EXPONENTIALLY CONSISTENT K-MEANS CLUSTERING ALGORITHM BASED ON KOLMOGROV-SMIRNOV TEST

Tiexing Wang[*]*, Donald J. Bucci Jr.*[†]*, Yingbin Liang*[⋆]*, Biao Chen*[*]*, Pramod K Varshney*[*]

[*]Department of EECS, Syracuse University, Syracuse, NY, 13244, USA
[†]Lockheed Martin - Advanced Technology Labs, Cherry Hill, NJ, 08002, USA
[⋆]Department of ECE, The Ohio State University, Columbus, OH 43210, USA
email:{twang17,bichen,varshney}@syr.edu, Donald.J.Bucci.Jr@lmco.com, liang.889@osu.edu

ABSTRACT

This paper studies clustering using a Kolmogorov-Smirnov based K-means algorithm. All data sequences are assumed to be generated by *unknown* continuous distributions. The pairwise KS distances of the distributions are assumed to be lower bounded by a certain positive constant. The convergence analysis of the proposed algorithms and upper bounds on the error probability are provided for both known and unknown number of clusters. More importantly, it is shown that the probability of error decays exponentially as the sample size of each data sequence goes to infinity, and the error exponent is only a function of the pairwise KS distances of the distributions. The analysis is validated by simulation results.

Index Terms— Kolmogorov-Smirnov distance, clustering, exponential consistency, probability of error, K-means algorithm.

1. INTRODUCTION

The goal of clustering is to group objects in such a way that the objects in the same cluster are similar. This paper aims to cluster sequences generated by *unknown* continuous distributions into classes based on the Kolmogorov-Smirnov (KS) distance such that each class contains all the sequences generated from the same distribution. The minimum pairwise KS distance of the distributions is assumed to be lower bounded away from 0. Furthermore, the number of distribution clusters is also of interest if it is not known *a priori*.

The unsupervised learning problem has been widely studied [1, 2]. If we view the data sequences as multivariate data, our problem can be solved by applying typical clustering methods, e.g. K-means clustering [3–5]. However, these approaches do not exploit the underlying generative model that these data sequences can possibly have in addition to being vectors, and hence the distance metric used in these approaches is mostly Euclidean distance or the distance induced by other vector norms. On the other hand, there are recent studies of anomaly detection problems [6–9], in which each data sequence consists of independently identically distributed (i.i.d.) samples generated by a certain unknown distribution. These studies exploited the generative model of the data in a similar way as the approach proposed in this paper, but with a different goal of detecting anomalous data sequences.

Among various distance functions used in clustering problems [10, 11], the KS distance/statistics is a suitable one given the data sequences. There are several works that are closely related to the approach presented here. In [12], the authors proposed an initialization method for the K-means algorithm. Given the number of clusters, at the initialization step, the first center is randomly chosen while the remaining centers are the observations that have the largest minimum distance to the previous centers. Another initialization approach is to randomly choose all the initial centers [11]. With an unknown number of clusters, the algorithm proposed in [13] assumed a maximum number of clusters. It began with one cluster containing all the observations and the cluster was split if the two-sample KS statistics between the center and any sequence exceeded the threshold determined by the significance level and the number of clusters was small. The algorithms in [11–13] were all validated by numerical results without carrying out an analysis of the probability of error.

The contribution of the present work is as follows. Given the number of clusters, the KS distance based K-means algorithm using the initialization method proposed in [12] is analyzed. With an unknown number of clusters, an algorithm capable of estimating the number of clusters and grouping the sequences is proposed and analyzed. The upper bounds on the probability of error for both cases are derived. The analysis helps establish both the convergence and the exponential consistency of the algorithms for both cases. Furthermore, meaningful bounds on the error exponents are established for both cases which turn out to be the same function of the lower bound of the pairwise KS distances of the distributions. It is also worth noting that the proposed algorithm for an unknown number of clusters works without the exact minimum pairwise KS distances of the distributions, and the analysis implies that with high probability a single iteration provides good enough estimate results given large sample size.

