Applications of concentration inequalities for statistical scoring and ranking problems

Nicolas Vayatis
ENS Cachan

Journées MAS 2012

-

Clermont-Ferrand, August 2012

Joint work with:

- Stéphan Clémençon (Telecom ParisTech)
- Gábor Lugosi (U. Pompeu Fabra)

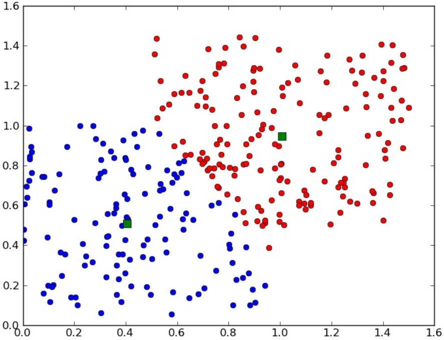
and

- Nicolas Baskiotis (UPMC),
- Marine Depecker (CEA-LIST),
- Sylvain Robbiano (Telecom ParisTech)

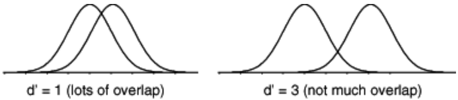
Motivations

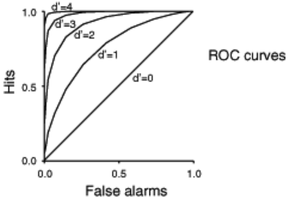
Example 1 - Bipartite ranking problem

- Observations: $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, +1\}$, $i = 1, \ldots, n$
- Infer an order over $\mathbb{R}^d$ where the $(+1)$s are above (-1) instances
- infer a scoring rule $s \; : \; \mathbb{R}^d \to \mathbb{R}$ from data with binary feedback



Example 1 (c'ed) - Evaluation metric: the ROC Curve





- ROC curve of scoring rule $s : \mathbb{R}^d \to \mathbb{R}$

$$t \in \mathbb{R} \mapsto \Big(\underbrace{P_- \{s(X) \geq t\}}_{\text{rate of false alarms}} \ , \ \underbrace{P_+ \{s(X) \geq t\}}_{\text{rate of hits}} \Big)$$

where $P_+ = \mathcal{L}(X \mid Y = +1)$ and $P_- = \mathcal{L}(X \mid Y = -1)$

Example 2 - Two-sample homogeneity test in $\mathbb{R}^d$

- $\mathcal{X}_k^+ = \{X_1^+, \ldots, X_k^+\}$ i.i.d. with distribution P_+ over $\mathbb{R}^d$
- $\mathcal{X}_m^- = \{X_1^-, \ldots, X_m^-\}$ i.i.d. with distribution P_- over $\mathbb{R}^d$
- Assume the two samples are independent
- Question: homogeneity testing with null assumption

$$\mathcal{H}_0 \ : \ P_+ = P_-$$

Example 2 (c'ed) - Connection with scoring

- Null assumption

$$\mathcal{H}_0 \ : \ P_+ = P_-$$

- Proposed strategy for $d > 1$: From multivariate homogeneity test to a collection of univariate tests
    - Consider $\mathcal{S}$ a class of scoring rules $s \ : \ \mathbb{R}^d \to \mathbb{R}$
    - Let

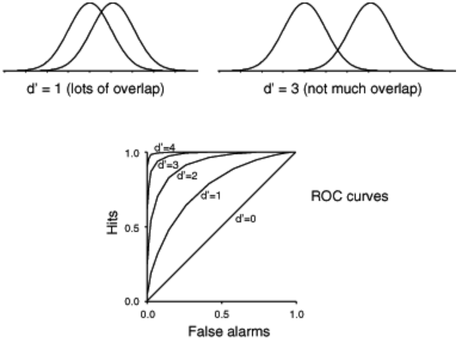
$$P_{s,+} = \mathcal{L}(s(X) \mid Y = +1) \text{ and } P_{s,-} = \mathcal{L}(s(X) \mid Y = -1)$$

    - For each $s \in \mathcal{S}$, consider homogeneity tests with null assumption

$$\mathcal{H}_{s,0} \ : \ P_{s,+} = P_{s,-}$$

    - Reject $\mathcal{H}_0$ if there exists an $s \in \mathcal{S}$ such that $\mathcal{H}_{s,0}$ is rejected
- Idea: find the most discriminative scoring rule s based on pretesting data

