Supplement to: Quantitative robustness of instance ranking problems

Tino Werner

1 Characterization of the OIBDP

Having a contaminated sample Z_n^m, we can rewrite the objective as

$$\begin{aligned}
L_{\lambda,n}(b,\beta,Z_n^m) := &\left[\frac{1}{n(n-1)}\sum\sum_{i\neq j,i,j\in I} L((Y_i-Y_j)(s_{b,\beta}(X_i)-s_{b,\beta}(X_j)))+J_\lambda(\beta)\right] \\
&+\frac{1}{n(n-1)}\sum\sum_{i<j,i,j\in I^0} L((Y_i^0-Y_j^0)(s_{b,\beta}(X_i)-s_{b,\beta}(X_j))) \\
&+\frac{1}{n(n-1)}\sum\sum_{i<j,i\in I^0,j\in I} L((Y_i^0-Y_j)(s_{b,\beta}(X_i^0)-s_{b,\beta}(X_j))) \\
&=: G_{\lambda,n}(\tilde\beta,Z_{n-m})+F_n(\tilde\beta,Z_m^0)+H_n(\tilde\beta,Z_{n-m},Z_m^0)
\end{aligned} \tag{1}$$

for $\tilde\beta=(b,\beta)$ and where Z_m^0 denotes the contaminated part of the sample, i.e., $Z_m^0=\{(X_i^0,Y_i^0),i=1,...,m\}$, Z_{n-m} is the clean part of the sample and I^0 denotes the indices of the contaminated instances w.r.t. Z_n, i.e., I^0 is an $m-$subset of $\{1,...,n\}$ such that $I\cup I^0=\{1,...,n\}$ and $I\cap I^0=\emptyset$ for the indices I of the clean instances w.r.t. Z_n. Note that Zhao et al. (2018) do not need the H_n-term since there are no interactions between clean and contaminated instances.

Before we state the definition of our OIBDP for ranking, we proceed along the same lines as Zhao et al. (2018) and study the effect of outliers. Analogously, we first consider a single outlier, wlog. $I^0=\{1\}$, i.e., we have the data set $\tilde Z_n=Z_{n-1}\cup\{Z_1^0\}$ where $Z_1^0=(X_1^0,Y_1^0)$ is some contaminated instance. Let

$$(\hat b,\hat\beta(\tilde Z_n))=\underset{(b,\beta):|b|<\infty}{\operatorname{argmin}}(L_{\lambda,n}(\tilde\beta,\tilde Z_n)).$$

Let the loss function satisfy $\lim_{u \to -\infty}(L(u)) = \infty$ for illustration. Then, for $||X_1^0|| \to \infty$ and $||Y_1^0|| \to \infty$, the loss diverges for each instance such that $(\hat{Y}_i - \hat{Y}_j)(Y_i - Y_j) \to -\infty$. Here, we have to ensure that $(s_{\hat{b},\hat{\beta}}(X_i) - s_{\hat{b},\hat{\beta}}(X_1^0))(Y_i - Y_1^0) \geq 0$ for all $i \neq 1$. These conditions correspond to the constraint set

$$\check{S}^+_{Z_1^0} := \bigcap \check{S}^+_{Z_1^0}(i), \quad \check{S}^+_{Z_1^0}(i) := \{\beta \mid (s_{b,\beta}(X_i) - s_{b,\beta}(X_1^0))(Y_i - Y_1^0) \geq 0\}$$

which mimicks the set $S^+_{Z_1^0}$ from (Zhao et al., 2018, p. 8), so that we can formulate the optimization problem as (Zhao et al., 2018, Eq. (3.3)) as

$$\min_{|b|<\infty, \beta \in \check{S}^+_{Z_1^0}} (G_{\lambda,n}(\tilde{\beta}, Z_{n-1}))$$

with $G_{\lambda,n}$ from Eq. 1. Clearly, the set $\check{S}^+_{Z_1^0}$ is a cone as an intersection of cones which is true since for $\beta \in \check{S}^+_{Z_1^0}(i)$ it holds that $c\beta \in \check{S}^+_{Z_1^0}(i)$ for any $c \geq 0$. For the case of m outliers, we similarly get the constraint set

$$\check{S}^+_{Z_m^0} = \bigcap_{i \in I^0} \check{S}^+_{Z_i^0}.$$

In the ranking setting, we get interesting insights into the constraint set due to the fact that even in the single outlier case, **the pairwise nature of ranking problems lets the outlier act globally**, in contrast to the classification setting from Zhao et al. (2018) where each outlier acts locally. We now distinguish between $X-$ and $Y-$outliers.

i) $Y-$outliers: If we have only outliers in the response, all X_i are maintained but some Y_i are contaminated such that one observes Y_i^0, i.e., the outlier set contains instances (X_i, Y_i^0) for $i \in I^0$. Let wlog. Y_1^0 be the only outlier. Then, letting $||Y_1^0|| \to \infty$ can cause the data to be (linearly) inrankable. To see why this is only possible but not guaranteed, consider the most simple case that $p = 1$. Let $Y_{i_1} \leq ... \leq Y_{i_n}$ for pair-wise different $i_j \in \{1, 2, .., n\}$. Having linearly rankable data, i.e., $Y_j > Y_i \Leftrightarrow X_j > X_i$, replacing Y_{i_1} with an extreme negative outlier, i.e., letting $Y_{i_1}^0 \to -\infty$, does not alter the linear rankability of the data, therefore, is not affecting the quality of the estimated coefficient so that its sign is maintained. The same holds for $Y_{i_n}^0 \to \infty$. On the other hand, for general p, if one of the intermediate responses is replaced by an extreme outlier, the set $\check{S}^+_{z_1^0}$ breaks down to $\{0_p\}$.

ii) $X-$outliers: Extreme $X-$outliers can have a similar effect. $X-$outliers solely affect the predictors, so the outliers are given by (X_i^0, Y_i) for $i \in I^0$, i.e., the attacker cannot alter the responses. Consider again the example with $p = 1$ as above and let $||X_{i_1}^0|| \to -\infty$. Again, this has no effect since any positive coefficient still produces a perfect ranking, similarly when replacing the regressor X_{i_n} corresponding to the largest response Y_{i_n} by an extreme positive outlier.

We can prove an analog to (Zhao et al., 2018, Thm. 2). In our work, the result is of lesser importance since we cannot conclude that the OIBDP always exists which Zhao et al. (2018) indeed can for their angular BDP. Assume that we have a bounded loss function, i.e., $L(u) \leq C_l < \infty$. Define

$$G^u_{\lambda,m}(\tilde{\beta}, Z_{n-m}) = G_{\lambda,n}(\tilde{\beta}, Z_{n-m}) + \frac{n(n-1)-(n-m)(n-m-1)}{n(n-1)} C_l,$$

as an analog to Zhao et al. (2018) with G_{λ_n} from Eq. 1. The second summand indicates the upper loss bound achieved due to the outliers, i.e., both due to the pairwise comparisons of outliers as well as due to the pairwise comparisons of outliers and non-outliers.

