A poor man's Wilks phenomenon

S. Boucheron[1] and P. Massart[2]

[1]Laboratoire de Probabilités et Modèles Aléatoires
Département de Mathématiques
Université Paris-Diderot

[2]Département de Mathématiques
Université Paris-Sud

3rd of July 2008

Early Motivations: Wilks phenomenon

Context: maximum likelihood estimation

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

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

Early Motivations: Wilks phenomenon

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

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

and

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

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

Early Motivations: Wilks phenomenon

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

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

and

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

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

Note more than 2mn

1. Dealing with a contrast minimization problem, statistical learning also relies on M-estimation
2. Empirical risk minimizer is well-defined
3.
4. The ingredients

Applications : model selection/identification

- Wilks phenomenon and model assessment
- Embedded models Θ m-dimensional submodel of $\Theta' \subseteq \mathbb{R}^{m+d}$
- If $X_1, \ldots, X_n, \ldots \sim_{\text{i.i.d.}} P_\theta, \theta \in \Theta$ then

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

- Akaike AIC criterion for model selection [1972]
- Other model selection criteria ...

Markov order identification [Csiszár IEEE IT 2002]

- Observations : sequences over a fixed finite alphabet $\mathcal{X}$
- Model k : Markov chains of order k (dimension $|\mathcal{X}|^k \times (|\mathcal{X}| - 1)$)
- Goal : Markov order identification
- If true model k^* and fixed $k > k^*$

$$\ell_n(\hat{\theta}_k) - \ell_n(\hat{\theta}_{k^*}) \rightsquigarrow \frac{1}{2}\chi^2_{d-d^*}$$

with $d - d^* = (|\mathcal{X}| - 1) \times (|\mathcal{X}|^k - |\mathcal{X}|^{k^*})$

- Order identification : Need uniformity over $k = O(\log n)$
- Consistency of penalized maximum log-likelihood : BIC

$$\tilde{k} = \arg\max \left\{ \ell_n(\hat{\theta}_k) - \frac{(|\mathcal{X}|-1)|\mathcal{X}|^k}{2} \log n \right\}$$

- Considering a growing family of models

Markov order identification [Csiszár IEEE IT 2002]

- Observations : sequences over a fixed finite alphabet $\mathcal{X}$
- Model k : Markov chains of order k (dimension $|\mathcal{X}|^k \times (|\mathcal{X}| - 1)$)
- Goal : Markov order identification
- If true model k^* and fixed $k > k^*$

$$\ell_n(\hat{\theta}_k) - \ell_n(\hat{\theta}_{k^*}) \rightsquigarrow \frac{1}{2}\chi^2_{d-d^*}$$

with $d - d^* = (|\mathcal{X}| - 1) \times (|\mathcal{X}|^k - |\mathcal{X}|^{k^*})$

- Order identification : Need uniformity over $k = O(\log n)$
- Consistency of penalized maximum log-likelihood : BIC

$$\tilde{k} = \arg\max \left\{ \ell_n(\hat{\theta}_k) - \frac{(|\mathcal{X}|-1)|\mathcal{X}|^k}{2} \log n \right\}$$

- Considering a growing family of models

Note more than 2mn

1. Generalized likelihood ratio in nested exponential models
2. Comparison of smooth parametric models
3. Asymptotic results
4. Seems to rely too much on likelihood inference
5. Known to fail under loss of identifiability

Generalizations: possible directions

1. Considering models of increasing dimensions
2. Beyond likelihood ratio inference
3. *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 $m \nearrow$ $\quad (\chi^2_p - \mathbb{E}[\chi^2_p])/\sqrt{2\mathbb{E}[\chi^2_p]} \rightsquigarrow \mathcal{N}(0,1)$
4. Generalization : when centered and scaled, log of ratio of maximum likelihoods $\rightsquigarrow$ non-degenerate random variable.

