PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Review

Cite this article: Gorban AN, Tyukin IY. 2018 Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Phil. Trans. R. Soc. A* **376**: 20170237. http://dx.doi.org/10.1098/rsta.2017.0237

Accepted: 4 January 2018

One contribution of 14 to a theme issue 'Hilbert's sixth problem'.

Subject Areas:
applied mathematics, artificial intelligence, statistical physics, statistics

Keywords:
measure concentration, extreme points, ensemble equivalence, Fisher's discriminant, linear separability

Author for correspondence:
I. Y. Tyukin
e-mail: I.Tyukin@le.ac.uk

Blessing of dimensionality: mathematical foundations of the statistical physics of data

A. N. Gorban[1] and I. Y. Tyukin[1,2]

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

IYT, 0000-0002-7359-7966

The concentrations of measure phenomena were discovered as the mathematical background to statistical mechanics at the end of the nineteenth/beginning of the twentieth century and have been explored in mathematics ever since. At the beginning of the twenty-first century, it became clear that the proper utilization of these phenomena in machine learning might transform the *curse of dimensionality* into the *blessing of dimensionality*. This paper summarizes recently discovered phenomena of measure concentration which drastically simplify some machine learning problems in high dimension, and allow us to correct legacy artificial intelligence systems. The classical concentration of measure theorems state that i.i.d. random points are concentrated in a thin layer near a surface (a sphere or equators of a sphere, an average or median-level set of energy or another Lipschitz function, etc.). The new *stochastic separation theorems* describe the 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 linear functionals for separation of points can be selected in the form of the linear Fisher's discriminant. All artificial intelligence systems make errors. Non-destructive correction requires separation of the situations (samples) with errors from the samples corresponding to correct behaviour by a simple and robust classifier. The stochastic separation theorems provide us with such classifiers and determine a non-iterative (one-shot) procedure for their construction.

THE ROYAL SOCIETY PUBLISHING

This article is part of the theme issue 'Hilbert's sixth problem'.

1. Introduction: five 'foundations', from geometry to probability, quantum mechanics, statistical physics and machine learning

It's not given us to foretell
How our words will echo through the ages,. . .

F. I. Tyutchev, English Translation by F. Jude

The sixth Hilbert problem was inspired by the 'investigations on the foundations of geometry' [1], i.e. by Hilbert's work *The foundations of geometry* [2], which firmly implanted the axiomatic method not only in the field of geometry but also in other branches of mathematics. The sixth problem proclaimed expansion of the axiomatic method beyond existing mathematical disciplines, into physics and further on.

The sixth problem sounds very unusual and not purely mathematical. This may be a reason why some great works which have been inspired by this problem have no reference to it. The most famous example is the von Neumann book [3] *Mathematical foundations of quantum mechanics*. John von Neumann was an assistant to Hilbert, and they worked together on the mathematical foundation of quantum mechanics. This work was obviously in the framework of the sixth problem, but this framework was not mentioned in the book.

In 1933, Kolmogorov answered the Hilbert challenge of axiomatization of the theory of probability [4]. He did not cite the sixth problem but explicitly referred to Hilbert's *Foundations of geometry* as the prototype for 'the purely mathematical development' of the theory. But Hilbert in his sixth problem asked for more, for 'a rigorous and satisfactory development of the method of the mean values in mathematical physics'. He had in mind statistical physics and 'in particular the kinetic theory of gases'. The sixth chapter of Kolmogorov's book contains a survey of some results obtained by Kolmogorov and Khinchin about independence and the law of large numbers, and the appendix includes a description of the 0-1 laws in probability. These are the first steps to a rigorous basis of 'the method of mean values'. Ten years later, in 1943, Khinchin published a book entitled *Mathematical foundations of statistical mechanics* [5]. This brought an answer to the sixth problem one step closer, but again without explicit reference to Hilbert's talk. The analogy between the titles of von Neumann's and Khinchin's books is obvious.

The main idea of statistical mechanics, in essence, can be called the *blessing of dimensionality*: if a system can be presented as a union of many weakly interacting subsystems, then, in the thermodynamic limit (when the number of such subsystems tends to infinity), the whole system can be described by relatively simple deterministic relations in the low-dimensional space of macroscopic variables. *More means less*—in very high-dimensional spaces many differences between sets and functions become negligible (vanish) and the laws become simpler. This point of view on statistical mechanics was developed mainly by Gibbs [6] (ensemble equivalence), but Khinchin made the following remark about this work: 'although the arguments are clear from the logical standpoint, they do not pretend to any analytical rigor' [5, p. 4], exactly in the spirit of Hilbert's request for 'a rigorous and satisfactory development'. The devil is in the detail: how should we define the thermodynamic limit and in which sense are the ensembles equivalent? For some rigorously formulated conditions, the physical statements become exact theorems.

Khinchin considered two types of background theorems: ergodic theorems and limit theorems for high-dimensional distributions. He claimed that the foundations of statistical mechanics should be a complete abstraction from the nature of the forces. Limit theorems use very general

properties of distributions in high dimension, indeed, but the expectations that ergodicity is a typical and universal property of smooth high-dimensional multi-particle Hamiltonian systems were not met [7]. To stress that the ergodicity problem is non-trivial, we have to refer to the Oxtoby–Ulam theorem about metric transitivity of a generic *continuous* transformation, which preserves volume [8] (we see that typical properties of continuous transformations differ significantly from typical properties of smooth transformations).

Various programmes proposed for the mathematical foundation of statistical mechanics were discussed, for example, by Dobrushin [9] and Batterman [10]. Despite the impressive proof of ergodicity of some systems (hyperbolic flows or some billiard systems, for example), the Jaynes point of view [11] on the role of ergodicity in the foundations of statistical mechanics now became dominant; the ergodic hypothesis is neither a necessary nor a sufficient condition for the foundation of statistical mechanics (Dobrushin [9] attributed this opinion to Lebowitz, while Jaynes referred to Gibbs [6], who, perhaps, 'did not consider ergodicity as relevant to the foundation of the subject' [11, p. 87]).

Through the efforts of many mathematicians, the limit theorems from probability theory and results about ensemble equivalence from the foundation of statistical physics were developed far enough to become the general theory of measure concentration phenomena. Three articles were especially important for our work [12–14]. Ledoux [15] gives an introduction into the mathematical theory of measure concentration. A simple geometric introduction into these phenomena was given by Ball [16].

Perhaps the simplest manifestation of measure concentration is the concentration of the volume of the high-dimensional ball near the sphere. Let $V_n(r)$ be a volume of the n-dimensional ball of radius r. It is useful to stress that the 'ball' here is not necessarily Euclidean and means the ball of *any* norm. Lévy [17] recognized this phenomenon as a very important property of geometry of high-dimensional spaces. He also proved that equidistributions in the balls are asymptotically equivalent in high dimensions to the Gaussian distributions with the same mean value of squared radius. Gibbs de facto used these properties for sublevel sets of energy to demonstrate equivalence of ensembles (microcanonical distribution on the surface of constant energy and canonical distribution in the phase space with the same mean energy).

Maxwell used the concentration of measure phenomenon in the following settings. Consider a rotationally symmetric probability distribution on the n-dimensional unit sphere. Then its orthogonal projection on a line will be a Gaussian distribution with small variance $1/n$ (for large n with high accuracy). This is exactly the Maxwellian distribution for one degree of freedom in a gas (and the distribution on the unit sphere is the microcanonical distribution of the kinetic energy of gas, when the potential energy is negligibly small). Geometrically, it means that if we look at the one-dimensional projections of the unit sphere, then the 'observable diameter' will be small, of the order of $1/\sqrt{n}$.

Lévy noticed that instead of orthogonal projections on a straight line we can use any η-Lipschitz function f (with $\|f(x) - f(y)\| \leq \eta\|x - y\|$). Let points x be distributed on a unit n-dimensional sphere with rotationally symmetric probability distribution. Then the values of f will be distributed 'not more widely' than a normal distribution around the mean value $\mathbf{E}_f$; for all $\varepsilon > 0$

$$\mathbf{P}(|f - \mathbf{E}_f| \geq \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{2c\eta^2}\right),$$

where c is a constant, $c \leq 9\pi^3$. Interestingly, if we use in this inequality the *median value* of f, M_f, instead of the mean, then the estimate of the constant c can be decreased: $c \leq 1$. From the statistical mechanics point of view, this Lévy lemma describes the upper limit of fluctuations in gas for an arbitrary observable quantity f. The only condition is the sufficient regularity of f (Lipschitz property).

Hilbert's sixth problem influenced this stream of research either directly (Kolmogorov and, perhaps, Khinchin among others) or indirectly, through the directly affected works. And it

continues to spread this influence to other areas, including high-dimensional data analysis, data mining and machine learning.