Example 2 (c'ed) - Test statistic based on ROC curve



- The case $P_{s,+} = P_{s,-}$ corresponds to the first diagonal ($d' = 0$)
- Use Wilcoxon rank statistics to assess discrepancy from the first diagonal

Main issues

- Optimal elements
- Variations along Example 1
  - Performance measures - summaries of ROC curves
  - Nature of feedback Y
  - Nature of sampling scheme (pointwise, pairwise, listwise)
- Empirical Risk Minimization principles and statistical theory
  - conditions for uniform convergence
  - consistency of M-estimators
  - (fast) rates of convergence?
- Design of efficient algorithms
- Meta-algorithms and aggregation principle

Optimality for bipartite ranking

ROC optimality = Neyman-Pearson theory

- ROC curve = **Power curve** of the test statistic $s(X)$ when testing

$$\mathcal{H}_0 : X \sim P_- \quad \text{against} \quad \mathcal{H}_1 : X \sim P_+$$

- Likelihood ratio $\phi(X)$ yields a **uniformly most powerful** test

$$\phi(X) = \frac{dP_+}{dP_-}(X) = \frac{1-p}{p} \times \frac{\eta(X)}{1-\eta(X)}.$$

with $p = \mathbb{P}\{Y = +1\}$, $\eta(x) = \mathbb{P}\{Y = 1+ \mid X = x\}$

- Set:

$$\mathcal{S}^* = \{T \circ \eta \mid T : [0,1] \to \mathbb{R} \;\; \text{strictly increasing}\} \,,$$

the class of ROC-optimal scoring rules

Representation of optimal scoring rules

- Note that if $U \sim \mathcal{U}([0,1])$

$$\forall x \in \mathcal{X}\ , \quad \eta(x) = \mathbb{E}\left(\mathbb{I}\{\eta(x) > U\}\right)$$

- If $s^* \in \mathcal{S}^*$, then:

$$\forall x \in \mathcal{X}\ , \quad s^*(x) = c + \mathbb{E}\left(w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right)$$

  for some:
  - $c \in \mathbb{R}$,
  - V continuous random variable in $[0,1]$
  - $w : [0,1] \to \mathbb{R}_+$ integrable.

- Optimal ranking amounts to recovering the level sets of η:

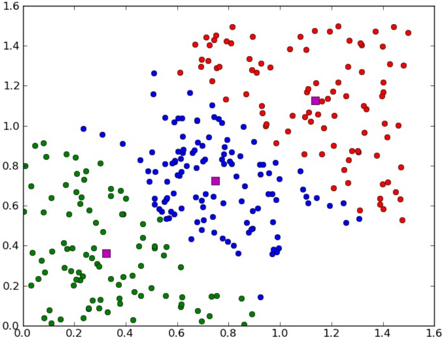
$$\{x : \eta(x) > q\}_{q \in (0,1)}$$

- Easier problem than regression function estimation!

Optimality for K-partite ranking

Ranking data with ordinal labels

- Observations: $(X_i, Y_i) \in \mathbb{R}^d \times \{1, 2, 3\}$, $i = 1, \ldots, n$



Optimal elements ($K > 2$)

- Consider feedback Y on vector X among K ordered classes
- Posterior distribution: $\forall j \in \{1, \ldots, K\}$, $\forall x \in \mathbb{R}^d$,

$$\eta_j(x) = \mathbb{P}(Y = j \mid X = x)$$

- An optimal element s^* satisfies the condition:

  $\forall l < k$, $\exists T_{l,k}$ strictly increasing such that:

$$s^* = T_{l,k} \circ \left(\frac{\eta_k}{\eta_l} \right)$$

  (optimality w.r.t. all bipartite subproblems)

- Equivalent to ROC-optimality in terms of ROC surface

Necessary and sufficient condition for optimality

- Requirement when scoring ordinal data with $K > 2$
- **Assumption.** For any $1 \leq l < k \leq K-1$, we have: for x, x',

$$\frac{\eta_{k+1}}{\eta_k}(x) < \frac{\eta_{k+1}}{\eta_k}(x') \Rightarrow \frac{\eta_{l+1}}{\eta_l}(x) < \frac{\eta_{l+1}}{\eta_l}(x')$$

