Statistics and Probability Letters 173 (2021) 109088

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality

Michael Naaman

Berkeley Research Group, 1800 M Street NW, Washington, D.C., United States of America

ARTICLE INFO

Article history:
Received 4 May 2020
Received in revised form 23 February 2021
Accepted 28 February 2021
Available online 6 March 2021

Keywords:
Machine learning
Empirical process
Hypothesis test
Non-parametric

ABSTRACT

We derive the tight constant in the multivariate version of the Dvoretzky–Kiefer–Wolfowitz inequality. The inequality is leveraged to construct the first fully non-parametric test for multivariate probability distributions including a simple formula for the test statistic. We also generalize the test under appropriate α-mixing conditions and describe applications of the tests to machine learning and representative sampling.

1. Introduction

The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality is a "famous" result, (Van Der Vaart (1998) pg. 268): for $n \geq 1$ and $t > 0$,

$$\Pr\left(\sup_{\theta\in\mathbb{R}} |F_n(\theta) - F(\theta)| > t\right) \leq 2e^{-2nt^2} \tag{1.1}$$

where F_n is the empirical cumulative distribution function (ecdf) and F is the expectation of F_n. In fact, the inequality implies uniform almost sure, a.s., convergence of the ecdf to its expectation:

$$\sup_{\theta\in\mathbb{R}^k} |F_n(\theta) - F(\theta)| \overset{a.s.}{\to} 0$$

which is "sometimes referred to as the fundamental theorem of mathematical statistics", see (Devroye et al. (1997) pg. 193).

In Kiefer and Wolfowitz (1958), the authors argue by way of counterexample that it is impossible for the multivariate DKW inequality to have the same functional form as the univariate case up to a scale factor. However, they made an optimization mistake by insisting an inequality was actually an equality. They went on to derive the limiting distribution for this upper bound, but Adler and Brown (1986) correctly observed the inequality noting that Kiefer and Wolfowitz (1958) calculated the probability distribution of the upper bound arguing the actual probability distribution seemed "hard to calculate" (Adler and Brown, 1986 pg. 10). However, we are able to derive the limiting probability distribution, which is the same limit as the univariate case, the Kolmogorov distribution.

In this paper, we derive a multivariate version of the DKW inequality.

$$\Pr\left(\sup_{\theta\in\mathbb{R}^k} |F_n(\theta) - F(\theta)| > t\right) \leq k(n+1)e^{-2nt^2} \tag{1.2}$$

E-mail address: mnaaman@thinkbrg.com.

https://doi.org/10.1016/j.spl.2021.109088

which is valid for all $t \geq 0$. Just as Massart (1990) showed two is the smallest possible constant in the univariate case, we will show $2k$ is the best possible constant in the multivariate case, if such a constant exists. Namely, we show that $n+1$ can be replaced with 2 and the inequality will hold for all sufficiently large n, see Theorem 3.2. Devroye (1977) derived an inequality similar to Eq. (1.2), but with a constant of $2e^2(2n)^k$ instead of $k(n+1)$.

Our results also correct an error in Dudley (1966), which argues that under certain regularity conditions,

$$\sqrt{n} \sup_{\theta \in \mathbb{R}^k} |F_n(\theta) - F(\theta)| \xrightarrow{d} \Pi_{l=1}^{k} K_l$$

where the limit is the product of independent random variables with the Kolmogorov distribution. Under the same regularity conditions as Dudley (1966), we show the limit is actually the maximum of independent Kolmogorov random variables as shown in Theorem 3.1.

Non-parametric statistical tests of multivariate probability distributions have proven elusive. Justel et al. (1997) implements the Rosenblatt transformation, Rosenblatt (1952), but this only applies to absolutely continuous measures and the estimation procedure is numerically cumbersome for larger dimensions, but the problem of testing multivariate probability distributions is still relevant.

Markatou and Sofikitou (2019) argued the Kolmogorov–Smirnov distance has the desirable property of being invariant to monotone transformations and the distance maximizes the difference between the power and size of the test, but a "fundamental drawback ... of the Kolmogorov–Smirnov distance is that there is no obvious extension of the distance and methods based on it to the multivariate case" (Markatou and Sofikitou, 2019 pg. 8). We provide the extension, which includes dependent data.

The layout of the paper is to define the relevant mathematical preliminaries that are used to prove the multivariate DKW inequality. Next, we solve the discontinuous optimization problem in closed form. We apply this solution to the two dimensional counterexample given in Kiefer and Wolfowitz (1958). We use our results to derive the limiting distribution under the regularity conditions given in Dudley (1966) and derive our non-parametric test for multivariate probability distributions. Finally, we generalize our main results.

