Neural Networks xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neural Networks

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

General stochastic separation theorems with optimal bounds

Bogdan Grechuk [a], Alexander N. Gorban [a,b,*], Ivan Y. Tyukin [a,b,c]

[a] Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK
[b] Lobachevsky University, Nizhni Novgorod, Russia
[c] Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Article history:
Received 11 October 2020
Received in revised form 8 January 2021
Accepted 29 January 2021
Available online xxxx

Keywords:
AI
Blessing of dimensionality
Curse of dimensionality
Concentration of measure
AI errors
Discriminant

ABSTRACT

Phenomenon of stochastic separability was revealed and used in machine learning to correct errors of Artificial Intelligence (AI) systems and analyze AI instabilities. In high-dimensional datasets under broad assumptions each point can be separated from the rest of the set by simple and robust Fisher's discriminant (is *Fisher separable*). Errors or clusters of errors can be separated from the rest of the data. The ability to correct an AI system also opens up the possibility of an attack on it, and the high dimensionality induces vulnerabilities caused by the same stochastic separability that holds the keys to understanding the fundamentals of robustness and adaptivity in high-dimensional data-driven AI. To manage errors and analyze vulnerabilities, the stochastic separation theorems should evaluate the probability that the dataset will be Fisher separable in given dimensionality and for a given class of distributions. Explicit and optimal estimates of these separation probabilities are required, and this problem is solved in the present work. The general stochastic separation theorems with optimal probability estimates are obtained for important classes of distributions: log-concave distribution, their convex combinations and product distributions. The standard i.i.d. assumption was significantly relaxed. These theorems and estimates can be used both for correction of high-dimensional data driven AI systems and for analysis of their vulnerabilities. The third area of application is the emergence of memories in ensembles of neurons, the phenomena of grandmother's cells and sparse coding in the brain, and explanation of unexpected effectiveness of small neural ensembles in high-dimensional brain.

1. Introduction: Data mining in post-classical world

Big data 'revolution' and the growth of the data dimension are commonplace. However, some implications of this growth are not so well known. In his 'millennium lecture', Donoho (2000) sought to present major 21st century challenges for data analysis. He described the multidimensional post-classical world where the number of attributes d (dimensionality of the dataspace) exceeds the sample size N:

$$d \gg N. \tag{1}$$

Of course, there are many practical tricks for handling data when the condition (1) holds. In such a situation, tools of the first choice are Principal Component Analysis with retaining of major components, the correlation transformation, that transforms the data set into its Gram matrix (the matrix of inner products or correlation coefficients between the data vectors), or their combination (for a case study see Moczko et al., 2016). These methods return the situation from (1) to $d \leq N$ but this is not the end of the story. For the non-classical effects, the inequality (1) is not necessary. Many such effects arise when

$$d \gg \log N. \tag{2}$$

Various examples of these effects are presented by Donoho and Tanner (2009), Gorban, Tyukin, Prokhorov et al. (2016), Kainen (1997) and Kainen and Kůrková (1993). High-dimensional data are very rarefied and have large data-free holes, even if the data sets are exponentially large (Kainen, 1997). Two effects are especially important:

- *Random vectors are quasiorthogonal*: N random vectors $\boldsymbol{x}_i$ ($i = 1, \ldots, N$) on a unit d-dimensional sphere are almost orthogonal: for a given $\varepsilon > 0$ and sufficiently large d, we can expect with high probability that $|(\boldsymbol{x}_i, \boldsymbol{x}_j)| < \varepsilon$ fo all $i \neq j$ when $d > A \ln N$ and A does not depend on N. (For various versions of exact formulations we refer to Gorban, Tyukin, Prokhorov et al. (2016) and Kainen and Kůrková (1993). Very recently, Kainen and Kůrková (2020) reviewed the concept of quasiorthogonal dimension and related notions.)

* Corresponding author at: Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK.
E-mail addresses: bg83@le.ac.uk (B. Grechuk), ag153@le.ac.uk (A.N. Gorban), it37@le.ac.uk (I.Y. Tyukin).

https://doi.org/10.1016/j.neunet.2021.01.034

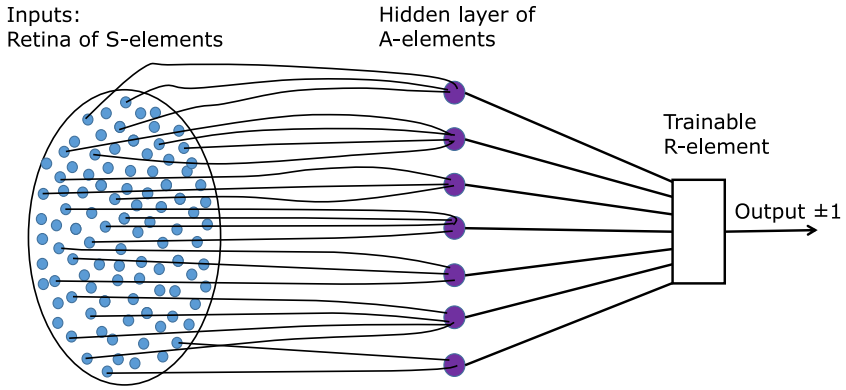

Fig. 1. Rosenblatt's elementary perceptron (Rosenblatt, 1962). *A*-and *R*-element are the classical threshold neurons. *R* element is trainable by the Rosenblatt algorithm, while *A*-elements should represent a sufficient collection of features.

- *Random points are extreme*: for a given $\varepsilon > 0$ with a probability $p > 1 - \varepsilon$ all N random points are vertices of their convex hull. It is sufficient that $d > B \ln N$ for some B that depends on ε only (for exact formulations we refer to Bárány & Füredi, 1988; Donoho & Tanner, 2009).

Of course, detailing these properties should include clarifying what 'random' means. A fairly regular distribution is usually assumed, while Donoho (2000) claimed that such 'blessing of dimensionality' effects hold for an unexpectedly wide class of distributions.

One more comment to (1), (2) is necessary: existence of many attributes does not mean large dimensionality of data. The naïve definition that dimensionality of data refers to how many attributes a dataset has leads to some confusions. Indeed, in the simplest example, when data are distributed along a straight line, data are one-dimensional despite large number of attributes. To distinguish between the number of attributes and the dimensionality of a dataset, the latter is often referred to as the "intrinsic dimensionality" of the data. Not the number of attributes but the dimensionality of data should be used in the definition of the post-classical world:

$$\dim(Dataset) \gg \log N. \tag{3}$$

Evaluation of the (intrinsic) dimensionality of data is a non-trivial problem discussed by many authors, and many approaches are used, ranged from classical Principal Component Analysis (PCA) (Jolliffe, 1993) and their generalizations (Gorban et al., 2008), to principal graphs and manifolds (Gorban & Zinovyev, 2010), and fractal dimension (Camastra, 2003). In recent review by Bac and Zinovyev (2020) the typology of these methods is proposed and a new family of methods based on the data separability properties is presented.

In the post-classical world, classical machine learning theory does not make much sense because it works near the limits of large N, when the law of large numbers and the central limit theorem can be used. The unlimited appetite of classical approaches for data is often considered as a 'curse of dimensionality'. But the properties (1), (2), or (3) themselves are neither a curse, nor a blessing, and can be beneficial. The idea of a 'blessing of dimensionality' was formulated by Kainen (1997), but some properties of the situations with (1) were exploited much earlier. In general situation, if the number of attributes $d \geq N - 1$, then any subsample is linearly separable from the rest of data. Therefore, Rosenblatt (1962, Theorem 1) used a non-linear extension of the set of attributes (*A*-elements, Fig. 1) to prove the omnipotence of elementary peceptrons in solving any classification problem (on a large training set, at least).

Other examples of post-classical phenomena are exponentially large sets of quasiorthogonal (almost orthogonal) random vectors we have already mentioned and stochastic separation in exponentially large datasets: with high probability, any sample point is linearly separable from other points and this separation could be performed by the simple and explicit Fisher discriminant (Gorban et al., 2018; Gorban & Tyukin, 2017; Gorban, Tyukin and Romanenko, 2016). This is a strengthening of the statements (Bárány & Füredi, 1988; Donoho, 2000; Donoho & Tanner, 2009) that random points are extreme ones. These properties were proven for sufficiently regular probability distribution or for products of large number of low-dimensional distributions. For other examples we refer to the book by Vershynin (2018).

The new characterization of post-classical data (3) captures one of the qualitative characterization of the post-classical world. Fundamental open questions, however, are:

1. Are there quantitatively accurate estimates of the boundary between the "classical" and the "post-classical" cases?
2. How these boundaries depend on statistical properties of the data?
3. If the "post-classical" limit always obeys $\log(N) \ll \dim(Dataset)$ or could have different forms such as $\log(N) \ll \dim(Dataset)^p$?

Answering these would allow us to determine applicability bounds for a host of relevant measure concentration-based algorithms in machine learning, including one-shot error correction and learning, randomized approximation, and prevention of vulnerabilities to attacks.

The present work aims to answer these questions. In Section 2 we introduce the stochastic separation phenomenon in detail and prove Theorem 1 that is a prototype of most stochastic separation theorems. Estimates given in this theorem can be improved for specific classes of distributions but it does not use the i.i.d. assumption at al. This major departure from the classical i.i.d. assumption in machine learning enables and justifies one-shot learning and AI correction algorithms in presence of concept drifts, sample dependencies, and non-stationarity.

Further in this work, we present such estimates for many practically important classes of probability distributions, in particular, for log-concave distributions and their convex combinations. In contrast to Theorem 1 and Corollary 1 of Section 2, these estimates are in many cases asymptotically sharp.

In Section 3 the previously known results are analyzed, including estimations for uniform distributions in a ball and a cube. In Section 4 we prove the stochastic separation theorems with estimates of separation probability and sample sizes for strongly log-concave distributions using the logarithmic Sobolev inequality and Poincare inequality. For special classes of distributions

stronger results are obtained, for example, for rotation invariant log-concave distributions including multivariate exponential distribution (Section 5). The known estimates for some distributions like uniform distribution in a ball and the standard normal distribution are significantly improved and optimal separation theorem for explicitly given distributions are found. Section 6 derives separation theorems for independent data from product distributions, while Section 7 generalizes some of these theorems to the case of dependent data relaxing the i.i.d. assumption. Section 8 briefly summarized the results, and in Section 9 we discuss what these estimates are for and present the main areas of applications.

2. Stochastic separation phenomenon

The 'post-classical' phenomenon of separability of random points from random sets in high dimensionality opens up the possibility for fast and non-iterative correction of errors of data-driven Artificial Intelligence (AI). Each situation of AI functioning is represented by a vector that combines inputs, internal signals and outputs of the AI system. If a situation with error can be separated by an explicit and simple functional (Fisher's discriminant, for example) from the known situations with correct functioning then this error can be corrected forever without destroying the existing skills (Gorban & Tyukin, 2018; Gorban, Tyukin and Romanenko, 2016). The corrector is a combination of the two-class classifier of situations ('AI error' versus 'correct functioning') with a modified decision rule for the 'error' class.

Below in this section, a prototype of most stochastic separation theorems is introduced.

Recall that the classical Fisher discriminant between two classes with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is separation of the classes by a hyperplane orthogonal to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ in the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = (\boldsymbol{x}, \boldsymbol{S}^{-1}\boldsymbol{y}),$$

where $(\cdot, \cdot)$ is the standard inner product and $\boldsymbol{S}$ is the average (or the weighted average) of the sample covariance matrix of these two classes. The classification rule is: if $\langle \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \boldsymbol{x} \rangle \geq \vartheta$ then $\boldsymbol{x}$ belongs to the first class, otherwise it belongs to the second class. The threshold ϑ should be chosen in such a way as to maximize the quality of classification evaluated by a preselected criterion.

Applications of stochastic separation theory consider separating a single point (error) or a small cluster of such points from a relatively large data set. Thus, $\boldsymbol{S}$ is by default the empiric covariance matrix of a large data set. Further on, assume that the dataset is preprocessed, this includes centralization (zero mean) and whitening. Whitening uses PCA to remove minor components and transform coordinates, making the empirical covariance matrix the identity matrix. After whitening, we get out of the situation described by the condition (1) but the conditions (2) or (3) can persist.

It is necessary to stress that the precise whitening in applications to high-dimensional datasets could be unavailable, and $\boldsymbol{S}$ may differ from $\mathbf{1}$. If $\boldsymbol{S}$ remains a well-conditioned matrix then this difference does not change qualitatively the separability properties. Analysis of the quantitative differences that may appear for non-isotropic $\boldsymbol{S}$ for some classes of probability distributions is presented in Section 4.2.

Presuming the described preprocessing with whitening, we take $\boldsymbol{S} = \mathbf{1}$ and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = (\boldsymbol{x}, \boldsymbol{y})$.

Definition 1. A point $\boldsymbol{x}$ is *Fisher separable* from a set $Y \subset \mathbb{R}^n$ with center $\boldsymbol{c} \in \mathbb{R}^n$ and threshold $\alpha \in (0, 1]$ if inequality

$$\alpha(\boldsymbol{x} - \boldsymbol{c}, \boldsymbol{x} - \boldsymbol{c}) > (\boldsymbol{x} - \boldsymbol{c}, \boldsymbol{y} - \boldsymbol{c}), \tag{4}$$

holds for all $\boldsymbol{y} \in Y$. If (4) does not hold for some $\boldsymbol{x}$ and $\boldsymbol{y}$, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ form an (ordered) $(\alpha, \boldsymbol{c})$-*inseparable pair* (see Fig. 2).

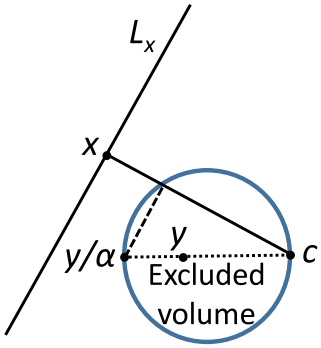

Fig. 2. Geometry of separation: $\alpha(\boldsymbol{x}, \boldsymbol{x}) > (\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{x}$ outside the outlined ball ('excluded volume') with diameter $\|y\|/\alpha$. Here, c is the origin (the data mean), L_x is the hyperplane orthogonal to x. If $\boldsymbol{x}$ belongs to the ball of excluded volume then $\boldsymbol{x}$ and $\boldsymbol{y}$ forms an ordered α-inseparable pair.

If $\boldsymbol{c} = \mathbf{0}$ is the origin, we will write $(\alpha, \mathbf{0})$-inseparable pair as just "α-inseparable pair", to simplify the notation. For a given $\boldsymbol{y}$, the set of such $\boldsymbol{x}$ that $\boldsymbol{x}, \boldsymbol{y}$ form an ordered α-inseparable pair is a ball given by inequality

$$\left\{ \boldsymbol{z} \;\middle|\; \left\| \boldsymbol{z} - \frac{\boldsymbol{y}}{2\alpha} \right\| \leq \frac{\|\boldsymbol{y}\|}{2\alpha} \right\}. \tag{5}$$

This is the ball of excluded volume from Fig. 2.

Two heuristic condition for the probability distribution are used in the stochastic separation theorems:

- The probability distribution has no heavy tails;
- The sets of small volume should not have large probability (what "small" and "large" mean should be strictly defined for different contexts).

In the following Theorem 1 the absence of heavy tails is formalized as the tail cut: the support of the distribution is the n-dimensional unit ball $\mathbb{B}_n$.

The absence of the sets of small volume but large probability is formalized in this theorem by the following inequality:

$$\rho(\boldsymbol{x}) < \frac{C}{r^n V_n(\mathbb{B}_n)}, \tag{6}$$

where ρ is the distribution density, $C > 0$ is an arbitrary constant, $V_n(\mathbb{B}_n)$ is the volume of the ball $\mathbb{B}_n$, and $1 > r > 1/(2\alpha)$. This inequality guarantees that the probability measure of each ball with the radius less than or equal to $1/(2\alpha)$ exponentially decays for $n \to \infty$. It should be stressed that the constant $C > 0$ is arbitrary but must not depend on n in asymptotic analysis for large n. Condition $1 > r > 1/(2\alpha)$ is possible only if $\alpha > 0.5$. Thus, the interval of possible α for Theorem 1 is $\alpha \in (0.5, 1]$.

