Kernel Stochastic Separation Theorems and Separability Characterizations of Kernel Classifiers

1st Ivan Y. Tyukin[1,2,3]
[1] *Department of Mathematics*
University of Leicester
[2] *Lobachevsky University*
[3] *St-Petersburg State Electrotechnical University*
[1] Leicester, United Kingdom
I.Tyukin@le.ac.uk

2nd Alexander N Gorban
University of Leicester
and Lobachevsky University
Leicester, UK, and Nizhni Novgorod, Russia
a.n.gorban@le.ac.uk

3rd Bogdan Grechuk
Department of Mathematics
University of Leicester
Leicester, United Kingdom
bg83@le.ac.uk

4th Stephen Green
Department of Mathematics
University of Leicester
Leicester, United Kingdom
slg46@le.ac.uk

***Abstract*—In this work we provide generalizations and extensions of stochastic separation theorems to kernel classifiers. A general separability result for two random sets is also established. We show that despite feature maps corresponding to a given kernel function may be infinite-dimensional, kernel separability characterizations can be expressed in terms of finite-dimensional volume integrals. These integrals allow to determine and quantify separability properties of an arbitrary kernel function. The theory is illustrated with numerical examples.**

***Index Terms*—Stochastic separation theorems, kernel stochastic separation theorems, kernel classifiers, artificial intelligence, machine learning**

I. Introduction

Kernel classifiers have long been recognized as a powerful tool for a broad range of classification problems [1], [2]. Not only they offer a natural extension of linear classifiers to the nonlinear ones, the Representer Theorem [3] states that kernel classifiers minimizing a wide range of risk functionals can often be expressed as kernel expansions over sample points. The latter property allows to expand the technology of support vector machines [4] to the realm of kernel classifiers and hence offers a computationally efficient way to construct classifiers with nonlinear decision boundaries.

Choosing a particular kernel for a given task at hand is recognized as a hard theoretical and computational problem [5]. Several approaches have been developed to date to address this problem, including the grid search algorithms [6], [7], automatic tuning of kernel parameters [8], genetic algorithms [9], and other heuristics [10]. These methods allow selection of optimal feature spaces via thorough and explicit statistical evaluation of kernel classifiers built over a family of kernel candidates.

The work was supported by Innovate UK Knowledge Transfer Partnership grant KTP010522 and by the Ministry of education and science of Russia (Project No. 14.Y26.31.0022).

In this work, we explore an alternative approach. Instead of repeatedly solving a given classification problem with a given family of kernel classifiers directly we investigate and assess relevant statistical properties of kernels and their corresponding feature maps. Our motivation stems from the seminal Cover's theorem [11], [12] suggesting that higher dimensionality of the feature maps relative to that of the original data may play a role in success of kernel classifiers, and other relevant body of work [13], [14], [15], [16], [17], [18], [19], [20], [21] on properties and geometry of high-dimensional spaces.

Links between dimensionality and separability have been explored in the literature on statistical learning theory through e.g. the concept of the Vapnik-Chervonenkis (VC) dimension [22], [23] measuring richness of classification rules which can be implemented by a classifier. Here we adapt stochastic separation theorems [14], [15], [16], [24], [25], [26] to kernel classifiers, provide their kernel generalizations, and use these results to derive computable separability measures for kernel classifiers, including for given samples of empirical data (cf. [27] exploring the notion of the local Rademacher complexity).

One of the outcomes of such generalization is an explicit characterization of kernel separability properties in terms of finite-dimensional volume integrals over domains determined by the kernel functions themselves. This suggests that even when kernel feature maps are infinite dimensional, separability properties of these maps can be expressed in terms of (finite) dimensionality of the space to which the original data belongs.

The paper is organized as follows. Section II contains necessary theoretical preliminaries and formal statement of the problem. In Section III we present new kernel stochastic separation theorems and derive separability characterizations for kernel classifiers, Section IV presents two numerical examples, and Section V concludes the paper.

