Empirical Bernstein Inequalities for U-Statistics

Authored by:

Liva Ralaivola
Thomas Peel
Sandrine Anthoine

Abstract

We present original empirical Bernstein inequalities for U-statistics with bounded symmetric kernels q. They are expressed with respect to empirical estimates of either the variance of q or the conditional variance that appears in the Bernstein-type inequality for U-statistics derived by Arcones [2]. Our result subsumes other existing empirical Bernstein inequalities, as it reduces to them when U-statistics of order 1 are considered. In addition, it is based on a rather direct argument using two applications of the same (non-empirical) Bernstein inequality for U-statistics. We discuss potential applications of our new inequalities, especially in the realm of learning ranking/scoring functions. In the process, we exhibit an efficient procedure to compute the variance estimates for the special case of bipartite ranking that rests on a sorting argument. We also argue that our results may provide test set bounds and particularly interesting empirical racing algorithms for the problem of online learning of scoring functions.

1 Paper Body

The motivation of the present work lies in the growing interest of the machine learning community for learning tasks that are richer than now well-studied classification and regression. Among those, we especially have in mind the task of ranking, where one is interested in learning a ranking function capable of predicting an accurate ordering of objects according to some attached relevance information. Tackling such problems generally implies the use of loss functions other than the 0-1 misclassification loss such as, for example, a misranking loss [6] or a surrogate thereof. For (x, y) and (x0 , y 0) two pairs from some space Z := X ? Y (e.g., X = Rd and Y = R) the misranking loss ‘rank and a surrogate convex loss ‘sur may be defined for a scoring function f ? Y X as: ‘rank (f, (x, y), (x0 , y 0)) := 1{(y?y0)(f (x)?f (x0))¡0} , sur
0
0

0
(1) 0
2
‘ (f, (x, y), (x , y)) := (1 ? (y ? y)(f (x) ? f (x)) .
(2)

Given such losses or, more generally, a loss ‘ : Y X ? Z ? Z ? R, and a training sample Z n = {(Xi , Yi)}ni=1 of independent copies of some random variable Z := (X, Y) distributed according to D, the learning task is to derive a function f ? X Y such that the expected risk R‘ (f) of f R‘ (f) := EZ,Z 0 ?D ‘(f, Z, Z 0) = EZ,Z 0 ?D ‘(f, (X, Y), (X 0 , Y 0)) 1

? ‘ (f, Z) is as small as possible. In practice, this naturally brings up the empirical estimate R n X 1 ? ‘ (f, Z) := R ‘(f, (Xi , Yi), (Xj , Yj)), n n(n ? 1)
(3)
i6=j

which is a U-statistic [6, 10]. ? ‘ (f, Z) is related to R‘ (f) and, more An important question is to precisely characterize how R n ? ‘ (f, Z) specifically, one may want to derive an upper bound on R‘ (f) that is expressed in terms of R n and other quantities such as a measure of the capacity of the class of functions f belongs to and the size n of Z n ? in other words, we may talk about generalization bounds [4]. Pivotal tools to perform such analysis are tail/concentration inequalities, which say how probable it is for a function of several independent variables to deviate from its expectation; of course, the sharper the concentration inequalities the more accurate the characterization of the relation between the empirical estimate and its expectation. It is therefore of the utmost importance to have at hand tail inequalities that are sharp; it is just as important that these inequalities rely as much as possible on empirical quantities. Here, we propose new empirical Bernstein inequalities for U-statistics. As indicated by the name (i) our results are Bernstein-type inequalities and therefore make use of information on the variance of the variables under consideration, (ii) instead of resting on some assumed knowledge about this variance, they only rely on empirical related quantities and (iii) they apply to U-statistics. Our new inequalities generalize those of [3] and [13], which also feature points (i) and (ii) (but not (iii)), while based on simple arguments. To the best of our knowledge, these are the first results that fulfill (i), (ii) and (iii); they may give rise to a few applications, of which we describe two in the sequel. The paper is organized as follows. Section 2 introduces the notations and briefly recalls the basics of U-statistics as well as tail inequalities our results are based upon. Our empirical Bernstein inequalities are presented in Section 3; we also provide an efficient way of computing the empirical variance when the U-statistics considered are based on the misranking loss ‘rank of (1). Section 4 discusses two applications of our new results: test set bounds for bipartite ranking and online ranking.