Generalizations: possible directions

1. Considering models of increasing dimensions
2. Beyond likelihood ratio inference
3. *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 $m \nearrow$ $(\chi^2_p - \mathbb{E}[\chi^2_p])/\sqrt{2\mathbb{E}[\chi^2_p]} \rightsquigarrow \mathcal{N}(0,1)$
4. Generalization : when centered and scaled, log of ratio of maximum likelihoods $\rightsquigarrow$ non-degenerate random variable.

1.
2.
3. This asymptotic pivotality property paves the way to non-trivial statistical applications.

Statistical learning setting

- Bounded contrats minimization
- $\mathcal{X} \times \mathcal{Y}$ endowed with unknown P,
- coordinate projections : X and Y.
  - Binary classification : $\mathcal{Y} = \{-1, 1\}$
  - Bounded regression : $\mathcal{Y} = [-b, b]$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$
  - Hard loss : $\ell(f(X), Y) = \mathbf{1}_{f(X) \neq Y}$
  - Hinge loss : $\ell(f(X), Y) = (1 - f(X)Y)_+$
- Risk of $f \in \mathcal{Y}^{\mathcal{X}}$ $R(f) = P\ell(f(X), Y) = \mathbb{E}_P \ell(f(X), Y)$

Statistical learning setting

- Assumption/notation : f^* minimizes $R(f) \in \mathcal{Y}^{\mathcal{X}}$
- Example : Bayes classifier in binary classification

$$f^*(x) = 2\mathbf{1}_{\mathbb{E}[Y|X]>0} - 1 = \operatorname{sign}(\mathbb{E}[\mathrm{Y} \mid \mathrm{X}])$$

- Goal : given a model $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$
  find $\bar{f} \in \mathcal{F}$ that minimizes risk $R(.)$ 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 $\mathcal{F}$

Statistical learning setting

- Assumption/notation : f^* minimizes $R(f) \in \mathcal{Y}^{\mathcal{X}}$
- Example : Bayes classifier in binary classification

$$f^*(x) = \mathbf{21}_{\mathbb{E}[Y|X]>0} - 1 = \mathrm{sign}(\mathbb{E}[Y \mid X])$$

- Goal : given a model $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ find $\bar{f} \in \mathcal{F}$ that minimizes risk $R(.)$ 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 $\mathcal{F}$

The settings considered in (**???**) and the references we are aware of, share a common feature: they are connected with density estimation in a Gaussian framework. They disregard robustness considerations.

This is in sharp contrast with the setting we are interested in: statistical learning (**??**).

For example, in binary classification, $\mathcal{X} \times \{-1, 1\}$ is endowed with a probability distribution P, the coordinate projections are denoted by X and Y.

The problem consists in finding a function f on $\mathcal{X}$ such that the risk $R(f) = P\{f(X) \neq Y\}$ is as small as possible starting from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ collected from an i.i.d. sample from P.

The best possible classifier f^*, the so-called Bayes classifier is defined from the regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ by $f^*(x) = \mathrm{sign}(\eta(x))$.

A learning algorithm typically looks for a good approximation $\hat{f}$ of the

Excess risks

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

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

- Control of excess risk/empirical excess risk :
  $\hookrightarrow$ control of increments of centered empirical process.

Control of increments of centered empirical process

- If random function ϕ_n satisfies

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

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

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

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

Controlling modulus of continuity of $\bar{R}_n(\cdot) - \bar{R}_n(\bar{f})$

- Loss class $\mathcal{H} = \{\ell(f(\cdot), \cdot),\ f \in \mathcal{F}\}$
  - $h(X, Y) = \ell(f(X), Y)$
  - $h^*(X, Y) = \ell(f^*(X), Y)$
  - $\bar{h}(X, Y) = \ell(\bar{f}(X), Y)$
- Complexity of the L_2 neighborhood of $\bar{h}$ in $\mathcal{H}$:

$$\sqrt{n}\mathbb{E}\left[\sup_{h\in\mathcal{F}, P(h-\bar{h})^2\le r^2} \left|(P_n - P)(h - \bar{h}\right|\right] \le \psi(r)$$

