Neural Networks 94 (2017) 255–259

Contents lists available at ScienceDirect

Neural Networks

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

# Stochastic separation theorems

A.N. Gorban [a,*], I.Y. Tyukin [a,b]

[a] Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK
[b] Department of Automation and Control Processes, Saint-Petersburg State Electrotechnical University, Saint-Petersburg, 197376, Russia

ARTICLE INFO

*Article history:*
Received 1 April 2017
Received in revised form 16 July 2017
Accepted 21 July 2017
Available online 31 July 2017

*Keywords:*
Fisher's discriminant
Random set
Measure concentration
Linear separability
Machine learning
Extreme point

ABSTRACT

The problem of non-iterative one-shot and non-destructive correction of unavoidable mistakes arises in all Artificial Intelligence applications in the real world. Its solution requires robust separation of samples with errors from samples where the system works properly. We demonstrate that in (moderately) high dimension this separation could be achieved with probability close to one by linear discriminants. Based on fundamental properties of measure concentration, we show that for $M < a\exp(bn)$ random $M$-element sets in $\mathbb{R}^n$ are linearly separable with probability $p$, $p > 1-\vartheta$, where $1 > \vartheta > 0$ is a given small constant. Exact values of $a, b > 0$ depend on the probability distribution that determines how the random $M$-element sets are drawn, and on the constant $\vartheta$. These *stochastic separation theorems* provide a new instrument for the development, analysis, and assessment of machine learning methods and algorithms in high dimension. Theoretical statements are illustrated with numerical examples.



## 1. Introduction

Artificial Intelligence (AI) systems make errors. They should be corrected without damage of existing skills. The *problem of non-destructive correction* arises in many areas of research and development, from AI to mathematical neuroscience, where the reverse engineering of the brain ability to learn on-the-fly remains a great challenge. It is very desirable that the corrector of errors is *non-iterative* (one-shot) because iterative re-training of a large system requires much time and resource and cannot be done immediately without impeding activity.

The non-destructive correction requires separation of the situations (samples) with errors from the samples corresponding to correct behaviour by a simple and robust classifier. Linear discriminants introduced by Fisher (1936) are simple, robust, require just the inverse covariance matrix of data, and may be easily modified for assimilation of new data. Rosenblatt (1962) revived the common interest in linear classifiers. His works sparked intensive scientific debate (Minsky & Papert, 1969) and gave rise to development of numerous crucial concepts such as e.g. Vapnik–Chervonenkis theory (Vapnik & Chervonenkis, 1971), learnability (Natarajan, 1989), and generalization capabilities of neural networks (Bousquet & Elisseeff, 2002; Vapnik, 2000). Linear functionals (adaptive summators) are basic building blocks of significantly more sophisticated AI systems such as e.g. multi-layer perceptrons, (Rumelhart, Hinton, & Williams, 1986), Convolutional Neural Networks (Le Cun & Bengio, 1995; LeCun, Bengio, & Hinton, 2015) and their derivatives. Much is known about linear functionals as "stand-alone" learning machines, including their generalization margins (Freund & Schapire, 1999; Vapnik, 2000) and numerous methods for their construction: linear discriminants and regression, perceptron learning, and Support Vector Machines (Vapnik, 1982) among others.

In this work, we demonstrate that in high dimensions and even for exponentially large samples, linear classifiers in their classical Fisher's form are powerful enough to separate errors from correct responses with high probability and to provide efficient solution to the non-destructive corrector problem. We prove that linear functionals, as learning machines, have surprising and, as far as we are concerned, new peculiar extremal properties: *in high dimension, with probability* $p > 1-\vartheta$ *and for* $M < a\exp(bn)$ *with* $a, b > 0$ *every point in random i.i.d. drawn* $M$*-element sets in* $\mathbb{R}^n$ *is linearly separable from the rest.* Moreover, the separating linear functional can be found explicitly, without iterations. This property holds for a broad set of relevant distributions, including products of probability measures with bounded support and equidistribution in a unit ball, providing mathematical foundations for one-trial correction of legacy AI systems (cf. Gorban, Romanenko, Burton, & Tyukin, 2016).

