Stochastic Separation Theorems: How Geometry May Help to Correct AI Errors

Alexander Gorban, Bogdan Grechuk, and Ivan Tyukin

1. Introduction

Recent years have seen explosive progress in data-driven artificial intelligence (AI) systems. Many decades of the development of mathematics underpinning statistical learning theory coupled with advancements in approximation theory, numerical analysis, technology, and computing gave rise to new-generation AI transforming our life. These systems show great promise in cancer diagnostics [MSG+20], they are a part of autonomous cars [22], automated face recognition and biometrics [KE21], image segmentation [SBKV+20], language processing and translation tools [DZS+22], and as such become our new reality. Availability of unprecedented volumes of data, citizens' expectations and participation are further driving this change.

Alexander Gorban holds a chair in applied mathematics at the University of Leicester. His email address is a.n.gorban@leicester.ac.uk.
Bogdan Grechuk is a lecturer of mathematics at the University of Leicester. His email address is bg83@leicester.ac.uk.
Ivan Tyukin is a professor of mathematical data science and modelling at King's College London. His email address is ivan.tyukin@kcl.ac.uk.

Communicated by Notices *Associate Editor Emilie Purvine.*

DOI: https://doi.org/10.1090/noti2599

New reality, however, brings new challenges. Uncertainties and biases are inherent within any empirical data. They enter production pipelines of data-driven AI and ripple through them causing errors. AI instabilities and adversarial examples—errors due to minor changes in data or structure—have recently been found in many advanced data-driven AI models. Moreover, mounting evidence suggests that these errors are in fact expected in such systems [THG20] and may not always be cured by larger volumes of data or better training algorithms [BHV21] as long as the AI architecture remains fixed.

This leads to the following question: if errors are inevitable in data-driven AI then how do we deal with them once they occur?

One way to address this imminent challenge is to equip an AI with an "error filter" or "error corrector" [GT18]. The function of the AI corrector is to learn from errors "on-the-job," supplementing the AI's initial training. Dynamic addition of AI correctors continuously extends AI architecture, adapts to data uncertainty [GGG+18], and enables AI to escape the stability barrier revealed in [BHV21]. When a new data arrives at AI input, the AI error corrector then decides if it is likely to cause an error, and if so, then reports. To do this, the filter uses some set I of attributes, such as, for example, internal latent representations of the input in AI decision space. To each attribute $i \in I$, the system assigns some weight w_i. For each new input, the system computes numerical values x_i of all attributes $i \in I$, and compares the weighted sum $\sum_{i\in I} w_i x_i$ with some threshold t to decide whether to report the input as an error.

However, how does the filter determine the weights w_i of all attributes? To do this, the filter is provided a training set of example inputs marked as correct and errors. Then the system tries to find weights w_i such that, ideally, all data in the training set are classified correctly. Moreover the system tries to ensure that all (or a large proportion of) future "unseen" inputs would be processed correctly too. In other words, the system seeks to learn the weights from some training data, and the error filter itself is therefore an example of a machine learning (ML) system.

Geometrically, any input is described by the values x_i of the attributes, and can therefore be represented as a point $x = (x_1, \dots, x_n)$ in the n-dimensional Euclidean space, where $n = |I|$ is the number of attributes. Then the criterion $\sum_{i\in I} w_i x_i \geq t$ for an input being an error defines a half-space, whose boundary is the hyperplane H defined by the equation $\sum_{i\in I} w_i x_i = t$. If we mark points corresponding to errors and correct AI behavior as red and blue, respectively, the machine learning task of error identification reduces to finding a hyperplane that separates the red points from the blue ones; see Figure 1.

Assume that such hyperplane H exists and we have started to use the filter with the corresponding weights w_i.

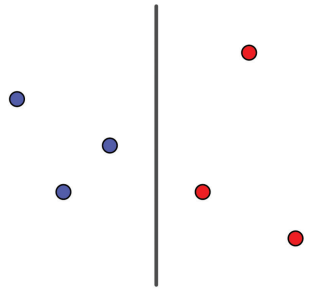

Figure 1. Separation of red and blue points by a hyperplane.

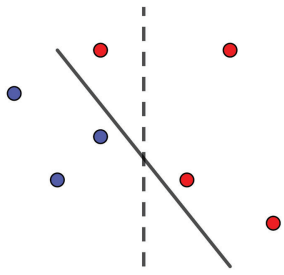

Figure 2. Retraining the system by recomputing a hyperplane.

Imagine, however, that a new input has arrived, which the filter classified as correct but the user marked as an error. In other words, the filter itself made an error. Of course, we would like the system to be able to learn from such errors and improve its performance in the future. An obvious way to do this is to add a new point to the training set and recompute the weights. This constitutes "retraining the system." Geometrically, this means that a new red point X appears on the "wrong" side of the hyperplane, so that we try to find a different hyperplane that separates all points correctly; see Figure 2. Obviously, it is not always possible to find such a hyperplane; see Figure 3. Moreover, even if it is possible, it may require substantial time to recompute all weights every time the filter makes an error.

Alternatively, one may use the following error-correction method, suggested in [GMT19]: separate a new red point X from the existing blue points by another hyperplane H' given by an equation $\sum_{i\in I} w'_i x_i = t'$; see Figure 3. After this, classify any new input as error if either $\sum_{i\in I} w_i x_i \geq t$ or $\sum_{i\in I} w'_i x_i \geq t'$.

