On the Properties of Variational Approximations of Gibbs Posteriors

Pierre Alquier

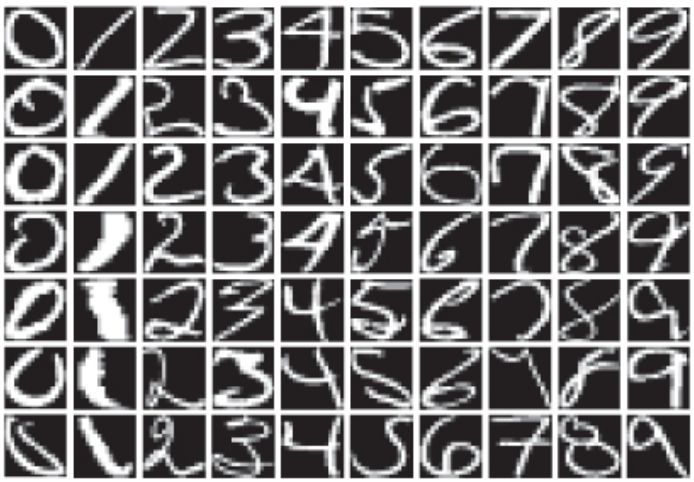
Computational and Statistical Trade-offs in Learning - IHES - 23/03/2016

Learning vs. estimation

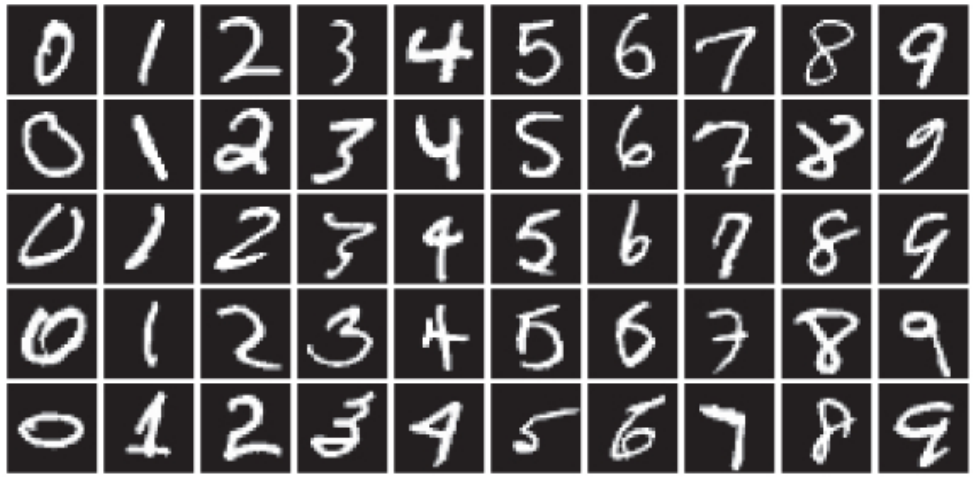
In many applications one would like to learn from a sample without being able to write the likelihood.

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

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

1st example : general bound for batch learning

Context :

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

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, O. (2007). *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of Lecture Notes-Monograph Series, IMS.

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

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

$$\geq 1-\varepsilon.$$

Catoni's bound for batch learning

Theorem

Catoni, O. (2007). *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of Lecture Notes-Monograph Series, IMS.

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

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

$$\geq 1-\varepsilon.$$

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*.

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

With π the uniform distribution on $\{\theta_1, \ldots, \theta_M\}$ 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 $\theta_1, \ldots, \theta_M$

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

$$\begin{aligned}
&\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \\
&\quad \leq \inf_{\rho=\delta_{\theta_i}} \left[\int R \mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right] \\
&\quad \leq \inf_{1 \leq i \leq M} \left[R(\theta_i) + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\log(M) + \log\left(\frac{2}{\varepsilon}\right) \right] \right]
\end{aligned}$$

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

With π the uniform distribution on $\{\theta_1, \ldots, \theta_M\}$ 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 $\lambda = \dfrac{\sqrt{2n\log(M)}}{B}$.

