P., Ridgway, J. & Chopin, N. (2015). On the Properties of Variational Approximations of Gibbs Posteriors. *JMLR*.

$$\forall \lambda > 0, \quad \mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \right.$$

$$\left. \leq \inf_{a \in \mathcal{A}} \left[\int R(\theta)\rho_a(\mathrm{d}\theta) + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho_a, \pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]\right\}$$

$$\geq 1 - \varepsilon.$$

A PAC-Bound for VB Approximation

Theorem

Alquier, P., Ridgway, J. & Chopin, N. (2015). On the Properties of Variational Approximations of Gibbs Posteriors. *JMLR*.

$$\forall \lambda > 0, \quad \mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_{a\in\mathcal{A}} \left[\int R(\theta)\rho_a(\mathrm{d}\theta) + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho_a, \pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]\right\} \geq 1 - \varepsilon.$$

$\dashrightarrow$ if we can derive a tight oracle inequality from this bound, we know that the VB approximation is "at no cost".

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference

2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior

3. Discussion

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r_n(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle\theta, x\rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r_n(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.
- Gaussian prior $\pi = \mathcal{N}(0, \vartheta I)$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[Y_i \neq f_\theta(X_i)]$.
- Gaussian prior $\pi = \mathcal{N}(0, \vartheta I)$.
- Gaussian approx. of the posterior :
  $\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \text{ s. pos. def.} \right\}$.

Application to a linear classification problem

- $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r_n(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.
- Gaussian prior $\pi = \mathcal{N}(0, \vartheta I)$.
- Gaussian approx. of the posterior : $\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \text{ s. pos. def.} \right\}$.

Optimization criterion :

$$\frac{\lambda}{n} \sum_{i=1}^n \Phi \left(\frac{-Y_i \langle X_i, \mu \rangle}{\sqrt{\langle X_i, \Sigma X_i \rangle}} \right) + \frac{\|\mu\|^2}{2\vartheta} + \frac{1}{2} \left(\frac{1}{\vartheta} \mathrm{tr}(\Sigma) - \log |\Sigma| \right)$$

using deterministic annealing and gradient descent.

Application of the main theorem

Corollary

Assume that, for $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle\theta, X\rangle \langle\theta', X\rangle) \leq c\|\theta - \theta'\|$ and take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

$$\mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\theta R(\theta) + \sqrt{\frac{d}{n}}\left[\log(4n\mathrm{e}^2) + c\right] + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}} \right\} \geq 1 - \varepsilon.$$

Application of the main theorem

Corollary

Assume that, for $\|\theta\| = \|\theta'\| = 1$,
$\mathbb{P}(\langle \theta, X\rangle \langle \theta', X\rangle) \leq c\|\theta - \theta'\|$ and take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

$$\mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_{\lambda}(\mathrm{d}\theta) \leq \inf_{\theta} R(\theta) + \sqrt{\frac{d}{n}}\left[\log(4n\mathrm{e}^2) + c\right] + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}} \right\} \geq 1 - \varepsilon.$$

N.B : under margin assumption, possible to obtain d/n rates...

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_{\lambda} \leq \inf_{\rho=\mathcal{N}(\theta,s^2 I)} \left[\int R\mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right].$$

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho=\mathcal{N}(\theta,s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right].$$

As $\pi = \mathcal{N}(0,\vartheta I)$ we have

$$\mathcal{K}(\rho,\pi) = \frac{1}{2}\left[M\left(\frac{s^2}{\vartheta} - 1 + \log\left(\frac{\vartheta}{s^2}\right)\right) + \frac{\|\theta_0\|^2}{\vartheta}\right].$$

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_\lambda$$
$$\leq \inf_{\rho=\mathcal{N}(\theta,s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right].$$

As $\pi = \mathcal{N}(0,\vartheta I)$ we have

$$\mathcal{K}(\rho,\pi) = \frac{1}{2}\left[M\left(\frac{s^2}{\vartheta} - 1 + \log\left(\frac{\vartheta}{s^2}\right)\right) + \frac{\|\theta_0\|^2}{\vartheta}\right].$$