A careful reader may have already noticed a limitation of this approach that appears to be fundamental: why did we assume that a point can be separated from all other points by a hyperplane? Obviously, if that point belongs

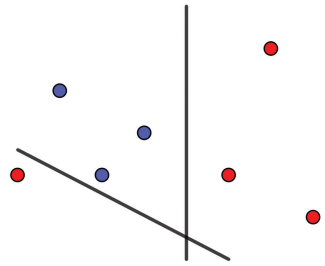

Figure 3. Separation of new red point by a different hyperplane.

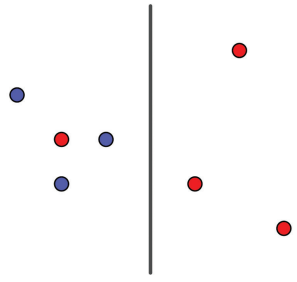

Figure 4. A red point not separable by a hyperplane.

to the convex hull[1] of other points, then a separating hyperplane does not exist and the method does not work; see Figure 4. For example, even if we have just 3 points X_1, X_2, X_3, then X_3 may lie in the interior of the line interval X_1X_2, and in this case it cannot be separated from X_1, X_2.

However, intuitively, the described case is in some sense "degenerate" and should not happen too often with real data. The best way to formalise this intuition is to use the language of probability theory, and ask what is the probability that the method would work for random data. This leads to a very nice problem that lies on the borderline of probability theory and geometry.

Problem 1. Given a set K of m random points in $\mathbb{R}^n$, what is the probability that each point $X \in K$ can be separated from all other points by a hyperplane? Equivalently, what is the probability that points in K are in convex position (in sense that each point $X \in K$ is a vertex of the convex hull of K)?

2. Sylvester's Problem

Problem 1 has a long history and goes back to at least the question asked by Sylvester in 1864: given 4 random points X, Y, Z, W on the plane, what is the probability p that they form a convex quadrilateral?

To address this question, it is convenient to introduce the following random variables. Let I_X be the random variable equal to 1 if point X is inside the triangle YZW and 0 otherwise. Let random variables I_Y, I_Z and I_W be defined similarly. Then random variable

$$I = I_X + I_Y + I_Z + I_W$$

counts the number of points that are inside the triangle formed by other points. Hence, $I = 0$ precisely if X, Y, Z, W form a convex quadrilateral, and this happens with probability p. With probability $1 - p$, $I = 1$. Thus, the expected value $\mathbb{E}[I] = 0 \cdot p + 1 \cdot (1 - p) = 1 - p$, and $p = 1 - \mathbb{E}[I]$.

This implies that to find p it suffices to find $\mathbb{E}[I]$. From the linearity of the expectation, and assuming that X, Y, Z, W are drawn independently from the same distribution,

$$\mathbb{E}[I] = \mathbb{E}[I_X] + \mathbb{E}[I_Y] + \mathbb{E}[I_Z] + \mathbb{E}[I_W] = 4\mathbb{E}[I_X].$$

Next, I_X is a random variable that takes values 0 or 1, and

$$\mathbb{E}[I_X] = 0 \cdot (1 - p_X) + 1 \cdot p_X = p_X,$$

where p_X is the probability that X lies inside triangle YZW.

If points X, Y, Z, W are selected independently and uniformly at random from the unit disk $\mathbb{D}$, then by the law of total expectation,

$$\mathbb{E}[I_X] = \mathbb{E}[\mathbb{E}[I_X|Y, Z, W]] = \mathbb{E}\left[\frac{A(YZW)}{A(\mathbb{D})}\right]$$

where A denotes the area. Hence, the problem reduces to determining the expected area of the triangle YZW. In 1867, Woolhouse determined that

$$\mathbb{E}\left[\frac{A(YZW)}{A(\mathbb{D})}\right] = \frac{35}{48\pi^2},$$

hence

$$p = 1 - \mathbb{E}[I] = 1 - 4\mathbb{E}[I_X] = 1 - \frac{35}{12\pi^2} = 0.704\ldots.$$

Of course, random points can be selected inside regions different from a disk. Sylvester also asked the same question in the following modified form. Let S be a convex body in the plane (that is, a compact convex set with nonempty interior) and choose four points from S independently and uniformly at random. What is the probability $p(4, S)$ that these points are the vertices of a convex quadrilateral? Further, for what S is this probability the smallest and the largest? The second question has been solved by Blaschke, who proved in 1917 that for all convex bodies S,

$$\frac{2}{3} = p(4, \mathbb{T}) \le p(4, S) \le p(4, \mathbb{D}) = 1 - \frac{35}{12\pi^2} = 0.704\ldots,$$

where $\mathbb{T}$ and $\mathbb{D}$ denotes a triangle and a disk in the plane, respectively.

[1] *Recall that the convex hull of set $K \subset \mathbb{R}^n$ is the intersection of all convex sets containing K.*

Sylvester's question can be asked for m points: if they are selected uniformly at random in a convex body S in the plane, what is the probability $p(m, S)$ that they form a convex m-gon?

