PAC-Bayesian Bounds and Aggregation: Introduction, and Algorithmic Issues

Pierre Alquier

CREST
ENSAE ParisTech
université PARIS-SACLAY

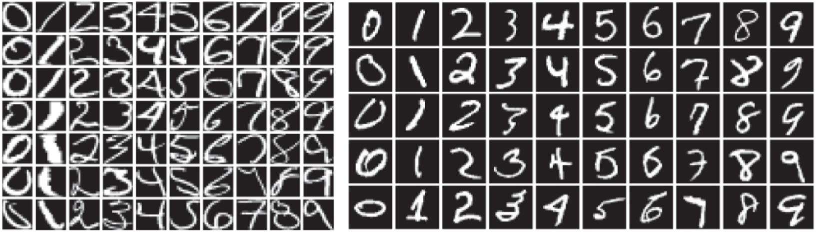
Statistics/Learning at Paris-Saclay - IHES - 08/01/2016

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 : $(X_1, Y_1), (X_2, Y_2), \ldots$
  $\rightarrow$ 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 : fixed design regression

Context :

- $X_1, \ldots, X_n$ deterministic ; $Y_i = f(X_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (say).

1st example : fixed design regression

Context :

- $X_1, \ldots, X_n$ deterministic ; $Y_i = f(X_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (say).
- any $(f_\theta(\cdot) = \langle \theta, g(\cdot) \rangle, \theta \in \mathbb{R}^p)$.

1st example : fixed design regression

Context :

- $X_1, \ldots, X_n$ deterministic ; $Y_i = f(X_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (say).
- any $(f_\theta(\cdot) = \langle \theta, g(\cdot) \rangle, \theta \in \mathbb{R}^p)$.
- $R(\theta) = \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - f_\theta(X_i)]^2$.

1st example : fixed design regression

Context :

- $X_1, \dots, X_n$ deterministic ; $Y_i = f(X_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (say).
- any $(f_\theta(\cdot) = \langle \theta, g(\cdot) \rangle, \theta \in \mathbb{R}^p)$.
- $R(\theta) = \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - f_\theta(X_i)]^2$.
- $r_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - f_\theta(X_i)]^2$.

1st example : fixed design regression

Context :

- $X_1, \ldots, X_n$ deterministic ; $Y_i = f(X_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (say).
- any $(f_\theta(\cdot) = \langle \theta, g(\cdot) \rangle, \theta \in \mathbb{R}^p)$.
- $R(\theta) = \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - f_\theta(X_i)]^2$.
- $r_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - f_\theta(X_i)]^2$.
- any prior π.

Dalalyan and Tsybakov's bound for EWA

Theorem

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

$$\forall \lambda \leq \frac{n}{4\sigma^2} : \quad \mathbb{E}\left\{ R\left[\int \theta \hat{\rho}_\lambda(\mathrm{d}\theta)\right]\right\} \leq \inf_{\rho}\left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right]$$

Dalalyan and Tsybakov's bound for EWA

Theorem

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

$$\forall \lambda \leq \frac{n}{4\sigma^2} : \quad \mathbb{E}\left\{ R\left[\int \theta \hat{\rho}_\lambda(\mathrm{d}\theta) \right] \right\} \leq \inf_\rho \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi) \right]$$

Based on previous work :

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

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

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

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

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

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

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

$$\leq \inf_{1\leq i\leq M}\left[\int R(\theta)\delta_{\theta_i}(\mathrm{d}\theta) + 4\sigma^2\mathcal{K}(\delta_{\theta_i},\pi)\right]$$

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

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

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

$$\leq \inf_{1\leq i\leq M}\left[\int R(\theta)\delta_{\theta_i}(\mathrm{d}\theta) + 4\sigma^2\mathcal{K}(\delta_{\theta_i},\pi)\right]$$

$$= \inf_{1\leq i\leq M}\left[R(\theta_i) + 4\sigma^2\log(M)\right].$$

Application : linear regression

With $\pi = \mathcal{N}(0, S^2 I_M)$,

$$\mathbb{E}\left\{ R\left[\int \theta \hat{\rho}_\lambda(\mathrm{d}\theta)\right]\right\} \leq \inf_{\rho=\mathcal{N}(\theta_0, s^2 I_M)} \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right].$$