- Noise conditions

$$\sup\left\{\left(P(h - h^*)^2\right)^{1/2} : P(h - h^*) \le r^2\right\} \le \omega(r)\,.$$

- Assumptions: $\psi, \omega \nearrow$, continuous ≥ 0, $\psi(x)/x, \omega(x)/x \searrow$ and $\psi(1), \omega(1) \ge 1$

Controlling modulus of continuity of $\bar{R}_n(\cdot) - \bar{R}_n(\bar{f})$

- Loss class $\mathcal{H} = \{\ell(f(\cdot),\cdot),\ f \in \mathcal{F}\}$
  - $h(X,Y) = \ell(f(X),Y)$
  - $h^*(X,Y) = \ell(f^*(X),Y)$
  - $\bar{h}(X,Y) = \ell(\bar{f}(X),Y)$
- Complexity of the L_2 neighborhood of $\bar{h}$ in $\mathcal{H}$:

$$\sqrt{n}\mathbb{E}\left[\sup_{h\in\mathcal{F}, P(h-\bar{h})^2\le r^2} |(P_n - P)(h-\bar{h}|\right] \le \psi(r)$$

- Noise conditions

$$\sup\left\{\left(P(h-h^*)^2\right)^{1/2} : P(h-h^*) \le r^2\right\} \le \omega(r)\,.$$

- Assumptions: $\psi, \omega \nearrow$, continuous ≥ 0, $\psi(x)/x, \omega(x)/x \searrow$ and $\psi(1), \omega(1) \ge 1$

1. The connexion between the richness of $\mathcal{F}$, the closeness of the Bayes classifier g to $\mathcal{F}$, the distribution of $|\eta(\cdot)|$ (the so-called noise conditions), and the distribution of the risk $R(\hat{f}) = P\mathbf{1}_{\hat{f}\neq Y}$ and more recently the excess risk $R(\hat{f}) - R(\bar{f}) = P\left(\mathbf{1}_{\hat{f}(X)\neq Y} - \mathbf{1}_{\bar{f}(X)\neq Y}\right)$ (where $\bar{f}$ minimizes the risk in $\mathcal{F}$) has been the subject of intense research during the last thirty-five years (**???????**).
2. Thanks to ideas borrowed from robust statistics (**?**), empirical process theory (**???**) and the theory of concentration inequalities (**????**), we have a rather precise idea of the aforementioned connexions (**?**).
3. The function ψ

$$\sqrt{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}, P(f-\bar{f})^2\le r^2} \left|(P - P_n)(f-\bar{f})\right|\right] \le \psi(r)$$

   aims at describing the richness of the L_2 neighborhood of $\bar{f}$ in $\mathcal{F}$
4. while the function ω aims at describing the so-called noise conditions:

$$\sup\left\{\left(P(f-g)^2\right)^{1/2} : R(f) - R(f^*) \le r^2\right\} \le \omega(r)\,.$$

5. In general ψ and ω are sublinear (see the definition of the class $\mathcal{C}_1$ below). Defining r_* as the positive root of the equation $\sqrt{n}r^2 = \phi(\omega(r))$, we will check that a function of r_*^2 upper-bounds the expected excess risk and the expected empirical risk.
6. Moreover as a by-product of the analysis it also avers that a function of r_* upper-bounds the expected value of $P(\bar{f} - \hat{f})^2$ and the expected value of $P_n(\bar{f} - \hat{f})^2$.
7. As a by product of the proofs, we establish that the tails of the distribution of those quantities are at worst exponentials.