978-1-7281-2009-6/$31.00

NOTATION

The following notational agreements are used throughout the text:

- $\mathbb{R}^n$ stands for the n-dimensional linear real vector space;
- $\mathbb{N}$ denotes the set of natural numbers;
- symbols $\boldsymbol{x} = (x_1, \ldots, x_n)$ will denote elements of $\mathbb{R}^n$;
- $(\boldsymbol{x}, \boldsymbol{y}) = \sum_k x_k y_k$ is the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$ is the standard Euclidean norm in $\mathbb{R}^n$;
- $\mathbb{B}_n$ denotes for the unit ball in $\mathbb{R}^n$ centered at the origin: $\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n | \ (\boldsymbol{x}, \boldsymbol{x}) \leq 1\}$;
- V_n is the n-dimensional Lebesgue measure, and $V_n(\mathbb{B}_n)$ is the volume of unit ball;
- if $\mathcal{Y}$ is a finite set then the number of elements in $\mathcal{Y}$ (cardinality of $\mathcal{Y}$) is denoted by $|\mathcal{Y}|$;
- if $\boldsymbol{x}$ is a random variable then $E[\boldsymbol{x}]$ is the expectation of $\boldsymbol{x}$.

II. PRELIMINARIES AND PROBLEM FORMULATION

Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be a set of vectors in $\mathbb{R}^n$ sampled from some distribution with a corresponding probability density function p. Each element of the sample is subjected to a transformation

$$\Phi : \mathbb{R}^n \to \mathbb{H},$$

called a *feature map*, mapping $\boldsymbol{x}_i \in \mathbb{R}^n$ into $\Phi(\boldsymbol{x}_i)$ in some Hilbert space $\mathbb{H}$. We shall assume that the feature map Φ is known. The process induces a new random variable, $\Phi(\boldsymbol{x})$, and a corresponding distribution. We suppose that the feature map Φ is such that

$$E[\Phi] = \int \Phi(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

exists. For the moment we assume that $E[\Phi] = 0$ and lift this technical assumption later in Section III-C.

For the given feature map $\Phi(\boldsymbol{x})$, a kernel function κ is defined as follows

$$\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \ \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)), \quad (1)$$

and a kernel classifier is the function

$$f : \mathbb{R}^n \to \mathbb{R}, \ f(\boldsymbol{x}) = \sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{x}) - b, \quad (2)$$

where $\alpha_j, b \in \mathbb{R}$, and $\boldsymbol{y}_j \in \mathbb{R}^n$. In the simplest binary classification setting, the classifier assigns positive values to elements $\boldsymbol{x} \in \mathbb{R}^n$ from the set corresponding to Class 1 and negative values to elements from set to Class 2.

Definition 1: A point $\boldsymbol{x} \in \mathbb{R}^n$ is kernel separable with the kernel function (1) from a set $\mathcal{Y} \subset \mathbb{R}^n$, if there exist $m \in \mathbb{N}$, α_j, $\boldsymbol{y}_j \in \mathbb{R}^n$, $j \in \{1, \ldots, m\}$ such that

$$\sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{x}) > \sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{y})$$

for all $\boldsymbol{y} \in \mathcal{Y}$.

Definition 2: A set $\mathcal{X} \subset \mathbb{R}^n$ is kernel separable with the kernel function (1) from a set $\mathcal{Y} \subset \mathbb{R}^n$, if there exist $m \in \mathbb{N}$, α_j, $\boldsymbol{y}_j \in \mathbb{R}^n$, $j \in \{1, \ldots, m\}$ such that

$$\sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{x}) > \sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{y})$$

for all $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{x} \in \mathcal{X}$.

Definition 3: A set $S \subset \mathbb{R}^n$ is kernel separable with the kernel function (1) if for each $\boldsymbol{x} \in S$ there exist $m \in \mathbb{N}$, α_j, $\boldsymbol{y}_j \in \mathbb{R}^n$, $j \in \{1, \ldots, m\}$ such that

$$\sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{x}) > \sum_{j=1}^{m} \alpha_j \kappa(\boldsymbol{y}_j, \boldsymbol{y})$$

for all $\boldsymbol{y} \in S$, $\boldsymbol{y} \neq \boldsymbol{x}$.

In addition to the notions of kernel separability, and similar to [16], we adopt the notion of Fisher separability to kernel classifiers.

Definition 4: A point $\boldsymbol{x} \in \mathbb{R}^n$ is Fisher separable with the kernel function (1) from a set $\mathcal{Y} \subset \mathbb{R}^n$, if

$$\kappa(\boldsymbol{x}, \boldsymbol{x}) > \kappa(\boldsymbol{x}, \boldsymbol{y}) \quad (3)$$

for all $\boldsymbol{y} \in \mathcal{Y}$. A set $S \subset \mathbb{R}^n$ is Fisher separable with the kernel function (1) if, for each $\boldsymbol{x} \in S$, (3) holds for all $\boldsymbol{y} \in S$, $\boldsymbol{y} \neq \boldsymbol{x}$.