Application : linear regression

With $\pi = \mathcal{N}(0, S^2 I_M)$,

$$\mathbb{E}\left\{R\left[\int \theta\hat{\rho}_\lambda(\mathrm{d}\theta)\right]\right\} \leq \inf_{\rho=\mathcal{N}(\theta_0, s^2 I_M)}\left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right].$$

As $\mathcal{K}(\rho,\pi) = \frac{1}{2}\left[M\left(\frac{s^2}{S^2} - 1 + \log\left(\frac{S^2}{s^2}\right)\right) + \frac{\|\theta_0\|^2}{S^2}\right]$ and (rough) calculations lead to $\int R(\theta)\rho(\mathrm{d}\theta) \leq R(\theta_0) + M^2\|g\|_\infty^2 s^2$,

Application : linear regression

With $\pi = \mathcal{N}(0, S^2 I_M)$,

$$\mathbb{E}\left\{R\left[\int \theta \hat{\rho}_\lambda(\mathrm{d}\theta)\right]\right\} \leq \inf_{\rho=\mathcal{N}(\theta_0, s^2 I_M)} \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right].$$

As $\mathcal{K}(\rho,\pi) = \frac{1}{2}\left[M\left(\frac{s^2}{S^2} - 1 + \log\left(\frac{S^2}{s^2}\right)\right) + \frac{\|\theta_0\|^2}{S^2}\right]$ and (rough) calculations lead to $\int R(\theta)\rho(\mathrm{d}\theta) \leq R(\theta_0) + M^2\|g\|_\infty^2 s^2$,

$$\mathbb{E}\left\{R\left[\int \theta \hat{\rho}_\lambda(\mathrm{d}\theta)\right]\right\} \leq \inf_{\theta_0 \in \mathbb{R}^M}\left\{R(\theta_0) + \frac{4M\sigma^2}{n}\log\left(\frac{S^2 Mn}{\mathrm{e}}\right) + \frac{1}{n}\left[\frac{\|\theta\|_0^2 + 1}{S^2} + \|g\|_\infty^2\right]\right\}.$$

2nd example : general bound for batch learning

Context :

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

2nd example : general bound for batch learning

Context :

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

2nd 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 $|\ell(\cdot,\cdot)| \leq B$.

2nd 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 $|\ell(\cdot,\cdot)| \leq B$.
- $r_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$.

2nd 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.

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

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

$$\geq 1 - \varepsilon.$$

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.

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

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*.

3rd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* other assumption than $|Y_i| \leq B$.

3rd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* other assumption than $|Y_i| \leq B$.
- any $(f_\theta, \theta \in \Theta)$, with $|f(\theta)(x)| \leq B$.

3rd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* other assumption than $|Y_i| \leq B$.
- any $(f_\theta, \theta \in \Theta)$, with $|f(\theta)(x)| \leq B$.
- 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* :

$$\mathcal{R} = \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 - \inf_{\theta} \sum_{t=1}^{T} (Y_t - f_\theta(X_t))^2.$$

3rd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* other assumption than $|Y_i| \leq B$.
- any $(f_\theta, \theta \in \Theta)$, with $|f(\theta)(x)| \leq B$.
- 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* :

$$\mathcal{R} = \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 - \inf_\theta \sum_{t=1}^{T} (Y_t - f_\theta(X_t))^2.$$

- at time t we can use as a proxy of the quality of θ : $r_{t-1}(\theta) = \sum_{\ell=1}^{t-1} (Y_\ell - f_\theta(X_\ell))^2$.

3rd example : online learning

- (X_1, Y_1), (X_2, Y_2), ... without *any* other assumption than $|Y_i| \leq B$.
- any $(f_\theta, \theta \in \Theta)$, with $|f(\theta)(x)| \leq B$.
- 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* :

$$\mathcal{R} = \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 - \inf_{\theta} \sum_{t=1}^{T} (Y_t - f_\theta(X_t))^2.$$

- at time t we can use as a proxy of the quality of θ : $r_{t-1}(\theta) = \sum_{\ell=1}^{t-1} (Y_\ell - f_\theta(X_\ell))^2$.
- any prior π.

Audibert / Gerchinovitz's bound for online learning

Fix $\lambda \leq \frac{1}{8B^2}$ and define, at each time t :

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

Audibert / Gerchinovitz's bound for online learning

Fix $\lambda \leq \frac{1}{8B^2}$ 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). Sparsity Regret Bounds for Individual Sequences in Online Linear Regression. *COLT'11*.