In 1995, Valtr solved this problem exactly for a parallelogram $\mathbb{L}$, and proved that

$$p(m, \mathbb{L}) = \left(\frac{\binom{2m-2}{m-1}}{m!} \right)^2 .$$

In 1996, Valtr also solved this problem for triangle $\mathbb{T}$, and showed that

$$p(m, \mathbb{T}) = \frac{2^m (3m-3)!}{(m-1)!^3 \, (2m)!}.$$

Using Stirling's approximation for the factorial, it is straightforward to prove that

$$\lim_{m\to\infty} \left(m^2 \sqrt[m]{p(m, \mathbb{T})} \right) = \frac{27}{2} e^2.$$

Because any convex body S in the plane can be sandwiched between two triangles, this implies the existence of universal constants $0 < c_1 < c_2 < \infty$ such that

$$c_1 \le m^2 \sqrt[m]{p(m, S)} \le c_2$$

for all m and all S. In fact, Bárány [Bár99] proved in 1999 that

$$\lim_{m\to\infty} \left(m^2 \sqrt[m]{p(m, S)} \right) = c(S)$$

for some constant $c(S)$ that depends on S. For example, $c(\mathbb{L}) = 16e^2$ for parallelogram, $c(\mathbb{T}) = \frac{27}{2} e^2$ for triangle, and $c(\mathbb{D}) = 2\pi^2 e^2$ for disk. In particular,

$$p(m, \mathbb{D}) \approx \left(\frac{2\pi^2 e^2}{m^2} \right)^m .$$

approaches 0 as $m \to \infty$ with super-exponential speed.

Can we have the exact (non-asymptotic) formulas for $p(m, \mathbb{D})$? In 1971, Miles derived the exact formula for $p(5, \mathbb{D})$:

$$p(5, \mathbb{D}) = 1 - \frac{305}{48\pi^2} = 0.356\ldots .$$

Finally, Marckert in 2017 derived exact (but somewhat complicated) formulas for $p(m, \mathbb{D})$ for an arbitrary m. For example, for $m = 6$,

$$p(6, \mathbb{D}) = 1 - \frac{305}{24\pi^2} - \frac{473473}{11520\pi^4} = 0.134\ldots .$$

The following table lists numerical values for $p(m, \mathbb{T})$, $p(m, \mathbb{L})$ and $p(m, \mathbb{D})$ for $4 \le m \le 7$.

m	4	5	6	7
$p(m, \mathbb{T})$	0.666...	0.305...	0.101...	0.0251...
$p(m, \mathbb{L})$	0.694...	0.340...	0.122...	0.0336...
$p(m, \mathbb{D})$	0.704...	0.356...	0.134...	0.039...

As expected, we see that the probabilities decrease fast even for small values of m. This is bad news for our machine learning application, because it shows that new points will most likely be in the convex hull of other points. However, *all these results are in the plane*, which corresponds to a (toy) machine learning system with just two attributes. Any real ML system has significantly more attributes, hence we should study Problem 1 in higher-dimensional spaces. In the next section we show that separability properties of random points in higher dimensions are dramatically different from those computed for our low-dimensional example.

3. The Effect of Higher Dimension

We start our analysis of Problem 1 in higher dimensions with a simple special case. Let $\mathbb{B}_n$ be the closed unit ball in $\mathbb{R}^n$. We first consider the case when points $X_1, \ldots, X_m \in \mathbb{B}_n$ are fixed, and $Y \in \mathbb{B}_n$ is selected uniformly at random in $\mathbb{B}_n$. In 1986, Elekes [Ele86] proved that for any m points $X_1, \ldots, X_m \in \mathbb{B}_n$, we have

$$\frac{\mathrm{Vol}(\mathrm{conv}(X_1, \ldots, X_m))}{\mathrm{Vol}(\mathbb{B}_n)} \le \frac{m}{2^n}, \tag{1}$$

where conv is the convex hull, and Vol denotes the n-dimensional volume. This implies that Y can be separated from $X_1, \ldots, X_m$ by a hyperplane with probability at least $1 - m/2^n$. This probability is greater than $1 - \delta$ provided that $m/2^n < \delta$, or

$$m < \delta 2^n. \tag{2}$$

Now assume that we select m points independently and uniformly at random in $\mathbb{B}_n$. Let E_i be the event that the point X_i is inside the convex hull of the remaining points. Then Elekes's theorem implies that the probability of E_i is at most $(m-1)2^{-n}$, and the probability of the event $E = \bigcup_{i=1}^m E_i$ is at most $m(m-1)2^{-n} < m^2 2^{-n}$. Hence with the probability greater than $1 - m^2 2^{-n}$ every point X_i is separable by a hyperplane from the remaining points. This probability is greater than $1 - \delta$ if $m^2 2^{-n} < \delta$, or

$$m < \sqrt{\delta}(\sqrt{2})^n. \tag{3}$$

The upper bound (3) was originally proved by Bárány and Füredi in 1988. Complementing this result, Bárány and Füredi also proved that, for all $n \ge 100$, the probability that

$$m = 20n^{3/4}(\sqrt{2})^n$$

independent uniformly distributed points in $\mathbb{B}_n$ are all vertices of their convex hull is less than $2e^{-10}$. Hence, the bound (3) is quite tight. In particular, the result is no longer true if $(\sqrt{2})^n$ in (3) is replaced by $(\sqrt{2} + \epsilon)^n$ for any $\epsilon > 0$.