Theorem 1 (*Gorban et al., 2018*). *Let $1 \geq \alpha > 1/2$, $1 > r > 1/(2\alpha)$, $1 > \delta > 0$, $Y \subset \mathbb{B}_n$ be a finite set, $|Y| < \delta(2r\alpha)^n/C$, and $\boldsymbol{x}$ be a randomly chosen point from a distribution in the unit ball with the bounded probability density $\rho(\boldsymbol{x})$. Assume that $\rho(\boldsymbol{x})$ satisfies inequality (6). Then with probability $p > 1 - \delta$ point $\boldsymbol{x}$ is Fisher-separable from Y with threshold α (4).*

Proof. The volume of the ball (5) does not exceed $V = \left(\frac{1}{2\alpha}\right)^n V_n(\mathbb{B}_n)$ for each $\boldsymbol{y} \in Y$. The probability that point $\boldsymbol{x}$ belongs to such a ball does not exceed

$$V \sup_{z \in \mathbb{B}_n} \rho(z) \leq C \left(\frac{1}{2r\alpha}\right)^n.$$

The probability that $\boldsymbol{x}$ belongs to the union of $|Y|$ such balls does not exceed $|Y|C\left(\frac{1}{2r\alpha}\right)^n$. For $|Y| < \delta(2r\alpha)^n/C$ this probability is smaller than δ and $p > 1 - \delta$. □

Table 1
The upper bound on $|Y|$ that guarantees separation of $\boldsymbol{x}$ from Y by Fisher's discriminant with probability 0.99 according to Theorem 1 for $\alpha = 0.8$, $r = 0.75$, $C = 1$ in various dimensions.

n	$\rho(\boldsymbol{x})/\rho_{\text{uniform}} \leq$	$\lvert Y\rvert \leq$
10	17.7	0.06
50	$1.7 \cdot 10^{6}$	91
100	$3.1 \cdot 10^{12}$	828,180
200	$9.7 \cdot 10^{24}$	$6.8 \cdot 10^{13}$
500	$2.9 \cdot 10^{62}$	$3.9 \cdot 10^{37}$
1000	$8.6 \cdot 10^{124}$	$1.5 \cdot 10^{77}$

Table 2
The upper bound on $|Y|$ that guarantees α-Fisher separability of Y with probability 0.99, according to Corollary 1 for $\alpha = 0.8$, $r = 0.75$, $C = 1$ in various dimensions.

n	$\rho(\boldsymbol{x})/\rho_{\text{uniform}} \leq$	$\lvert Y\rvert \leq$
10	17.7	0.25
50	$1.7 \cdot 10^{6}$	9.54
100	$3.1 \cdot 10^{12}$	910
200	$9.7 \cdot 10^{24}$	$8.2 \cdot 10^{6}$
500	$2.9 \cdot 10^{62}$	$6.2 \cdot 10^{18}$
1000	$8.6 \cdot 10^{124}$	$3.9 \cdot 10^{38}$

Remark 1. Note that:

- The finite set Y in Theorem 1 is just a finite subset of the ball $\mathbb{B}_n$ without any assumption of its randomness. We only used the assumption about distribution of $\boldsymbol{x}$.
- The distribution of $\boldsymbol{x}$ may deviate significantly from the uniform distribution in the ball $\mathbb{B}_n$. Moreover, this deviation may grow with dimension n as a geometric progression:

  $$\rho(\boldsymbol{x})/\rho_{\text{uniform}} \leq C/r^n,$$

  where $\rho_{\text{uniform}} = 1/V_n(\mathbb{B}_n)$ is the density of uniform distribution and $1/(2\alpha) < r < 1$ (assuming that $1/2 < \alpha \leq 1$).

Example 1. Let $\alpha = 0.8$, $r = 0.75$, $C = 1$, $\delta = 0.01$. Table 1 shows the upper bounds on $|Y|$ given by Theorem 1 in various dimensions n if the ratio $\rho(\boldsymbol{x})/\rho_{\text{uniform}}$ is bounded by the geometric progression $1/r^n$.

For example, for $n = 100$, we see that for any set with $|Y| < 828{,}180$ points in the unit ball, and any distribution whose density ρ deviates from the uniform one by a factor at most $3.1 \cdot 10^{12}$, a random point from this distribution is Fisher-separable from all points in Y with 99% probability.

In the following Definition we consider separation of each points of a set from all other points by Fisher discriminant.

Definition 2. A finite set $Y \subset \mathbb{R}^n$ is *Fisher separable* with center $\boldsymbol{c} \in \mathbb{R}^n$ and threshold $\alpha \in (0, 1]$, or $(\alpha, \boldsymbol{c})$-Fisher separable in short, if inequality

$$\alpha(\boldsymbol{x} - \boldsymbol{c}, \boldsymbol{x} - \boldsymbol{c}) > (\boldsymbol{x} - \boldsymbol{c}, \boldsymbol{y} - \boldsymbol{c}),$$

holds for all $\boldsymbol{x}, \boldsymbol{y} \in Y$ such that $\boldsymbol{x} \neq \boldsymbol{y}$.

If $\boldsymbol{c} = \boldsymbol{0}$ is the origin, we will write $(\alpha, \boldsymbol{0})$-Fisher separable set as just "α-Fisher separable", to simplify the notation. From Theorem 1 we obtain the following corollary.

Corollary 1. *If $Y \subset \mathbb{B}_n$ is a random set $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{|Y|}\}$ and for each j the conditional distributions of vector $\boldsymbol{y}_j$ for any given positions of the other $\boldsymbol{y}_k$ in $\mathbb{B}_n$ satisfy the same conditions as the distribution of $\boldsymbol{x}$ in Theorem 1, then the probability of the random set Y to be α-Fisher separable can be easily estimated:*

$$p \geq 1 - |Y|^2 C \left(\frac{1}{2r\alpha}\right)^n.$$

So, $p > 0.99$ if $|Y| < (1/10)\, C^{-1/2}(2r\alpha)^{n/2}$.

For this estimate, elements of Y should not be i.i.d. random vectors and each of them can have its own distribution but with the same restrictions (with support in a ball and inequality (6)). The flight from i.i.d assumption in machine learning is recognized as an important problem (Kůrková, 2019). The measure concentration phenomena can provide an instrument for avoiding this assumption (Gorban et al., 2018; Kůrková & Sanguineti, 2019).

Example 2. Let $\alpha = 0.8$, $r = 0.75$, $C = 1$. Table 2 shows the upper bounds on $|Y|$ given by Corollary 1 in various dimensions n if the ratio $\rho(\boldsymbol{x})/\rho_{\text{uniform}}$ is bounded by the geometric progression $1/r^n$.

For example, for $n = 100$, we see that for any distribution whose density ρ deviates from the uniform one by a factor at most $3.1 \cdot 10^{12}$, any set with $|Y| < 910$ points from this distribution is Fisher-separable with 99% probability. In dimension $n = 200$, we may deviate from the uniform density by a factor $9.7 \cdot 10^{24}$ and still separate over 8 millions points.

In the post-classical world correction of AI errors is possible by separation of situations with errors from the situations of correct functioning. This can be done because the intrinsically high-dimensional data are very 'rarefied'. At the same time, the possibility of repairing AI is closely related to the possibility of its attack. The specific post-classical vulnerabilities and new types of attacks were identified recently (Tyukin et al., 2020). The exact line between the classic world of 'condensed' data and the post-classic world of rarefied data is important for both analyzing AI fixes and fixing AI vulnerability to attacks.

Theorem 1 and Corollary 1 ensure us that if the probability distributions have no heavy tails and sets of relatively small volume cannot have high probability, then the exponentially large sets are Fisher separable. Nevertheless, the presented estimates are far from being optimal, and sharp estimations of probabilities and sample sizes are very desirable.

The formalization of the idea 'no heavy tails' does not require a bounded distribution support. Below, we show that the exponential asymptotics at infinity will be fast enough to constructively describe the phenomenon of stochastic separation. For this purpose, we use the class *log-concave distributions*, but already the first estimate shows that the asymptotics of $|Y|$ guaranteeing Fisher separability for such a general class of distributions are nonexponential: the boundary $|Y|$ that guarantees separability with a fixed probability, grows with dimension n as $a\exp(b\sqrt{n})$ (Theorem 3). It is demonstrated that this estimate cannot be significantly improved (Example 4). Exponential asymptotic is proved for a narrower class, *strongly log-concave distributions* (Section 4). The most prominent member of this family is the normal distribution. The separability properties for the normal distribution are studied in detail in Section 5.1.

The general stochastic separation theorems are proven for convex combinations of strongly log-concave distributions (Theorem 11). The conditions of Theorem 11 formalize both no heavy tails condition (through strong log-concavity) and no small sets with high probability condition. In some sense, the generality of this theorem is sufficient for most of practical purposes, but in specific cases, for narrower classes and selected distributions the estimates can be much better than for a wide general class. Therefore, we explore additional classes like product distributions (data with independent attributes) (Section 6), rotation invariant distributions (Section 5) and some special examples: uniform distributions in a ball or in a cube and normal distribution. For data

Table 3
The upper bounds on M that guarantees Fisher separability of M i.i.d. points from uniform distribution in an n-dimensional ball with probability $p > 0.99$ for $\alpha = 0.6, 0.8$ and 1, according to Theorem 2.

	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
$n = 10$	0.35	1.48	4.52
$n = 50$	13.5	17,927	$4.7 \cdot 10^6$
$n = 100$	1287	$2.2 \cdot 10^9$	$1.6 \cdot 10^{14}$
$n = 200$	$1.1 \cdot 10^7$	$3.6 \cdot 10^{19}$	$1.8 \cdot 10^{29}$
$n = 500$	$8.8 \cdot 10^{18}$	$1.5 \cdot 10^{50}$	$2.5 \cdot 10^{74}$
$n = 1000$	$5.5 \cdot 10^{38}$	$1.6 \cdot 10^{101}$	$4.6 \cdot 10^{149}$

with independent attributes, the dependent samples are studied (Section 7). A short guide on proven theorems is presented in Section 8, and in Section 9 we briefly discuss the application of the stochastic separation theorems in machine learning and neuroscience.

3. Analysis of known stochastic separation theorems

Let us focus on Fisher separability because Fisher discriminants are robust and can be created by simple, explicit and one-shot rule. The results of Bárány and Füredi (1988) and Donoho and Tanner (2009) about linear separability remain beyond the scope of this analysis.

Gorban and Tyukin (2017) proved that if M points are selected independently uniformly at random in the unit ball in $\mathbb{R}^n$, then they are 1-Fisher separable with high probability, provided that M is bounded by some exponential function of n. A simple version of this result was later proved[1] in Gorban et al. (2018).

Theorem 2 (*Gorban et al., 2018*)**.** *Let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ be i.i.d points from uniform distribution in a ball. For any $\delta > 0$, if*

$$M < \sqrt{2\delta}(2\alpha)^{n/2} = \sqrt{2\delta}\exp\left(\frac{1}{2}\log(2\alpha)n\right), \tag{7}$$

then set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is α-Fisher separable with probability greater than $1 - \delta$.

The estimate (7) grows exponentially fast in n provided that $\alpha > 1/2$.

Example 3. Let $\delta = 0.01$. Table 3 shows the upper bounds on M in Theorem 2 for $\alpha = 0.6, 0.8$ and 1 in various dimensions n.

For example, for $n = 100$, we see that over 2 billions points from the uniform distribution in the unit ball are Fisher-separable at level $\alpha = 0.8$ with probability greater than 99%.

Of course, uniform distribution in a ball is a very special case, and separation theorems have been proved for various other families of distributions. We say that density $\rho : \mathbb{R}^n \to [0, \infty)$ of random vector $\boldsymbol{x}$ (and the corresponding probability distribution) is *log-concave*, if set $D = \{z \in \mathbb{R}^n \mid \rho(z) > 0\}$ is convex and $g(z) = -\log(\rho(z))$ is a convex function on D. We say that ρ is whitened, or *isotropic*, if $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$, and

$$\mathbb{E}[(\boldsymbol{x}, \theta)^2] = 1 \qquad \forall \theta \in \mathbb{S}^{n-1}, \tag{8}$$

where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. Eq. (8) is equivalent to the statement that the variance–covariance matrix for the components of $\boldsymbol{x}$ is the identity matrix. This can be achieved by linear transformation of the data during the pre-processing step, therefore this assumption is not restrictive.

[1] The proof in Gorban et al. (2018) is presented for $\alpha = 1$, but the argument works for general α.

Theorem 3 (*Gorban et al., 2018, Corollary 2*)**.** *Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from an isotropic log-concave distribution in $\mathbb{R}^n$. Then set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is 1-Fisher separable with probability greater than $1 - \delta$, $\delta > 0$, provided that*

$$M \le ae^{b\sqrt{n}}, \tag{9}$$

where $a > 0$ and $b > 0$ are constants, depending only on δ.

The following Example demonstrates that $\sqrt{n}$ in (9) cannot be replaced by $n^{0.5+\epsilon}$ for any $\epsilon > 0$, even if points are selected from a product distribution with identical log-concave components.

Example 4. Let $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$ denote the l_1 norm of $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. Let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ be i.i.d points from (isotropic log-concave) distribution in $\mathbb{R}^n$ with density

$$\rho(\boldsymbol{x}) = 2^{-n/2}e^{-\sqrt{2}\cdot\|\boldsymbol{x}\|_1}.$$

For any $\alpha \in (0, 1]$, $a > 0$, $b > 0$, and $\epsilon > 0$, if

$$M \ge a\exp\left(b \cdot n^{0.5+\epsilon}\right), \tag{10}$$

then set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is *not* α-Fisher separable with probability tending to 1 as $n \to \infty$.

Detail. The probability that any two i.i.d. points $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ from the given distribution are *not* α-Fisher separable is bounded by

$$\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \le (\boldsymbol{x}, \boldsymbol{y})] \ge \mathbb{P}[(\boldsymbol{x}, \boldsymbol{x}) \le (\boldsymbol{x}, \boldsymbol{y})] = \mathbb{P}\left[\sum_{i=1}^n (x_i y_i - x_i^2) \ge 0\right] = \mathbb{P}\left[\sum_{i=1}^n z_i \ge n\right], \tag{62}$$

where $z_i = x_i y_i - x_i^2 + 1$, $i = 1, \dots, n$ are i.i.d. random variables with zero mean. Next,

$$\mathbb{P}\left[\sum_{i=1}^n z_i \ge n\right] \ge \mathbb{P}[z_1 \ge n] \cdot \mathbb{P}\left[\sum_{i=2}^n z_i \ge 0\right] = \mathbb{P}[z_1 \ge n]\left(\frac{1}{2} + o(1)\right), \tag{65}$$

where the last equality follows from central limit theorem, and $o(1)$ is the quantity which goes to 0 as $n \to \infty$. Further,

$$\frac{1}{2}\mathbb{P}[z_1 \ge n] \ge \frac{1}{2}\mathbb{P}\left[\sqrt{n} \le x_1 \le 2\sqrt{n}\right] \times \mathbb{P}\left[3\sqrt{n} \le y_1\right] = \frac{1}{8}e^{-4\sqrt{2}\sqrt{n}}(1 + o(1)). \tag{68}$$

Because M points can be divided into $M/2$ independent pairs, the probability that all these pairs are α-Fisher separable is at most

$$\left(1 - \frac{1}{8}e^{-4\sqrt{2}\sqrt{n}}(1 + o(1))\right)^{M/2}, \tag{71}$$

and the last expression vanishes as $n \to \infty$ if (10) holds. □

Example 4 demonstrates that, to recover exponential dependence of M from n, one must consider subclasses of log-concave distributions.

We say that density $\rho : \mathbb{R}^n \to [0, \infty)$ is strongly log-concave with constant $\gamma > 0$, or γ-SLC in short, if $g(z) = -\log(\rho(z))$ is strongly convex, that is, $g(z) - \frac{\gamma}{2}\|z\|$ is a convex function on D.

Theorem 4 (*Gorban et al., 2018, Corollary 4*)**.** *Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from an isotropic γ-SLC distribution in $\mathbb{R}^n$. Then set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is 1-Fisher separable with probability greater than $1 - \delta$, $\delta > 0$, provided that*

$$M \le ae^{bn},$$

where $a > 0$ and $b > 0$ are some constants, which depends on δ and γ.

Table 4
The lower bound for probability that 100,000 points in the "randomly perturbed data" (Theorem 5) are Fisher separable for various dimension n and noise bound ϵ.

	$\epsilon = 1/10$	$\epsilon = 1/5$	$\epsilon = 1/2$
$n = 500$	<0	<0	<0
$n = 1000$	<0	<0	0.9998
$n = 2000$	<0	<0	$1 - 5.8 \cdot 10^{-18}$
$n = 5000$	<0	0.95	$1 - 1.2 \cdot 10^{-57}$
$n = 10000$	<0	$1 - 5 \cdot 10^{-13}$	$1 - 1.3 \cdot 10^{-123}$
$n = 20000$	0.96	$1 - 8 \cdot 10^{-35}$	$1 - 2.2 \cdot 10^{-255}$

Separation theorems have also been proved for various families of distributions which are not log-concave. As an example, consider "randomly perturbed data" model (Example 2 in Gorban et al. (2018)). For a fixed $\epsilon \in (0, 1)$, let $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_M$ be the set of M arbitrary (non-random) points inside the ball with radius $1 - \epsilon$ in $\mathbb{R}^n$. Let $\boldsymbol{x}_i$, $i = 1, 2, \dots, M$ be a point, selected uniformly at random from a ball with center $\boldsymbol{y}_i$ and radius ϵ. We think about $\boldsymbol{x}_i$ as "perturbed" version of $\boldsymbol{y}_i$. In this model, we say that set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is α-*Fisher separable* if

$$\alpha(\boldsymbol{x}_i - \boldsymbol{y}_i, \boldsymbol{x}_i - \boldsymbol{y}_i) > (\boldsymbol{x}_i - \boldsymbol{y}_i, \boldsymbol{x}_j - \boldsymbol{y}_i),$$

holds for all $i = 1, 2, \dots, M$ and $j = 1, 2, \dots, M$ such that $i \neq j$.