2 2.1

Background Notation

The following notation will hold from here on. Z is a random variable of distribution D taking values in Z := X ? Y; Z 0 , Z1 , . . . , Zn are independent copies of Z and Z n := {Zi = (Xi , Yi)}ni=1 and Z p:q := {Zi }qi=p . m Am

n denotes the set An := {(i1 , . . . , im) : 1 ? i1 6= . . . 6= im ? n} , with 0 ? m ? n.

Finally, a function q : Z m ? R is said to be symmetric if the value of q(z) = q(z1 , . . . , zm) is independent of the order of the zi ?s in z. 2.2

U-statistics and Tail Inequalities

?q (Z) defined as Definition 1 (U-statistic, Hoeffding [10]). The random variable U n X 1 ?q (Z) := U q(Zi1 , . . . , Zim), n —Am n— m i?An

is a U-statistic of order m with kernel q, when q : Z m ? R is a measurable function on Z m . ?q (Z); in addition, EZ U ?q (Z) is a lowest Remark 1. Obviously, EZ m q(Z1 , . . . , Zm) = EZ n U n n n variance estimate of EZ m q(Z1 , . . . , Zm) based on Z n [10]. Also, reusing some notation from the ? ‘ (f, Z) of Eq. (3) is a U-statistic of order 2 with kernel qf (Z, Z 0) := ‘(f, Z, Z 0). introduction, R n Remark 2. Two peculiarities of U-statistics that entail a special care are the following: (i) they are sums of identically distributed but dependent variables: special tools need be resorted to in order ?q (Z) from Eq, and (ii) from an to deal with these dependencies to characterize the deviation of U n algorithmic point of view, their direct computations may be expensive, as it scales as O(nm); in Section 3, we show for the special case of bipartite ranking how this complexity can be reduced. 2

m=2 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0 101
m=10 ?Bernstein ?Hoeffding
102
103
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
m=2 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0 101
?Bernstein ?Hoeffding
102
103
104
m=10 ?Bernstein ?Hoeffding
102
103
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0 101
?Bernstein ?Hoeffding
102
103

Figure 1: First two plots: values of the right-hand size of (5) and (6), for Duni and kernel qm for m = 2 and m = 10 (see Example 1) as functions of n. Last two plots: same for DBer (0.15). We now recall three tail inequalities (Eq. (5), (6), (7)) that hold for U-statistics with symmetric and bounded kernels q. Normally, these inequalities make explicit use of the length qmax ? qmin of the range [qmin , qmax] of q. To simplify the reading, we will consider without loss of generality that q has range [0, 1] (an easy way of retrieving the results for bounded q is to consider q/kqk?). One key quantity that appears in the original versions of tail inequalities (5) and (6) below is bn/mc, the integer part of the ratio n/m ? this quantity might be thought of as the effective number of

data. To simplify the notation, we will assume that n is a multiple of m and, therefore, bn/mc = (n/m). ?q , [11].). Hoeffding proved the following: Theorem 1 (First order tail inequality for U

n o

?q (Z 0) ? U ?q (Z) ? ? ? 2 exp ?(n/m)?2 , ?? ¿ 0, PZ n EZ 0n U (4) n n Hence ?? ? (0, 1], with probability at least 1 ? ? over the random draw of Z n : s

2 1 0 ?

0 ?q (Z) ? ln . EZ n Uq (Z n) ? U n (n/m) ?

(5)

To go from the tail inequality (4) to the bound version (5), it suffices to make use of the elementary inequality reversal lemma (Lemma 1) provided in section 3, used also for the bounds given below. ?q , [2, 11]). Hoeffding [11] and, later, Arcones [2] refined Theorem 2 (Bernstein Inequalities for U the previous result in the form of Bernstein-type inequalities of the form

n o 2

?q (Z 0) ? U ?q (Z) ? ? ? a exp ? (n/m)? ?? ¿ 0, PZ n EZ 0n U , n n 2?q,m + bm ? For Hoeffding, a = 2, ?q,m = ?2q where, ?2q is the variance of q(Z1 , . . . , Zm) and bm = 2/3. Hence, ?? ? (0, 1], with probability at least 1 ? ?: s

2?2q 2 2 2 0 ?

