Concentration Inequalities for Two-Sample Rank Processes with Application to Bipartite Ranking

Myrto Limnios[1], Stephan Clémençon[2], Nicolas Vayatis[1]

[1]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190 Gif-sur-Yvette, France

[2]Telecom Paris, LTCI, Institut Polytechnique de Paris; 19 place Marguerite Perey, Palaiseau, 91120, France

Contributions

- Generalize univariate two-sample *linear rank statistics* to the multivariate setting with generic feature space.
- Provide nonasymptotic theoretical guarantees on this collection of statistics, defined as R-processes, when indexed by classes of *scoring* functions.
- Apply to Bipartite Ranking by considering the R-processes as empirical performance criteria summarizing the ROC curve.

Notations and Framework

- Let $p \in (0,1)$ be the 'theoretical' fraction of the first sample. For $N \geq 1/p$, $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$.
- Let two independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i \leq n}$, $\{\mathbf{Y}_j\}_{j \leq m}$, drawn from $G(dt)$, $H(dt)$ respectively, defined on the sample probability space, valued in $\mathcal{X}$, *e.g.* $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$. Define the mixture *c.d.f.* by $F = pG + (1-p)H$.
- Let $\mathcal{S} = \{s : \mathcal{X} \mapsto \mathbb{R}$ measurable$\}$ be the class *scoring functions* that maps observations into $\mathbb{R}$. Suppose $\mathcal{S}_0 \subset \mathcal{S}$ has finite VC-dimension $\mathcal{V}$.
- Let $M > 0$. For all $s \in \mathcal{S}_0$, the random variables $s(\mathbf{X})$ and $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathcal{W}^{2,\infty}$-norms. The image distributions are denoted by G_s and H_s.
- Denote by Ψ the likelihood ratio defined by $\Psi : x \in \mathcal{X} \mapsto dG/dH(x)$.

Relation to the ROC curve

The ROC curve as the graph of a càd-làg nondecreasing mapping:

$$\mathrm{ROC}(s,.) : \alpha \in [0,1] \mapsto 1 - G_s \circ H_s^{-1}(1-\alpha) \ . \tag{1}$$

W_ϕ-ranking performance criteria as summaries of the ROC curve:

$$W_\phi(s) = \frac{1}{p}\int_0^1 \phi(u)du - \frac{1-p}{p}\int_0^1 \phi\left(p(1-\mathrm{ROC}(s,\alpha)) + (1-p)(1-\alpha)\right) \ d\alpha \ . \tag{2}$$

Optimality

1. The optimal class maximizing W_ϕ is: $\mathcal{S}^* = \{s^* = T \circ \Psi \mid T : [0,1] \to \mathbb{R}$ strictly increasing$\}$.
2. Under the assumptions on ϕ: $W_\phi(s) \leq W_\phi(\Psi) = W_\phi(s^*) = W_\phi^*$ for any $s \in \mathcal{S}$, $s^* \in \mathcal{S}^*$.

Generalization Error Bound

[Cor. 7 in [3]]. Let $\hat{s}$ be an empirical W_ϕ-ranking performance maximizer over the class $\mathcal{S}_0$, *i.e.* $\hat{s} \in \mathrm{argmax}_{s \in \mathcal{S}_0} \widehat{W}^\phi_{n,m}(s)$. Under the assumptions, for any $\delta \in (0,1)$, there exist constants C_1, $C_3 > 0$, $C_2 \geq 4$ depending on ϕ, $\mathcal{V}$ and $C_4 > 0$ depending on ϕ, we have with probability at least $1 - \delta$:

$$W_\phi^* - W_\phi(\hat{s}) \leq 2C_3\sqrt{\frac{\log(C_2/\delta)}{pN}} + \left(W_\phi^* - \sup_{s \in \mathcal{S}_0} W_\phi(s)\right) , \tag{3}$$

as soon as $C_1/\min(1, \sqrt{p^3N}, \sqrt{p^2(1-p)N^3}) \leq C_3\sqrt{\log(C_2/\delta)} \leq C_4\sqrt{pN}\min(1-p, 1/N)$.

