Machine learning theory for time series

Exponential inequalities for nonstationary Markov chains

Pierre Alquier

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

CIMFAV seminar
January 16, 2019

CIMFAV
CENTRO DE INVESTIGACIÓN Y MODELAMIENTO DE FENÓMENOS ALEATORIOS - VALPARAÍSO

Universidad de Valparaíso CHILE

Generic machine learning problem

Main ingredients :

Generic machine learning problem

Main ingredients :

- observations : $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- A loss function ℓ
  $\rightarrow$ $\ell(y' - y)$ incurred by predicting y' while the truth is y.

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- A loss function ℓ
  $\rightarrow$ $\ell(y' - y)$ incurred by predicting y' while the truth is y.
- the risk $R(\theta)$

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- A loss function ℓ
  $\rightarrow$ $\ell(y' - y)$ incurred by predicting y' while the truth is y.
- the risk $R(\theta)$
  $\rightarrow$ $R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X) - Y)]$. Not observable.

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- A loss function ℓ
  $\rightarrow$ $\ell(y' - y)$ incurred by predicting y' while the truth is y.
- the risk $R(\theta)$
  $\rightarrow$ $R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X) - Y)]$. Not observable.
- an empirical proxy $r(\theta)$ for $R(\theta)$

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- A loss function ℓ
  $\rightarrow$ $\ell(y' - y)$ incurred by predicting y' while the truth is y.
- the risk $R(\theta)$
  $\rightarrow$ $R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X) - Y)]$. Not observable.
- an empirical proxy $r(\theta)$ for $R(\theta)$
  $\rightarrow$ for example $r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i) - Y_i)$.
- empirical risk minimizer $\hat{\theta}$

Generic machine learning problem

Main ingredients :

- observations : (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)
  $\rightarrow$ usually i.i.d from an unknown distribution P...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- A loss function ℓ
  $\rightarrow$ $\ell(y' - y)$ incurred by predicting y' while the truth is y.
- the risk $R(\theta)$
  $\rightarrow$ $R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X) - Y)]$. Not observable.
- an empirical proxy $r(\theta)$ for $R(\theta)$
  $\rightarrow$ for example $r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i) - Y_i)$.
- empirical risk minimizer $\hat{\theta}$
  $\rightarrow$ $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\ r(\theta)$.

Sub-gamma random variables

Definition

T is said to be sub-gamma iff $\exists (v,w)$ such that $\forall k \geq 2$,

$$\mathbb{E}\left(|T|^k\right) \leq \frac{k! v w^{k-2}}{2}.$$

Sub-gamma random variables

Definition

T is said to be sub-gamma iff $\exists(v, w)$ such that $\forall k \geq 2$,

$$\mathbb{E}\left(|T|^k\right) \leq \frac{k! v w^{k-2}}{2}.$$

Examples :

- $T \sim \Gamma(a, b)$, holds with $(v, w) = (ab^2, b)$.

Sub-gamma random variables

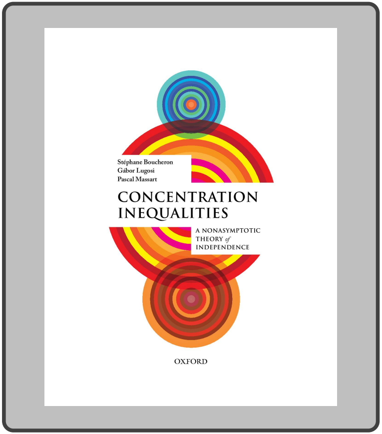
Definition

T is said to be sub-gamma iff $\exists (v,w)$ such that $\forall k \geq 2$,

$$\mathbb{E}\left(|T|^k\right) \leq \frac{k! v w^{k-2}}{2}.$$

Examples :

- $T \sim \Gamma(a,b)$, holds with $(v,w) = (ab^2, b)$.
- any Z with $\mathbb{P}(|Z| \geq t) \leq \mathbb{P}(|T| \geq t)$.



Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v, w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) \leq \exp\left(\frac{n v \zeta^2}{2(1 - w\zeta)}\right).$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v, w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) \leq \exp\left(\frac{nv\zeta^2}{2(1 - w\zeta)}\right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v, w)-sub-gamma, then