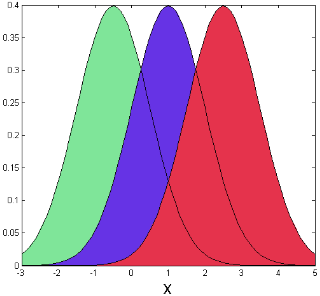
- In particular, under the assumption, the regression function

$$\eta(x) = \mathbb{E}(Y \mid X = x) = \sum_{k=1}^{K} k\eta_k(x)$$

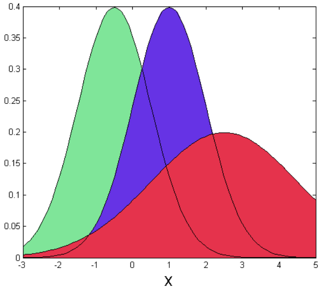
is optimal.

Example and counterexample

Here $d = 1$, $K = 3$
with GREEN = class 1 / BLUE = class 2 / RED = class 3



(a) Assumption satisfied
$m_1 < m_2 < m_3$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$

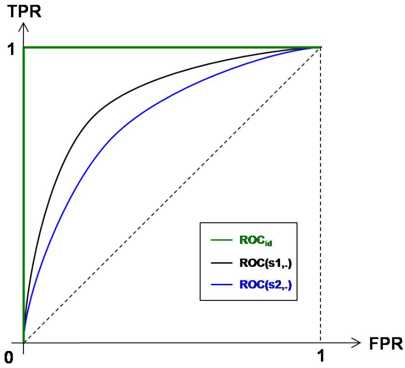


(b) Assumption not satisfied
$m_1 < m_2 < m_3$, $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_3^2 = 2$

Empirical summaries of ROC curve

Performance measures in the bipartite case ($K = 2$)

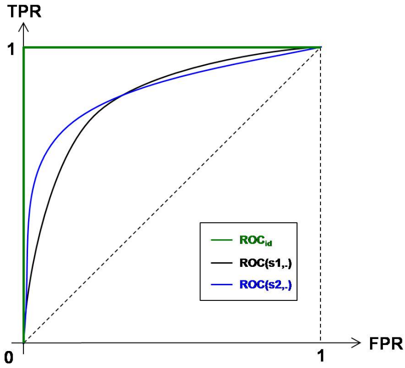
- Curves:
  - **ROC curve**
  - (Precision-Recall curve)
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



ROC curves.

Performance measures in the bipartite case ($K = 2$)

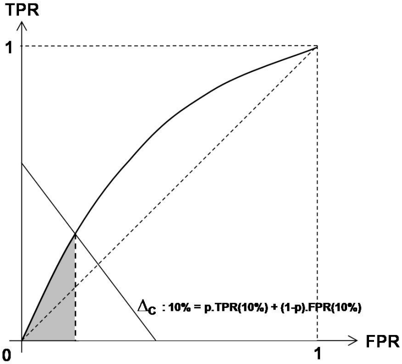
- Curves:
  - **ROC curve**
  - (Precision-Recall curve)
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



ROC curves.

Performance measures in the bipartite case ($K = 2$)

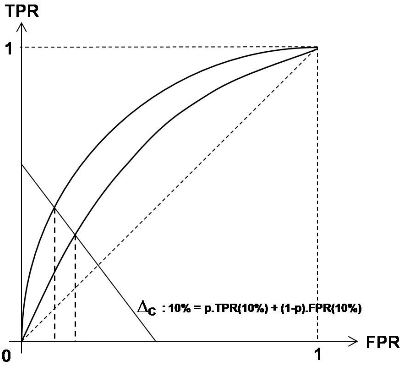
- Curves:
  - **ROC curve**
  - (Precision-Recall curve)
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



Partial AUC.

Performance measures in the bipartite case ($K = 2$)

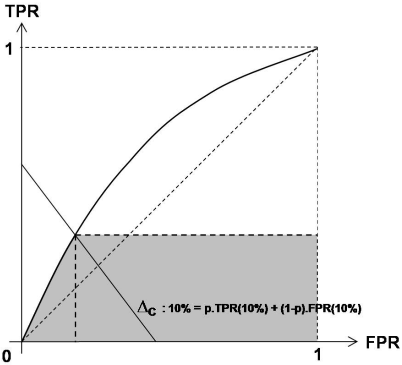
- Curves:
  - **ROC curve**
  - (Precision-Recall curve)
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



Inconsistency of Partial AUC.