Theorem 1.1. *Let $\beta \neq 0_p$ and let the loss function be decreasing with $\lim_{u \to \infty}(L(u)) = 0$ and $\lim_{u \to -\infty}(L(u)) = C_l < \infty$. Then the estimator does not break down in the sense of the population resp. sample OIBDP for ranking if and only if*

$$\min_{\tilde{\beta}_1 \in \Delta^+_{BL}} (G^u_{\lambda,n}(\tilde{\beta}_1, Z_{n-m})) < \min_{\tilde{\beta}_2 \in \Delta^-_{BL}} (G_{\lambda,n}(\tilde{\beta}_2, Z_{n-m})) \tag{2}$$

for $S^-_\cap$ as in Def. 4, $S^+_\cap = \mathbb{R}^p \setminus S^-_\cap$ and $\Delta^+_{BL} = \{(b, \beta) \mid \beta \in S^+_\cap, |b| < \infty\}$ and Δ^-_{BL} analogously.

Proof. *We argue along the same lines Zhao et al. (2018) with some modifications but for making the proof self-contained, we detail out the steps. We restrict ourselves to the population setting since in the sample setting, one just has to replace the original coefficient by the coefficient estimated on the original data. We define the following set of outliers:*

$$\check{Z}^m_0(X, Y, \beta) := \{(X^0_i, Y^0_i) \mid X^0_{ij} = X_{ij} + c_{ij} \ \forall j : \beta_j \geq 0, X^0_{ij} = X_{ij} - c_{ij} \ \forall j : \beta_j < 0,$$
$$0 < c_{ij} < c_{kj} \ \forall i < k \ \forall j, |X^0_{ij}| > \max_i(|X_{ij}|) \ \forall i, Y^0_i > Y^0_k \ \forall i < k, \max(Y^0_i) < \min(Y_i)\}.$$

Graphically, this set is easily understood and is depicted for an example in Fig. 1 in the proof of Lemma 2 for $p = 1$. By construction of the X^0_i, we proceed along the cone where $x\beta$ is increasing ensuring that the magnitude of each component X^0_{ij} exceeds the magnitude of each X_{ij} and that the X^0_i are different. However, the responses are defined such that they are descending with i while the original coefficient β would lead to an ascending ordering, i.e., the ordering is reverted and on top of that, the ordering of each outlier compared with each original observation is reverted.

Consequently, for any $\tilde{\beta}_2 \in \Delta_{BL}^-$, we have

$$L_{\lambda,n}(\tilde{\beta}_2, \check{Z}_n) = \min_{Z_m^0}(L_n(\tilde{\beta}_2, \tilde{Z}_n)) = G_{\lambda,n}(\tilde{\beta}_2, Z_{n-m})$$

where $\check{Z}_n = Z_{n-m} \cup \check{Z}_m^0$ for $\check{Z}_m^0 \in \check{Z}_m^0(X,Y,\beta)$ and $\tilde{Z}_n = Z_{n-m} \cup Z_m^0$ for any outlier set Z_m^0 since any such $\tilde{\beta}_2$ reverts the ordering on the original data but makes perfect predictions for all pairs of outliers and all pairs with one outlier and one original response. We cannot guarantee that any $\tilde{\beta}_1 \in \Delta_{BL}^+$ achieves the worst-case loss for the components H_n and F_n, but by construction, for such a given $\tilde{\beta}_1$, there definitely exists an outlier set $\check{Z}_m^0$ such that

$$L_{\lambda,n}(\tilde{\beta}_1, \check{Z}_n) = \sup_{Z_m^0}(L_n(\tilde{\beta}_1, \tilde{Z}_n)) = G_{\lambda,n}^u(\tilde{\beta}_1, Z_{n-m}).$$

Although we cannot guarantee that a worst-case outlier set exists such that every coefficient that does not satisfy the breakdown criterion suffers supremal loss, for now we can only conclude that the estimator does not breakdown if

$$\min_{\tilde{\beta}_1 \in \Delta_{BL}^+}(G_{\lambda,n}(\tilde{\beta}_1, Z_{n-m})) < \min_{\tilde{\beta}_2 \in \Delta_{BL}^-}(G_{\lambda,n}(\tilde{\beta}_2, Z_{n-m})).$$

Now, we are ready to prove the stated equivalence.
i) *Assume that the estimator does not break down, i.e., the computed estimator $\hat{\beta}_\lambda(\tilde{Z}_n)$ is contained in Δ_{BL}^+ for any outlier set Z_m^0. Then, due to the fact that for any $\tilde{\beta}_1 \in \Delta_{BL}^+$, there exist an outlier set such that $\tilde{\beta}_1$ suffers the maximal loss $G_{\lambda,n}^u(\tilde{\beta}_1, Z_{m-n})$, it follows that the inequality in Eq. 2 indeed holds.*
ii) *Assume that inequality 2 holds. The property $L_{\lambda,n}(\tilde{\beta}_1, \check{Z}_n) \leq G_{\lambda,n}^u(\tilde{\beta}_1, Z_{n-m})$ obviously holds. The statement $L_{\lambda,n}(\tilde{\beta}_2, Z_n) \geq L_{\lambda,n}(\tilde{\beta}_2, \check{Z}_n)$ for $\tilde{Z}_n = Z_{n-m} \cup Z_m^0$ holds by construction for any outlier set Z_m^0 still holds so that we conclude*

$$\min_{\tilde{\beta}_1 \in \Delta_{BL}^+}(L_{\lambda_n}(\tilde{\beta}, \check{Z}_n)) \leq \min_{\tilde{\beta}_1 \in \Delta_{BL}^+}(G_{\lambda_n}(\tilde{\beta}_1, Z_{n-m})) < \min_{\tilde{\beta}_2 \in \Delta_{BL}^-}(G_{\lambda,n}(\check{\beta}_2, Z_{n-m})) \leq \min_{\tilde{\beta}_2 \in \Delta_{BL}^-}(L_{\lambda,n}(\tilde{\beta}_2, \tilde{Z}_n))$$

where the strict inequality holds by assumption and where the last inequality holds since any outlier set from the worst case outlier set $\check{Z}_0^m(X,Y,\beta)$ lets no pair of an outlier and an original response suffer any loss for any coefficient from Δ_{BL}^- which is not guaranteed by general outlier sets. Therefore, the estimator does not break down since a coefficient from Δ_{BL}^+ will be optimal, i.e., achieve the minimum loss.