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v, w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) \leq \exp\left(\frac{nv\zeta^2}{2(1 - w\zeta)}\right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v, w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\Big[R(\theta) - r(\theta) > t\Big] = \mathbb{P}\Big\{\exp\left[s\left(R(\theta) - r(\theta)\right)\right] > \exp(st)\Big\}$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v, w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) \leq \exp\left(\frac{nv\zeta^2}{2(1 - w\zeta)}\right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v, w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\Big[R(\theta) - r(\theta) > t\Big] \leq \mathbb{E} \exp\Big[s\left(R(\theta) - r(\theta)\right) - st\Big]$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v, w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) \leq \exp\left(\frac{n v \zeta^2}{2(1 - w\zeta)}\right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v, w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\Big[R(\theta) - r(\theta) > t\Big] \leq \mathbb{E} \exp\left[\frac{s}{n}\left(\sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) - st\right]$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v,w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E}\, T_i] \right) \leq \exp\left(\frac{nv\zeta^2}{2(1 - w\zeta)} \right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v,w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\Big[R(\theta) - r(\theta) > t \Big] \leq \mathbb{E} \exp\left[\frac{s}{n} \left(\textstyle\sum_{i=1}^{n} [T_i - \mathbb{E}\, T_i] \right) - st \right]$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v,w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E}\, T_i] \right) \leq \exp\left(\frac{nv\zeta^2}{2(1-w\zeta)} \right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v,w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\Big[R(\theta) - r(\theta) > t \Big] \leq \exp\left[\frac{vs^2}{2(n - ws)} - st \right]$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v, w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E} T_i]\right) \leq \exp\left(\frac{nv\zeta^2}{2(1 - w\zeta)}\right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v, w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\Big[R(\theta) - r(\theta) > t\Big] \leq \exp\left[\frac{vs^2}{2(n - ws)} - st\right]$$

Bernstein's inequality

Theorem

Let $T_1, \dots, T_n$ be i.i.d and (v,w)-sub-gamma random variables. Then, $\forall \zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E}\, T_i]\right) \leq \exp\left(\frac{nv\zeta^2}{2(1-w\zeta)}\right).$$

Consequence in ML : put $T_i = -\ell(f_\theta(X_i) - Y_i)$ and assume T_i is (v,w)-sub-gamma, then for any $s > 0$,

$$\mathbb{P}\left[|R(\theta) - r(\theta)| > t\right] \leq 2 \exp\left[\frac{vs^2}{2(n - ws)} - st\right]$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta, \, |R(\theta) - r(\theta)| > t\right] \;=\; \mathbb{P}\left[\bigcup_{\theta\in\Theta}\left\{|R(\theta) - r(\theta)| \geq t\right\}\right]$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq \sum_{\theta\in\Theta} \mathbb{P}\left[|R(\theta) - r(\theta)| \geq t\right]$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta, \ |R(\theta) - r(\theta)| > t\right] \quad \leq \quad 2|\Theta| \exp\left(\frac{vs^2}{2(n - ws)} - st\right)$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \ \leq \ 2|\Theta| \exp\left(\frac{vs^2}{2(n - ws)} - st\right)$$

On the complement,

$$R(\hat{\theta}) \ \leq \ r(\hat{\theta}) + t$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq 2|\Theta| \exp\left(\frac{vs^2}{2(n-ws)} - st\right)$$

On the complement, $\forall\theta_0$,

$$R(\hat{\theta}) \leq r(\hat{\theta}) + t$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq 2|\Theta| \exp\left(\frac{vs^2}{2(n - ws)} - st\right)$$

On the complement, $\forall\theta_0$,

$$R(\hat{\theta}) \leq r(\theta_0) + t$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq 2|\Theta| \exp\left(\frac{vs^2}{2(n-ws)} - st\right)$$

On the complement, $\forall\theta_0$,

$$R(\hat{\theta}) \leq R(\theta_0) + 2t$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq 2|\Theta| \exp\left(\frac{vs^2}{2(n - ws)} - st\right)$$

On the complement,

$$R(\hat{\theta}) \leq \min_{\theta\in\Theta} R(\theta) + 2t$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq \underbrace{2|\Theta| \exp\left(\frac{vs^2}{2(n-ws)} - st\right)}_{=\alpha}$$

On the complement,

$$R(\hat{\theta}) \leq \min_{\theta\in\Theta} R(\Theta) + \frac{vs}{n-ws} + \frac{2\log\frac{2|\Theta|}{\alpha}}{s}$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq \underbrace{2|\Theta| \exp\left(\frac{vs^2}{2(n-ws)} - st\right)}_{=\alpha}$$

On the complement, for
$s = [n/(2w)] \wedge \sqrt{(n/v)\log(2|\Theta|/\alpha)}$,

$$R(\hat{\theta}) \leq \min_{\theta\in\Theta} R(\Theta) + \frac{vs}{n - ws} + \frac{2\log\frac{2|\Theta|}{\alpha}}{s}$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists\theta \in \Theta,\ |R(\theta) - r(\theta)| > t\right] \leq \underbrace{2|\Theta| \exp\left(\frac{vs^2}{2(n-ws)} - st\right)}_{=\alpha}$$

On the complement,

we obtain :