2. Mathematical preliminaries

Recall, an Iverson bracket is a function that returns one if the proposition is true and zero if the proposition is false. For example, $[1 = 0] = 0$ because $1 \neq 0$. If two functions f and g have the same domain, then whenever we write $[f = g]$, then the proposition is true if $f(x) = g(x)$ for every x in the domain. In an abuse of notation, we will refer to a sequence of vectors as x_i where the lth component of the vector is given by $x_i^{(l)}$. A minus superscript, $x_i^{(-l)}$, refers to the $k-1$-dimensional vector of components that are not $x_i^{(l)}$ for some l.

We assume all probability spaces are of the form $(\Omega, \mathcal{B}, P)$ where Ω is the sample space. The empirical cdf (cumulative distribution function) for a sequence of random vectors, x_i, is given by

$$F_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} [x_i \leq \theta]$$

where $x_i \leq t$ is true if and only if every component of x_i is no larger than the corresponding component of t. Whenever we write $F_{nl}(t^{(l)})$ or $F_l(t^{(l)})$, then we are referring to the corresponding marginal of $F_n(t)$ and $F(t)$. Let $D_n^+ = \sup_{\theta \in \mathbb{R}^k} (F_n(\theta) - EF_n(\theta))$, $D_n^- = \sup_{\theta \in \mathbb{R}^k} (EF_n(\theta) - F_n(\theta))$, and $D_n = D_n^+ \vee D_n^-$. As above, $D_n(l)$, $D_n^+(l)$, and $D_n^-(l)$ all refer to the supremum of the corresponding marginal. For any cdf, F, take

$$F(-t) = \sup_{\delta > 0} F(t - \delta)$$

so that $F(-t) = \Pr(x < t)$. Recall, $\vee$ is the maximum and $\wedge$ is the minimum. As noted in Kiefer and Wolfowitz (1958) and Massart (1990), it is only necessary to prove the continuous case as any distribution with atoms will be stochastically smaller, so we simply note in passing that we may assume F is continuous without loss of generality wherever appropriate. If X is the n by k data matrix, then $X(j)$ is the matrix that has been row sorted on the jth column so that the jth column is descending.

We will also generalize our results by relaxing the independence assumption. Following Merlevède et al. (2009), let $\{X_i\}_{t \in \mathbb{Z}, n \in \mathbb{N}}$ be a sequence of k-dimensional random vectors. Let $F_{-\infty}^{l}$ and $F_{l+\tau}^{\infty}$ be the two σ-fields of events $\sigma(\ldots, X_l)$ and $\sigma(X_{l+\tau}, \ldots)$.

The strong mixing coefficients, $\alpha(n)$, for the sequence are given by

$$\alpha(n) = \sup_{k \geq 1} \sup_{F_1 \in F_{-\infty}^{k}, F_2 \in F_{k+n}^{\infty}} |\Pr(F_1 \cap F_2) - \Pr(F_1)\Pr(F_2)|$$

and when we write $\alpha(n) = O(c_n)$, then $\alpha(n)/c_n$ is bounded by some positive real number for all sufficiently large n. In order to avoid confusion with sequence of significance levels, α_n, we will always write $\alpha(n)$ to refer to mixing coefficients.

3. Main results

In the univariate case for data with a continuous cdf, it is well known, Massart (1990), that $\sqrt{n}D_n \xrightarrow{d} K$ and $\sqrt{n}D_n^- \xrightarrow{d} W$,

$$\Pr(K > t) = 2\sum_{k}^{\infty}(-1)^{k-1}e^{-2k^2t^2} \tag{3.1}$$

$$\Pr(W > t) = e^{-2t^2} \tag{3.2}$$

where K and W are random variables that follow the Kolmogorov and Weibull distribution, respectively, and $\Pr(\sqrt{n}D_n^- > t) \leq e^{-2t^2}$ for any $t \geq \sqrt{\ln(2)/2}$.

In order to derive our results for the multivariate case, we will proceed by solving the following optimization problem,

$$D_n = \left(\sup_{\theta\in\mathbb{R}^k} F_n(\theta) - F(\theta)\right) \vee \left(-\inf_{\theta\in\mathbb{R}^k} F_n(\theta) - F(\theta)\right)$$

for arbitrary k, which results in any asymmetry between D_n^+ and D_n^-, which will turn out to be irrelevant, at least asymptotically.