The following table shows, in various dimensions n, the upper bounds for m in (2) and (3) with $\delta = 0.01$, ensuring the separability with 99% probability.

n	Upper bound in (2)	Upper bound in (3)
10	10.24	3.2
30	$1.07 \cdot 10^7$	3276
50	$1.12 \cdot 10^{13}$	$3.35 \cdot 10^6$
100	$1.26 \cdot 10^{28}$	$1.12 \cdot 10^{14}$

We see that in dimension $n = 30$, a random point is separable from millions of other points with probability over 99%, and thousands of random points are all separable. In dimension $n = 50$, millions of points all become separable. In other words, if we select 3 million uniformly random points in ball $B_{50} \subset \mathbb{R}^{50}$, then with probability over 99% they are all vertices of their convex hull. This observation is in sharp contrast with our low-dimensional intuition.

This effect is not limited to the uniform distribution in the unit ball $\mathbb{B}_n$. In fact, when we say "uniform distribution in the unit ball," we actually mean a *family* of distributions, one for each dimension: the uniform distribution on the interval $[-1, 1]$ in $\mathbb{R}^1$, the uniform distribution in the the disk $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \le 1\}$ in $\mathbb{R}^2$, and so on. In the theorems below, the dimension will not be fixed but will be a variable, and in this case we need to consider a family

$$\mathbb{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_n, \dots\}$$

of probability measures, where $\mathbb{P}_n$ denotes the probability measure on $\mathbb{R}^n$.

Definition 1. [GGG+18] The family of joint distributions of points $X_1, \dots, X_m$ in $\mathbb{R}^n$ has **SmAC property** if there exist constants $\epsilon > 0$, $A > 0$, and $B \in (0, 1)$, such that for every positive integer n, any convex set $S \in \mathbb{R}^n$ such that

$$\frac{\mathrm{Vol}(S)}{\mathrm{Vol}(\mathbb{B}_n)} \le \epsilon^n,$$

any index $i \in \{1, 2, \dots, m\}$, and any points $Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_m$ in $\mathbb{R}^n$, we have

$$\mathbb{P}(X_i \in \mathbb{B}_n \setminus S \mid X_j = Y_j, \forall j \ne i) \ge 1 - AB^n. \qquad (4)$$

Condition (4) says that, with probability exponentially close to 1, a random point lies inside the unit ball, but outside of any convex set of exponentially small volume. In other words, SmAC property holds for the distributions without (i) heavy tails and (ii) sharp peaks in sets with exponentially small volume. Indeed, any bounded or light-tailed distribution can, after appropriate shift and rescaling, be located essentially inside $\mathbb{B}_n$, while for heavy-tailed distributions there is a significant probability that $X_i \notin \mathbb{B}_n$, hence (4) fails. The name SmAC is an abbreviation of "SMeared Absolute Continuity" and comes from analogy with absolute continuity: the absolute continuity means that the sets of zero measure have zero probability, and the SmAC condition requires that convex sets with exponentially small volume should not have high probability.

The theorem below states that if a family of distributions has the SmAC property, then exponentially many points are in convex position with high probability.

Theorem 1. [GGG+18] *Let $\{X_1, \dots, X_m\}$ be a set of random points in $\mathbb{R}^n$ from a distribution satisfying the SmAC property. Let $\delta \in (0, 1)$ be fixed. Then there exists constants $a > 0$ and $c > 1$ such that if $m < ac^n$ then points $\{X_1, \dots, X_m\}$ are in convex position with probability greater than $1 - \delta$.*

The SmAC condition is very general and holds for a large variety of distributions. As an illustration, consider a special case of i.i.d. data. If probability measures $\mathbb{P}_n$ in family $\mathbb{P}$ have support in the unit ball $\mathbb{B}_n$ and density ρ_n, then the SmAC condition holds provided that

$$\frac{\rho_n(x)}{\rho_{\mathrm{uni}}(x)} \le CR^n, \qquad \forall x \in \mathbb{B}_n \qquad (5)$$

where $C > 0$ and $R > 0$ are some constants independent of the dimension, and $\rho_{\mathrm{uni}}(x)$ is the density of the uniform distribution in $\mathbb{B}_n$. In other words, the density $\rho_n(x)$ is allowed to differ from the uniform density by an exponentially large factor, and the exponent R must be a constant independent of n but can be arbitrarily large.

For example, let A_n be a bounded measurable set in $\mathbb{R}^n$. Then it is not difficult to see that (5) is true for the uniform distribution in (a possibly scaled and shifted) A_n provided that

$$\frac{\mathrm{diam}(A_n)}{\sqrt[n]{\mathrm{Vol}(A_n)}} \le R\sqrt{n} \qquad (6)$$

for some constant $R < \infty$. In particular, if A_n is the unit cube in $\mathbb{R}^n$, then $\mathrm{Vol}(A_n) = 1$, $\mathrm{diam}(A_n) = \sqrt{n}$, and (6) holds with $R = 1$. Hence Theorem 1 implies that exponentially many points selected uniformly at random from the unit cube are in convex position with high probability.

4. Computing Separating Hyperplanes

Under SmAC condition, exponentially many random points $X_1, \dots, X_m$ in $\mathbb{R}^n$ are linearly separable with high probability: for each $i \in 1, \dots, m$, there exists a hyperplane H passing through X_i such that all other points are on the same side from H. If $(x_{j1}, \dots, x_{jn})$ are the coordinates of point X_j, $j = 1, \dots, m$, then we can explicitly find H by solving the quadratic program

$$\min_{c_1, \dots, c_n, v} \|c\|^2, \quad \text{subject to} \qquad (7)$$

$$\sum_{k=1}^{n} c_k x_{jk} + v \le -1, \quad j \ne i; \quad \sum_{k=1}^{n} c_k x_{ik} + v = 1.$$

If $c^* = (c_1^*, \dots, c_n^*, v^*)$ is the solution to (7), then

$$H = \left\{(x_1, \dots, x_n) \in \mathbb{R}^n : 1 - v^* = \sum_{k=1}^{n} c_k^* x_k\right\}.$$

The above program is a version of the well-known maximal-margin classifier or a support vector machine.