Performance measures in the bipartite case ($K = 2$)

- Curves:
  - **ROC curve**
  - (Precision-Recall curve)
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



Local AUC.

Case 1 - Area Under an ROC Curve (AUC)

- For any scoring function s, define the AUC as:

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X^-) < s(X^+)\}$$

where $X^+ \sim P_+$ and $X^- \sim P_-$ are independent

- Empirical AUC $=$ U-statistic (Mann-Whitney)

$$\widehat{\mathrm{AUC}}(s) = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} \mathbb{I}\{(s(X_j^-) < s(X_i^+)\}$$

- Connection to rank statistics (Wilcoxon):

$$km\widehat{\mathrm{AUC}}(s) + k(k+1)/2 = \sum_{i=1}^{k} \mathrm{Rank}(s(X_i^+))$$

Case (2-a): Learning-to-rank criteria

- Average precision:

$$\widehat{W}(s) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n + 1 - \mathrm{Rank}(s(X_i^+))}$$

- The top-@u%

$$\widehat{W}(s) = \sum_{i=1}^{k} \mathbb{I}\{\mathrm{Rank}(s(X_i^+))/(n+1) > u\}$$

- Discounted Cumulative Gain

$$\widehat{W}(s) = \sum_{i=1}^{k} \frac{1}{\log_2(\mathrm{Rank}(s(X_i^+)) + 1)}$$

Case (2 -b): Generic setup = Linear rank statistics

- W-ranking functional

$$\widehat{W}_{k,m}(s) = \sum_{i=1}^{k} \phi\left(\frac{\mathrm{Rank}(s(X_i^+))}{k+m+1}\right) \; , \; \forall s \in \mathcal{S}.$$

- Score-generating function (SGF) $\phi : [0,1] \to [0,1]$ nondecreasing
  - $\phi(x) = x \Rightarrow$ empirical AUC
  - $\phi(x) = x \, \mathbb{I}\{x \geq 1-u\} \Rightarrow$ (empirical) Local AUC
  - $\phi(x) = x^p$, $p > 1 \Rightarrow$ p-norm push
  - $\phi(x) = c\left(\left(n+1\right)x\right) \cdot \mathbb{I}_{\{x \geq k/(n+1)\}} \Rightarrow$ DCG
  - smooth ϕ

('RankBoost' by Freund *et al.* - JMLR, 2003)(Agarwal *et al.* - JMLR, 2005) (CLV - COLT, 2005 & AoS, 2008)(CV - JMLR, 2007) (Rudin - JMLR, 2006) (Cossock and Zhang - COLT 2006)(CV - NIPS, 2008)

Rates of convergence for M-estimation
-
Main technical arguments

Baseline - ERM in statistical learning theory (1)

- ERM with target criterion $L(s) = \mathbb{E}\ell(s, Z)$ and ℓ loss function

$$\widehat{s}_n = \arg\min_{s \in \mathcal{S}} \widehat{L}_n(s) := \frac{1}{n} \sum_{i=1}^{n} \ell(s, Z_i)$$

with $Z_i = (X_i, Y_i)$ i.i.d. and $\mathcal{S}$ collection of candidate decision rules

- Second-order analysis: Talagrand's concentration inequality

$$L(\widehat{s}_n) - \inf_{s \in \mathcal{S}} L(g) \leq 2\mathbb{E}\left\{\sup_{s \in \mathcal{S}} |\widehat{L}_n(s) - L(s)|\right\} + \dots$$
$$\dots \sqrt{\frac{2\left(\sup_{s \in \mathcal{S}} \tau(s)\right) \log(1/\delta)}{n}} + c\frac{\log(1/\delta)}{n}$$

where $\tau(s)$ is a variance term, with probability at least $1 - \delta$

Baseline - ERM in statistical learning theory (2)

- Brick 1 - Complexity control, e.g. Vapnik-Chervonenkis inequality:

$$\mathbb{E}\left\{\sup_{s\in\mathcal{S}}\left|\widehat{L}_n(s)-L(s)\right|\right\}\leq c\sqrt{\frac{V}{n}}$$

  where V is the VC dimension of the class $\mathcal{S}$

- Brick 2- Variance control assumption with $\alpha\in(0,1]$, $L^*=\inf L$

$$\tau(s)\leq C\left(L(s)-L^*\right)^{\alpha}\ ,\quad \forall s$$

  $\Rightarrow$ Fast rates of convergence (Mammen-Tsybakov): excess risk in $n^{-1/(2-\alpha)}$