Lemma 3.1. *Let x_i be a sequence of i.i.d. k-dimensional random vectors so that the data, X, is an n by k matrix. Take $X(j)$ to be the matrix that has been row sorted on the jth column in a descending fashion for $1 \leq j \leq k$. Suppose F is the cdf of the joint probability distribution, then*

$$D_n^+ = \max_{1\leq i\leq n}\max_{1\leq j\leq k}\left(F_n(z_i(j)) - F(z_i(j))\right)[nF_n(z_i(j)) = i] \tag{3.3}$$

where $z_i^{(l)}(j) = \max_{m\leq i} x_m^{(l)}(j)$ for $1 \leq i \leq n$ and $1 \leq l, j \leq k$.

Proof. First, assume F is continuous. If $D_n = 0$, then all of columns have equal elements and the result is trivially true, so assume $D_n^+ > 0$. Since we have continuous marginals, then

$$z_1(j) \leq z_2(j) \leq \cdots \leq z_n(j)$$

where none of the inequalities are equalities w.p.1. because the continuous marginal will have unique order statistics w.p.1., so all ties will be broken when we row sort the data matrix on column j. Since there are exactly i rows of $Z(j)$ that are no larger than $z_i(j)$, then $F_n(z_i) = i/n$ because z_i is no smaller than exactly i rows in $X(j)$.

Now, let us rewrite the optimization problem,

$$D_n^+ = \max_{1\leq i\leq n}\sup_{t\in\mathbb{R}^k, nF_n(t)=i}\left(\frac{i}{n} - F(t)\right)$$

and consider each optimization problem separately. For any fixed i, $(i/n - F(t))$ is maximized by making $F(t)$ as small as possible because $D_n^+ > 0$, but the constraint must also be satisfied. For any column j, we must have $t \geq z_i(j)$ or else $nF_n(t) < i$ is not feasible, so assume $t = z_i(j)$ at least for the jth column, so we just need to check the other columns,

$$D_n^+ = \max_{1\leq i\leq n}\max_{1\leq j\leq k}\left(F_n(z_i(j)) - F(z_i(j))\right)[nF_n(z_i(j)) = i] \tag{3.4}$$

which finishes the result for the continuous case. If F is not continuous, then there can be ties in $Z(j)$, so $F_n(z_i(j)) \geq i/n$, but we can impose this restriction directly with $[nF_n(z_i(j)) = i]$. This ensures we only consider the subset of rows such that $F_n(z_i(j)) = i/n$. Effectively, there will be a jump in the running maximums whenever there is a tie, so it does not matter how ties are broken, if it all. □

For continuous F, if D_n^- and D_n^+ were symmetric optimization problems we would expect,

$$D_n^- = \max_{1\leq i\leq n}\max_{1\leq j\leq k} F(z_i(j)) - F_n(z_i(j))$$

but this cannot be true because $F_j(z_i^{(j)}(j)) > F(z_i(j))$ with $F_{nj}(z_i^{(j)}(j)) = i/n$.

Lemma 3.2. *For any sequence of i.i.d. k-dimensional random variables, x_i, the*

$$\Pr\left(\max_{1\leq j\leq k} D_n^-(j) > t\right) = \Pr\left(\sup_{\theta\in\mathbb{R}^k}\left(F(\theta) - F_n(\theta)\right) > t\right) \leq ke^{-2nt^2} \tag{3.5}$$

for every $t \geq 0$, $n \geq 1$ and $k \geq 2$.

Proof. Assume F is continuous without loss of generality, and rewrite the optimization problem,

$$D_n^- = \max_{1 \le i \le n} \sup_{t \in \mathbb{R}^k, nF_n(t)=i} \left(F(t) - \frac{i}{n} \right)$$

where $(F(t) - i/n)$ is now maximized by making t as large as possible while satisfying the constraint for any fixed i. Now, fix a column, j, then we must have $[t > z_{i+1}(j)] = 0$ or else $F_n(t) > i/n$, so choose some $\delta > 0$ sufficiently small so that $z_{i+1} - \delta > z_i$, but this gives

$$F_j(z_{i+1}^{(j)}(j) - \delta) > F(z_{i+1}(j) - \delta)$$

with $F_{nj}(z_{i+1}^{(j)}(j) - \delta) = F_n(z_i(j) - \delta) = i/n$, which gives

$$F_j(-z_{i+1}^{(j)}(j)) - i/n > F(-z_{i+1}(j)) - i/n$$

for fixed i and j. If there is an $l \neq j$ such that $F_{nl}(-z_{i+1}^{(l)}(j)) = i/n$, then this case will be covered when we do the same calculation for column l, $z_i(k)$. Since the jth marginal is equivalent to setting $t^l = \infty$ for $l \neq j$ and we must have $[t > z_{i+1}(j)] = 0$, then the supremum can be calculated by checking each marginal. Letting $\delta \to 0$, we have