A problem of data fusion in multiagent systems has clear similarity to the problem of non-destructive correction. According to Forney, Pearl, and Bareinboim (2017), data collected by different agents may not be naively combined due to changes in the context, and special procedures for their assimilation without damage of gained skills are needed. The proven stochastic separation effects

* Corresponding author.
*E-mail addresses:* ag153@le.ac.uk (A.N. Gorban), it37@le.ac.uk (I.Y. Tyukin).

http://dx.doi.org/10.1016/j.neunet.2017.07.014
0893-6080/

can be used to approach this problem. They also shed light on the possible origins of remarkable selectivity to stimuli observed in-vivo in the real brain (Quian Quiroga, Reddy, Kreiman, Koch, & Fried, 2005).

## 2. Preliminaries

### 2.1. Notation

Throughout the text, $\mathbb{R}^n$ is the $n$-dimensional linear real vector space. Unless stated otherwise, symbols $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,n})$ denote elements of $\mathbb{R}^n$, and $(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_k x_{i,k} x_{j,k}$ is the inner product of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in $\mathbb{R}^n$. Symbol $\mathbb{B}_n$ stands for the unit ball in $\mathbb{R}^n$ centred at the origin: $\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n |\ (\boldsymbol{x}, \boldsymbol{x}) \leq 1\}$.

### 2.2. Linear separability of sets

**Definition.** A set $S \subset \mathbb{R}^n$ is *linearly separable* if for each $x \in S$ there exists a linear functional $l$ such that $l(x) > l(y)$ for all $y \in S, y \neq x$.

Recall that $x \in \mathbb{R}^n$ is an *extreme point* of a convex compact $K$ if there exist no points $y, z \in K, y \neq z$ such that $x = (y + z)/2$. The basic examples of linearly separable sets are extreme points of a convex compacts: vertices of convex polyhedra or points on the $n$-dimensional sphere. The sets of extreme points of a compact may be not linearly separable as is demonstrated by simple 2D examples (Simon, 2011).

**Proposition 1.** *Every compact linearly separable set $S \subset \mathbb{R}^n$ is a set of extreme points of a convex compact $K = \mathrm{conv}S$.*

The proof follows immediately from the previous definitions, the Krein–Milman theorem (Simon, 2011) (its finite-dimensional form was known to Minkovsky) and classical theorems about separation of a point from a convex set.

If $n+1$ points in $\mathbb{R}^n$ do not belong to a hyperplane then they are vertices of a simplex and are, obviously, linearly separable. If the lengths of the edges are bounded then the volume of the simplex decreases with $n$ not slower than $1/n!$. Therefore, we can expect that for a sufficiently regular distribution of points, a random point does not belong to the simplex, and $n + 2$ independently chosen random points are also linearly separable, with high probability, which tends to 1 as $1 - c/n!\,(n \to \infty, c > 0)$. This fast convergence allows us to hypothesize that even a large random finite set is linearly separable in high dimension with high probability, if the distribution is regular enough. We prove this statement below for i.i.d. random points from equidistributions in a ball and a cube, and from distributions that are products of measures with bounded support.

## 3. Main results

Let us start from the equidistribution in the unit ball $\mathbb{B}_n$ in $\mathbb{R}^n$. The probability $p$ that a random point belongs to a layer $\mathbb{B}_n \setminus r\mathbb{B}_n$ $(0 < r < 1)$ between spheres of radius 1 and of radius $r$ is $p = 1 - r^n$. Let us take a unit vector $\boldsymbol{v}$. The probability that the projection of a random vector $\boldsymbol{x}$ on $\boldsymbol{v}$, $(\boldsymbol{x}, \boldsymbol{v})$, exceeds $r$ can be estimated from above by half of the ratio of volumes of balls of radii $\rho = \sqrt{1 - r^2}$ and 1 (see Fig. 1 with $\varepsilon = 1 - r$): $\mathbf{P}((\boldsymbol{x}, \boldsymbol{v}) > r) \leq 0.5\rho^n$.