0 ?q (Z) ? ln + ln . (6) EZ n Uq (Z n) ? U n (n/m) ? 3(n/m) ? For Arcones, a = 4, ?q,m = m?q2 where ?q2 is the variance of EZ2 ,...,Zm q(Z1 , Z2 , . . . , Zm) (this is a function of Z1) and bm = 2m+3 mm?1 + (2/3)m?2 . ?? ? (0, 1], with probability at least 1 ? ?: s

2m?q2 4 bm 4 0 ?

0 ?q (Z) ? ln + ln . (7) EZ n Uq (Z n) ? U n (n/m) ? (n/m) ? With a slight abuse, we will now refer to Eq. (5), (6) and (7) as tail inequalities. In essence, these ?q (Z). are confidence intervals at level 1 ? ? for EZ m q(Z m) = EZ n U n Remark 3. Eq. (7) is based on the so-called Hoeffding decomposition of U-statistics [11]. It provides a more accurate Bernstein-type inequality than that of Eq. (6), as m?q2 is known to be smaller than ?2q (see [16]). However, for moderate values of n/m (e.g. n/m ¡ 105) and reasonable values of ? (e.g. ? = 0.05), the influence of the log terms might be such that the advantage of (7) over (6) goes unnoticed. Thus, we detail our results focusing on an empirical version of (6). Example 1. To illustrate how the of the variance information provides smaller confidence interQuse m vals, consider the kernel qm := i=1 zi and two distributions Duni and DBer (p). Duni is the uniform distribution on [0, 1], for which ?2 = 31m ? 41m . DBer (p) is the Bernoulli distribution with parameter p ? [0, 1], for which ?2 = pm (1 ? pm). Figure 1 shows the behaviors of (6) and (5) for various values of m as functions of n. Observe that the variance information renders the bound smaller. 3

3

Main Results

This section presents the main results of the paper. We first introduce the inequality reversal lemma, which allows to transform tail inequalities into upper

bounds (or confidence intervals), as in (5)-(7). Lemma 1 (Inequality Reversal lemma). Let X be a random variable and a, b ¿ 0, c, d ? 0 such that

b?2 ?? ¿ 0, PX (—X— ? ?) ? a exp ? , (8) c + d? then, with probability at least 1 ? ? r —X— ?

c a d a ln + ln . b ? b ?

(9)

Proof. Solving for ? such that the right hand side of (8) is equal to ? gives: r

1 a a a ?= d ln + d2 ln2 + 4bc ln . 2b ? ? ? ? ? ? Using a + b ? a + b gives an upper bound on ? and provides the result. 3.1

Empirical Bernstein Inequalities

Let us now define the empirical variances we will use in our main result. ? 2 be the U-statistic of order 2m defined as: Definition 2. Let ? q

? 2q (Z) := ? n

X 2 1 q(Zi1 , . . . , Zim) ? q(Zim+1 , . . . , Zi2m) , 2m —An — 2m

(10)

i?An

and ? ?q2 be the U-statistic of order 2m ? 1 defined as: ? ?q2 (Z n) :=

1 —A2m?1 — n

X

q(Zi1 , Zi2 , . . . , Zim)q(Zi1 , Zim+1 , . . . , Zi2m?1),

(11)

i?A2m?1 n

It is straightforward to see that (cf. the definitions of ?2q in (6) and ?q2 in (7)) ? 2q (Z) = ?2q , EZ n ? n

and

EZ n ??q2 (Z n) = ?q2 + E2Z m q(Z1 , . . . , Zm).

We have the following main result. Theorem 3 (Empirical Bernstein Inequalities/Bounds). With probability at least 1 ? ? over Z n , s

? 2q 2? 4 5 4

0 ? 0 ?q (Z) ? ln + ln . (12) EZ n Uq (Z n) ? U n (n/m) ? (n/m) ? And, also, with probability at least 1 ? ?, (bm is the same as in (7)) s ?

2m? ?q2 8 5 m + bm 8 0 ?

0 ? E) ? U (Z) ? ln + ln . U (Z Zn q n q n (n/m) ? (n/m) ?