Theorem 5 (*Gorban et al., 2018, Theorem 7*)**.** *Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M random points in the "randomly perturbed" model with parameter $\epsilon > 0$. For any ϑ such that $\frac{1}{\sqrt{n}} < \vartheta < 1$, set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is 1-Fisher separable with probability at least*

$$1 - \frac{2M^2}{\vartheta\sqrt{n}}\left(\sqrt{1-\vartheta^2}\right)^{n+1} - M\left(\frac{2\vartheta}{\epsilon}\right)^n. \tag{11}$$

In particular, set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is 1-Fisher separable with probability at least $1-\delta$, $\delta > 0$, provided that $M < ab^n$, where a, b are constants depending only on δ and ϵ.

In Theorem 5 we can, for any fixed ϵ, n and M, select $\vartheta \in \left(n^{-1/2}, 1\right)$ to maximize the lower bound (11) for the probability. The optimal ϑ can be easily found numerically.

Example 5. Let $\vartheta = \vartheta(\epsilon, n, M)$ be such that (11) is maximized. Table 4 shows the lower bound for the probability that $M = 100{,}000$ points in the "randomly perturbed data" model are 1-Fisher separable, for various values of n and ϵ.

We see that Theorem 5 is starting to give meaningful results only if the dimension n is rather large, and the smaller ϵ the large dimension we need. This is not much surprising taking into account that the bounds in Table 4 are valid for an arbitrary set of M points in the n-dimensional ball with radius $1 - \epsilon$ and the perturbations make this random finite set closer to the i.i.d. sample from the uniform distribution in the unit ball. In the limit $\epsilon \to 1$ this randomly perturbed set turns into such an i.i.d. sample.

Our final example concerns i.i.d. random points from a product distribution in a unit cube $U_n = [0, 1]^n$.

Theorem 6 (*Gorban & Tyukin, 2017, Corollary 2*)**.** *Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from a product distribution in a unit cube. Let $\boldsymbol{c} \in U_n$ be an arbitrary (non-random) point. Then set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is $(1, \boldsymbol{c})$-Fisher separable with probability greater than $1 - \delta$, $\delta > 0$, provided that*

$$(M+1)^2 < \frac{\delta}{3}\exp(0.5n\sigma_0^4), \tag{12}$$

where σ_0 is the minimal standard deviation of a component distribution.

Theorems 2–6 are proved in works by Gorban et al. (2018) and Gorban and Tyukin (2017) based on the following general principle, which, however, was not formulated explicitly. The high-dimensional stochastic separation theorems are formulated for the classes of distributions in $\mathbb{R}^n$ for all sufficiently large n. For these classes, the probability that two random points are $(\alpha, \boldsymbol{c})$-Fisher inseparable is estimated from above by some function $f(n, \alpha)$. After that, further estimates of M and probabilities of separability of sets are constructed from this function, $f(n, \alpha)$. Let us formulate this principle explicitly.

Theorem 7. *Let $\mathcal{F}$ be a family of M-point distributions in $\mathbb{R}^n$, $F \subset \mathbb{R}^n$ be a random M-point set chosen according to some distribution in $\mathcal{F}$, $\boldsymbol{c} \in \mathbb{R}^n$, $\delta \in (0, 1)$, and $I \subset (0, 1]$. Assume that there exists a function $f(n, \alpha)$ such that for any two points $\boldsymbol{x} \in F$ and $\boldsymbol{y} \in F$*

$$\mathbb{P}[\alpha(\boldsymbol{x}-\boldsymbol{c}, \boldsymbol{x}-\boldsymbol{c}) \leq (\boldsymbol{x}-\boldsymbol{c}, \boldsymbol{y}-\boldsymbol{c})] \leq f(n, \alpha), \quad \alpha \in I,\ n = 1, 2, \dots \tag{13}$$

and

$$M < \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\delta}{f(n, \alpha)}}. \tag{14}$$

Then, for all n and $\alpha \in I$, the expected number of $(\alpha, \boldsymbol{c})$-inseparable pairs in F is less than δ. In particular, set F is $(\alpha, \boldsymbol{c})$-Fisher separable with probability greater than $1 - \delta$.

Proof. If $I(i, j)$ is the indicator function for the event that pair $\boldsymbol{x}_i, \boldsymbol{x}_j$ is $(\alpha, \boldsymbol{c})$-inseparable. Then the expected number of $(\alpha, \boldsymbol{c})$-inseparable pairs is

$$\mathbb{E}\left[\sum_{i \neq j} I(i, j)\right] = \sum_{i \neq j} \mathbb{E}[I(i, j)] \leq \sum_{i \neq j} f(n, \alpha) = M(M-1)f(n, \alpha) < \delta,$$

where the last inequality follows from (14).

If set F would be $(\alpha, \boldsymbol{c})$-Fisher separable with probability $p \leq 1 - \delta$, then the expected number E of $(\alpha, \boldsymbol{c})$-inseparable pairs would be

$$E \geq p \cdot 0 + (1 - p) \cdot 1 \geq \delta,$$

which is a contradiction. Here, the first inequality follows from the fact that the number of (α, c)-inseparable pairs is integer hence it is at least 1. □

If $\boldsymbol{c} = \boldsymbol{0}$ is the origin, inequality (13) simplifies to

$$\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \leq (\boldsymbol{x}, \boldsymbol{y})] \leq f(n, \alpha), \quad \alpha \in I,\ n = 1, 2, \dots \tag{15}$$

A sufficient condition for (14) is the simpler estimate

$$M \leq \sqrt{\frac{\delta}{f(n, \alpha)}}. \tag{16}$$

We will always use (16) in place of (14) unless we aim for the exact (necessary and sufficient) bound for M. In particular, Theorem 2 follows from Theorem 7 with (16) and inequality

$$\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \leq (\boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{2}(2\alpha)^{-n}, \quad \alpha \in (0, 1],\ n = 1, 2, \dots \tag{17}$$

which holds as equality for $\alpha = 1$, see Gorban et al. (2018). Theorems 3 and 4 are proved in the same way. This implies the following corollary.

Corollary 2. *The conclusion "set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is 1-Fisher separable with probability greater than $1-\delta$" in Theorems 2–6 can be replaced by a stronger conclusion that the expected number of inseparable pairs in this set is less than δ.*

Table 5
The upper bound of M (20) that guarantees Fisher separability of a M-element i.i.d. sample from an isotropic γ-SLC distribution in $\mathbb{R}^n$ for $\gamma = 0.6, 0.8$ and 1 with probability $p > 0.99$ for various dimensions n.

	$\gamma = 0.6$	$\gamma = 0.8$	$\gamma = 1$
$n = 10$	0.12	0.15	0.18
$n = 50$	1.71	5.56	18
$n = 100$	61	692	7974
$n = 200$	92,783	$1.2 \cdot 10^7$	$1.8 \cdot 10^9$
$n = 500$	$4.3 \cdot 10^{14}$	$1.1 \cdot 10^{20}$	$2.7 \cdot 10^{25}$
$n = 1000$	$7 \cdot 10^{30}$	$4.7 \cdot 10^{41}$	$3.2 \cdot 10^{52}$

This stronger conclusion is important for practical purposes because it prevents a scenario when we have many (maybe exponentially many in n) inseparable pairs with probability δ.

The proof of Theorem 7 implies that the bound (14) is in fact *necessary and sufficient* condition in the i.i.d case.

Corollary 3. *Let $\boldsymbol{c} \in \mathbb{R}^n$ be fixed, $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from an arbitrary distribution in $\mathbb{R}^n$. Let $f(n, \alpha) := \mathbb{P}[\alpha(\boldsymbol{x}-\boldsymbol{c}, \boldsymbol{x}-\boldsymbol{c}) \le (\boldsymbol{x}-\boldsymbol{c}, \boldsymbol{y}-\boldsymbol{c})]$, where the probability does not depend on the choice of $\boldsymbol{x} \in F$ and $\boldsymbol{y} \in F$. Then the expected number of α-inseparable pairs in F is less than δ if and only if inequality (14) holds.*

For example, the fact that inequality (17) is an equality for $\alpha = 1$ implies the following optimal separation result.

Corollary 4. *Let $\alpha = 1$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be the set of i.i.d points from uniform distribution in a ball. For any $\delta > 0$, the expected number of 1-inseparable pairs from F is less than δ if and only if*

$$M < \frac{1}{2} + \sqrt{\frac{1}{4} + \delta 2^{n+1}}. \tag{18}$$

In particular, (18) implies that F is 1-Fisher separable with probability greater than $1 - \delta$.

In Section 5.3, we prove a version of Corollary 4 for any $\alpha \in (0, 1]$.

The disadvantage of Theorems 3–5 is that constants a and b in the bounds for M are not explicitly given. In Theorem 6, the upper bound for M is explicit but impractical in the important case if the dimension n is measured in hundreds rather than in thousands.

Example 6. For $\delta = 0.01$ (which corresponds to 99% confidence), $n = 500$, and $\sigma_0 = 0.5$ (the maximal possible standard deviation for distribution with $[0, 1]$ support), (12) holds provided $M < 141.7$.

In practise, however, datasets often have much more than 141 point, but Fisher separability still holds. This motivates the search for stochastic Fisher separability theorems with better bounds.

In this paper we obtain separation theorems for various classes of log-concave and product distributions with *explicit* bounds on M. Moreover, we will aim to provide as good bounds as possible, ideally the optimal ones. In addition to better bounds, we also relax the i.i.d assumption.

In the i.i.d. case, Corollary 3 implies that, if we can calculate the probability in (13) *exactly*, then (14) provides the optimal (necessary and sufficient) bound for M. This exact bound, however, is usually quite complicated, based on some integral expressions, and in such cases we will aim for simpler asymptotically tight bounds. We will write

$f(n) \sim g(n)$

if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. We say that function $g(n)$ is asymptotically tight lower (respectively, upper) bound for $f(n)$ if $f(n) \ge g(n)$ (respectively, $f(n) \le g(n)$) and $f(n) \sim g(n)$. If $f(n, \alpha)$ in (13) is the asymptotically tight upper bound for the probability in question, then (14) and (16) provide asymptotically tight upper bounds for M.

If one can prove (13) with $f(n, \alpha) = ae^{-2bn}$ for some constants a, b depending on α, one get (16) with bound $M \le \sqrt{\frac{\delta}{a}} e^{bn}$. If $f(n, \alpha) = ae^{-2bn}$, then $b = -\frac{\log(f(n,\alpha)/a)}{2n}$. In general, the last expression may depend on n, and we define

$$b(\alpha) = b_f(\alpha) := \lim_{n\to\infty} -\frac{\log(f(n, \alpha))}{2n} = -\frac{1}{2} \lim_{n\to\infty} \log \sqrt[n]{f(n, \alpha)}. \tag{19}$$

Let $\mathcal{G}$ be the set of all functions $f(n, \alpha)$ for which (13) holds. We say that separation Theorem 7 has optimal exponent if $b_f(\alpha) \ge b_g(\alpha)$ for all $g \in \mathcal{G}$. Obviously, if bound in (13) is asymptotically tight, it also has optimal exponent, but not vice versa. For non-optimal separation theorems the exponent $b(\alpha)$ is a good way to measure the "quality" of the theorem. We show that in all our non-optimal theorems the exponents differ from optimal by a factor less than 2.

4. Separation theorems for strongly log-concave distributions

4.1. Separation of i.i.d. data from isotropic strongly log-concave distribution

This Section proves the following explicit versions of Theorem 4.

Theorem 8. *Let $\delta > 0, \alpha \in (0, 1], \gamma > 0$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from an isotropic γ-SLC distribution in $\mathbb{R}^n$. If*

$$M < \sqrt{\frac{\delta}{2}} \exp\left(\frac{\alpha^2(\gamma n - 1)}{4(1+\alpha)^2}\right),$$

then the expected number of α-inseparable pairs in F is less than δ. In particular, set F is α-Fisher separable with probability greater than $1 - \delta$.

Theorem 9. *Let $\delta > 0, \alpha \in (0, 1], \gamma > 0$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from an isotropic γ-SLC distribution in $\mathbb{R}^n$. If $n > \frac{1+2\alpha^2}{\gamma\alpha^2}$ and*

$$M < \sqrt{\delta}\left(\frac{\alpha^2}{(1+\alpha^2)^{3/2}}\sqrt{2\pi(\gamma n - 1)} \exp\left(-\frac{\alpha^2(\gamma n - 1)}{2(1+\alpha^2)}\right) + \exp\left(-\frac{\alpha^2(\gamma n - 1)}{2}\right)\right)^{-1/2}, \tag{20}$$

then the expected number of α-inseparable pairs in F is less than δ. In particular, set F is α-Fisher separable with probability greater than $1 - \delta$.

Theorem 9 provides a less restrictive upper bound for M for large n, while the upper bound in Theorem 8 is substantially simpler.

Example 7. Let $\delta = 0.01$ and $\alpha = 1$. Table 5 shows the upper bounds on M in Theorem 9 for $\gamma = 0.6, 0.8$ and 1 in various dimensions n.

For example, for $n = 200$, we see that 12 millions points from a strictly log-concave distribution with $\gamma = 0.8$ are 1-Fisher-separable with probability greater than 99%.

The exponent $b(\alpha)$ defined in (19) is

$$b(\alpha) = \frac{\alpha^2\gamma}{4(1+\alpha)^2}$$

for Theorem 8, and

$$b(\alpha) = \frac{\alpha^2\gamma}{4(1+\alpha^2)}$$

for Theorem 9. For example, if $\gamma = 1$ (which is the case for normal distribution) and $\alpha = 1$, the exponents are $b = \frac{1}{16} = 0.0625$ and $b = \frac{1}{8} = 0.125$, respectively. The optimal exponent for normal distribution is given in Theorem 12 and is equal to $\frac{1}{4}\log(2) = 0.173..$, hence exponent in Theorem 9 cannot be improved more than by a factor $2\log(2) = 1.386....$

The first step in the proof of Theorems 8 and 9 is the following estimate.

Proposition 1. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two i.i.d. points from an isotropic γ-SLC distribution. Then*

$$\mathbb{P}[\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})] \le \mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}]. \tag{21}$$

Proof. Theorem 5.2 in the book of Ledoux (2001) states that, if random vector $\boldsymbol{z}$ follows a γ-SLC distribution, then logarithmic Sobolev inequality

$$\mathbb{E}[f^2(\boldsymbol{z})\log f^2(\boldsymbol{z})] - \mathbb{E}[f^2(\boldsymbol{z})]\mathbb{E}[\log f^2(\boldsymbol{z})] \le \frac{2}{\gamma}\mathbb{E}[\|\nabla f(\boldsymbol{z})\|^2] \tag{22}$$

holds for every locally Lipschitz function f on $\mathbb{R}^n$. By Ledoux (2001, Theorem 5.3), this implies that inequality

$$\mathbb{P}[g(\boldsymbol{z}) \ge E[g(\boldsymbol{z})] + r] \le e^{-\gamma r^2/2} \tag{23}$$

holds for every $r \ge 0$ and every 1-Lipschitz function g on $\mathbb{R}^n$.

Assuming that $\boldsymbol{x} \ne \boldsymbol{0}$ is fixed, and applying (23) to 1-Lipschitz function $g(\boldsymbol{y}) = \frac{(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}\|}$ with $\mathbb{E}[g(\boldsymbol{y})] = \frac{(\boldsymbol{x},\mathbb{E}[\boldsymbol{y}])}{\|\boldsymbol{x}\|} = \frac{(\boldsymbol{x},\boldsymbol{0})}{\|\boldsymbol{x}\|} = 0$ and $r = \alpha\|\boldsymbol{x}\|$, we obtain

$$\mathbb{P}[\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})] = \mathbb{P}\left[\frac{(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}\|} \ge 0 + \alpha\|\boldsymbol{x}\|\right] \le e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} \tag{24}$$

Now let $\boldsymbol{x}$ and $\boldsymbol{y}$ be both random, and let I be the indicator function of the event $\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})$. Then

$$\mathbb{P}[\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})] = \mathbb{E}[I] = \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{\boldsymbol{y}}[I|\boldsymbol{x}]] \le \mathbb{E}_{\boldsymbol{x}}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}],$$

where the second equality follows from independence of $\boldsymbol{x}$ and $\boldsymbol{y}$, and the inequality follows from (24). □

The next proposition provides an easy estimate for the right-hand side of (21).

Proposition 2. *Let $\boldsymbol{x}$ be points from an isotropic γ-SLC distribution. Then*

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}] \le 2\exp\left(-\frac{\gamma\alpha^2}{2(1+\alpha)^2}\mu^2\right), \tag{25}$$

where $\mu = \mathbb{E}[\|\boldsymbol{x}\|]$.