**Theorem 1.** *Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be a set of $M$ i.i.d. random points from the equidistribution in the unit ball $\mathbb{B}_n$, $0 < r < 1$. Then*

$$\mathbf{P}\left(\|\boldsymbol{x}_M\| > r \text{ and } \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_M}{\|\boldsymbol{x}_M\|}\right) < r \text{ for all } i \neq M\right) \geq 1 - r^n - 0.5(M-1)\rho^n; \tag{1}$$

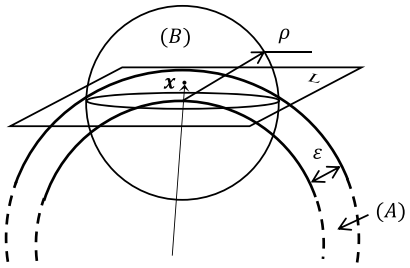


**Fig. 1.** Point $\boldsymbol{x}$ belongs to a spherical layer $(A)$ of thickness $\varepsilon$. The data are centralized and the centre of the spheres from $A$ is the origin. Hyperplane $L$ is orthogonal to vector $\boldsymbol{x}$ and tangent to the internal sphere of $A$. $L$ cuts an upper spherical cap from $A$ and separates $\boldsymbol{x}$ from the data points which belong to the external sphere of $A$ but do not belong to that cap. The cap is included into the upper half of ball $(B)$. The centre of $B$ is intersection of the radius $x$ with the internal sphere of the layer $A$.

$$\mathbf{P}\left(\|\boldsymbol{x}_j\| > r \text{ and } \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|}\right) < r \text{ for all } i, j,\ i \neq j\right) \geq 1 - Mr^n - 0.5M(M-1)\rho^n; \tag{2}$$

$$\mathbf{P}\left(\|\boldsymbol{x}_j\| > r \text{ and } \left(\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|}\right) < r \text{ for all } i, j,\ i \neq j\right) \geq 1 - Mr^n - M(M-1)\rho^n. \tag{3}$$

The proof is based on the independence of random points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, on the geometric picture presented in Fig. 1, and on an elementary inequality $\mathbf{P}(A_1 \& A_2 \& \ldots \& A_m) \geq 1 - \sum_i (1 - \mathbf{P}(A_i))$ for any events $A_1, \ldots, A_m$. In Fig. 1 we should take $\varepsilon = 1 - r$ and the external radius of the spherical layer $A$ is 1. Ball (1997) provides more geometric details of concentration of the volume of high-dimensional balls. In (3) we estimate the probability that the cosine of the angles between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ does not exceed $r$. Gorban, Tyukin, Prokhorov, and Sofeikov (2016) analysed the asymptotic behaviour of these estimations for small $r$. The idea of almost orthogonal bases was introduced by Kainen and Kůrková (1993) and used efficiently by Kůrková and Sanguineti (2007) for estimation of the cardinality of $\varepsilon$-nets in compact convex subsets of Hilbert spaces including the sets of functions computable by perceptrons.

The following corollary gives simple estimates of exponential growth of the maximal possible $M$ for which inequalities (1) and (2) hold with a given probability value.

**Corollary 1.** *Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be a set of $M$ i.i.d. random points from the equidistribution in the unit ball $\mathbb{B}_n$ and $0 < r, \vartheta < 1$. If*

$$M < 2(\vartheta - r^n)/\rho^n, \tag{4}$$

*then $\mathbf{P}((\boldsymbol{x}_i, \boldsymbol{x}_M) < r\|\boldsymbol{x}_M\| \text{ for all } i = 1, \ldots, M-1) > 1 - \vartheta$. If*

$$M < (r/\rho)^n \left(-1 + \sqrt{1 + 2\vartheta\rho^n/r^{2n}}\right), \tag{5}$$

*then $\mathbf{P}((\boldsymbol{x}_i, \boldsymbol{x}_j) < r\|\boldsymbol{x}_i\| \text{ for all } i, j = 1, \ldots, M,\ i \neq j) \geq 1 - \vartheta$.*

*In particular, if inequality (5) holds then the set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is linearly separable with probability $p > 1 - \vartheta$.*

A weaker and simpler estimate (sufficient condition) follows immediately from (5):

$$\vartheta/M^2 > r^n + 0.5\rho^n. \tag{6}$$

**Remark 1.** According to (6) the pre exponential factor in the estimate for $M^2$ may be chosen as $\vartheta$, while the exponent depends on $r$ only. For example, for $r = 1/\sqrt{2}$ the simple sufficient condition (6) gives $M^2 < \frac{2}{3}\vartheta 2^{n/2}$. For $\vartheta = 0.01$ (or specificity 99%) and $n = 100$ we get $M < 2{,}740{,}000$.