Then

$$\int R \mathrm{d}\rho \leq R(\theta) + \int 2c\|u-\theta\|\rho(\mathrm{d}u) \leq R(\theta) + 2c\sqrt{M}\sigma.$$

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_{\lambda} \leq \inf_{\rho=\mathcal{N}(\theta,s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right].$$

As $\pi = \mathcal{N}(0,\vartheta I)$ we have

$$\mathcal{K}(\rho,\pi) = \frac{1}{2}\left[M\left(\frac{s^2}{\vartheta} - 1 + \log\left(\frac{\vartheta}{s^2}\right)\right) + \frac{\|\theta_0\|^2}{\vartheta}\right].$$

Then

$$\int R \mathrm{d}\rho \leq R(\theta) + \int 2c\|u-\theta\|\rho(\mathrm{d}u) \leq R(\theta) + 2c\sqrt{M}\sigma.$$

Chose adequate values for λ, ϑ and s^2 to conclude.

Test on real data

Dataset	Covariates	VB	SMC	SVM
Pima	7	21.3	22.3	30.4
Credit	60	33.6	32.0	32.0
DNA	180	23.6	23.6	20.4
SPECTF	22	06.9	08.5	10.1
Glass	10	19.6	23.3	4.7
Indian	11	25.5	26.2	26.8
Breast	10	1.1	1.1	1.7

Table – Comparison of misclassification rates (%). Last column : kernel-SVM with radial kernel. The hyper-parameters λ and ϑ are chosen by cross-validation.

Convexification of the loss

Can replace the 0/1 loss by a convex surrogate at "no" cost :

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*.

Convexification of the loss

Can replace the 0/1 loss by a convex surrogate at "no" cost :

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*.

- $R(\theta) = \mathbb{E}[(1 - Yf_\theta(X))_+]$ (hinge loss).
- $r_n(\theta) = \frac{1}{n}\sum_{i=1}^n (1 - Y_i f_\theta(X_i))_+$.
- Gaussian approx. : $\mathcal{F} = \left\{\mathcal{N}(\mu, \sigma^2 I), \mu \in \mathbb{R}^d, \sigma > 0\right\}$.

Convexification of the loss

Can replace the 0/1 loss by a convex surrogate at "no" cost :

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*.

- $R(\theta) = \mathbb{E}[(1 - Yf_\theta(X))_+]$ (hinge loss).
- $r_n(\theta) = \frac{1}{n}\sum_{i=1}^n (1 - Y_i f_\theta(X_i))_+$.
- Gaussian approx. : $\mathcal{F} = \left\{\mathcal{N}(\mu, \sigma^2 I), \mu \in \mathbb{R}^d, \sigma > 0\right\}$.

--→ the following criterion (which turns out to be convex !) :

$$\frac{1}{n}\sum_{i=1}^n (1 - Y_i \langle \mu, X_i \rangle) \Phi\left(\frac{1 - Y_i \langle \mu, X_i \rangle}{\sigma \|X_i\|_2}\right)$$

$$+\frac{1}{n}\sum_{i=1}^n \sigma \|X_i\| \varphi\left(\frac{1 - Y_i \langle \mu, X_i \rangle}{\sigma \|X_i\|_2}\right) + \frac{\|\mu\|_2^2}{2\vartheta} + \frac{d}{2}\left(\frac{\vartheta}{\sigma^2} - \log \sigma^2\right).$$

Application of the main theorem

Optimization with stochastic gradient descent on a ball of radius M. On this ball, the objetive function is L-Lipschitz. After k step, we have the approximation $\tilde{\rho}_\lambda^{(k)}$ of the posterior.