Theorem

With probability at least $1-\alpha$,

$$R(\hat{\theta}) \leq \min_{\theta\in\Theta} R(\theta) + 2\sqrt{\frac{v \log\frac{2|\Theta|}{\alpha}}{n}} \vee \frac{2w \log\frac{2|\Theta|}{\alpha}}{n}.$$

Finite set of predictors and union bound

$$\mathbb{P}\left[\exists \theta \in \Theta, \, |R(\theta) - r(\theta)| > t\right] \leq \underbrace{2|\Theta| \exp\left(\frac{vs^2}{2(n-ws)} - st\right)}_{=\alpha}$$

On the complement,

we obtain :

Theorem

With proba. at least $1-\alpha$, with α not rediculously small,

$$R(\hat{\theta}) \leq \min_{\theta\in\Theta} R(\theta) + 2\sqrt{\frac{v \log \frac{2|\Theta|}{\alpha}}{n}}$$

Infinite parameter set

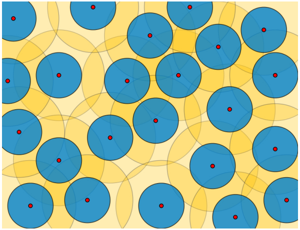
Θ compact $\Rightarrow \exists$ a finite $\Theta^{(\varepsilon)}$ such that

$$\forall \theta \in \Theta,\ \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$

Infinite parameter set (image from Wikipedia)

Θ compact $\Rightarrow \exists$ a finite $\Theta^{(\varepsilon)}$ such that

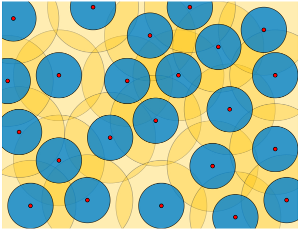
$$\forall \theta \in \Theta,\ \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$



Infinite parameter set

Θ compact $\Rightarrow \exists$ a finite $\Theta^{(\varepsilon)}$ such that

$$\forall \theta \in \Theta, \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$



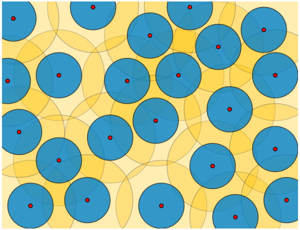
Assume $\theta \mapsto \ell(f_\theta(X) - Y)$ is a.s. L-Lipschitz w.r.t $\delta(\cdot, \cdot)$, then

$$R(\hat{\theta}) \leq \min_{\theta \in \Theta} R(\theta) + 2L\varepsilon + 2\sqrt{\frac{v \log \frac{2|\Theta^{(\varepsilon)}|}{\alpha}}{n}}$$

Infinite parameter set

Θ compact $\Rightarrow$ $\exists$ a finite $\Theta^{(\varepsilon)}$ such that

$$\forall \theta \in \Theta,\ \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$



Example : in the finite dimensional case, there is Θ_ε with

$$|\Theta^{(\varepsilon)}| \lesssim \frac{1}{\varepsilon^d}.$$

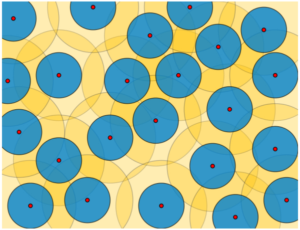
Assume $\theta \mapsto \ell(f_\theta(X) - Y)$ is a.s. L-Lipschitz w.r.t $\delta(\cdot, \cdot)$, then

$$R(\hat{\theta}) \leq \min_{\theta \in \Theta} R(\theta) + 2L\varepsilon + 2\sqrt{\frac{vd \log \frac{2}{\varepsilon \alpha}}{n}}$$

Infinite parameter set

Θ compact $\Rightarrow \exists$ a finite $\Theta^{(\varepsilon)}$ such that

$$\forall \theta \in \Theta, \ \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$



Example : in the finite dimensional case, there is Θ_ε with

$$|\Theta^{(\varepsilon)}| \lesssim \frac{1}{\varepsilon^d}.$$

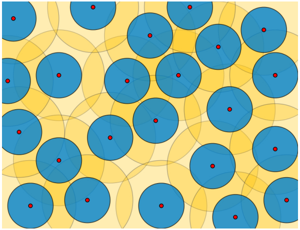
Assume $\theta \mapsto \ell(f_\theta(X) - Y)$ is a.s. L-Lipschitz w.r.t $\delta(\cdot, \cdot)$, then

$$R(\hat{\theta}) \leq \min_{\theta \in \Theta} R(\theta) + 2L\varepsilon + 2\sqrt{\frac{vd \log \frac{2}{\varepsilon \alpha}}{n}}, \ \varepsilon = \sqrt{\frac{d}{n}}$$

