Ranking: Consistency and Rates

Dany Haddad

July 2020

1 Introduction

This is a work in progress document discussing the consistency of common convex surrogate functions for ranking as well as the convergence rate of the corresponding empirical risk. First we summarize the main points of Duchi et al. in this regard [8, 9] focusing on the uniform convergence guarantees. Then we discuss some relevant research directions, in partictular, we are interested in achieving fast convergence rates for ranking applications. [1]

2 Literature Review

2.1 The Asymptotics of Ranking Algorithms

2.1.1 Overview

The authors first introduce the ranking problem and describe a pairwise loss function, equivalent to the pairwise 0-1 ranking loss. They then define several forms of consistency for ranking problems and unify them under some conditions (theorem 1). Surprisingly, common surrogates for the pairwise 0-1 ranking loss function are shown to not be consistent.

The next section discusses how aggregating the pairwise preferences into lists allows us to define consistent and tractable surrogate loss functions. The key idea is that the surrogates must be order preserving. [2]

Given that our new surrogate loss functions consist of an aggregation step, it is natural to consider the aggregation in batches. In particular, the authors describe the loss function as a sum of U-statistics. However, this is seems mostly to simplify the analysis rather than to leverage the fact that U-statistics

[1] I also point out confusion and questions I have in the footnotes.

[2] The authors describe inconsistency as arising due to a lack of complete preference information. This point is not entirely clear to me since the proof of inconsistency seems to depend on the variability in preference information rather than the lack of information. See Ravikumar et al. [1] for a discussion on how a lack of normalization leads to "non-robustness".

are UMVUEs. [3] The authors then prove a ULLN for the risk functional they define and show that it is consistent for the underlying listwise losses.

2.2 Preliminaries

The ranking problem can be described generally as ordering a list of m items for each query q in the query space Q (which we assume is countable). The is equivalent to providing a real valued score for each item in the list.

Duchi et al. [9] consider a general setup where the item preference information is received in an arbitrary form and then transformed into a problem specific structure space by an aggregation function. For example, the preference information Y_i could be a matrix of pairwise preferences and the aggregation function s is either a simple average, or a transformation into a list of item scores. Y_i could also be the interactions a user had with a list of items (click-stream data). We also consider the case where s performs no aggregation at all. Consider the 0-1 pairwise ranking loss:

$$L(\alpha, A) = \sum_{i<j} A_{ij} 1(\alpha_i \leq \alpha_j) + \sum_{i>j} A_{ij} 1(\alpha_i < \alpha_j) \tag{1}$$

where $A = s(\{Y_i\}_{i=1}^k)$. Note that this is different from the typical 0-1 pairwise ranking loss in that the error $\alpha_i = \alpha_j$ is penalized only once. [4]

Assumption A. The distribution of preference information μ_q^k converges weakly to some limiting distribution mu_q for every query q where k is the amount of preference aggregation. So, for $S_k = s(\{Y_i\}_{i=1}^k) \sim \mu_q^k$ as $k \to \infty$:

$$\mu_q^k \to^d \mu_q \tag{2}$$

Assumption B. The loss function L is bounded in $[0, 1]$ and continuous.

Given the loss function L we define the loss over a distribution of preference information μ as:

$$\ell(\alpha, \mu) = \int L(\alpha, s) d\mu(s) \tag{3}$$

For the case where the preference information Y_i is a matrix of pairwise preferences and the aggregation function is the average, then this reduces to:

[3] The authors mention that they introduce the U-statistics based risk functional to be able to analyze the setup without overly detailed knowledge of the surrogate loss. But they already assume the surrogate is lipschitz continuous and bounded. What more information would we need to conduct the analysis without taking the U-statistics route?

[4] They mention that this is a technical detail, but it's not clear to me where in the proofs it is important. Most likely it has to do with the inconsistency results.

$ell(\alpha, \mu) = L(\alpha, Y_{i,j}^{\mu})$ where $Y_{i,j}^{\mu}$ is the average preference of item i over item j according to the distribution μ.

We define the ranking risk of our scoring function $f : Q \to R^m$ as:

$$R(f) = \sum_{q}^{Q} p_q \ell(f(q), \mu_q) \tag{4}$$

Where p_q is the prior probability of query q. Without loss of generality, let p_q be decreasing with the query index q.

Similarly, we define the surrogate loss and surrogate risk as:

$$\ell_{\varphi}(\alpha, \mu) = \int \varphi(\alpha, s) d\mu(s) \tag{5}$$

$$R_{\varphi}(f) = \sum_{q} p_q \ell_{\varphi}(f(q), \mu_q) \tag{6}$$

2.2.1 Inconsistency Results

2.2.2 Proving a ULLN for the U-statistics Based Risk Functional

Define the U-statistics based empirical risk:

$$\hat{R}_{n,\varphi}(f) = \sum_{q} \frac{\hat{n}_q}{n} \frac{1}{\binom{\hat{n}_q}{k}} \sum_{I_q \in \mathcal{I}_q} \varphi(f(q), s(\{Y_i\}_{I_q})) \tag{7}$$

where $\hat{n}_q$ is the frequency of query q in a sample of size n and $\mathcal{I}_q$ is the set of all combinations of indexes of samples corresponding to query q.