It is clear that Fisher separability with a given kernel function automatically implies kernel separability and as such is a stronger property. Note also that the notion of Fisher separability can be further extended to the notion of Fisher separability with a threshold $\gamma \in [0, 1)$, cf. [16], by replacing (3) with

$$\gamma \kappa(\boldsymbol{x}, \boldsymbol{x}) > \kappa(\boldsymbol{x}, \boldsymbol{y}).$$

This generalization allows to meaningfully pose the separability problem for kernels like $\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}/\|\boldsymbol{x}\|, \boldsymbol{y}/\|\boldsymbol{y}\|)$ for which the first part of Definition 4 always holds true.

Having introduced relevant notions, we are now ready to present main results of the contribution.

III. KERNEL STOCHASTIC SEPARATION THEOREMS

A. Kernel separability of points

Our first result is provided in Theorem 1

Theorem 1: Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M \in \mathbb{R}^n$ be given, and let $\boldsymbol{x}$ be drawn from a distribution with the probability density function $p(\boldsymbol{x}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M)$. Then $\boldsymbol{x}$ is Fisher separable with the kernel function (1) from the set $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$ with probability at least

$$1 - \sum_{i=1}^{M} \int_{\kappa(\boldsymbol{x}, \boldsymbol{x}) - \kappa(\boldsymbol{x}, \boldsymbol{y}_i) \leq 0} p(\boldsymbol{x}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M) d\boldsymbol{x}. \quad (4)$$

Proof of Theorem 1. Consider events

$$A_i : \ \boldsymbol{x} \text{ is Fisher separable from } \boldsymbol{y}_i$$

It is clear that

$$P(\text{not } A_i) = \int_{\kappa(\boldsymbol{x}, \boldsymbol{x}) - \kappa(\boldsymbol{x}, \boldsymbol{y}_i) \leq 0} p(\boldsymbol{x}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M) d\boldsymbol{x}.$$

Recall that

$$P(A_1 \& A_2 \& \cdots \& A_M) \geq 1 - \sum_{i=1}^{M} P(\text{not } A_i).$$

Combining the last two observations we can conclude that the probability that $\boldsymbol{x}$ is separable from all $\boldsymbol{y}_i$ is bounded from below by the expression in (4). $\square$

Corollary 1: Suppose that assumptions of Theorem 1 hold. Let us further assume that there is a $\lambda \in (0,1)$, an $L \in \mathbb{R}_{>0}$, and a function $\alpha : \mathbb{N} \to \mathbb{R}$ such that

$$\int_{\kappa(\boldsymbol{x},\boldsymbol{x})-\kappa(\boldsymbol{x},\boldsymbol{y})\leq 0} p(\boldsymbol{x}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_M)d\boldsymbol{x} \leq L\lambda^{\alpha(n)} \tag{5}$$

for all $\boldsymbol{y} \in \mathbb{R}^n$. Then $\boldsymbol{x}$ is Fisher separable with the kernel function (1) from the set $\mathcal{Y} = \{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_M\}$ with probability at least

$$1 - ML\lambda^{\alpha(n)}.$$

It has been shown in [16] that for the feature maps $\Phi(\boldsymbol{x}) = \boldsymbol{x}$ and $p(\cdot|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_M)$ defined on $\mathbb{B}_n$ and bounded from above by $L/V_n(\mathbb{B}_n)$, condition (5) holds with

$$\lambda = \frac{1}{2},\ \alpha(n) = n.$$

Corollary 2: Consider the set $S = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M\}$ in which $\boldsymbol{x}_i \in \mathbb{R}^n$, $i = 1,\ldots,M$ are random i.i.d. vectors. Let $p : \mathbb{R}^n \to \mathbb{R}$ be the corresponding probability density function, and let there exist $\lambda \in (0,1)$, $L \in \mathbb{R}_{>0}$, and a function $\alpha : \mathbb{N} \to \mathbb{R}$ such that

$$\int_{\kappa(\boldsymbol{x},\boldsymbol{x})-\kappa(\boldsymbol{x},\boldsymbol{y})\leq 0} p(\boldsymbol{x})d\boldsymbol{x} \leq L\lambda^{\alpha(n)} \tag{6}$$

for all $\boldsymbol{y} \in \mathbb{R}^n$. Then the set S is Fisher separable with the kernel function (1) with probability at least

$$1 - (M-1)ML\lambda^{\alpha(n)}.$$

Theorem 1 and Corollaries 1, 2 extend stochastic separation theorems to kernel classifiers. Note that dimensionality N of the space to which the feature map, $\Phi(\cdot)$, maps original data points $\boldsymbol{x}$ need not be finite. And yet, according to these results, the probabilities of kernel separability in these cases can still be estimated via integration in finite dimensional spaces using e.g. (5), (6).

