



# A poor man's Wilks phenomenon

## A Wilks phenomenon in classification

S. Boucheron<sup>1</sup> and P. Massart<sup>2</sup>

<sup>1</sup>Laboratoire de Probabilités et Modèles Aléatoires  
Département de Mathématiques  
Université Paris-Diderot

<sup>2</sup>Département de Mathématiques  
Université Paris-Sud

19th of October 2007



# Motivations: Wilks phenomenon

Context: maximum likelihood estimation

- $(P_\theta, \theta \in \Theta \subseteq \mathbb{R}^m)$ : distributions over  $\mathcal{X}$
- $\forall \theta, p_\theta$  : density of  $P_\theta$  w.r.t.  $\mu$ .
- Sample  $x_1, \dots, x_n, x_i \in \mathcal{X}$ ,
- $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(x_i)$ .
- Assumption:  $\mu^{\otimes n}$ -a.s.  $\exists \hat{\theta} \in \Theta$  such that

$$\ell_n(\hat{\theta}) = \sup_{\theta \in \Theta} \sum_{i=1}^n \log p_\theta(x_i).$$

- If model "smooth enough", and  $X_1, \dots, X_n, \dots \sim_{i.i.d.} P_\theta$ , then

$$2 \left( \ell_n(\hat{\theta}) - \ell_n(\theta) \right) \rightsquigarrow \chi_m^2$$

and

$$2nD(P_\theta, P_{\hat{\theta}}) \rightsquigarrow \chi_m^2$$

- $\left( \ell_n(\hat{\theta}) - \ell_n(\theta) \right)$  excess empirical risk
- $2nD(P_\theta, P_{\hat{\theta}})$  excess risk



# Applications

- Embedded models  $\Theta$   $m$ -dimensional submodel of  $\Theta' \subseteq \mathbb{R}^{m+d}$
- If  $X_1, \dots, X_n, \dots \sim_{\text{i.i.d.}} P_\theta, \theta \in \Theta$  then

$$2 \left( \ell_n(\hat{\theta}') - \ell_n(\hat{\theta}) \right) \rightsquigarrow \chi_d^2$$

- Akaike AIC criterion for model selection [1972]
- Csiszár [IEEE IT 2002] : Markov order identification  
Consistency of penalized maximum log-likelihood BIC  
Considering a growing family of models



# Generalizations

## Possible directions

- Considering models of increasing dimensions
- Beyond likelihood ratio inference
- *Generalized likelihood ratio statistics and Wilks phenomenon* by Fan, Zhang & Zhang, AoS, 2001
  - Nonparametric Gaussian regression model where the parameter space is a Sobolev ball
  - Testing whether regression function is affine against Sobolev ball
  - Maximum likelihood estimator in Sobolev ball tends to have  $\nearrow$  dimension,
  - As  $m \nearrow (x_p^2 - \mathbb{E}[x_p^2]) / \sqrt{2\mathbb{E}[x_p^2]} \rightsquigarrow \mathcal{N}(0, 1)$
- Generalization I : When centered and scaled, the difference between the maximum log-likelihoods  $\rightsquigarrow$  non-degenerate random variable.

# Statistical learning

- $\mathcal{X} \times \mathcal{Y}$  endowed with unknown  $P$ ,
- coordinate projections :  $X$  and  $Y$ .
  - Binary classification :  $\mathcal{Y} = \{-1, 1\}$
  - Bounded regression :  $\mathcal{Y} = [-b, b]$
- Loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ 
  - Hard loss :  $\ell(f(X), Y) = \mathbf{1}_{f(X) \neq Y}$
  - Hinge loss :  $\ell(f(X), Y) = (1 - f(X)Y)_+$
- Risk of  $f \in \mathcal{Y}^{\mathcal{X}}$   $R(f) = P\ell(f(X), Y) = \mathbb{E}_P\ell(f(X), Y)$
- Assumption/notation :  $g$  minimizes  $R(f) \in \mathcal{Y}^{\mathcal{X}}$   
 Example : Bayes classifier in binary classification  
 $g(x) = 2\mathbf{1}_{\mathbb{E}[Y|X] > 0} - 1 = \text{sign}(\mathbb{E}[Y | X])$
- **Goal** : given a model  $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$  find  $\bar{g} \in \mathcal{F}$  that minimizes **risk**  $R(\cdot)$  over  $\mathcal{F}$
- **Recipes** : minimize **empirical risk**  $R_n(f) = P_n\ell(f(X), Y) = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i)$
- Assumption/notation :  $\hat{f}$  minimizes empirical risk over  $\mathcal{F}$

