Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-24 (2016) 064–069

The Blessing of Dimensionality: Separation Theorems in the Thermodynamic Limit ⋆

Alexander N. Gorban * **Ivan Yu. Tyukin** ** **Ilya Romanenko** ***

* *University of Leicester, Department of Mathematics, UK (e-mail: ag153@le.ac.uk).*
** *University of Leicester, Department of Mathematics, UK (e-mail: I.Tyukin@le.ac.uk)*
*** *Apical LTD, UK (e-mail: ilya@apical.co.uk)*

Abstract: We consider and analyze properties of large sets of randomly selected (i.i.d.) points in high dimensional spaces. In particular, we consider the problem of whether a single data point that is randomly chosen from a finite set of points can be separated from the rest of the data set by a linear hyperplane. We formulate and prove stochastic separation theorems, including: 1) with probability close to one a random point may be separated from a finite random set by a linear functional; 2) with probability close to one for every point in a finite random set there is a linear functional separating this point from the rest of the data. The total number of points in the random sets are allowed to be exponentially large with respect to dimension. Various laws governing distributions of points are considered, and explicit formulae for the probability of separation are provided. These theorems reveal an interesting implication for machine learning and data mining applications that deal with large data sets (big data) and high-dimensional data (many attributes): simple linear decision rules and learning machines are surprisingly efficient tools for separating and filtering out arbitrarily assigned points in large dimensions.

Keywords: Measure concentration, separation theorems, big data, machine learning.

1. INTRODUCTION

Curse of dimensionality is a widely known problem. The term, originally introduced by R. Bellman in relation to complications occurring in dynamic programming (Bellman, 1957), has now become a common name for issues of both theoretical and computational nature arising in high dimensions. An example of one particular issue is vastness off high-dimensional spaces. Indeed, suppose that we were to fill in a 100-dimensional unit cube with an extremely coarse accuracy of just 2 measurements or samples per each dimension. It turns out that we are highly unlikely to complete this task due to that the amount of data one would have to acquire and transmit is $2^{100} \approx 10^{30}$. If we partition this hypercube into a union of disjoint cubes with the edge length equal to 1/2 and imagine this object as an abstract storage in which each smaller hypercube is an individual storage cell and use it to store data then almost all of the 2^{100} pieces will remain empty for any dataset currently available.

One the other hand, certain tasks become much simpler in high dimensions. Statistical physics gives a great example of such simplification. Existence of entropy in macroscopic physics is a manifestation of the ensemble equivalence, that is an essentially high-dimensional phenomenon (Gibbs, 1960 [1902]). In particular, equidistribution on a sphere (microcanonical ensemble) in high dimensions is equivalent to equidistribution in the corresponding ball inside the sphere and, at the same time, to the normal distribution (canonical distribution). The notions of order and disorder and the possibility of order/disorder separation appear also because of special multidimensional *measure concentration* phenomena (Gorban, 2007).

Perhaps, the first mathematician who recognized both beauty and utility of these phenomena (several decades after Maxwell, Boltzmann, Gibbs and Einstein) was Paul Lévy. He named them 'concentration phenomena' and described them in detail in his seminal book (Lévy, 1951). The first example was: *concentration of the volume of a ball near its surface (sphere)*. Let $V_n(1)$ be a volume of a unit ball in $\mathbb{R}^n$. The volume of a ball of radius r in $\mathbb{R}^n$ is $r^n V_n(1)$. Hence, the volume of a ball of radius $1-\varepsilon$ is $(1-\varepsilon)^n V_n(1)$. For small ε, $(1-\varepsilon)^n \approx \exp(-\varepsilon n)$. The fraction of the volume of the unit ball in the ε-vicinity of the sphere is

$$1-(1-\varepsilon)^n \approx 1-\exp(-\varepsilon n). \tag{1}$$

This fraction asymptotically approaches 1 when $n \to \infty$.

Another important example is *waist concentration*: the surface area of a high-dimensional sphere is concentrated in a small vicinity of its equator (for general theory see (Gromov, 2003)). Therefore, with probability close to 1 two random vectors from an equidistribution on a multidimensional sphere are almost orthogonal with cosine of the angle between them close to zero. Moreover, n such vectors in $\mathbb{R}^n$ with probability close to 1 form an almost orthogonal basis. The number of pairwise almost orthogonal vectors may be much bigger than the dimension

⋆ The work is partially supported by Innovate UK, Technology Strategy Board, Knowledge Transfer Partnership grant KTP009890

2405-8963 © 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.10.755

n. In particular, it has been shown in (Gorban et al., 2016) that if one randomly chooses $M \leq N_n$ vectors on an n-dimensional sphere, where the bound N_n is dependent on the dimension n, then with probability close to 1 they all are pairwise almost orthogonal. The value of N_n grows exponentially with n. For these theorems, see (Gorban et al., 2016). Existence of such almost orthogonal bases was shown in (Kainen and Kurkova, 1993).