At the turn of the millennium, Donoho [18] gave a lecture about the main problems of high-dimensional data analysis with the impressive subtitle: 'The curses and blessings of dimensionality'. He used the term *curse of dimensionality* 'to refer to the apparent intractability of systematically searching through a high-dimensional space, the apparent intractability of accurately approximating a general high-dimensional function, the apparent intractability of integrating a high-dimensional function' [18, p. 1]. To describe the blessing of dimensionality, he referred to the concentration of measure phenomenon, 'which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated' [18, p. 1]. Anderson *et al*. [19] characterized some manifestations of this phenomenon as 'The more, the merrier'.

In 1997, Kainen [20] described the phenomenon of blessing of dimensionality, illustrated it with numerous examples and suggested connections with geometric phenomena in high-dimensional spaces.

The claim of Donoho's talk was similar to Hilbert's talk and he cited this talk explicitly: 'My personal research experiences, cited above, convince me of Hilbert's position, as a long run proposition, operating on the scale of centuries rather than decades' [18, p. 28]. The role of Hilbert's sixth problem in the analysis of the curse and blessing of dimensionality was not mentioned again.

The blessing of dimensionality and the curse of dimensionality are two sides of the same coin. For example, the typical property of a random finite set in a high-dimensional space is: the squared distance of these points to a selected point are, with high probability, close to the average (or median) squared distance. This property drastically simplifies the expected geometry of data (blessing) [21,22], but, at the same time, makes the similarity search in high dimensions difficult and even useless (curse) [23].

Extension of the sixth Hilbert problem to data mining and machine learning is a challenging task. There exists no unified general definition of machine learning. Most classical texts consider machine learning through formalization and analysis of a set of standardized tasks [24–26]. Traditionally, these tasks are:

— classification—learning to predict a categorical attribute using values of given attributes on the basis of given examples (supervised learning);
— regression—learning to predict numerical attributes using values of given attributes on the basis of given examples (supervised learning);
— clustering—joining of similar objects in several clusters (unsupervised learning) [27];
— various data approximation and reduction problems: linear and nonlinear principal components [28], principal graphs [29], independent components [30], etc. (clustering can also be considered as a data approximation problem [31]);
— probability distribution estimation.

For example, Cucker & Smale [24] considered the least squares regression problem. This is the problem of the best approximation of an unknown function $f : X \to Y$ from a random sample of pairs $(x, y) \in X \times Y$. Selection of 'the best' regression function means minimization of the mean square error deviation of the observed y from the value $f(x)$. They use the concentration inequalities to evaluate the probability that the approximation has a given accuracy.

It is important to mention that the Cucker–Smale approach was inspired in particular by J. von Neumann: 'We try to write in the spirit of H. Weyl and J. von Neumann's contributions to the foundations of quantum mechanics' [24, p. 4]. The J. von Neumann book [3] was one step in the realization of Hilbert's sixth problem programme, as we know. Therefore, the Cucker–Smale 'mathematical foundation of learning' is a grandchild of the sixth problem. This is the fourth 'foundation' (after Kolmogorov, von Neumann and Khinchin). Indeed, it was an attempt to give a rigorous development of what they 'have found to be the central ideas of learning

theory' [24, p. 4]. This problem statement follows Hilbert's request for 'rigorous and satisfactory development of the method of mean values' [1, p. 454], but this time the development was done for machine learning instead of mathematical physics.

Cucker and Smale followed Gauss and proved that the least squares solution enjoys remarkable statistical properties, i.e. it provides the *minimum variance estimate* [24]. Nevertheless, non-quadratic functionals are employed as solutions to many problems: to enhance robustness, to avoid oversensitivity to outliers, to find sparse regression with exclusion of non-necessary input variables, etc. [25,26]. Even non-convex quasinorms and their tropical approximations are used efficiently to provide sparse and robust learning results [32]. Vapnik [26] defined a formalized fragment of machine learning using minimization of a *risk functional* that is the mathematical expectation of a general loss function.

Gromov [33] proposed a radically different concept of ergosystems which function by building their 'internal structure' out of the 'raw structures' in the incoming flows of signals. The essential mechanism of ergosystem learning is goal free and independent of any reinforcement. In a broad sense, loosely speaking, in this concept 'structure' = 'interesting structure' and learning of structure is goal free and should be considered as a structurally interesting process.

There are many other approaches and algorithms in machine learning that use some specific ideas from statistical mechanics: annealing, spin glasses, etc. (see, for example, [34]) and randomization. It was demonstrated recently that the assignment of random parameters should be data dependent to provide the efficient and universal approximation property of the randomized learner model [35]. Various methods for evaluation of the output weights of the hidden nodes after random generation of new nodes were also tested [35]. Swarm optimization methods for learning with random regeneration of the swarm (virtual particles) after several epochs of learning were developed in 1990 [36]. Sequential Monte Carlo methods for learning neural networks were elaborated and tested [37]. A comprehensive overview of the classical algorithms and modern achievements in stochastic approaches to neural networks was performed by Scardapane & Wang [38].

In our paper, we do not discuss these ideas; instead, we focus on a deep and general similarity between high-dimensional problems in learning and statistical physics. We summarize some phenomena of measure concentration which drastically affect machine learning problems in high dimension.

2. Waist concentration and random bases in machine learning

After the classical works of Fisher [39] and Rosenblatt [40], linear classifiers have been considered as the inception of data analytics and machine learning (e.g. [26,41,42], and references therein). The mathematical machinery powering these developments is based on the concept of linear separability.

Definition 2.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of $\mathbb{R}^n$. Recall that a linear functional l on $\mathbb{R}^n$ separates $\mathcal{X}$ and $\mathcal{Y}$ if there exists a $t \in \mathbb{R}$ such that

$$l(\boldsymbol{x}) > t > l(\boldsymbol{y}) \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{y} \in \mathcal{Y}.$$

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

If $\mathcal{X} \subset \mathbb{R}^n$ is a set of measurements or data samples that are labelled as 'class 1', and $\mathcal{Y}$ is a set of data labelled as 'class 2', then a functional l separating $\mathcal{X}$ and $\mathcal{Y}$ is the corresponding linear classifier. The fundamental question, however, is whether such functionals exist for the given $\mathcal{X}$ and $\mathcal{Y}$, and if the answer is 'yes' then how to find them?

It is well known that if (i) $\mathcal{X}$ and $\mathcal{Y}$ are disjoint, (ii) the cardinality, $|\mathcal{X} \cup \mathcal{Y}|$, of $\mathcal{X} \cup \mathcal{Y}$ does not exceed $n + 1$, and (iii) elements of $\mathcal{X} \cup \mathcal{Y}$ are in a general position, then they are vertices of a simplex. Hence, in this setting, there is always a linear functional l separating $\mathcal{X}$ and $\mathcal{Y}$.

Rosenblatt's α-perceptron [40] used a population of linear threshold elements with random synaptic weights (A-elements) as the layer before an R-element, that is, a linear threshold element which learns iteratively (authors of some papers and books called the R-elements 'perceptrons' and lose the complex structure of α-perceptron with a layer of random A-elements). The randomly initiated elements of the first layer can undergo selection of the most relevant elements.

According to Rosenblatt [40], any set of data vectors becomes linear separable after transformation by the layer of A-elements, if the number of these randomly chosen elements is sufficiently large. Therefore, the perceptron can solve any classification problem, where classes are defined by pointing out examples (ostensive definition). But this 'sufficiently large' number of random elements depends on the problem and may be large, indeed. It can grow for a classification task proportionally to the number of the examples. The perceptron with a sufficiently large number of A-elements can approximate binary-valued functions on finite domains with arbitrary accuracy. Recently, the bounds on errors of these approximations are derived [43]. It is proved that, unless the number of network units grows faster than any polynomial of the logarithm of the size of the domain, a good approximation cannot be achieved for almost any uniformly randomly chosen function. The results are obtained by the application of concentration inequalities.

The method of random projections became popular in machine learning after the Johnson–Lindenstrauss lemma [44], which states that relatively large sets of m vectors in a high-dimensional Euclidean space $\mathbb{R}^d$ can be linearly mapped into a space of much lower dimension n with approximate preservation of distances. This mapping can be constructed (with high probability) as a projection on n random basis vectors with rescaling of the projection with a factor $\sqrt{d}$ [45]. Repeating the projection $O(m)$ times and selecting the best of them, one can achieve the appropriate accuracy of the distance preservation. The number of points m can be exponentially large with n ($m \leq \exp(cn)$).

Two unit random vectors in high dimension are almost orthogonal with high probability. This is a simple manifestation of the so-called *waist concentration* [13]. A high-dimensional sphere is concentrated near its equator. This is obvious: just project a sphere onto a hyperplane and use the concentration argument for a ball on the hyperplane (with a simple trigonometric factor). This seems highly non-trivial, if we ask: near which equator? The answer is: near each equator. This answer is obvious because of rotational symmetry, but it seems to be counterintuitive.