Additional ingredient: projection argument

- $Z_1, \ldots, Z_n$ independent random variables
- $T = T(Z_1, \ldots, Z_n)$ be a square integrable statistic
- Hájek projection

$$\widehat{T} = \sum_{i=1}^{n} \mathbb{E}[T \mid Z_i] - (n-1)\mathbb{E}(T)$$

- We have:

$$\mathbb{E}[\widehat{T}] = \mathbb{E}[T]$$

  and

$$\mathbb{E}[(\widehat{T} - T)^2] = \mathbb{E}[(T - \mathbb{E}[T])^2] - \mathbb{E}[(\widehat{T} - \mathbb{E}[\widehat{T}])^2]$$

Structure of U-Statistics - Hoeffding's decomposition

- General definition of a U-statistic: $Z_1, ..., Z_n$ i.i.d. r.v., f kernel

$$U_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} f(Z_i, Z_j)$$

- Hoeffding's decomposition

$$U_n(f) = \mathbb{E}(U_n(f)) + 2T_n(f) + W_n(f)$$

  where

  - $T_n(f) = \frac{1}{n} \sum_{i=1}^{n} h(Z_i)$ (empirical average of i.i.d.)
  - $h(z) = \mathbb{E} f(Z_1, z)$
  - $W_n(f)$ = degenerate U-statistic (remainder term)

- Degenerate U-statistic W_n with kernel $\tilde{h}$ is such that:

$$\mathbb{E}(\tilde{h}(Z_1, Z_2) \mid Z_1) = 0 \quad \text{a.s.}$$

Main results

In this talk

1. AUC maximization - U-statistic case
2. Finding the best - Signed rank statistic case (with non-smooth SGF)
3. Maximizing general ranking criteria - the case of smooth SGF

Main results

-

1. The U-Statistic case

Bipartite ranking as pairwise classification

- Pairwise classification error $L(s)$

$$L(s) = \mathbb{P}\{(Y - Y') \cdot (s(X) - s(X')) < 0\}$$

- Ranking error and AUC:

$$\mathrm{AUC}(s) = 1 - \frac{1}{2p(1-p)} L(s)$$

- Maximization of AUC $=$ Minimization of pairwise classification error

Empirical Ranking Risk Minimization

- Data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
- Empirical criterion for ranking:

$$L_n(s) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}\{(Y_i - Y_j) \cdot (s(X_i) - s(X_j)) < 0\}$$

- M-estimator over a class $\mathcal{S}$ of scoring rules

$$\widehat{s}_n = \arg\min_{s \in \mathcal{S}} L_n(s)$$

- $L_n(s)$ is a U-statistic

The U-Statistic case

- Sampling of $Z_i = (X_i, Y_i)$ i.i.d. over $\mathbb{R}^d \times \{-1, +1\}$
- U-process indexed by scoring rule $s \in \mathcal{S}$

$$\widehat{U}_n(s) - U(s) = \frac{1}{n(n-1)} \sum_{i<j} q_s(Z_i, Z_j),$$

- Kernel:

$$q_s(z, z') = \mathbb{I}\{(y - y') \cdot (s(x) - s(x')) < 0\} - \mathbb{I}\{(y - y') \cdot (s^*(x) - s^*(x')) < 0\}$$

- Key quantity: take Z and Z' i.i.d.

$$h_s(z) = \mathbb{E}\{q_s(z, Z')\} - \mathbb{E}\{q_s(Z, Z')\}$$

Insights for rates-of-convergence results

- Leading term T_n is an empirical process
  - handled by Talagrand's concentration inequality
  - involves "standard" complexity measures:

$\Rightarrow$ Variance control involves the function h

- Exponential inequality for degenerate U-processes
  - VC classes - exponential inequality by Arcones and Giné (AoP1993)
  - general case - a new moment inequality

$\Rightarrow$ additional complexity measures

Fast Rates - VC Case

Theorem

Assume we have:

- *the class $\mathcal{S}$ of scoring rules is a VC major class with dimension V*
- *for all $s \in \mathcal{S}$,*

$$Var(h_s(Z)) \leq c\,(U(s) - U^*)^{\alpha} \qquad \textbf{(V)}$$

with some constants $c > 0$ and $\alpha \in [0, 1]$.