Corollary

Assume $\|X\| \leq c_x$ a.s., take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

$$\mathbb{P}\Bigg\{ \int R(\theta)\tilde{\rho}_\lambda^{(k)}(\mathrm{d}\theta) \leq \inf_\theta R(\theta) + \frac{LM}{\sqrt{1+k}} + \frac{c_x}{2}\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right) + \frac{\frac{c_x^2+1}{2c_x} + 2c_x \log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}} \Bigg\} \geq 1-\varepsilon.$$

The PACVB package (James Ridgway)

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

PACVB: Variational Bayes (VB) Approximation of Gibbs Posteriors with Hinge Losses

Variational Bayesian approximations of Gibbs measures with hinge losses for classification and ranking.

Version:	1.1
Depends:	Rcpp, MASS
LinkingTo:	Rcpp, RcppArmadillo, BH
Published:	2016-02-04
Author:	James Ridgway
Maintainer:	James Ridgway <james.ridgway at bristol.ac.uk>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	yes
CRAN checks:	PACVB results

Downloads:

Reference manual:	PACVB.pdf
Package source:	PACVB_1.1.tar.gz
Windows binaries:	r-devel: PACVB_1.1.zip, r-release: PACVB_1.1.zip, r-oldrel: PACVB_1.1.zip
OS X Snow Leopard binaries:	r-release: PACVB_1.1.tgz, r-oldrel: not available
OS X Mavericks binaries:	r-release: PACVB_1.1.tgz

Application to collaborative filtering

Application to collaborative filtering

Collaborative filtering as matrix completion

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

Collaborative filtering as matrix completion

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	? ? ?
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

Collaborative filtering as matrix completion

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	👍
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

1-bit matrix completion

Object of interest : an $m_1 \times m_2$ matrix M, values in

$$\{👎, 👍\} = \{-1, +1\}.$$

1-bit matrix completion

Object of interest : an $m_1 \times m_2$ matrix M, values in

{, } $= \{-1, +1\}$.

Entries $X_1 = (i_1, j_1), \ldots, (i_n, j_n)$ i.i.d from a distribution P, and $Y_\ell = M_{X_\ell}$.

1-bit matrix completion

Object of interest : an $m_1 \times m_2$ matrix M, values in

$\{$  ,  $\} = \{-1, +1\}.$

Entries $X_1 = (i_1, j_1), \dots, (i_n, j_n)$ i.i.d from a distribution P, and $Y_\ell = M_{X_\ell}$.

Usual assumption : $\mathrm{rank}(M) = r \ll \min(m_1, m_2)$.

Prior specification

Prior π

$$\underbrace{M}_{m_1 \times m_2} = \underbrace{L}_{m_1 \times K} \underbrace{R^T}_{K \times m_2},$$

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k), \quad \frac{1}{\gamma_k} \sim \Gamma(a, b).$$

Prior specification

Prior π

$$\underbrace{M}_{m_1 \times m_2} = \underbrace{L}_{m_1 \times K} \underbrace{R^T}_{K \times m_2},$$

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k), \quad \frac{1}{\gamma_k} \sim \Gamma(a, b).$$

Empirical hinge risk :

$$r(L, R) = \frac{1}{n} \sum_{\ell=1}^{n} \left(1 - Y_\ell (LR^T)_{X_\ell}\right)_+ .$$

Prior specification

Prior π

$$\underbrace{M}_{m_1 \times m_2} = \underbrace{L}_{m_1 \times K} \underbrace{R^T}_{K \times m_2},$$

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k), \quad \frac{1}{\gamma_k} \sim \Gamma(a, b).$$

Empirical hinge risk :

$$r(L, R) = \frac{1}{n} \sum_{\ell=1}^{n} \left(1 - Y_\ell (LR^T)_{X_\ell}\right)_+ .$$

Gibbs posterior : $\hat{\rho}_\lambda(L, R) = \exp\left[-\lambda r(L, R)\right] \pi(L, R)$.