Thus, if we select 2,700,000 i.i.d. points from an equidistribution in a unit ball in $\mathbb{R}^{100}$ then with probability $p > 0.99$ *all* these points will be vertices of their convex hull.

Estimates similar to (3), (5), and (6) are useful for the equidistribution of the normalized data on a unit sphere too. This is because they not only establish the fact of separability but also specify separation margins.

Consider a product distribution in an $n$-dimensional unit cube. Let the coordinates of a random point, $X_1, \dots, X_n$ ($0 \le X_i \le 1$) be independent random variables with expectations $\overline{X}_i$ and variances $\sigma_i^2 > \sigma_0^2 > 0$. Let $\overline{\boldsymbol{x}}$ be a vector with coordinates $\overline{X}_i$. For large $n$, this distribution is concentrated in a relatively small vicinity of a sphere with an *arbitrary* centre $\boldsymbol{c}$ with coordinates $c_i$ and radius $R$, where

$$R^2 = \mathbf{E}\left(\sum_i (X_i - c_i)^2\right) = \sum_i \sigma_i^2 + \|\overline{\boldsymbol{x}} - \boldsymbol{c}\|^2. \tag{7}$$

Concentration near the spheres with different centres implies concentration in the vicinity of their intersection (an example of the 'waist concentration' Gromov, 2003). The vicinity of the spheres, where the distribution is concentrated, can be estimated by the Hoeffding inequality (Hoeffding, 1963). Let $Y_1, \dots, Y_n$ be independent bounded random variables: $0 \le Y_i \le 1$. The empirical mean of these variables is defined as $\overline{Y} = \frac{1}{n}(Y_1 + \dots + Y_n)$. Then

$$\begin{aligned} &\mathbf{P}\left(\overline{Y} - \mathbf{E}\left[\overline{Y}\right] \ge t\right) \le \exp\left(-2nt^2\right); \\ &\mathbf{P}\left(\left|\overline{Y} - \mathbf{E}\left[\overline{Y}\right]\right| \ge t\right) \le 2\exp\left(-2nt^2\right). \end{aligned} \tag{8}$$

Let us take $Y_i = (X_i - c_i)^2$. Consider the centres located in the cube, $0 \le c_i \le 1$. Then $0 \le Y_i \le 1$ and $\mathbf{E}\left[\overline{Y}\right] = \frac{1}{n}R^2$. In particular, if $c_i = \overline{X}_i$ then $\mathbf{E}\left[\overline{Y}\right] = \frac{1}{n}R_0^2$ (the minimal possible value), where $R_0^2 = \sum_i \sigma_i^2 \ge n\sigma_0^2$. In general, $n\sigma_0^2 \le R^2 \le n$.

With probability $p > 1 - 2\exp\left(-2nt^2\right)$ a random point $\boldsymbol{x}$ belongs to the spherical layer ($\delta = nt/R_0^2$, $t = \delta R_0^2/n$):

$$1 - \delta \le \|\boldsymbol{x} - \overline{\boldsymbol{x}}\|^2/R_0^2 \le 1 + \delta. \tag{9}$$

Consider $M$ i.i.d. points $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ from the product distribution. With probability $p > 1 - 2M\exp\left(-2nt^2\right)$ they all belong to the spherical layer (9). Therefore, with this probability we return to the situation presented in Fig. 1 with internal radius $\sqrt{1-\delta}R_0$ and external radius $\sqrt{1+\delta}R_0$. The difference from the equidistribution in the ball is that the volume of the ball is concentrated near the external sphere, while the distribution in the layer (9) is concentrated around the sphere $\|\boldsymbol{x} - \overline{\boldsymbol{x}}\|^2 = R_0^2$.

The radius of ball $B$ is defined by $\rho^2 = (1+\delta)R_0^2 - (1-\delta)R_0^2 = 2\delta R_0^2$. The concentration radius (7) for the spheres concentric with the ball $B$ (Fig. 1) is defined by $R^2 = R_0^2 + (1-\delta)R_0^2 = (2-\delta)R_0^2$. Therefore, a random point does not belong to the ball $B$ with probability $p > 1 - \exp\left(-2n\tau^2\right)$, where $\tau = \frac{1}{n}(R^2 - \rho^2) = \frac{1}{n}(2 - 3\delta)R_0^2$. Thus, we get the following statement.