Proof. For every $t > 0$,

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}] = \mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} \mid \|\boldsymbol{x}\| > t]\mathbb{P}[\|\boldsymbol{x}\| > t] + \mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} \mid \|\boldsymbol{x}\| \le t]\mathbb{P}[\|\boldsymbol{x}\| \le t].$$

Now,

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} \mid \|\boldsymbol{x}\| > t] \le \mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}t^2} \mid \|\boldsymbol{x}\| > t] = e^{-\frac{\gamma\alpha^2}{2}t^2},$$

$$\mathbb{P}[\|\boldsymbol{x}\| > t] \le 1,$$

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} \mid \|\boldsymbol{x}\| \le t] \le \mathbb{E}[1 \mid \|\boldsymbol{x}\| \le t] = 1,$$

and

$$\mathbb{P}[\|\boldsymbol{x}\| \le t] = \mathbb{P}[\mu - \|\boldsymbol{x}\| \ge \mu - t] \le e^{-\gamma(\mu-t)^2/2},$$

where the last inequality is an application of (23) to 1-Lipschitz function $g(\boldsymbol{x}) = \mu - \|\boldsymbol{x}\|$. Hence,

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}] \le e^{-\frac{\gamma\alpha^2}{2}t^2} + e^{-\gamma(\mu-t)^2/2}.$$

Applying the last inequality with $t = \frac{\mu}{\alpha+1}$, we get the result. □

The next Proposition provides an estimate for $\mu = \mathbb{E}[\|\boldsymbol{x}\|]$.

Proposition 3. *Let $\boldsymbol{x}$ be a points from an isotropic γ-SLC distribution and let $\mu = \mathbb{E}[\|\boldsymbol{x}\|]$. Then*

$$\mu^2 \ge n - \frac{1}{\gamma}. \tag{26}$$

Proof. As remarked by Ledoux (2001, p. 92), the logarithmic Sobolev inequality (22) implies Poincare inequality

$$Var[f(\boldsymbol{x})] \le \frac{1}{\gamma}\mathbb{E}[\|\nabla f(\boldsymbol{x})\|^2].$$

Applying it with $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$, we obtain

$$\mathbb{E}[\|\boldsymbol{x}\|^2] - \mu^2 = Var[\|\boldsymbol{x}\|] \le \frac{1}{\gamma}\|\nabla(\|\boldsymbol{x}\|)\| = \frac{1}{\gamma}.$$

Because $\mathbb{E}[\|\boldsymbol{x}\|^2] = n$ for isotropic distributions, this implies (26). □

Proof of Theorem 8. The combination of (21), (25), and (26) implies that (15) holds with

$$f_\gamma(n,\alpha) = 2\exp\left(-\frac{\gamma\alpha^2}{2(1+\alpha)^2}\left(n - \frac{1}{\gamma}\right)\right),$$

and Theorem 8 follows from Theorem 7. □

To prove Theorem 9, we need an improved version of Proposition 2.

Proposition 4. *Let $\boldsymbol{x}$ be a points from an isotropic γ-SLC distribution. Then*

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}] \le \frac{\alpha^2}{(1+\alpha^2)^{3/2}}\sqrt{2\pi\gamma}\,\mu\exp\left(-\frac{\gamma\alpha^2\mu^2}{2(1+\alpha^2)}\right) + \exp\left(-\frac{\gamma\alpha^2\mu^2}{2}\right), \tag{27}$$

where $\mu = \mathbb{E}[\|\boldsymbol{x}\|]$.

Proof. Because $e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}$ takes value between 0 and 1,

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}] = \int_0^1 \mathbb{P}\left[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} > z\right]dz.$$

We have

$$p(z) := \mathbb{P}\left[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2} > z\right] = \mathbb{P}\left[\|\boldsymbol{x}\| < \sqrt{-\frac{2\log z}{\gamma\alpha^2}}\right] = \mathbb{P}\left[\mu - \|\boldsymbol{x}\| > \mu - \sqrt{-\frac{2\log z}{\gamma\alpha^2}}\right]$$

If $z \geq z_0 := e^{-\frac{1}{2}\gamma\alpha^2\mu^2}$, then $r := \mu - \sqrt{-\frac{2\log z}{\gamma\alpha^2}} \geq 0$, and (23) with 1-Lipschitz function $g(\boldsymbol{x}) = \mu - \|\boldsymbol{x}\|$ yields

$$p(z) \leq q(z) := \exp\left(-\frac{\gamma}{2}\left(\mu - \sqrt{-\frac{2\log z}{\gamma\alpha^2}}\right)^2\right).$$

We also have a trivial estimate $p(z) \leq 1$ for $z \leq z_0$, which implies

$$\int_0^1 p(z)dz \leq \int_0^{z_0} 1\,dz + \int_{z_0}^1 q(z)\,dz = z_0 + I,$$

where $I = \int_{z_0}^1 q(z)\,dz$. Integration in Mathematica returns

$$I = \frac{\alpha^2}{2(1+\alpha^2)^{3/2}} \exp\left(-\frac{1}{2}\gamma(2+\alpha^2)\mu^2\right)(S_1 + S_2),$$

where

$$S_1 = -2\sqrt{1+\alpha^2}(e^{\gamma\mu^2} - e^{\frac{1}{2}(1+\alpha^2)\gamma\mu^2}),$$

and

$$S_2 = \exp\left(\frac{(2+2\alpha^2+\alpha^4)\gamma\mu^2}{2(1+\alpha^2)}\right)\sqrt{2\pi\gamma}\mu\left(\phi\left(\sqrt{\frac{\gamma}{2+2\alpha^2}}m\right) + \phi\left(\alpha^2\sqrt{\frac{\gamma}{2+2\alpha^2}}m\right)\right),$$

where $\phi(y) := \frac{1}{\sqrt{\pi}}\int_{-y}^{y} e^{-t^2}dt$. Inequality $\alpha \leq 1$ implies that $S_1 \leq 0$. Using this and the fact that $\phi(y) \leq 1$ for all y, we get an estimate

$$\int_0^1 p(z)dz \leq z_0 + I \leq z_0 + \frac{\alpha^2}{2(1+\alpha^2)^{3/2}} \exp\left(-\frac{1}{2}\gamma(2+\alpha^2)\mu^2\right) \times \exp\left(\frac{(2+2\alpha^2+\alpha^4)\gamma\mu^2}{2(1+\alpha^2)}\right)\sqrt{2\pi\gamma}\mu \cdot 2,$$

which simplifies to (27). □

Proof of Theorem 9. Let us consider the left-hand side of (27) as a function $f(\mu)$ of μ and show that f is a decreasing function. The second term is clearly decreasing, while the first term is decreasing if the derivative of $\mu\exp\left(-\frac{\gamma\alpha^2\mu^2}{2(1+\alpha^2)}\right)$ is negative, which holds if $\mu^2 > \frac{1+\alpha^2}{\gamma\alpha^2}$. The last inequality follows from condition $n > \frac{1+2\alpha^2}{\gamma\alpha^2}$ and Proposition 3.

Because $f(\mu)$ is a decreasing function, and $\mu \geq n - \frac{1}{\gamma}$ by Proposition 3, we have

$$\mathbb{E}[e^{-\frac{\gamma\alpha^2}{2}\|\boldsymbol{x}\|^2}] \leq f\left(n - \frac{1}{\gamma}\right).$$

This together with (21) implies that (15) holds with $f_\gamma(n,\alpha) = \delta(R_n)^{-2}$, where R_n is the right-hand side of (20). Then (20) follows from Theorem 7. □

Remark 2. In fact, the only place when we have used that the underlying distribution is γ-SLC is the assertion that Sobolev inequality (22) holds. Hence, the condition that the distribution is γ-SLC in Theorems 8 and 9 can be relaxed to the condition that the distribution is isotropic, log-concave, and such that (22) holds.

4.2. Some generalizations

This section provides some generalizations of Theorem 8. We first consider the case when the data are independent, but

- the data are not identically distributed, and
- the distributions for the data points are strongly log-concave but not necessarily isotropic.

Theorem 10. *Let $\delta > 0$, $\alpha \in (0,1]$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M independent random points in $\mathbb{R}^n$. Let $\boldsymbol{x}_i$ follow a γ_i-SLC distribution with $\gamma_i > 0$, with expectation $x_i^0 = E[\boldsymbol{x}_i]$ and norm expectation $\mu_i = \mathbb{E}[\|\boldsymbol{x}_i\|]$. Assume that inequality*

$$\|x_j^0\| < \alpha\mu_i \tag{28}$$

holds for every pair $1 \leq i, j \leq M$. If

$$M < \sqrt{\frac{\delta}{2}}\exp\left(\min_{i,j}\left(\frac{\gamma_i\gamma_j(\mu_i\alpha - \|x_j^0\|)^2}{2(\sqrt{\gamma_j}\alpha + \sqrt{\gamma_i})^2}\right)\right), \tag{29}$$

then the expected number of α-inseparable pairs in F is less than δ. In particular, set F is α-Fisher separable with probability greater than $1-\delta$.

Proof. We will use inequality (23), which is valid for every γ-SLC distribution, not necessarily isotropic. Fix some indices i and j. Define

$$t = \frac{\sqrt{\gamma_i}\mu_i + \sqrt{\gamma_j}\|x_j^0\|}{\sqrt{\gamma_j}\alpha + \sqrt{\gamma_i}}. \tag{30}$$

Then

$$\mu_i - t = \frac{\sqrt{\gamma_j}(\alpha\mu_i - \|x_j^0\|)}{\sqrt{\gamma_j}\alpha + \sqrt{\gamma_i}} > 0, \tag{31}$$

where the inequality follows from (28).

We have

$$\mathbb{P}[\alpha(\boldsymbol{x}_i,\boldsymbol{x}_i) \leq (\boldsymbol{x}_i,\boldsymbol{x}_j)] \leq \mathbb{P}[\|\boldsymbol{x}_i\| \leq t] + \mathbb{P}[\alpha(\boldsymbol{x}_i,\boldsymbol{x}_i) \leq (\boldsymbol{x}_i,\boldsymbol{x}_j) \text{ conditional to } \|\boldsymbol{x}_i\| \geq t]. \tag{32}$$

Applying (23) to 1-Lipschitz function $g(\boldsymbol{x}_i) = \mu_i - \|\boldsymbol{x}_i\|$, we obtain

$$\mathbb{P}[\|\boldsymbol{x}_i\| \leq t] = \mathbb{P}[\mu_i - \|\boldsymbol{x}_i\| \geq \mu_i - t] \leq e^{-\gamma_i(\mu_i - t)^2/2}. \tag{33}$$

Let us now estimate the second term in (32). Assume that $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i\| \geq t$ is fixed. Applying (23) to 1-Lipschitz function $g(\boldsymbol{x}_j) = \frac{(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\|\boldsymbol{x}_i\|}$ with $\mathbb{E}[g(\boldsymbol{x}_j)] = \frac{(\boldsymbol{x}_i,x_j^0)}{\|\boldsymbol{x}_i\|}$ and $r = \alpha\|\boldsymbol{x}_i\| - \frac{(\boldsymbol{x}_i,x_j^0)}{\|\boldsymbol{x}_i\|}$, we obtain

$$\mathbb{P}[\alpha(\boldsymbol{x}_i,\boldsymbol{x}_i) \leq (\boldsymbol{x}_i,\boldsymbol{x}_j)] = \mathbb{P}\left[\frac{(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\|\boldsymbol{x}_i\|} \geq \frac{(\boldsymbol{x}_i,x_j^0)}{\|\boldsymbol{x}_i\|} + r\right] \leq e^{-\frac{\gamma_j}{2}r^2}, \tag{66}$$

provided that $r > 0$. In fact,

$$r = \alpha\|\boldsymbol{x}_i\| - \frac{(\boldsymbol{x}_i,x_j^0)}{\|\boldsymbol{x}_i\|} \geq \alpha t - \frac{\|\boldsymbol{x}_i\|\cdot\|x_j^0\|}{\|\boldsymbol{x}_i\|} = \alpha t - \|x_j^0\| = \frac{\sqrt{\gamma_i}}{\sqrt{\gamma_j}}(\mu_i - t), \tag{68}$$

where the last equality follows from (30). This implies that $r > 0$, and also that

$$\mathbb{P}[\alpha(\boldsymbol{x}_i,\boldsymbol{x}_i) \leq (\boldsymbol{x}_i,\boldsymbol{x}_j)] \leq \exp\left(-\frac{\gamma_j}{2}r^2\right) \leq \exp\left(-\frac{\gamma_j}{2}\left(\frac{\sqrt{\gamma_i}}{\sqrt{\gamma_j}}(\mu_i - t)\right)^2\right) = e^{-\gamma_i(\mu_i-t)^2/2}. \tag{71}$$

Because this inequality holds for every fixed $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i\| \geq t$, it implies that

$$\mathbb{P}[\alpha(\boldsymbol{x}_i,\boldsymbol{x}_i) \leq (\boldsymbol{x}_i,\boldsymbol{x}_j) \text{ conditional to } \|\boldsymbol{x}_i\| \geq t] \leq e^{-\gamma_i(\mu_i-t)^2/2}.$$

Combining this with (33) and (32), we obtain

$$\mathbb{P}[\alpha(\boldsymbol{x}_i,\boldsymbol{x}_i) \leq (\boldsymbol{x}_i,\boldsymbol{x}_j)] \leq 2e^{-\gamma_i(\mu_i-t)^2/2} \leq 2\exp\left(-\min_{i,j}\left(\frac{\gamma_i\gamma_j(\mu_i\alpha - \|x_j^0\|)^2}{2(\sqrt{\gamma_j}\alpha + \sqrt{\gamma_i})^2}\right)\right), \tag{34}$$

where we have used (31). The last bound holds for any pair of indices i, j, and application of Theorem 7 finishes the proof. □

Repeating the proof of Proposition 3, we get that

$$\mu_i^2 \geq \mathbb{E}[\|\boldsymbol{x}_i\|^2] - \frac{1}{\gamma_i}.$$

Writing $\boldsymbol{x}_i = (x_i^1, \dots, x_i^n)$ component-wise, we obtain that

$$\mathbb{E}[\|\boldsymbol{x}_i\|^2] = \sum_{k=1}^{n} \mathbb{E}[(x_i^k)^2].$$

Hence, if we assume that

- All γ_i are bounded from below by some constant independent of n, and
- The averages $\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[(x_i^k)^2]$ are bounded from below by some constant independent of n, and
- Ratios $\frac{\|x_j^0\|}{\alpha\mu_i}$ are bounded from above by some constant $\beta < 1$ independent from n,

then the bound in (29) grows exponentially in n.

Next we consider the case when the data are i.i.d. but follow the distribution which is a mixture of γ-SLC distributions.

Theorem 11. *Let $\delta > 0$, $\alpha \in (0, 1]$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points in $\mathbb{R}^n$, which follow the distribution with density*

$$f(x) = \sum_{i=1}^{k} \beta_i f_i(x),$$

where $\beta_i \geq 0$ are coefficients such that $\sum_{i=1}^{k} \beta_i = 1$, and f_i are densities of γ_i-SLC distributions with $\gamma_i > 0$. Let x_i^0 and μ_i be the expectation and norm expectation, respectively, of a random vector following distribution with density f_i. Assume that inequality

$$\|x_j^0\| < \alpha\mu_i$$

holds for every pair $1 \leq i, j \leq k$. If

$$M < \sqrt{\frac{\delta}{2}} \exp\left(\min_{i,j}\left(\frac{\gamma_i\gamma_j(\mu_i\alpha - \|x_j^0\|)^2}{2(\sqrt{\gamma_j}\alpha + \sqrt{\gamma_i})^2}\right)\right),$$

then the expected number of α-inseparable pairs in F is less than δ. In particular, set F is α-Fisher separable with probability greater than $1 - \delta$.

Proof. Let $\Omega \subset \mathbb{R}^{2n}$ be the set of points $(x_1, \dots, x_n, y_1, \dots, y_n) \in \mathbb{R}^{2n}$ such that $\alpha\sum_{k=1}^{n} x_k^2 \leq \sum_{k=1}^{n} x_k y_k$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in F$,

$$\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \leq (\boldsymbol{x}, \boldsymbol{y})] = \int_\Omega f(x)f(y)dxdy =$$
$$\int_\Omega \left(\sum_{i=1}^{k} \beta_i f_i(x)\right)\left(\sum_{j=1}^{k} \beta_j f_j(y)\right)dxdy =$$
$$\sum_{i=1}^{k}\sum_{j=1}^{k} \beta_i\beta_j \int_\Omega f_i(x)f_j(y)dxdy \leq \sum_{i=1}^{k}\sum_{j=1}^{k} \beta_i\beta_j B = B,$$

where

$$B = 2\exp\left(-\min_{i,j}\left(\frac{\gamma_i\gamma_j(\mu_i\alpha - \|x_j^0\|)^2}{2(\sqrt{\gamma_j}\alpha + \sqrt{\gamma_i})^2}\right)\right)$$

is the right-hand side of (34). Then application of Theorem 7 finishes the proof. □

A straightforward combination of Theorems 10 and 11 allows to treat even more general case when the data are independent but not identically distributed, and the distribution of each data point is a mixture of log-concave ones, but the notation become messy so we omit the details.