(13)

Proof. We provide the proof of (12) for the upper bound of the confidence interval; the same reasoning carries over to prove the lower bound. The proof of (13) is very similar. ? 2q : First, let us call Q the kernel of ? 2

Q(Z1 , . . . , Z2m) := (q(Z1 , . . . , Zm) ? q(Zm+1 , . . . , Z2m)) . 4

Q is of order 2m, has range [0, 1] but it is not necessarily symmetric. An equivalent symmetric ? 2q is Qsym : kernel for ? X 2 1 Qsym (Z1 , . . . , Z2m) := q(Z?(1) , . . . , Z?(m)) ? q(Z?(m+1) , . . . , Z?(2m)) (2m)! ??P2m

where Pm is the set of all the permutations over {1, . . . , m}. This kernel is symmetric (and has range [0, 1]) and Theorem 2 can be applied to bound ?2 as follows: with prob. at least 1 ? ? s 2 2 0 0 2 2 2 ? (Z) ? ? ? (Z) + 2V(Qsym

) ln 2 + ? = EZ 02m Qsym (Z 2m) = EZ 0n ? ln , q q n n (n/2m) ? 3(n/2m) ? where V(Qsym) is the variance of Qsym . As Qsym has range [0, 1],

V(Qsym) = EQ2sym ? E2 Qsym ? EQ2sym ? EQsym = ?2 , and therefore

s 2

? ?

? 2q (Z) ? n

+

2 4 2 4?2 ln + ln . (n/m) ? 3(n/m) ?

(To establish (13) we additionally use ? ?q2 (Z n) ? ?q2). ? 2 p Following the approach of [13], we introduce ?2 ? (m/n) ln(2/?) and we get ? 2 p 2 7 ? 2q (Z) + ln , ?2 ? (m/n) ln(2/?) ? ? n 3(n/m) ? p ? ? ? and taking the square root of both side, using 1 + 7/3 ¡ 3 and a + b ? a + b again gives s q ? 1 2 2 2 ? ? ? ?q (Z n) + 3 ln . (n/m) ? ?q (Z 0)?U ?q (Z)—, and plug in the latter equation, adjusting We now apply Theorem 2 to bound —EZ 0n U n n ? to ?/2 so the obtained inequality still holds with probability 1 ? ?. Bounding appropriate constants gives the desired result. Remark 4. In addition to providing an empirical Bernstein bound for U-statistics based on arbitrary bounded kernels, our result differs from that of Maurer and Pontil [13] by the way we derive it. Here, we apply the same tail inequality twice, taking advantage of the fact that estimates for the variances we are interested in are also U-statistics. Maurer and Pontil use a tail inequality on self bounded random variables and do not explicitly take advantage of the estimates they use being U-statistics. 3.2

Efficient Computation of the Variance Estimate for Bipartite Ranking

We have just showed how empirical Bernstein inequalities can be derived for U-statistics. The estimates that enter into play in the presented results are U-statistics with kernels of order 2m (or 2m ? 1), meaning that a direct approach to practically compute them would scale as O(n2m) (or O(n2m?1)). This scaling might be prohibitive as soon as n gets large. ? 2q (a similar reasoning carries over for Here, we propose an efficient way of evaluating the estimate ? 2 ? ?q) in the special case where Y = {?1, +1} and the kernel qf induces the misranking loss (1): qf ((x, y), (x0 , y 0)) := 1{(y?y0)(f (x)?f (x0))¡0} , ?f ? RX , which is a symmetric kernel of order m = 2 with range [0, 1]. In other words, we address the bipartite ranking problem. We have the following result. ? 2q). ?n, the computation of Proposition 1 (Efficient computation of ? f 2 X 1 ? q (zn) = 1 ? 1 ? f {(yi1 ?yi2)(f (xi1)?f (xi2))¡0} {(yi3 ?yi4)(f (xi3)?f (xi4))¡0} —A4n — 4 i?An

can be performed in O(n ln n). 5

? 2q (zn). To simplify the reading, we Proof. We simply provide an algorithmic way to compute ? f replace i1 , i2 , i3 , i4 by i, j, k, l, respectively. We also drop the normalization factor —A4n —?1 (hence the use of ? instead of = in the first line below). We have ? 2q (zn) ? ? f