Benchmark: VC classes under gentle noise

- Examples: half-spaces in $\mathbb{R}^d$
- Classification. Hard loss. $\ell(y, y') = \mathbf{1}_{y \neq y'}$
- VC classes with dimension V
- Random classification noise $|\mathbb{E}[Y \mid X]| = \beta$
  $Y = \mathrm{sign}(\eta(X))$ with probability β
- $\omega(r) = \frac{r}{\sqrt{\beta}}$
  $\hookrightarrow$ If $P(h - h^*) \leq r^2$ then $P(h - h^*)^2 \leq \frac{r^2}{\beta}$
- $\psi(r) = Cr\sqrt{V(1 + \log(1 \vee r^{-1}))}$
  V : ako model dimension

Deviation for excess risk

- $(P, \ell, \mathcal{F})$ learning task
- r_*, solution of $\sqrt{n} r^2 = \psi(2\omega(r))$.
- $\exists \kappa_1, \kappa_2, \kappa_3 \geq 1$, w.p. $\geq 1 - 2\delta$:

$$\begin{aligned} & \max\left(R(\widehat{f}) - R(\bar{f}), R_n(\bar{f}) - R_n(\widehat{f})\right) \\ \leq \quad & \kappa_1 L(\bar{f}) + \kappa_2 r_*^2 + \kappa_3 r_*^2 \log\frac{1}{\delta} \\ & \max\left(\mathbb{E}[R(\widehat{f}) - R(\bar{f})], \mathbb{E}[R_n(\bar{f}) - R_n(\widehat{f})]\right) \\ \leq \quad & \kappa_1 L(\bar{f}) + (\kappa_2 + \kappa_3) r_*^2 \, . \end{aligned}$$

- Tools: Peeling (Huber, 1967), Vapnik & Chervonenkis, Talagrand's concentration inequality, Mammen & Tsybakov AoS 2000, Koltchinskii (AoS 2006), Massart et al. (Toulouse 2000 , Saint-Flour 2003, AoS 2006), Bartlett and Mendelson (PTRF 2004)...

Learning VC classes

Toy problem from Kearns et al.,
Machine Learning, 1997

- VC-dimension of $\mathcal{F}$: 1600
- $R(f^*) = .2$
- $\omega(r) = \frac{r}{\sqrt{\beta}}$
- $n = 20000$
- 1000 trials, , $\frac{\beta}{2} = .3$,
- $\mathbb{E}[R(f^*) - R(\hat{f})] \approx 956$.
- Sample variance : 784.
- Blue line : Gamma(1165, 1.21)

Objectives: poor man Wilks phenomenon

- Understand the moments of EER

$$nP_n(\bar{h} - \hat{h}) = n \sup_{h \in \mathcal{H}} P_n(\bar{h} - h)$$

- Relate bounds on expectation, variance and ...
- Hook: EER is a supremum of empirical process

$$\mathbb{E}\left[nP_n(\bar{h} - h)\right] \leq 0$$

$$\mathbb{E}\left[n \sup_{h \in \mathcal{H}} P_n(\bar{h} - h)\right] \geq 0$$

- Proving concentration inequalities for $n \sup_{h \in \mathcal{H}} P_n(\bar{h} - h)$

Efron-Stein estimates of variance

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

$$\begin{aligned} V_+ &= \sum_{i=1}^n \mathbb{E}\left[(Z - Z_i')_+^2 \mid X_1, ..., X_n\right] \\ V &= \sum_i (Z - Z_i)^2 \ . \end{aligned}$$

- Efron-Stein inequalities:

$$\mathrm{Var}[Z] \leqslant \mathbb{E}[V_+] \leqslant \mathbb{E}[V] \, .$$

Efron-Stein estimates of variance

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

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

- Efron-Stein inequalities:

$$\mathrm{Var}[Z] \leqslant \mathbb{E}[V_+] \leqslant \mathbb{E}[V] .$$

1. Let Z denote now a (square-integrable) function of a sequence of independent random variables $(X_1, X_2, \ldots, X_n)$, that is $Z = h(X_1, \ldots, X_n)$ for some function h. Let $X'_1, \ldots, X'_n$ be distributed as $X_1, \ldots, X_n$ and be independent from $X_1, \ldots, X_n$. For each i in $1, \ldots, n$, let $Z'_i = h(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots X_n)$. And for each i in $1, \ldots, n$, let $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Let h_i denote a function of $n-1$ arguments and let $Z_i = h_i(X_1, \ldots, X_{i-1}, X_{i+1}, X_n) = h_i(X^{(i)})$.
2. The jackknife estimates of variance are