2. SYSTEM MODEL

2.1. Clustering Problem

Suppose there are M data sequences $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ to be clustered. Each sequence $\mathbf{x}_i$ consists of i.i.d. samples generated by one of K distinct distributions in the set $\mathcal{P} = \{p_1, \ldots, p_K\}$, and is said to belong to cluster k if it is gener-

This material is based upon work supported in part by the Defense Advanced Research Projects Agency under Contract No. HR0011-16-C-0135 and by the Dynamic Data Driven Applications Systems (DDDAS) program of AFOSR under grant number FA9550-16-1-0077.

ated by p_k. We further assume that

$$\min_{k \neq k'} d_{KS}(p_k, p_{k'}) > D_{KS}, \tag{1}$$

where $d_{KS}(p_k, p_{k'})$ is defined in (2). The goal is to group all data sequences belonging to the same cluster together.

An error occurs if and only if the sequences generated by different distributions are assigned to the same cluster, or sequences generated by the same distribution are assigned to more than one cluster. Denote by P_e the probability of error of a clustering algorithm. The algorithm is said to be consistent if $\lim_{n\to\infty} P_e = 0$, where n is the sample size. The algorithm is said to be exponentially consistent if

$$\lim_{n\to\infty} -\frac{1}{n} \log P_e > 0.$$

We are also interested in characterizing the error exponent.

2.2. Preliminaries of KS Distance

Suppose $\mathbf{x} = \{x_1, \ldots, x_n\}$ is generated by the distribution p. Then the empirical cumulative distribution function (c.d.f.) induced by $\mathbf{x}$ is given by

$$F_{\mathbf{x}}(a) = \frac{1}{n} \sum_{i=1}^{n} 1_{[-\infty, a]} x_i,$$

where $1_{[-\infty,x]}$ is the indicator function. Let the true c.d.f. of p evaluated at a be $F_p(a)$. Define the KS distance as

$$d_{KS}(\star, *) = \sup_{a \in \mathbb{R}} |F_\star(a) - F_*(a)|, \tag{2}$$

where the arguments of the function can be either sequences or distributions. A well-known upper bound on the convergence rate of the KS distance between the empirical and true c.d.f. was given by Dvoretzky-Kiefer-Wolfowitz, and later refined by Massart in [14] into the following Lemma.

Lemma 2.1. *[14] Suppose* $\mathbf{x}$ *is generated by* p *and* $F_{\mathbf{x}}(a)$ *is the corresponding empirical c.d.f.. Then*

$$P\left(\sup_{a \in \mathbb{R}} \left| F_{\mathbf{x}}(a) - F_p(a) \right| > \epsilon \right) \leq 2 \exp\big(-2n\epsilon^2 \big).$$

The following lemmas are extensions of Lemma 2.1:

Lemma 2.2. *Suppose* $\mathbf{x}$ *and* $\mathbf{z}$ *are generated by* p_1*, and* $\mathbf{y}$ *is generated by* p_2*. Then,*

$$P\left(d_{KS}(\mathbf{x}, \mathbf{z}) > d_{KS}(\mathbf{y}, \mathbf{z}) \right) \leq 6 \exp\big(-\frac{n d_{KS}^2(p_1, p_2)}{8} \big).$$

Lemma 2.3. *Let* $\mathbf{x} \sim p_1$ *and* $\mathbf{y} \sim p_2$*. Then*

$$P\left(d_{KS}(\mathbf{x}, \mathbf{y}) < \frac{d_{KS}(p_1, p_2)}{2} \right) \leq 4 \exp\big(-\frac{n d_{KS}^2(p_1, p_2)}{8} \big).$$

Lemma 2.4. *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *are generated by* p*. Then*

$$P\left(d_{KS}(\mathbf{x}, \mathbf{y}) > \epsilon \right) \leq 4 \exp\big(-\frac{n\epsilon^2}{2} \big).$$

The proofs of Lemmas 2.2 - 2.4 will be provided in a forthcoming paper. Lemmas 2.1 - 2.4 lead to a reasonable KS based K-means algorithm which is to group all the sequences that are close to each other in terms of KS distances.