Infinite parameter set

Θ compact $\Rightarrow \exists$ a finite $\Theta^{(\varepsilon)}$ such that

$$\forall \theta \in \Theta,\ \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$



Example : in the finite dimensional case, there is Θ_ε with

$$|\Theta^{(\varepsilon)}| \lesssim \frac{1}{\varepsilon^d}.$$

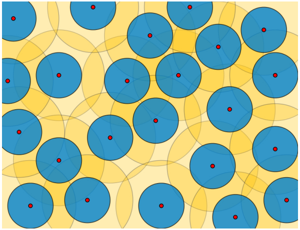
Assume $\theta \mapsto \ell(f_\theta(X) - Y)$ is a.s. L-Lipschitz w.r.t $\delta(\cdot, \cdot)$, then

$$R(\hat{\theta}) \leq \min_{\theta \in \Theta} R(\theta) + \sqrt{\frac{d}{n}} \left(2L + 2\sqrt{v \log \frac{2n}{d\alpha}} \right)$$

Infinite parameter set

Θ compact $\Rightarrow \exists$ a finite $\Theta^{(\varepsilon)}$ such that

$$\forall \theta \in \Theta,\ \exists \theta' \in \Theta^{(\varepsilon)} \text{ with } \delta(\theta, \theta') \leq \varepsilon.$$



Example : in the finite dimensional case, there is Θ_ε with

$$|\Theta^{(\varepsilon)}| \lesssim \frac{1}{\varepsilon^d}.$$

With the loss is L-Lipschitz, with proba. at least $1-\alpha$,

$$R(\hat{\theta}) \leq \min_{\theta \in \Theta} R(\theta) + \sqrt{\frac{d}{n}} \left(2L + 2\sqrt{v \log \frac{2n}{d\alpha}} \right).$$

Model selection

Now we consider $\Theta_1, \dots, \Theta_M$ with estimators $\hat{\theta}_1, \dots \hat{\theta}_M$:

Model selection

Now we consider $\Theta_1, \dots, \Theta_M$ with estimators $\hat{\theta}_1, \dots \hat{\theta}_M$:

$$\mathbb{P}\left[\exists m, \exists \theta \in \Theta_m, \ |R(\theta) - r(\theta)| > t_m\right] \leq \alpha.$$

Model selection

Now we consider $\Theta_1, \dots, \Theta_M$ with estimators $\hat{\theta}_1, \dots \hat{\theta}_M$:

$$\mathbb{P}\left[\exists m, \exists \theta \in \Theta_m, \, |R(\theta) - r(\theta)| > t_m\right] \leq \alpha.$$

Define $\hat{m} = \arg\min_m [r(\hat{\theta}_m) + t_m]$, similar derivations lead to

Model selection

Now we consider $\Theta_1, \dots, \Theta_M$ with estimators $\hat{\theta}_1, \dots \hat{\theta}_M$:

$$\mathbb{P}\left[\exists m, \exists \theta \in \Theta_m, \ |R(\theta) - r(\theta)| > t_m\right] \leq \alpha.$$

Define $\hat{m} = \operatorname*{argmin}_m [r(\hat{\theta}_m) + t_m]$, similar derivations lead to

With proba. at least $1 - \alpha$,

$$R(\hat{\theta}) \leq \min_{1 \leq m \leq M} \left\{ \min_{\theta \in \Theta_m} R(\theta) + \sqrt{\frac{d_m}{n}} \left(2L + 2\sqrt{v \log \frac{2nM}{d\alpha}} \right) \right\}.$$

Going further

Improvements, extensions...

- removing $\log(n)$ by refinment of the ε-net structure.

Going further

Improvements, extensions...

- removing $\log(n)$ by refinment of the ε-net structure.
- faster rates : $\sqrt{d/n}$ becomes d/n thanks to a better analysis of v under the **Bernstein condition**.

Going further

Improvements, extensions...

- removing $\log(n)$ by refinment of the ε-net structure.
- faster rates : $\sqrt{d/n}$ becomes d/n thanks to a better analysis of v under the **Bernstein condition**.
- relaxing the sub-gamma assumption.

Going further

Improvements, extensions...

- removing $\log(n)$ by refinment of the ε-net structure.
- faster rates : $\sqrt{d/n}$ becomes d/n thanks to a better analysis of v under the **Bernstein condition**.
- relaxing the sub-gamma assumption.
- more flexible way to measure the complexity of Θ : **PAC-Bayesian bounds**.

Extension to time series

Machine learning studied for time series with various techniqes. Asymptotic study in the mixing case :

I. Steinwart, D. Hush, C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 2009.

Extension to time series

Machine learning studied for time series with various techniqes. Asymptotic study in the mixing case :

I. Steinwart, D. Hush, C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 2009.

In order to extend the previous (non-asymptotic) approach to non-independent observations, exponential inequalities (Hoeffding, Bernstein, etc.) required.