We also note that, since Theorem 1 and Corollaries 1 concern Fisher separability, these bounds apply to kernel separability too (cf. Definitions 1 – 3). Moreover, all statements derived so far could be generalized to cases when the functions κ themselves are not kernels.

An interesting question is if these results can be extended to characterise separability of two random sets. As we shall see below, the above formalism can be generalized to answer this question too.

B. Kernel separability of two random sets

Consider two random sets $\mathcal{X} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M\}$ and $\mathcal{Y} = \{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K\}$. Let there be a process (e.g. a learning algorithm) which, for the given $\mathcal{X}$, $\mathcal{Y}$ or their subsets, produces a function

$$f(\cdot) = \sum_{i=1}^{d} \alpha_i \kappa(\boldsymbol{z}_i, \cdot),\ \alpha_j \in \mathbb{R}.$$

The vectors $\boldsymbol{z}_i$, $i = 1,\ldots,d$ are supposed to be known. Furthermore, we suppose that the function f is such that

$$f(\boldsymbol{y}_j) > \sum_{m,k=1}^{d} \alpha_m \alpha_k \kappa(\boldsymbol{z}_m, \boldsymbol{z}_k) \tag{7}$$

for all $\boldsymbol{y}_j \in \mathcal{Y}$. In other words, if we denote $\boldsymbol{w} = \sum_{i=1}^{d} \alpha_i \Phi(\boldsymbol{z}_i)$, the following holds true:

$$(\boldsymbol{w},\boldsymbol{w}) < (\boldsymbol{w}, \Phi(\boldsymbol{y}_i)) \text{ for all } i = 1,\ldots,K. \tag{8}$$

Note that since the $\mathcal{Y}, \mathcal{X}$ are random, it is natural to expect that the vector $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_d)$ is also random. The following statement can now be formulated:

Theorem 2: Consider sets $\mathcal{X}$ and $\mathcal{Y}$. Let $p_\alpha(\boldsymbol{\alpha})$ be the probability density function associated with the random vector $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}$ satisfies condition (7) with probability 1. Then the set $\mathcal{X}$ is kernel separable with the kernel function (1) from the set $\mathcal{Y}$ with probability at least

$$1 - \sum_{i=1}^{M} \int_{H(\boldsymbol{\alpha},\boldsymbol{x}_i)\leq 0} p_\alpha(\boldsymbol{\alpha})d\boldsymbol{\alpha}, \tag{9}$$

where

$$H(\boldsymbol{\alpha},\boldsymbol{x}_i) = \sum_{k,m=1}^{d} \alpha_k \alpha_m \kappa(\boldsymbol{z}_k,\boldsymbol{z}_m) - \sum_{m=1}^{d} \alpha_m \kappa(\boldsymbol{z}_m,\boldsymbol{x}_i).$$

Proof of Theorem 2. The proof of the theorem is similar to that of Theorem 1. Consider events

$$A_i :\ (\boldsymbol{w},\boldsymbol{w}) > (\boldsymbol{w}, \Phi(\boldsymbol{x}_i)).$$

Events A_i are equivalent to that $H(\boldsymbol{\alpha},\boldsymbol{x}_i) > 0$. Eq. (9) provides a lower bound for the probability that all these events hold. Recall that vectors $\boldsymbol{\alpha}$ satisfy (8), and hence

$$\begin{aligned}\sum_{m=1}^{d} \alpha_m \kappa(\boldsymbol{z}_m,\boldsymbol{x}_i) =& (\boldsymbol{w}, \Phi(\boldsymbol{x}_i)) \\ &< (\boldsymbol{w}, \Phi(\boldsymbol{y}_j)) = \sum_{m=1}^{d} \alpha_m \kappa(\boldsymbol{z}_m, \boldsymbol{y}_j)\end{aligned}$$

for all $\boldsymbol{x}_i \in \mathcal{X}$ and $\boldsymbol{y}_j \in \mathcal{Y}$ with probability at least (9). The statement now follows immediately from Definition 2. $\square$

Theorem 2 generalizes earlier k-tuple separation theorems [24], [28] in that it applies to a much broader class of classifiers and is not limited to a particular set of distributions. Similar to Corollaries 1, 2, one can establish conditions linking dimensionality of the vector $\boldsymbol{\alpha}$ with the probability of

separation. An example of such condition could be existence of $L > 0$, $\lambda \in (0,1)$ and a function $\beta : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ such that

$$\int_{H(\boldsymbol{\alpha},\boldsymbol{y})\leq 0} p_\alpha(\boldsymbol{\alpha})d\boldsymbol{\alpha} \leq L\lambda^{\beta(d,n)}$$

for any $\boldsymbol{y} \in \mathbb{R}^n$.