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

and

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

3. Note that the last quantity is a bona fide estimator while the first one is just $X^{(i)}$-measurable. The Efron-Stein inequalities (**?**) assert that the jackknife estimates of variance are upper-bounds:

$$\mathrm{Var}[Z] \leqslant \mathbb{E}[V_+] \leqslant \mathbb{E}[V] . \tag{1}$$

Let us now recall how the Efron-Stein inequalities can be used to upper-bound suprema of bounded centered empirical processes.
We may define Z_i as

$$Z_i = \sup_{f \in \mathcal{F}} \sum_{j \leqslant n, j \neq i} f(X_j)$$

and Z'_i as

Variance bounds for empirical excess risk

- Let $\mathcal{F}, \mathcal{H}, f^*, \bar{f}, \bar{h}, L, \rho, \hat{f}, \ldots$ be defined as usual
- Assumption: the loss functions $h = \ell(f(\cdot), \cdot)$, $f \in \mathcal{F}$ are $[0,1]$-valued.
- $\hat{h}_n$ minimizer of $P_n h$
- Consequences of Efron-Stein inequalities :

$$\begin{aligned}
&\mathrm{Var}\left[nP_n(\bar{h} - \widehat{h}_n)\right] \\
&\quad \leq \quad 2n\mathbb{E}\left[\left((P_{n-1} - P)(\bar{h} - \widehat{h}_{n-1})\right)\right] + 2n\mathbb{E}\left[P\left(\bar{h} - \widehat{h}_n\right)^2\right] \\
&\qquad \textit{and} \\
&\mathrm{Var}\left[nP_n(\bar{h} - \widehat{h}_n)\right] \\
&\quad \leq \quad 2n\left(\mathbb{E}\left[P_n(\bar{h} - \widehat{h}_n)^2\right] + \mathbb{E}\left[P(\bar{h} - \widehat{h}_n)^2\right]\right) .
\end{aligned}$$

Notice ...

Bounds

$$\begin{aligned}
&\mathrm{Var}\left[nP_n(\bar{h}-\widehat{h}_n)\right] \\
&\quad\leq \quad 2n\mathbb{E}\left[\left((P_{n-1}-P)(\bar{h}-\widehat{h}_{n-1})\right)\right]+2n\mathbb{E}\left[P\left(\bar{h}-\widehat{h}_n\right)^2\right]
\end{aligned}$$

and

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

to be compared with

$$\mathrm{Var}\left[n\left(R_n(\bar{f})-R_n\left(\hat{f}\right)\right)\right]\leq n\mathbb{E}\left[\sup_{h\in\mathcal{H}}P_n(\bar{h}-h)^2\right]+n\sup_{h\in\mathcal{H}}P(\bar{h}-h)^2$$

Notice ...

Bounds

$$\begin{aligned}\mathrm{Var}\left[nP_n(\bar{h}-\hat{h}_n)\right] &\leq 2n\mathbb{E}\left[\left((P_{n-1}-P)(\bar{h}-\hat{h}_{n-1})\right)\right]+2n\mathbb{E}\left[P\left(\bar{h}-\hat{h}_n\right)^2\right]\end{aligned}$$

and

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

to be compared with

$$\mathrm{Var}\left[n\left(R_n(\bar{f})-R_n\left(\hat{f}\right)\right)\right] \leq n\mathbb{E}\left[\sup_{h\in\mathcal{H}} P_n(\bar{h}-h)^2\right]+n\sup_{h\in\mathcal{H}} P(\bar{h}-h)^2$$