□

2 Proofs and additional examples

2.1 Outliers and breakdown for ranking with linear scoring functions

Proof (Proof of Lem. 1). ***a)*** *Since $J_\lambda(\beta) \to \infty$ as $||\beta|| \to \infty$, we just have to show that there exists a β with $||\beta|| < \infty$ such that the objective is finite. This is true since $\beta = 0_p$ leads to a finite loss and $J_\lambda(0_p) = 0$, so there definitely exists an optimizer of $L_{\lambda,n}$ with finite norm, disregarding if we have Z_n or Z_n^m.*
b) *Recall that ranking loss functions are based on the product of the differences of the responses resp. the fitted responses. Clearly, having infinite values, we face the problem that it is impossible to reasonably define something like "$\infty - \infty$" which would arise if for example $\beta = (\infty, -\infty)$ and $X_i = (1, 1)$. However, the indicator loss function in the hard ranking loss can be simply rewritten as*

$$I(\{\{Y_i > Y_j\} \cap \{\hat{Y}_i < \hat{Y}_j\}\} \cup \{\{Y_i < Y_j\} \cap \{\hat{Y}_i > \hat{Y}_j\}\}),$$

so the loss will always be computable. If $||\beta|| = \infty$ but if $\hat{Y}_i = s_{b,\beta}(X_i)$ and $\hat{Y}_j = s_{b,\beta}(X_j)$ are computable, we can indeed get an infinite norm solution. For example, consider univariate predictors so that $\beta = \infty$ leads to $s_{b,\beta}(x) = \pm\infty$ for $\mathrm{sign}(x) = \pm 1$ (let $\mathrm{sign}(0) := 0$). Then, if the signs coincide, which is true if for example all $X_i > 0$, we just get the loss $L(0)$ as for random guessing which proves the statement since there is no evidence that there exists a coefficient with finite norm which can beat each coefficient with infinite norm.

□

Remark 2.1. *The potential incomputability of the scores for infinite coefficient components is a severe problem. Therefore, we propose to treat the whole ranking model as random guessing (i.e., $\beta = 0_p$) if there exists any i such that $s_{b,\beta}(X_i)$ is not computable.*

2.2 Hard ranking

The following example has been announced in Rem. 9.

Example 2.1. *We already thought of an outlier scheme with more dependencies where, here wlog. $p = 2$, we do not use $X^{(3)} = (X_1' + 2, X_2')$ but $X^{(3)} = (X_1' + 1, X_2' + 1)$, i.e., the outliers form the edges of a square. Then these four outliers already enable two comparisons along both axes which required five outliers using the proposed outlier scheme in the proof of Thm. 2. The next iteration would be to build a square with vertex length 2 where the outliers define the edges, the mid-points of the vertices and the* *middle point of the square which leads to nine comparisons along each axis. In general, we would have k^2 outliers and $k^2(k-1)/2$ axis-wise comparisons. The general strategy would construct a $p-$dimensional hypercube grid with k^p outliers, leading to $k^p(k-1)/2$ axis-wise comparisons. This strategy can be beneficial for small n, for example in the case $p=3$ and $k=2$, we have 4 comparisons along each axis using only 8 outliers while our strategy before would require 12 outliers to beat this (9 outliers would only lead to 3 comparisons along each axis). However, it is easily revealed that this strategy does not work for larger n, for example in the case $p=2$ and $k=5$, we had 25 outliers and 50 comparisons, but with the strategy before, 24 outliers would already enable 66 comparisons. Maybe future research is able to nevertheless reveal a better strategy than ours proposed in the proof of Thm. 2.*

Proof (Proof of Cor. 2)**.** ***i)*** *We prove the statement by setting $k := cn$ for some $c \in]0,1]$. The condition for a breakdown is then*

$$\frac{cn(cn+1)}{2} \overset{!}{>} \frac{(n-pcn-1)(n-pcn-2)}{2}$$
$$\iff c^2n^2 + cn \overset{!}{>} n^2 - 2pcn^2 - 3n + p^2c^2n^2 + 3pcn + 2$$
$$\iff n^2(c^2 + 2pc - p^2c^2 - 1) \overset{!}{>} n(3pc - c - 3) + 2 \overset{n>0}{\iff} n(c^2 + 2pc - p^2c^2) \overset{!}{>} 3pc - c - 3 + \frac{2}{n} \tag{3}$$

and since $3pc - c - 3$ is fixed and $2/n \to 0$ asymptotically, we only have to guarantee that the bracket is positive, i.e.,

$$c^2 + 2pc - p^2c^2 > 0 \implies c = \begin{cases} \frac{-1-p}{1-p^2} \\ \frac{1-p}{1-p^2} = \frac{1}{p+1} \end{cases}$$

where the first case is not meaningful since it contradicts $c > 0$. Therefore, we asymptotically set $c^* = 1/(p+1)$ and get

$$m^* = 1 + c^*np = 1 + \frac{p}{n(1+p)} \iff \frac{m^*}{n} = \frac{1}{n} + \frac{p}{1+p}$$

which asymptotically equals $p/(p+1)$ as stated.

ii)-iii) *The second statement is obvious since for any such sequence $(b_n)_n$ eventually leads to a diverging number $p(n)$ of variables with $p(n)/(1+p(n)) \to 1$. Note that in the last statement in the last line in Eq. 3, the right hand side cannot grow indefinitely with p since $pc < 1$ (due to the requirement that $n - pcn - 2 \geq 0$), so the same computations as for static p hold. The third statement already has been discussed for the fixed $p \geq n$.*

□

The following example provides finite-sample upper bounds for the OIBDP for hard ranking with the indicator loss in selected scenarios.

Example 2.2. *We simulate the BDP for different p and for a sequence of values for $n > p$:*