C. Kernels with $E[\Phi(\boldsymbol{x})] \neq 0$

Theorems 1, 2, their corollaries, and Fisher separability notions in Definition 4 have been produced under the simplifying assumption that $E[\Phi(\boldsymbol{x})] = 0$. These statements, however, can be straightforwardly generalized to more general settings with

$$E[\Phi(\boldsymbol{x})] = \overline{\Phi}.$$

The generalization can be achieved by replacing kernel functions $\kappa(\boldsymbol{x},\boldsymbol{y}) = (\Phi(\boldsymbol{x}),\Phi(\boldsymbol{y}))$ with

$$\begin{aligned}\tilde{\kappa}(\boldsymbol{x},\boldsymbol{y}) = (\Phi(\boldsymbol{x}) - \overline{\Phi}, \Phi(\boldsymbol{y}) - \overline{\Phi}) = \kappa(\boldsymbol{x},\boldsymbol{y})\\ - \int p(\boldsymbol{x})\kappa(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x} - \int p(\boldsymbol{y})\kappa(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y}\\ + \int\int p(\boldsymbol{x})p(\boldsymbol{y})\kappa(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}\end{aligned}$$

in relevant expressions. In practice, $E[\Phi(\boldsymbol{x})]$ can be replaced with the sample mean over a finite sample $\mathcal{X} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M\}$ leading to the following approximations of $\tilde{\kappa}(\boldsymbol{x},\boldsymbol{y})$:

$$\begin{aligned}\tilde{\kappa}(\boldsymbol{x},\boldsymbol{y}) = \kappa(\boldsymbol{x},\boldsymbol{y})\\ - \frac{1}{N}\sum_{i=1}^{M}(\kappa(\boldsymbol{x}_i,\boldsymbol{y}) + \kappa(\boldsymbol{x},\boldsymbol{x}_i)) + \frac{1}{N^2}\sum_{i,j=1}^{M}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j).\end{aligned}$$

D. Kernel separability measure

One of the outcomes of our theoretical results is the kernel separability characterization expressed e.g. in terms of the upper bound $L\lambda^{\alpha(n)}$ on the integral

$$\int_{\tilde{\kappa}(\boldsymbol{x},\boldsymbol{x})-\tilde{\kappa}(\boldsymbol{x},\boldsymbol{y})\leq 0} p(\boldsymbol{x})d\boldsymbol{x}$$

in the left-hand side of (6). The bound allows to determine how well a particular kernel $\kappa(\cdot,\cdot)$ separates points in samples from a given distribution. The smaller the value of λ is and the faster the function $\alpha(\cdot)$ grows with n the better is the kernel's separability (as per Definition 4). Direct derivation of λ, $\alpha(\cdot)$, requires knowledge or at least bounds on the probability density functions p, p_α.

In practice, however, the probability density functions are rarely known. On the other hand, it is not unrealistic to expect that some data sample $\mathcal{X}$ from the distribution might be available. If this is the case then one may replace evaluation of the above integral with

$$\omega(\boldsymbol{y},n) = \frac{|\{\boldsymbol{x}\in\mathcal{X}\subset\mathbb{R}^n \ |\boldsymbol{x}\neq\boldsymbol{y},\ \tilde{\kappa}(\boldsymbol{x},\boldsymbol{x}) - \tilde{\kappa}(\boldsymbol{x},\boldsymbol{y}) \leq 0\}|}{|\mathcal{X}|}.$$

Empirical kernel separability measure, can then be defined as the average

$$\Omega_a(n) = \frac{\sum_{i=1}^{|\mathcal{X}|}\omega(\boldsymbol{x}_i,n)}{|\mathcal{X}|}$$

or

$$\Omega_{\max}(n) = \max_{\boldsymbol{x}_i\in\mathcal{X}}\omega(\boldsymbol{x}_i,n). \tag{10}$$

In the next section we show how these measures can be employed to characterize and compare different kernel functions with respect to their ability to separate points in random sets.