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

Proof

- $R_n(\bar{f})-R_n(\hat{f})$ 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 $\bar{h}$

Sketch of proof

- First Efron-Stein inequality,
- $$(Z - Z_i') \leq \left(\left(\bar{h} - \hat{h} \right) (X_i, Y_i) - \left(\bar{h} - \hat{h} \right) (X_i', Y_i') \right) ,$$

  $\hookrightarrow$

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

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

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

Deviation for excess risk

- $(P, \ell, \mathcal{F})$ learning task
- r_*, solution of $\sqrt{n} r^2 = \psi(2\omega(r))$.
- $\exists \kappa_1, \kappa_2, \kappa_3 \geq 1$, w.p. $\geq 1 - 2\delta$:

$$\begin{aligned} & \max\left(R(\widehat{f}) - R(\bar{f}), R_n(\bar{f}) - R_n(\widehat{f})\right) \\ \leq \quad & \kappa_1 L(\bar{f}) + \kappa_2 r_*^2 + \kappa_3 r_*^2 \log\frac{1}{\delta} \\ & \max\left(\mathbb{E}[R(\widehat{f}) - R(\bar{f})], \mathbb{E}[R_n(\bar{f}) - R_n(\widehat{f})]\right) \\ \leq \quad & \kappa_1 L(\bar{f}) + (\kappa_2 + \kappa_3) r_*^2 \, . \end{aligned}$$

- Tools: Peeling (Huber, 1967), Vapnik & Chervonenkis, Talagrand's concentration inequality, Mammen & Tsybakov AoS 2000, Koltchinskii (AoS 2006), Massart et al. (Toulouse 2000 , Saint-Flour 2003, AoS 2006), Bartlett and Mendelson (PTRF 2004)...

Back to variance

- $(P, \ell, \mathcal{F})$: learning task.
- $\psi, \omega \in \mathcal{C}_1$ complexity and the noise
- Let r_* denote the positive solution of

$$\sqrt{n} r^2 = \psi(2\omega(r)) \,.$$

- $\exists \kappa_4$ such that

$$\mathrm{Var}\left[n(R_n(\bar{f}) - R_n(\widehat{f})) \right] \leq n\kappa_4 \left(\omega^2 \left(r_* \right) + \omega^2 \left(\sqrt{L(\bar{f})} \right) \right)$$

VC Classes

- VC classes under random classification noise ($L(\bar{f}) = 0$)
- $\hookrightarrow r_*^2 \leq C^2 \left(\left(\frac{V(1+\log(nh^2/V))}{n\beta} \right) \wedge \sqrt{\frac{V}{n}} \right)$
- $\omega^2(r_*) \leq C^2 \left(\left(\frac{V(1+\log(n\beta^2/V))}{n\beta^2} \right) \wedge \sqrt{\frac{V}{n\beta^2}} \right)$
-

$$\mathbb{E}\left[n(R_n(\bar{f}) - R_n(\widehat{f})) \right]$$

$$\leq \quad (\kappa_2 + \kappa_3) \left(C^2 \left(\left(\frac{V(1 + \log(n\beta^2/V))}{\beta} \right) \wedge \sqrt{nV} \right) \right)$$

$$\mathrm{Var}\left[n(R_n(\bar{f}) - R_n(\widehat{f})) \right]$$

$$\leq \quad \kappa_4 C^2 \left(\left(\frac{V(1 + \log(nh^2/V))}{\beta^2} \right) \wedge \sqrt{\frac{nV}{\beta^2}} \right)$$

Ingredients for Bernstein-like inequalities

- Z satisfies a Bernstein inequality with paramters V and c

$$\mathbb{P}\{Z \geq t\} \leq \exp\left(-\kappa \min\left(\frac{t^2}{v}, \frac{t}{c}\right)\right)$$

- Recentered $\Gamma(p, c)$ random variable satisfy Berstein inequalities
- If

$$\|Z\|_q \leq \sqrt{vq} + cq$$

for $q \geq 2$ then Z satisfies a Bernstein inequality.