5. Separation theorems for rotation invariant log-concave distributions

Assume that points in $\mathbb{R}^n$ are selected from distribution whose density $\hat{\rho} : \mathbb{R}^n \to \mathbb{R}_+$, where $\mathbb{R}_+ = [0, \infty)$, is rotation invariant, that is,

$$\hat{\rho}(x) = C_n\rho(\|x\|), \quad \forall x \in \mathbb{R}^n \tag{35}$$

for some function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$, where the factor C_n is selected such that the density integrates to 1. In fact,

$$C_n = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{n/2}}\left(\int_0^\infty r^{n-1}\rho(r)dr\right)^{-1},$$

where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the gamma function.

This section derives separation theorems for such distributions. We start with optimal separation theorem for the most famous example of rotation invariant distribution, the standard normal one.

5.1. The standard normal distribution

For the standard normal distribution, the following result is presented in the conference paper (Grechuk, 2019, Corollary 6).

Theorem 12. *Let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ be i.i.d points from the standard normal distribution. For any $\delta > 0$, if*

$$M < \sqrt{\delta}\exp\left(\frac{1}{4}\log(1 + \alpha^2)n\right) = \sqrt{\delta}(1 + \alpha^2)^{n/4}, \tag{36}$$

then set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is α-Fisher separable with probability greater than $1 - \delta$.

Theorem 12 follows from Theorem 7 and estimate

$$\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \leq (\boldsymbol{x}, \boldsymbol{y})] \leq (1 + \alpha^2)^{-n/2}$$

for i.i.d. points $\boldsymbol{x}$ and $\boldsymbol{y}$ from the standard normal distribution. Here, we derive the *exact* expression for $\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \leq (\boldsymbol{x}, \boldsymbol{y})]$.

From rotation invariance, we may assume that $\boldsymbol{x} = (\|\boldsymbol{x}\|, 0, 0, \dots, 0)$. Then

$$\mathbb{P}[\alpha(\boldsymbol{x}, \boldsymbol{x}) \leq (\boldsymbol{x}, \boldsymbol{y})] = \mathbb{P}[\alpha\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{x}\|y_1] =$$
$$\mathbb{P}[0 < y_1]\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{y_1} \leq \frac{1}{\alpha}\right] = \frac{1}{2}\mathbb{P}\left[\frac{\|\boldsymbol{x}\|^2/n}{y_1^2} \leq \frac{1}{n\alpha^2}\right], \tag{71}$$

where y_1 is the first component of $\boldsymbol{y}$, which follows the standard normal distribution. The sum of squares of k independent standard normal random variables follows the chi-squared distribution $\chi(k)$ with degree k. Hence, $\frac{\|\boldsymbol{x}\|^2/n}{y_1^2}$ is the ratio of two independent random variables from chi-squared distributions with degrees n and 1, respectively, scaled by their degrees. This ratio is known to follow so-called F-distribution $F(d_1, d_2)$ with parameters $d_1 = n$ and $d_2 = 1$. The cumulative distribution function of F-distribution is

$$F(x; d_1, d_2) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right),$$

where $I_z(a, b)$ is the cumulative distribution function of beta distribution, also known as regularized incomplete beta function. It is given by

$$I_z(a, b) = \frac{B_z(a, b)}{B(a, b)}, \tag{37}$$

Table 6
The upper bound of M (39) that guarantees α-Fisher separability of an M-element i.i.d. sample from the standard normal distribution with probability $p > 0.99$ for various α and dimensions.

	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
$n = 10$	1.19	1.45	1.99
$n = 50$	14	164	2075
$n = 100$	794	93,806	$1.4 \cdot 10^7$
$n = 200$	$2 \cdot 10^6$	$2.6 \cdot 10^{10}$	$5.6 \cdot 10^{14}$
$n = 500$	$2.6 \cdot 10^{16}$	$4.2 \cdot 10^{26}$	$2.6 \cdot 10^{37}$
$n = 1000$	$1.5 \cdot 10^{33}$	$3.6 \cdot 10^{53}$	$1.3 \cdot 10^{75}$

where $B_z(a,b) = \int_0^z t^{a-1}(1-t)^{b-1}dt$ is the incomplete beta function, and

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

is the beta function.

Hence,

$$\mathbb{P}[\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})] = \frac{1}{2}F\left(\frac{1}{n\alpha^2}; n, 1\right) = \frac{1}{2}I_{\frac{1}{1+\alpha^2}}\left(\frac{n}{2}, \frac{1}{2}\right) \quad (38)$$

With Theorem 7, this implies the following optimal separation result.

Theorem 13. *Let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ are i.i.d points from the standard normal distribution. For any $\delta > 0$, the expected number of α-inseparable pairs from set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is less than δ if and only if*

$$M < \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{2\delta}{I_{\frac{1}{1+\alpha^2}}\left(\frac{n}{2}, \frac{1}{2}\right)}}. \quad (39)$$

In particular, (39) implies that F is α-Fisher separable with probability greater than $1-\delta$.

Example 8. Let $\delta = 0.01$. Table 6 shows the upper bounds on M in Theorem 13 for $\alpha = 0.6, 0.8$ and 1 in various dimensions n.

For example, for $n = 100$, we see that 14 millions points from the standard normal distribution are 1-Fisher-separable with probability greater than 99%. In dimension $n = 200$, millions of points become Fisher separable even at level $\alpha = 0.6$.

The following proposition establishes asymptotic behavior of (38) as n goes to infinity.

Proposition 5. *For every $a > 0$, $b \in (0,1)$ and $z \in (0,1)$, we have*

$$I_z(a,b) \le \frac{z^a(1-z)^{b-1}a^{b-1}}{\Gamma(b)}, \quad (40)$$

and the bound is asymptotically tight if b and z are fixed but $a \to \infty$, in sense that ratio of the right and left sides converges to 1. In particular,

$$\frac{1}{2}I_{\frac{1}{1+\alpha^2}}\left(\frac{n}{2}, \frac{1}{2}\right) \le \sqrt{\frac{1+\alpha^2}{2\pi n\alpha^2}}(1+\alpha^2)^{-n/2}, \quad (41)$$

and the bound is asymptotically tight if α is fixed and $n \to \infty$.

Proof. Wendel (1948) proved that for every $a > 0$ and $b \in (0,1)$ $\frac{\Gamma(a+b)}{a^b\Gamma(a)} \le 1$, and $\lim_{a\to\infty}\frac{\Gamma(a+b)}{a^b\Gamma(a)} = 1$. This is equivalent to

$$B(a,b) \ge \Gamma(b)a^{-b} \quad \text{and} \quad \lim_{a\to\infty}\frac{B(a,b)}{\Gamma(b)a^{-b}} = 1. \quad (42)$$

Asymptotic expansion (Lopez & Sesma, 1999) implies that

$$B_x(a,b) \le \frac{x^a(1-x)^{b-1}}{a},$$

and $\lim_{a\to\infty}\frac{B_x(a,b)a}{x^a(1-x)^{b-1}} = 1$. Hence, inequality (40) holds and is asymptotically tight. Applying it with $z = \frac{1}{1+\alpha^2}$, $a = n/2$, $b = 1/2$, and using the fact that $\Gamma(1/2) = \pi$, we get (41). □

Theorem 13 and Proposition 5 imply the following corollary, which provides a simple but asymptotically tight estimate for M.

Corollary 5. *Let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ are i.i.d points from the standard normal distribution. For any $\delta > 0$, if*

$$M < \sqrt[4]{\frac{2\pi n\alpha^2\delta^2}{1+\alpha^2}}(1+\alpha^2)^{n/4}, \quad (43)$$

then the expected number of α-inseparable pairs in set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is less than δ. In particular, (43) implies that set F is α-Fisher separable with probability greater than $1-\delta$.

The bound in (43) is weaker than in (39), but the ratio of these bounds converges to 1 if α is fixed and n goes to infinity. In particular, the optimal exponent (19) for the standard normal distribution is

$$b(\alpha) = \frac{1}{4}\log(1+\alpha^2).$$

Corollary 5 is an improvement over Theorem 12 if $n > \frac{1+\alpha^2}{2\pi\alpha^2}$.

Example 9. If $\delta = 0.01$, $\alpha = 0.9$ and $n = 100$, then (36) in Theorem 12 reduces to $M < 276,671$, (43) in Corollary 5 reduces to $M < 1,132,950$, while the optimal bound (39) in Theorem 13 is $M < 1,141,060$.

5.2. Optimal separation theorem for explicitly given distribution

This section establishes optimal separation theorem if the rotation invariant distribution is not necessary standard normal but is explicitly given.

Proposition 6. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two points selected independently from rotation invariant distributions with the same center. Then*

$$\mathbb{P}[\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})] = \frac{\alpha}{B(\frac{n-1}{2}, \frac{1}{2})}\int_0^{1/\alpha}(1-\alpha^2t^2)^{\frac{n-3}{2}}\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le t\right]dt, \quad (44)$$

where $B(.)$ is the beta function.

Proof. Note that

$$\mathbb{P}[\alpha(\boldsymbol{x},\boldsymbol{x}) \le (\boldsymbol{x},\boldsymbol{y})] = \mathbb{P}[\alpha\|\boldsymbol{x}\|^2 \le \|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|\cos\beta] = \mathbb{P}[\alpha t \le \cos\beta],$$

where $t = \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|}$, and β is an angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. By spherical invariance, the last probability is equal to the ratio of the area of the hyperspherical cap with angle $\beta_0 = \arccos(\alpha t) = \arcsin(\sqrt{1-\alpha^2t^2})$ to the area of the whole hypersphere, provided that $t \le \frac{1}{\alpha}$. By Li (2011), this ratio is equal to $\frac{1}{2}I_{\sin^2\beta_0}\left(\frac{n-1}{2}, \frac{1}{2}\right)$, where $I_z(a,b)$ is given by (37). Hence,

$$\mathbb{P}[\alpha t \le \cos\beta] = \frac{1}{2}\int_0^{1/\alpha} I_{1-\alpha^2t^2}\left(\frac{n-1}{2}, \frac{1}{2}\right)u(t)dt,$$

where u is the density of the distribution of $\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|}$.

Using the formula for density of beta distribution, we get

$$\frac{d}{dt}I_{1-\alpha^2t^2}\left(\frac{n-1}{2},\frac{1}{2}\right)= \frac{(1-\alpha^2t^2)^{\frac{n-1}{2}-1}(1-(1-\alpha^2t^2))^{1/2-1}}{B(\frac{n-1}{2},\frac{1}{2})}\frac{d}{dt}(1-\alpha^2t^2)= \frac{-2\alpha(1-\alpha^2t^2)^{\frac{n-3}{2}}}{B(\frac{n-1}{2},\frac{1}{2})},$$

where $B(a,b)=\int_0^1 z^{a-1}(1-z)^{b-1}dz$ is the beta function. Hence, integration by parts yields (44). □

If $\boldsymbol{x}$ has density given by (35), then $\|\boldsymbol{x}\|$ has density given by $C_n r^{n-1}\rho(r)$, where $C_n=\left(\int_0^\infty r^{n-1}\rho(r)dr\right)^{-1}$ is the normalization constant. Hence,

$$\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|}\le t\right]=h(n,t):=\frac{\int_0^\infty y^{n-1}\rho(y)dy\int_0^{ty}x^{n-1}\rho(x)dx}{\left(\int_0^\infty r^{n-1}\rho(r)dr\right)^2}, \tag{45}$$

where ρ is defined in (35). Hence, Proposition 6 in combination Theorem 7 implies the following optimal separation theorem.

Theorem 14. *Let $\delta>0$, $\alpha\in(0,1]$, and let $F=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_M\}$ be a set of M i.i.d. random points from a rotation invariant distribution in $\mathbb{R}^n$. Then the expected number of α-inseparable pairs from set F is less than δ if and only if*

$$M<\frac{1}{2}+\sqrt{\frac{1}{4}+\frac{\delta}{p(n,\alpha)}}, \quad \text{where } p(n,\alpha)=\frac{\alpha}{B(\frac{n-1}{2},\frac{1}{2})}\int_0^{1/\alpha}(1-\alpha^2t^2)^{\frac{n-3}{2}}h(n,t)dt, \tag{46}$$

and $h(n,t)$ is defined in (45). In particular, (46) implies that F is α-Fisher separable with probability greater than $1-\delta$.

We next apply Theorem 14 to some famous rotation invariant distributions.

5.3. Uniform distribution in a ball

We may assume that the ball has radius 1. Uniform distribution in the unit ball is given by (35) with $\rho(r)=1, 0\le r\le 1$. Substituting this into (45) and integrating, we get

$$h(n,t)=\begin{cases}t^n/2, & 0\le t\le 1\\ 1-t^{-n}/2, & 1\le t.\end{cases}$$

Hence $p(n,\alpha)$ in (46) is given by:

$$p(n,\alpha)=\frac{\alpha}{2B(\frac{n-1}{2},\frac{1}{2})}\left(\int_0^1(1-\alpha^2t^2)^{\frac{n-3}{2}}t^n dt+\int_1^{1/\alpha}(1-\alpha^2t^2)^{\frac{n-3}{2}}(2-t^{-n})dt\right). \tag{47}$$

Note that the answer may be written down explicitly using hypergeometric functions, but we find it more convenient to work with the integral expression.

With Theorem 14, this implies the following result.

Theorem 15. *Let $\alpha\in(0,1]$, and let points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_M$ be i.i.d points from uniform distribution in a ball. For any $\delta>0$, the expected number of α-inseparable pairs from set $F=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_M\}$ is less than δ if and only if $M<\frac{1}{2}+\sqrt{\frac{1}{4}+\frac{\delta}{p(n,\alpha)}}$, where $p(n,\alpha)$ is given by (47). In particular, in this case F is α-Fisher separable with probability greater than $1-\delta$.*

To find the asymptotic growth of (47) as $n\to\infty$, we will use the Laplace's method. Informally, it states that, if function $h(t)$ has a unique maximum on $[a,b]$ attained at $t=c$, and $\phi(c)\ne 0$, then, for large n, the value of the integral

$$I(n)=\int_a^b\phi(t)e^{nh(t)}dt \tag{48}$$

depends mainly on $\phi(c)$ and the behavior of $h(t)$ is the neighborhood of c. We can then replace in (48) $\phi(t)$ by $\phi(c)$ and $h(t)$ by its Taylor expansion at $t=c$ up to the first non-zero term, and integrate. We get

$$I(n)\sim\phi(c)e^{nh(c)}\sqrt{\frac{2\pi}{n|h''(c)|}} \tag{49}$$

if $a<c<b$, $h'(c)=0$, $h''(c)\ne 0$,

$$I(n)\sim\phi(c)e^{nh(c)}\sqrt{\frac{\pi}{2n|h''(c)|}}$$

if $c=a$ or $c=b$, and $h'(c)=0$, $h''(c)\ne 0$, and

$$I(n)\sim\frac{\phi(c)e^{nh(c)}}{n|h'(c)|} \tag{50}$$

if $c=a$ or $c=b$, and $h'(c)\ne 0$. We refer to Wong (2001, Theorem 1, p. 58) for a formal statement and proof.

Applying this method to (47), we get the following estimate.

Proposition 7. *Let $p(n,\alpha)$ be given by (47) and $n>3$.*

I. If $0<\alpha<\frac{\sqrt{2}}{2}$, then

$$p(n,\alpha)\le q(n,\alpha):=\sqrt{\frac{1}{2\pi}}\frac{1}{\alpha(1-2\alpha^2)}\frac{n^{3/2}}{(n-3)^2}(1-\alpha^2)^{\frac{n+3}{2}}, \quad \text{and} \quad p(n,\alpha)\sim q(n,\alpha), \tag{51}$$

II. If $\frac{\sqrt{2}}{2}<\alpha\le 1$, then

$$p(n,\alpha)\le\frac{1}{2}(2\alpha)^{-n}, \quad \text{and} \quad p(n,\alpha)\sim\frac{1}{2}(2\alpha)^{-n} \tag{52}$$

(with equality for $\alpha=1$).