We include conditions I and II as well as assumptions D, E and F from Duchi et al. [9] in the statement of theorem 4.

Theorem 4. For $\phi \in [0, B_n]$ bounded and L_n-lipshitz over $\mathcal{F}_n$, if there exists some $\rho, \beta > 0$ such that:

$$p_q = O(q^{-\beta-1}) \tag{8}$$

and the expected surrogate loss at aggegation level k satisfies for all f, k, q, n:

$$\left| \mathbb{E}_q \varphi(f(q), s(\{Y_i\}^k)) - \lim_{k' \to \infty} \mathbb{E}_q \varphi(f(q), s(\{Y_i\}^{k'})) \right| = O(B_n k^{-\rho}) \tag{9}$$

with $B_n / k_n^{\rho} = o(1)$ and $k_n B_n^{(1+\beta)/\beta} = o(n)$ and in addition we have the following constrain on the covering numbers:

$$k_n B_n \sqrt{\log\left(N(\frac{\epsilon}{4L_n})\right)} = o(\sqrt{n}) \tag{10}$$

then we have the uniform convergence guarantee:

$$\sup_{f \in \mathcal{F}_n} \left| \hat{R}_{n,\varphi}(f) - R_\varphi(f) \right| \to^p 0 \tag{11}$$

The proof for this uniform law of large numbers can be broken down by decomposing the uniform deviation into two terms and controlling each of them:

$$\sup_{f \in \mathcal{F}_n} \left| \hat{R}_{n,\varphi}(f) - R_\varphi(f) \right| \leq \sup_{f \in \mathcal{F}_n} \left| \hat{R}_{n,\varphi}(f) - R_{n,\varphi}(f) \right| + \sup_{f \in \mathcal{F}_n} |R_{n,\varphi}(f) - R_\varphi(f)| \tag{12}$$

where we denote:[5]

$$R_{\varphi,n} = \sum_q \sum_{l=0}^{n} l \mathbb{P}_n / n(\hat{n}_q = l) \mathbb{E}(\varphi(f(q), s(\{Y_i\}^{\min(l,k)})) | Q = q) \tag{13}$$

The first term (I) on the RHS of inequality 12 is the deviation of the U-statistic based estimate $\hat{R}_{n,\varphi}(f)$ from it's expectation $R_{n,\varphi}(f)$. The second term (II) is the deviation of $R_{n,\varphi}(f)$ from it's limit $R_\varphi(f)$ as $n, k \to \infty$. The second term is controlled by lemma 8 of the supplementary material [9] while the first term is controlled by lemma 10.

Lemma 8 uses the conditions on the the boundedness of ϕ and the aggregation level k_n as well as a covering number argument to give a deterministic bount on the second term above.

Lemma 10 shows that the first term is $o_p(1)$ by arguing that the risk functional $\hat{R}_{n,\varphi}(f)$ satisfies a bounded differences inequality. Applying McDiarmid's inequality as well as a union bound involving the covering number gives the following bound, yielding the claim:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_n} \left| \hat{R}_{n,\varphi}(f) - R_{n,\varphi}(f) \right| \geq \epsilon \right) \leq 2 \exp\left(\log N\left(\frac{\epsilon}{4L_n}, n \right) - \frac{n\epsilon^2}{32 k^2 B_n^2} \right) \tag{14}$$

where $N\left(\frac{\epsilon}{4L_n}, n\right)$ gives the covering number of $\mathcal{F}_n$. [6]

[5] The factor of $1/n$ is missing from the paper but is used in the supplementary material.

[6] My understanding is that this gives a rate worse than $O_p(1/\sqrt{n})$ since the concentration inequality involves the covering number which is dependent on ϵ. Unless for the function class under consideration this can be bound uniformly independent of *epsilon*, this is worse than the typical rate.

2.2.3 Potential improvements

The U-statistics formulation is totally ignored in controlling term (I), leading to a rate slower than $O_p(1/\sqrt{n})$. For the case of bipartite ranking, Clemencon et al. [6] derive the typical $O_p(q/sqrtn)$ rate of convergence of the empirical risk minimizer to the optimal optimizer. The approach involves making a low noise assumption (bound variance of the risk in terms of the expectation) and applying a Hoeffding decomposition of the U-statistic based risk. The decomposed terms are controlled by applying the decoupling technique for degenerate U-statistics.

However, their results are more relevant in a model selection context since they deal directly with the non-convex pairwise 0-1 ranking loss. Clemencon et al. [6] also mention (in Remark 7) that it would be interesting to derive fast rates for convex surrogates, but that it would be quite technical. They point to Blanchard et al. [3] for a similar derivation for the binary classification case.