# Excess risks

- Model bias  $R(\bar{g}) - R(g)$
- Excess risk  $R(\hat{f}) - R(\bar{g})$
- Excess empirical risk  $R_n(\bar{g}) - R_n(\hat{f})$
- Notation:  $\bar{R}_n(f) = R_n(f) - R(f)$

$$\bar{R}_n(\bar{g}) - \bar{R}_n(\hat{f}) = \overbrace{R(\hat{f}) - R(\bar{g})}^{\text{Excess risk}} + \overbrace{R_n(\bar{g}) - R_n(\hat{f})}^{\text{Excess empirical risk}}$$

- Control of excess risk/empirical excess risk : control of increments of centered empirical process.
- If random function  $\phi_n$  satisfies

$$\forall f \in \mathcal{F} \quad |\bar{R}_n(f) - \bar{R}_n(\bar{g})| \leq \phi_n(R(f))$$

looking for largest value of  $R(f)$  that satisfies

$$R(f) - R(\bar{g}) \leq \phi_n(R(f))$$

↔ upper bound on  $R(\hat{f}) - R(\bar{g})$

Goal : controlling modulus of continuity of  $\bar{R}_n(\cdot) - \bar{R}_n(\bar{g})$

- Complexity of the  $L_2$  neighborhood of  $\bar{g}$  in  $\mathcal{F}$  :

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}, P(\ell(f(X), Y) - \ell(\bar{g}(X), Y))^2 \leq r^2} |\bar{R}_n(f) - \bar{R}_n(\bar{g})| \right] \leq \psi(r)$$

- Noise conditions

$$\sup \left\{ \left( P(\ell(f(X), Y) - \ell(g(X), Y))^2 \right)^{1/2} : R(f) - R(g) \leq r^2 \right\} \leq \omega(r).$$

- Assumptions :  $\psi$  and  $\omega$  in  $\mathcal{C}_1$

$\mathcal{C}_1$  : non-decreasing, continuous functions  $\psi$  from  $\mathbb{R}_+^{\mathbb{R}_+}$ , such that  $\psi(x)/x$  is non-increasing and  $\psi(1) \geq 1$

- $r_*$  : positive solution of  $nr^2 = \psi(\omega(r))$



- $K > 1, \delta > 0,$   
 $r(\delta)$  positive solution of

$$r^2 = K \left( 4(1 + \epsilon) \frac{\psi(2\omega(r))}{\sqrt{n}} + \frac{\omega(r)}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} + \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \frac{\log \frac{1}{\delta}}{n} \right)$$

if  $r(\delta) > r_{\text{cr}}$ , then, with probability  $> 1 - 2\delta$ ,

$$\forall f \in \mathcal{F} \quad \frac{K-1}{K} (R(f) - R(\bar{g})) + (R_n(\bar{g}) - R_n(f)) \leq \frac{1}{K} (R(\bar{g}) - R(g) + r^2(\delta)).$$



With probability larger than  $1 - 2\delta$ :



$$R(\hat{f}) - R(\bar{g}) \leq \frac{1}{K-1} \left( R(\bar{g}) - R(g) + 128K^2(1+\epsilon)^2 r_*^2 \right) + \frac{2K}{K-1} \left( Kr_*^2 + \frac{1}{n} \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \right) \log \frac{1}{\delta}$$

$$R_n(\bar{g}) - R_n(\hat{f}) \leq \frac{1}{K} \left( R(\bar{g}) - R(g) + 128K(1+\epsilon)^2 r_*^2 \right) + 2 \left( Kr_*^2 + \frac{1}{n} \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \right) \log \frac{1}{\delta}.$$

- For  $q \geq 1$

$$\|R(\hat{f}) - R(\bar{g})\|_q \leq \frac{1}{K-1} \left( R(\bar{g}) - R(g) + 128K^2(1+\epsilon)^2 r_*^2 \right) + \frac{2K}{K-1} \left( Kr_*^2 + \frac{1}{n} \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \right) (\Gamma(q+1))^{1/q}$$

$$\|R_n(\bar{g}) - R_n(\hat{f})\|_q \leq \frac{1}{K} \left( R(\bar{g}) - R(g) + 128K(1+\epsilon)^2 r_*^2 \right) + 2 \left( Kr_*^2 + \frac{1}{n} \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \right) (\Gamma(q+1))^{1/q}.$$