This quadratic program has m constraints and $n+1$ variables. Worst-case computational complexity of solving this problem scales as $O(\max(n+1,m)\min(n+1,m)^2)$ [Cha07]. When m is potentially exponentially large in n, the worst-case complexity grows exponentially with n.

The other issue with finding separating hyperplanes through solving (7) is that this approach requires full knowledge of all points X_j, $j=1,\dots,m$. Whilst such knowledge might be available in some tasks, it is hardly practical in the task of *correcting AI errors*. In this context, X_i represents an AI "error" that has already been detected and is to be removed, and X_j, $j\neq i$ stand for "correct or expected" past and possibly future AI behavior. The fact that some or all X_j are unknown makes solving (7) hardly possible. The question, therefore, is:

Problem 2. How to construct H separating X_i from the remaining points without knowing their positions?

In the next sections we show that, for appropriately high dimension n and under some mild assumptions, there are simple closed-form expressions defining hyperplanes separating X_i from X_j, $i\neq j$ with probability close to 1.

4.1. **One-shot separability: Fisher separability.** In order to develop the intuition for Problem 2, let us return to the simplest example, when the points are selected uniformly at random from the n-dimensional unit ball $\mathbb{B}_n$. Any hyperplane H through X_i divides $\mathbb{B}_n$ into pieces with volumes $V_1\leq V_2$. To maximize the chance that hyperplane H separates X_i from all other points, we aim to select H such that volume V_1 is the minimal possible. The optimal choice of H is the hyperplane orthogonal to OX_i, where O is the centre of $\mathbb{B}_n$; see Figure 5. If A is the event that point X_j belongs to the piece with volume V_1, then a straighforward calculation shows that

$$\mathbb{P}(A)=\mathbb{E}[I_A]=\mathbb{E}[\mathbb{E}[I_A|X_j]]=\mathbb{E}[R^n]=\frac{1}{2^{n+1}}, \quad (8)$$

where I_A is the indicator function of the event A, the second equality is the law of total expectation, the third equality follows from the fact that $I_A|X_j$ is equal to 1 if and only if X_i belongs to a ball with radius $R=|OX_j|/2$, and the last equality follows from the fact that R is a random variable with cdf $\mathbb{P}[R\leq r]=\mathbb{P}[|OX_j|\leq 2r]=(2r)^n$, $\quad 0\leq r\leq 1/2$.

Now, if we have m i.i.d. points from $\mathbb{B}_n$, there are $m(m-1)$ ordered pairs of points. Hence, the probability that we can find some pair X_i, X_j such that the corresponding event A happens is at most $m(m-1)2^{-(n+1)}<m^2 2^{-(n+1)}$. This probability is less than δ provided that

$$m<\sqrt{2\delta}(\sqrt{2})^n.$$

Remarkably, this bound is even less restrictive than (3), while the conclusion is stronger: not only are this many points in convex position with probability greater than $1-\delta$, but in fact each point X_i can be separated from the other ones by the *specific* hyperplane tangent to OX_i, which is independent from other points, and can be constructed exponentially faster than solving the program (7).

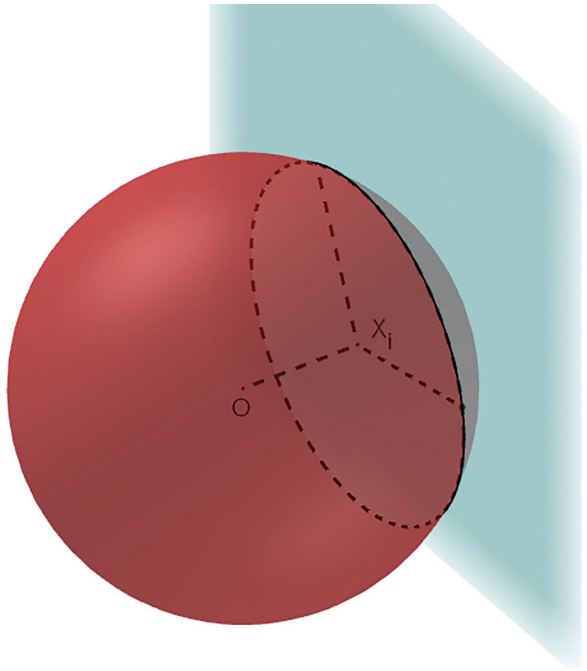

Figure 5. One-shot separability in a sphere.

It turns out that this simple idea to choose hyperplane H tangent to OX_i solves Problem 2 for surprisingly many families of distributions, and is known as Fisher separability [GGG+18].

Definition 2. A point $X\in\mathbb{R}^n$ is **Fisher-separable** from $Y\in\mathbb{R}^n$ with threshold $\alpha\in(0,1]$ if

$$\alpha\|X\|^2\geq\sum_{i=1}^n x_i y_i. \quad (9)$$

We say that X is Fisher-separable from a finite set $F\in\mathbb{R}^n$ with threshold α if (9) holds for all $Y\in F$.