Then, with probability larger than $1 - \delta$:

$$U(\widehat{s}_n) - U^* \leq 2\left(\inf_{s \in \mathcal{S}} U(s) - U^*\right) + C\left(\frac{V \log(n/\delta)}{n}\right)^{1/(2-\alpha)}$$

Margin condition - Bipartite Ranking

- Question: Sufficient condition for Assumption **(V)**

$$\forall s \in \mathcal{S}, \quad \mathrm{Var}(h_s(Z)) \leq c\,(U(s) - U^*)^\alpha \qquad ?$$

- Wich assumptions on $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$?
- Noise Assumption **(NA)**
  There exist constants $c > 0$ and $\alpha \in [0, 1]$ such that :

$$\forall x \in \mathcal{X}, \quad \mathbb{E}(|\eta(x) - \eta(X)|^{-\alpha}) \leq c\,.$$

- Sufficient condition for **(NA)** with $\alpha < 1$

  $\eta(X)$ absolutely continuous on $[0, 1]$ with bounded density

Remainder Term

Degenerate U-process

Consider $\tilde{q}_s$ a class of degenerate kernels, indexed by $\mathcal{S}$, and

$$\tilde{W}_n = \sup_{s \in \mathcal{S}} \left| \sum_{i,j} \tilde{q}_s(Z_i, Z_j) \right|$$

Additional Complexity Measures

$\epsilon_1, \ldots, \epsilon_n$ i.i.d. Rademacher random variables

Complexity measures:

$$(1) \quad Z_\epsilon = \sup_{s \in \mathcal{S}} \left| \sum_{i,j} \epsilon_i \epsilon_j \tilde{q}_s(Z_i, Z_j) \right|$$

$$(2) \quad U_\epsilon = \sup_{s \in \mathcal{S}} \sup_{\alpha : \|\alpha\|_2 \le 1} \sum_{i,j} \epsilon_i \alpha_j \tilde{q}_s(Z_i, Z_j)$$

$$(3) \quad M_\epsilon = \sup_{s \in \mathcal{S}} \max_{k=1\ldots n} \left| \sum_{i=1}^{n} \epsilon_i \tilde{q}_s(Z_i, Z_k) \right|$$

Moment Inequality

Theorem

If $\tilde{W}_n$ is a degenerate U-process, then there exists a universal constant $C > 0$ such that for all n and $q \geq 2$,

$$\left(\mathbb{E}\tilde{W}_n^q\right)^{1/q} \leq C\left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M_\epsilon + n) + q^{3/2}n^{1/2} + q^2\right)$$

- Main tools: symmetrization, decoupling and concentration inequalities
- Related work: Adamczak (AoP, 2006), Arcones and Giné (AoP, 1993), Giné, Latala and Zinn (HDP II, 2000), Houdré and Reynaud-Bouret (SIA, 2003), Major (PTRF, 2006)

Control of the Degenerate Part

Corollary

With probability $1-\delta$,

$$\tilde{W}_n \le C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M_\epsilon \log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n}\right)$$

VC case

$$\mathbb{E}Z_\epsilon \le CnV \ , \qquad \mathbb{E}U_\epsilon \le Cn\sqrt{V} \ , \qquad \mathbb{E}_\epsilon M_\epsilon \le C\sqrt{Vn}$$

Hence, with probability $1-\delta$

$$\tilde{W}_n \le \frac{1}{n}\left(V + \log(1/\delta)\right)$$

Main results
-
2. Finding the best

Finding the best

- Denote by $F_s^{-1}(1-u)$ the $(1-u)$-quantile of $s(X)$
- Take sets of the form:

$$C_{s,u} = \{x \in \mathbb{R}^d \mid s(x) > F_s^{-1}(1-u)\}$$

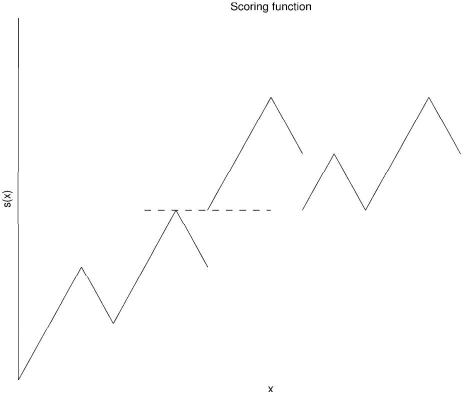
  where s real-valued scoring rule

- Empirical risk:

$$\widehat{W}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \cdot (s(X_i) - \widehat{F}_s^{-1}(1-u)) < 0\}.$$