References

[1] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[2] S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. In *Advances in Neural Information Processing Systems*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009.

[3] S. Clémençon, M. Limnios, and N. Vayatis. Concentration Inequalities for Two-Sample Rank Processes with Application to Bipartite Ranking. Apr. 2021.

[4] J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Stat.*, 39:325–346, 1968.

[5] A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Related work

- *Linear rank statistics* were initially introduced in semi/nonparametric univariate framework by [5], [4].
- First generalization of the univariate *Mann-Whitney-Wilcoxon* rank statistic applied to hypothesis testing in [2].
- Empirical risk minimization of bivariate loss function has been shown to be equivalent with empirical maximization of the R-statistic associated with ([1]).

Definitions

The *two-sample W_ϕ-ranking performance measure* is defined by:

$$W_\phi(s) = \mathbb{E}[\phi(F_s(s(\mathbf{X})))] \ , \tag{4}$$

where we define by $\phi : [0,1] \mapsto \mathbb{R}$ the *score-generating function* and suppose to be fixed, nondecreasing and twice continuously differentiable. The empirical counterpart based on the two samples $\{\mathbf{X}_i\}_{i \leq n}$, $\{\mathbf{Y}_j\}_{j \leq m}$ is:

$$\widehat{W}^\phi_{n,m}(s) = \sum_{i=1}^n \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right) \ , \tag{5}$$

where $\mathrm{Rank}(t) = \sum_{i=1}^n \mathbb{I}\{s(\mathbf{X}_i) \leq t\} + \sum_{j=1}^m \mathbb{I}\{s(\mathbf{Y}_j) \leq t\} = N\widehat{F}_{s,N}(t)$. The smooth version *via* kernel regularization is:

$$\widetilde{W}^\phi_{n,m,h}(s) = \sum_{i=1}^n (\phi \circ \widehat{F}_{s,N,h})(s(\mathbf{X}_i)) \ , \tag{6}$$

where $\widehat{F}_{s,N,h}(t) = (1/N)\sum_{i=1}^n \kappa\left((t - s(\mathbf{X}_i))/h\right) + (1/N)\sum_{j=1}^m \kappa\left((t - s(\mathbf{Y}_j))/h\right)$ as the smooth version of $\widehat{F}_{s,N}$.

Gradient Ascent Algorithm and Experiments

Algorithm 1: Gradient Ascent for smooth W_ϕ criteria maximization wrt parametric classes $\mathcal{S}_0(\Theta)$

Data: Independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i \leq n}$ and $\{\mathbf{Y}_j\}_{j \leq m}$.
Input: Score-generating function ϕ, kernel K, bandwidth $h > 0$, number of iterations $T \geq 1$, step size $\eta > 0$.
Result: Scoring rule $s_{\widehat{\theta}_{n,m}}(z)$.

```
Choose the initial point θ^(0) in Θ;
for t = 0, ..., T − 1 do
    compute the gradient estimate ∇_θ( W̃^φ_{n,m,h}(s_{θ^(t)}) ) ;
    update the parameter θ^(t+1) = θ^(t) + η∇_θ( W̃^φ_{n,m,h}(s_{θ^(t)}) ) ;
end
Set θ̂_{n,m} = θ^(T).
```