With probability larger than  $1 - 4\delta$ :

$$\begin{aligned} & \frac{1}{4} P \left( \hat{f} - \bar{g} \right)^2 \\ & \leq \frac{4K}{K-1} \left( \omega^2 \left( \sqrt{R(\bar{g}) - R(g)} \right) \right. \\ & \quad \left. + 32K^2(1+\epsilon)^2 \omega^2(r_*) + \left( 2K^2 \omega^2(r_*) + 2K^2 \omega^2 \left( \frac{3+4\epsilon}{12n\epsilon} \right) \right) \log \frac{1}{\delta} \right) \end{aligned}$$

$$\begin{aligned} & \frac{1}{4} P_n \left( \hat{f} - \bar{g} \right)^2 \\ & \leq \frac{K+1}{K} \left( \omega^2 \left( \sqrt{\frac{1}{K-1} (R(\bar{g}) - R(g))} \right) \right. \\ & \quad \left. + \frac{128K^2(1+\epsilon)^2}{K-1} \omega^2(r_*) \right. \\ & \quad \left. + \frac{2K}{K-1} \omega^2 \left( \sqrt{\left( Kr_*^2 + \frac{1}{n} \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \right)} \right) \log \frac{1}{\delta} \right) \\ & \quad + r^2(\delta). \end{aligned}$$

# Proofs

Massart and Nédélec, AoS, 2005

- $\mathcal{F}$  countable index set.
- $L \in \mathbb{R}_+^{\mathcal{F}}$ ,
- Assumption :  $\exists g \in \mathcal{F}$ ,  $L(g) = \inf_{f \in \mathcal{F}} L(f)$ .
- $\mathcal{B}(r) = \{f : f \in \mathcal{F}, L(f) \leq r^2\}$ .
- $Z$  : stochastic process indexed by  $\mathcal{F}$ .
- Assumption:  $\exists \psi \in \mathcal{C}_1$

$$\forall r, \quad \mathbb{E} \left[ \sup_{f \in \mathcal{B}(r)} |Z(f) - Z(g)| \right] \leq \psi(r).$$

- 

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{r^2}{r^2 + L(f)} |Z(f) - Z(g)| \right] \leq 4\psi(r).$$

- See also Giné & Koltchinskii 2006, Giné, Koltchinskii, and Wellner 2003

## Proofs (II)

### Concentration inequality for suprema of bounded centered empirical processes

Talagrand, 1996, ..., Bousquet 2002.

- $X_1, \dots, X_n \sim_{\text{i.i.d.}} X$ .
- $\sigma^2 = \sup_{h \in \mathcal{H}} \text{Var}[h(X)]$ .
- $b = \sup_h \|h(X) - \mathbb{E}[h(X)]\|_\infty$ .
- 

$$Z = \sup_{h \in \mathcal{H}} \sum_{i=1}^n (h(X_i) - \mathbb{E}[h(X)]) = n \sup_{h \in \mathcal{H}} (P_n - P)h$$

- Let  $v = 2b\mathbb{E}[Z] + n\sigma^2$ .
- 

$$\mathbb{P} \left\{ Z \geq \mathbb{E}[Z] + \sqrt{2v \log \frac{1}{\delta}} + \frac{b}{3} \log \frac{1}{\delta} \right\} \leq \delta.$$

## Proofs (III)

- Let  $\mathcal{H}$  denote a countable class of measurable functions over  $\mathcal{X}$ .
- Assumptions
  - $\sup_{x \in \mathcal{X}} |h(x) - g(x)| \leq 1$ .
  - $L$  a non-negative function on  $\mathcal{H}$ , which achieves its minimum at  $g \in \mathcal{H}$ .
  - Let  $\mathcal{B}(r) = \{h : h \in \mathcal{H}, L(h) \leq r^2\}$ .
  - Let  $\rho$  be some non-negative mapping on  $\mathbb{R}_+$  such that for every  $h \in \mathcal{B}(r)$ :

$$P(h(X) - g(X))^2 \leq \rho^2(r).$$

- $\exists \psi \in \mathcal{C}_1$

$$\sqrt{n} \mathbb{E} \left[ \sup_{h \in \mathcal{H}, \mathbb{E}[(h-g)^2] \leq r^2} (P_n - P)(h - g) \right] \leq \psi(r).$$