We call vectors $\boldsymbol{x}$, $\boldsymbol{y}$ from Euclidean space $\mathbb{R}^n$ *ε-orthogonal* if $|(\boldsymbol{x}, \boldsymbol{y})| < \varepsilon$ ($\varepsilon > 0$). Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be i.i.d. random vectors distributed uniformly (rotationally invariant) on the unit sphere in Euclidean space $\mathbb{R}^n$. Then the distribution of their inner product satisfies the inequality (see, for example, [16] or [46] and compare with the Maxwellian and Lévy's lemma):

$$\mathbf{P}(|(\boldsymbol{x}, \boldsymbol{y})| < \varepsilon) \geq 1 - 2\exp(-\tfrac{1}{2}n\varepsilon^2).$$

Proposition 2.2. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be i.i.d. random vectors distributed uniformly (rotationally invariant) on the unit sphere in Euclidean space $\mathbb{R}^n$. For*

$$\boxed{N < e^{\varepsilon^2 n/4}\left[\ln\left(\frac{1}{1-\vartheta}\right)\right]^{1/2}} \tag{2.1}$$

all vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are pairwise ε-orthogonal with probability $P > 1 - \vartheta$ [46].

There are two consequences of this statement: (i) in high dimension there exist exponentially many pairwise almost orthogonal vectors in $\mathbb{R}^n$ and (ii) N random vectors are ε-orthogonal with high probability $P > 1 - \vartheta$ even for exponentially large N (2.1). Existence of exponentially large ε-orthogonal systems in high-dimensional spaces was discovered in 1993 by Kainen & Kůrková [47]. They introduced the notion of *quasiorthogonal dimension*, which was immediately used in the problem of random indexing of high-dimensional data [21]. The fact that an exponentially large random set consists of pairwise ε-orthogonal vectors with high probability was demonstrated in [46] and used for analysis of the data approximation problem in random

bases. We show that not only do such ε-orthogonal sets exist, but also that they are typical in some sense.

N randomly generated vectors $\boldsymbol{x}_i$ will be almost orthogonal to a given data vector $\boldsymbol{y}$ (the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ will be close to $\pi/2$ with probability close to 1). Therefore, the coefficients in the approximation of $\boldsymbol{y}$ by a linear combination of $\boldsymbol{x}_i$ could be arbitrarily large and the approximation problem will be ill-conditioned, with high probability. The following alternative is proved for approximation by random bases.

— Approximation of a high-dimensional data vector by linear combinations of randomly and independently chosen vectors requires (with high probability) generation of exponentially large 'bases', if we would like to use bounded coefficients in linear combinations.
— If arbitrarily large coefficients are allowed, then the number of randomly generated elements that are sufficient for approximation is even less than dimension. We have to pay for such a reduction in the number of elements by ill-conditioning of the approximation problem.

We have to choose between a well-conditioned approximation problem in exponentially large random bases and an ill-conditional problem in relatively small (moderate) random bases. This dichotomy is fundamental, and it is a direct consequence of the waist concentration phenomenon. In what follows, we will formally present another concentration phenomenon, stochastic separation theorems [48,49], and outline their immediate applications in artificial intelligence (AI) and neuroscience.

3. Stochastic separation theorems and their applications in artificial intelligence systems

(a) Stochastic separation theorems

Existence of a linear functional that separates two finite sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ is no longer obvious when $|\mathcal{X} \cup \mathcal{Y}| \gg n$. A possible way to answer both questions could be to cast the problem as a constrained optimization problem within the framework of, for example, support vector machines [26]. The issue with this approach is that theoretical worst-case estimates of computational complexity for determining such functions are of the order $O(|\mathcal{X} \cup \mathcal{Y}|^3)$ (for quadratic loss functions); *a posteriori* analysis of experiments on practical use cases, however, suggests that the complexity could be much smaller than $O(|\mathcal{X} \cup \mathcal{Y}|^3)$ and becomes linear or even sublinear in $|\mathcal{X} \cup \mathcal{Y}|$ [50].

This apparent discrepancy between the worst-case estimates and *a posteriori* evaluation of computational complexities can be resolved if concentration effects are taken into account. If the dimension n of the underlying topological vector space is large, then random finite but exponentially large in n samples are linearly separable, with high probability, for a range of practically relevant classes of distributions. Moreover, we show that the corresponding separating functionals can be derived using Fisher linear discriminants [39]. Computational complexity of the latter is linear in $|\mathcal{X} \cup \mathcal{Y}|$. It can be made sublinear too if proper sampling is used to estimate corresponding covariance matrices. As we have shown in [49], the results hold for i.i.d. random points from equidistributions in a ball, a cube, and from distributions that are products of measures with bounded support. The conclusions are based on stochastic separation theorems for which the statements for relevant classes of distributions are provided below.

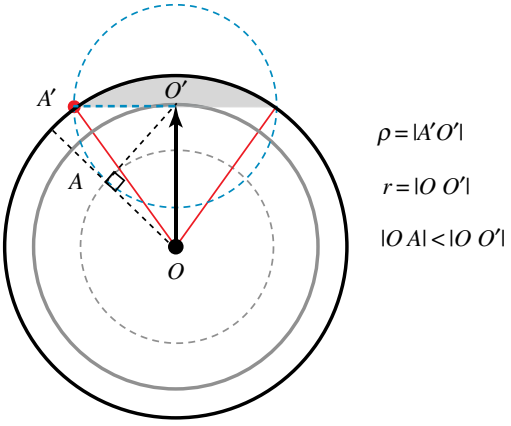

Figure 1. Illustration to theorem 3.1. (Online version in colour.)

Theorem 3.1 (equidistribution in $\mathbb{B}_n(1)$ [48,49]). *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 $\mathbb{B}_n(1)$. Let $0 < r < 1$, and $\rho = \sqrt{1 - r^2}$. Then*

$$\mathbf{P}\left(\|\boldsymbol{x}_M\| > r \text{ and } \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_M}{\|\boldsymbol{x}_M\|}\right) < r \text{ for all } i \neq M\right) \geq 1 - r^n - 0.5(M-1)\rho^n, \tag{3.1}$$

$$\mathbf{P}\left(\|\boldsymbol{x}_j\| > r \text{ and } \left(\boldsymbol{x}_i, \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|}\right) < r \text{ for all } i, j,\ i \neq j\right) \geq 1 - Mr^n - 0.5M(M-1)\rho^n \tag{3.2}$$

and

$$\mathbf{P}\left(\|\boldsymbol{x}_j\| > r \text{ and } \left(\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|}\right) < r \text{ for all } i, j,\ i \neq j\right) \geq 1 - Mr^n - M(M-1)\rho^n. \tag{3.3}$$

The proof of the theorem can be illustrated with figure 1. The probability that a single element, $\boldsymbol{x}_M$, belongs to the difference $\mathbb{B}_n(1)\backslash\mathbb{B}_n(r)$ of two n-balls centred at O is not smaller than $1 - r^n$. Consider the hyperplane

$$l(\boldsymbol{x}) = r, \text{ where } l(\boldsymbol{x}) = \left(\boldsymbol{x}, \frac{\boldsymbol{x}_M}{\|\boldsymbol{x}_M\|}\right).$$

This hyperplane partitions the unit ball $\mathbb{B}_n(1)$ centred at O into two disjoint subsets: the spherical cap (shown as a grey shaded area in figure 1) and the rest of the ball. The element $\boldsymbol{x}_M$ is in the shaded area and is on the line containing the vector OO'. The volume of this spherical cap does not exceed the volume of the half-ball of radius ρ centred at O' (the ball $\mathbb{B}_n(\rho)$ is shown as a dashed circle centred at O' in the figure). Recall that

$$\mathbf{P}(A_1 \& A_2 \& \ldots \& A_m) \geq 1 - \sum_i (1 - \mathbf{P}(A_i)) \text{ for any events } A_1, \ldots, A_m. \tag{3.4}$$

This ensures that (3.1) holds. Applying the same argument to all elements of the set $\mathcal{S}$ results in (3.2). Finally, to show that (3.3) holds, observe that the length of the segment OA on the tangent line to the sphere $\mathbb{S}_{n-1}(\rho)$ centred at O' is always smaller than $r = |OO'|$. Hence the cosine of the angle between an element from $(\mathbb{B}_n(1)\backslash\mathbb{B}_n(r))\backslash\mathbb{B}_n(\rho)$ and the vector OO' is bounded from above by $\cos(\angle(OA', OO')) = r$. The estimate now follows from (3.4).