X

X

(qf (zi , zj) ? qf (zk , zl))2 =

i,j,k,l i6=j6=k6=l

qf2 (zi , zj) ? 2qf (zi , zj)qf (zk , zl) + qf2 (zk , zl) ,
i,j,k,l i6=j6=k6=l
= 2(n ? 2)(n ? 3)
X
qf (zi , zj) ? 2
i,j
X
qf (zi , zj)qf (zk , zl) since
2 qf =qf qf (z,z)=0
.
i,j,k,l i6=j6=k6=l

The first term of the last line is proportional to the well-known Wilcoxon-Mann-Whitney statistic [9]. There exist efficient ways (O(n ln n)) to compute it, based on sorting the values of the f (xi)?s. We show how to deal with the second term, using sorting arguments as well. Note that !2 X
qf (zi , zj)qf (zk , zl) =
X
?4
qf (zi , zj)
i,j
i,j,k,l i6=j6=k6=l
X
qf (zi , zj)qf (zi , zk) ? 2
X
qf2 (zi , zj).
i,j
i6=j6=k

P We have subtracted from the square of i,j qf (zi , zj) all the products qf (zi , zj)qf (zk , zl) such that exactly one of the variables appears both in qf (zi , zj) and qf (zk , zl), which happens when i = k, i = l, j = k, j = l; using the symmetry of qf then provides the second term (together with the factor 4). We also have subtracted all the products qf (zi , zj)qf (zk , zl) where i = k and j = l or i = l and j = k, in which case the product reduces to qf2 (zi , zj) (hence the factor 2) ? this gives the last term. P Thus (using qf2 = qf), defining R(zn) := ij qf (zi , zj) and doing some simple calculations: ? ? X 1 2 2 ? q (zn) = ??2R (zn) + 2(n ? 5n + 8)R(zn) + 8 ? qf (zi , zj)qf (zi , zk)? f —A4n —

(14)

i6=j6=k

The only term that now requires special care is the last one (which is proportional to ? ?q2f (zn)). Recalling that qf (zi , zj) = 1{(yi ?yj)(f (xi)?f (xj))¡0} , we observe that

yi = ?1, yj = yk = +1 and f (xi) ¿ f (xj), f (xk), or qf (zi , zj)qf (zi , zk) = 1 ? yi = +1, yj = yk = ?1 and f (xi) ¡ f (xj), f (xk).

(15)

Let us define E + (i) and E ? (i) as E + (i) := {j : yj = ?1, f (xj) ¿ f (xi)} , and E ? (i) := {j : yj = +1, f (xj) ¡ f (xi)} . ? + ? and their sizes ?+ i := —E (i)—, and ?i := —E (i)—.

For i such that yi = 1, ?+ i is the number of negative instances that have been scored higher than xi by f . From (15), we see that the contribution of i to the last term of (14) corresponds to the number ? + + ?+ i (?i ? 1) of ordered pairs of indices in E (i) (similarly for ?i , with yi = ?1). Henceforth: X

qf (zi , zj)qf (zi , zk) =

i6=j6=k

X

+ ?+ i (?i ? 1) +

i:yi =+1

X

? ?? i (?i ? 1).

i:yi =?1

A simple way to compute the first sum (on i such that yi = +1) is to sort and visit the data by descending order of scores and then to incrementally compute the ?+ i ?s and the corresponding sum: when a negative instance is encountered, ?+ is incremented by 1 and when a positive instance is i + visited, ?+ (? ? 1) is added to the current sum. An identical reasoning works for the second sum. i i ? q is therefore that of sorting the scores, which has cost O(n ln n). The cost of computing ? f

4

Applications and Discussion

Here, we mention potential applications of the new empirical inequalities we have just presented. 6

Bernstein vs Hoeffding

Banana dataset 0.5 2

?Hoeffding ?Bernstein

0.4

1

0.3

0

0.2

-1

0.1 -2

-1

0

1

0 102

2

103

104

Figure 2: Left: UCI banana dataset, data labelled +1 (?1) in red (green). Right: half the confidence interval of the Hoeffding bound and that of the empirical Bernstein bound as functions of ntest . 4.1