- Let  $K$  denote a number larger than 1. Let  $\delta > 0$  be a positive number.
- $r(\delta)$  : positive solution of

$$r^2 = K \left( 4(1 + \epsilon) \frac{\psi(\rho(r))}{\sqrt{n}} + \rho(r) \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \frac{\log \frac{1}{\delta}}{n} \right),$$

- with probability larger than  $1 - 2\delta$ , for all  $h \in \mathcal{H}$

$$|(P - P_n)(h - g)| \leq \frac{L(h) + r^2(\delta)}{K}.$$

# Efron-Stein estimates of variance

Efron & Stein AoS 1981, Steele, AoS, 1986

- $Z = h(X_1, X_2, \dots, X_n)$ , (independent R.V)
- Let  $X'_1, \dots, X'_n \sim X_1, \dots, X_n$  and independent from  $X_1, \dots, X_n$ .
- For each  $i \in \{1, \dots, n\}$ 
  - $Z'_i = h(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$ .
  - $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ .
  - $h_i$ : a function of  $n - 1$  arguments
  - $Z_i = h_i(X_1, \dots, X_{i-1}, X_{i+1}, X_n) = h_i(X^{(i)})$ .
- Jackknife estimates of variance:

$$V_+ = \sum_{i=1}^n \mathbb{E} \left[ (Z - Z'_i)_+^2 \mid X_1, \dots, X_n \right]$$

and

$$V = \sum_i (Z - Z_i)^2 .$$

- Efron-Stein inequalities:

$$\text{Var}[Z] \leq \mathbb{E}[V_+] \leq \mathbb{E}[V] .$$

# General moment bounds

B., Bousquet, Lugosi & Massart, AoP, 2005

- Assuming
  - $(X_1, \dots, X_n)$  independent random variables
  - $Z = F(X_1, \dots, X_n)$
  - $X'_1, \dots, X'_n$ , independent copies of  $X_1, \dots, X_n$ .
  - $Z'_i = F(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$ .
  - $V_+ = \sum_{i=1}^n \mathbb{E}' \left[ (Z - Z'_i)_+^2 \right]$ .
- for any  $q \geq 2$ :

$$\|(Z - \mathbb{E}[Z])_+\|_q \leq \sqrt{3q \|V_+\|_{q/2}} = \sqrt{3q} \|\sqrt{V_+}\|_q.$$

- Assuming  $\exists M$  a random variable satisfying  $(Z - Z'_i)_+ \leq M$  for all  $i \leq n$ ,
- for all  $q \geq 2$

$$\|(Z - \mathbb{E}[Z])_-\|_q \leq \sqrt{5q} \left( \|\sqrt{V_+}\|_q \vee \|M\|_q \right).$$

# Variance bounds for empirical excess risk

- Let  $\mathcal{F}$ ,  $g$ ,  $L$ ,  $\rho$ ,  $\hat{f}$  be defined as usual
- Assumption: the loss functions  $\ell(f(\cdot), \cdot)$ ,  $f \in \mathcal{F}$  are  $[0, 1]$ -valued.
- $\hat{f}_n$  minimizer of empirical risk when  $(X_n, Y_n)$  removed from sample
- 

$$\begin{aligned} \text{Var}[n(R_n(\bar{g}) - R_n(\hat{f}))] &\leq 2n \left( \mathbb{E}[R_n(\bar{g}) - R_n(\hat{f})] + \rho^2 \left( \sqrt{\mathbb{E}[L(\hat{f}_n)]} + \mathbb{E}[L(\hat{f}_n)] \right) \right) \\ &\leq 6n\rho^2(Cr_*) \end{aligned}$$

- Proof
  - $R_n(\bar{g}) - R_n(\hat{f})$  is a supremum of bounded (non-centered) empirical process
  - Refrain from using

$$\mathbb{E} \left[ (\bar{h}(X) - \hat{h}(X))^2 \mid X^{(n)} \right] \leq \sup_{h \in \mathcal{H}} \mathbb{E} \left[ (\bar{h}(X) - h(X))^2 \right]$$

- take advantage on bounds on the  $L_2$  distance between  $\hat{h}$  and  $\bar{h}$