The question is: how do we know that X_i is Fisher-separable from X_j, $i\neq j$? An answer to this question follows from the next statement.

Proposition 1 ([GGG+18]). *Let $\alpha\in(1/2,1]$, $1>\delta>0$, let X be drawn from a distribution supported on $\mathbb{B}_n$ whose probability density satisfies (5) with some $C>0$ and $R\in(1,2\alpha)$, and let Y be a finite set in $\mathbb{B}_n$ with*

$$|Y|\leq\delta\left(\frac{2\alpha}{R}\right)^n\frac{1}{C}.$$

Then the point X is Fisher-separable from the set Y with probability at least $1-\delta$.

Several interesting observations stem immediately from Proposition 1. It appears that construction of separating hyperplanes does not always require complete knowledge

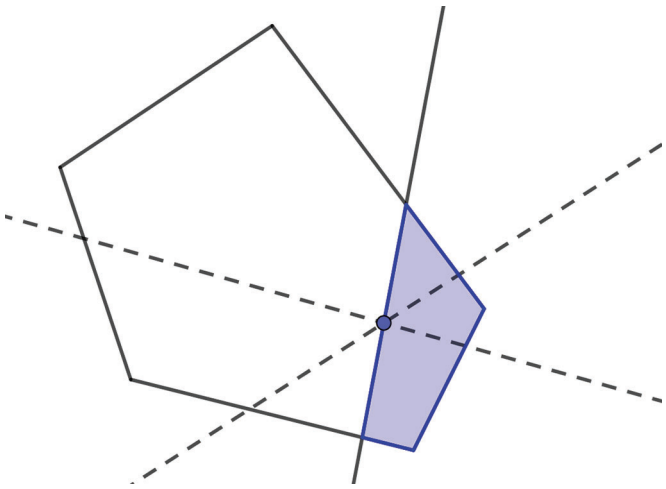

Figure 6. One-shot separability: we select a hyperplane that divides the probability measure into two maximally unequal parts.

of sets that are being separated. Some rough information such as the value of the point O, the fact that all X_j, $j \neq i$ are in a unit ball centered at O, and that X_i is drawn from a SmAC distribution suffices. The resulting hyperplane H separates X_i from X_j, $j \neq i$ with probability at least $1-\delta$, and with some guaranteed margin $(1-\alpha)\|X_i\|$. Note, however, that this margin is not necessarily maximal as requested by program (7).

It turns out that Fisher separability for exponentially many points holds for many important families of distributions, including rotation invariant log-concave distributions and product distributions whose components have bounded support or very fast-decaying tails [GGT21]. At the same time, there are examples of product distributions with identical log-concave components for which this is no longer true [GGT21]. It is hence natural to ask if and how similar simple solutions could be derived for such distributions with "heavier" tails.

4.2. **One-shot separability: general case.** Now we formulate the same idea in general. Let $\mathbb{P}_n$ be an arbitrary probability measure in $\mathbb{R}^n$, and let $X \in \mathbb{R}^n$ be an arbitrary point. Problem 2 asks to construct a hyperplane separating X from other m points selected at random from $\mathbb{P}_n$ without knowing their positions. Every hyperplane divides $\mathbb{R}^n$ into two half-spaces, say H_1 and H_2, whose probability measures are $p_1 = \mathbb{P}_n(H_1)$, and $p_2 = \mathbb{P}_n(H_2)$, respectively. We would like X and the remaining m points to belong to different subspaces, say $X \in H_1$, and other m points to belong H_2. The probability of the latter event is $p_2^m = (1-p_1)^m$. This probability is maximized if p_1 is minimized. Hence, the idea is to select the halfspace containing X whose probability measure is minimal; see Figure 6. Formally, let $\mathbb{H}_X$ be the set of halfspaces of $\mathbb{R}^n$ containing X, let

$$\phi(\mathbb{P}_n, X) = \inf_{H \in \mathbb{H}_X} \mathbb{P}_n(H) \tag{10}$$

be the minimal measure of a halfspace containing X, and let $H^*(X)$ be the minimizer[2] in (10). Function $\phi(\mathbb{P}_n, X)$ is known as Tukey's halfspace depth.

The probability that $H^*(X)$ separates x from m points is $(1-\phi(\mathbb{P}_n, X))^m$. We would like this probability to be greater than a given constant $1-\delta$ even if m grows exponentially fast with n. To ensure this, $\phi(\mathbb{P}_n, X)$ should decrease exponentially fast with n. This may not be the case for all X: for example, if $\mathbb{P}_n$ is the uniform distribution in the ball, and X is the center of the ball, then $\phi(\mathbb{P}_n, X) = 1/2$. However, there is a hope that $\phi(\mathbb{P}_n, X)$ decreases fast on average, for random point X. In other words, we need expected value

$$c(\mathbb{P}_n) = \mathbb{E}[\phi(\mathbb{P}_n, X)]$$

to decrease exponentially fast with n.

Definition 3. Let $\mathbb{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_n, \dots\}$ be a family of probability measures, where $\mathbb{P}_n$ is the probability measure on $\mathbb{R}^n$. We say that $\mathbb{P}$ has **exponential one-shot separability** if

$$c(\mathbb{P}_n) \leq a_{\mathbb{P}}(c_{\mathbb{P}})^n$$

for some constants $a_{\mathbb{P}} < \infty$, $c_{\mathbb{P}} \in (0,1)$.