Test Set Bounds

A direct use of the empirical Bernstein inequalities is to draw test set bounds. In this scenario, a sample Z n is split into a training set Z train := Z 1:ntrain of ntrain data and a hold-out set Z test := Z ntrain +1:n of size ntest . Z train is used to train a model f that minimizes an empirical risk based on a U-statistic inducing loss (such as in (1) or (2)) and Z test is used to compute a confidence interval on the expected risk of f . For instance, if we consider the bipartite ranking problem, the loss is ‘rank , the corresponding kernel is qf (Z, Z 0) = ‘rank (f, Z, Z 0), and, with probability at least 1 ? ? s ? 2q (Z) ln(4/?) 4? 10 4 test f ? ‘rank (f, Z) + R‘rank (f) ? R + ln , (16) test ntest ntest ? ? 2q (Z) is naturally the empirical variance of qf computed on Z . where ? test test f Figure 2 displays the behavior of such test set bounds as ntest grows for the UCI banana dataset. To produce this plot, we have learned a linear scoring function f (?) = hw, ?i by minimizing X 2 ?kwk2 + (1 ? (Yi ? Yj)hw, Xi ? Xj i) i6=j

for ? = 1.0. Of course, a purely linear scoring function would not make it possible to achieve good ranking accuracy so we in fact work in the reproducing kernel hilbert space associated with the Gaussian kernel k(x, x0) = exp(?kx?x0 k2 /2). We train our scoring function on ntrain = 1000 data points and evaluate the test set bound on ntest = 100, 500, 1000, 5000, 10000 data points. Figure 2 (right) reports the size of half the confidence interval of the Hoeffding bound (5) and that of the empirical Bernstein bound given in (16). Just as in the situation described in Example 1, the use of variance information gives rise to smaller confidence intervals, even for moderate sizes of test sets. 4.2

Online Ranking and Empirical Racing Algorithms

Another application that we would like to describe is online bipartite ranking. Due to space limitation, we only provide the main ideas on how we think our empirical tail inequalities and the efficient computation of the variance estimates we propose might be particularly useful in this scenario. First, let us precise what we mean by online bipartite ranking. Obviously, this means that Y = {?1, +1} and that the loss of interest is ‘rank . In addition, it means that given a training set Z = {Zi := (Xi , Yi)}ni=1 the learning procedure will process the data of Z incrementally to give rise to hypotheses f1 , f2 , . . . , fT . As ‘rank entails a kernel of order m = 2, we assume that n = 2T and that we process the data from Z pairs by pairs, i.e. (Z1 , Z2) are used to learn f1 , (Z3 , Z4) and f1 are used to learn f2 and, more generally, (Z2t?1 , Z2t) and ft?1 are used to produce ft (there exist more clever ways to handle the data but this goes out of the scope of the present paper). We do not specify any learning algorithm but we may imagine trying to minimize a penalized empirical risk based on the surrogate loss ‘sur : if linear functions f (?) = hw, ?i are considered and a penalization 7

like kwk2 is used then the optimization problem to solve is of the same form as in the batch case: X 2 ?kwk2 + (1 ? (Yi ? Yj)hw, Xi ? Xj i) , i6=j

but is solved incrementally here. Rank-1 update formulas for inverses of matrices easily provide means to incrementally solve this problem as new data arrive (this is the main reason why we have mentioned this surrogate function).

As evoked by [5], a nice feature of online learning is that the expected risk of hypothesis ft can be estimated on the n ? 2t examples of Z it was not trained on. Namely, when 2? data have been processed, there exist ? hypotheses f1 , . . . , f? and, for t ¡ ? , with probability at least 1 ? ?: s

? 2q (Z 2? 5 4 2t:2?) ln(4/?)

f ? ‘rank (ft , Z + ln . R‘rank (ft) ? R 2t:2?)) ? ? ?t ? ?t ? If one wants to have these confidence intervals to simultaneously hold for all t and all ? with probability 1 ? ?, basic computations to calculate the number of pairs (t, ?), with 1 ? t ¡ ? ? n show that it suffices to adjust ? to 4?/(n + 1)2 . Hence, with probability at least 1 ? ?: ?1 ? t ¡ ? ? n, s

? 2q (Z 4? 10 n+1 2t:2?) ln((n + 1)/?)