$$D_n^- = \sup_{t \in \mathbb{R}^k} F(-t) - F_n(-t) = \max_{0 \le i \le n-1} \max_{1 \le j \le k} F_j(-z_{i+1}^{(j)}(j)) - F_{nj}(-z_{i+1}^{(j)}(j))$$

and $z_i^{(j)}(j)$ is simply the ith order statistic of the jth column of the data matrix $X(j)$, which gives $D_n^- = \max_{1 \le j \le k} D_n^-(j)$. An upper bound can be constructed with Boole's inequality

$$\Pr(D_n^- > t) \le \sum_{j=1}^{k} \Pr(D_n^-(j) > t)$$

but $\Pr(D_n^-(j) > t) \le e^{-2nt^2}$ for $t > \sqrt{\ln(2)/2n}$, see Massart (1990). Nevertheless, ke^{-2nt^2} is an upper bound for all t because $ke^{-2nt^2} \ge 1$ for $t \le \sqrt{\ln(2)/2n}$ and $k > 1$. □

Dudley (1966) asserted that whenever the columns of the data matrix are independent, independent components case, then

$$\sqrt{n} D_n \xrightarrow{d} \Pi_{l=1}^{k} K_l$$

where $\Pi_{l=1}^{k} K_i$ is the product of independent Kolmogorov random variables. However, this seems suspicious because $\sqrt{n} D_n \ge \max_j \sqrt{n} D_n(j) \xrightarrow{d} \max_l K_l$, even when the components are not independent.

Theorem 3.1. *For any sequence of i.i.d. k-dimensional random variables with continuous marginals,*

$$\sqrt{n} \sup_{\theta \in \mathbb{R}^k} |F_n(\theta) - F(\theta)| \xrightarrow{d} \max_{1 \le j \le k} K_j \tag{3.6}$$

$$\Pr(\max_{1 \le j \le k} K_j) < 2ke^{-2t^2} \tag{3.7}$$

for $t \ge 0$ *and* $\Pr(K_j > t) = 1 - F_K(t)$.

Proof. Starting with Eq. (3.7),

$$\Pr(\max_{1 \le j \le k} K_j > t) \le \sum_{j=1}^{k} \Pr(K_j > t)$$

by Boole's inequality because the probability measures the likelihood that at least one K_j is greater than t. In fact, the inequality is strict because equality can only hold in the pairwise mutually exclusive case, e.g. all of the circles in a Venn diagram must be disjoint.

By Slutsky's theorem,

$$\max_{1 \le j \le k} D_n^+(j)/D^+ \xrightarrow{p} 1$$

is a sufficient condition for Eq. (3.6). Since $F_{nj}(z_i^{(j)}(j)) = i/n$, then it follows from Lemma 3.1 that

$$D_n^+ = \max_{1 \le i \le n} \max_{1 \le j \le k} \left(F_{nj}(z_i^{(j)}(j)) - F(z_i(j)) \right) \tag{3.8}$$

where $z_i^{(l)}(j) = \max_{m \le i} x_m^{(l)}(j)$ for $1 \le i \le n$ and $1 \le l, j \le k$. Now, make a change of variable so that index i refers to the ith row of the original data matrix X. Namely, the new variable $\tilde{z}_{in}$ satisfies the following two properties

$$\tilde{z}_{in}^{(j)}(j) = x_i^{(j)}$$
$$\tilde{z}_{in}(j) = z_l(j)$$

for some $1 \le l \le k$. The transformation is a permutation of the i index that represents the original location in the data matrix, $x_i^{(j)}$, is the component in the ith row and jth column of the unsorted data matrix, X.

Even though, $\tilde{z}_{in}^{(j)}(j) = x_i^{(j)}$ is fixed as more data rows are added, the other components will not be. In fact, whenever a new data row is added, there is a non-zero probability, $F_j(x_i^{(j)})$, that the new observation will be less than $x_i^{(j)}$. Whenever this event occurs, then $\tilde{z}_{in+1}^{(-j)}(j) = \tilde{z}_{in}^{(-j)}(j) \vee x_{n+1}^{(-j)}$ and $\tilde{z}_{in+1}^{(-j)}(j) = \tilde{z}_{in}^{(-j)}(j)$ if $x_{n+1}^{(j)} > x_i^{(j)}$, so we must have $F(x_i^{(j)}, \tilde{z}_{in}^{(-j)}(j)) \xrightarrow{p} F_j(x_i^{(j)})$.