**Theorem 2.** *Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be i.i.d. random points from the product distribution in a unit cube, $0 < \delta < 2/3$. Then*

$$\begin{aligned} &\mathbf{P}\left(1 - \delta \le \frac{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2}{R_0^2} \le 1 + \delta \text{ and } \right. \\ &\quad \left. \left(\frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}}}{R_0}, \frac{\boldsymbol{x}_M - \overline{\boldsymbol{x}}}{\|\boldsymbol{x}_M - \overline{\boldsymbol{x}}\|}\right) < \sqrt{1-\delta} \text{ for all } i, j,\ i \ne M\right) \\ &\quad \ge 1 - 2M\exp\left(-2\delta^2 R_0^4/n\right) - (M-1)\exp\left(-2R_0^4(2-3\delta)^2/n\right); \end{aligned} \tag{10}$$

$$\begin{aligned} &\mathbf{P}\left(1 - \delta \le \frac{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2}{R_0^2} \le 1 + \delta \text{ and } \right. \\ &\quad \left. \left(\frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}}}{R_0}, \frac{\boldsymbol{x}_j - \overline{\boldsymbol{x}}}{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|}\right) < \sqrt{1-\delta} \text{ for all } i, j,\ i \ne j\right) \\ &\quad \ge 1 - 2M\exp\left(-2\delta^2 R_0^4/n\right) \\ &\qquad - M(M-1)\exp\left(-2R_0^4(2-3\delta)^2/n\right). \end{aligned} \tag{11}$$

When the value of delta is chosen as $\delta = 0.5$ and $R_0$ is replaced with its estimate from below, $R_0^2 \ge n\sigma_0^2$, inequalities (10) and (11) result in the following estimates:

$$\begin{aligned} &\mathbf{P}\left(\frac{1}{2} \le \frac{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2}{R_0^2} \le \frac{3}{2} \text{ and } \right. \\ &\quad \left. \left(\frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}}}{R_0}, \frac{\boldsymbol{x}_M - \overline{\boldsymbol{x}}}{\|\boldsymbol{x}_M - \overline{\boldsymbol{x}}\|}\right) < \sqrt{1-\delta} \text{ for all } i, j,\ i \ne M\right) \\ &\quad \ge 1 - 3M\exp\left(-0.5n\sigma_0^4\right); \end{aligned} \tag{12}$$

$$\begin{aligned} &\mathbf{P}\left(\frac{1}{2} \le \frac{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2}{R_0^2} \le \frac{3}{2} \text{ and } \right. \\ &\quad \left. \left(\frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}}}{R_0}, \frac{\boldsymbol{x}_j - \overline{\boldsymbol{x}}}{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|}\right) < \sqrt{1-\delta} \text{ for all } i, j,\ i \ne j\right) \\ &\quad \ge 1 - M(M+1)\exp\left(-0.5n\sigma_0^4\right). \end{aligned} \tag{13}$$

**Corollary 2.** *Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be i.i.d. random points from the product distribution in a unit cube and $0 < \vartheta < 1$. If*

$$M < \frac{1}{3}\vartheta\exp\left(0.5n\sigma_0^4\right), \tag{14}$$

*then with probability $p > 1 - \vartheta$*

$$0.5 \le \frac{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2}{R_0^2} \le 1.5 \text{ and } \left(\frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}}}{R_0}, \frac{\boldsymbol{x}_M - \overline{\boldsymbol{x}}}{\|\boldsymbol{x}_M - \overline{\boldsymbol{x}}\|}\right) < \frac{\sqrt{2}}{2}$$
*for all $i, j$, $i \ne M$.*

*If*

$$(M+1)^2 < \frac{1}{3}\vartheta\exp\left(0.5n\sigma_0^4\right), \tag{15}$$

*then with probability $p > 1 - \vartheta$*

$$0.5 \le \frac{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2}{R_0^2} \le 1.5 \text{ and } \left(\frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}}}{R_0}, \frac{\boldsymbol{x}_j - \overline{\boldsymbol{x}}}{\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|}\right) < \frac{\sqrt{2}}{2}$$
*for all $i, j$, $i \ne j$.*

*In particular, if inequality (15) holds then the set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is linearly separable with probability $p > 1 - \vartheta$.*