Extension to time series

Machine learning studied for time series with various techniqes. Asymptotic study in the mixing case :

I. Steinwart, D. Hush, C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 2009.

In order to extend the previous (non-asymptotic) approach to non-independent observations, exponential inequalities (Hoeffding, Bernstein, etc.) required. These inequalities require some assumption on the dependence of the series : Markov, mixing, weak dependence, martingale,...

An example on Markov chains

An example on Markov chains

ELSEVIER

Stochastic Processes and their Applications

Volume 125, Issue 1, January 2015, Pages 60-90

Deviation inequalities for separately Lipschitz functionals of iterated random functions

Jérôme Dedecker [a], Xiequan Fan [b]

An example on Markov chains

ELSEVIER
Stochastic Processes and their Applications
Volume 125, Issue 1, January 2015, Pages 60-90

Deviation inequalities for separately Lipschitz functionals of iterated random functions

Jérôme Dedecker [a], Xiequan Fan [b]

Let $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and$(X_t)_{t\geq 1}$ be the Markov chain

$$X_t = F(X_{t-1}, \varepsilon_t).$$

An example on Markov chains

ELSEVIER

Stochastic Processes and their Applications
Volume 125, Issue 1, January 2015, Pages 60-90

Deviation inequalities for separately Lipschitz functionals of iterated random functions

Jérôme Dedecker [a], Xiequan Fan [b]

Let $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and$(X_t)_{t\geq 1}$ be the Markov chain

$$X_t = F(X_{t-1}, \varepsilon_t).$$

Assume, for $\rho \in [0,1)$ and $C > 0$,

$$\mathbb{E}\left[d\left(F(x,\varepsilon_1), F(x',\varepsilon_1)\right)\right] \leq \rho d(x,x')$$
$$d(F(x,y), F(x,y')) \leq C\delta(y,y').$$

Objective

$$X_t = F(X_{t-1}, \varepsilon_t).$$

Objective

$$X_t = F(X_{t-1}, \varepsilon_t).$$

In this case, we study one step ahead prediction :

$$r(\theta) = \frac{1}{n-1} \sum_{i=2}^{n} \ell(f_\theta(X_{i-1}) - X_i),$$
$$R(\theta) = \mathbb{E}[\ell(f_\theta(X_1) - X_2)].$$

Objective

$$X_t = F(X_{t-1}, \varepsilon_t).$$

In this case, we study one step ahead prediction :

$$r(\theta) = \frac{1}{n-1} \sum_{i=2}^{n} \ell(f_\theta(X_{i-1}) - X_i),$$
$$R(\theta) = \mathbb{E}[\ell(f_\theta(X_1) - X_2)].$$

Define

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} r(\theta).$$

Notations in DF15

$$\text{Define } G_{X_1}(x) = \int d(x, x') P_{X_1}(dx'),$$

$$G_{\varepsilon}(y) = \int C\delta(y, y') P_{\varepsilon}(dy').$$

Notations in DF15

$$\text{Define } G_{X_1}(x) = \int d(x, x') P_{X_1}(dx'),$$

$$G_{\varepsilon}(y) = \int C\delta(y, y') P_{\varepsilon}(dy').$$

Assumption : for any $k \geq 2$,

$$\mathbb{E}\left[G_{X_1}(X_1)^k\right] \leq \frac{k!}{2} V_1 M^{k-2}, \text{ and}$$

$$\mathbb{E}\left[G_{\varepsilon}(\varepsilon)^k\right] \leq \frac{k!}{2} V_2 M^{k-2}.$$

Notations in DF15

$$
\text{Define } G_{X_1}(x) = \int d(x, x') P_{X_1}(dx'),
$$
$$
G_{\varepsilon}(y) = \int C\delta(y, y') P_{\varepsilon}(dy').
$$

Assumption : for any $k \geq 2$,

$$
\mathbb{E}\left[G_{X_1}(X_1)^k \right] \leq \frac{k!}{2} V_1 M^{k-2}, \text{ and}
$$
$$
\mathbb{E}\left[G_{\varepsilon}(\varepsilon)^k \right] \leq \frac{k!}{2} V_2 M^{k-2}.
$$

$$
\text{Define } \mathcal{V} = \frac{V_1 + V_2}{1 - \rho^2}, \; \delta = \frac{1-\rho}{M}.
$$

Dedecker and Fan's inequality

Theorem (Dedecker & Fan 2015)