General moment bounds B., Bousquet, Lugosi & Massart, AoP, 2005

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

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

- Assuming $\exists M$ r.v. with $(Z - Z'_i)_+ \leq M \; \forall i \leq n$,
- for all $q \geq 2$

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

Main statement

A Bernstein-like inequality for excess empirical risk.
Let $Z = nP_n\left(\bar{h} - \widehat{h}_n\right)$. For $q \geqslant 2$.

$$\begin{aligned} &\|Z - \mathbb{E}[Z]\|_q \\ &\quad \leq \sqrt{n\kappa_5'}\left(\omega\left(\sqrt{L(\bar{f})}\right) + \omega(r_*)\right) q^{1/2} + \sqrt{n\kappa_6'}\omega(r_*)q\,. \end{aligned}$$

Deviation inequalities for L_2 distances

- $\exists \kappa_5$ and κ_6 such that for $q \geq 2$

$$\left\| P\left(\widehat{h} - \bar{h}\right)^2 \right\|_q \vee \left\| P_n\left(\widehat{h} - \bar{h}\right)^2 \right\|_q$$
$$\leq \quad \kappa_5 \left(\omega^2 \left(\sqrt{L(\bar{f})} \right) + \omega^2 \left(r_* \right) \right) + \kappa_6 \omega^2(r_*) q \, .$$

- Argument: the same as for deriving deviation inequalities for excess risk.
- Work on $\left\{ (\bar{h} - h)^2 \ : \ h \in \mathcal{H} \right\}$
  - Risk: expectation !
  - Bounded process ...
  - $P\left((h - \bar{h})^2 \right)^2 \leq \omega^2 \left(\sqrt{L(h)} \right)$
  - Use contraction principle to get a convenient complexity function

Sketch of proof

- Back to variance bounds:
  $V_+ \leq 2n \left(P_n(\bar{h} - \widehat{h}_n)^2 + P(\bar{h} - \widehat{h}_n)^2 \right)$.
- for $q \geq 2$:

$$
\begin{aligned}
&\|(Z - \mathbb{E}[Z])_+\|_q \\
&\quad \leq \quad \sqrt{3q} \left\| \sqrt{2n \left(P_n(\bar{h} - \widehat{h}_n)^2 + P(\bar{h} - \widehat{h}_n)^2 \right)} \right\|_q \\
&\quad \leq \quad \sqrt{6nq} \left(\sqrt{\left\| P_n(\bar{h} - \widehat{h}_n)^2 \right\|_{q/2}} + \sqrt{\left\| P(\bar{h} - \widehat{h}_n)^2 \right\|_{q/2}} \right)
\end{aligned}
$$

Plugging the bounds on L_2 distances

$$
\begin{aligned}
&\|(Z - \mathbb{E}[Z])_+\|_q \\
&\quad \leq \quad 2\sqrt{6n\kappa_5} \left(\omega \left(\sqrt{L(\bar{f})} \right) + \omega(r_*) \right) q^{1/2} + 2\sqrt{3n\kappa_6}\omega(r_*)q \, .
\end{aligned}
$$

Learning VC classes

Toy problem from Kearns et al.,
Machine Learning, 1997

- VC-dimension of $\mathcal{F}$: 1600
- $R(f^*) = .2$
- $\omega(r) = \frac{r}{\sqrt{\beta}}$
- $n = 20000$
- 1000 trials, , $\frac{\beta}{2} = .3$,
- $\mathbb{E}[R(f^*) - R(\hat{f})] \approx 956$.
- Sample variance : 784.
- Blue line : Gamma(1165, 1.21)