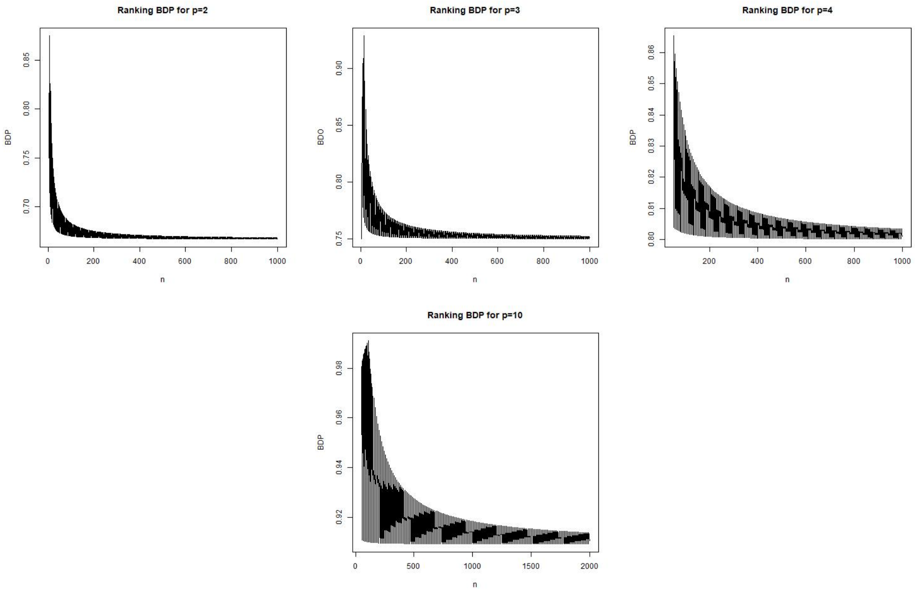

Figure 1: Plots of upper bound for the BDPs for ranking for different p

2.3 Expected OIBDP

Let us outline an artificial case where the outlier flexibility is severely hindered so that the expected OIBDP is better suited.

Example 2.3. *Note that we implicitly assumed* ***open*** *regressor resp. responses spaces when constructing the outlier set. If we have compact regressor or response sets, there is no guarantee that the outlier scheme is applicable. To illustrate this setting for $p = 1$, let again $\beta > 0$ and let the original data be linearly rankable according to this coefficient. More precisely, assume the (very artificial) case that n is even and $n/2$ points are given by $(\max(\mathcal{X}), \max(\mathcal{Y}))$ and the other half of the points at the respective minima. Assuming a bounded loss function, wlog. the indicator loss function, we have no choice but to replace one of these clusters completely by keeping the regressor value but by moving the response to the other extremum of the response space. Now, the usual outlier* *scheme that we already introduced is no longer applicable. The only chance we have is to start by replacing wlog. the whole upper cluster by $n/2$ outliers according to the scheme $(X_i^0, Y_i^0) = (\max(\mathcal{X}) - \epsilon_i, \min(\mathcal{Y}) + \epsilon_i)$ for $\epsilon_1 > \epsilon_2 > ... > \epsilon_k > 0$ with ϵ_1 being small enough to let the first outlier be contained in the open interior of $\mathcal{X} \times \mathcal{Y}$. This strategy is depicted in Fig. 2 where we jittered the points at the left corner only to make them visible.*

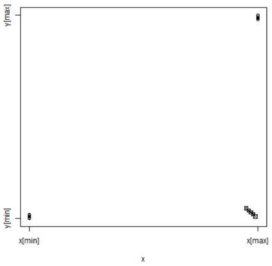

Figure 2: Worst-case outliers for $p = 1$

Tedious algebra reveals an asymptotic BDP of $p^2/(p^2 + 0.5)$.

2.4 Hard binary and hard $d-$partite ranking problems

Proof (Proof of Cor. 3)**.** *Follows the same steps as in the proof of Thm. 2 with the difference that we cannot explicitly produce $Y-$outliers on the score scale (since it is unobservable) but only on the response scale. For illustration, set $p = 2$. We cannot produce extreme $Y-$outliers but we indeed can produce extreme $X-$outliers, so let wlog. $\beta_1, \beta_2 > 0$. We again use a starting point $X' = (\max_i(X_{i1}), \max_i(X_{i2}))$ and set $Y' = 1$. Then, let $Y^{(1)}, Y^{(2)} = -1$ and $X^{(1)} = (X_1' + c_1, X_2')$, $X^{(2)} = (X_1', X_2' + c_2)$. Letting $c_1, c_2 \to \infty$ will induce an unbounded loss when comparing (X', Y') with each of the both outliers, so $\beta_1, \beta_2 < 0$ is enforced. Clearly, this will also produce losses when comparing the original data pairs and the pairs with one original data point and one outlier, but they are finite.*

This strategy clearly also holds for $p = 1$ and $p > 2$ where a complete sign-reversal is again no longer guaranteeable for $p \geq n$. For $p = n - 1$, we can use the last remaining original data point as starting point for the construction if its response is negative and an analogous construction otherwise.

□

Remark 2.2. *Due to the ordering of the classes in $d-$partite ranking problems, a similar approach can be executed to show that the upper bound for the OIBDP for ranking in this setting is identical. One just has to set $Y' = d$ and set $Y^{(1)} = Y^{(2)} = 1$. Letting the respective predictor components diverge, the breakdown should be achievable for any setting where Y' has to be a higher label than $Y^{(1)}$ and $Y^{(2)}$, but however, using the extreme classes is the most logical configuration.*

Proof (Proof of Lem. 3)**.** *Due to the discrete observable response space, we cannot apply the outlier scheme the we suggested in Fig. 1. We first need to identify the configuration of the original responses that supports the true coefficient most. Therefore, we wlog. assume that $\beta > 0$ is the true coefficient. Then, for even n, the worst-case original data configuration (from the view of the attacker) is composed by a set of $n/2$ instances with $X_i > 0$ and $Y_i = 1$ and a set of $n/2$ instances with $X_j < 0$ and $Y_j = -1$. For uneven n, either "half" contains $\lceil n/2 \rceil$ resp. $\lfloor n/2 \rfloor$ instances. Since we do not have a classification problem where one usually classifies all instances with a score greater than zero as class 1 instance and vice versa, we do not necessarily have to consider $X = 0$ as "boundary", we do it just for the sake of easiness.*

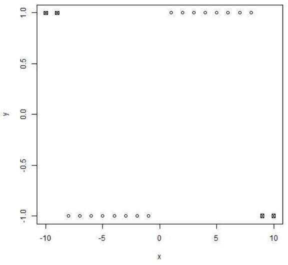

Figure 3: Worst-case outliers for $p = 1$

Now, starting by replacing either the $m < n/2$ rightmost resp. leftmost instances (w.r.t. X) by outliers that keep the regressor value but switch the response sign would not lead to a breakdown if one only has access to the binary outcome, see Fig. 3. This is easily seen assuming $n (mod\ 4) = 0$. If the $n/4$ rightmost instances are replaced as suggested, each $\beta > 0$ will produce exactly $(n/4)^2$ misrankings which arise from comparing each of the $n/4$ original instances on the right half with the $n/4$ outliers. In contrast, any $\beta < 0$ will produce $n^2/8$ misrankings by comparing each of the leftmost $n/2$ instances with the $n/4$ remaining instances on the right half. Note that one cannot compare the *instances on the left half with the outliers since they all have the same response value. Replacing more than $n/4$ instances will even supply the original coefficient more whence we considered the case that around the half of the instances belong to either class as the worst case.*

Therefore, a reasonable outlier scheme is to start by replacing the leftmost and the rightmost instance simultaneously by outliers as illustrated in Fig. 3, i.e., by switching the sign of the response. Note that due to the binary response and the indicator loss function, $Y-$outliers and $X-$outliers can be regarded as being equivalent, so it suffices only to use $Y-$outliers, keeping the original regressor values. This first step induces, for even n, exactly $n/2$ misrankings by comparing the leftmost outliers with every instance of class -1 and additionally $(n/2-1)$ misrankings by comparing the remaining original instances on the right half with the rightmost outlier for $\beta > 0$ (the sum will also be $(n-1)$ for uneven n). In contrast, the remaining $(n/2-1)$ original instances on the left half lead to $(n/2-1)(n/2-1)$ misrankings by comparing them with the $(n/2-1)$ original instances on the right half for $\beta < 0$ (resp. $(\lfloor n/2 \rfloor - 1)(\lceil n/2 \rceil - 1)$ ones).