Consider a separately Lipschitz function $f : \mathcal{X}^n \to \mathbb{R}$:

$$|f(x_1,\dots,x_n) - f(x'_1,\dots,x'_n)| \leq \sum_{t=1}^{n} d(x_t,x'_t).$$

Then, for any $s \in [0,\delta^{-1})$,

$$\mathbb{E}\left[e^{\pm s\{f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)]\}}\right] \leq \exp\left(\frac{(n-1)s^2\mathcal{V}}{2(1-s\,\delta)}\right).$$

Consequences of DF15 for prediction

$$\mathbb{E}\left[e^{\pm s\{f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)]\}}\right] \leq \exp\left(\frac{(n-1)s^2\mathcal{V}}{2(1-s\,\delta)}\right).$$

Consequences of DF15 for prediction

$$\mathbb{E}\left[e^{\pm s\{f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)]\}}\right] \leq \exp\left(\frac{(n-1)s^2\mathcal{V}}{2(1-s\,\delta)}\right).$$

$$\text{Take } f(X_1,\dots,X_n) = \frac{1}{L}\sum_{i=2}^{n} \ell(f_\theta(X_{i-1}) - X_i).$$

Consequences of DF15 for prediction

$$\mathbb{E}\left[e^{\pm s\{f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)]\}}\right] \leq \exp\left(\frac{(n-1)s^2\mathcal{V}}{2(1-s\,\delta)}\right).$$

$$\text{Take } f(X_1,\dots,X_n) = \frac{1}{L}\sum_{i=2}^{n} \ell(f_\theta(X_{i-1})-X_i).$$

Then for any $0 \leq s < (n-1)/(L(1+\rho)\delta)$,

$$\mathbb{P}\Big[|R(\theta)-r(\theta)| > t\Big] \leq 2\exp\left(\frac{s^2(1+\rho)^2L^2\mathcal{V}}{2(n-1)-2s(1+\rho)\delta L} - st\right).$$

Learning theorem for Markov chains

Assume $|\Theta^{(\varepsilon)}| \leq \varepsilon^{-d}$.

Theorem

As soon as $n \geq 1 + 4\delta^2 d \log(Ln)/\mathcal{V}$ we have, with probability at least $1-\alpha$,

$$R(\hat{\theta}) \leq \inf_{\theta\in\Theta} R(\theta) + C_1\sqrt{\frac{d\log(Ln)}{n-1}} + C_2\frac{\log\left(\frac{4}{\alpha}\right)}{\sqrt{n-1}} + \frac{C_3}{n},$$

where $C_1 = 4(1+\rho)L\sqrt{\mathcal{V}}$, $C_2 = 2(1+\rho)L\sqrt{\mathcal{V}} + 2\delta$ and $C_3 = 3[G_\varepsilon(0) + G_{X_1}(0)]/(1-\rho) + \mathcal{V}/(2\delta)$.

Other works on ML & TS

Study of $X_t = F(\varepsilon_t; X_{t-1}, X_{t-1}, \dots)$ in

P. Alquier, O. Wintengerger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 2012.

based on Rio's version of Hoeffding's inequality

E. Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *CRAS*, 2000.

Rates in $\sqrt{\frac{d}{n}}$.

Fast rates

Rates in $\dfrac{d}{n}$ for quadratic ℓ in

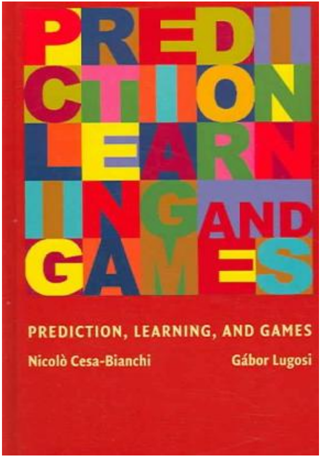
P. Alquier, X. Li, O. Wintengerger. Prediction of time series by statistical learning : general losses and fast rates.. *Dependence Modeling*, 2013.

based on Samson's version of Bernstein's inequality for φ-mixing processes

P.-M. Samson. Concentration of measure inequalities for markov chains and φ-mixing processes. *The Annals of Probability*, 2000.

Online prediction approach

The online prediction approach provides tools to aggregate predictors without stochastic assumptions on the data.

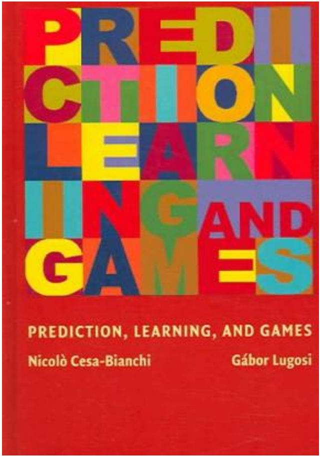


Online prediction approach

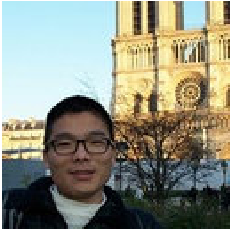
The online prediction approach provides tools to aggregate predictors without stochastic assumptions on the data.