Variational approximation

Here, family of approximation : $\rho_a = \rho_{(\mathcal{L},\mathcal{R},S,\Sigma,\alpha,\beta)}$

$$L_{i,k} \text{ indep. } \mathcal{N}(\mathcal{L}_{i,k}, S_{i,k}),\ R_{i,k} \text{ indep. } \mathcal{N}(\mathcal{R}_{i,k}, \Sigma_{i,k}),$$

$$\frac{1}{\gamma_k} \text{ indep. } \Gamma(\alpha_k, \beta_k).$$

Variational approximation

Here, family of approximation : $\rho_a = \rho_{(\mathcal{L},\mathcal{R},S,\Sigma,\alpha,\beta)}$

$$L_{i,k} \text{ indep. } \mathcal{N}(\mathcal{L}_{i,k}, S_{i,k}),\ R_{i,k} \text{ indep. } \mathcal{N}(\mathcal{R}_{i,k}, \Sigma_{i,k}),$$

$$\frac{1}{\gamma_k} \text{ indep. } \Gamma(\alpha_k, \beta_k).$$

In this case, the $\int r \mathrm{d}\rho_a$ is not tractable but we prove that

$$\forall a \in \mathcal{A}, \quad \int r \mathrm{d}\rho_a + \frac{\mathcal{K}(\rho_a, \pi)}{\lambda} \leq r\left(\mathcal{L}\mathcal{R}^T\right) + \mathcal{B}_\lambda(a)$$

for some known and tractable $\mathcal{B}_\lambda(a)$.

Definition

$$\tilde{\rho} = \arg\min_{\rho_a} r\left(\mathcal{L}\mathcal{R}^T\right) + \mathcal{B}_\lambda(a).$$

Application of the general result

Theorem

Cottet, V. & Alquier, P. (2018). 1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation. *Machine Learning*.

With proba. at least $1-\varepsilon$ on the sample,

$$\mathbb{P}_{(L,R)\sim\tilde{\rho},(i,j)\sim P}[\mathrm{sign}((LR^T)_{i,j}) \neq M_{i,j}] \leq \mathcal{C}\frac{r(m_1+m_2)\log(n)}{n}$$

for some (known) $\mathcal{C}>0$.

Application of the general result

Theorem

Cottet, V. & Alquier, P. (2018). 1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation. *Machine Learning*.

With proba. at least $1 - \varepsilon$ on the sample,

$$\mathbb{P}_{(L,R)\sim\tilde{\rho},(i,j)\sim P}[\mathrm{sign}((LR^T)_{i,j}) \neq M_{i,j}] \leq \mathcal{C}\frac{r(m_1+m_2)\log(n)}{n}$$

for some (known) $\mathcal{C} > 0$.

- in practice, blockwise coordinate optimization with gradient descent gives good results to compute $\tilde{\rho}$.

Application of the general result

Theorem

Cottet, V. & Alquier, P. (2018). 1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation. *Machine Learning*.

With proba. at least $1 - \varepsilon$ on the sample,

$$\mathbb{P}_{(L,R)\sim\tilde{\rho},(i,j)\sim P}[\mathrm{sign}((LR^T)_{i,j}) \neq M_{i,j}] \leq \mathcal{C}\frac{r(m_1+m_2)\log(n)}{n}$$

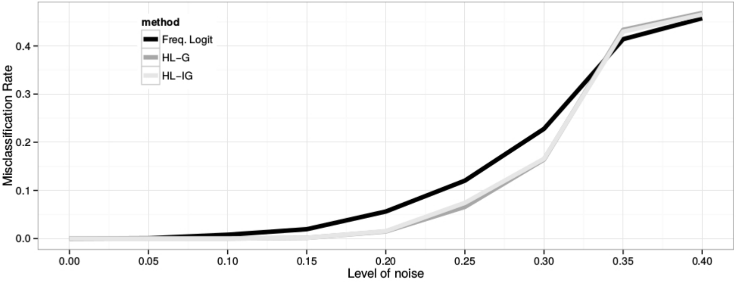
for some (known) $\mathcal{C} > 0$.