Other examples of concentration phenomena include (but are not limited to) famous Johnson and Lindenstrauss result (Johnson and Lindenstrauss, 1984), error bound in machine learning (Vapnik, 2000), and function approximation (Pao et al., 1994), (Rahimi and Recht, 2008a), (Rahimi and Recht, 2008b). This contribution aims to further explore advantages offered by concentration phenomena in machine learning and data analysis applications.

Here we consider the problem of what is the probability that a point or a set of points are separable from a given set, i.e. the data, by linear functionals and their cascades. Notwithstanding the relevance of this problem for machine learning, separation theorems are important tools in convex geometry and functional analysis. The famous Hahn-Banach separation theorem for real spaces states that if L is a locally convex topological vector space space, X is compact, and Y is closed set and X, Y are convex, then there exists a continuous linear functional l on V such that $l(x) < t < s < l(y)$ for some real numbers t, s and for all $x \in X$ and $y \in Y$.

We formulate and prove the *stochastic separation theorem* which demonstrates that in a high dimensional finite i.i.d. sample with high probability every point can be linearly separated from the set of all other points. The cardinality of the sample for which the theorem holds is allowed to be exponential in dimension. This and other results of our present contribution reveal surprising and rather unexpected benefits offered by high-dimensionality to theory and practice of high dimensional data mining. Of course, high dimensionality of data may cause difficulties in analysis of data, for example, in organization of similarity search (Pestov, 2000). This is the curse of dimensionality. But at the same time concentration phenomena may bring a range of rewards too, including linear separability. We can call this effect *blessing of dimensionality* (cf. (Anderson et al., 2014), (Chen et al., 2013)).

NOTATION

Throughout the paper the following notational agreements are used.

- $\mathbb{R}$ denotes the field of real numbers;
- $\mathbb{N}$ is the set of natural numbers;
- $\mathbb{R}^n$ stands for the n-dimensional linear space over the field of reals; unless stated otherwise symbol n is reserved to denote dimension of the underlying linear space;
- let $x \in \mathbb{R}^n$, then $\|x\|$ is the Euclidean norm of x: $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$;
- $B_n(R)$ denotes a n-ball of radius R centered at 0: $B_n(R) = \{x \in \mathbb{R}^n | \ \|x\| \leq R\}$;
- $\mathcal{V}(\Xi)$ is the Lebesgue volume of $\Xi \subset \mathbb{R}^n$;
- $\mathcal{M}$ is an i.i.d. sample equidistributed in $B_n(1)$;
- M is the number of points in $\mathcal{M}$, or simply the cardinality of the set $\mathcal{M}$.

2. SEPARATION OF A POINT FROM A FINITE RANDOM SET IN HIGH DIMENSIONS

Here and throughout the paper our basic example is an equidistribution on the unit ball $B_n(1)$ in $\mathbb{R}^n$.

Definition 1. Let X and Y be subsets of $\mathbb{R}^n$. We say that a linear functional l on $\mathbb{R}^n$ separates X and Y if there exists a $t \in \mathbb{R}$ such that

$$l(x) > t > l(y) \ \forall \ x \in X, \ y \in Y.$$

Let $\mathcal{M}$ be an i.i.d. sample drawn from the equidistribution on the unit ball $B_n(1)$. We begin with evaluating the probability that a single element x randomly and independently selected from the same equidistribution can be separated from $\mathcal{M}$ by a linear functional. This probability, denoted as $P_1(\mathcal{M}, n)$, is estimated in the theorem below.

Theorem 2. Consider an equidistribution in a unit ball $B_n(1)$ in $\mathbb{R}^n$, and let $\mathcal{M}$ be an i.i.d. sample from this distribution. Then

$$\begin{aligned} P_1(\mathcal{M}, n) &\geq \max_{\varepsilon \in (0,1)} (1 - (1-\varepsilon)^n) \left(1 - \frac{\rho(\varepsilon)^n}{2}\right)^M, \\ \rho(\varepsilon) &= (1 - (1-\varepsilon)^2)^{\frac{1}{2}} \end{aligned} \tag{2}$$

Proof of Theorem 2. The proof of the theorem is contained mostly in the following lemma

