On the Properties of Variational Approximations in Statistical Learning.

Pierre Alquier

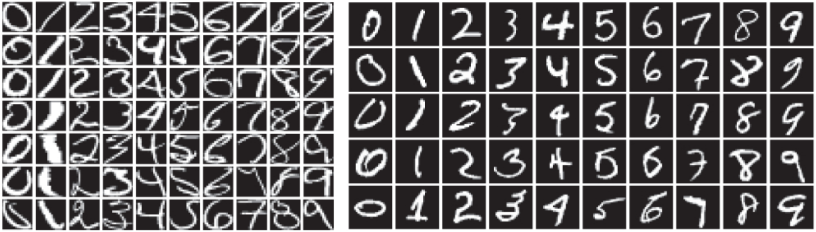
UCD Dublin - Statistics Seminar - 29/10/15

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... In this talk, $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

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... In this talk, (X_1, Y_1), ..., (X_n, Y_n) i.i.d.
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.

Typical machine learning problem

Main ingredients :

- observations object-label : $(X_1, Y_1), (X_2, Y_2), \ldots$
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request... In this talk, $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
- 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... In this talk, (X_1, Y_1), ..., (X_n, Y_n) i.i.d.
- 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... In this talk, (X_1, Y_1), ..., (X_n, Y_n) i.i.d.
- 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 $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$ (classification error). In this talk $R(\theta) = \mathbb{E}[\ell(Y, f_\theta(X))]$. We want to minimize $R(\theta)$. But note that it is unknown in practice.

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... In this talk, (X_1, Y_1), ..., (X_n, Y_n) i.i.d.
- 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 $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$ (classification error). In this talk $R(\theta) = \mathbb{E}[\ell(Y, f_\theta(X))]$. We want to minimize $R(\theta)$. But note that it is unknown in practice.
- 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... In this talk, (X_1, Y_1), ..., (X_n, Y_n) i.i.d.
- 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 $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$ (classification error). In this talk $R(\theta) = \mathbb{E}[\ell(Y, f_\theta(X))]$. We want to minimize $R(\theta)$. But note that it is unknown in practice.
- an empirical proxy $r(\theta)$ for this criterion of success :
  $\rightarrow$ here $r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))$.

Empirical risk minimization (ERM)

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} r(\theta).$$

Empirical risk minimization (ERM)

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} r(\theta).$$

Theorem (Vapnik and Chervonenkis, in the 70's)

Vapnik, V. (1998). *Statistical Learning Theory*, Springer.

Classification setting. Let d_Θ denote the VC-dim. of Θ.

$$\mathbb{P}\left\{ R(\hat{\theta}_n) \leq \inf_{\theta\in\Theta} R(\theta) + 4\sqrt{\frac{d_\Theta \log(n+1) + \log(2)}{n}} + \sqrt{\frac{\log(2/\varepsilon)}{2n}} \right\} \geq 1-\varepsilon.$$

ERM with linear classifiers

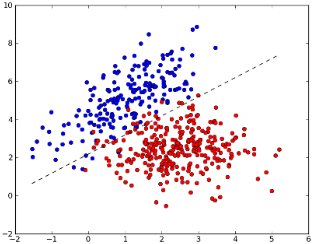


Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

ERM with linear classifiers

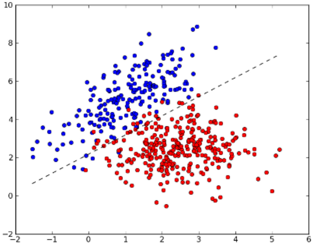


Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

Here $d_\Theta = 3$, $n = 500$.

ERM with linear classifiers

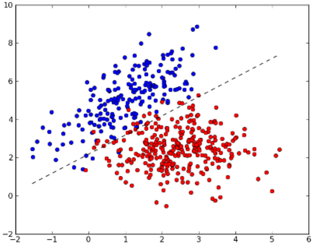


Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

Here $d_\Theta = 3$, $n = 500$. With probability at least 90%,

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 0.842.$$

ERM with linear classifiers

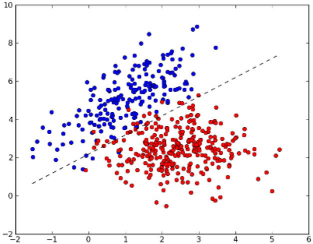


Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

Here $d_\Theta = 3$, $n = 500$. With probability at least 90%,

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 0.842.$$

With $n = 5000$ we would have

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 0.301.$$

The PAC-Bayesian approach : origins

Idea : combine these tools with a prior π on Θ.

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.*

"A PAC performance guarantee theorem applies to a broad class of experimental settings. A Bayesian correctness theorem applies to only experimental settings consistent with the prior used in the algorithm. However, in this restricted class of settings the Bayesian learning algorithm can be optimal and will generally outperform PAC learning algorithms. (...) The PAC-Bayesian theorems and algorithms (...) attempt to get the best of both PAC and Bayesian approaches by combining the ability to be tuned with an informal prior with PAC guarantees that hold in all i.i.d experimental settings."

The PAC-Bayesian approach

EWA / pseudo-posterior / Gibbs estimator / ...

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

The PAC-Bayesian approach

EWA / pseudo-posterior / Gibbs estimator / ...

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

Theorem - for a bounded loss $\ell \leq B$.