3. KNOWN NUMBER OF CLUSTERS

The clustering algorithm for known K is summarized in Algorithms 1 and 2. The method proposed in [12] is used for center initialization. The initial K centers can be chosen sequentially such that the center of the k-th cluster is the sequence that has the largest minimum KS distance to the previous $k-1$ centers. Given the centers, each sequence is assigned to the cluster for which the sequence has the minimum KS distance to the center. For a cluster, a sequence is assigned as the center if the sum of its KS distances to all sequences in the cluster is the smallest. The algorithm continues until the clustering result converges.

Algorithm 1 KS-based initialization given K

Input: $\{\mathbf{x}_j\}_{j=1}^M$, number of clusters K.
Output: Partition set $\{C_k\}_{k=1}^K$.
{Center initialization}
Arbitrarily choose one $\mathbf{x}_{j_1}$ as $\mathbf{c}_1$.
for $k = 2$ to K **do**
  $\mathbf{c}_k \leftarrow \arg\max_{\mathbf{x}_j} \big(\min_{l \in \{1,\ldots,k-1\}} d_{KS}(\mathbf{x}_j, \mathbf{c}_l) \big)$
end for
{Cluster initialization}
Set $C_k \leftarrow \emptyset$ for $1 \leq k \leq K$.
for $j = 1$ to M **do**
  $C_k \leftarrow C_k \cup \{\mathbf{x}_j\}$, where

$$k = \arg \min_{k \in \{1,\ldots,K\}} d_{KS}(\mathbf{x}_j, \mathbf{c}_k)$$

end for
Return $\{C_k\}_{k=1}^K$

Algorithm 2 KS based clustering given K

Input: $\{\mathbf{x}_j\}_{j=1}^M$, number of clusters K.
Output: Partition set $\{C_k\}_{k=1}^K$.
Initialization: $\{C_k\}_{k=1}^K$ by Algorithm 1.
Method:
while not converge **do**
  {Center update}
  for $k = 1$ to K **do**
    $\mathbf{c}_k \leftarrow \arg\min_{\mathbf{x}_j \in C_k} \sum_{\mathbf{x}_{j'} \in C_k} d_{KS}(\mathbf{x}_j, \mathbf{x}_{j'})$
  end for
  {Cluster update}
  for $j = 1$ to M **do**
    if $\mathbf{x}_j \in C_{k'}$ and $d_{KS}(\mathbf{x}_j, \mathbf{c}_k) < d_{KS}(\mathbf{x}_j, \mathbf{c}_{k'})$ **then**
      $C_k \leftarrow C_k \cup \{\mathbf{x}_j\}$ and $C_{k'} \leftarrow C_{k'} \setminus \{\mathbf{x}_j\}$.
    end if
  end for
end while
Return $\{C_k\}_{k=1}^K$

Theorem 3.1. *Under Assumption* (1)*, Algorithm 2 converges after finite number of iterations and the error probability after*

T iterations is upper bounded by

$$P_e \leq 2M(K^2+3(T+1)(K-1))\exp\big(-\frac{nD_{KS}^2}{8}\big).$$

Sketch of Proof. Let C_l^t be the l-th cluster obtained at the t-th ($t>0$) iteration and $\mathbf{c}_l^t$ be the corresponding center. Moreover, let C_l^0 be the l-th cluster obtained at the initialization step and $\mathbf{c}_l^0$ be the corresponding center. Then for $t \geq 1$, we have

$$\sum_{l=1}^{K}\sum_{\mathbf{x}_j\in C_l^{t-1}} d_{KS}(\mathbf{x}_j,\mathbf{c}_l^{t-1}) \geq \sum_{l=1}^{K}\sum_{\mathbf{x}_j\in C_l^{t-1}} d_{KS}(\mathbf{x}_j,\mathbf{c}_l^{t}),$$
$$\sum_{l=1}^{K}\sum_{\mathbf{x}_j\in C_l^{t-1}} d_{KS}(\mathbf{x}_j,\mathbf{c}_l^{t}) \geq \sum_{l=1}^{K}\sum_{\mathbf{x}_j\in C_l^{t}} d_{KS}(\mathbf{x}_j,\mathbf{c}_l^{t}). \tag{3}$$

The equalities of (3) hold only if $C_l^{t-1}=C_l^t$ for all l. Therefore the algorithm converges in finite numbers of iterations.