Proof. For the coefficient before the integral in (47), (42) implies that

$$\frac{\alpha}{2B(\frac{n-1}{2},\frac{1}{2})}\le\frac{\alpha\sqrt{n}}{2\sqrt{2\pi}}, \quad \text{and} \quad \frac{\alpha}{2B(\frac{n-1}{2},\frac{1}{2})}\sim\frac{\alpha\sqrt{n}}{2\sqrt{2\pi}}. \tag{53}$$

The first integral $I_1(n):=\int_0^1(1-\alpha^2t^2)^{\frac{n-3}{2}}t^n dt$ in (47) can be written in the form (48) with $h(t)=\frac{1}{2}\log(1-\alpha^2t^2)+\log(t)$, $n:=n-3$, $\phi(t)=t^3$. If $\frac{\sqrt{2}}{2}<\alpha\le 1$, $h(t)$ attains maximum on $(0,1]$ at point $t=c:=\frac{1}{\sqrt{2}\alpha}$, $h(c)=-\log(2\alpha)$, $h'(c)=0$, $h''(c)=-8\alpha^2$, $\phi(c)=(\sqrt{2}\alpha)^{-3}$, and (49) implies that

$$I_1(n)\sim(\sqrt{2}\alpha)^{-3}e^{(n-3)(-\log(2\alpha))}\sqrt{\frac{2\pi}{8(n-3)\alpha^2}}\sim\sqrt{\frac{2\pi}{n\alpha^2}}(2\alpha)^{-n}. \tag{54}$$

If $0<\alpha<\frac{\sqrt{2}}{2}$, $h(t)$ attains maximum on $(0,1]$ at point $t=1$, $h(1)=\frac{1}{2}\log(1-\alpha^2)$, $h'(1)=\frac{1-2\alpha^2}{1-\alpha^2}$, and inequalities $h(t)\le h(1)+h'(1)(t-1)$, $0<t\le 1$, $\phi(t)=t^3\le 1$, $0<t\le 1$

imply that

$$I_1(t) \le \int_0^1 \exp\left[(n-3)(h(1)+h'(1)(t-1))\right]dt = \frac{(1-\alpha^2)^{\frac{n-1}{2}}}{(1-2\alpha^2)(n-3)}\left(1-\exp\left(-\frac{(1-2\alpha^2)(n-3)}{1-\alpha^2}\right)\right) \le \frac{(1-\alpha^2)^{\frac{n-1}{2}}}{(1-2\alpha^2)(n-3)}, \tag{55}$$

and (50) implies that

$$I_1(t) \sim \frac{(1-\alpha^2)^{\frac{n-1}{2}}}{(1-2\alpha^2)(n-3)}. \tag{56}$$

The second integral $I_2(n) := \int_1^{1/\alpha}(1-\alpha^2t^2)^{\frac{n-3}{2}}(2-t^{-n})$ in (47) can be estimated similarly. Inequalities $\frac{1}{2}\log(1-\alpha^2t^2) \le \frac{1}{2}\log(1-\alpha^2) - \frac{\alpha^2}{1-\alpha^2}(t-1)$ and $-\log(t) \ge -(t-1)$ imply that

$$I_2(t) \le \int_1^{1/\alpha} \exp\left[(n-3)\left(\frac{\log(1-\alpha^2)}{2} - \frac{\alpha^2(t-1)}{1-\alpha^2}\right)\right] \times \left(2-e^{-n(t-1)}\right)dt = (1-\alpha^2)^{\frac{n-1}{2}}\left(\frac{2}{\alpha^2(n-3)} - \frac{1}{n-3\alpha^2} + \frac{\exp\left(-\frac{n-3\alpha^2}{\alpha+\alpha^2}\right)}{n-3\alpha^2} - \frac{\exp\left(-\frac{\alpha(n-3)}{1+\alpha}\right)}{\alpha^2(n-3)}\right) \le (1-\alpha^2)^{\frac{n-1}{2}}\left(\frac{2}{\alpha^2(n-3)} - \frac{1}{n-3\alpha^2}\right) \le (1-\alpha^2)^{\frac{n-1}{2}}\frac{(2-\alpha^2)n}{\alpha^2(n-3)^2}, \tag{57}$$

where the second inequality follows from the facts that $n-3\alpha^2 \ge \alpha^2(n-3)$ and $\exp\left(-\frac{n-3\alpha^2}{\alpha+\alpha^2}\right) \le \exp\left(-\frac{\alpha(n-3)}{1+\alpha}\right)$. Moreover, (50) implies that the first inequality in (57) is asymptotically tight, and the asymptotic tightness of the second and third inequalities in (57) is straightforward, hence

$$I_2(t) \sim (1-\alpha^2)^{\frac{n-1}{2}}\frac{(2-\alpha^2)n}{\alpha^2(n-3)^2}. \tag{58}$$

If $0 < \alpha < \frac{\sqrt{2}}{2}$, the combination of (53), (55), and (57) yields

$$p(n,\alpha) \le \frac{\alpha\sqrt{n}}{2\sqrt{2\pi}}\frac{(1-\alpha^2)^{\frac{n-1}{2}}n}{(n-3)^2}\left(\frac{1}{1-2\alpha^2}\frac{n-3}{n} + \frac{2-\alpha^2}{\alpha^2}\right) \le q(n,\alpha),$$

where the second inequality follows from substituting 1 instead of $\frac{n-3}{n}$ and simplifying. The $p(n,\alpha) \sim q(n,\alpha)$ part of (51) follow from (53), (56), and (58).

The inequality in (52) follows from (17). For $\frac{\sqrt{2}}{2} < \alpha \le 1$, the $\sim$ part of (52) follows from (53), (54), and (58). □

We conjecture that factor $\frac{n^{3/2}}{(n-3)^2}$ in (51) can be improved to a simpler factor $\sqrt{n}$, which would allow to remove the condition $n > 3$, but this improvement is negligible for large n, and the $\sim$ part of (51) implies that asymptotically non-negligible improvement is impossible, and bound (51) is essentially the best possible if $0 < \alpha < \frac{\sqrt{2}}{2}$. Similarly, bound (52) is essentially the best possible if $\frac{\sqrt{2}}{2} < \alpha \le 1$.

Theorem 15 and Proposition 7 imply the following corollary.

Table 7
The upper bound of M (59) that guarantees α-Fisher separability with probability $p > 0.99$ of M i.i.d. points sampled from uniform distribution in a ball, for various α and dimensions.

	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
$n = 10$	0.25	0.34	0.2
$n = 50$	8.9	60	350
$n = 100$	400	19,491	$1.9 \cdot 10^6$
$n = 200$	642,645	$1.6 \cdot 10^9$	$4.8 \cdot 10^{13}$
$n = 500$	$1.9 \cdot 10^{15}$	$7.1 \cdot 10^{23}$	$5.2 \cdot 10^{35}$
$n = 1000$	$9.4 \cdot 10^{30}$	$1.4 \cdot 10^{48}$	$2.2 \cdot 10^{72}$

Corollary 6. *Let $\alpha \in \left(0, \frac{1}{\sqrt{2}}\right)$, $n > 3$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be the set of M i.i.d points from the uniform distribution in a ball. For any $\delta > 0$, if*

$$M \le \sqrt[4]{2\pi}\sqrt{\delta\alpha(1-2\alpha^2)}\frac{n-3}{n^{3/4}}\left(\frac{1}{1-\alpha^2}\right)^{\frac{n+3}{4}}, \tag{59}$$

then the expected number of α-inseparable pairs in F is less than δ. In particular, (59) implies that F is α-Fisher separable with probability greater than $1-\delta$.

Proposition 7 implies that the bound for M in Corollary 6 is asymptotically tight, and has the advantage of being a simple explicit formula. For $\alpha \ge \frac{1}{\sqrt{2}}$, (52) implies that an asymptotically tight bound is given in Theorem 2.

Example 10. Let $\delta = 0.01$. Table 7 shows the upper bounds on M in Corollary 6 for $\alpha = 0.5, 0.6$ and 0.7 in various dimensions n.

For example, for $n = 200$, we see that 642,645 points from the uniform distribution in the unit ball are Fisher-separable at level $\alpha = 0.5$ with probability greater than 99%. For comparison, with the same parameters, the bound (7) in Theorem 2 reduces to $M < 0.14$. The optimal bound in Theorem 15 gives $M < 661,243$.

The results of this section can be straightforwardly extended to the case when the points in $\mathbb{R}^n$ are selected from the uniform distribution in a spherical layer, that is, from the distribution (35) with

$$\rho(r) = \begin{cases} \frac{1}{1-R}, & R \le r \le 1, \\ 0, & \text{otherwise}, \end{cases}$$

where $R \in (0,1)$ is a parameter. Then $h(n,t)$ in (45) is given by

$$h(n,t) = \begin{cases} 0, & t \le R \\ \frac{t^{-n}(R^n-t^n)^2}{2(1-R^n)^2}, & R < t \le 1, \\ 1-\frac{t^n(R^n-t^{-n})^2}{2(1-R^n)^2}, & 1 \le t < 1/R, \\ 1, & 1/R < t. \end{cases} \tag{60}$$

With Theorem 14, this implies that if M i.i.d. points are selected from this distribution, then the expected number of α-inseparable pairs is less than δ if and only if

$$M < \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\delta}{p(n,\alpha)}},$$

where $p(n,\alpha)$ is given by (46) with $h(n,t)$ given by (60).

Sidorov and Zolotykh (2020) provided simpler (although non-optimal) versions of the separation theorems for spherical layers, with asymptotic analysis and numerical experiments.

5.4. Multivariate exponential distribution

By multivariate exponential distribution in $\mathbb{R}^n$ we will mean rotation invariant distribution such that $\rho(\|\boldsymbol{x}\|)$ in (35) is equal

Table 8
The upper bounds on M (61) that guarantees α-Fisher separability of M i.i.d. points from exponential distribution with probability $p > 0.99$ for various α and dimensions.

	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
$n = 10$	0.65	0.81	1.06
$n = 50$	7.6	43	249
$n = 100$	218	6662	203,805
$n = 200$	154,501	$1.3 \cdot 10^8$	$1.1 \cdot 10^{11}$
$n = 500$	$4.1 \cdot 10^{13}$	$7.6 \cdot 10^{20}$	$1.6 \cdot 10^{28}$
$n = 1000$	$3.8 \cdot 10^{27}$	$1.1 \cdot 10^{42}$	$4.8 \cdot 10^{56}$

to $\exp(-\|\boldsymbol{x}\|)$. In this case, the distribution of $\|\boldsymbol{x}\|$ is the standard Gamma distribution with n degrees of freedom, and, for i.i.d. $\boldsymbol{x}$ and $\boldsymbol{y}$, ratio $\frac{\boldsymbol{x}}{\boldsymbol{y}}$ follows beta prime distribution, that is,

$$\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le t\right] = I_{\frac{t}{1+t}}(n, n),$$

where $I_z(a, b)$ is the regularized incomplete beta function defined in (37). Hence, Theorem 14 implies the following result.

Theorem 16. *Let $\alpha \in (0, 1]$, and let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ be i.i.d points from exponential distribution in $\mathbb{R}^n$. For any $\delta > 0$, the expected number of α-inseparable pairs from set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is less than δ if and only if*

$$M < \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\delta}{p(n, \alpha)}}, \tag{61}$$

where $p(n, \alpha) = \dfrac{\alpha}{B(\frac{n-1}{2}, \frac{1}{2})} \displaystyle\int_0^{1/\alpha} (1 - \alpha^2 t^2)^{\frac{n-3}{2}} I_{\frac{t}{1+t}}(n, n) dt.$

In particular, (61) implies that F is α-Fisher separable with probability greater than $1 - \delta$.

Example 11. Let $\delta = 0.01$. Table 8 shows the upper bounds on M in Theorem 16 for $\alpha = 0.6$, 0.8 and 1 in various dimensions n. For example, for $n = 100$, we see that over 200,000 points from the multivariate exponential distribution are Fisher-separable at level $\alpha = 1$ with probability greater than 99%. In dimension $n = 200$, over 150,000 points from the same distribution become Fisher-separable at level $\alpha = 0.6$.

The growth of factor $I_{\frac{t}{1+t}}(n, n)$ in (61) is described by the following proposition.

Proposition 8. *For any $t > 0$ and $n \ge 1$,*

$$I_{\frac{t}{1+t}}(n, n) = \begin{cases} \frac{1}{2} I_{\frac{4t}{(1+t)^2}}(n, \frac{1}{2}), & 0 < t \le 1, \\ 1 - \frac{1}{2} I_{\frac{4t}{(1+t)^2}}(n, \frac{1}{2}), & 1 \le t. \end{cases} \tag{62}$$

In particular,

$$I_{\frac{t}{1+t}}(n, n) \le \frac{1}{2\sqrt{\pi n}} \frac{1+t}{1-t} \left(\frac{4t}{(1+t)^2}\right)^n, \quad 0 < t < 1, \tag{63}$$

and this upper bound is asymptotically tight if t is fixed and $n \to \infty$.

Proof. If $0 < t \le 1$, then $z = \frac{t}{1+t} \le \frac{1}{2}$, and

$$B_z(n, n) = \int_0^z u^{n-1}(1-u)^{n-1} du = \int_0^{4z(1-z)} (s/4)^{n-1} \frac{ds}{4\sqrt{1-s}} = 4^{-n} B_{4z(1-z)}\left(n, \frac{1}{2}\right),$$

where the second inequality is the change of variables $s = 4u(1-u)$. Next,

$$B(n, n) = \int_0^1 u^{n-1}(1-n)^{n-1} dt = 2\int_0^{1/2} u^{n-1}(1-n)^{n-1} dt = 2B_{\frac{1}{2}}(n, n) = 4^{-n} B\left(n, \frac{1}{2}\right),$$

hence

$$I_z(n, n) = \frac{B_z(n, n)}{B(n, n)} = \frac{4^{-n} B_{4z(1-z)}\left(n, \frac{1}{2}\right)}{4^{-n} B\left(n, \frac{1}{2}\right)} = I_{4z(1-z)}\left(n, \frac{1}{2}\right),$$

which with $z = \frac{t}{1+t}$ implies the first line of (62). The second line follows from the first one and the identity $I_z(a, b) = 1 - I_{1-x}(b, a)$.

For $0 < t < 1$, (40) with $a = n$, $b = 1/2$, and $z = \frac{4t}{(1+t)^2}$ implies that

$$I_{\frac{t}{1+t}}(n, n) = \frac{1}{2} I_{\frac{4t}{(1+t)^2}}(n, \frac{1}{2}) \le \left(\frac{4t}{(1+t)^2}\right)^n \left(1 - \frac{4t}{(1+t)^2}\right)^{-1/2} \frac{n^{-1/2}}{2\Gamma(1/2)},$$

which simplifies to the right-hand side of (63). □

The next proposition establishes asymptotic growth of $p(n, \alpha)$ in (61) as $n \to \infty$.

Proposition 9. *Let $p(n, \alpha)$ be given by (61). Then*

$$p(n, \alpha) \sim \frac{\sqrt{1 + 5\alpha^2 + (1+\alpha^2)\sqrt{1+8\alpha^2}}}{2\alpha\sqrt{\pi n}\sqrt[4]{1+8\alpha^2}} \left(\frac{4\sqrt{2}\alpha(\sqrt{1+8\alpha^2} - 1)}{(\sqrt{1+8\alpha^2} + 4\alpha^2 - 1)^{3/2}}\right)^n. \tag{64}$$

Proof. Proposition 8 together with obvious bound $I_{\frac{t}{1+t}}(n, n) \le 1$ implies that the integral in (61) is bounded by

$$I_n := \int_0^{1/\alpha} (1 - \alpha^2 t^2)^{\frac{n-3}{2}} I_{\frac{t}{1+t}}(n, n) dt \le J_n + L_n, \tag{65}$$

where

$$J_n := \int_0^1 (1 - \alpha^2 t^2)^{\frac{n-3}{2}} \min\left\{\frac{1}{2\sqrt{\pi n}} \frac{1+t}{1-t} \left(\frac{4t}{(1+t)^2}\right)^n, 1\right\} dt,$$
$$L_n := \int_1^{1/\alpha} (1 - \alpha^2 t^2)^{\frac{n-3}{2}} dt.$$

Because for $n \ge 3$ we have $J_n \le 1$ and $L_n \le \frac{1}{\alpha} - 1$, and (63) is asymptotically tight, (65) is also asymptotically tight by dominated convergence theorem.

Integral J_n can be written in the form (48) with

$$\phi_n(t) := (1 - \alpha^2 t^2)^{-3/2} \min\left\{\frac{1}{2\sqrt{\pi n}} \frac{1+t}{1-t}, \left(\frac{(1+t)^2}{4t}\right)^n\right\}$$

and

$$h(t) := \log\left(\sqrt{1 - \alpha^2 t^2} \frac{4t}{(1+t)^2}\right).$$

Function $h(t)$ attains maximum on $(0, 1)$ at

$$t_0 = t_0(a) := \frac{\sqrt{1+8a^2} - 1}{4a^2}$$

and by (49)

$$J_n \sim \phi_n(t_0)\sqrt{\frac{2\pi}{n|h''(t_0)|}}e^{nh(t_0)} \sim$$

$$(1-\alpha^2 t_0^2)^{-3/2}\frac{1}{n}\frac{1+t_0}{1-t_0}\sqrt{\frac{1}{2|h''(t_0)|}}e^{nh(t_0)},$$

where the second $\sim$ follows from the fact that $\frac{1}{2\sqrt{\pi n}}\frac{1+t_0}{1-t_0} < \left(\frac{(1+t_0)^2}{4t_0}\right)^n$ for large n.

Similarly, by (50)

$$L_n \sim \frac{1-\alpha^2}{n\alpha^2}\left(\sqrt{1-\alpha^2}\right)^{(n-3)}.$$

Because $e^{h(t_0)} > \sqrt{1-\alpha^2}$ for all $0 < \alpha \le 1$, $I_n \sim J_n + L_n \sim J_n$ as $n \to \infty$. This together with (61) and (53) implies that

$$p(n,\alpha) \sim \frac{\alpha\sqrt{n}}{\sqrt{2\pi}}(1-\alpha^2 t_0^2)^{-3/2}\frac{1}{n}\frac{1+t_0}{1-t_0}\sqrt{\frac{1}{2|h''(t_0)|}}e^{nh(t_0)},$$

which simplifies to (64). □

For $\alpha = 1$, Theorem 16 and Proposition 9 imply the following corollary.