Assuming that for step k, one has k outliers on each side, i.e., a total of $2k$ outliers, one gets the requirement for k^* stated in Eq. 5.1. Asymptotically, we assume $k = cn$ and easily conclude that $c^* = \frac{1}{2} - \sqrt{1/8}$, so the asymptotic BDP is

$$\frac{m^*}{n} = 2c^* = 1 - \sqrt{\frac{1}{2}}.$$

We already argued in Rem. 2.2 that outlier strategies for binary ranking problems are also applicable to $d-$partite ranking problems. Although one would have more flexible outlier schemes if the instance labels are diverse enough, for example, by considering ascending classes on the right half and descending classes on the left half which enables to produce more misrankings for the original coefficient by taking outlier-outlier-pairs into account, we would be essentially be in the same setting as in the binary ranking problem if one considers the "least favorable" configuration of the original data where the instances on the left half belong to class 1 and the ones on the right half to class d, making the upper bound of the OIBDP for the binary ranking problem also a sharp bound for the $d-$partite ranking problem.

□

Proof (Proof of Thm. 3)**.** *Let us illustrate the proof for $p = 2$. Similarly as in the univariate case in Lem. 3, it does not suffice to generate axis-wise outliers along one direction (i.e., either for very large or very small $X_{\cdot,j}$ for axis j) but one has to generate outliers on both sides. More precisely, again assuming a starting point X^*, for each k one* *has to produce one axis-wise outlier on axis j where the $j-th$ variable is greater than the $j-th$ variable for all other data and where the response is -1 (wlog. let again $\beta_j > 0$ for all j) and one outlier where the $j-th$ variable is lower than the $j-th$ variable for all other data with response 1. This leads to $m = 1+2pk$ outliers per iteration. On each axis, there are k outliers on each side, leading to $k(k+1)$ misrankings since the starting point X^* either has response 1 or -1, leading to kl additional misrankings. In contrast, the sign-reverted coefficient potentially causes misrankings between all remaining $(n-2pk-1)$ original data points, so the formula 5.2 is proven.*

Clearly, if no remaining data points would be available, an early stopping strategy is applicable, i.e., it would suffice to let the starting point have response 1 and to only consider one axis-wise outlier with larger regressor value and response -1, so for $p \leq n-1$, the BDP always exists.

Again, since there essentially is no difference in the robustness of bipartite and $d-$partite ranking problems as already discussed in Rem. 2.2 and the proof of Lemma 3, the results directly transfer to $d-$partite ranking problems.

□

Proof (Proof of Cor. 4)**.** *Statement i) follows using some algebra as in similar statements before, ii) is true since the coefficient in i) converges to 1 for growing p and iii) has already been discussed.*

□

2.5 Localized ranking problems

Proof (Proof of Cor. 5)**.** *If the ranking loss is unbounded, we can directly use Thm. 1 with the only difference that due to the localization on the top K instances, the number n in Thm. 1 must be replaced by K. Then the first part follows.*

As for the case of an unbounded classification loss function, we propose a similar construction as for the hard ranking case. There, we needed one starting point outlier and based on this outlier, we constructed our outlier scheme. Here, we use one of the clean instances of class -1 as starting point. Based on this starting point outlier, we proceed similarly as in the outlier scheme from Thm. 1 by constructing outliers whose predictor vector differs from the predictor vector of the starting point by one component. W.l.o.g., for $\beta_j > 0$ for all j, the starting point has the predictor vector X'. Then, in iteration 1, the outlier j has the predictor vector $(X_1', ..., X_{j-1}', X_j' - c_j, X_{j+1}', ..., X_p')$ and response 1. When moving $c_j \to \infty$, the component $\hat{\beta}_j$ is eventually forced to sign-switch *to a negative sign since otherwise, the loss would diverge. Note that In contrast to the ranking part where the starting point outlier already counts to the K top instances, we can generate K axis-wise outliers in the classification part since the starting point outlier is one of the bottom instances, so the BDP exists unless $p > K$.*

□

Proof (Proof of Lem. 4)**.** ***i)*** *Since we concentrate on the true top K instances, we have to use the outlier scheme depicted in Fig. 4 if $\beta > 0$ is the true coefficient (for $\beta < 0$, the responses of the m rightmost instances would be moved upwards).*

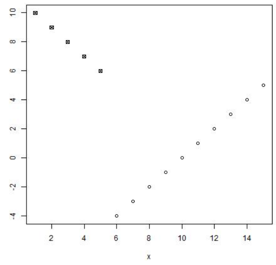

Figure 4: Worst-case outliers for the localized ranking problem for $p = 1$

Then, any coefficient $\beta > 0$ will misclassify m instances, i.e., the outliers, while any coefficient $\beta < 0$ will misclassify $(K - m)$ instances. As for the ranking part, any coefficient $\beta > 0$ will produce both misrankings on the outliers as well as on every pair of an outlier an a clean top-$K-$instance, leading to $m(m-1)/2 + m(K-m)$ misrankings, whereas any coefficient $\beta < 0$ will produce misrankings on the non-outlier top-$K-$instances, i.e., $(K-m)(K-m-1)/2$ misrankings. Therefore, the statement in Eq. 6.1 in the minimum-brackets is true. Moreover, if we had $m = K$ outliers, the loss for $\beta < 0$ would be zero and since any positive coefficient produces a loss greater than zero, indicating that the BDP is not larger than K/n which proves the outer minimum operator in Eq. 6.1.

ii) *The only difference in the case $K > n/2$ compared to the case $K \leq n/2$ is that the broken coefficient does not necessarily cause $(K - m)$ misclassifications due to the fact that more than the half of the instances are labeled as class 1 instances. Consider the concrete example that $n = 100$, $K = 70$ and $m = 10$, following the scheme in Fig. 4. Then $\beta < 0$ does not lead to 60 misclassifications but just to 30 misclassifications, i.e., the rightmost $n - K = 30$ instances. However, if we had $K = 52$, we would not make*