Lemma 3. Let $\mathcal{M}$ be an i.i.d. sample from an equidistribution on a unit ball $B_n(1)$. Let $x \in \mathbb{R}^n$ be a point inside the ball with $1 > \|x\| > 1 - \varepsilon > 0$. Then

$$P\left(y \in \mathcal{M} \,\middle|\, \left\langle \frac{x}{\|x\|}, y \right\rangle < 1 - \varepsilon \right) \geq 1 - \frac{\rho(\varepsilon)^n}{2}. \tag{3}$$

Proof of Lemma 3. Recall that (Lévy, 1951): $\mathcal{V}(B_n(r)) = r^n \mathcal{V}(B_n(1))$ for all $n \in \mathbb{N}$ $r > 0$. The point x is inside the spherical cap $C_n(\varepsilon)$:

$$C_n(\varepsilon) = B_n(1) \cap \left\{ \xi \in \mathbb{R}^n \,\middle|\, \left\langle \frac{x}{\|x\|}, \xi \right\rangle > 1 - \varepsilon \right\}. \tag{4}$$

The volume of this cap can be estimated from above (see Fig. 1) as

$$\mathcal{V}(C_n(\varepsilon)) \leq \frac{1}{2} \mathcal{V}(B_n(1)) \rho(\varepsilon)^n. \tag{5}$$

The probability that the point $y \in \mathcal{M}$ is outside of $C_n(\varepsilon)$ is equal to $1 - \mathcal{V}(C_n(\varepsilon))/\mathcal{V}(B_n(1))$. Estimate (3) now immediately follows from (5). $\square$

Let us now return to the proof of the theorem. If x is selected independently from the equidistribution on $B_n(1)$ then the probabilities that $x = 0$ or that it is on the boundary of the ball are 0. Let $x \neq 0$ be in the interior of $B_n(1)$. According to Lemma 3, the probability that a linear functional l separates x from a point $y \in \mathcal{M}$ is larger than $1 - 1/2\rho(\varepsilon)^n$. Given that points of the set $\mathcal{M}$ are i.i.d. in accordance to the equidistribution on $B_n(1)$, the probability that l separates x from $\mathcal{M}$ is no smaller than $(1 - 1/2\rho(\varepsilon)^n)^M$.

On the other hand

$$P(x \in B_n(1) | 1 > \|x\| > 1 - \varepsilon) = (1 - (1-\varepsilon)^n).$$

Given that x and $y \in \mathcal{M}$ are independently drawn from the same equidistribution and that the probabilities of

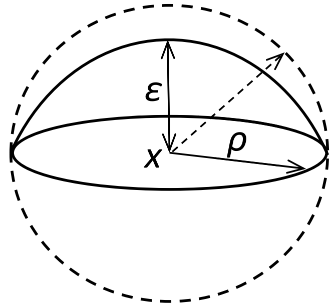

Fig. 1. A test point x and a spherical cap on distance ε from the surface of unit sphere. Escribed sphere of radius ρ is showed by dashed line.

randomly selecting the point x exactly on the boundary of $B_n(1)$ or in its centre are zero, we can conclude that

$$P_1(\mathcal{M}, n) \geq (1-(1-\varepsilon)^n)(1-1/2\rho(\varepsilon)^n)^M. \quad (6)$$

Finally, noticing that (6) holds for all $\varepsilon \in (0,1)$ including the value of ε maximizing the rhs of (6), we can conclude that (2) holds true too. □

Remark 4. For $\rho(\varepsilon)^n$ small (i.e. $\rho(\varepsilon)^n \ll 1$) the term $\left(1-\frac{\rho^n}{2}\right)^M$ can be approximated as $(1-\rho^n)^M \approx e^{-M\frac{\rho^n}{2}}$. Thus estimate (2) becomes

$$P_1(\mathcal{M}, n) \gtrsim \max_{\varepsilon\in(0,1)} (1-(1-\varepsilon)^n)e^{-M\frac{\rho^n}{2}},\ \rho(\varepsilon)^n \ll 1. \quad (7)$$

To see how large the probability $P_1(\mathcal{M}, n)$ could become for already rather modest dimensions, e.g. for $n = 50$, we estimate the right-hand side of (7) from below by letting $\varepsilon = 1/5$ and $\rho = 3/5$:

$$P_1(\mathcal{M}, 50) \gtrsim 0.99998 \exp(-4\times 10^{-12} M).$$

For $M \leq 10^9$ this estimate gives $P_1(M, 50) \gtrsim 0.996$. Thus in dimension 50 (and higher) a random point is linearly separable from a random set of 10^9 points with probability 0.996.