According to theorem 3.1, the probability that a single element $\boldsymbol{x}_M$ from the sample $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is linearly separated from the set $\mathcal{S}\backslash\{\boldsymbol{x}_M\}$ by the hyperplane $l(x) = r$ is at least

$$1 - r^n - 0.5(M-1)(1 - r^2)^{n/2}.$$

This probability estimate depends on both $M = |\mathcal{S}|$ and dimensionality n. An interesting consequence of the theorem is that if one picks a probability value, say $1 - \vartheta$, then the maximal

possible values of M for which the set $\mathcal{S}$ remains linearly separable with probability that is no less than $1-\vartheta$ grows at least exponentially with n. In particular, the following holds.

Corollary 3.2. *Let $\{x_1,\dots,x_M\}$ be a set of M i.i.d. random points from the equidistribution in the unit ball $\mathbb{B}_n(1)$. Let $0<r,\vartheta<1$, and $\rho=\sqrt{1-r^2}$. If*

$$M<\frac{2(\vartheta-r^n)}{\rho^n}, \tag{3.5}$$

then $\mathbf{P}((x_i,x_M)<r\|x_M\| \text{ for all } i=1,\dots,M-1)>1-\vartheta$. If

$$M<\left(\frac{r}{\rho}\right)^n\left(-1+\sqrt{1+\frac{2\vartheta\rho^n}{r^{2n}}}\right), \tag{3.6}$$

then $\mathbf{P}((x_i,x_j)<r\|x_i\| \text{ for all } i,j=1,\dots,M,\ i\neq j)\geq 1-\vartheta$.

In particular, if inequality (3.6) holds, then the set $\{x_1,\dots,x_M\}$ is linearly separable with probability $p>1-\vartheta$.

The linear separability property of finite but exponentially large samples of random i.i.d. elements is not restricted to equidistributions in $\mathbb{B}_n(1)$. As has been noted in [22], it holds for equidistributions in ellipsoids as well as for the Gaussian distributions. Moreover, it can be generalized to product distributions in a unit cube. Consider, for example, the case when coordinates of the vectors $x=(X_1,\dots,X_n)$ in the set $\mathcal{S}$ are independent random variables X_i, $i=1,\dots,n$ with expectations $\bar{X}_i$ and variances $\sigma_i^2>\sigma_0^2>0$. Let $0\leq X_i\leq 1$ for all $i=1,\dots,n$. The following analogue of theorem 3.1 can now be stated.

Theorem 3.3 (product distribution in a cube [49]). *Let $\{x_1,\dots,x_M\}$ be i.i.d. random points from the product distribution in a unit cube. Let*

$$R_0^2=\sum_i\sigma_i^2\geq n\sigma_0^2$$

and $0<\delta<\frac{2}{3}$. Then

$$\mathbf{P}\left(1-\delta\leq\frac{\|x_j-\bar{x}\|^2}{R_0^2}\leq 1+\delta \text{ and } \left(\frac{x_i-\bar{x}}{R_0},\frac{x_M-\bar{x}}{\|x_M-\bar{x}\|}\right)<\sqrt{1-\delta} \text{ for all } i,j,\ i\neq M\right)$$
$$\geq 1-2M\exp\left(-\frac{2\delta^2R_0^4}{n}\right)-(M-1)\exp\left(-\frac{2R_0^4(2-3\delta)^2}{n}\right) \tag{3.7}$$

and

$$\mathbf{P}\left(1-\delta\leq\frac{\|x_j-\bar{x}\|^2}{R_0^2}\leq 1+\delta \text{ and } \left(\frac{x_i-\bar{x}}{R_0},\frac{x_j-\bar{x}}{\|x_j-\bar{x}\|}\right)<\sqrt{1-\delta} \text{ for all } i,j,\ i\neq j\right)$$
$$\geq 1-2M\exp\left(-\frac{2\delta^2R_0^4}{n}\right)-M(M-1)\exp\left(-\frac{2R_0^4(2-3\delta)^2}{n}\right). \tag{3.8}$$

The proof is based on concentration inequalities in product spaces [14,51]. Numerous generalizations of theorems 3.1 and 3.3 are possible for different classes of distributions, for example for weakly dependent variables, etc.

Linear separability, as an inherent property of datasets in high dimension, is not necessarily confined to cases whereby a linear functional separates a single element of a set from the rest. Theorems 3.1 and 3.3 can be generalized to the case of separating m-tuples, $m>1$ too. An example of such generalization is provided in the next theorem.

Theorem 3.4 (separation of m-tuples [52]). *Let $\mathcal{X}=\{x_1,\dots,x_M\}$ and $\mathcal{Y}=\{x_{M+1},\dots,x_{M+k}\}$ be i.i.d. samples from the equidistribution in $\mathbb{B}_n(1)$. Let $\mathcal{Y}_c=\{x_{M+r_1},\dots,x_{M+r_m}\}$ be a subset of m elements*

from $\mathcal{Y}$ *such that*

$$\beta_2(m-1) \le \sum_{r_j,\, r_j \neq r_i} (\boldsymbol{x}_{M+r_i}, \boldsymbol{x}_{M+r_j}) \le \beta_1(m-1) \quad \textit{for all } i = 1, \ldots, m. \tag{3.9}$$

Then

$$\mathbf{P}(\exists \textit{ a linear functional separating } \mathcal{X} \textit{ and } \mathcal{Y}_c) \ge \max_{\varepsilon \in (0,1)} (1-(1-\varepsilon)^n)^m \left(1 - \frac{\Delta(\varepsilon, m)^{\frac{n}{2}}}{2}\right)^M, \tag{3.10}$$

where

$$\Delta(\varepsilon, m) = 1 - \frac{1}{m}\left(\frac{(1-\varepsilon)^2 + \beta_2(m-1)}{\sqrt{1+(m-1)\beta_1}}\right)^2,$$

subject to:

$$(1-\varepsilon)^2 + \beta_2(m-1) > 0, \quad 1 + (m-1)\beta_1 > 0.$$

The separating linear functional is again the inner product, and the separating hyperplane can be taken in the form [52]

$$l(\boldsymbol{x}) = r, \quad \text{where } l(\boldsymbol{x}) = \left(\boldsymbol{x}, \frac{\bar{\boldsymbol{y}}}{\|\bar{\boldsymbol{y}}\|}\right), \quad r = \frac{1}{\sqrt{m}}\left(\frac{(1-\varepsilon)^2 + \beta_2(m-1)}{\sqrt{1+(m-1)\beta_1}}\right) \tag{3.11}$$

and ε is the maximizer of the nonlinear program on the right-hand side of (3.10), and $\bar{\boldsymbol{y}} = (1/m)\sum_{i=1}^m \boldsymbol{x}_{M+r_i}$. To see this, observe that $\|\boldsymbol{x}_{M+r_i}\| \ge 1-\varepsilon$, $\varepsilon \in (0,1)$, for all $i = 1, \ldots, m$, with probability $(1-(1-\varepsilon)^n)^m$. With this probability the following estimate holds:

$$\left(\frac{\bar{\boldsymbol{y}}}{\|\bar{\boldsymbol{y}}\|}, \boldsymbol{x}_{M+r_i}\right) \ge \frac{1}{m\|\bar{\boldsymbol{y}}\|}((1-\varepsilon)^2 + \beta_2(m-1)).$$

Hence

$$\frac{1}{m}(1+(m-1)\beta_1) \ge (\bar{\boldsymbol{y}}, \bar{\boldsymbol{y}}) \ge \frac{1}{m}((1-\varepsilon)^2 + \beta_2(m-1))$$

and $l(\boldsymbol{x})$ in (3.11) is the required functional (figure 1).

If the elements of $\mathcal{Y}_c$ are uncorrelated, i.e. the values of $\beta_1(m-1), \beta_2(m-1)$ are small, then the distance from the spherical cap induced by linear functional (3.11) to the centre of the ball decreases as $O(1/\sqrt{m})$. This means that the lower-bound probability estimate in (3.10) is expected to decrease too. On the other hand, if the elements of $\mathcal{Y}_c$ are all positively correlated, i.e. $1 \ge \beta_1 > \beta_2 > 0$, then one can derive a lower-bound probability estimate which does not depend on m.

The peculiar properties of data in high dimension, expressed in terms of linear separability, have several consequences and applications in the realm of AI and machine learning; examples are provided in the next sections.