$n - K = 48$ misclassifications but just $K - m = 42$ ones. Since $n - K < K - m$ iff $K > (n + m)/2$, Eq. 6.2 follows.

iii) *Once $m > n - K$, we cannot misclassify m instances with the coefficient of the original sign but only $(n - K)$ ones (the leftmost $(n - K)$ ones in Fig. 4). The same is true for the broken coefficient where the rightmost $(n - K)$ instances in Fig. 4 are misclassified. This lets the classification loss be equal for both cases and reduces the problem to the hard ranking problem on $Best_K$.*

□

Corollary 2.1. ***i)*** *Asymptotically, a fixed K would lead to a BDP of zero. For $K = n$, we get the asymptotic BDP of $1 - \sqrt{0.5}$ as for hard ranking.*
ii) *For $K = K(n) := dn$ for $d \in]0, d_0]$ with $d_0 \approx 0.6352578$, we can conclude that for $m = cn$, we have*

$$c^* = 2 - d - \sqrt{4 - 6d + 5d^2/2}$$

which takes values in $]0, 0.270514]$ and is strictly monotonically increasing w.r.t. d.
iii) *For $K = K(n) := dn$ for $d \in [d_0, d_1]$ with $d_1 \approx 0.773455$, we can conclude that for $m = cn$, we have*

$$c^* = 1 - \sqrt{-1 + 4d - 5d^2/2}$$

which takes values in $[0.2265413, 0.270514]$ and has its minimum at d_1.
iv) *For $K = K(n) := dn$ for $d \in [d_1, 1]$, we have the asymptotic BDP $d(1 - \sqrt{0.5})$.*

Proof. *Statement i) is trivial. The formulae in statements ii) and iii) can be easily computed but we have to explain the value d_0. As we have seen in Lem. 4, it depends on m whether the asymptotic BDP corresponding to Eq. 6.1 or to Eq. 6.2 applies, depending on $\min(K - m, n - K)$. Graphically, we search for the first intersection of both BDPs in dependence of d (note that the second intersection is given at $d = 1$). A numerical evaluation delivers the value d_0 above, see the black curve (asymptotic BDP from ii)) and the blue curve (asymptotic BDP from iii) which is only valid for at least $K > n/2$, therefore the growth for decreasing d can be ignored) in Fig. 5. Note that at d_0, the asymptotic BDP for both cases is given by around 0.270514 which equals $2(d_0 - 0.5)$ and can be explained by our argumentation in the proof of Lemma 4 that the larger m is in the case $K > n/2$, the lower is the number of misclassified instances for the broken coefficient, so according to the formula in Eq. 6.2, we switch between both asymptotics once $n - dn > dn - cn$, i.e., once $c > 2(d - 0.5)$. For statement iv), we similarly have to* *find d for which $dn > n - cn$ holds (i.e., $K > n - m$). Again, by numerical evaluation where we search for the intersection of the purple curve and the black line in Fig. 5 (for illustration, we also added the graph of $d \mapsto 1 - d$ (red line) which intersects the blue curve at the same point) which is the case at d_1 where the BDP is exactly $1 - d_1$, so increasing d from here will cause the regime switch in the classification part of the right hand side, which results in the combined BDP curve in the right part of Fig. 5.*

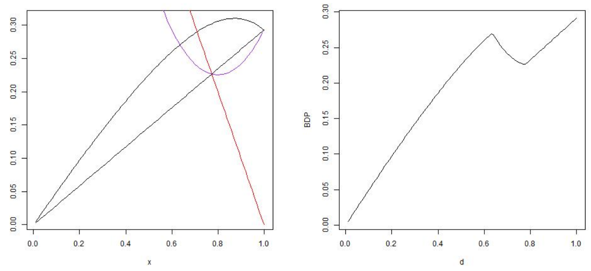

Figure 5: Asymptotic BDP for $K \leq n/2$ and $K > n/2$

□

Proof (Proof of Thm. 4)**.** *Along the same lines as the proofs of Thm. 2 and Lemma 4.*

□

Corollary 2.2. *For the localized ranking problem where the localized ranking loss is computed on $Best_K$ and where both the classification and the ranking loss are indicator functions, we asymptotically conclude that*
i) *the BDP is zero for K and p being fixed,*
ii) *the BDP for $c \leq 0.5$ tends to*

$$pc^*, \quad c^* = \frac{4p - 3pd}{p^2 - 1} - \sqrt{\frac{16p^2 - 28p^2d + 12p^2d^2 + 4d - 3d^2}{(p^2 - 1)^2}}$$

and for $c > 0.5$, it tends to the same quantity provided that $c \leq 2(d - 0.5)$, to

$$pc^*, \quad c^* = \frac{2p - pd}{p^2 - 1} - \sqrt{\frac{4p^2d - 4p^2d^2 - 8d + 5d^2 + 4}{(p^2 - 1)^2}}$$

for $c \in [2(d - 0.5), 1 - d]$ and to $dp/(p + 1)$ otherwise. For $p \to \infty$, the break-even point is given by $d_0 \approx 0.6923$ and the second break-even point tends to 1 for growing p.
iii) *the BDP tends to the asymptotic BDP for hard ranking, i.e., $p/(p + 1)$, for $d = 1$.*
iv) *the BDP does not exist for $p = p(n)$ with $p(n)/n \to b \geq d$.*

Proof. *Statement i) is trivial, statement iv) has already been discussed (since the BDP converges to 1 for $d \to 1$, there is no second regime-switching point as in the univariate case), the formulae in ii) can be easily computed and iii) follows directly. As for d_0, see Fig. 6 for illustration where the black curve corresponds to the first formula in ii) and the red curve to the second formula.*

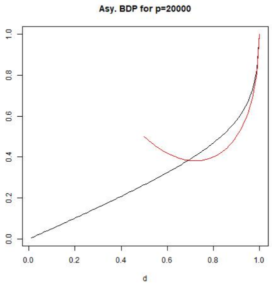

Figure 6: Asymptotic BDP for $K \leq n/2$ and $K > n/2$

□

2.5.1 Localized continuous ranking problems on $\widehat{Best_K}$

Alternatively, we can localize the ranking loss on $\widehat{Best_K}$, i.e., the goal is to ensure that the instances that are predicted to be at the top of the list are ranked correctly, although these instances may not be the true top instances. We provide similar results.