Note that at each iteration, centers are incorrectly chosen only if the clustering result is incorrect at the most recent step. We first prove that the probability of error of Algorithm 1 is exponentially consistent. Let E_k be the event that $\mathbf{c}_k^0$ is incorrectly chosen for $k\geq 2$ while $\mathbf{c}_l^0$ for all $l\in\{1,\dots,k-1\}$ are correctly chosen. Without loss of generality, assume that $\mathbf{c}_l^0$ for $l=1,\dots,k-1$ are generated by $p_1,\dots,p_{k-1}$, respectively. Then E_k is the event that the sequence with the largest minimum KS distance to $\{\mathbf{c}_l^0\}_{l=1}^{k-1}$ is actually generated by one of the distributions in $\{p_l\}_{l=1}^{k-1}$. By Lemma 2.1, Lemma 2.3 and the union bound, the probability of error at the initialization step is bounded as

$$P\Big(\cup_{k=1}^{K} E_k\Big) \leq 2MK^2\exp\big(-\frac{nD_{KS}^2}{8}\big). \tag{4}$$

Let H^0 be the event that the error occurs at the clustering initialization step while the center initialization is correct. Then

$$H^0 \subset \cup_{l=1}^{K}\cup_{l'\neq l}\cup_{j\,:\,\mathbf{x}_j\sim p_l}\{d_{KS}(\mathbf{x}_j,\mathbf{c}_l^0)\geq d_{KS}(\mathbf{x}_j,\mathbf{c}_{l'}^0)\ :\ \mathbf{c}_l^0\sim p_l,\ \mathbf{c}_{l'}\sim p_{l'}\}.$$

Then by the union bound and Lemma 2.2, we have

$$P(H^0) \leq 6M(K-1)\exp\big(-\frac{nD_{KS}^2}{8}\big). \tag{5}$$

Let H^T be the event that incorrect clustering occurs at the T-th ($T\geq 1$) cluster update step while clustering results at the $1,\dots,(T-1)$-th steps are correct. Then $P(H^t)$ has the same upper bound as $P(H^0)$. Therefore, by (4), (5) and the union bound, we have

$$P_e = P\Big(\big(\cup_{k=1}^{K}E_k\big)\cup\big(\cup_{t=0}^{T}H^t\big)\Big)$$
$$\leq 2M(K^2+3(T+1)(K-1))\exp\big(-\frac{nD_{KS}^2}{8}\big). \quad \square$$

4. UNKNOWN NUMBER OF CLUSTERS

Besides the error that could occur in Algorithm 2, it is possible that the number of distributions is incorrectly estimated if the number of clusters is unknown. However, by Lemma 2.4, with high probability the two centers generated by the same distribution are close to each other. This is the premise of the clustering approach with unknown number of clusters, in particular, the merging of cluster centers.

The proposed approach is summarized in Algorithms 3 and 4. There are two major differences between Algorithms 3 and 4 and Algorithms 1 and 2. First, the center initialization step keeps selecting centers until all the sequences are close to the existing center. Second, an additional Merge Step in Algorithm 4 helps to combine centers that have small KS distances to each other.

Algorithm 3 KS-based initialization with unknown K

Input: $\{\mathbf{x}_j\}_{j=1}^{M}$.
Output: Partition sets $\{C_k\}_{k=1}^{\hat{K}}$.
$\{$Center initialization$\}$
Arbitrarily choose one $\mathbf{x}_{j_1}$ as $\mathbf{c}_1$ and set $\hat{K}=1$.
while $\max_{\mathbf{x}_j}\big(\min_{i\in\{1,\dots,\hat{K}\}} d_{KS}(\mathbf{x}_j,\mathbf{c}_i)\big) > \frac{D_{KS}}{2}$ **do**
  $\mathbf{c}_{\hat{K}+1} \leftarrow \arg\max_{\mathbf{x}_j}\big(\min_{i\in\{1,\dots,\hat{K}\}} d_{KS}(\mathbf{x}_j,\mathbf{c}_i)\big)$
  $\hat{K}\leftarrow\hat{K}+1$
end while
Cluster initialization specified in Algorithm 1.
Return $\{C_k\}_{k=1}^{\hat{K}}$