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

Gerchinovitz, S. (2011). *PhD Thesis, Univ. Paris Sud.*

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning and Games*, Cambridge.

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

Bibliographical remarks (1/2)

- **Improved rates for specific losses** : exp-convexity, etc. For quadratic loss, see the bound by

Dalalyan, A. & Tsybakov, A. (2008). Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity. *Machine Learning*.

Bibliographical remarks (1/2)

- **Improved rates for specific losses** : exp-convexity, etc. For quadratic loss, see the bound by

Dalalyan, A. & Tsybakov, A. (2008). Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity. *Machine Learning*.

- Other works on **aggregation theory** :

Leung, G. and Barron, A. (2006). Information Theory and Mixing Least-Square Regressions. *IEEE Trans. on Information Theory*.

Bibliographical remarks (2/2)

- Link with **decision theory and Bayesian statistics** :

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

Bayesian interpretation : "pseudo-likelihood".


Bissiri, P., Holmes, C. and Walker, S. (2013). Fast learning Rates in Statistical Inference through Aggregation. *Preprint*.


Grünwald, P. D. & van Ommen, T. (2013). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Preprint*.

Bibliographical remarks (2/2)

- Link with **decision theory and Bayesian statistics** :

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

Bayesian interpretation : "pseudo-likelihood".


Bissiri, P., Holmes, C. and Walker, S. (2013). Fast learning Rates in Statistical Inference through Aggregation. *Preprint*.


Grünwald, P. D. & van Ommen, T. (2013). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Preprint*.

- Link with **posterior concentration of the posterior** :


Ghoshal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence Rates of Posterior Distributions. *Annals of Statistics*.

(different objectives, but computation often similar).

Reminder : EWA

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

Reminder : EWA

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

Depending on the setting, we have to

- sample from $\hat{\rho}_\lambda$,
- compute $\int f_\theta(\cdot)\hat{\rho}_\lambda(\mathrm{d}\theta)$.

A natural idea : MCMC methods

Langevin Monte-Carlo :

Dalalyan, A. and Tsybakov, A. (2011). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Science*.

Markov Chain Monte-Carlo :

Alquier, P. & Biau, G. (2013). Sparse Single-Index Model. *Journal of Machine Learning Reseach*.

A natural idea : MCMC methods

Langevin Monte-Carlo :

Dalalyan, A. and Tsybakov, A. (2011). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Science*.

Markov Chain Monte-Carlo :

Alquier, P. & Biau, G. (2013). Sparse Single-Index Model. *Journal of Machine Learning Reseach*.

However : very hard to prove the convergence of the algorithm. Usually not possible to provide guarantees after a finite number of steps. See however

Joulin, A. & Ollivier, Y. (2010). Curvature, Concentration, and Error Estimates for Markov Chain Monte Carlo. *The Annals of Probability*.

Dalalyan, A. (2014). Theoretical Guarantees for Approximate Sampling from a Smooth and Log-Concave Density. *Preprint*.

Variational Bayes methods

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

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

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

Variational Bayes methods

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

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

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 :

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

Variational Bayes methods

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

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

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 :

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

Theoretical guarantees on the approximation ?

VB in PAC-Bayesian 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 PAC-Bayesian 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. *Preprint, accepted for publication in 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. *Preprint, accepted for publication in JMLR*.

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

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

$$\geq 1 - \varepsilon.$$

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

Proof (1/4)

Hoeffding's inequality :

$$\mathbb{E} \exp \left\{ \lambda \left[R(\theta) - r(\theta) \right] \right\} \leq \exp \left(\frac{\lambda^2}{2n} \right)$$

Proof (1/4)

Hoeffding's inequality :

$$\mathbb{E} \exp \left\{ \lambda \left[R(\theta) - r(\theta) \right] \right\} \leq \exp \left(\frac{\lambda^2}{2n} \right)$$

that is

$$\mathbb{E} \exp \left\{ \lambda \left[R(\theta) - r(\theta) - \frac{\lambda}{2n} \right] + \log \varepsilon \right\} \leq \varepsilon.$$