3. EXTREME POINTS OF A RANDOM FINITE SET

So far we have discussed the question of separability of a single random point x, drawn i.i.d. from the equidistribution on $B_n(1)$, from a random i.i.d. sample $\mathcal{M}$ drawn from the same distribution. In practice, however, the data or a training set are given or fixed. It is thus important to know if the "point" linear separability property formulated in Theorem 2 persists (in one form or another) when the test point x belongs to the sample $\mathcal{M}$ itself. In particular, the question is if the probability $P_M(\mathcal{M}, n)$ that each point $y \in \mathcal{M}$ is linearly separable from $\mathcal{M}\backslash\{y\}$ is close to 1 in high dimensions? If such a property does hold then one could conclude that in high dimensions with probability close to 1 all points of $\mathcal{M}$ are the vertices (extreme points) of the convex hull of $\mathcal{M}$ and none of $y \in \mathcal{M}$ is a convex combination of other points. The fact that this is indeed the case follows from Theorem 5

Theorem 5. Consider an equidistribution in a unit ball $B_n(1)$ in $\mathbb{R}^n$, and let $\mathcal{M}$ be an i.i.d. sample from this distribution.Then

$$P_M(\mathcal{M}, n) \geq$$
$$\max_{\varepsilon\in(0,1)} \left[(1-(1-\varepsilon)^n)\left(1-(M-1)\frac{\rho(\varepsilon)^n}{2}\right)\right]^M. \quad (8)$$

Proof of Theorem 5. Let $P : \mathcal{F} \to [0,1]$ be a probability measure and $A_i \in \mathcal{F}$, $i = 1, \ldots, M$. It is well-known that

$$P(A_1 \vee A_2 \vee \ldots \vee A_M) \leq \sum_{i=1}^{M} P(A_i) \quad (9)$$

The probability that a test point y is in the ε-vicinity of the boundary of $B_n(1)$ is $1-(1-\varepsilon)^n$. Fix $y \in \mathcal{M}$ and construct spherical caps $C_n(\varepsilon)$ for each element in $\mathcal{M}\backslash\{y\}$ as specified by (4) but with x replaced by the corresponding points from $\mathcal{M}\backslash\{y\}$. According to (9), the probability that y is in any of these caps is no larger than $(M-1)\frac{\rho(\varepsilon)^n}{2}$. Hence the probability that a point $y \in \mathcal{M}$ is separable from $\mathcal{M}\backslash\{y\}$ is larger or equal to $(1-(1-\varepsilon)^n)(1-(M-1)\frac{\rho(\varepsilon)^n}{2})$. Given that points of $\mathcal{M}$ are drawn independently and that there are exactly M points in $\mathcal{M}$, the probability that every single point is linearly separable from the rest satisfies (8). □

Remark 6. Note that employing (9) one can obtain another estimate of P_M:

$$P_M(\mathcal{M}, n) \geq 1 - M(1 - P_1(\mathcal{M}, n)).$$

Similar to the example discussed in the end of the previous section, let us evaluate the right-hand side of (8) for some fixed values of n and M. If $n = 50$, $M = 1000$, and $\varepsilon = 1/5$, $\rho = 3/5$ then this estimate gives: $P_M(\mathcal{M}, 50) > 0.985$.

4. OTHER DISTRIBUTIONS

The nature of phenomenon described is universal and does not depend on many details of the distribution. In some sense, it should be just essentially high-dimensional. The assumption about independence and identical distribution (i.i.d.) of points in $\mathcal{M}$ could be also relaxed to weak dependence between points and not much different distributions of them. Due to space limitations cannot review all these generalizations here. Nevertheless, we will give one example: equidistribution in an ellipsoid.

Let points in the set $\mathcal{M}$ be selected by independent trials taken from the equidistribution in a n-dimensional ellipsoid. Without loss of generality, we present this ellipsoid in the orthonormal eigenbasis:

$$E_n = \{x \in \mathbb{R}^n | \sum_{i=1}^{n} \frac{x_i^2}{c_i^2} \leq 1\}, \quad (10)$$

where rc_i are the semi-principal axes. The linear transformation $(x_1, \ldots, x_n) \mapsto \left(\frac{x_1}{c_1}, \ldots, \frac{x_n}{c_n}\right)$ transforms the ellipsoid into the unit ball. The volume of every set in the new coordinates scales with the factor $1/\prod_i c_i$. Therefore the ratio of two volumes does not change, and the equidistribution in the ellipsoid is transformed into the equidistribution in the unit ball. Hyperplanes are transformed into hyperplanes and the property of linear separability is not affected by a nonsingular linear transformation. Thus, the following corollaries from Theorems 2, 5 hold.