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

Audibert / Gerchinovitz's bound for online learning

Fix $\lambda \leq \frac{1}{8B^2}$ 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). Sparsity Regret Bounds for Individual Sequences in Online Linear Regression. *COLT'11*.

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

Based on a result with general loss to be found in

Audibert, J.-Y. (2009). Fast learning Rates in Statistical Inference through Aggregation. *Annals of Statistics*.

Bibliographical remarks (1/2)

"Catoni's type bound" : under the name "PAC-Bayesian bounds", many authors including Langford, Seeger, Meir, Cesa-Bianchi, Li, Jiang, Tanner, Laviolette, Guedj, sorry for not being exhaustive, see the papers for more references !

Bibliographical remarks (1/2)

"Catoni's type bound" : under the name "PAC-Bayesian bounds", many authors including Langford, Seeger, Meir, Cesa-Bianchi, Li, Jiang, Tanner, Laviolette, Guedj, sorry for not being exhaustive, see the papers for more references !

"Dalalyan-Tsybakov's type" bound : under the name "Exponentially Weighted Aggregation", Golubev, Suzuki, Montuelle, Le Pennec, Robbiano, Salmon...

Bibliographical remarks (1/2)

"Catoni's type bound" : under the name "PAC-Bayesian bounds", many authors including Langford, Seeger, Meir, Cesa-Bianchi, Li, Jiang, Tanner, Laviolette, Guedj, sorry for not being exhaustive, see the papers for more references !

"Dalalyan-Tsybakov's type" bound : under the name "Exponentially Weighted Aggregation", Golubev, Suzuki, Montuelle, Le Pennec, Robbiano, Salmon...

Related to other works on aggregation : Vovk, Rissanen, Abramovitch, Nemirovski, Yang, Rigollet, Lecué, Bellec, Michel, Gaïffas...

Bibliographical remarks (2/2)

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

Bayesian interpretation : $\exp\left[-\lambda r(\theta)\right] =$ "pseudo-likelihood".

Bibliographical remarks (2/2)

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

Bayesian interpretation : $\exp\left[-\lambda r(\theta)\right] =$ "pseudo-likelihood".

Decision theory and Bayesian statistics : more authors advocate the use of $\hat{\rho}_\lambda$: Miller, Dunson...

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)

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

Bayesian interpretation : $\exp\left[-\lambda r(\theta)\right] =$ "pseudo-likelihood".

Decision theory and Bayesian statistics : more authors advocate the use of $\hat{\rho}_\lambda$: Miller, Dunson...

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*.

Asymptotic study of Bayesian estimators : Ghosh, Ghoshal, van der Vaart, Gassiat, Rousseau, Castillo... different from PAC-Bayes but most calculations are 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 $\int \theta \hat{\rho}_{\lambda}(\mathrm{d}\theta)$.

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 :

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

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

VB in PAC-Bayesian framework

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

Then :

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

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*.

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

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

$$\geq 1-\varepsilon.$$

A PAC-Bound for VB Approximation

Theorem

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

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

--→ if we can derive a tight oracle inequality from this bound, we know that the VB approximation is sensible !

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 :

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

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...

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

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

Dataset	Convex VB	VB	SMC	SVM
Pima	21.8	21.3	22.3	30.4
Credit	27.2	33.6	32.0	32.0
DNA	4.2	23.6	23.6	20.4
SPECTF	19.2	06.9	08.5	10.1
Glass	26.1	19.6	23.3	4.7
Indian	26.2	25.5	26.2	26.8
Breast	0.5	1.1	1.1	1.7

Table: Comparison of misclassification rates (%), including the convexified version of VB.

Convergence graphs

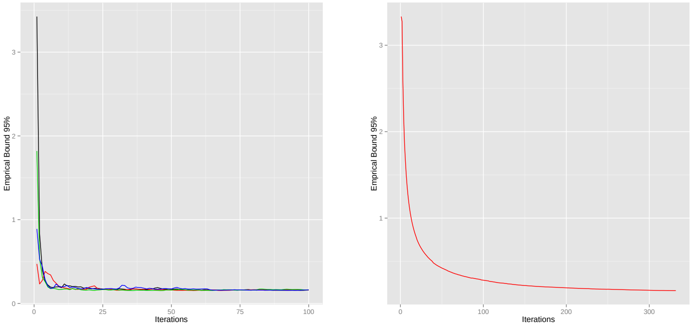


Figure: Stochastic gradient descent, Pima and Adult datasets.

Thanks & best wishes for 2016 !