Corollary 7. *Let points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ are i.i.d points from exponential distribution in $\mathbb{R}^n$. For any $\delta > 0$, if*

$$M < \sqrt{\frac{\delta}{p(n,1)}} \sim \sqrt{\delta}\sqrt[4]{\pi n}\left(\frac{\sqrt[4]{27}}{2}\right)^n,$$

then the expected number of 1-inseparable pairs in set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is less than δ. In particular, set F is 1-Fisher separable with probability greater than $1-\delta$.

In particular,

$$b(1) = \log\left(\frac{\sqrt[4]{27}}{2}\right) = 0.1308\dots,$$

where $b(\alpha)$ is defined in (19). For comparison, for uniform distribution in a ball $b(1) = \frac{1}{2}\log 2 = 0.3465...$, while for the standard normal distribution $b(1) = \frac{1}{4}\log 2 = 0.1732...$.

5.5. General log-concave rotation invariant distribution

This section derives separation theorems for arbitrary rotation invariant distribution. We start with the following easy result.

Theorem 17. *Let $\delta > 0$, $\alpha \in (1/2, 1]$, and let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from a rotation invariant log-concave distribution in $\mathbb{R}^n$. If*

$$M < \sqrt{\frac{\delta}{2}}\exp\left(n\frac{(2\alpha-1)^2}{8(2\alpha+1)^2}\right), \tag{66}$$

then the expected number of α-inseparable pairs in set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is less than δ. In particular, set F is α-Fisher separable with probability greater than $1-\delta$.

Proof. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two i.i.d points from the given distribution. Inequality (4) can be rewritten as

$$\left\|\boldsymbol{x} - \frac{\boldsymbol{y}}{2\alpha}\right\| \le \left\|\frac{\boldsymbol{y}}{2\alpha}\right\|,$$

that is, $\boldsymbol{x}$ belongs to a ball of radius $\left\|\frac{\boldsymbol{y}}{2\alpha}\right\|$. For every fixed $t > 0$, this may happen if either

(i) $\|\boldsymbol{y}\| > t$, or
(ii) $\boldsymbol{x}$ belongs to a ball of radius at most $\frac{t}{2\alpha}$.

We will prove that, for $t = \frac{4\alpha}{2\alpha+1}\mu$, where $\mu = \mathbb{E}[\|\boldsymbol{x}\|] = \mathbb{E}[\|\boldsymbol{y}\|]$, both these possibilities can happen with probability at most $e^{-n\frac{(2\alpha-1)^2}{4(2\alpha+1)^2}}$. With (16), this will imply (66).

As observed by Bobkov (2010, p. 328), if random variable $\boldsymbol{x}$ has density given by (35) with log-concave ρ, then $\|\boldsymbol{x}\|$ has log-concave distribution of order n (that is, has density of the form $q(r) = r^{n-1}\rho(r)$ for log-concave ρ). According to Bobkov (2010, Corollary 3.2), this implies that, for any $h \in [0.1]$,

$$\mathbb{P}[\|\boldsymbol{x}\| - \mu \ge h\mu] \le e^{-nh^2/4} \tag{67}$$

and

$$\mathbb{P}[\mu - \|\boldsymbol{x}\| \ge h\mu] \le e^{-nh^2/4}. \tag{68}$$

With $h = \frac{2\alpha-1}{2\alpha+1}$, (67) implies that probability of (i) is at most $e^{-n\frac{(2\alpha-1)^2}{4(2\alpha+1)^2}}$, while (68) implies that

$$\mathbb{P}\left[\|\boldsymbol{x}\| \le \frac{2}{2\alpha+1}\mu\right] \le e^{-n\frac{(2\alpha-1)^2}{4(2\alpha+1)^2}}.$$

In other words, the probability that $\boldsymbol{x}$ belongs to a ball B of radius $\frac{t}{2\alpha}$ centered at origin is at most $e^{-n\frac{(2\alpha-1)^2}{4(2\alpha+1)^2}}$. However, because the density $\hat{\rho}$ is rotation invariant and log-concave, we have $\hat{\rho}(x) \ge \hat{\rho}(y)$ for every $x \in B$ and $y \notin B$, hence shifting the ball cannot increase the probability for a point to belong to it. □

The bound (66) in Theorem 17 is simple and explicit. For example, for $\alpha = 1$ it reduces to

$$M < \sqrt{\frac{\delta}{2}}\exp\left(\frac{n}{72}\right). \tag{69}$$

However, the bound is far from being optimal, and the Theorem is not applicable for $\alpha \le \frac{1}{2}$. We next prove a separation theorem with more complicated but better bound. It also applies to a broader class of distributions, because it does not require for ρ in (35) to be non-increasing.

Theorem 18. *Let $\delta > 0$, $\alpha \in (0, 1]$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from a distribution in $\mathbb{R}^n$ given by (35) with log-concave ρ. If*

$$M < \sqrt{\delta} f(n,\alpha)^{-1/2}, \tag{70}$$

where $f(n,\alpha)$ is an explicit function defined in formulas (71)–(73) below, then the expected number of α-inseparable pairs in set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ is less than δ. In particular, (70) implies that set F is α-Fisher separable with probability greater than $1-\delta$.

Proof. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be any i.i.d. points from the given distribution. Let us derive an upper bound for $\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le t\right]$ for any $t \in (0, 1/\alpha)$. Let $q(.)$ be the density for absolute value distribution. We have

$$\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le t\right] = \int_0^\infty q(x)dx\int_{x/t}^\infty q(y)dy = \int_0^\infty q(x)dx \cdot \mathbb{P}\left[\|\boldsymbol{y}\| \ge \frac{x}{t}\right].$$

We claim that

$$\mathbb{P}\left[\|\boldsymbol{y}\| \ge \frac{x}{t}\right] \le \psi_{n,t}(x) := \begin{cases} 1, & x \le t, \\ \exp\left(-\frac{n(x-t)^2}{4t^2}\right), & t \le x \le 2t \\ \exp\left(-\frac{nx}{8t}\right), & 2t \le x. \end{cases} \tag{71}$$

Indeed, the first line in (71) is trivial. If $t \le x \le 2t$, then, applying (67) with $\mu = 1$ and $h = \frac{x-t}{t}$, we get the second line in (71). Further, equation (3.9) in the cited work (Bobkov, 2010) states that

$$\mathbb{P}[\|\boldsymbol{y}\| \ge h\mu] \le \exp\left(-\frac{nh}{8}\right), \quad h \ge 2.$$

Applying this with $\mu = 1$ and $h = \frac{x}{t}$, we get the third line in (71). With (71),

$$\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le t\right] \le \int_0^\infty q(x)\psi_{n,t}(x)dx = \int_0^\infty \mathbb{P}[\|\boldsymbol{x}\| \le x](-\psi'_{n,t}(x))dx,$$

where the last equality is integration by parts. Because $\psi_{n,t}(x)$ is non-increasing, $-\psi'_{n,t}(x)$ is non-negative, and $\mathbb{P}[\|\boldsymbol{x}\| \le x]$ can be bounded by

$$\mathbb{P}[\|\boldsymbol{x}\| \le x] \le g_n(x) := \begin{cases} \exp\left(-\frac{n(1-x)^2}{4}\right), & x \le 1, \\ 1, & x \ge 1, \end{cases} \tag{72}$$

where the first line in (72) follows from (68) with $\mu = 1$ and $h = 1 - x$. Hence,

$$\mathbb{P}\left[\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le t\right] \le \phi(t,n) := \int_0^\infty g_n(x)(-\psi'_{n,t}(x))dx.$$

Applying this bound to (44), we get

$$p \le f(n,\alpha) := \frac{\alpha}{B(\frac{n-1}{2}, \frac{1}{2})}\int_0^{1/\alpha}(1-\alpha^2t^2)^{\frac{n-3}{2}}\phi(t,n)dt, \tag{73}$$

and (70) follows from (16). □

The function $f(n,\alpha)$ in (70) is complicated but explicit and, for any specific values of n and α, can be easily computed in any package like Mathematica. In particular, we verified in Mathematica that

$$\frac{-\log f(n,1)}{n} \ge 0.14, \quad 1 \le n \le 4000.$$

This together with Theorem 18 implies the following Corollary.

Corollary 8. *Let $\delta > 0$, and let $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ be a set of M i.i.d. random points from a distribution in $\mathbb{R}^n$ given by (35) with log-concave ρ. If $1 \le n \le 4000$ and*

$$M < \sqrt{\delta}\exp(0.07n) \tag{74}$$

then the expected number of 1-inseparable pairs in set F is less than δ. In particular, (74) implies that set F is 1-Fisher separable with probability greater than $1 - \delta$.

If $n > 4000$, then we can use (69) and get the bound much higher than needed for any practical purposes. However, for smaller n, Corollary 8 is a significant improvement comparing to (69).

Example 12. Let $\alpha = 1$ and $\delta = 0.01$.

(a) If $n = 4001$, then (69) reduces to $M < 96,158,590,065,160,622,896,817$;
(b) If $n = 400$, then (69) reduces to $M \le 18$, while (74) reduces to $M \le 144,625,706,429$;
(c) If $n = 200$, (74) still gives a reasonable bound $M \le 120,260$.

Example 13. Let $\alpha = 1$ and $\delta = 0.01$. Table 9 shows the upper bounds on M in Corollary 8 in various dimensions n.

Corollary 7 demonstrates that constant 0.07 in Corollary 8 is within a factor less than 2 from being optimal.

Table 9
The upper bounds on M in Corollary 8 in various dimensions n for $\alpha = 1$ and $\delta = 0.01$.

n	$M \le$
10	0.2
50	3.3
100	109
200	120,260
500	$1.5 \cdot 10^{14}$
1000	$2.5 \cdot 10^{29}$

6. Improved bounds for product distributions in the unit cube

6.1. The general case

In this section we assume the following.

(a) all points in a finite set F are chosen independently;
(b) points in F are not necessary identically distributed, but have the same mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$;
(c) for each $\boldsymbol{x} = (x_1, \dots, x_n) \in F$, components $x_1, \dots, x_n$ are independent and have $[0, 1]$ support;
(d) there are no point $\boldsymbol{x} \in F$ such that $\mathbb{P}[\boldsymbol{x} = \boldsymbol{\mu}] = 1$.

From (c), F is a subset of the unit cube $U_n = [0,1]^n$. From (b), $E[x_i] = \mu_i$ for all $i = 1, \dots, n$ and for all $\boldsymbol{x} \in F$. Let

$$\sigma_0^2 = \min_{\boldsymbol{x}\in F}\left(\frac{1}{n}\sum_{i=1}^n Var[x_i]\right),$$

that is, the minimal value of average variance of the components. From (d), $\sigma_0^2 > 0$.

Fix any point $\boldsymbol{c} = (c_1, \dots, c_n) \in U_n$, and any pair $\boldsymbol{x}, \boldsymbol{y} \in F$. Let

$$z_i = (x_i - c_i)(y_i - c_i) - \alpha(x_i - c_i)^2, \quad i = 1, \dots, n. \tag{75}$$

Inequality (13) reduces to

$$\mathbb{P}\left[\sum_{i=1}^n z_i \ge 0\right] \le f(n,\alpha).$$

From (a) and (c) it follows that all random variables z_i are independent. Next,

$$E[z_i] = E[(x_i - c_i)(y_i - c_i)] - \alpha E[(x_i - c_i)^2].$$

By independence, $E[(x_i - c_i)(y_i - c_i)] = E[(x_i - c_i)]E[(y_i - c_i)] = (\mu_i - c_i)^2$, and $E[z_i] = (\mu_i - c_i)^2 - \alpha E[(x_i - c_i)^2] = (1-\alpha)(\mu_i - c_i)^2 - \alpha Var[x_i - c_i]$. Hence,

$$E\left[\sum_{i=1}^n z_i\right] = (1-\alpha)\sum_{i=1}^n(\mu_i - c_i)^2 - \alpha\sum_{i=1}^n Var[x_i] \le$$

$$\le (1-\alpha)\sum_{i=1}^n(\mu_i - c_i)^2 - n\alpha\sigma_0^2 = -nt,$$

where

$$t := \alpha\sigma_0^2 - (1-\alpha)\frac{1}{n}\sum_{i=1}^n(\mu_i - c_i)^2. \tag{76}$$

Note that t is guaranteed to be positive if either (i) α is sufficiently close to 1, or (ii) $\boldsymbol{c} = \boldsymbol{\mu}$.

The following Proposition established bounds on z_i.

Proposition 10. *Let $c_i' := \max\{c_i, 1 - c_i\}$ for all i, and $f(c) := -c + c^2(1-\alpha)$.*

(i) *if* $\alpha \geq 0.5$, *then*

$$-c_i' + (c_i')^2(1-\alpha) \leq z_i \leq \frac{(c_i')^2}{4\alpha}.$$

In particular, $-\alpha \leq z_i \leq \frac{1}{4\alpha}$ *for all i;*

(ii) *if* $\alpha \leq 0.5$, *then*

$$\min\{f(c_i), f(1-c_i)\} \leq z_i \leq (1-\alpha)(c_i')^2.$$

In particular, $-\frac{1}{4(1-\alpha)} \leq z_i \leq 1-\alpha$ *for all i;*

(iii) *if* $c_i = \frac{1}{2}$ *for all i, then* $-\frac{1}{2} + \frac{1}{4}(1-\alpha) \leq z_i \leq \frac{1}{16\alpha}$ *for all i if* $\alpha \geq 0.5$ *and* $-\frac{1}{2} + \frac{1}{4}(1-\alpha) \leq z_i \leq \frac{1}{4}(1-\alpha)$ *for all i if* $\alpha \leq 0.5$.

Proof. For each fixed c_i and y_i, z_i in (75) is maximized if $x_i = \frac{y_i - c_i}{2\alpha} + c_i$, resulting in $z_i = \frac{(y_i-c_i)^2}{4\alpha}$. The last expression is maximized if y_i is either 0 or 1, with maximum equal to $\frac{(c_i')^2}{4\alpha}$. This bound is tight if $\alpha \geq 0.5$. If $\alpha < 0.5$, then critical point $\frac{y_i-c_i}{2\alpha} + c_i$ lies outside of $[0, 1]$ and (75) is maximized if x_i is either 0 or 1, resulting in bound $(1-\alpha)(c_i')^2$.

Similarly, z_i in (75) is minimized when either $x_i = 1$ and $y_i = 0$ or vice versa, resulting in bound $\min\{f(c_i), f(1-c_i)\} \leq z_i$. Because $f(c) \geq -\frac{1}{4(1-\alpha)}$ for all c, bound $-\frac{1}{4(1-\alpha)} \leq z_i$ follows. If $\alpha \geq 0.5$, then f is monotone decreasing on $[0, 1]$, hence $\min\{f(c_i), f(1-c_i)\} = f(c_i') = -c_i' + (c_i')^2(1-\alpha) \geq f(1) = -\alpha$. □

Let $S_n = \sum_{i=1}^n z_i$. By Hoeffding's inequality (Hoeffding, 1963), (Boucheron et al., 2013, Theorem 2.8),

$$\mathbb{P}[S_n \geq 0] \leq \mathbb{P}[S_n - E[S_n] \geq nt] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

provided that $t > 0$, where $[a_i, b_i]$ is the support of random variable z_i. Applying Proposition 10 to bound $b_i - a_i$, we get the following result.

Theorem 19. *Assume that (a)–(d) hold. Let* $\delta > 0$, $\alpha \in (0, 1]$, *and let* $\boldsymbol{c}$ *be an arbitrary point inside unit cube* $[0, 1]^n$ *such that t in (76) is positive. Let* $c_i' := \max\{c_i, 1-c_i\}$ *and* $g(c) := \min\{-c + c^2(1-\alpha), -(1-c) + (1-c)^2(1-\alpha)\}$. *If* $\alpha \geq 0.5$ *and*

$$M < \sqrt{\delta}\exp\left(\frac{n^2t^2}{\sum_{i=1}^n \left(c_i' - (c_i')^2(1-\alpha) + \frac{(c_i')^2}{4\alpha}\right)^2}\right),$$

or $\alpha \leq 0.5$ *and*

$$M < \sqrt{\delta}\exp\left(\frac{n^2t^2}{\sum_{i=1}^n \left((1-\alpha)(c_i')^2 - g(c_i)\right)^2}\right),$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *is* $(\alpha, \boldsymbol{c})$*-Fisher separable with probability greater than* $1 - \delta$.

For $\alpha = 1$, we get the following corollary

Corollary 9. *Assume that (a)–(d) hold. Let* $\delta > 0$, *and let* $\boldsymbol{c}$ *be an arbitrary point inside unit cube* $[0, 1]^n$. *If*

$$M < \sqrt{\delta}\exp\left(\frac{16}{25}n\sigma_0^4\right), \tag{77}$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *is* $(1, \boldsymbol{c})$*-Fisher separable with probability greater than* $1 - \delta$.

Example 14. With $\delta = 0.01$, $n = 500$, and $\sigma_0 = 0.5$, (same values as in Example 6) (77) reduces to $M < 48, 516, 519$.

By selecting $\boldsymbol{c}$ being the center of the cube, we can improve the bound further.