Corollary 7. Let $\mathcal{M}$ be formed by a finite number of i.i.d. trials taken from the equidistribution in a n-dimensional ellipsoid E_n (10), and let x be a test point drawn independently from the same distribution. Then x can be separated from $\mathcal{M}$ by a linear functional with probability $P_1(\mathcal{M}, n)$:

$$P_1(\mathcal{M}, n) \geq (1 - (1 - \varepsilon)^n) \left(1 - \frac{\rho(\varepsilon)^n}{2}\right)^M .$$

Corollary 8. Let $\mathcal{M}$ be formed by a finite number of i.i.d. trials taken from the equidistribution in a n-dimensional ellipsoid E_n. With probability

$$P_M(\mathcal{M}, n) \geq \left[(1 - (1 - \varepsilon)^n) \left(1 - (M-1)\frac{\rho(\varepsilon)^n}{2}\right)\right]^M$$

each point y from the random set $\mathcal{M}$ can be separated from $\mathcal{M}\backslash\{y\}$ by a linear functional.

5. SEPARATION BY SMALL NEURAL NETWORK

Separation of points and sets by linear functionals is exactly their separation by one-neuron systems (or elementary perceptrons (Rosenblatt, 1962)). At this point, machine learning and functional analysis perspectives coincide but the next step, the multineuron systems, is specific to machine learning. Consider a topological real vector space L. For every continuous linear functional l on L and a real number θ we define an open *elementary neural predicate* $l(x) > \theta$ and a closed elementary predicate $l(x) \geq \theta$ ('open' and 'closed' are sets defined by predicates). The elementary neural predicate is true (=1) or false (=0) depending on whether the correspondent inequality holds. The negation of an open elementary neural predicate is a closed elementary neural predicate (and converse). A compound predicate is a Boolean expression composed of elementary predicates.

Definition 9. A set $X \subset L$ is k-neuron separable from a set $Y \subset L$ if there exist k elementary neural predicates $P_1, \ldots, P_k$ and a composed of them compound predicate, a Boolean function $B(x) = B(P_1(x), \ldots, P_k(x))$, such that $B(x)$ =true for $x \in X$ and $B(x)$ =false for $x \in Y$.

If X consist of one point then the compound predicate in the definition of k-neuron separability can be selected in a form of conjunction of k elementary predicates $B(x) = P_1(x)\&\ldots\&P_k(x)$):

Proposition 10. Let $X = \{z\}$ and X is k-neuron separable from $Y \subset L$. Then there exist k elementary neural predicates $P_1, \ldots, P_k$ such that the Boolean function $B(x) = P_1(x)\&\ldots\&P_k(x)$ is true on X and false on Y.

Proof. Assume that there exist k elementary neural predicates $Q_1, \ldots, Q_k$ and a Boolean function $B(Q_1, \ldots, Q_k)$ such that B is true on x and false on Y. Represent B in a disjunctive normal form as a disjunction of conjunctive clauses:

$$B = C_1 \vee C_2 \vee \ldots \vee C_N,$$

where each C_i has a form $C_i = R_1\&R_2\&\ldots\&R_k$ and $R_i = Q_i$ or $R_i = \neg Q_i$.

At least one conjunctive clause $C_i(x)$ is true because $B(x)$ =true. On the other hand, $C_i(y)$ =false for all $y \in Y$ and $i = 1, \ldots, N$ because their disjunction $B(y)$=false. Let $C_i(x)$=true. Then C_i separates x from Y and other conjunctive clauses are not necessary in B. We can take $B = C_i = R_1\&R_2\&\ldots\&R_k$, where $R_i = Q_i$ or $R_i = \neg Q_i$. Finally, we take $P_i = R_i$. □

In contrast to Hahn-Banach theorems of linear separation, the k-neuron separation of sets does not require any sort of convexity. Weak compactness is sufficient. Assume that continuous linear functionals separate points in L. Let $Y \subset L$ be a weakly compact set (Schaefer, 1999) and $x \notin Y$.

Proposition 11. x is k-neuron separable from Y for some k.