IV. Examples

A. Polynomial kernels for an equidistribution in the $[-1,1]^n$ cube

In the first group of experiments we considered behavior of polynomial kernels in a synthetic test in which the original data samples are generated from equidistributions in the unit cubes $[-1,1]^n$ of varying dimension n. The kernel functions were chosen as follows

$$\kappa(\boldsymbol{x},\boldsymbol{y}) = ((\boldsymbol{x},\boldsymbol{y})+1)^p, \tag{11}$$

where parameter p took values in the set $\{1,2,3\}$. Observe that for $p=1$ the centralized kernel $\tilde{\kappa}$ reduces to standard inner product $(\boldsymbol{x},\boldsymbol{y})$, and for quadratic kernels, $p=2$, the feature map is

$$\begin{aligned}\Phi(\boldsymbol{x}) =& (x_1^2,\ldots,x_n^2,\sqrt{2}x_1x_2,\ldots,\sqrt{2}x_{n-1}x_n,\\ &\sqrt{2}x_1,\ldots,\sqrt{2}x_n,1).\end{aligned} \tag{12}$$

For each given value of n we generated a sample of $M = 10^3$ vectors and calculated the value of $\Omega_{\max}(n)$ (see Eq. (10)) from a sub-sample of 10^2 points chosen randomly from this sample. The outcomes of this process for different values of p are summarized in Fig. 1.

According to Fig. 1, separability of quadratic kernels is higher that that of the original Fisher discriminants. This is hardly surprising given that the dimensionality of the quadratic feature map, $n(n+1)/2+n+1$, is significantly higher than that of the original space, n.

Unexpectedly, these experiments reveal that re-weighting of features, e.g. via the whitening transformation, may produce higher performance gains than choosing a kernel with a higher-dimensional feature map (black triangled lines in Fig. 1 vs squared magenta lines). Note, however, that such re-weighting generates a different inner product and hence corresponds to a kernel function that is different from the ones specified in (11).

B. Polynomial kernels for the bottlenecks of Inception model

In this example we investigated and tested kernel separability in the setting when the original features are bottle necks of a pre-trained convolutional neural network. In particular, we considered a network trained to distinguish ten digits in American Sign Language. The network was an Inception deep neural network model [29] whose architecture and training process has been described in detail in [28]. The model was trained[1] on ten sets of images corresponding to the American Sign Language pictures for 0-9. Each set contained 1000

[1] https://www.tensorflow.org/tutorials/image_retraining

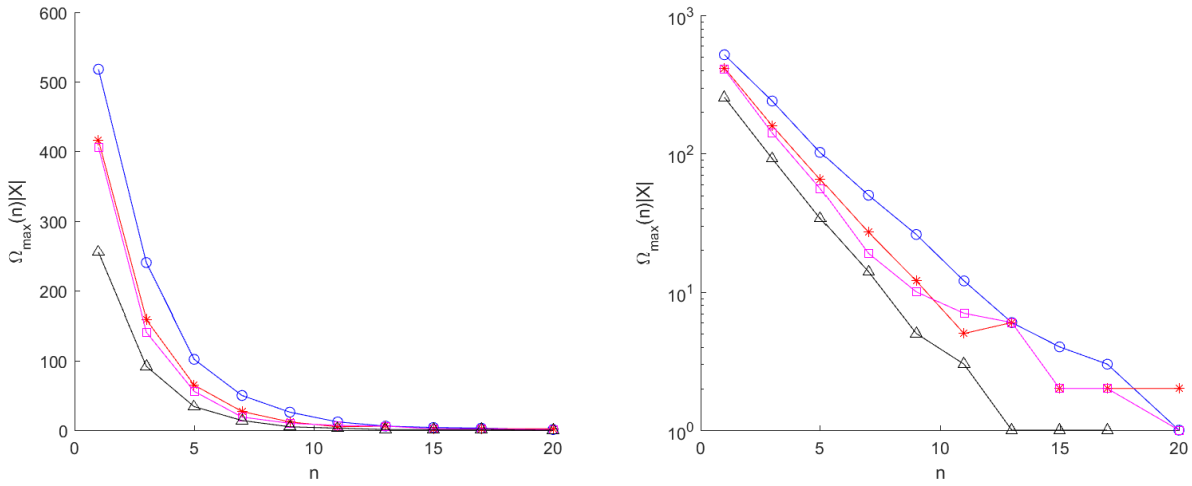

Fig. 1. Separability characterizations for polynomial kernels with $p = 1, 2, 3$. Blue circled line corresponds to $p = 1$, red starred line corresponds to $p = 2$, and magenta squares show performance of the chosen polynomial kernel with $p = 3$. Black line with triangle markers corresponds to Fisher discriminants built over quadratic feature maps followed by centralization and whitening. Left panel shows original data, and right panel shows the same data but in the logarithmic scale.

unique images consisting of profile shots of the persons hand, along with 3/4 profiles and shots from above and below. The states $\boldsymbol{x}_i$ are the vectors containing the values of pre-softmax layer bottlenecks of size n for however many neurons are in the penultimate layer.