C. Giraud, F. Roueff, A. Sanchez-Perez. Aggregation of predictors for nonstationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes. *The Annals of Statistics*, 2015.

takes advantage of this approach to predict time varying AR processes.



Nonstationary Markov chains



P. Alquier, P. Doukhan, X. Fan.
Exponential inequalities for nonstationary Markov Chains.
Preprint arxiv :1808.08811, 2018.

Non-stationary Markov chains

We now assume

$$X_t = F_t(X_{t-1}, \varepsilon_t),$$

Non-stationary Markov chains

We now assume

$$X_t = F_t(X_{t-1}, \varepsilon_t),$$

$$\sup_t \mathbb{E}\left[d\left(F_t(x, \varepsilon_1), F_t(x', \varepsilon_1)\right)\right] \leq \rho d(x, x')$$

$$\sup_t d(F_t(x, y), F_t(x, y')) \leq C\delta(y, y').$$

Non-stationary Markov chains

We now assume

$$X_t = F_t(X_{t-1}, \varepsilon_t),$$

$$\sup_t \mathbb{E}\left[d\left(F_t(x,\varepsilon_1), F_t(x',\varepsilon_1)\right)\right] \leq \rho d(x,x')$$
$$\sup_t d(F_t(x,y), F_t(x,y')) \leq C\delta(y,y').$$

Example 1 : time-varying AR(1)

$$X_t = a_t X_{t-1} + \varepsilon_t, \ \sup_t |a_t| \leq \rho.$$

Non-stationary Markov chains

We now assume

$$X_t = F_t(X_{t-1}, \varepsilon_t),$$

$$\sup_t \mathbb{E}\left[d\left(F_t(x,\varepsilon_1), F_t(x',\varepsilon_1)\right)\right] \leq \rho d(x,x')$$
$$\sup_t d(F_t(x,y), F_t(x,y')) \leq C\delta(y,y').$$

Example 1 : time-varying AR(1)

$$X_t = a_t X_{t-1} + \varepsilon_t, \ \sup_t |a_t| \leq \rho.$$

Example 2 : T-periodic AR(1)

$$X_t = a_{t[T]} X_{t-1} + \varepsilon_t, \ \max_{1 \leq t \leq T} |a_t| \leq \rho.$$

Example : T-periodic AR(1)

4-periodic AR(1), $(a_1, a_2, a_3, a_4) = (0.8, 0.5, 0.9, -0.7)$.

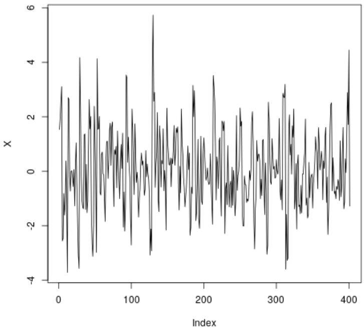


Figure – Simulated data.

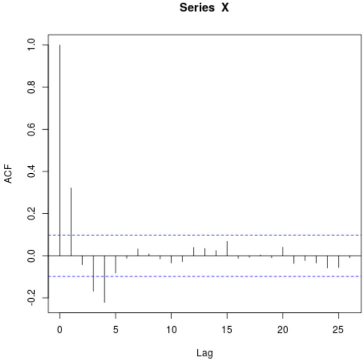


Figure – Autocorrelations.

Bernstein's inequality

Theorem (ADF18)

Assume that, for any $k \geq 2$,

$$\mathbb{E}\left[G_{X_1}(X_1)^k\right] \leq \frac{k!}{2} V_1 M^{k-2}, \text{ and}$$
$$\mathbb{E}\left[G_{\varepsilon}(\varepsilon)^k\right] \leq \frac{k!}{2} V_2 M^{k-2}.$$

Consider a separately Lipschitz function $f : \mathcal{X}^n \to \mathbb{R}$. For any $s \in [0, \delta^{-1})$,

$$\mathbb{E}\left[e^{\pm s\{f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)]\}}\right] \leq \exp\left(\frac{(n-1)s^2 \mathcal{V}}{2(1-s\,\delta)}\right).$$

Problem : estimation of the (best) period

From now, assume that

$$X_t = f_t^*(X_{t-1}) + \varepsilon_t$$

(not necessarily periodic, but we hope so).

Problem : estimation of the (best) period

From now, assume that

$$X_t = f_t^*(X_{t-1}) + \varepsilon_t$$

(not necessarily periodic, but we hope so).