Proof. For each $y \in Y$ there exists a continuous linear functional l_y on L such that $l_y(x) = l_y(y) + 2$ because continuous linear functionals separate points in L. Inequality $l_y(z) < l_y(x) - 1$ defines an open half-space $L_y^< = \{z \mid l_y(z) < l_y(y) + 1\}$, and $y \in L_y^<$. The collection of sets $\{L_y^< \mid y \in Y\}$ forms an open cover of Y. Each set $L_y^<$ is open in weak topology and Y is weakly compact, hence there exists a final covering of Y by sets $L_y^<$: $Y \subset \bigcup_{i=1,\ldots,k} L_{y_i}^<$ for some finite subset $\{y_1, \ldots, y_k\} \subset Y$. The inequality $l_{y_i}(x) > l_{y_i}(y_i) + \alpha$ holds for all $i = 1, \ldots, k$ and $\alpha < 2$. Let us select $1 < \alpha < 2$ and take elementary neural predicates $P_i^\alpha(x) = (l_{y_i}(x) > l_{y_i}(y_i) + \alpha)$. The conjunction $P_1^\alpha\&\ldots\&P_k^\alpha$ is true on x. Each point $y \in Y$ belongs to $L_{y_i}^<$ for at least one $i = 1, \ldots, k$. For this i, $P_i^\alpha(y)$=false. Therefore, $P_1^\alpha\&\ldots\&P_k^\alpha$ is false on Y. Hence, this conjunction separates x from Y. □

Moreover, We can select two numbers $1 < \beta < \alpha < 2$. Both $P_1^\alpha\&\ldots\&P_k^\alpha$ and $P_1^\beta\&\ldots\&P_k^\beta$ separate x from Y.

The theorem about k-neuron separation of weakly compact sets follows from Proposition 11. Let X and Y be weakly compact subsets of L and continuous linear functionals separate points in L

Theorem 12. X is k-neuron separable from Y for some k.

Proof. For each $x \in X$ apply the construction from the proof of Proposition 11 and construct the conjunction $B_x = P_1^\alpha\&\ldots\&P_k^\alpha$, which separates x from Y. Every set of the form $F_x = \{z \in L \mid B_x(z) = \text{true}\}$ is intersection of a finite number of open half-spaces and, therefore, is open in weak topology. The collection of sets $\{F_x \mid x \in X\}$ covers X. There exist a final cover of X by sets F_x:

$$X \subset \bigcup_{j=1,\ldots,g} F_{x_j}$$

for some finite set $\{x_1, \ldots, x_l\} \subset X$. The composite predicate that separates X from Y can be chosen in the form $B(z) = B_{x_1}(z) \vee \ldots \vee B_{x_l}(z)$. □

6. TWO-NEURON SEPARATION IN FINITE SETS

So far we have provided estimates of the probabilities that a single linear classifier or a learning machine can separate a given point from the rest of data and showed that two disjoint weakly compact subsets of a topological vector space can be separated by small networks of perceptrons. Let us now see how employing small networks may improve probabilities of separation of a point from the rest of the

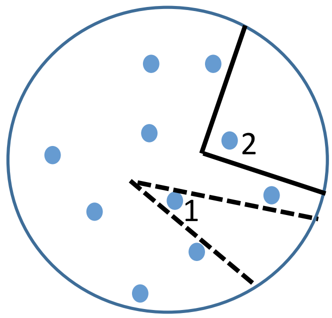

Fig. 2. Two-neuron separation in finite sets. Every point can be separated by a sufficiently acute angle or highly correlated neurons: Point 1 is separated from other points by an acute angle (dashed lines). Point 2 is separated by a right angle (non-correlated neurons), but cannot be separated by a linear functional (i.e. by a straight line).

data in high dimensions. In particular, we will consider the case of a two-neuron separation in which the network is a simple cascade comprised of 2 perceptrons followed by a conjunction operation.

Before, however, going any further we need to somewhat clarify and adjust the notion of separability of a point from a finite data set by a network so that the question makes any practical and theoretical sense. Consider, for instance, the problem of separating a test point by just two perceptrons. If one projects the data onto a 2d plane so that projections of the test point and any other point from the rest of the data do not coincide then the problem always has a solution. This is illustrated with the diagram in Fig. 2. According to this diagram any given point from an arbitrary but finite data set could be cut out from the rest of the data by just two lines that intersect at a sufficiently acute angle. Thus two hyperplanes $\{x \mid l_1(x) = \theta_1\}$ and $\{x \mid l_2(x) = \theta_2\}$ whose projections onto the 2d plane are exactly these two lines already constitute the two-neuron separating cascade.

The problem with this solution is that if the acute angle determined by the inequalities $l_1(x) > \theta_1$, $l_2(x) > \theta_2$ is small then the coefficients of the corresponding linear functionals $l_1(x)$ and $l_2(x)$, i.e. synaptic weights of the neurons, are highly correlated. Their Pearson correlation coefficient is close to -1. This property implies that small changes in the coefficients of one perceptron should be accounted for in the coefficients of the other. Failing to do so may result in that the desired separation will be lost. Further to this, finding separations of the form shown in Fig. 2 requires prior knowledge of the data geometry, or at least that of its projections onto a suitable 2d plane.