- in practice, blockwise coordinate optimization with gradient descent gives good results to compute $\tilde{\rho}$.
- in the paper, extention for noisy observations.

Simulation study

Comparison with the logistic regression approach with nuclear norm penalization from

J. Laffond, O. Klopp, E. Moulines & J. Salmon (2014). Probabilistic low-rank matrix completion on finite alphabets. *NIPS*.



1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

Reminder : concentration of the tempered posterior

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

Theorem

For $1/2 \leq \alpha < 1$, for any sequence (ε_n) such that

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

we have

$$\mathbb{E}\left[\int h^2(P_\theta, P_{\theta_0}) \pi_\alpha(\mathrm{d}\theta | X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

Analysis of VB approx.

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

Theorem

Alquier, P. & Ridgway, J. (2017). Concentration of Tempered Posteriors and of their Variational Approximations. *Preprint arXiv*.

Fix $1/2 \leq \alpha < 1$. Assume that for the sequence (ε_n) there is $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta_0}, P_\theta) \rho_n(\mathrm{d}\theta) \leq \varepsilon_n \text{ and } \mathcal{K}(\rho_n, \pi) \leq \varepsilon_n.$$

$$\text{Then } \mathbb{E}\left[\int h^2(P_\theta, P_{\theta_0}) \tilde{\pi}_\alpha(\mathrm{d}\theta | X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha} \varepsilon_n.$$

Further work (1/2)

- our paper contains applications to various statistical models (logistic regression, nonparametric regression estimation).

Further work (1/2)

- our paper contains applications to various statistical models (logistic regression, nonparametric regression estimation).
- our paper also contains results for the misspecified case where the true distribution of the X_i does not belong to $(P_\theta, \theta \in \Theta)$.

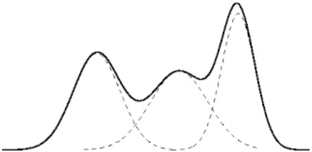
Further work (1/2)

- our paper contains applications to various statistical models (logistic regression, nonparametric regression estimation).
- our paper also contains results for the misspecified case where the true distribution of the X_i does not belong to $(P_\theta, \theta \in \Theta)$.
- the case $\alpha = 1$ ("proper" Bayesian inference) is not covered by our paper. It was recently analyzed by

F. Zhang & C. Gao (2017). Convergence rates of variational posterior distributions. *Preprint arXiv*.

This requires additional assumptions and does not cover the misspecified case.

Further work (2/2)



- according to a recent survey (Blei *et al*), one of the most popular applications of VB is to mixture models. The upper bound is also used for model selection. Blei states that there is no justification to this. Badr-Eddine Chérief-Abdellatif since proved this is consistent.

D. Blei, A. Kucukelbir & J. D. McAuliffe (2017). Variational inference : A review for statisticians. *Journal of the American Statistical Association*.

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

1. Introduction : Learning with PAC-Bayes Bounds
   - A PAC-Bayesian Bound for Batch Learning
   - A PAC-Bayesian Bound for Online Learning
   - Bayesian inference
2. Variational Approximation of the Posterior
   - Analysis of VB approximations of Gibbs posteriors
   - Applications : classification, collaborative filtering
   - Analysis of VB approximations of the Tempered Posterior
3. Discussion

Issue 1 : distortion of the posterior

- we proved that the VB approx. concentrates at the optimal rate but what about the closeness of the approx. to the true posterior ?

Issue 1 : distortion of the posterior

- we proved that the VB approx. concentrates at the optimal rate but what about the closeness of the approx. to the true posterior ?
- example : it is well known by practitioners that VB tends to reduce the variance of the posterior.

Issue 1 : distortion of the posterior

- we proved that the VB approx. concentrates at the optimal rate but what about the closeness of the approx. to the true posterior ?
- example : it is well known by practitioners that VB tends to reduce the variance of the posterior.
- is it possible to control the variance distortion ?