The estimates (14), (15) are far from being optimal and can be improved. The main message here is their exponential dependence on $n$: the upper boundary of $M$ can grow with $n$ exponentially. Numerical experiments show that the equidistribution in cube is not worse, from the practical point of view, than the uniform distribution in a ball. To illustrate this, we empirically assessed linear separability of samples drawn from equidistributions in the unit $n$-cubes. For selected values of $n$ from the set $\{1, \dots, 5{,}000\}$ we generated 100 samples $S$ of $M = 20{,}000$ random points from $[0, 1]^n$. For each sample, a sub-sample $\underline{S} \subset S$ of $N = 4{,}000$ points was randomly chosen, and for each point $\boldsymbol{x}_i$ in this sub-sample linear functionals $l(\boldsymbol{x}) = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}, \boldsymbol{x} - \overline{\boldsymbol{x}}) - \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}\|^2$ were constructed. Sings of $l(\boldsymbol{x}_j)$, $\boldsymbol{x}_j \in S$, $\boldsymbol{x}_j \ne \boldsymbol{x}_i$ were calculated, and the numbers $N_-$ of instances when $l(\boldsymbol{x}_j) < 0$ where recorded. Empirical frequencies $N_-/N$ were then derived. Outcomes of this experiment are summarized in Fig. 2. These experiments demonstrate that the probability that a randomly selected point in a sample is linearly separable from the rest could be significantly higher than the simple exponential estimates provided. This, however, is not surprising as the estimates are based on the values of means and variances, and do not take into account other quantitative properties of the sample distribution.

In general position, a set of $n$ points in $\mathbb{R}^{n-1}$ is linearly separable. Therefore, if $n - 1$ or less points from $\mathcal{M} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{M-1}\}$ are not separated from $\boldsymbol{x}$ by the hyperplane $L$ (Fig. 1) then they can be separated by an additional hyperplane orthogonal to $L$. This means that $\boldsymbol{x}$ can be separated from the whole $\mathcal{M}$ by a conjunction of two

linear inequalities, $(\bullet, \boldsymbol{x}/\|\boldsymbol{x}\|) > r \,\&\, (\bullet, \boldsymbol{y}) > q$, for some $0 < r < 0$, $q > 0$, and $\boldsymbol{y}, (\boldsymbol{y}, \boldsymbol{x}) = 0$. This system can be considered as a cascade of two independent neurons (Gorban et al., 2016). The probability of such a *two-neuron separability* is higher than of linear separability. (Compare inequality (16) in the following theorem to (1).)

**Theorem 3.** *Let $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be a set of $M$ i.i.d. random points from the equidistribution in the unit ball $\mathbb{B}_n$, $0 < r < 1$. Then*

$$\begin{aligned}\mathbf{P}\left(\|\boldsymbol{x}_M\| > r \,\&\, \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_M}{\|\boldsymbol{x}_M\|}\right) < r \text{ for at least } M-n \text{ points } \boldsymbol{x}_i \in S\right)\\ \geq (1-r^n)(1-0.5\rho^n)^{M-1}\\ \times \left(1 - \frac{1}{n!}\left(\frac{0.5(M-n)\rho^n}{1-0.5\rho^n}\right)^n\right)\exp\left[\frac{0.5(M-n)\rho^n}{1-0.5\rho^n}\right],\end{aligned} \tag{16}$$

where $\rho = \sqrt{1-r^2}$.

For $r = 1/\sqrt{2}$, $n = 100$, and $M = 2,74 \cdot 10^6$, (16) gives: $\mathbf{P}\left(\|\boldsymbol{x}_M\| > r \& \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_M}{\|\boldsymbol{x}_M\|}\right) < r \text{ for at least } M-n \ \boldsymbol{x}_i \in S\right) \geq 1-\theta$ with $\theta < 5 \cdot 10^{-14}$. The probability stays close to 1 for much larger values of $M$, as setting $M = 7 \cdot 10^{16}$ results in the estimate: $\mathbf{P}\left(\|\boldsymbol{x}_M\| > r \& \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_M}{\|\boldsymbol{x}_M\|}\right) < r \text{ for at least } M-n \ \boldsymbol{x}_i \in S\right) \geq 1-\theta$ with $\theta < 5 \cdot 10^{-9}$.

