Objective: Derive optimal dissimilarity criteria by tailoring ranking random processes in the two-sample problem.

Two-sample problem: Let \( \{X_i\}_{i \leq N}, \{Y_i\}_{i \leq N} \) be observations drawn from two samples of unknown probability distribution. By ranking the first sample’s data amongst the pooled sample, being able to distinguish possible differences between both distributions.

**Contributions**

- Express empirical performance criteria as linear rank statistics by using Hajek projection and Hoeffding decomposition of U-statistics techniques.
- Analyse concentration properties of this novel class of linear rank processes when it is generalized with unknown scoring generating function and optimized over the class of measurable scoring functions.
- In-depth understanding of both global and local dissimilarities criteria for the two-sample problem.
- Apply linear rank processes in the two-sample problem and nonparametric homogeneity tests in high dimension.

**Notations and Framework**

- Let \( X \sim G, Y \sim H \) two independent absolute continuous r.v. in the probability space \( (\mathcal{X}, \mathcal{P}(\mathcal{X})) \) and consider \( \{X_i\}_{i \leq N}, \{Y_i\}_{i \leq N} \) its realizations a.s. p. proportion of \( X \) in the pooled sample. Denote by \( F \) the c.d.f. of the pooled sample s.t. \( F := pG + (1-p)H \).
- Let \( S \) by the major class of scoring functions s.t. \( S := \{s: \mathcal{X} \to \mathbb{R} \} \) measurable that maps observations into the real line where its natural relation order can be used. \( S \) has a VC-dimension denoted by \( V \).
- Once observations are mapped with \( s \in S \), denote by \( G_s \) (resp. \( H_s, F_s \)) the c.d.f. of \( s(X) \) (resp. \( s(Y) \), \( F = pG_s + (1-p)H_s \)).
- Denote by \( p \) the likelihood ratio defined by \( p : x \in \mathcal{X} \to \frac{G(s(x))}{H(s(x))} \).

**Choice of the score-generating function \( \phi \)**

<table>
<thead>
<tr>
<th>Scoring-generating function</th>
<th>Empirical Ranking process</th>
<th>Related Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \phi := \mathbb{I} )</td>
<td>( \hat{W}<em>{n,m}(s) = \frac{1}{N+1} \sum</em>{i=1}^{N} \sum_{j=1}^{m} \mathbb{I}(s(Z_i) \leq s(X_{ij})) )</td>
<td>Mann-Whitney-Wilcoxon [1]</td>
</tr>
<tr>
<td>( \phi : u \to u : \mathbb{I}(u \geq 0), u \in [0,1) )</td>
<td>( \hat{W}<em>{n,m}(s) = \frac{1}{N+1} \sum</em>{i=1}^{N} \sum_{j=1}^{m} \mathbb{I}(\text{Rank}(s(X_{ij})) \geq \text{Rank}(s(Y_{ij}))) )</td>
<td>Local AUC, concentrates the decision rule on the &quot;best&quot; instances [3]</td>
</tr>
<tr>
<td>( \phi : u \to u^q )</td>
<td>( \hat{W}<em>{n,m}(s) = \frac{1}{N+1} \sum</em>{i=1}^{N} \sum_{j=1}^{m} \text{Rank}(s(X_{ij}))^q )</td>
<td>Related to ( \phi )-norm push [6]</td>
</tr>
</tbody>
</table>

Table: Examples of different choices of scoring generating functions

**Optimality**

Let \( n, m \in \mathbb{N}^* \), express:

\[
\hat{W}_{n,m}(s) = \frac{1}{N+1} \sum_{i=1}^{N} \sum_{j=1}^{m} \mathbb{I}(s(Z_i) \leq s(X_{ij})) , \quad \forall s \in S.
\]

where \( \hat{F}_{X|Y} \) is the empirical c.d.f. of the scored pooled sample.

A widely used tool for measuring the performance of a scoring function \( s \) is the ROC curve defined by:

\[
\text{ROC}(s, .) : \alpha \in [0,1] \mapsto 1-G_s \cdot H_{s^{-1}}(1-\alpha)
\]

Goal: Interpret the \( R \)-processes as optimal unbiased two-sample statistic through the ROC functional curve.

Consider \( S^* := \{ s^*: T \circ \phi : T: [0,1] \to \mathbb{R} \text{ strictly increasing} \} \).

Proposition: Assume that the score-generating function \( \phi \) is strictly increasing. Then, we have:

\[
\forall s \in S^* , \quad W_n(s) \leq W_n(\phi) .
\]

Moreover \( W_n(\phi) = W_n(\phi^*) \) for any \( s^* \in S^* \).