Now, rewrite D_n^+ in terms of the transformed variables, $\tilde{z}_{in}^{(j)}(j)$,

$$D_n^+ = \max_{1 \le i \le n} \max_{1 \le j \le k} \left(F_{nj}(x_i^{(j)}) - F(\tilde{z}_{in}(j)) \right) = F_{nj_n}(x_{i_n}^{(j_n)}) - F(\tilde{z}_{i_n n}(j_n))$$

where $F_{nj_n}(x_{i_n}^{(j_n)}) - F(\tilde{z}_{i_n n}(j_n)) \le \max_{1 \le j \le k} D_n^+(j)$ is some sequence of maximizers. Since $F(\tilde{z}_{i_n n}(j_n))/F_{j_n}(x_{i_n}^{(j_n)}) \xrightarrow{p} 1$, then $max_{1 \le j \le k} D_n(j)/D_n \xrightarrow{p} 1$. □

Theorem 3.1 can be used to calculate the limiting probability distribution whenever the components are independent,

$$\lim_{n \to +\infty} \Pr(\sqrt{n} D_n > t) = 1 - (F_K(t))^k \le 1 - (R(1 - 2e^{-2t^2}))^k$$

where R is the ramp function, F_k is the Kolmogorov cdf, and the inequality follows from $1 - F_K(t) \le 2e^{-2t^2}$. The term after the inequality, $1 - (R(1 - 2e^{-2t^2}))^k$, might be a least upper bound in the space of absolutely continuous measures.

Since $\left(1 - 2e^{-2t^2}\right)^k = 1 - 2ke^{-2t^2}(1 + o(e^{-t^2}))$, then

$$\frac{\Pr\left(\sqrt{n} \sup_{\theta \in \mathbb{R}^k} |F_n(\theta) - F(\theta)| > t\right)}{2ke^{-2t^2}} \to 1$$

as $t \to \infty$ whenever the components are also independent. Since any constant less than $2k$ will not be possible for some sufficiently large n, then $2k$ is the smallest possible constant for the multivariate DKW inequality, which generalizes the result in Massart (1990).

Theorem 3.2. *For any sequence of i.i.d. k-dimensional random variables, x_i with $k > 1$, then*

$$\Pr(D_n > t) < 2ke^{-2nt^2} \tag{3.9}$$

for all $t \ge 0$ and n sufficiently large.

Proof. Without loss of generality we can restrict attention to the continuous case. By Theorem 3.1, $\lim \Pr(D_n > t) < 2ke^{-2t^2}$. Since the limit inequality is strict, there is some finite n, that works for every $t \in (0, 1)$. This means there exists some $N_*(F)$ such that

$$\Pr(D_n > t) < 2ke^{-2nt^2} \tag{3.10}$$

for all $n \ge N_*(F)$. □

Now, we show how our method works for the flawed counterexample in Kiefer and Wolfowitz (1958). Suppose u_i is i.i.d. uniform on $[0, 1]$ and take

$$G_n(s, t) = \sum_{i=1}^{n} [u_i \le s][1 - u_i \le t]$$
$$G(s, t) = \Pr(u \le s, 1 - u \le t)$$

and it immediately follows that we have $D_n(1) \overset{w.p.1.}{=} D_n(2)$. In Kiefer and Wolfowitz (1958), the authors argue that

$$\sup_{s,t} |G_n(s, t) - G(s, t)| = D_n^+(1) + D_n^-(1) \tag{3.11}$$

where $D_n^+(1) = \sup_x G_{n1} - x$ and $D_n^-(1) = \sup x - G_{n1}(-x)$; however, the equality should have been an inequality. Their argument is completed by noting the following

$$\lim \Pr\left(\sqrt{n}\left(D_n^+ + D_n^-\right) > r\right) \approxeq 8r^2e^{-2r^2}$$

for large r, which implies there is no simple version of the multivariate DKW inequality; this incorrect result is even in textbooks, see (Serfling (1980) pg. 61).

Now, we prove that the counterexample fails. Since the uniform cdf is continuous, then

$$u_{(1)} < u_{(2)} < \cdots < u_{(n)}$$

where the order statistics are strict with probability one. One may take z_i to be

$$z_i(1) = (u_{(i)}, 1 - u_{(1)})$$
$$z_i(2) = (u_{(n)}, 1 - u_{(n+1-i)})$$

which gives

$$D_n^+ = \max_{1\leq i\leq n}(G_n(z_i(1)) - G(z_i(1))) \vee (G_n(z_i(2)) - G(z_i(2)))$$

which can be rewritten as $D_n^+ = (D_n^+(1) + u_{(1)}) \vee (D_n^+(2) + 1 - u_{(n)})$. Since $D_n^-(2) = \sup_s(1 - G_{2n}(-s)) - (1 - G_2(s))$, then $D_n^-(2) = D_n^+(1)$ resulting in

$$\sup_{s,t} |G_n(s, t) - G(s, t)| \xrightarrow{d} K$$

where K is a Kolmogorov random variable. This means the equality in Kiefer and Wolfowitz (1958) and Eq. (3.11) does not hold. Moreover, the probability distribution in the example corresponds to the smallest possible copula, $R(u + v - 1)$, which is a singular measure concentrated on a 1-dimensional subspace.