(b) Correction of legacy artificial intelligence systems

Legacy AI systems, i.e. AI systems that have been deployed and are currently in operation, are becoming more and more widespread. Well-known commercial examples are provided by global multi-nationals, including Google, IBM, Amazon, Microsoft and Apple. Numerous open-source legacy AIs have been created to date, together with dedicated software for their creation (e.g. Caffe [53], MXNet [54], Deeplearning4j [55] and Tensorflow [56] packages). These AI systems require significant computational and human resources to build. Regardless of resources spent, virtually any AI and/or machine learning-based systems are likely to make a mistake. Real-time correction of these mistakes by retraining is not always viable due to the resources involved. AI retraining is not necessarily desirable either, because AI's performance after retraining may not always be guaranteed to exceed that of the previous performance. We can, therefore, formulate the technical requirements for the correction procedures. The corrector should: (i) be simple; (ii) not change the skills of the legacy system; (iii) allow fast non-iterative learning; and (iv) allow correction of new mistakes without destroying previous corrections.

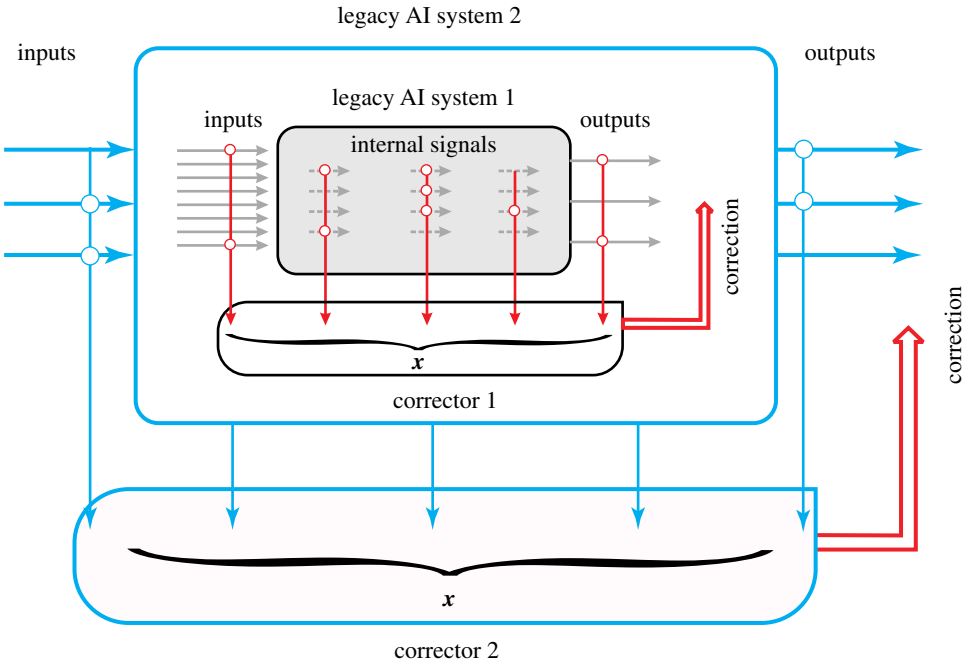

Figure 2. Cascade of AI correctors. (Online version in colour.)

A possible remedy to this issue is the AI correction method [22] based on stochastic separation theorems. Suppose that at a time instance t values of signals from inputs, outputs and the internal state of a legacy AI system could be combined together to form a single measurement object, $\boldsymbol{x} = (x_1, \ldots, x_n)$. All n entries in this object are numerical values, and each measurement $\boldsymbol{x}$ corresponds to a relevant decision of the AI system at time t. Over the course of the system's existence a set $\mathcal{S}$ of such measurements is collected. For each element in the set $\mathcal{S}$ a label 'correct' or 'incorrect' is assigned, depending on external evaluation of the system's performance. Elements corresponding to 'incorrect' labels are then filtered out and dealt with separately by an additional subsystem, a corrector. The process is illustrated in figure 2. In this diagram, the original legacy AI system (shown as legacy AI system 1) is supplied with a corrector altering its responses. The combined new AI system can in turn be augmented by another corrector, leading to a cascade of AI correctors (figure 2).

If distributions modelling elements of the set $\mathcal{S}$ are, for example, an equidistribution in a ball or an ellipsoid, a product of measures distribution, a Gaussian, etc. then

— Theorems 3.1–3.4 guarantee that construction of such AI correctors can be achieved using mere linear functionals.
— These linear functionals admit closed-form formulae (Fisher linear discriminant) and can be determined in a non-iterative way.
— The availability of explicit closed-form formulae in the form of a Fisher discriminant offers major computational benefits as it eliminates the need to employ iterative and more computationally expensive alternatives, such as support vector machines.
— If a cascade of correctors is employed, the performance of the corrected system drastically improves [22].

The results, perhaps, can be generalized to other classes of distributions that are regular enough to enjoy the stochastic separability property.

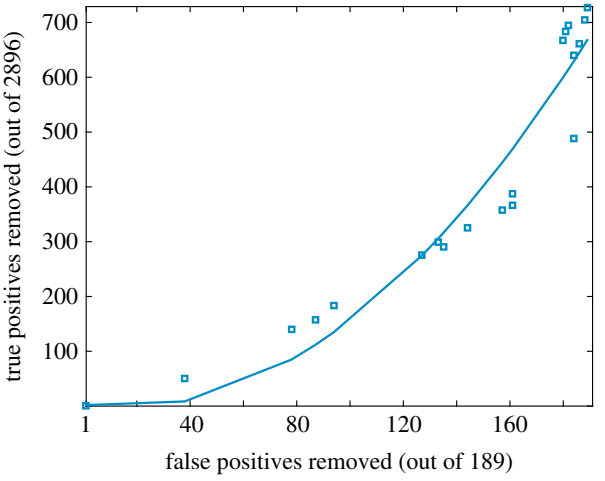

Figure 3. True positives removed as a function of false positives removed by a single-functional corrector [22]. (Online version in colour.)

The corrector principle has been demonstrated in [22] for a legacy AI system in the form of a convolutional neural network trained to detect pedestrians in images. AI errors were set to be false positives, and the corrector system had to remove labelled false positives by a single linear functional. A detailed description of the experiment is provided in [22], and a performance snapshot is shown in figure 3. Dimensionality n of the vectors $\boldsymbol{x}$ was 2000. As we can see from figure 3, single linear functionals are capable of removing several errors of a legacy AI without compromising the system's performance. Note that AI errors, i.e. false positives, were chosen at random and have not been grouped or clustered to take advantage of positive correlation. (The definition of clusters could vary [27].) As the number of errors to be removed grows, performance starts to deteriorate. This is in agreement with our theoretical predictions (theorem 3.4).

(c) Knowledge transfer between artificial intelligence systems

Legacy AI correctors can be generalized to a computational framework for automated AI knowledge transfer whereby labelling of the set $\mathcal{S}$ is provided by an external AI system. AI knowledge transfer has been the focus of growing attention during the last decade [57]. Application of stochastic separation theorems to AI knowledge transfer was proposed in [52], and the corresponding functional diagram of this automated set-up is shown in figure 4. In this set-up a student AI, denoted as AI_s, is monitored by a teacher AI, denoted as AI_t. Over a period of activity system AI_s generates a set $\mathcal{S}$ of objects $\boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^n$. The exact composition of the set $\mathcal{S}$ depends on the task at hand. If the outputs from AI_s differ from those of AI_t for the same input, then an error is registered in the system. Objects $\boldsymbol{x} \in \mathcal{S}$ associated with errors are combined into the set $\mathcal{Y}$. The process gives rise to two disjoint sets:

$$\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \quad \mathcal{X} = \mathcal{S} \backslash \mathcal{Y} \quad \text{and} \quad \mathcal{Y} = \{\boldsymbol{x}_{M+1}, \ldots, \boldsymbol{x}_{M+k}\}.$$

Having created these two sets, knowledge transfer from AI_t to AI_s can now be organized in accordance with algorithm 1. Note that data regularization and whitening are included in the preprocessing step of algorithm 1. The algorithm can be used for AI correctors too. Similar to AI correction, AI knowledge transfer also be cascaded. Specific examples and illustrations of AI knowledge transfer based on stochastic separation theorems are discussed in [52].

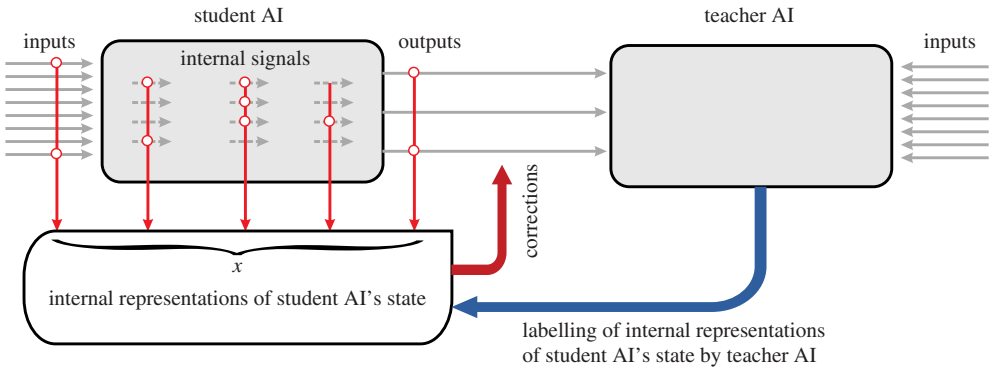

Figure 4. AI knowledge transfer. (Online version in colour.)