Consequence: The optimal scoring function \( s^* \in S \) for the homogeneity two-sample problem is the solution of the empirical maximization of the \( R \)-process \( \{W_{n,m}(s)\}_{s \in S} \).

**Related work**

- **Linear rank statistics:** were initially introduced in semi/nonparametric univariate framework by [7], [5].
- **Empirical risk minimization:** Bivariate loss function has been shown to be equivalent with empirical maximization of the \( R \)-statistic associated with [22].
- **Hypothesis testing:** has been widely studied in univariate and mostly parametric framework.
- **Homogeneity testing:** for the two-sample problem has recently gained interest for multivariate distribution-free settings, especially through the work of Gretton [4] by introducing the Maximum Mean Discrepancy.

**Linear rank processes**

Definitions: The \( W \)-ranking performance measure for two samples is defined by:

\[
W_n(s) = E(\phi(F_s(s(X)))) , \quad \forall s \in S .
\]

Let \( n, m \in \mathbb{N}^* \), the empirical \( W \)-ranking performance measure for two samples \( \{X_i\}_{i \leq N}, \{Y_i\}_{i \leq N} \) has the following empirical risk function:

\[
\hat{W}_{n,m}(s) = \frac{1}{N+1} \sum_{i=0}^{N} \mathbb{I}(\text{Rank}(s(X_{ij})) \geq \text{Rank}(s(Y_{ij}))) , \quad \forall s \in S .
\]

The function \( \phi : [0,1] \to [0,1] \) is called the score-generating function of the rank process \( \{\hat{W}_{n,m}(s)\}_{s \in S} \). It is supposed to be fixed, nondecreasing and continuously twice differentiable.

**Optimality**

Let \( n, m \in \mathbb{N}^* \), express:

\[
\hat{W}_{n,m}(s) = \frac{1}{N+1} \sum_{i=1}^{N} \sum_{j=1}^{m} \mathbb{I}(s(Z_i) \leq s(X_{ij})) , \quad \forall s \in S.
\]

where \( \hat{F}_{X|Y} \) is the empirical c.d.f. of the scored pooled sample.

A widely used tool for measuring the performance of a scoring function \( s \) is the ROC curve defined by:

\[
\text{ROC}(s, .) : \alpha \in [0,1] \mapsto 1-G_s \cdot H_{s^{-1}}(1-\alpha)
\]

Goal: Interpret the \( R \)-processes as optimal unbiased two-sample statistic through the ROC functional curve.

Consider \( S^* := \{ s^*: T \circ \phi : T: [0,1] \to \mathbb{R} \text{ strictly increasing} \} \).

Proposition: Assume that the score-generating function \( \phi \) is strictly increasing. Then, we have:

\[
\forall s \in S^* , \quad W_n(s) \leq W_n(\phi) .
\]

Moreover \( W_n(\phi) = W_n(\phi^*) \) for any \( s^* \in S^* \).

Consequence: The optimal scoring function \( s^* \in S \) for the homogeneity two-sample problem is the solution of the empirical maximization of the \( R \)-process \( \{W_{n,m}(s)\}_{s \in S} \).

**Linearization**

Proposition: Let \( S_0 \subset S \) be a VC-major class of functions and suppose \( \phi \) as in definition. Then:

\[
\hat{W}_{n,m}(s) = n\hat{W}_n(s) + \hat{V}_m(s) + O(1) , \quad \forall s \in S_0
\]

up to a centering term for the random process, for \( n, m \),

\[
\hat{V}_m(s) = \sum_{i=1}^{N} \hat{F}_m(s(X_i)) + \sum_{i=1}^{m} \hat{F}_m(s(Y_i))
\]

where \( \hat{F}_m : x \in \mathcal{X} \to \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}(s(Y_j) \geq s(x)) \).

**Uniform bound**

**Theorem:** Under the same assumptions, at \( \phi \) fixed, let the empirical \( W \)-ranking performance maximizer \( \hat{s}_{n,m} = \arg\max_{s \in S} \hat{W}_{n,m}(s) \). We have, for any \( \delta \in (0,1) \),

\[
W_{n,m}(\hat{s}_{n,m}) \leq n \sqrt{\frac{1}{N} \log(2/\delta)} + \frac{1}{N} \log(2/\delta)
\]

for some universal positive constants \( \kappa, k_1, k_2 \), depending on \( n, m \), bounds of \( \phi \) and its derivatives, \( V(S_0) \). (\( \kappa, k_1, k_2 \)) depend also on \( \delta \).