In this section, we overview our recent results that establish exponential one-shot separability for a large class of product distributions, and discuss a conjecture that this property holds for all log-concave distributions.

Let us now be a bit more formal. We say that density $\rho_n : \mathbb{R}^n \to [0,\infty)$ of random vector $X = (x_1, \dots, x_n)$ (and the corresponding probability measure $\mathbb{P}_n$) is *log-concave*, if set

$$D = \{z \in \mathbb{R}^n \mid \rho_n(z) > 0\}$$

is convex and $g(z) = -\log(\rho_n(z))$ is a convex function on D. For example, the uniform distribution in an arbitrary convex body is log-concave. Let C be the variance-covariance matrix of X, that is, matrix with components $c_{ij} = \mathrm{Cov}(x_i, x_j)$. Because the log-concavity of $\mathbb{P}_n$ and the definition of $c(\mathbb{P}_n)$ are invariant under invertible linear transformations, we may assume that $\mathbb{E}[X] = 0$ and $C = I_n$ is the $n \times n$ identity matrix. Such distributions are called isotropic. Quantity

$$L_{\mathbb{P}_n} = \left(\sup_{z \in \mathbb{R}^n} \rho_n(z)\right)^{1/n}$$

is called the isotropic constant of $\mathbb{P}_n$. Very recently, Brazitikos, Giannopoulos, and Pafis [BGP22] proved that

$$c(\mathbb{P}_n) \leq \exp\left(-\frac{an}{L_{\mathbb{P}_n}}\right) \tag{11}$$

[2] *Each halfspace can be identified with its normal unit vector, the set of all such vectors is a compact set, hence there must be a halfspace that achieves the minimum.*

for some absolute constant $a > 0$. A famous conjecture in convex geometry predicts that

$$L_{\mathbb{P}_n} \geq \epsilon \tag{12}$$

for some constant $\epsilon > 0$ independent from the dimension. This conjecture has been made in 1986 by Jean Bourgain [Bou86] in the form that "There exists a universal constant $\epsilon > 0$ (independent from n) such that for any convex set K of unit volume in $\mathbb{R}^n$, there exists a hyperplane H such that the $(n-1)$-dimensional volume of the section $K \cap H$ is bounded below by ϵ," and since then is known as the Hyperplane conjecture. It turns out that this conjecture is equivalent to (12), and in fact has many other equivalent formulations. Recently, Chen [Che21] made a breakthrough and proved that

$$L_{\mathbb{P}_n} \geq n^{-f(n)}$$

for some function f tending to 0 as $n \to \infty$. Even more recently, Klartag and Lehec [KL22] improved this to $L_{\mathbb{P}_n} \geq b(\log n)^{-5}$ for some absolute constant $b > 0$. In combination with (11), a full proof of conjecture (12) would imply that any family of log-concave probability measures has exponential one-shot separability.

Our next example is a family of product distributions. Specifically, for each n, let $\mathbb{P}_n$ be the the product measure of one-dimensional probability measures $\mu_{1,n}, \dots, \mu_{n,n}$. For any distribution μ on $\mathbb{R}$, define

$$\psi_\mu(x) = \inf_{c \in \mathbb{R}} \mathbb{E}[\exp(c(Z-x))],\ c_\mu = \mathbb{E}[\psi_\mu(X)],$$

where Z and X are random variables with distribution μ. Then we have proved [GGT] that $\mathbb{P}_n$ has exponential one-shot separability provided that $c_\mu < 1$ for each component distribution μ. This property holds for a large variety of distributions. For example, we have the following sufficient condition [GGT].

Proposition 2. *Let Z be a random variable with distribution μ. Assume that Z is non-constant and $M_Z(t) := \mathbb{E}[e^{tZ}] < \infty$ for some $t \neq 0$. Then $c_\mu < 1$.*

When our data are non-negative, Proposition 2 implies the following corollary.

Corollary 1. *Let Z be a non-constant non-negative random variable with distribution μ. Then $c_\mu < 1$.*

For log-concave distributions, we have the following explicit and uniform upper bound [GGT].

Proposition 3. *For any log-concave probability distribution μ on $\mathbb{R}$, we have*

$$c_\mu < 1 - 2 \cdot 10^{-5}.$$

Proposition 3 implies the following result.

Theorem 2. *Let $\mathbb{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_n, \dots\}$ be a family of product distributions such that all component distributions are log-concave. Then $\mathbb{P}$ has exponential one-shot separability (see Definition 3) with parameters $a_\mathbb{P} = 1$ and $c_\mathbb{P} < 1 - 2 \cdot 10^{-5}$.*

We did not try to optimize the upper bound for c_μ in Proposition 3. Instead, we pose the problem of finding the optimal upper bound as an open question. Specifically, if $\mathcal{F}$ is the class of all log-concave distributions on $\mathbb{R}$, then what is the value of

$$c_\mathcal{F} = \sup_{\mu \in \mathcal{F}} c_\mu?$$

Proposition 3 provides the upper bound $c_\mathcal{F} \leq 1 - 2 \cdot 10^{-5} < 1$. On the other hand, example of Laplace distribution shows that

$$c_\mathcal{F} \geq \frac{3}{4} + \frac{e}{16} \int_1^\infty \frac{e^{-t}}{t} dt = 0.7872\dots.$$

While the upper bound is clearly non-optimal, it may be that $c_\mathcal{F}$ is equal to the lower bound.