It was shown in [28] that these bottlenecks can be used to construct error correcting cascades for such systems, and thanks to stochastic separation theorems, such cascades can be derived using mere Fisher discriminants. The higher is the dimension of the bottlenecks $\boldsymbol{x}_i$ the larger is the probability that the error correcting cascades are successful. It is therefore interesting to see if employing kernels in place of the original linear classifiers could potentially improve error correction performance. To assess this, we projected original bottle necks $\boldsymbol{x}_i$ on n principal components, and constructed higher-dimensional representations of the projected data using quadratic kernel feature map (12). This was followed by centering and whitening transformation. For the new feature vectors defined in this way, we derived $\Omega_{\max}(n)$ for n ranging from 1 to 94. Results are shown in Fig. 2.

The figure suggests that using kernels induced by quadratic feature maps may offer superior point separability properties as compared to the original feature vectors.

V. Conclusion

In this work we presented an extension of the framework of stochastic separation theorems to arbitrary kernel classifiers. Separability criteria emerging from these generalizations reduce to finite-dimensional volume integrals despite the fact that the feature maps corresponding to relevant kernels may be infinite-dimensional. In addition, we formulated a general separability result for two random sets. The latter result assumes some prior knowledge of the distribution of the weights of the classifier.

These results allowed us to produce empirical kernel separability characterizations for arbitrary kernel functions. The application of these new characterizations has been illustrated with two case study examples. These examples showed that if an additional whitening and re-weighting of features are allowed then point separability performance of the induced kernel may be drastically improved. We have not, however, investigated generalization capabilities of such kernel classifiers and their derivatives for the problem of AI error correction [28], [24]. This will be the subject of our future work.

Acknowledgment

The work was supported by Innovate UK Knowledge Transfer Partnership grant KTP010522 and by the Ministry of education and science of Russia (Project No. 14.Y26.31.0022).

References

[1] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, pp. 1171–1220, 2008.

[2] R. Herbrich, *Learning kernel classifiers: theory and algorithms*, 2001.

[3] B. Schölkopf and A. J. Herbrich, R.and Smola, "A generalized representer theorem," in *International conference on computational learning theory*. Springer, 2001, pp. 416–426.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 2000.

[5] K.-P. Wu and S.-D. Wang, "Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space," *Pattern Recognition*, vol. 42, no. 5, pp. 710–717, 2009.

[6] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[7] N. Ancona, G. Cicirelli, E. Stella, and A. Distante, "Object detection in images: run-time complexity and parameter selection of support vector machines," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2. IEEE, 2002, pp. 426–429.

[8] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 131–159, 2002.

[9] S. Lessmann, R. Stahlbock, and S. F. Crone, "Genetic algorithms for support vector machine model selection." 2006.