Proof (1/4)

Hoeffding's inequality :

$$\mathbb{E} \exp \left\{ \lambda \left[R(\theta) - r(\theta) \right] \right\} \leq \exp \left(\frac{\lambda^2}{2n} \right)$$

that is

$$\mathbb{E} \exp \left\{ \lambda \left[R(\theta) - r(\theta) - \frac{\lambda}{2n} \right] + \log \varepsilon \right\} \leq \varepsilon.$$

Integrate w.r.t π + Fubini :

$$\mathbb{E} \int \exp \left\{ \lambda \left[R(\theta) - r(\theta) - \frac{\lambda}{2n} \right] + \log \varepsilon \right\} \pi(\mathrm{d}\theta) \leq \varepsilon.$$

Proof (2/4)

$$\mathbb{E} \int \exp\left\{ \lambda \left[R(\theta) - r(\theta) - \frac{\lambda}{2n} \right] + \log \varepsilon \right\} \pi(\mathrm{d}\theta) \leq \varepsilon.$$

Lemma

$$\log \int \exp(h) \mathrm{d}\pi = \sup_{\rho} \left[\int h \mathrm{d}\pi - \mathcal{K}(\rho, \pi) \right]$$

Proof (2/4)

$$\mathbb{E} \int \exp\left\{ \lambda \left[R(\theta) - r(\theta) - \frac{\lambda}{2n} \right] + \log \varepsilon \right\} \pi(\mathrm{d}\theta) \leq \varepsilon.$$

Lemma

$$\log \int \exp(h) \mathrm{d}\pi = \sup_{\rho} \left[\int h \mathrm{d}\pi - \mathcal{K}(\rho, \pi) \right]$$

$$\mathbb{E} \exp \sup_{\rho} \left\{ \lambda \int \left[R(\theta) - r(\theta) - \frac{\lambda}{2n} \right] \rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \pi) + \log \varepsilon \right\} \leq \varepsilon.$$

Proof (2/4)

$$\mathbb{E}\int \exp\left\{\lambda\left[R(\theta)-r(\theta)-\frac{\lambda}{2n}\right]+\log\varepsilon\right\}\pi(\mathrm{d}\theta)\leq\varepsilon.$$

Lemma

$$\log\int\exp(h)\mathrm{d}\pi=\sup_{\rho}\left[\int h\mathrm{d}\pi-\mathcal{K}(\rho,\pi)\right]$$

$$\mathbb{E}\exp\sup_{\rho}\left\{\lambda\int\left[R(\theta)-r(\theta)-\frac{\lambda}{2n}\right]\rho(\mathrm{d}\theta)-\mathcal{K}(\rho,\pi)+\log\varepsilon\right\}\leq\varepsilon.$$

Markov's inequality :

$$\mathbb{P}\left(\forall\rho,\int R\mathrm{d}\rho\leq\int r\mathrm{d}\rho+\frac{\lambda}{2n}+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{1}{\varepsilon}\right)}{\lambda}\right)\geq 1-\varepsilon.$$

Proof (3/4)

$$\forall \rho, \int R \mathrm{d}\rho \leq \int r \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}$$

Proof (3/4)

$$\forall \rho, \int R \mathrm{d}\rho \leq \int r \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}$$

We remind that $\hat{\rho}_\lambda$ is the minimizer of this upper bound, so :

$$\int R \mathrm{d}\hat{\rho}_\lambda \leq \inf_\rho \left[\int r \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right].$$

Proof (3/4)

$$\forall \rho, \int R \mathrm{d}\rho \leq \int r \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}$$

We remind that $\hat{\rho}_\lambda$ is the minimizer of this upper bound, so :

$$\int R \mathrm{d}\hat{\rho}_\lambda \leq \inf_\rho \left[\int r \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right].$$

Restrict your attention to $\rho \in \mathcal{F} = \{\rho_a, a \in \mathcal{A}\}$:

Empirical bound

$$\forall a \in \mathcal{A}, \int R \mathrm{d}\rho_a \leq \inf_{a \in \mathcal{A}} \left[\int r \mathrm{d}\rho_a + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho_a, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right].$$

Proof (4/4)

End of the proof : in a similar way, we prove that with proba. at least $1-\varepsilon$,

$$\forall \rho, \int r \mathrm{d}\rho \leq \int R \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}.$$

Proof (4/4)

End of the proof : in a similar way, we prove that with proba. at least $1-\varepsilon$,

$$\forall \rho, \int r \mathrm{d}\rho \leq \int R \mathrm{d}\rho + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}.$$

So :

$$\begin{aligned}
\int R \mathrm{d}\tilde{\rho}_\lambda &\leq \int r \mathrm{d}\rho_{\hat{a}_\lambda} + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho_{\hat{a}_\lambda},\pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \\
&= \inf_{a\in\mathcal{A}} \left[\int r \mathrm{d}\rho_a + \frac{\lambda}{2n} + \frac{\mathcal{K}(\rho_a,\pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right] \\
&\leq \inf_{a\in\mathcal{A}} \left[\int R \mathrm{d}\rho_a + \frac{\lambda}{n} + 2\frac{\mathcal{K}(\rho_a,\pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right].
\end{aligned}$$

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

$\dashrightarrow$ 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

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

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

How to use PACVB ?

```
> X
     [,1]        [,2]      [,3]
[1,]    1 -1.48290060 0.3974124
[2,]    1 -1.05599316 0.2554146
[3,]    1  0.63464838 1.5370450
[4,]    1 -0.36583539 1.5540228
[5,]    1  0.08339866 0.9395758
...
> Y
[1]  1 -1  1  1  1...
```

How to use PACVB ?

```
 > Sol = GDHinge(X,Y,lambda=25)
 > Sol
$m
          [,1]         [,2]       [,3]
[1,] 2.223396 -0.02744416 -1.205612

$s
[1] -1.406072

$bound
[1] 0.7990907
```

Other models

Also provided in the paper :

- a complete analysis of ranking through linear score functions,

Other models

Also provided in the paper :

- a complete analysis of ranking through linear score functions,
- a sketch of the analysis of matrix factorization (the theory is not complete yet).

Other models

Also provided in the paper :

- a complete analysis of ranking through linear score functions,
- a sketch of the analysis of matrix factorization (the theory is not complete yet).

Current work in progress :

- improved minimization procedures to include in the package (James Ridgway),

Other models

Also provided in the paper :

- a complete analysis of ranking through linear score functions,
- a sketch of the analysis of matrix factorization (the theory is not complete yet).

Current work in progress :

- improved minimization procedures to include in the package (James Ridgway),
- extension to other PAC-Bayes bounds (online bound, Dalalyan & Tsybakov bound)...

A (useless ?) bound in the online case

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta)$$

$$\tilde{\rho}_{\lambda,t} = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \hat{\rho}_{\lambda,t}) \text{ and } \tilde{Y}_t = \int f_\theta(X_t)\tilde{\rho}_{\lambda,t}(\mathrm{d}\theta).$$

A (useless ?) bound in the online case

$$\hat{\rho}_{\lambda,t}(\mathrm{d}\theta) \propto \exp[-\lambda r_{t-1}(\theta)]\pi(\mathrm{d}\theta)$$

$$\tilde{\rho}_{\lambda,t} = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \hat{\rho}_{\lambda,t}) \text{ and } \tilde{Y}_t = \int f_\theta(X_t)\tilde{\rho}_{\lambda,t}(\mathrm{d}\theta).$$

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

A (useless ?) bound in the online case

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

$$\Delta_t(\mathcal{F}) = \inf_{\rho \in \mathcal{F}} \left\{ \int \sum_{h=1}^{t} \ell(Y_h, f_\theta(X_h)) \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\} - \inf_{\rho} \left\{ \int \sum_{h=1}^{t} \ell(Y_h, f_\theta(X_h)) \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}.$$

Thank you !