Since the marginals do not uniquely identify the joint distribution, D_n^- cannot be used as a multivariate test, but a test can be constructed because D_n^+ will become relevant under the alternative as it will detect differences between copulas.

Theorem 3.3. *Suppose the desired significance level is α, then take $d(\alpha, k) = \sqrt{-\ln(\frac{\alpha}{2k})/2}$. For any sequence of i.i.d. k-dimensional random variables, x_i with cdf F, then*

$$\Pr\left(\sqrt{n} \sup_{\theta\in\mathbb{R}^k} |F_n(\theta) - G(\theta)| > d(\alpha, k)\right) \leq \alpha \tag{3.12}$$

for sufficiently large n whenever the null is true, $\sup_\theta |G(\theta) - F(\theta)| = 0$, and $\sqrt{n}D_n \overset{a.s.}{\to} +\infty$ under the alternative, $\sup_\theta |G(\theta) - F(\theta)| > 0$.

Proof. The case when the null is true follows from Theorem 3.2. Under the alternative, $\sqrt{n}D_n^+ \overset{a.s.}{\to} +\infty$ unless $G(\theta) \geq F(\theta)$ for all $\theta \in \mathbb{R}^k$, but in this case at least one of the marginals of G will be different and we will have $\sqrt{n}D_n^- \overset{a.s.}{\to} +\infty$. □

4. Generalizations

The solution of the discontinuous optimization problem has nothing to do with the underlying probability measures, which can be leveraged to generalize the DKW inequality. The key insight is that the supremum always occurs on a discrete set, which is developed below.

Lemma 4.1. *For any sequence of independent k-dimensional random variables, x_i,*

$$\Pr\left(\sup_{\theta\in\mathbb{R}^k} |F_n(\theta) - EF_n(\theta)| > t\right) \leq k(n + 1)e^{-2nt^2} \tag{4.1}$$

for $t \geq 0$ and every $n, k \geq 1$.

Proof. Since $\Pr(\sqrt{n}D_n > t) \leq P(\sqrt{n}D_n^+ > t) + ke^{-2t^2}$, then it will be sufficient to show that

$$\Pr\left(\max_{1\leq i\leq n} \max_{1\leq j\leq k} (F_n(z_i(j)) - EF_n(z_i(j))) > t\right) \leq nke^{-2nt^2}$$

for $t > 0$ and we can assume the cdf, F, is continuous without loss of generality.

Let $s = (s_1, s_2, \ldots)$ be an arbitrary sequence of k-dimensional vectors with components that are extended real numbers and assume there are k versions of s, which are indexed by $s_i(j)$. Since s_i is fixed we have n averages that are converging to zero almost surely, so

$$F_n(s_i) - EF_n(s_i) = \frac{1}{n}\sum_{j=1}^{n} \left[x_j \leq s_i\right] - E\left[x_j \leq s_i\right]$$

where $1 \leq i \leq n$, which is the average of independent random variables. By Boole's inequality,

$$\Pr\left(\max_{1\leq i\leq n}\max_{1\leq j\leq k}\left(F_n(s_i(j)) - EF_n(s_i(j))\right) > t\right) \leq nk\Pr\left(\left(F_n(s_i(j)) - EF_n(s_i(j))\right) > t\right)$$

for any (i, j) pair, which results in

$$\Pr\left(\max_{1\leq i\leq n}\max_{1\leq j\leq k}\left(F_n(s_i(j)) - EF_n(s_i(j))\right) > t\right) \leq nke^{-2nt^2}$$

by Hoeffding's inequality, Hoeffding (1963). Since the inequality holds for any sequence of vectors, then it also holds for any realization of the data sequence, $x = (x_1, x_2, \ldots)$. □

A subtle consequence of Lemma 4.1 is that it is still valid when $k >> n$, so statistical tests are still available. Candes and Tao (2007) discuss this important case in the context of the Dantzig selector.

Note, $EF_n(t)$ in Lemma 4.1 can be replaced with any cdf, G_n, as long as $[EF_n = EG_n] = 1$ for all n. For example, one may replace EF_n with $\prod_{j=1}^{k} F_{nj}$ to construct a test for independent coordinates.

More generally, the optimization results, Lemmas 3.1 and 3.2, can be used to measure the distance between any two data sets over some set of k columns. This is particularly useful in machine learning where data sets are split into a training group and a comparison group.