Theorem 4.1. *Under the assumption* (1)*, the error probability of Algorithm 3 and 4 is upper bounded by*

$$P_e \leq \Big(4M^2(K+1)+6M(K-1)(T+1)+4TK^2\Big)\exp\big(-\frac{nD_{KS}^2}{8}\big).$$

Sketch of Proof. Note that the merge step only happens a finite number of times. Thus, convergence of the algorithm can be proved in a way similar to the previous case.

Next we analyze the probability of error. An error occurs at the center initialization step if and only if the following two events which is denoted by G_1 and G_2 happen:

1. Initialization finds K_1 centers that are drawn from K_2 ($<K$) distributions, i.e. sequences generated by different distributions are close to each other.
2. Initialization finds K_3 ($>K$) centers that are drawn from K distributions, i.e. sequences generated by the same distribution have large KS distances.

By Lemmas 2.3 and 2.4, we have

$$P\big(G_1\cup G_2\big) \leq 4M^2(K+1)\exp\big(-\frac{nD_{KS}^8}{8}\big). \tag{6}$$

Let D^t be the event that incorrect merges occur at the t-th ($t\geq 1$) merge step with correct clustering results before that.

Algorithm 4 KS based clustering with unknown K

Input: $\{\mathbf{x}_j\}_{j=1}^{M}$, number of clusters $\hat{K}$.
Output: Partition set $\{C_k\}_{k=1}^{\hat{K}}$.
Initialization: $\{C_k\}_{k=1}^{\hat{K}}$ by Algorithm 3.
Method:
while not converge **do**
  Center update step specified in Algorithm 2
  {Merge Step}
  for $k_1, k_2 \in \{1, \ldots, \hat{K}\}$ and $k_1 \neq k_2$ **do**
    if $d_{KS}(\mathbf{c}_{k_1}, \mathbf{c}_{k_2}) \leq \frac{D_{KS}}{2}$ **then**
      if $\sum_{\mathbf{x}_j \in C_{k_1}} d_{KS}(\mathbf{c}_{k_2}, \mathbf{x}_j) < \sum_{\mathbf{x}_j \in C_{k_2}} d_{KS}(\mathbf{c}_{k_1}, \mathbf{x}_j)$ **then**
        $\mathcal{C}_{k_2} \leftarrow \mathcal{C}_{k_1} \cup \mathcal{C}_{k_2}$ and delete $\mathbf{c}_{k_1}$ and $\mathcal{C}_{k_1}$.
      else
        $\mathcal{C}_{k_1} \leftarrow \mathcal{C}_{k_1} \cup \mathcal{C}_{k_2}$ and delete $\mathbf{c}_{k_2}$ and $\mathcal{C}_{k_2}$.
      end if
      $\hat{K} \leftarrow \hat{K} - 1$.
    end if
  end for
  Cluster update step specified in Algorithm 2.
end while
Return $\{C_k\}_{k=1}^{\hat{K}}$

Then D^t is a subset of the event that sequences generated by different distributions have KS distance no more than $\frac{D_{KS}}{2}$. By Lemma 2.3, we have

$$P\big(D^T\big) \leq 4K^2 \exp\big(-\frac{nD_{KS}^2}{8}\big). \tag{7}$$

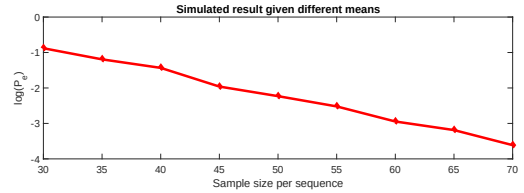
Moreover, H^t is a subset of the event that incorrect assignments occur at the t-th ($t \geq 1$) Assignment Step while the t-th Merge Step provides correct centers. One can easily show that $P\big(\cup_{t=0}^{T} H^t\big)$ has the same upper bound as (5). By (5) - (7), we have

$$P_e \leq \Big(4M^2(K+1) + 6M(T+1)(K-1) + 4TK^2\Big) \exp\big(-\frac{nD_{KS}^2}{8}\big). \quad \square$$