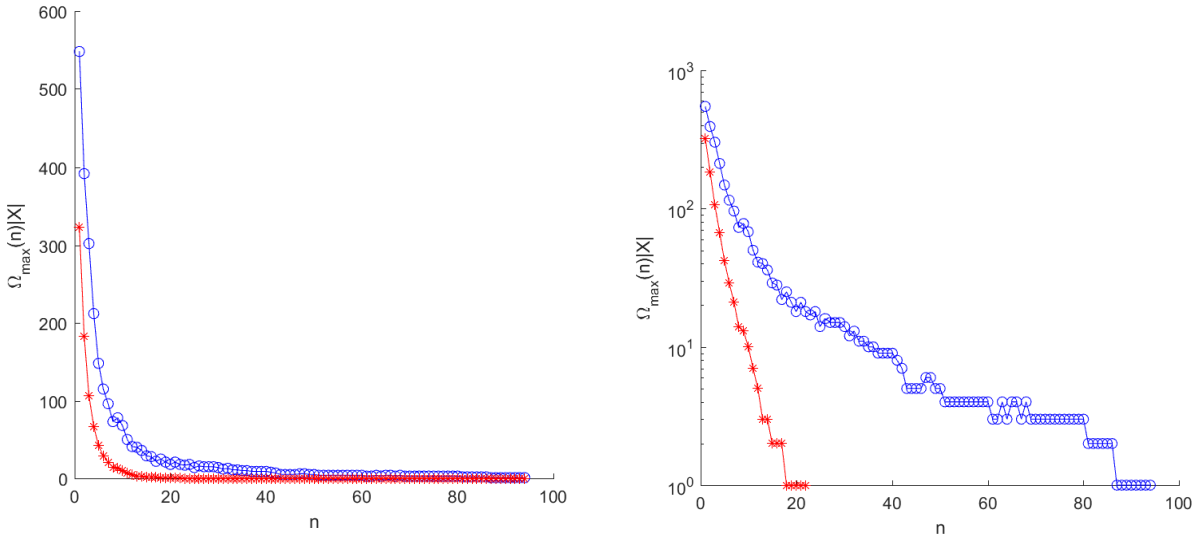

Fig. 2. Separability of the feature vectors sampled from Inception bottlenecks. Blue circled lines show $\Omega_{\max}(n)|\mathcal{X}|$ as a function of the number n of the principal components retained. Red starred line shows $\Omega_{\max}(n)|\mathcal{X}|$ for kernels with quadratic, centered, and whitened feature map (12). Left panel shows original data, and right panel shows the same data but in the logarithmic scale.

[10] M. Varewyck and J.-P. Martens, "A practical approach to model selection for support vector machines with a gaussian kernel," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 2, pp. 330–340, 2011.

[11] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.

[12] S. S. Haykin, *Neural networks and learning machines*. Pearson Upper Saddle River, 2009, vol. 3.

[13] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[14] A. Gorban and I. Tyukin, "Stochastic separation theorems," *Neural Networks*, vol. 94, pp. 255–259, 2017.

[15] A. Gorban, I. Tyukin, and I. Romanenko, "The blessing of dimensionality: Separation theorems in the thermodynamic limit," 09 2016, a talk given at TFMST 2016, 2nd IFAC Workshop on Thermodynamic Foundations of Mathematical Systems Theory. September 28-30, 2016, Vigo, Spain. [Online]. Available: https://arxiv.org/abs/1610.00494v1

[16] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of ai systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303–322, 2018.

[17] P. Kainen and V. Kurkova, "Quasiorthogonal dimension of euclidian spaces," *Appl. Math. Lett.*, vol. 6, no. 3, pp. 7–10, 1993.

[18] P. C. Kainen, "Utilizing geometric anomalies of high dimension: When complexity makes computation easier," in *Computer Intensive Methods in Control and Signal Processing*. Springer, 1997, pp. 283–294.

[19] A. Gorban, I. Tyukin, D. Prokhorov, and K. Sofeikov, "Approximation with random bases: Pro et contra," *Information Sciences*, vol. 364–365, pp. 129–145, 2016.

[20] A. Gorban and I. Tyukin, "Blessing of dimensionality: mathematical foundations of the statistical physics of data," *Philosophical Transactions of the Royal Society A*, vol. 376, p. 20170237, 2018.

[21] M. Gromov, "Isoperimetry of waists and concentration of maps," *GAFA, Geomteric and Functional Analysis*, vol. 13, pp. 178–215, 2003.

[22] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*. Springer, 2015, pp. 11–30.

[23] V. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[24] I. Y. Tyukin, A. N. Gorban, K. Sofeikov, and I. Romanenko, "Knowledge transfer between artificial intelligence systems," *Frontiers of Neurorobotics*, vol. 12, Article 49, 2018.

[25] A. Gorban, R. Burton, I. Romanenko, and I. Tyukin, "One-trial correction of legacy AI systems and stochastic separation theorems," *Information Sciecnes*, vol. 484, pp. 237–254, 2019.

[26] I. Tyukin, A. Gorban, S. Green, and D. Prokhorov, "Fast construction of correcting ensembles for legacy artificial intelligence systems: Algorithms and a case study," *Information Sciecnes*, vol. 485, pp. 230–247, 2019.

[27] C. Cortes, M. Kloft, and M. Mohri, "Learning kernels using local Rademacher complexity," in *Advances in neural information processing systems*, 2013, pp. 2760–2768.

[28] I. Y. Tyukin, A. N. Gorban, S. Green, and D. Prokhorov, "Fast construction of correcting ensembles for legacy artificial intelligence systems: Algorithms and a case study," *arXiv preprint arXiv:1810.05593*, 2018.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.