Let $(f_\theta, \theta \in \Theta)$ be a set of predictors $\mathcal{X} \to \mathcal{X}$, define $\mathcal{H}(\varepsilon)$ as $\log |\Theta^{(\varepsilon)}|$. Put, for any T and $\theta_{1:T} = (\theta_1, \dots, \theta_T) \in \Theta^T$:

$$r_n(\theta_{1:T}) = \frac{1}{n-1} \sum_{i=2}^{n} \ell(f_{\theta_{i[T]}}(X_{i-1}) - X_i)$$

$$R(\theta_{1:T}) = \mathbb{E}[r_n(\theta_{1:T})]$$

Problem : estimation of the (best) period

From now, assume that

$$X_t = f_t^*(X_{t-1}) + \varepsilon_t$$

(not necessarily periodic, but we hope so).

Let $(f_\theta, \theta \in \Theta)$ be a set of predictors $\mathcal{X} \to \mathcal{X}$, define $\mathcal{H}(\varepsilon)$ as $\log |\Theta^{(\varepsilon)}|$. Put, for any T and $\theta_{1:T} = (\theta_1, \dots, \theta_T) \in \Theta^T$:

$$r_n(\theta_{1:T}) = \frac{1}{n-1} \sum_{i=2}^{n} \ell(f_{\theta_{i[T]}}(X_{i-1}) - X_i)$$

$$R(\theta_{1:T}) = \mathbb{E}[r_n(\theta_{1:T})] \in \left[\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\ell(f_{\theta_{t[T]}}(X_{t-1}) - X_t)] \pm \frac{C_0 T}{n-1} \right]$$

if actually $f_t^* = f_{t[T]}^*$, where $C_0 = L(1+\rho) \left[\frac{G_\varepsilon(0)}{1-\rho} + G_{X_1}(0) \right]$.

Estimators

Estimation for a given period T :

$$\hat{\theta}_{1:T} = (\hat{\theta}_1, \ldots, \theta_T) = \underset{\theta_{1:T}=(\theta_1,\ldots,\theta_T)}{\operatorname{argmin}} r_n(\theta_{1:T}).$$

Period selection :

$$\hat{T} = \underset{1 \leq T \leq T_{\max}}{\operatorname{argmin}} \left[r_n(\hat{f}_{1:T}) + \frac{C_1}{2} \sqrt{\frac{T\mathcal{H}(\frac{1}{Ln})}{n-1}} \right]$$

where C_1 is as in the stationary case : $C_1 = 4(1+\rho)L\sqrt{\mathcal{V}}$.

Analysis of the estimators

Theorem (ADF18)

As soon as $n \geq 1 + 4\delta^2 T_{\max}\mathcal{H}(\frac{1}{Ln})/\mathcal{V}$, with probability at least $1-\alpha$,

$$R(\hat{\theta}_{1:\hat{T}}) \leq \inf_{1\leq T\leq T_{\max}} \inf_{\theta_{1:T}\in\Theta^T} \Bigg[R(\theta_{1:T}) \\ + C_1\sqrt{\frac{T\mathcal{H}(\frac{1}{Ln})}{n-1}} + C_2\frac{\log\left(\frac{4T_{\max}}{\alpha}\right)}{\sqrt{n-1}} + \frac{C_3}{n} \Bigg].$$

Analysis of the estimators

Theorem (ADF18)

As soon as $n \geq 1 + 4\delta^2 T_{\max}\mathcal{H}(\frac{1}{Ln})/\mathcal{V}$, with probability at least $1-\alpha$,

$$R(\hat{\theta}_{1:\hat{T}}) \leq \inf_{1\leq T\leq T_{\max}} \inf_{\theta_{1:T}\in\Theta^T} \Bigg[R(\theta_{1:T}) + C_1\sqrt{\frac{T\mathcal{H}(\frac{1}{Ln})}{n-1}} + C_2\frac{\log\left(\frac{4T_{\max}}{\alpha}\right)}{\sqrt{n-1}} + \frac{C_3}{n}\Bigg].$$

In practice, C_1, C_2 and C_3 are too large and ρ is not known anyway... we recommend to use the slope heuristic here.

Slope heuristic

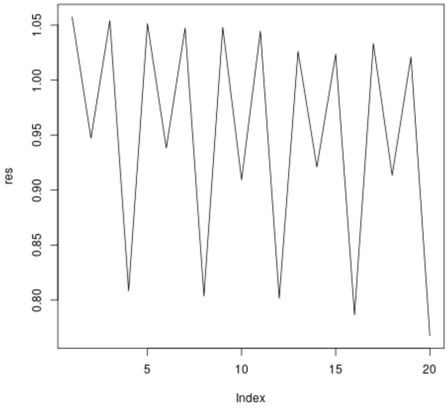


Figure – Empirical risk as a function of T.

Thank you !