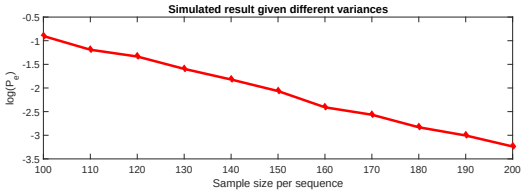
5. SIMULATION RESULT

In this section, we provide some simulation results. All the sequences are generated by $\mathcal{N}(\mu_i, 1)$ and $\mathcal{N}(0, \sigma_i)$, where $\mu_i = i - 1$ and $\sigma_i = 2^{i-1}$ for $i = 1, \ldots, 5$. Each distribution generates three sequences. Simulation results for known and unknown number of clusters are shown in Figs. 1 and 2, respectively. One can observe from the figures that for each case $\log(P_e)$ is a linear function of the sample size, i.e. P_e is exponentially consistent.

Furthermore, the comparison of theoretical error exponents and simulated ones are summarized in Table 1. The result shows that there is a gap between the theoretical results and the empirical ones. However, this does not necessarily mean that the theoretical lower bound of the error exponents is always loose. Since the presented bounds work for arbitrary distributions, there may exist some distributions for which the empirical result is close to the bound.

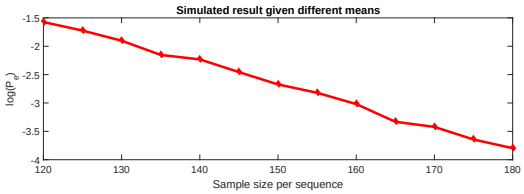


(a) Gaussian distributions with different means

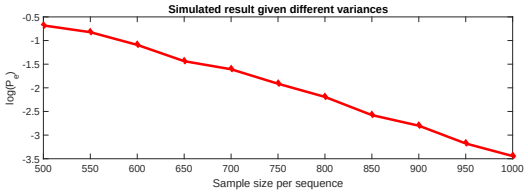


(b) Gaussian distributions with different variances

Fig. 1: Performance of Algorithm 2



(a) Guassian distributions with different means



(b) Guassian distributions with different variances

Fig. 2: Performance of Algorithm 4

Table 1: Comparison of the error exponents

	different means	different variances
theoretical result	0.0183	0.0032
Algorithm 2	0.0683	0.0234
Algorithm 4	0.0371	0.0055

6. REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge, 2012.

[3] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, pp. 129–137, Mar. 1982.

[4] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, 2000.

[5] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 24, pp. 881–892, Jul. 2002.

[6] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, pp. 4066–4082, July 2014.

[7] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear complexity exponentially consistent tests for outlying sequence detection," *arXiv preprint*, 2017.

[8] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data streams," *IEEE Trans. Signal Process.*, to be published.

[9] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of geometric structures over networks," *IEEE Trans. Signal Process.*, vol. 65, pp. 5034–5046, 2017.

[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[11] R. Moreno-Sáez and L. Mora-López, "Modelling the distribution of solar spectral irradiance using data mining techniques," *Environmental Modelling and Software*, vol. 53, no. Supplement C, pp. 163 – 172, 2014.

[12] I. Katsavounidis, C. C. Jay Kuo, and Zhen Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Process. Lett.*, vol. 1, no. 10, pp. 144–146, Oct. 1994.

[13] L. Mora-López and J. Mora, "An adaptive algorithm for clustering cumulative probability distribution functions using the kolmogorov-Smirnov two-sample test," *Expert Syst. with Applicat.*, vol. 42, no. 8, pp. 4016 – 4021, 2015.

[14] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *The Ann. of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990.