If the cdf, F, is continuous, then Lemma 4.1 follows almost immediately from an observation made in Dvoretzky et al. (1956, pg. 644): taking the supremum over the reals, $\mathbb{R}^k$, is the same as taking the supremum over the rationals, which is countable, giving

$$\Pr(D_n > t) = \sup_{k\in\mathbb{N}, t_k\in\mathbb{R}^k}\Pr\left(\max_{1\leq i\leq k}\left|F_n\left(t_k^{(i)}\right) - F\left(t_k^{(i)}\right)\right| > t\right)$$

for $t > 0$ and D_n refers to the univariate case. However, this bound can be improved upon using our results because we know that the supremum occurs at some element in a set with n points, so we must also have

$$\Pr(D_n > t) = \sup_{t_n\in\mathbb{R}^n}\Pr\left(\max_{1\leq i\leq n}\left|F_n\left(t_n^{(i)}\right) - F\left(t_n^{(i)}\right)\right| > t\right)$$

which is the same point made in the proof of Lemma 4.1.

In the proof, an independence assumption was only needed so that Hoeffding's inequality could be applied. This allows us to generalize the test by using well known concentration inequalities, but first we need a technical lemma.

Lemma 4.2. *Let x_i be a 1-dimensional sequence of centered random variables. Suppose x_i is α-mixing with $\alpha(n) \leq e^{-2cn}$, where $c \leq 1/2$, and $\sup_i |x_i| \leq 1$. Take δ_n to be any non-decreasing, positive, integer sequence satisfying $1 \leq \delta_n < n$, then, whenever $c\delta_n \geq 1$ and $n \geq 4$,*

$$\Pr\left(\left|\sum_{i=\delta_n}^{n+\delta_n-1} x_i\right| \geq nt\right) \leq \exp\left(\frac{-nt^2}{4C_{\delta_n}\left(1+\ln n\right)}\right)$$

where $C_{\delta_n} = 14.2 + (25.8 + 2/\ln 2)\,\delta_n + 8\delta_n^2$.

Proof. It follows from Corollary 12 in Merlevède et al. (2009) that, for any $n \geq 4$,

$$\Pr\left(|S_n| \geq nt\right) \leq \exp\left(\frac{-nt^2}{4D\ln n\left(1+\frac{t}{cD\ln n}\right)}\right)$$

where $D = 6.2\,(1 + 8\alpha_*) + (1 + 2/\ln 2)\,c + 8/c^2$, $S_n = \sum_{i=1}^{n} x_i$, and $\alpha_* = \sum_{i=1}^{+\infty}\alpha(i)$. Since the bound holds whenever $t > 1$, then we can assume $t \leq 1$, so

$$\Pr\left(|S_n| \geq nt\right) \leq \exp\left(\frac{-nt^2}{4D\left(1+\ln n\right)}\right)$$

because $cD > 1$ and $t < 1$, implies $t/(cD) \leq 1$. Since $\alpha_* \leq \alpha(1) + \int_1^{\infty}\alpha(x)dx$, then we can integrate the upper bound of α, which leads to $\alpha_* \leq (1 + 1/(2c))$. The result follows by replacing x_i with x_{δ_n-1+i}, which has a mixing coefficient no larger than that of x_i. □

To clarify, $\sum_{i=\delta_n}^{n+\delta_n-1} x_i$, represents the average after the first $\delta_n - 1$ observations have been removed from the analysis. If $\delta_n = 1$, then no observations are removed, so n is the number of remaining observations after we remove the first $\delta_n - 1$, see Naaman (2016) for a simulation example.

Now, that we have a bound for averages, we can derive a DKW-type inequality that holds for any sufficiently large n.

Lemma 4.3. *Suppose x_i is a k-dimensional sequence of α-mixing random variables satisfying $\alpha(n) = O(e^{-\lambda n})$ for some $\lambda > 0$. Take δ_n to be any non-decreasing, positive, integer sequence satisfying $\sqrt{n}/\delta_n \to \infty$ and $\delta_n \to \infty$, then*

$$\Pr\left(\sup_{\theta\in\mathbb{R}^k}\left|\sum_{i=\delta_n}^{n+\delta_n-1}[x_i \leq \theta] - E\left[x_i \leq \theta\right]\right| \geq nt\right) \leq k\left(n+\delta_n\right)^2 \exp\left(\frac{-nt^2}{32\delta_n^2(1+\ln n)}\right)$$

for $t \geq 0$ and n is sufficiently large.