## Variance bounds (II)

- Let  $\mathcal{F}$ ,  $g$ ,  $L$ ,  $\rho$ ,  $\hat{r}$  as usual
- Assumption : all functions in loss class are  $[0, 1]$ -valued.
- 

$$\begin{aligned}
 & \text{Var} \left[ n \left( R_n(\bar{g}) - R_n(\hat{r}) \right) \right] \\
 & \leq 2n \left( \mathbb{E} \left[ P_n(\bar{h} - \hat{h}_n)^2 \right] + \mathbb{E} \left[ P(\bar{h} - \hat{h}_n)^2 \right] \right) \\
 & \leq 2n \left( 2\mathbb{E} \left[ P(\bar{h} - \hat{h}_n)^2 \right] + (32(1 + \epsilon)^2)r_*^2 + 2 \left( \frac{\rho^2(r_*)}{r_*^2} + \left( \frac{1}{3} + \frac{1}{4\epsilon} \right) \frac{1}{n} \right) \right).
 \end{aligned}$$

## Sketch of proof

- First Efron-Stein inequality,



$$(Z - Z'_i) \leq \left( (\bar{h} - \hat{h})(X_i, Y_i) - (\bar{h} - \hat{h})(X'_i, Y'_i) \right),$$

↪

$$V_+ \leq 2n \left( P_n(\bar{h} - \hat{h})^2 + P(\bar{h} - \hat{h})^2 \right).$$

- Taking expectations over  $X_1, \dots, X_n$ :

$$\begin{aligned} \text{Var} \left[ nP_n(\bar{h} - \hat{h}) \right] &\leq \mathbb{E}[V_+] \\ &\leq 2n\mathbb{E} \left[ P_n(\bar{h} - \hat{h})^2 \right] + 2n\mathbb{E} \left[ P(\bar{h} - \hat{h})^2 \right]. \end{aligned}$$



## Main statement

Let  $Z = n \left( R_n(\bar{g}) - R_n(\hat{f}_n) \right)$ .

For some universal constants  $C$  and  $C'$ :

For  $q \geq 2$ .

$$\|Z - \mathbb{E}[Z]\|_q \leq \sqrt{n} \left[ C \sqrt{q\omega^2 \left( \sqrt{L(\bar{g})} \vee r_* \right)} + C' q\omega(r_*) \right]$$

where  $L(\bar{g}) = R(\bar{g}) - R(g)$

## Sketch of proof

- Use

$$V_+ \leq 2n \left( P_n(\bar{h} - \hat{h})^2 + P(\bar{h} - \hat{h})^2 \right).$$

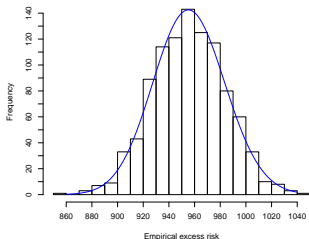
- and for any  $q \geq 2$ :

$$\|(Z - \mathbb{E}[Z])_+\|_q \leq \sqrt{3q \|V_+\|_{q/2}} = \sqrt{3q} \left\| \sqrt{V_+} \right\|_q.$$

## VC classes under gentle noise

- Classification. Hard loss.  $\ell(y, y') = \mathbf{1}_{y \neq y'}$
- VC classes with dimension  $V$
- Random classification noise  $|\mathbb{E}[Y | X]| = h$
- $\omega(r) = \sqrt{1/hr}$
- $\psi(r) = Cr\sqrt{V(1 + \log(1 \vee r^{-1}))}$
- $\hookrightarrow r_*^2 \leq C^2 \left( \left( \frac{V(1 + \log(nh^2/V))}{nh} \right) \wedge \frac{V}{n} \right)$
- $\omega^2(r_*) \leq C^2 \left( \left( \frac{V(1 + \log(nh^2/V))}{n} \right) \wedge \frac{Vh}{n} \right)$
- $\hookrightarrow$  in range  $V \leq nh^2$   
Variance of excess empirical risk mostly depends on  $V$

# Learning VC classes under random classification noise



- Classification problem with efficient ERM algorithm (Kearns, Mansour, Ng & Ron, Machine Learning, 1997)
- Classification loss
- VC-dimension of  $\mathcal{F}$ : 1600
- $\psi(r) =$
- $R(g) = .2$
- $\omega(r) = \frac{r}{\sqrt{\text{gap}}}$
- $n = 20000$
- 1000 trials,  $\text{gap} = .3$ ,
- Average value and median of empirical excess risk = 956.
- Sample variance : 784.
- Blue line : Gamma(1165, 1.21)