Let us now analyse the problem of separation of a random i.i.d. finite sample $\mathcal{M}$ drawn from an equidistribution in $B_n(1)$ from a point x drawn independently from the same distribution by two non-correlated neurons. More formally, we are interested in the probability $\mathcal{P}_1(\mathcal{M}, n)$ that a two-neuron cascade with uncorrelated synaptic weights separates x from $\mathcal{M}$. An estimate of this probability is provided in the next theorem.

Theorem 13. Consider an equidistribution in a unit ball $B_n(1)$ in $\mathbb{R}^n$, and let $\mathcal{M}$ be an i.i.d. sample from this distribution. Then

$$\mathcal{P}_1(\mathcal{M}, n) \geq \max_{\varepsilon\in(0,1)} (1-(1-\varepsilon)^n)\times$$
$$\left(1-\frac{\rho(\varepsilon)^n}{2}\right)^M e^{(M-n+1)\left[\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right]}\times \qquad (11)$$
$$\left(1-\frac{1}{n!}\left((M-n+1)\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right)^n\right).$$

Proof. Observe that in the case of general position, a single neuron (viz. linear functional) separates $n+1$ points with probability 1. This means that if no more than $n-1$ points from $\mathcal{M}$ are in the spherical cap $C_n(\varepsilon)$ corresponding to the test point then the second perceptron whose weights are orthogonal to the first one will filter out these additional spurious $n-1$ points with probability 1.

Let p_c be the probability that a point from $\mathcal{M}$ falls within the spherical cap $C_n(\varepsilon)$. Then the probability that only up to $n-1$ points of $\mathcal{M}$ will be in the cap $C_n(\varepsilon)$ is

$$\mathcal{P}(M,n) = \sum_{k=0}^{n-1}\binom{M}{k}(1-p_c)^{M-k}p_c^k$$

Observe that $\mathcal{P}(M,n)$, as a function of p_c, is monotone and non-increasing on the interval $[0,1]$, with $\mathcal{P}(M,n)=0$ at $p_c=1$ and $\mathcal{P}(M,n)=1$ at $p_c=0$. Hence taking estimate (5) into account one can conclude:

$$\mathcal{P}(M,n) \geq \sum_{k=0}^{n-1}\binom{M}{k}\left(1-\frac{\rho(\varepsilon)^n}{2}\right)^{M-k}\left(\frac{\rho(\varepsilon)^n}{2}\right)^k.$$

Noticing that

$$\sum_{k=0}^{n-1}\binom{M}{k}\left(1-\frac{\rho(\varepsilon)^n}{2}\right)^{M-k}\left(\frac{\rho(\varepsilon)^n}{2}\right)^k =$$
$$\left(1-\frac{\rho(\varepsilon)^n}{2}\right)^M\sum_{k=0}^{n-1}\binom{M}{k}\left(\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right)^k,$$

and bounding $\binom{M}{k}$ from above and below as $\frac{(M-n+1)^k}{k!} \leq \binom{M}{k} \leq \frac{M^k}{k!}$ for $0 \leq k \leq n-1$ we obtain

$$\mathcal{P}(\mathcal{M},n) \geq \left(1-\frac{\rho(\varepsilon)^n}{2}\right)^M\sum_{k=0}^{n-1}\frac{1}{k!}\left(\frac{(M-n+1)\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right)^k.$$

Invoking Taylor's expansion of e^x at $x=0$ with the Lagrange remainder term:

$$e^x = \sum_{k=0}^{n-1}\frac{x^k}{k!}+\frac{x^n}{n!}e^{\xi},\ \xi\in[0,x],$$

we can conclude that

$$\sum_{k=0}^{n-1}\frac{x^k}{k!} \geq e^x\left(1-\frac{x^n}{n!}\right)$$

for all $x \geq 0$. Hence

$$\mathcal{P}(\mathcal{M},n) \geq \left(1-\frac{\rho(\varepsilon)^n}{2}\right)^M e^{(M-n+1)\left[\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right]}\times$$
$$\left(1-\frac{1}{n!}\left((M-n+1)\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right)^n\right).$$

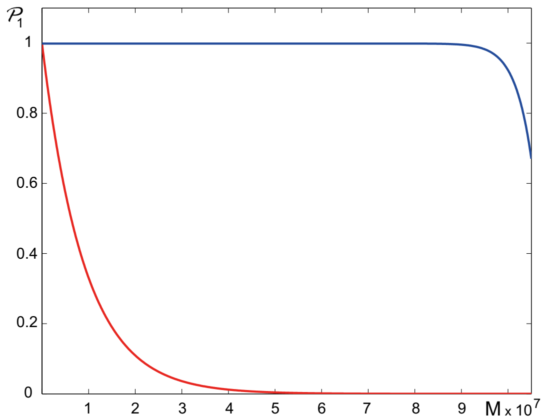

Fig. 3. Illustration to Theorem 13. Blue line shows an estimate of the rhs of (11) as a function of M at $\varepsilon = 1/5$, $\rho(\varepsilon) = 3/5$, and $n = 30$. Red line depicts an estimate of the rhs of (2) as a function of M for the same values of ε, $\rho(\varepsilon)$, and n.