f ? ‘rank (ft , Z + ln . (17) R‘rank (ft) ? R 2t:2?)) ? ? ?t ? ?t ? We would like to draw the attention of the reader on two features: one has to do with statistical considerations and the other with algorithmic ones. First, if the confidence intervals simultaneously hold for all t and all ? as in (17), it is possible, as the online learning process goes through, to discard the hypotheses ft which have their lower bound (according to (17)) on R‘rank (ft) that is higher than the upper bound (according to (17) as well) on R‘rank (ft0) for some other hypothesis ft0 . This corresponds to a racing algorithm as described in [12]. Theoretically analyzing the relevance of such a race can be easily done with the results of [14], which deal with empirical Bernstein racing, but for non-U-statistics. This full analysis will be provided in a long version of the present paper. Second, it is algorithmically possible to preserve some efficiency in computing the various variance estimates through the online learning process: these computations rely on sorting arguments, and it is possible to take advantage of structures like binary search trees such as AVL trees, that are precisely designed to efficiently maintain and update sorted lists of numbers. The remaining question is whether it is possible to have shared such structures to summarize the sorted lists of scores for various hypotheses (recall that the scores are computed on the same data). This will be the subject of further research.

5 Conclusion

We have proposed new empirical Bernstein inequalities designed for U-statistics. They generalize the empirical inequalities of [13] and [3] while they merely result from two applications of the same non-empirical tail inequality for U-statistics. We also show how, in the bipartite ranking situation, the empirical variance can be efficiently computed. We mention potential applications, with illustrative results for the case of test set bounds in the realm of bipartite ranking. In addition to the possible extensions discussed in the previous section, we wonder whether it is possible to draw similar empirical inequalities for other types of rich statistics such as, e.g., linear rank statistics [8]. Obviously, we plan to work on establishing generalization bounds derived from the new concentration inequalities presented. This would require to carefully define a sound notion of capacity for U-statistic-based classes of functions (inspired, for example, from localized Rademacher complexities). Such new bounds would be compared with those proposed in [1, 6, 7, 15] for the bipartite ranking and/or pairwise classi-

fication problems. Finally, we also plan to carry out intensive simulations ?in particular for the task of online ranking? to get even more insights on the relevance of our contribution. Acknowledgments This work is partially supported by the IST Program of the EC, under the FP7 Pascal 2 Network of Excellence, ICT-216886-NOE. LR is partially supported by the ANR project ASAP. 8

2 References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization Bounds for the Area under the ROC Curve. Journal of Machine Learning Research, 6:393?425, 2005. [2] M. A. Arcones. A bernstein-type inequality for u-statistics and u-processes. Statistics & probability letters, 22(3):239?247, 1995. [3] J.-Y. Audibert, R. Munos, and C. Szepesv?ari. Tuning bandit algorithms in stochastic environments. In ALT ?07: Proceedings of the 18th international conference on Algorithmic Learning Theory, pages 150?165, Berlin, Heidelberg, 2007. Springer-Verlag. [4] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. ESAIM. P&S, 9:323?375, 2005. [5] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of online learning algorithms. IEEE Transactions on Information Theory, 50(9):2050?2057, 2004. [6] S. Cl?emenc?on, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of u -statistics. The Annals of Statistics, 36(2):844?874, April 2008. [7] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 4:933?969, 2003. [8] J. H?ajek and Z. Sid?ak. Theory of Rank Tests. Academic Press, 1967. [9] J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology, 143(1):29?36, April 1982. [10] W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematical Statistics, 19(3):293?325, 1948. [11] W. Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301):13?30, 1963. [12] O. Maron and A. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In Adv. in Neural Information Processing Systems NIPS 93, pages 59?66, 1993. [13] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In COLT 09: Proc. of The 22nd Annual Conference on Learning Theory, 2009. [14] V. Mnih, C. Szepesv?ari, and J.-Y. Audibert. Empirical bernstein stopping. In ICML ?08: Proceedings of the 25th international conference on Machine learning, pages 672?679, New York, NY, USA, 2008. ACM. [15] C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. Journal of Machine Learning Research, 10:2193?2232, Oct 2009. [16] R. J. Serfling. Approximation theorems of mathematical statistics. J. Wiley & Sons, 1980.

9