Issue 1 : distortion of the posterior

- we proved that the VB approx. concentrates at the optimal rate but what about the closeness of the approx. to the true posterior ?
- example : it is well known by practitioners that VB tends to reduce the variance of the posterior.
- is it possible to control the variance distortion ?
- controversial : it is also well known that Bayesian "credibility intervals" can be misleading.

Issue 2 : convergence of the optimization algorithm

- in the case of classification with hinge loss, we obtain a convex minimization problem. This also happens for logistic regression.

Issue 2 : convergence of the optimization algorithm

- in the case of classification with hinge loss, we obtain a convex minimization problem. This also happens for logistic regression.
- but in many other settings, the VB approximation is defined by a non-convex minimization problem. In this case, the convergence is an open issue in general. E.g : mixture models.

Issue 2 : convergence of the optimization algorithm

- in the case of classification with hinge loss, we obtain a convex minimization problem. This also happens for logistic regression.
- but in many other settings, the VB approximation is defined by a non-convex minimization problem. In this case, the convergence is an open issue in general. E.g : mixture models.
- the work on matrix completion relies on alternate optimisation of $d(M, UV)$ w.r.t U and V. The problem is convex in U, in V, but not in (U, V). Still, recent work gives hope that this procedure might converge :

R. Ge, J. D. Lee & T. Ma (2016). Matrix Completion has No Spurious Local Minimum. *NIPS*.

Issue 3 : online variational approximations

- online algorithms (like OGA) are sometimes used to compute the variational approximation. This is called online variational inference by

C. Wang, J. Paisley & D. Blei (2011). Online variational inference for the hierarchical Dirichlet process. *AISTATS*.

Issue 3 : online variational approximations

- online algorithms (like OGA) are sometimes used to compute the variational approximation. This is called online variational inference by

C. Wang, J. Paisley & D. Blei (2011). Online variational inference for the hierarchical Dirichlet process. *AISTATS*.

- but a more challenging question is to extend variational approximations to approximate EWA in the online setting (extend the result by [Audibert, 2016]) : it is well known that apart in the case of a finite number of predictors, EWA is not feasible in practice...

Issue 3 : reminder

Fix $\lambda > 0$ and define, at each time t,

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta) \text{ and } \hat{Y}_t = \int f_\theta(X_t)\hat{\rho}_{\lambda,t}(\mathrm{d}\theta).$$

Issue 3 : reminder

Fix $\lambda > 0$ and define, at each time t,

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta) \text{ and } \hat{Y}_t = \int f_\theta(X_t)\hat{\rho}_{\lambda,t}(\mathrm{d}\theta).$$

Theorem [Consequence of Audibert, 2006]

$$\sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t) \leq \inf_\rho \left\{ \int \sum_{t=1}^{T} \ell(Y_t, f_\theta(X_t))\rho(\mathrm{d}\theta) + \frac{\lambda T B^2}{2} + \frac{\mathcal{K}(\rho,\pi)}{\lambda} \right\}.$$

Issue 3 : extension of VB

Many definitions are possible. For example :

C. V. Nguyen, T. D. Bui, Y. Li & R. E. Turner (2017). Online Variational Bayesian Inference : Algorithms for Sparse Gaussian Processes and Theoretical Bounds. *ICML*.

propose

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta) \text{ and } \tilde{\rho}_{\lambda,t} = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}\left(\rho, \hat{\rho}_{\lambda,t}\right).$$

Interesting, but might computationally expensive, and there is no accurate theoretical analysis.

Issue 3 : extension of VB

Many definitions are possible. For example :

C. V. Nguyen, T. D. Bui, Y. Li & R. E. Turner (2017). Online Variational Bayesian Inference : Algorithms for Sparse Gaussian Processes and Theoretical Bounds. *ICML*.

propose

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta) \text{ and } \tilde{\rho}_{\lambda,t} = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}\left(\rho, \hat{\rho}_{\lambda,t}\right).$$

Interesting, but might computationally expensive, and there is no accurate theoretical analysis.

We work currently on an alternative approach with Badr-Eddine.

Thank you !