(d) Grandmother cells, memory and high-dimensional brain

Stochastic separation theorems are a generic phenomenon, and their applications are not limited to AI and machine learning systems. An interesting consequence of these theorems for neuroscience has been discovered and presented in [58]. Recently, it has been shown that in humans new memories can be learnt very rapidly by supposedly individual neurons from a limited number of experiences [59]. Moreover, neurons can exhibit remarkable selectivity to complex stimuli, the evidence that has led to debates around the existence of the so-called 'grandmother' and 'concept' cells [60–62], and their role as elements of a declarative memory. These findings suggest that not only can the brain learn rapidly but also it can respond selectively to 'rare' individual stimuli. Moreover, experimental evidence indicates that such a cognitive functionality can be delivered by single neurons [59–61]. The fundamental questions, hence, are: How is this possible? and What could be the underlying functional mechanisms?

It has been shown in [58] that stochastic separation theorems offer a simple answer to these fundamental questions. In particular, extreme neuronal selectivity and rapid learning can already be explained by these theorems. Model-wise, the explanation of extreme selectivity is based on a conventional and widely accepted phenomenological generic description of neural response to stimulation. Rapid acquisition of selective responses to multiple stimuli by single neurons is ensured by classical Hebbian synaptic plasticity [63].

4. Conclusion

Twenty-three Hilbert's problems created important 'focus points' for the concentration of efforts of mathematicians for a century. The sixth problem differs significantly from the other twenty-two problems. It is very far from being a purely mathematical problem. It seems to be impossible to imagine its 'final solution'. The sixth problem is a 'programmatic call' [64], and it works:

— We definitely know that the sixth problem had great influence on the formulation of the mathematical foundation of quantum mechanics [3] and on the development of axiomatic quantum field theory [65].
— We have no doubt (but the authors have no direct evidence) that the sixth problem has significantly affected research in the foundation of probability theory [4] and statistical mechanics [5].
— The modern theory of measure concentration phenomena has direct relations to the mathematical foundations of probability and statistical mechanics, uses results of Kolmogorov and Khinchin (among others), and definitely helps to create 'a rigorous and satisfactory development of the method of the mean values. . .' [1, p. 454].
— Some of the recent attempts at a rigorous approach to machine learning [24] used parts of the sixth problem programme [3] as a prototype for their conceptual approach.

Algorithm 1 AI knowledge transfer/correction [52].

(i) **Preprocessing**

(a) *Centring*. For the given set $\mathcal{S}$, determine the set average, $\bar{\boldsymbol{x}}(\mathcal{S})$, and generate sets $\mathcal{S}_c$

$$\mathcal{S}_c = \{\boldsymbol{x} \in \mathbb{R}^n \,|\boldsymbol{x} = \boldsymbol{\xi} - \bar{\boldsymbol{x}}(\mathcal{S}),\ \boldsymbol{\xi} \in \mathcal{S}\},$$
$$\mathcal{Y}_c = \{\boldsymbol{x} \in \mathbb{R}^n \,|\boldsymbol{x} = \boldsymbol{\xi} - \bar{\boldsymbol{x}}(\mathcal{S}),\ \boldsymbol{\xi} \in \mathcal{Y}\}.$$

(b) *Regularization*. Determine covariance matrices $\mathrm{Cov}(\mathcal{S}_c)$, $\mathrm{Cov}(\mathcal{S}_c \backslash \mathcal{Y}_c)$ of the sets $\mathcal{S}_c$ and $\mathcal{S}_c \backslash \mathcal{Y}_c$. Let $\lambda_i(\mathrm{Cov}(\mathcal{S}_c))$, $\lambda_i(\mathrm{Cov}(\mathcal{S}_c \backslash \mathcal{Y}_c))$ be their corresponding eigenvalues, and $h_1, \ldots, h_n$ be the eigenvectors of $\mathrm{Cov}(\mathcal{S}_c)$. If some of $\lambda_i(\mathrm{Cov}(\mathcal{S}_c))$, $\lambda_i(\mathrm{Cov}(\mathcal{S}_c \backslash \mathcal{Y}_c))$ are zero or if the ratio $\frac{\max_i\{\lambda_i(\Sigma(\mathcal{S}_c))\}}{\min_i\{\lambda_i(\Sigma(\mathcal{S}_c))\}}$ is too large, project $\mathcal{S}_c$ and $\mathcal{Y}_c$ onto an appropriately chosen set of $m < n$ eigenvectors, $h_{n-m+1}, \ldots, h_n$:

$$\mathcal{S}_r = \{\boldsymbol{x} \in \mathbb{R}^n \,|\boldsymbol{x} = H^T\boldsymbol{\xi},\ \boldsymbol{\xi} \in \mathcal{S}_c\},$$
$$\mathcal{Y}_r = \{\boldsymbol{x} \in \mathbb{R}^n \,|\boldsymbol{x} = H^T\boldsymbol{\xi},\ \boldsymbol{\xi} \in \mathcal{Y}_c\},$$

where $H = \begin{pmatrix} h_{n-m+1} \cdots h_n \end{pmatrix}$ is the matrix consisting of m significant principal components of $\mathcal{S}_c$.

(c) *Whitening*. For the centred and regularized dataset $\mathcal{S}_r$, derive its covariance matrix, $\mathrm{Cov}(\mathcal{S}_r)$, and generate whitened sets

$$\mathcal{S}_w = \{\boldsymbol{x} \in \mathbb{R}^m \,|\boldsymbol{x} = \mathrm{Cov}(\mathcal{S}_r)^{-\frac{1}{2}}\boldsymbol{\xi},\ \boldsymbol{\xi} \in \mathcal{S}_r\},$$
$$\mathcal{Y}_w = \{\boldsymbol{x} \in \mathbb{R}^m \,|\boldsymbol{x} = \mathrm{Cov}(\mathcal{S}_r)^{-\frac{1}{2}}\boldsymbol{\xi},\ \boldsymbol{\xi} \in \mathcal{Y}_r\}.$$

(ii) **Knowledge transfer**

(a) *Clustering*. Pick $p \geq 1$, $p \leq k$, $p \in \mathbb{N}$, and partition the set $\mathcal{Y}_w$ into p clusters $\mathcal{Y}_{w,1}, \ldots \mathcal{Y}_{w,p}$ so that elements of these clusters are, on average, pairwise positively correlated. That is, there are $\beta_1 \geq \beta_2 > 0$ such that:

$$\beta_2(|\mathcal{Y}_{w,i}| - 1) \leq \sum_{\xi \in \mathcal{Y}_{w,i} \backslash \{\boldsymbol{x}\}} (\boldsymbol{\xi}, \boldsymbol{x}) \leq \beta_1(|\mathcal{Y}_{w,i}| - 1) \text{ for any } \boldsymbol{x} \in \mathcal{Y}_{w,i}.$$

(b) *Construction of auxiliary knowledge units*. For each cluster $\mathcal{Y}_{w,i}$, $i = 1, \ldots, p$, construct separating linear functionals ℓ_i and thresholds c_i:

$$\ell_i(\boldsymbol{x}) = \left(\frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}, \boldsymbol{x}\right),$$
$$\boldsymbol{w}_i = \left(\mathrm{Cov}(\mathcal{S}_w \backslash \mathcal{Y}_{w,i}) + \mathrm{Cov}(\mathcal{Y}_{w,i})\right)^{-1} \left(\bar{\boldsymbol{x}}(\mathcal{Y}_{w,i}) - \bar{\boldsymbol{x}}(\mathcal{S}_w \backslash \mathcal{Y}_{w,i})\right),$$
$$c_i = \min_{\boldsymbol{\xi} \in \mathcal{Y}_{w,i}} \left(\frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}, \boldsymbol{\xi}\right),$$

where $\bar{\boldsymbol{x}}(\mathcal{Y}_{w,i})$, $\bar{\boldsymbol{x}}(\mathcal{S}_w \backslash \mathcal{Y}_{w,i})$ are the averages of $\mathcal{Y}_{w,i}$ and $\mathcal{S}_w \backslash \mathcal{Y}_{w,i}$, respectively. The separating hyperplane is $\ell_i(\boldsymbol{x}) = c_i$.

(c) *Integration*. Integrate auxiliary knowledge units into decision-making pathways of AIs. If, for an $\boldsymbol{x}$ generated by an input to AIs, any of $\ell_i(\boldsymbol{x}) \geq c_i$ then report $\boldsymbol{x}$ accordingly (swap labels, report as an error, etc.)