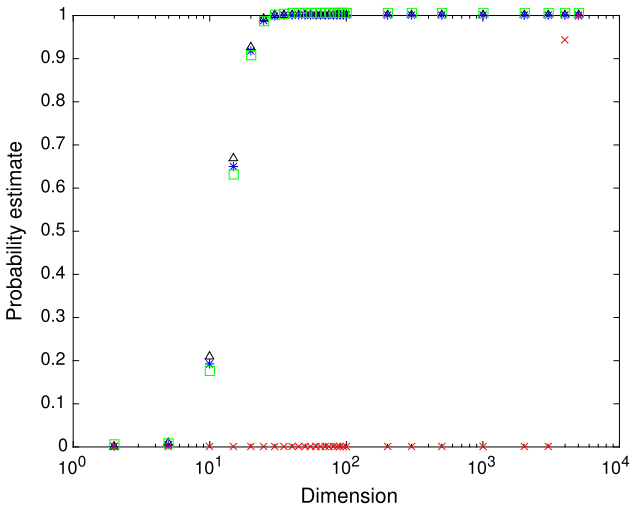


**Fig. 2.** Estimates of probabilities that a random point in a sample of 20000 points i.i.d. drawn from an equidistribution in the unit cube $[0, 1]^n$ in $\mathbb{R}^n$ is separable from the remaining points in the sample as a function of dimension $n$. Blue stars, black triangles, and green squares, are the means, maxima, and minima of $N_-/N$ over all 100 samples for each value of $n$. Red crosses show estimates (12).

## 4. Conclusion

Classical measure concentration theorems state that random points are concentrated in a thin layer near a surface (a sphere, an average or median level set of energy or another Lipschitz function, etc.). The *stochastic separation theorems* describe thin structure of these thin layers: the random points are not only concentrated in a thin layer but are all linearly separable from the rest of the set even for exponentially large random sets. The estimates are produced for two classes of distributions in high dimension: for equidistributions in balls or ellipsoids or for the product distributions with compact support (i.e. for the case when coordinates are bounded independent random variables). Numerous generalizations are possible, for example:

- Relax the requirement of independent coordinates in Theorem 2 to that of weakly dependent vector-valued variables;
- Instead of equidistributions, consider distributions with strongly log-concave probability densities;
- Use various simple and robust nonlinear classifiers like small neural cascades (compare to Theorem 3), algebraic classifiers and other families. For these generalizations, the VC dimension is expected to play the role similar to dimension $n$ in Theorems 1 and 2.

Stochastic separation Theorems 1–3 are important for synthesis and one-shot correction of AI systems. For example, inequalities (1) and (10) evaluate the probability that a randomly selected point $\boldsymbol{x}_M$ is linearly separable from all other $M-1$ points by the linear functional $l(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{x}_M - \bar{\boldsymbol{x}})$. This separation is needed to correct a mistake of a legacy AI system without any re-learning and modification of existing skills (Gorban et al., 2016). Such measure concentration effects reveal the hidden geometric background of the reported success of randomized neural networks models (Scardapane & Wang, 2017).

Stochastic separation theorems can simplify high-dimensional data analysis and generate the 'blessing of dimensionality' (Gorban, Tyukin, & Romanenko, 2016). For example, according to (6), in a dataset with 100 attributes and $M < 2.7 \cdot 10^6$ samples we should not be surprised to observe the linear separability of *each* sample from the rest of the database by the inequalities $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle < \sqrt{\frac{1}{2}\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}$ $(i \neq j)$ in the Mahalanobis inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = (\boldsymbol{x}, S^{-1}\boldsymbol{y})$, where $S$ is the empirical covariance matrix. The Mahalanobis inner product is used for 'whitening', i.e. for transformation of the data cloud into the spherical form. Of course, these attributes should not be highly correlated and the empiric covariance matrix should be invertible.

We analysed separation of random points from random sets. This is the problem of single *correction* of a legacy AI system. The question of generalizability of this correction is of great practical importance. It leads to a problem of *separation of two random sets*. A simple series of generalizations can be immediately produced from Theorems 1–3 for separation of an $M$-element random set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subset \mathbb{R}^n$ from a $k$-element one $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ for $k < n$. For this purpose, we can consider a linear space $E = \text{span}\{\boldsymbol{y}_i - \boldsymbol{y}_1 \mid i = 2, \ldots, k\}$ and study separation of a point from an $M$-element set in the projection onto the quotient space $\mathbb{R}^n/E$. If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are independent then separation would likely be limited to sets of small cardinality $k < n$. If, in contrast, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are pair-wise positively correlated then we can expect that a single functional would separate them from $S$, with reasonable probability even for some $k \geq n$. This naturally gives rise to generalization of corrections.