Lemma 2.1. *Let $p = 1$. For the localized continuous ranking problem optimizing*

$$\frac{n-K}{n}\frac{2}{n}\sum_{i \in Best_K} I(\mathrm{rk}(s_\beta(X_i)) > K) + \frac{2}{n(n-1)} \sum\sum_{i<j,i,j \in \widehat{Best_K}} I((s_\beta(X_i) - s_\beta(X_j))(Y_i - Y_j) < 0), \tag{4}$$

the sample and population OIBDP for ranking
i) *is given by*

$$\frac{\check{m}}{n}, \quad \check{m} = \min\left(K, \min\left\{k \;\middle|\; \frac{n-K}{n} \cdot \frac{2(K-m)}{n} + \frac{(K-m)(K-m-1)}{2n(n-1)} < \frac{n-K}{n} \cdot \frac{2m}{n}\right\}\right) \tag{5}$$

for $K \leq (n+m)/2$,

ii) *is given by*

$$\begin{aligned}\frac{\check{m}}{n}, \quad \check{m} = \min\{k \;\Big|\; & \frac{n-K}{n}\cdot\frac{2(n-K)}{n} + \frac{(K-m)(K-m-1)}{2n(n-1)} \\ & < \frac{n-K}{n}\cdot\frac{2m}{n} + \frac{1}{n(n-1)}\left[\frac{m(m-1)}{2} + m(K-m)\right]\}\end{aligned} \tag{6}$$

for $K \in [(n+m)/2, n-m]$,
iii) *is given by Eq. 4.1 in Lemma 2 where* n *in the definition of* $\check{m}$ *is replaced by* K *for* $K \geq n-m$.

Proof. *The situation here is inherently different from the case that the ranking performance is computed on* $Best_K$. *We have to distinguish carefully between the two outlier schemes in Fig. 1 and Fig. 4.*

Let us start with the case that $K < n/2$. *The misclassification rate is obviously not affected by localizing the ranking performance on* $\widehat{Best_K}$, *so the formulae from Lemma 4 remain valid. As for the misrankings, if we consider the outlier scheme as in Fig. 4, we will not produce any misranking for the original coefficient. This is true since for* $\beta > 0$, *the rightmost instances are predicted to be the best ones and the ordering of their responses is correctly predicted as ascending. In contrast, any negative coefficient produces a complete inversion of the ordering of the remaining* $(K-m)$ *original instances, i.e.,* $(K-m)(K-m-1)/n$ *misrankings. Let us now consider the outlier scheme as in Fig. 1. First note that the number of necessary outliers to produce a breakdown cannot exceed* K *since we can achieve a zero loss for the broken coefficient using the outlier scheme in Fig. 4 while the loss for the coefficient of the original sign is greater than zero due to the classification part. Now, the outliers according to Fig. 1 cause* $m(m-1)/2$ *misrankings on the* m *rightmost instances and additional* $m(K-m)$ *misrankings for any pairs of an outlier and one of the remaining rightmost* $(K-m)$ *instances. Any negative coefficient however again produces* $(K-m)(K-m-1)/2$ *misrankings on the intermediate* $(K-m)$ *instances. On the other hand, while the original coefficient only makes* m *misclassifications, the broken coefficient misclassifies the maximum number of* K *instances.*

Now, we have to argue which of the proposed outlier schemes applies. The answer is that it depends on K. *Still assuming* $K \leq (m+n)/2$, *we can observe that the classification loss is constant for the outlier scheme from Fig. 1 for any* $\beta < 0$ *for only a small additional ranking loss for* $\beta > 0$. *We deduct that for* $K \leq (n+m)/2$, *one should use the outlier scheme as in Fig. 4 (asymptotically, it can be shown by numerical evaluation that the required number of outliers is always larger for the outlier scheme as* *in Fig. 1 for $K = dn$ for $d \leq d_0 \approx 0.692291$) and for $K > (n+m)/2$, we should use the outlier scheme as in Fig. 1. This proves formula 5 and part i) as well as formula 6 and part ii) where the dependence on K has already been discussed in the proof of Lemma 4.*

Finally, note that once $K \geq m-n$, the broken coefficient will only misclassify $(n-K)$ instead of m instances, so the misclassification loss is equal for $\beta > 0$ and $\beta < 0$. Since only the ranking part remains which is the same as in the hard ranking setting with n replaced by K, statement iii) is valid.

□

Corollary 2.3. ***i)*** *Asymptotically, a fixed K would lead to a BDP of zero. For $d = 1$, we get the asymptotic BDP of $1 - \sqrt{0.5}$ as for hard ranking.*
ii) *For $K = K(n) := dn$ for $d \in]0, d_0]$ for $d_0 \approx 0.5774659$, we can conclude that for $m = cn$, we have*

$$c^* = 4 - 3d - \sqrt{16 - 28d + 12d^2}$$

which takes values in $]0, 0.30993]$ and is strictly monotonically increases with d.
iii) *For $K = K(n) := dn$ for $d \in [d_0, d_1]$ with $d_1 \approx 0.773455$, we can conclude that for $m = cn$, we have*

$$c^* = 1 - \sqrt{-1 + 4d - 5d^2/2}$$

which takes values in $[0.2265413, 0.30993]$ and has its minimum at d_1.
iv) *For $K = K(n) := dn$ for $d \in [d_1, 1]$, we have the asymptotic BDP $d(1 - \sqrt{0.5})$.*

Proof. *Along the same lines as the proof of the Cor. 2.1. The asymptotic BDP in dependence of d is depicted in Fig. 7.*

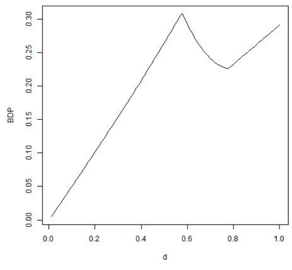

Figure 7: Asymptotic BDP for $K \leq n/2$ and $K > n/2$

□

Theorem 2.1. *Let wlog. the indicator loss functions for both the classification and the ranking part be used and let $p \geq 2$. Then, the upper bound for the OIBDP for localized ranking with the optimization problem as in Eq. 4 is given by*

$$\frac{m^*}{n}, \quad m^* = \min(1 + pk^*, K),$$

$$k^* = \min \left\{ k \;\middle|\; \frac{n-K}{n} \frac{2(K-pk-1)}{n} + \frac{(K-1-pk)(K-2-pk)}{2n(n-1)} < \frac{n-K}{n} \frac{2(1+pk)}{n} \right\} \tag{7}$$

for the case $K \leq (n+m)/2$. For $K > (n+m)/2$, we have $m^* = 1 + pk^*$ where

$$k^* = \min \left\{ k \;\middle|\; \frac{n-K}{n} \frac{2(n-K)}{n} + \frac{(K-1-pk)(K-2-pk)}{2n(n-1)} < \frac{n-K}{n} \frac{2(1+pk)}{n} + \frac{k(k+1)}{2n(n-1)} \right\} \tag{8}$$

and for $K > n - m$, we get the same k^* as in Eq. 4.2 in Thm. 2. This quantity always exists for $p \leq K - 1$.