Proof. Take $k = 1$. If $\alpha(n) = O(e^{-\lambda n})$, then $\alpha(n) = o(e^{-2cn})$ for some $2c < \lambda$ and we must have

$$\alpha(n) \leq e^{-2cn}$$

for some $c \leq 1/2$. So the moving average eventually satisfies the mixing condition. Since $w_i = [x_i \leq \theta] - E\left[x_i \leq \theta\right]$ is centered and bounded by one with a mixing coefficient no larger than that of x_i, then the previous lemma applies,

$$\Pr\left(\sup_{\theta\in\mathbb{R}}\left|\sum_{i=\delta_n}^{n+\delta_n-1}[x_i \leq \theta] - E\left[x_i \leq \theta\right]\right| \geq nt\right) \leq$$
$$(n+\delta_n-1)\exp\left(\frac{-nt^2}{4C_{\delta_n}\left(1+\ln n\right)}\right)$$

whenever $c\delta_n \geq 1$. Since $C_{\delta_n}/8\delta_n^2 \to 1$, then we can increase the bound slightly be replacing $n+\delta_n-1$ with $n+\delta_n$ so that the result eventually holds. The multivariate case is handled along the same lines as Lemma 4.1. □

One can generalize this result to polynomial mixing rates, $\alpha(n) = O(n^{-p})$, for $p > 1$ using Eq. 2.23 in Rio (2013); however, this method does not yield an exponential bound, so it is not presented.

5. Conclusions

We have shown that $2k$ is the smallest possible constant that can exist for the multivariate DKW inequality. It would seem that such a constant exists, Theorem 3.2 holds for all n, but the proof eludes us. We have derived the first non-parametric test for multivariate probability distributions. In fact, we believe our results are particularly useful in machine learning because it allows one to measure the statistical distance between a training data set and a comparison set. One might also use the test to measure the distance between a subsample and the sample, which could be the entire data set. When this distance is small, the subsample is statistically indistinguishable from the sample and one might conclude the subsample is representative of the sample.

References

Adler, R., Brown, L., 1986. Tail behaviour for suprema of empirical processes. Ann. Probab. 14 (1), 1–30.

Candes, E., Tao, T., 2007. The dantzig selector: statistical estimation when p is much larger than n. Ann. Math. Stat. 35 (6), 2313 – 2351.

Devroye, L., 1977. A uniform bound for the deviation of empirical distribution functions. J. Multivariate Anal. 7 (4), 594–597.

Devroye, L., Györfi, L., Lugosi, G., 1997. A Probabilistic Theory of Pattern Recognition. Springer New York.

Dudley, R., 1966. Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. Illinois J. Math. 10 (1), 109–126. http://dx.doi.org/10.1215/ijm/1256055206.

Dvoretzky, A., Kiefer, J., Wolfowitz, J., 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Stat. 27 (3), 642–669. http://dx.doi.org/10.1214/aoms/1177728174.

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58 (301), 13–30.

Justel, A., Peña, D., Zamar, R., 1997. A multivariate Kolmogorov-Smirnov test of goodness of fit. Statist. Probab. Lett. 35 (3), 251–259.

Kiefer, J., Wolfowitz, J., 1958. On the deviations of the empiric distribution function of vector chance variables. Trans. Amer. Math. Soc. 87, 173–186.

Markatou, M., Sofikitou, E.M., 2019. Statistical distances and the construction of evidence functions for model adequacy. Front. Ecol. Evol. 7, 447. http://dx.doi.org/10.3389/fevo.2019.00447, URL: https://www.frontiersin.org/article/10.3389/fevo.2019.00447.

Massart, P., 1990. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab. 18 (3), 1269–1283.

Merlevède, F., Peligrad, M., Rio, E., 2009. Bernstein inequality and moderate deviations under strong mixing conditions. In: Houdré, C., Koltchinskii, V., Mason, D.M., Peligrad, M. (Eds.), High Dimensional Probability V: The Luminy Volume. In: Collections, Volume 5, Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp. 273–292. http://dx.doi.org/10.1214/09-IMSCOLL518.

Naaman, M., 2016. Almost sure hypothesis testing and a resolution of the Jeffreys-Lindley paradox. Electron. J. Statist. 10 (1), 1526–1550.

Rio, E., 2013. Inequalities and limit theorems for weakly dependent sequences. In: 3rd Cycle. France. Working Paper, pp. 1–175, URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.599.2366&rep=rep1&type=pdf.

Rosenblatt, M., 1952. Remarks on a multivariate transformation. Ann. Math. Stat. 23 (3), 470–472. http://dx.doi.org/10.1214/aoms/1177729394.

Serfling, R., 1980. Approximation Theorems of Mathematical Statistics. John Wiley Sons New York.

Van Der Vaart, A.W., 1998. Asymptotic Statistics. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.