The reported extreme separation capabilities of linear functionals offer new insights into the Grandmother cell or concept cell phenomena that are broadly reported in neuroscience (Quian Quiroga et al., 2005; Viskontasa, Quian Quiroga, & Fried, 2009). The essence of the phenomenon is that some neurons in the human brain respond unexpectedly selective to particular persons or objects. Strikingly, not only is the brain able to respond selectively to "rare" individual stimuli but also such selectivity can be learnt very rapidly from a limited number of experiences (Ison, Quian Quiroga, & Fried, 2015). The question is: Why small ensembles of neurons may deliver such a sophisticated functionality reliably? Stochastic separation Theorems 1–3 provide a possible answer. If we accept that (a) linear functionals followed by nonlinear threshold-modulated response as phenomenological models of cells whose activity was measured, (b) the number of inputs converging to these cells is large enough, and (c) they are statistically independent, then extreme selectivity of responses of such models follows immediately from Theorem 2.

## Acknowledgements

Authors are grateful to M. Gromov and S. Utev for seminal questions and comments. The work was partially supported by Innovate UK (KTP009890 and KTP010522).

## References

Ball, K. (1997). An elementary introduction to modern convex geometry. *Flavors of Geometry*, *31*, 1–58.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 499–526.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Machine Learning*, *7*, 179–188.

Forney, A., Pearl, J., & Bareinboim, E. (2017). Counterfactual data-fusion for online reinforcement learners. A talk at the 1st Workshop on Transfer in Reinforcement Learning at AAMAS 2017, May 8–9, 2017, São Paulo, Brazil, Technical Report R-471, UCLA, June 2017. ftp://ftp.cs.ucla.edu/pub/stat_ser/r471.pdf.

Freund, Y., & Schapire, R. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, *37*, 277–296.

Gorban, A. N., Romanenko, I., Burton, R., & Tyukin, I. (2016). One-trial correction of legacy AI systems and stochastic separation theorems, arXiv:1610.00494 [stat.ML].

Gorban, A. N., Tyukin, I., Prokhorov, D., & Sofeikov, K. (2016). Approximation with random bases: Pro et contra. *Information Sciences*, *364–365*, 129–145.

Gorban, A. N., Tyukin, I. Y., & Romanenko, I. (2016). The blessing of dimensionality: Separation theorems in the thermodynamic limit. *IFAC-PapersOnLine*, *49*(24), 64–69.

Gromov, M. (2003). Isoperimetry of waists and concentration of maps. *GAFA, Geometric and Functional Analysis*, *13*, 178–215.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *301*, 13–30.

Ison, M., Quian Quiroga, R., & Fried, I. (2015). Rapid encoding of new memories by individual neurons in the human brain. *Neuron*, *87*, 220–230.

Kainen, P., & Kůrková, V. (1993). Quasiorthogonal dimension of euclidian spaces. *Applied Mathematics Letters*, *6*, 7–10.

Kůrková, V., & Sanguineti, M. (2007). Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets. *Discrete Applied Mathematics*, *155*, 1930–1942.

Le Cun, Y., & Bengio, T. (1995). Convolutional networks for images, speech, and time series. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 255–258). Cambridge, MA: MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Minsky, M. L., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Natarajan, B. (1989). On learning sets and functions. *Machine Learning*, *4*, 67–97.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books.

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. McClelland, & The PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 25–40). Cambridge, MA: MIT Press.

Scardapane, S., & Wang, D. (2017). Randomness in neural networks: an overview. *WIREs Data Mining and Knowledge Discovery*, *7*, e1200. http://dx.doi.org/10.1002/widm.1200.

Simon, B. (2011). *Cambridge tracts in mathematics*: *vol. 187*. *Convexity, an analytic viewpoint*. Cambridge, UK: Cambridge University Press.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. Springer-Verlag.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer-Verlag.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, *16*, 264–280.

Viskontasa, I., Quian Quiroga, R., & Fried, I. (2009). Human medial temporal lobe neurons respond preferentially to personally relevant images. *Proceedings of the National Academy of Sciences*, *120*, 21329–21334.