— The modern idea of *blessing of dimensionality* in high-dimensional data analysis [18,19, 48] is, in essence, an extension and further development of ideas from the mathematical foundations of statistical mechanics.

The classical measure concentration theorems state that random points in a highly dimensional data distribution are concentrated in a thin layer near an average or median-level set of a Lipschitz function. The stochastic separation theorems describe the fine structure of these thin layers: the random points are all linearly separable from the rest of the set even for exponentially large random sets. Of course, for all these concentration and separation theorems the probability distribution should be 'genuinely' high dimensional. Equidistributions in balls or ellipsoids or the products of distributions with compact support and non-vanishing variance are simple examples of such distributions. Various generalizations are possible.

For which dimensions does the blessing of dimensionality work? This is a crucial question. The naive point of view that the dimension of data is just a number of coordinates is wrong. This is the dimension of the data space, where data are originally situated. The notion of *intrinsic* dimension of data is needed [66,67]. The situation when the number of data points N is less (or even much less) than the dimension d of the data space is not new. Moreover, Donoho [18] considered the property $d > N$ as a generic case in the 'post-classical world' of data analysis. In such a situation, we really explore data on a $d-1$-dimensional plane and should modestly reduce our high-dimensional claim. The projection of data on that plane can be performed by various methods. We can use as new coordinates projections of points on the known data points or Pearson's correlation coefficients, when they are suitable; for example, when the data points are fragments of time series or large spectral images, etc. In these new coordinates the data table becomes a square matrix and further dimensionality reduction could be performed using principal component analysis (PCA), or its nonlinear versions like principal manifolds [28] or neural auto-encoders [68].

A standard example can be found in [69]: the initial data space consisted of fluorescence diagrams and had dimension 5.2×10^5. There were 62 data points, and a combination of correlation coordinates with PCA showed intrinsic dimension 4 or 5. For the selection of relevant principal components the Kaiser rule, the broken stick models or other heuristical or statistical methods can be used [70].

A similar preprocessing ritual is helpful even in more 'classical' cases when $d < N$. The correlation (or projection) transformation is not essential here, but the formation of relevant features with dimension reduction is important. If after model reduction and *whitening* (transformation of coordinates to get the unit covariance matrix, step i.c in algorithm 1) the new dimension $D \gtrsim 100$, then for $\lesssim 10^6$ data points we can expect that the stochastic separation theorems work with probability greater than 99%. Thus the separation of errors with Fisher's linear discriminant is possible, and many other 'blessing of dimensionality benefits' are achievable. Of course, some additional hypotheses about the distribution functions are needed for a rigorous proof, but there is practically no chance to check them *a priori* and the validation of the whole system *a posteriori* is necessary. In smaller dimensions (for example, less than 10), nonlinear data approximation methods can work well, capturing the intrinsic complexity of data, like principal graphs do [29,71].

We have an alternative: either essentially high-dimensional data with thin shell concentrations, stochastic separation theorems and efficient linear methods, or essentially low-dimensional data with efficient complex nonlinear methods. There is a problem of the 'no man's land' in between. To explore this land, we can extract the most interesting low-dimensional structure and then consider the residual as an essentially high-dimensional random set, which obeys stochastic separation theorems. We do not currently know a theoretically justified efficient approach to this area, but here we should say, following Hilbert: 'Wir müssen wissen, wir werden wissen' (We must know, we shall know).

Data accessibility. This article has no additional data.

Authors' contributions. Both authors made substantial contributions to the conception, proof of the theorems, analysis of applications, drafting the article, revising it critically and the final approval of the version to be published.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by Innovate UK grant nos KTP009890 and KTP010522. I.Y.T. was supported by the Russian Ministry of Education and Science, projects 8.2080.2017/4.6 (assessment and computational support for knowledge transfer algorithms between AI systems) and 2.6553.2017/BCH Basic Part.

References

1. Hilbert D. 1902 Mathematical problems. *Bull. Am. Math. Soc.* **8**, 437–479. (doi:10.1090/S0002-9904-1902-00923-3)
2. Hilbert D. 1902 *The foundations of geometry*. La Salle, IL: Open Court Publishing Company. See http://www.gutenberg.org/ebooks/17384.
3. Von Neumann J. 1955 *Mathematical foundations of quantum mechanics*. Princeton, NJ: Princeton University Press. (English translation from German Edition, Springer, Berlin, 1932.)
4. Kolmogorov AN. 1956 *Foundations of the theory of probability*. New York, NY: Chelsea Publ. (English translation from German edition, Springer, Berlin, 1933.)
5. Khinchin AY. 1949 *Mathematical foundations of statistical mechanics*. New York, NY: Courier Corporation. (English translation from the Russian edition, Moscow, Leningrad, 1943.)
6. Gibbs GW. 1960 [1902] *Elementary principles in statistical mechanics, developed with especial reference to the rational foundation of thermodynamics*. New York, NY: Dover Publications.
7. Markus L, Meyer KR. 1974 *Generic Hamiltonian dynamical systems are neither integrable nor ergodic*. Memoirs of Amer. Math. Soc., vol. 144. Providence, RI: American Mathematical Society. (doi:10.1090/memo/0144)
8. Oxtoby JC, Ulam SM. 1941 Measure-preserving homeomorphisms and metrical transitivity. *Ann. Math.* **42**, 874–920. (doi:10.2307/1968772)
9. Dobrushin RL. 1997 A mathematical approach to foundations of statistical mechanics. *Atti dei Convegni Lincei – Accademia Nazionale dei Lincei* **131**, 227–244. See http://www.mat.univie.ac.at/~esiprpr/esi179.pdf.
10. Batterman RW. 1998 Why equilibrium statistical mechanics works: universality and the renormalization group. *Philos. Sci.* **65**, 183–208. (doi:10.1086/392634)
11. Jaynes ET. 1967 Foundations of probability theory and statistical mechanics. In *Delaware seminar in the foundations of physics* (ed. M Bunge), pp. 77–101. Berlin, Germany: Springer. (doi:10.1007/978-3-642-86102-4_6)
12. Giannopoulos AA, Milman VD. 2000 Concentration property on probability spaces. *Adv. Math.* **156**, 77–106. (doi:10.1006/aima.2000.1949)
13. Gromov M. 2003 Isoperimetry of waists and concentration of maps. *Geom. Funct. Anal.* **13**, 178–215. (doi:10.1007/s00039-009-0703-1)
14. Talagrand M. 1995 Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathematiques de l'IHES* **81**, 73–205. (doi:10.1007/BF02699376)
15. Ledoux M. 2001 *The concentration of measure phenomenon*. Mathematical Surveys & Monographs, no. 89. Providence, RI: AMS. (doi:10.1090/surv/089)
16. Ball K. 1997 An elementary introduction to modern convex geometry. In *Flavors of geometry* (ed. S Levy), vol. 31, pp. 1–58. Cambridge, UK: MSRI Publications. See http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4601.
17. Lévy P. 1951 *Problèmes concrets d'analyse fonctionnelle*. Paris, France: Gauthier-Villars.
18. Donoho DL. 2000 High-dimensional data analysis: the curses and blessings of dimensionality. In *AMS Math Challenges of the 21st Century, Los Angeles, CA, 6–11 August 2000.* See http://statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf.
19. Anderson J, Belkin M, Goyal N, Rademacher L, Voss J. 2014 The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. *J. Mach. Learn. Res.: Workshop Conf. Proc.* **35**, 1–30. See http://proceedings.mlr.press/v35/anderson14.pdf.
20. Kainen PC. 1997 Utilizing geometric anomalies of high dimension: when complexity makes computation easier. In *Computer-intensive methods in control and signal processing: the curse of dimensionality* (eds M Kárný, K Warwick), pp. 283–294. New York, NY: Springer. (doi:10.1007/978-1-4612-1996-5_18)
21. Hecht-Nielsen R. 1994 Context vectors: general-purpose approximate meaning representations self-organized from raw data. In *Computational intelligence: imitating life* (eds J Zurada, R Marks, C Robinson), pp. 43–56. New York, NY: IEEE Press.
22. Gorban AN, Romanenko I, Burton R, Tyukin I. 2016 One-trial correction of legacy AI systems and stochastic separation theorems. (https://arxiv.org/abs/1610.00494)
23. Pestov V. 2013 Is the *k*-NN classifier in high dimensions affected by the curse of dimensionality? *Comput. Math. Appl.* **65**, 1427–1437. (doi:10.1016/j.camwa.2012.09.011)