Bartlett et al. [2] derive fast rates for convex surrogates in the context of binary classification. They also demonstrate via simulation that their derived rates are representative. Clemencon et al. [6] show how the results from this work transfer over to the case of bipartite ranking but do not derive fast rates.

The U-statistics based empirical risk defined in this work (7) resembles a bagging approach and might be analyzed in a similar way. The objective function of the bipartite ranking problem [6] (equivalently, AUC maximization) is naturally framed as an optimization of a U-statistic. The ranking problem discussed here is more general, bipartite ranking does not consider the impact of a query context nor does it consider graded relevance levels.

2.3 Ranking and Empirical Minimization of U-Statistics

Clèmençon et al. [6] explore fast rates for empirical risk minimization of U-statistics based loss functions of order 2. In particular, they show that under suitable noise conditions, empirical risk minimization of the bipartite ranking loss gives fast rates dependent on the exponent in the satisfied noise condition. The low-noise condition is similar to that defined elsewhere (see Boucheron et al. [4]) adapted for the U-statistics case:

$$\mathrm{Var}(h_r(X, Y)) \leq c\Lambda(r)^\alpha$$

where c is some positive constant and $\alpha \in [0, 1]$ is the coefficient determining the low-noise behavior. h_r is the kernel defined in the Hoeffding decomposition of the empirical excess risk. Let $\Lambda(r)$ be the excess risk of a pairwise ranking function r:

$$h_r(x, y) = \mathbb{E}(\mathbb{I}((y - Y)r(x, X) < 0) - \mathbb{I}((y - Y)r^*(x, X) < 0)) - \Lambda(r) \quad (15)$$

2.3.1 Main results

Theorem 5 gives conditions under which we can achieve fast rates of convergence. The proof relies on 3 pieces:

- The Hoeffding decomposition of the empirical excess risk into a sum of iid terms and a degenerate U-statistic of order 2.
- A moment inequality for U-processes of order 2 (Theorem 11) to show that the degenerate component of the decomposition is small with high probability.
- A bound on the excess risk in the iid case (Theorem 8.3 from Massart [10]).[7]

Note that the moment inequality of Theorem 11 is largely based on the moment inequalities from Boucheron et al. [5]. The proof is essentially application of the bounds present there in addition to a few applications of decoupling and un-decoupling steps (see de la Peña et al. chapter 3 [7]).

The bound in Theorem 5 is stated in terms of expectations of Rademacher averages and chaoses; the corollaries show some examples on how to bound these terms for the case of VC classes.

Questions

- Can we extend the results of Clemencon et al. [6] to listwise ranking with aggregation (as framed by Duchi)? The next step would be to apply a Hoeffding decomposition to the empirical risk 7 for $k = 3$ and see if the terms can be controlled similarly.
  - First we must determine an analogous condition to Assumption 4 from [6] for the multi-partite case.
- What is the next step in deriving fast rates for the case of bipartite ranking with a surrogate? Can Proposition 12 from Clèmençon et al. [6] be tightened?
  - Clèmençon et al. mention that Blanchard et al. [3] have explored this for binary classification.

[7] It's still unclear to me how exactly Theorem 8.3 of Massart is applied. Is the moment bound used directly? In that case, how do we get δ in the statement of theorem 5 from $\delta/2$ in the proof? If the tail bound is used, then isnt Theorem 5 missing a constant term ($\kappa\epsilon_*^2$ in Massart)?

- How can we unify the notions of aggregation from Duchi and normalization from Ravikumar?
- Is the model selection problem more interesting? In particular for the case of listwise ranking in an online setting.
- From the conclusion of Duchi: investigate aggregation functions that yield non-point distributions as $k \to \infty$. For example, scaling the Thurstone–Mosteller least-squares solutions by $\sqrt{k}$ to achieve asymptotic normality could lead to more robust solutions.
- Ravikumar et al. [1] stress the importance of normalizing the relevance labels by the maximum possible NDCG to ensure 'robustness' to less likely but influential samples. Can this normalization be estimated with an EM type or other iterative approach?

References

[1] On NDCG Consistency of Listwise Ranking Methods.

[2] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[3] Gilles Blanchard, Gabor Lugosi, and Nicola Vayatis. On the rates of convergence of regularized boosting classifiers.

[4] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[5] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560, 2005.

[6] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[7] Víctor H de la Peña and Evarist Giné. Decoupling: From Dependence to Independence. 1999.

[8] John Duchi. On the Consistency of Ranking Algorithms.

[9] John C. Duchi, Lester Mackey, and Michael I. Jordan. The asymptotics of ranking algorithms. *The Annals of Statistics*, 41(5):2292–2323, 2013.

[10] Pascal Massart. Concentration Inequalities and Model Selection. 2007.