Proof. *Along the same lines as the proofs of Thm. 2 and Lemma 2.1.*

□

Corollary 2.4. *For the localized ranking problem where the localized ranking loss is computed on $\widehat{Best_K}$ and where both the classification and the ranking loss are indicator functions, we asymptotically conclude that*
i) *the BDP is zero for K and p being fixed,*
ii) *the BDP for $c \leq 0.5$ tends to*

$$pc^*, \quad c^* = \frac{4-3d}{p} - \sqrt{\frac{16 - 28d + 12d^2}{p^2}}$$

and for $c > 0.5$, it tends to the same quantity provided that $c \leq 2(d - 0.5)$, to

$$pc^*, \quad c^* = \frac{2p - pd}{p^2 - 1} - \sqrt{\frac{4p^2 d - 4p^2 d^2 - 8d + 5d^2 + 4}{(p^2-1)^2}}$$

for $c \in [2(d-0.5), 1]$. For $p \to \infty$, the break-even point is given by $d_0 \approx 0.6923$.
iii) *the BDP tends to the asymptotic BDP for hard ranking, i.e., $p/(p+1)$, for $d = 1$.*
iv) *the BDP does not exist for $p = p(n) = b_n n$ with $b_n \to b \geq d$.*

Proof. *Similar as in Cor. 2.2. Note that the first formula in ii) tends to the respective one in Cor. 2.2 for $p \to \infty$ while the second formulae are already equal.*

□

2.6 Other ranking problems

Proof (Proof of Cor. 6)**.** *Follows the same argumentation as the corresponding corollaries before. The starting point can be an arbitrary instance from the bottom of the list, so the axis-wise outliers (with a response tending to infinity) enforce the respective coefficients to switch their sign in order to keep the classification loss low.*

□

Proof (Proof of Thm. 5)**.** *Statement a) is trivial since the number of misclassifications is then lower for the coefficient with the opposite sign than for the original coefficient. Statement b) is only a coarse bound which cannot be tightened due to the missing ranking loss part in the weak ranking problems. The only opportunity that we have is to use an instance from the bottom of the list and to place axis-wise outliers with sufficiently large responses so that they are at the top of the list around this starting point. In the worst case, we have to start from an original instance from the bottom of the list and replace all K true top instances with such axis-wise outliers. Statement c) is obvious.*

□

2.7 The sample OIBDP

In principle, we could directly transfer the results from this paper that restricted themselves to the population OIBDPs to the sample OIBDP setting. However, the sample OIBDP heavily relies on the quality of the estimator $\hat{\beta}$ on the clean data set (note that the sample angular BDP from Zhao et al. (2018) faces the same issue). Let us first recapitulate an important definition that can be found for example in Bühlmann and Van De Geer (2011).

Definition 2.1. *Assume that S^0 is the true set of variables and $\hat{S}$ is the set of parameters selected by the model selection procedure. The model selection procedure is* ***variable selection consistent*** *if $P(\hat{S} = S^0) \longrightarrow 1$ for $n \to \infty$.*

Variable selection consistency is a strict assumption. Theoretical results often cover a relaxed property, the so-called screening property (Bühlmann and Van De Geer (2011)) which indicates that the set of selected variables contains the set of true variables asymptotically, so variable selection consistency is the special case of equality of the two sets. Note that even sophisticated algorithms like L_2-Boosting (Bühlmann and Yu (2003), Bühlmann (2006)) fail to be variable selection consistent (Vogt (2018)). If the model selection procedure therefore has only the screening property, it essentially selects too many variables, making the ranking problem artificially more robust. We do not think that this is reasonable since this seeming robustness would result from the deficiencies of the applied model selection algorithm. Therefore, we only consider variable selection consistent model selection which enables the following asymptotic result.

Corollary 2.5. *Let $q = q(n)$ such that $q(n)$ with $q(n)/n \to b \in]0,1[$ resp $q(n)/n \to b < d$ for $K(n)/n \to d \in]0,1]$. Then the asymptotic sample order-inversal breakdown point for all non-localized ranking problems resp. localized ranking problems considered in this work exists provided that the estimated coeffient is computed using a variable selection consistent procedure.*

Note that variable selection consistency is an asymptotic property, so for fixed n, even a procedure that satisfies this property can produce an estimated coefficient which selects irrelevant variables or which misses relevant variables. Summarizing, we believe that sample versions of BDPs are not very informative and should always be considered with caution.

2.8 Outlook: SVM-type approaches

Proof (Proof of Lem. 5)**.** *The statement is easily seen since $(\hat{\alpha}^*, \hat{\alpha})$ obviously satisfies the constraints. The first sum of the objective does not change, also the third sum does not change by switching the sign of the two factors. The negation of the second sum due to the sign switch of the responses is compensated by the sign switch of the coefficient differences, so the value of the objective for the solution $(\hat{\alpha}^*, \hat{\alpha})$ on the manipulated data is identical to the value of the objective of the solution $(\hat{\alpha}, \hat{\alpha}^*)$ on the clean data and since a minimum is attained, $(\hat{\alpha}^*, \hat{\alpha})$ is optimal.*

□

Proof (Proof of Cor. 9)**.** *Similarly as in the proof of Lemma 5, let $\hat{\alpha}$ be a solution of the corresponding dual optimization problem which is given in (Herbrich et al., 1999, Eq. (68)). The objective function invokes a double sum where factors $Y_i Y_j$ appear (note that $Y_i \in \{\pm 1\}$ for all i in their work). However, switching the sign of all responses will not affect the objective function and therefore keep the solution. Due to the linear expansion of the weights given in (Herbrich et al., 1999, Eq. (69)), all summands that form the weights are sign-switched, so the whole weight coefficient is sign-reverted since the features stay untouched. The same is true if the kernelized SVMs which are computed by maximizing the objective function given in (Herbrich et al., 1999, Eq. (75)), so by the analog expansion of the weights, implicitly given in (Herbrich et al., 1999, Eq. (79)), a breakdown is achieved.*

□

References

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media, Berlin, Heidelberg.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: Regression and Classification. *Journal of the American Statistical Association*, 98(462):324–339.

Herbrich, R., Graepel, T., and Obermayer, K. (1999). *Regression models for ordinal data: A machine learning approach*. Citeseer.

Vogt, M. (2018). On the differences between L_2-boosting and the lasso. *arXiv preprint arXiv:1812.05421*.

Zhao, J., Yu, G., and Liu, Y. (2018). Assessing robustness of classification using angular breakdown point. *Annals of statistics*, 46(6B):3362.