Finally, given that the probability that the test point x is in the ε-vicinity of the boundary of $B_n(1)$ is at least $(1-(1-\varepsilon)^n)$ and that x is independently selected form the same eqidistribution as the set $\mathcal{M}$, we obtain $\mathcal{P}_1(\mathcal{M}, n) \geq (1-(1-\varepsilon)^n)\mathcal{P}(\mathcal{M}, n)$. This in turn implies that (11) holds. □

At the first glance estimate (11) looks more complicated as compared to e.g. (2). Yet, it differs from the latter by mere two factors. The first factor

$$\left(1 - \frac{1}{n!}\left((M-n+1)\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right)^n\right)$$

is close to 1 for $(M-n+1)\frac{\rho(\varepsilon)^n}{2} < 1$ and n sufficiently large. The second factor:

$$e^{(M-n+1)\left[\frac{\frac{\rho(\varepsilon)^n}{2}}{1-\frac{\rho(\varepsilon)^n}{2}}\right]},$$

is more important. It compensates for the decay of the probability of separation due to the term $(1-\frac{\rho(\varepsilon)^n}{2})^M$ keeping the rhs of (11) close to 1 over large interval of values of M. Significance and extent of this effect is illustrated with Fig. 3. The probability $\mathcal{P}_M(\mathcal{M}, n)$ that *each* point from $\mathcal{M}$ can be separated from other points by two uncorrelated neurons can be estimated like in Remark 6: $\mathcal{P}_M(\mathcal{M}, n) \geq 1 - M(1-\mathcal{P}_1(\mathcal{M}, n))$.

7. CONCLUSION

In this work we discussed and analyzed properties of large sets of randomly drawn i.i.d. data in high dimensional space. We showed that high dimensionality of data can play a major and positive role in various machine learning and data analysis tasks, including problems of separation, filtration, and selection. In contrast to naive intuition suggesting that high dimensionality of data more often than not brings in additional complexity and uncertainty, our findings reveal that high data dimensionality may constitute a valuable blessing too. This blessing, formulated here in the form of several stochastic separation theorems, offers several new insights for big data analysis.

After Maxwell and Gibbs, statistical thermodynamics is the theory of measure concentration phenomena in physics. For over a century the spectrum of its applications was limited mostly to various problems in physics. Here we highlight a new and important domain of applications for this classical filed. This is statistical thermodynamics of big data. All our results belong to this area.

REFERENCES

Anderson, J., Belkin, M., Goyal, N., Rademacher, L., and Voss, J. (2014). The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 35, 1–30.

Bellman, R.E. (1957). *Dynamic programming.* Princeton University Press, Princeton, NJ.

Chen, D., Cao, X., Wen, F., and Sun, J. (2013). Blessing of dimensionality: High dimensional feature and its efficient compression for face verification. In *CVPR 2013.*

Gibbs, J. (1960 [1902]). *Elementary Principles in Statistical Mechanics, developed with especial reference to the rational foundation of thermodynamics.* Dover Publications, New York.

Gorban, A. (2007). Order-disorder separation: Geometric revision. *Physica A*, 374, 85–102.

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

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

Johnson, W. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26, 189–206.

Kainen, P. and Kurkova, V. (1993). Quasiorthogonal dimension of Euclidian spaces. *Appl. Math. Lett.*, 6(3), 7–10.

Lévy, P. (1951). *Problèmes concrets d'analyse fonctionnelle.* Gauthier-Villars, Paris, second edition.

Pao, Y., Park, G., and Sobajic, D. (1994). Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*, 6(2), 163–180.

Pestov, V. (2000). On the geometry of similarity search: dimensionality curse and concentration of measure. *Inform. Process. Lett.*, 73(1–2), 47–51.

Rahimi, A. and Recht, B. (2008a). Uniform approximation of functions with random bases. In *Proceedings of the 46th IEEE Annual Allerton Conference on Communication, Control and Computing*, 555–561.

Rahimi, A. and Recht, B. (2008b). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 2, 1316–1323.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books.

Schaefer, H. (1999). *Topological Vector Spaces.* Springer, New York.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory.* Springer-Verlag.