24. Cucker F, Smale S. 2002 On the mathematical foundations of learning. *Bull. Am. Math. Soc.* **39**, 1–49. (doi:10.1090/S0273-0979-01-00923-5)
25. Friedman J, Hastie T, Tibshirani R. 2009 *The elements of statistical learning*. New York, NY: Springer. (doi:10.1007/978-0-387-84858-7)
26. Vapnik V. 2000 *The nature of statistical learning theory*. New York, NY: Springer. (doi:10.1007/978-1-4757-3264-1)
27. Xu R, Wunsch D. 2008 *Clustering*. Hoboken, NJ: John Wiley & Sons. (doi:10.1002/9780470382776)
28. Gorban AN, Kégl B, Wunsch D, Zinovyev A (eds). 2008 *Principal manifolds for data visualisation and dimension reduction*. Lect. Notes Comput. Sci. Eng., vol. 58. Berlin, Germany: Springer. (doi:10.1007/978-3-540-73750-6)
29. Gorban AN, Zinovyev A. 2010 Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. Neural Syst.* **20**, 219–232. (doi:10.1142/S0129065710002383)
30. Hyvärinen A, Oja E. 2000 Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 41–430. (doi:10.1016/S0893-6080(00)00026-5)
31. Mirkin B. 2012 *Clustering: a data recovery approach*. Boca Raton, FL: CRC Press. (doi:10.1201/b13101)
32. Gorban AN, Mirkes EM, Zinovyev A. 2016 Piece-wise quadratic approximations of arbitrary error functions for fast and robust machine learning. *Neural Netw.* **84**, 28–38. (doi:10.1016/j.neunet.2016.08.007)
33. Gromov M. 2011 *Structures, learning and ergosystems*, chapters 1–4, 6. Bures-sur-Ivette, France: IHES. See http://www.ihes.fr/~gromov/PDF/ergobrain.pdf.
34. Engel A, Van den Broeck C. 2001 *Statistical mechanics of learning*. Cambridge, UK: Cambridge University Press.
35. Wang D, Li M. 2017 Stochastic configuration networks: fundamentals and algorithms. *IEEE Trans. Cybern.* **47**, 3466–3479. (doi:10.1109/TCYB.2017.2734043)
36. Gorban AN. 1990 *Training neural networks*. Moscow, Russia: USSR-USA JV 'ParaGraph'. (doi:10.13140/RG.2.1.1784.4724)
37. De Freitas N, Andrieu C, Højen-Sørensen P, Niranjan M, Gee A. 2001 Sequential Monte Carlo methods for neural networks. In *Sequential Monte Carlo methods in practice* (eds A Doucet, N de Freitas, N Gordon), pp. 359–379. New York, NY: Springer. (doi:10.1007/978-1-4757-3437-9_17)
38. Scardapane S, Wang D. 2017 Randomness in neural networks: an overview. *WIREs Data Mining Knowl. Discov.* **7**, e1200. (doi:doi.org/10.1002/widm.1200)
39. Fisher RA. 1936 The use of multiple measurements in taxonomic problems. *Ann. Hum. Genet.* **7**, 179–188. (doi:10.1111/j.1469-1809.1936.tb02137.x)
40. Rosenblatt F. 1962 *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Washington, DC: Spartan Books. See http://www.dtic.mil/docs/citations/AD0256582.
41. Duda RD, Hart PE, Stork DG. 2012 *Pattern classification*. New York, NY: John Wiley and Sons.
42. Aggarwal CC. 2015 *Data mining: the textbook*. Berlin, Germany: Springer. (doi:10.1007/978-3-319-14142-8)
43. Kůrková V, Sanguineti M. 2017 Probabilistic lower bounds for approximation by shallow perceptron networks. *Neural Netw.* **91**, 34–41. (doi:10.1016/j.neunet.2017.04.003)
44. Johnson WB, Lindenstrauss J. 1984 Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* **26**, 189–206. (doi:10.1090/conm/026/737400)
45. Dasgupta S, Gupta A. 2003 An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms* **22**, 60–65. (doi:10.1002/rsa.10073)
46. Gorban AN, Tyukin I, Prokhorov D, Sofeikov K. 2016 Approximation with random bases: pro et contra. *Inf. Sci.* **364–365**, 129–145. (doi:10.1016/j.ins.2015.09.021)
47. Kainen P, Kůrková V. 1993 Quasiorthogonal dimension of Euclidian spaces. *Appl. Math. Lett.* **6**, 7–10. (doi:10.1016/0893-9659(93)90023-G)
48. Gorban AN, Tyukin IY, Romanenko I. 2016 The blessing of dimensionality: separation theorems in the thermodynamic limit. *IFAC-PapersOnLine* **49**, 64–69. (doi:10.1016/j.ifacol.2016.10.755)
49. Gorban AN, Tyukin IY. 2017 Stochastic separation theorems. *Neural Netw.* **94**, 255–259. (doi:10.1016/j.neunet.2017.07.014)
50. Chapelle O. 2007 Training a support vector machine in the primal. *Neural Comput.* **19**, 1155–1178. (doi:10.1162/neco.2007.19.5.1155)

51. Hoeffding W. 1963 Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **301**, 13–30. (doi:10.1080/01621459.1963.10500830)
52. Tyukin IY, Gorban AN, Sofeikov K, Romanenko I. 2017 Knowledge transfer between artificial intelligence systems. (https://arxiv.org/abs/1709.01547)
53. Jia Y. 2013 Caffe: an open source convolutional architecture for fast feature embedding. See http://caffe.berkeleyvision.org/.
54. Chen T, Li M, Li Y, Lin M, Wang N, Xiao T, Xu B, Zhang C, Zhang Z. 2015 MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. See https://github.com/dmlc/mxnet.
55. Team DD. 2016 Deeplearning4j: open-source distributed deep learning for the JVM. Apache Software Foundation License 2.0. See http://deeplearning4j.org.
56. Abadi M *et al.* 2015 TensorFlow: large-scale machine learning on heterogeneous systems. See https://www.tensorflow.org/.
57. Buchtala O, Sick B. 2007 Basic technologies for knowledge transfer in intelligent systems. In *Proc. of the IEEE Symp. on Artificial Life, Honolulu, HI, 1–5 April 2007, ALIFE'07*, pp. 251–258. New York, NY: IEEE Press. (doi:10.1109/ALIFE.2007.367804)
58. Tyukin IY, Gorban AN, Calvo C, Makarova J, Makarov VA. 2017 High-dimensional brain. A tool for encoding and rapid learning of memories by single neurons. (https://arxiv.org/abs/1710.11227)
59. Ison MJ, Quian Quiroga R, Fried I. 2015 Rapid encoding of new memories by individual neurons in the human brain. *Neuron* **87**, 220–230. (doi:10.1016/j.neuron.2015.06.016)
60. 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. (doi:10.1038/nature03687)
61. Viskontas IV, Quian Quiroga R, Fried I. 2009 Human medial temporal lobe neurons respond preferentially to personally relevant images. *Proc. Natl Acad. Sci. USA* **106**, 21 329–21 334. (doi:10.1073/pnas.0902319106)
62. Quian Quiroga R. 2012 Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* **13**, 587–597. (doi:10.1038/nrn3251)
63. Oja E. 1982 A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267–273. (doi:10.1007/BF00275687)
64. Corry L. 1997 David Hilbert and the axiomatization of physics (1894–1905). *Arch. Hist. Exact Sci.* **51**, 83–198. (doi:10.1007/BF00375141)
65. Wightman AS. 1976 Hilbert's sixth problem: mathematical treatment of the axioms of physics. In *Mathematical developments arising from Hilbert problems* (ed. FE Browder). Proc. of Symp. in Pure Mathematics, XXVIII, pp. 147–240. Providence, RI: AMS. (doi:10.1090/pspum/028.1/0436800)
66. Kégl B. 2003 Intrinsic dimension estimation using packing numbers. In *Advances in neural information processing systems '15* (*NIPS 2002*) (eds S Thrun, LK Saul, B Schölkopf), pp. 697–704. Cambridge, MA: MIT Press.
67. Levina E, Bickel PJ. 2005 Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems '17* (*NIPS 2004*) (eds LK Saul, Y Weiss, L Bottou), pp. 777–784. Cambridge, MA: MIT Press.
68. Bengio Y. 2009 Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2**, 1–127. (doi:10.1561/2200000006)
69. Moczko E, Mirkes EM, Ceceres C, Gorban AN, Piletsky S. 2016 Fluorescence-based assay as a new screening tool for toxic chemicals. *Sci. Rep.* **6**, 33922. (doi:10.1038/srep33922)
70. Cangelosi R, Goriely A. 2007 Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* **2**, 2. (doi:10.1186/1745-6150-2-2)
71. Zinovyev A, Mirkes E. 2013 Data complexity measured by principal graphs. *Comput. Math. Appl.* **65**, 1471–1482. (doi:10.1016/j.camwa.2012.12.009)