5. Conclusions

A phenomenon known as curse of dimensionality states that many methods and techniques that are efficient in low dimension become infeasible is high dimension. Stochastic separation theorems are examples of the opposite phenomenon, blessing of dimensionality, which states that some aspects become easier in higher dimensions. The theorems state that if we have m random points in $\mathbb{R}^n$, then, with high probability, every point can be separated from all others by a hyperplane. This is true even if the number of points grows exponentially fast with dimension.

While being interesting from purely mathematical perspective, stochastic separation theorems could be a stepping stone for the development of much-needed error correcting mechanisms [GGG+18], algorithms capable of learning from just few examples [GGM+21], approaching the challenge of continuous learning without catastrohpic forgetting in machine learning and AI, and to produce new notions of data dimension [GMT19]. The theorems imply that if the number of attributes is moderately high, AI errors may be corrected by adding simple linear correctors, that are fast, easy to compute and implement, and do not destroy existing functionality of the system. The simplest corrector is based on Fisher separability discussed in Section 4.1. Deeper one-shot separation theorems discussed in Section 4.2 make the method applicable even for distributions for which Fisher separability fails.

References

[22] *Safe driving cars*, Nat. Mach. Intell. **4** (2022), 95–96.

[Bár99] Imre Bárány, *Sylvester's question: the probability that n points are in convex position*, Ann. Probab. **27** (1999), no. 4, 2020–2034. MR1742899

[BHV21] A. Bastounis, A. C. Hansen, and V. Vlačić, *The extended Smale's 9th problem–on computational barriers and paradoxes in estimation, regularisation, computer-assisted proofs and learning*, arXiv preprint arXiv:2110.15734 (2021).
[Bou86] J. Bourgain, *On high-dimensional maximal functions associated to convex bodies*, Amer. J. Math. **108** (1986), no. 6, 1467–1476. MR868898
[BGP22] S. Brazitikos, A. Giannopoulos, and M. Pafis, *Half-space depth of log-concave probability measures*, arXiv preprint arXiv:2201.11992 (2022).
[Cha07] Olivier Chapelle, *Training a support vector machine in the primal*, Neural Comput. **19** (2007), no. 5, 1155–1178. MR2309267
[Che21] Yuansi Chen, *An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture*, Geom. Funct. Anal. **31** (2021), no. 1, 34–61. MR4244847
[DZS+22] I. Drori, S. Zhang, R. Shuttleworth, L. Tang, and A. Lu et al., *A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level*, Proc. of the Nat. Acad. Sci. **119** (2022), no. 32, e2123433119, available at `https://www.pnas.org/doi/pdf/10.1073/pnas.2123433119`.
[Ele86] G. Elekes, *A geometric inequality and the complexity of computing volume*, Discrete Comput. Geom. **1** (1986), no. 4, 289–292. MR866364
[GGG+18] A. N. Gorban, A. Golubkov, B. Grechuk, E. M. Mirkes, and I. Y. Tyukin, *Correction of AI systems by linear discriminants: probabilistic foundations*, Inform. Sci. **466** (2018), 303–322. MR3847955
[GGM+21] A. N. Gorban, B. Grechuk, E. M. Mirkes, S. V. Stasenko, and I. Y. Tyukin, *High-dimensional separability for one-and few-shot learning*, Entropy **23** (2021), no. 8, 1090.
[GGT] A. N. Gorban, B. Grechuk, and I. Y. Tyukin, *One-shot separation theorems*, In preparation.
[GMT19] A. N. Gorban, V. A. Makarov, and I. Y. Tyukin, *The unreasonable effectiveness of small neural ensembles in high-dimensional brain*, Physics of Life Reviews **29** (2019), 55–88.
[GT18] A. N. Gorban and I. Y. Tyukin, *Blessing of dimensionality: mathematical foundations of the statistical physics of data*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376** (2018), no. 2118, 20170237.
[GGT21] B. Grechuk, A. N. Gorban, and I. Y. Tyukin, *General stochastic separation theorems with optimal bounds*, Neural Networks **138** (2021), 33–56.
[KE21] N. Khan and M. Efthymiou, *The use of biometric technology at airports: The case of customs and border protection (cbp)*, International Journal of Information Management Data Insights **1** (2021), no. 2, 100049.
[KL22] Bo'az Klartag and Joseph Lehec, *Bourgain's slicing problem and kls isoperimetry up to polylog*, arXiv preprint arXiv:2203.15551 (2022).
[MSG+20] S.M. McKinney, M. Sieniek, V. Godbole, et al., *International evaluation of an AI system for breast cancer screening*, Nature **577** (2020), no. 7788, 89–94.
[SBKV+20] H. Seo, M. Badiei Khuzani, V. Vasudevan, C. Huang, and H. Ren et al., *Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications*, Medical Physics **47** (2020), no. 5, e148–e167.
[THG20] I. Y. Tyukin, D. J. Higham, and A. N. Gorban, *On adversarial examples and stealth attacks in artificial intelligence systems*, 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–6.


Alexander Gorban


Bogdan Grechuk


Ivan Tyukin

Credits

Opening image is courtesy of metamorworks via Getty.
Figures 1–6 are courtesy of the authors.
Photo of Alexander Gorban is courtesy of Alexander Gorban.
Photo of Bogdan Grechuk is courtesy of Bogdan Grechuk.
Photo of Ivan Tyukin is courtesy of Ivan Tyukin.