Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online 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

$$1-(1-\varepsilon)^n \approx 1-\exp(-\varepsilon n). \qquad (1)$$

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

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

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

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). □

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