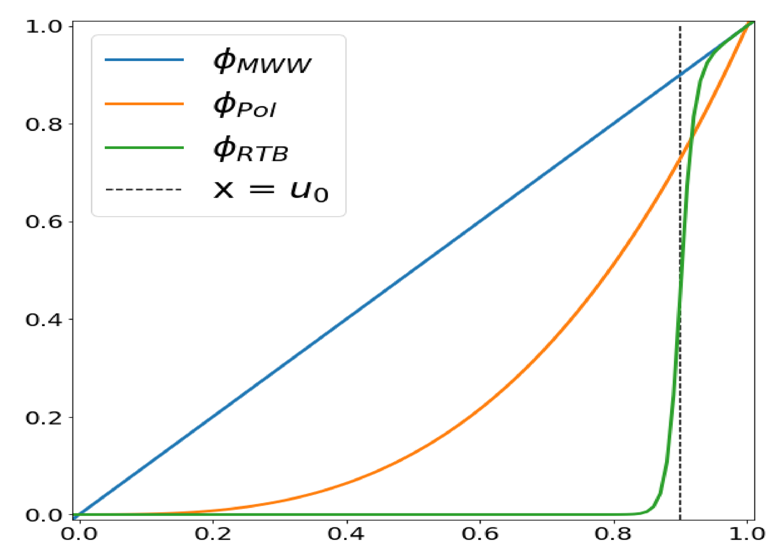
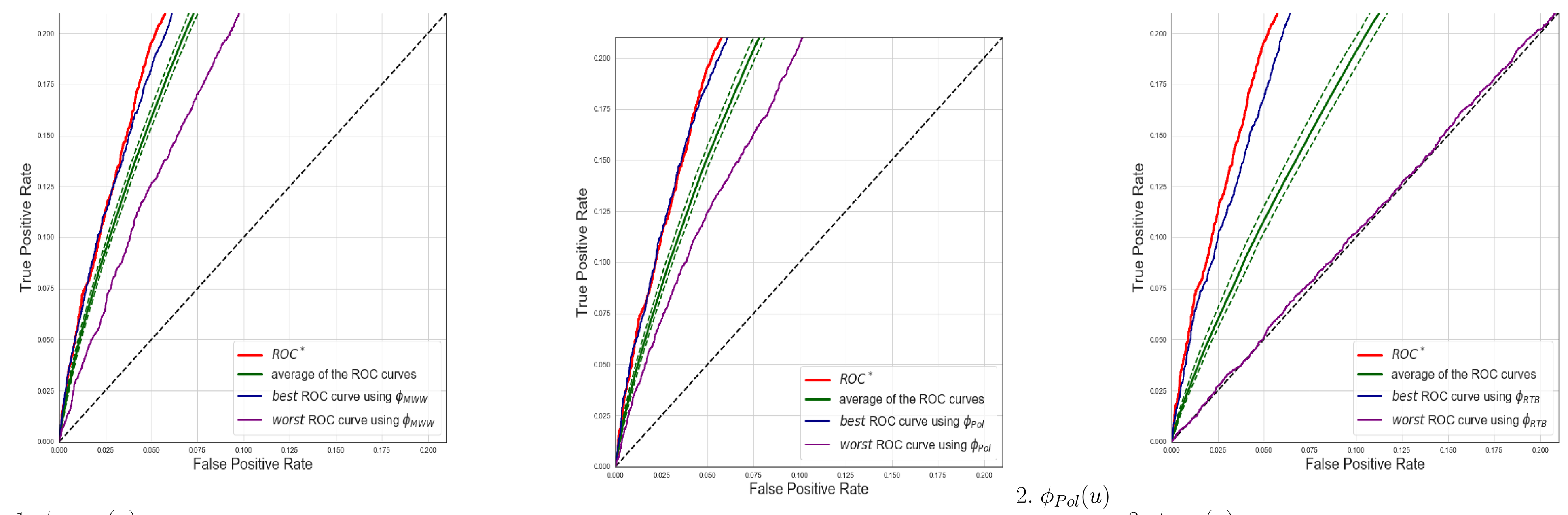


Figure: $\phi_{MWW}(u) = u$ in blue, $\phi_{Pol}(u) = u^3$ in orange, $\phi_{RTB}(u) = u\mathbb{I}\{u \geq u_0\}$ in green, vertical line at $x = u_0$ in black.



1. $\phi_{MWW}(u)$ 2. $\phi_{Pol}(u)$ 3. $\phi_{RTB}(u)$

Figure: Empirical and average ROC curves for the location model with $\mathbf{X} \sim \mathcal{N}_d(\varepsilon 1_d, \Sigma)$, $\mathbf{Y} \sim \mathcal{N}_d(1_d, \Sigma)$ $(\varepsilon = 0.20)$ with linear $\mathcal{S}_0$.