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

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

Another point of view

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

Provides decision theoretic reason to use

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

instead of

$$\pi(\mathrm{d}\theta|(X_1, Y_1), \ldots, (X_n, Y_n)) \propto \mathcal{L}(\theta)\pi(\mathrm{d}\theta).$$

- The likelihood $\mathcal{L}(\theta)$ might be too complicated or not even available ;
- We might think it's safer to replace it by a *robust* loss function (Huber...).

Bibliographical remarks

PAC-Bayesian bounds : many authors including Langford, Seeger, Meir, Cesa-Bianchi, Li, Jiang, Tanner, Laviolette, sorry for not being exhaustive, see the papers for more references !

Bibliographical remarks

PAC-Bayesian bounds : many authors including Langford, Seeger, Meir, Cesa-Bianchi, Li, Jiang, Tanner, Laviolette, sorry for not being exhaustive, see the papers for more references !

Related to other works on aggregation : Barron, Vovk, Rissanen, Abramovitch, Nemirovski, Yang, Zhang, Rigollet, Lecué, Bellec, Suzuki...

Bibliographical remarks

PAC-Bayesian bounds : many authors including Langford, Seeger, Meir, Cesa-Bianchi, Li, Jiang, Tanner, Laviolette, sorry for not being exhaustive, see the papers for more references !

Related to other works on aggregation : Barron, Vovk, Rissanen, Abramovitch, Nemirovski, Yang, Zhang, Rigollet, Lecué, Bellec, Suzuki...

Related work on misspecification in Bayesian statistics : the "safe Bayes rule" of

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

Reminder : pseudo-posterior

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

Reminder : pseudo-posterior

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

Depending on the setting, we have to

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

Reminder : pseudo-posterior

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

Depending on the setting, we have to

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

How to do it ?

A natural idea : MCMC methods for PAC-Bayes

Langevin Monte-Carlo :

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

A natural idea : MCMC methods for PAC-Bayes

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*.

Guedj, B. & Alquier, P. (2013). PAC-Bayesian Estimation and Prevision in Sparse Additive Models. *Electronic Journal of Statistics*.

A natural idea : MCMC methods for PAC-Bayes

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*.

Guedj, B. & Alquier, P. (2013). PAC-Bayesian Estimation and Prevision in Sparse Additive Models. *Electronic Journal of Statistics*.

However : usually not possible to provide guarantees after a finite number of steps. See however

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 $\tilde{\pi}(\theta)$:

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

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

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 $\tilde{\pi}(\theta)$:

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

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

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

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

Example : Gaussian approximation

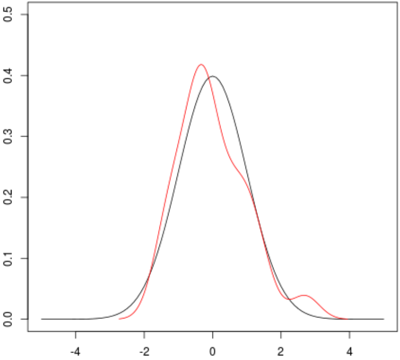


Table: The true posterior and the best Gaussian approximation.

VB in PAC-Bayesian framework

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

Then :

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

VB in PAC-Bayesian framework

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

Then :

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

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*.

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

A PAC-Bound for VB Approximation

Theorem

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

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

--→ if the infimum on the right is small enough, VB approximation is "at no cost".

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

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

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 $\pi = \mathcal{N}(0, \vartheta I)$.

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 : $F_\lambda(\mu, \Sigma) =$

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

Application of the main theorem

Corollary

Assume that, for $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle \theta, X\rangle \langle \theta', X\rangle < 0) \leq c\|\theta - \theta'\|$ and take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

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

Application of the main theorem

Corollary

Assume that, for $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle \theta, X\rangle \langle \theta', X\rangle < 0) \leq c\|\theta - \theta'\|$ and take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

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

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

Implementation : deterministic annealing

Algorithm 1 Deterministic annealing

Input $(\lambda_t)_{t \in [0,T]}$ a sequence of temperature

Init. Set $\mu = 0$ and $\Sigma = \vartheta I_d$, the values minimizing KL-divergence for $\lambda = 0$

Loop t=1,...,T

a. $\mu^{\lambda_t}, \Sigma^{\lambda_t}$ = Minimize $F^{\lambda_t}(m, \Sigma)$ using some local optimization routine (gradient descent) with initial points $\mu^{\lambda_{t-1}}, \Sigma^{\lambda_{t-1}}$

b. Break if the empirical bound increases.

End Loop

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

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

(One more) test on real data

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

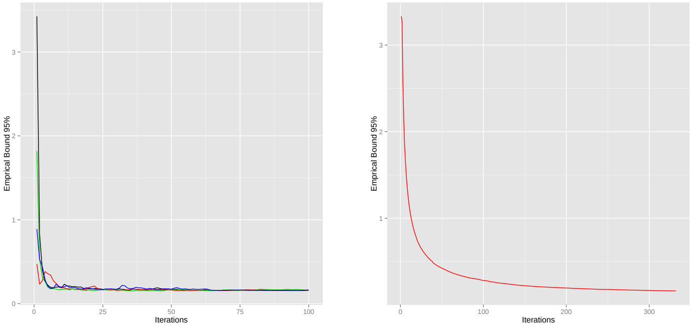


Figure: Stochastic gradient descent, Pima and Adult datasets.