- Conditions for consistency and (fast) rates:
  - class of scoring functions neither too flat nor too steep
  - behavior of η around $F_\eta^{-1}(1-u)$

Typical scoring functions over the real line



- Left and right derivatives uniformly bounded over the class $\mathcal{S}$

Signed rank statistics

- Take $Z_1, \ldots, Z_n$ i.i.d.
- $\Phi \ : \ [0,1] \to [0,1]$ (score generating function)
- $R_i^+ = \text{rank}(|Z_i|)$

Definition

The statistic

$$\sum_{i=1}^{n} \Phi\left(\frac{R_i^+}{n+1}\right) \ \mathrm{sgn}(Z_i)$$

is a linear signed rank statistic.

Structure of the empirical risk

Notations:

- $K(s,u) = \mathbb{E}\left(Y\ \mathbb{I}\{s(X) \leq F_s^{-1}(1-u)\}\right)$
- $\widehat{K}_n(s,u) = \dfrac{1}{n}\sum_{i=1}^{n} Y_i\ \mathbb{I}\{s(X_i) \leq \widehat{F}_s^{-1}(1-u)\}$

We have:

- $W(s) = 1 - p + K(s,u)$
- $\widehat{W}_n(s) = \dfrac{m}{n} + \widehat{K}_n(s,u)$ where $m = \sum_{i=1}^{n} \mathbb{I}\{Y_i = -1\}$

Observe

For fixed s and u, the statistic $\widehat{K}_n(s,u)$ is a linear signed rank statistic.

Koul's argument - Hoeffding's-type decomposition

Notations:

$$Z_n(s,u) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - K'(s,u) \right) \mathbb{I}\{s(X_i) \leq F_s^{-1}(1-u)\} - K(s,u) + uK'(s,u) \, ,$$

where $K'(s,u) = K'_u(s,u)$.

Proposition

We have, for all s and $u \in [0,1]$:

$$\widehat{K}_n(s,u) = K(s,u) + Z_n(s,u) + \Lambda_n(s) \ .$$

with

$$\Lambda_n(s) = O_{\mathbb{P}}(n^{-1}) \ \text{ as } \ n \to \infty \ .$$

Rates of convergence

- Under VC major class assumption, regular rate of the order $n^{-1/2}$
- Under margin condition:

  $\Rightarrow$ Fast rate of the order $n^{-2/3}$
- Question: weaker assumptions? Faster rates? Lower bounds?

Main results
-
3. Smooth case

Smooth case

- Consider W-ranking functional with ϕ twice continuously differentiable on $[0,1]$

$$\widehat{W}_{k,m}(s) = \sum_{i=1}^{k} \phi\left(\frac{\mathrm{Rank}(s(X_i^+))}{k+m+1}\right), \ \forall s \in \mathcal{S}.$$

- Set F_s^+ (resp. F_s^-) the cdf of $s(X^+)$ (resp. $s(X^-)$)
- We set $\Phi_s(x) = \phi(F_s^+(s(x))) + p \int_{s(x)}^{+\infty} \phi'(F_s^+(u)) dF_s^-(u)$ for all $x \in \mathbb{R}^d$.
- Let $\mathcal{S}$ be a VC major class of functions. Then, we have: $\forall s \in \mathcal{S}$,

$$\widehat{W}_{k,m}(s) = \widehat{V}_k(s) + \widehat{R}_{k,m}(s),$$

where

$$\widehat{V}_{k,m}(s) = \sum_{i=1}^{k} \Phi_s(X_i^+)$$

and $\widehat{R}_n(s) = O_{\mathbb{P}}(1)$ as $n \to \infty$ uniformly over $s \in \mathcal{S}$.

Open problems

1. U-statistic case: Fast rates for convex surrogate loss functions?
2. Finding the best instances: Beyond the $n^{-2/3}$-rate?
3. Smooth case: Fast rates?

And beyond...

- Generic arguments for R-processes?
- General complexity measures for the control of R-processes?
- (Too) Many other questions left...