Corollary 10. *Assume that (a)–(d) hold. Let* $\delta > 0$, *and let* $\boldsymbol{c} = (\frac{1}{2}, \ldots, \frac{1}{2})$ *be the center of unit cube* $[0, 1]^n$. *If*

$$M < \sqrt{\delta}\exp\left(\frac{256}{81}n\sigma_0^4\right), \tag{78}$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *is* $(1, \boldsymbol{c})$*-Fisher separable with probability greater than* $1 - \delta$.

In contrast to (77), bound (78) may be practical in dimension $n = 100$.

Example 15. If $\delta = 0.01$ and $n = 100$, then, even with maximal possible $\sigma_0 = 0.5$, (77) reduces to $M < 5.5$. In contrast, (78) with these parameters gives $M < 37, 901, 503$.

In larger dimensions, bound (78) may be practical for (slightly) lower σ_0.

Example 16. If $\delta = 0.01$, and $\sigma_0 = \frac{1}{2\sqrt{3}}$ (the standard deviation of uniform distribution on $[0, 1]$), then, even with $n = 1000$, (77) reduces to $M < 8.5$. In contrast, (78) with these parameters gives $M < 340, 283, 178$.

If $\alpha < 1$, it is convenient to apply Theorem 19 with $\boldsymbol{c} = \boldsymbol{\mu}$. In this case t in (76) is guaranteed to be positive, and bounds in Proposition 10(i),(ii) imply the following result.

Corollary 11. *Assume that (a)–(d) hold. Let* $\delta > 0$, $\alpha \in (0, 1]$. *If*

$$M < \sqrt{\delta}\exp\left(\frac{16\alpha^4}{(1+4\alpha^2)^2}n\sigma_0^4\right), \quad \alpha \geq 0.5,$$

or

$$M < \sqrt{\delta}\exp\left(\frac{16(1-\alpha)^2\alpha^2}{(1+4(1-\alpha)^2)^2}n\sigma_0^4\right), \quad \alpha \leq 0.5,$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *is* $(\alpha, \boldsymbol{\mu})$*-Fisher separable with probability greater than* $1 - \delta$.

Example 17. With $\delta = 0.01$, $n = 500$, and $\sigma_0 = 0.5$, and $\alpha = 0.9$, Corollary 11 is applicable if $M < 8, 411, 607$.

6.2. The mean-centered distributions

In this section we consider a special case when $\boldsymbol{\mu} = (\frac{1}{2}, \ldots, \frac{1}{2})$ is the center of the unit cube. In this case, Theorem 19 with Proposition 10 (iii) implies that set F is $(\alpha, \boldsymbol{\mu})$-Fisher separable with probability greater than $1 - \delta$ provided that

$$M < \sqrt{\delta}\exp\left(\frac{256\alpha^4}{(1+2\alpha)^4}n\sigma_0^4\right), \quad \alpha \geq 0.5, \tag{79}$$

or

$$M < \sqrt{\delta}\exp\left(4\alpha^2 n\sigma_0^4\right), \quad \alpha \leq 0.5.$$

With $\alpha = 1$, (79) reduces to (78). It is practical if σ_0 is close to its maximal value 0.5, but, because of factor σ_0^4, quickly becomes useless if σ_0 decreases.

Example 18. If $\delta = 0.01$, and $\sigma_0 = 0.2$, then, even with $n = 1000$, (78) reduces to $M < 15.7$.

The theorem below uses Bernstein inequality to derive an alternative bound with better dependence of σ_0.

Theorem 20. *Assume that (a)–(d) hold, and assume that* $\boldsymbol{\mu} = (\frac{1}{2}, \ldots, \frac{1}{2})$ *is the center of unit cube* $[0, 1]^n$. *For any* $\delta > 0$, $\alpha \in (0, 1]$, *if*

$$M < \sqrt{\delta}\exp\left(\frac{12\alpha^2}{12\alpha^2 + 13}n\sigma_0^2\right), \quad \alpha \geq 0.5,$$

or

$$M < \sqrt{\delta}\exp\left(\frac{3\alpha^2}{2\alpha^2+\alpha+3}n\sigma_0^2\right), \quad \alpha \le 0.5,$$

then set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ *is* $(\alpha, \boldsymbol{\mu})$*-Fisher separable with probability greater than* $1-\delta$.

Proof. Bernstein inequality (Boucheron et al., 2013, p. 36) states that, if $S_n = \sum_{i=1}^n z_i$ is the sum of independent random variables with finite variance such that $z_i \le b$ for some $b > 0$ with probability 1 for all $i = 1, 2, \dots, n$, then, for any $T > 0$,

$$\mathbb{P}[S_n - E[S_n] \ge T] \le \exp\left(-\frac{T^2}{2(v+bT/3)}\right), \tag{80}$$

where $v = \sum_{i=1}^n E[z_i^2]$. With z_i given by (75), $c_i = \mu_i = 1/2$, and notation $\bar{x}_i = x_i - 1/2$, $\bar{y}_i = y_i - 1/2$,

$$E[z_i] = E[\bar{x}_i]E[\bar{y}_i] - \alpha E[\bar{x}_i^2] = -\alpha E[\bar{x}_i^2].$$

Let $\sigma_i^2 = E[\bar{x}_i^2]$ be the variance of x_i, and $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ be the average variance of the components of $\boldsymbol{x}$. Then $E[S_n] = \sum_{i=1}^n E[z_i] = -n\alpha\sigma_x^2$. Also,

$$E[z_i^2] = E[\bar{x}_i^2]E[\bar{y}_i^2] - 2\alpha E[\bar{x}_i^3]E[\bar{y}_i] + \alpha^2 E[\bar{x}_i^4].$$

Because $\bar{y}_i$ has support $[-1/2, 1/2]$, $E[\bar{y}_i] = 0$, and $f(y) = y^2$ is a convex function, $E[\bar{y}_i^2]$ is maximal if $\bar{y}_i$ takes values $\pm 1/2$ with equal chances, and $E[\bar{y}_i^2] \le 1/4$. Next, denoting $u_i = \bar{x}_i^2$, we note that $E[u_i] = \sigma_i^2$, support of u_i is $[0, 1/4]$, hence $E[u_i^2]$ is maximal if u_i takes values 0 and 1/4 with probabilities $1 - 4\sigma_i^2$ and $4\sigma_i^2$, respectively. Hence, $E[\bar{x}_i^4] = E[u_i^2] \le (1/4)^2 4\sigma_i^2 = \sigma_i^2/4$. This implies that $E[z_i^2] \le \sigma_i^2(1/4) + \alpha^2(\sigma_i^2/4) = (1+\alpha^2)\sigma_i^2/4$. Hence, $v = \sum_{i=1}^n E[z_i^2] \le \frac{1+\alpha^2}{4}\sum_{i=1}^n \sigma_i^2 = \frac{1+\alpha^2}{4}n\sigma_x^2$.

By Proposition 10(iii), $z_i \le b$ for all i, where $b = \frac{1}{16\alpha}$ if $\alpha \ge 0.5$ and $b = \frac{1}{4}(1-\alpha)$ if $\alpha \le 0.5$.

Hence, for $\alpha \ge 0.5$, (80) implies that

$$\mathbb{P}[S_n \ge 0] = \mathbb{P}[S_n - E[S_n] \ge n\alpha\sigma_x^2] \le$$

$$\le \exp\left(-\frac{(n\alpha\sigma_x^2)^2}{2((1+\alpha^2)n\sigma_x^2/4 + n\alpha\sigma_x^2/48\alpha)}\right) =$$

$$= \exp\left(\frac{-24\alpha^2}{12\alpha^2+13}n\sigma_x^2\right) \le \exp\left(\frac{-24\alpha^2}{12\alpha^2+13}n\sigma_0^2\right).$$

For $\alpha \ge 0.5$, similar calculation gives

$$\mathbb{P}[S_n \ge 0] \le \exp\left(\frac{-6\alpha^2}{2\alpha^2+\alpha+3}n\sigma_0^2\right)$$

Combining these bounds with (16), we obtain the desired result. □

For $\alpha = 1$, this gives the following corollary.

Corollary 12. *Assume that (a)–(d) hold, and assume that* $\boldsymbol{\mu} = (\frac{1}{2}, \dots, \frac{1}{2})$ *is the center of unit cube* $[0,1]^n$. *For any* $\delta > 0$, *if*

$$M < \sqrt{\delta}\exp\left(\frac{12}{25}n\sigma_0^2\right) \tag{81}$$

then set $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ *is* $(1, \boldsymbol{\mu})$*-Fisher separable with probability greater than* $1-\delta$.

This bound is better than (78), provided that $\frac{12}{25}n\sigma_0^2 > \frac{256}{81}n\sigma_0^4$, or $\sigma_0 < \frac{9\sqrt{3}}{40} \approx 0.39$.

Example 19. If $\delta = 0.01$, $\sigma_0 = 0.2$, and $n = 1000$ (the same parameters as in Example 18) (81) reduces to $M < 21,799,877$.

How close these bounds to being optimal? If each point in F is distributed uniformly among vertices of the cube, then 2 points $\boldsymbol{x}$ and $\boldsymbol{y}$ are not Fisher separable if and only if they coincide, which may happen with probability 2^{-n}. Hence, Fisher separability of a set of M points holds with probability $1-\delta$ for

$$M \approx \sqrt{\frac{\delta}{2^{-n}}} = \sqrt{\delta}\exp\left(\frac{\log 2}{2}n\right) \approx \sqrt{\delta}\exp(0.35n).$$

In this example, $\sigma_0 = 0.5$, and (78) gives bound $\sqrt{\delta}\exp\left(\frac{16}{81}n\right) \approx \sqrt{\delta}\exp(0.2n)$. Note that the coefficients in these estimates differ less than by the factor of 2.

Corollary 12 follows from two-point bound (13) with $f(n,\alpha) = \exp\left(-\frac{24}{25}n\sigma_0^2\right)$. Can we significantly improve the constant here, or the dependence from σ_0? Consider two points $\boldsymbol{x}$ and $\boldsymbol{y}$, such that all components x_i of $\boldsymbol{x}$ take values 0, 1/2, 1 with probabilities $2\sigma_0^2$, $1-4\sigma_0^2$, $2\sigma_0^2$, respectively, and all components y_i of $\boldsymbol{y}$ take values 0,1 with equal chances. Then z_i given by (75) take values $-1/2$ and 0 with probabilities $2\sigma_0^2$ and $1-2\sigma_0^2$, respectively, hence $\boldsymbol{x}$ and $\boldsymbol{y}$ are not Fisher separable with probability $p_n = (1-2\sigma_0^2)^n$. For small σ_0, $p_n \approx \exp(-2n\sigma_0^2)$, hence the quadratic dependence on σ_0 cannot be improved, and the coefficient $\frac{24}{25}$ cannot be improved to any value higher than 2.

6.3. Better bounds if the product distribution is known

The results in Sections 6.1 and 6.2 are valid for the whole family of product distributions satisfying certain conditions. This Section studies the case when the data distribution is explicitly given. In this case, we can deduce improved estimates from Chernoff's inequality. Our first result is for general product distributions, not necessary bounded in the unit cube.

Theorem 21. *Let points* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ *be i.i.d points from an arbitrary but explicitly given product distribution* G *in* $\mathbb{R}^n$. *For any* $\delta > 0$, *let*

$$M < \sqrt{\delta}\exp(\gamma_n), \tag{82}$$

where

$$\gamma_n = \frac{1}{2}\sup_{\lambda \ge 0}\left(-\sum_{i=1}^n \log E\left[e^{\lambda(x_i y_i - \alpha x_i^2)}\right]\right), \tag{78}$$

where x_i *and* y_i *are independent random variables distributed as* i*th component of* G. *Then set* $F = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ *is* $(\alpha, \boldsymbol{0})$*-Fisher separable with probability greater than* $1-\delta$.

Proof. With $\boldsymbol{c} = \boldsymbol{0}$, (75) simplifies to

$$z_i = x_i y_i - \alpha x_i^2, \quad i = 1, \dots, n,$$

where x_i and y_i are independent and distributed as ith component of G. Points $\boldsymbol{x}$ and $\boldsymbol{y}$ are not Fisher separable if $S \ge 0$, where $S = \sum_{i=1}^n z_i$.

Chernoff's inequality (Boucheron et al., 2013, p. 21) states that, for any random variable S, and any real number t,

$$\mathbb{P}[S \ge t] \le \exp[-\psi_S^*(t)],$$

where

$$\psi_S^*(t) = \sup_{\lambda \ge 0}(\lambda t - \psi_S(\lambda)),$$

where $\psi_S(\lambda) = \log(E[e^{\lambda S}])$. If $S = \sum_{i=1}^n z_i$ for independent random variables z_i,

$$e^{\psi_S(\lambda)} = E[e^{\lambda\sum_{i=1}^n z_i}] = E\left[\prod_{i=1}^n e^{\lambda z_i}\right] = \prod_{i=1}^n E[e^{\lambda z_i}],$$

hence

$$\psi_S(\lambda) = \log\left(\prod_{i=1}^{n} E[e^{\lambda z_i}]\right) = \sum_{i=1}^{n} \log E\left[e^{\lambda(x_i y_i - \alpha x_i^2)}\right],$$

and $\psi_S^*(0) = 2\gamma_n$. Hence, Chernoff's inequality with $t = 0$ implies that

$$\mathbb{P}[S \geq 0] \leq \exp[-\psi_S^*(0)] = \exp[-2\gamma_n],$$

and (82) follows from (16). □

Corollary 13. *If all components of G in Theorem 21 have the same distribution, then estimate (82) simplifies to*

$$M < \sqrt{\delta}\exp(\gamma n), \tag{83}$$

where

$$\gamma = \frac{1}{2}\sup_{\lambda\geq 0}\left(-\log E\left[e^{\lambda(xy-\alpha x^2)}\right]\right),$$

where x and y are independent random variables distributed as a component of G. In particular, if the component distribution has density f, then

$$\gamma = \frac{1}{2}\sup_{\lambda\geq 0}\left(-\log\left[\int_{\mathbb{R}^2} e^{\lambda(xy-\alpha x^2)} f(x)f(y)dxdy\right]\right).$$

It follows from the proof of Theorem 21 and Cramer's theorem (Pham, 2007, Theorem 2.1) that the exponent γ in (83) is the best possible. However, estimate (83) maybe non-optimal in lower order terms. Below we give a formula for the asymptotically best possible upper bound for M in Corollary 13.

Let λ^* be the (unique) minimizer of $E\left[e^{\lambda(xy-\alpha x^2)}\right]$, and let

$$c^* := \frac{d}{d\lambda}\left(\frac{E\left[(xy-\alpha x^2)e^{\lambda(xy-\alpha x^2)}\right]}{E\left[e^{\lambda(xy-\alpha x^2)}\right]}\right)\Bigg|_{\lambda=\lambda^*}$$

The exact asymptotic growth of the probability $\mathbb{P}[S \geq 0]$ in Theorem 21 is given by Petrov (1965, Theorem 1)

$$\mathbb{P}[S \geq 0] = \frac{\exp[-2\gamma n]}{\lambda^*\sqrt{c^*}\sqrt{2\pi n}}(1+o(1)),$$

hence the exact asymptotic estimate for M in Corollary 13 is

$$M < \sqrt{\delta}\sqrt{\lambda^*}\sqrt[4]{2\pi c^* n}\exp(\gamma n)(1+o(1)).$$

We can see that estimate (83) differs from the optimal one by $\sqrt{\lambda^*}\sqrt[4]{2\pi c^* n}$ term. However, the advantages of estimate (83) are simplicity and the absence of $(1+o(1))$ term.

We now apply Corollary 13 to some special cases. First, Corollary 13 specialized to the standard normal distribution implies Theorem 12. As another example, we apply Corollary 13 to the uniform distribution in a cube.

Corollary 14. *Let $\alpha = 1$ and points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ are i.i.d points from uniform distribution in a cube with center $\boldsymbol{\mu}$. For any $\delta > 0$, if*

$$M < \sqrt{\delta}\exp(\gamma n), \tag{84}$$

where

$$\gamma = \frac{1}{2}\sup_{\lambda\geq 0}\left(-\log\left[\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{\lambda(xy-x^2)}dxdy\right]\right) = 0.23319\ldots$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is $(1, \boldsymbol{\mu})$-Fisher separable with probability greater than $1-\delta$.

For the unit cube, $\sigma_0^2 = \frac{1}{12}$, and Corollary 12 implies $(1, \boldsymbol{\mu})$-Fisher separability with probability greater than $1-\delta$ provided that

$$M < \sqrt{\delta}\exp\left(\frac{n}{25}\right) \tag{85}$$

Table 10
The upper bounds on M (85) in Corollary 14 in various dimensions n for $\alpha = 1$ and $\delta = 0.01$.

n	$M \leq$
10	1.02
50	11,578
100	$1.3 \cdot 10^9$
200	$1.7 \cdot 10^{19}$
500	$4.3 \cdot 10^{49}$
1000	$1.8 \cdot 10^{100}$

We can see that (84) is a substantial improvement over (85). This is because (84) works for uniform distribution only, while (85) works for *any* product distribution in the unit cube with $\sigma_0^2 = \frac{1}{12}$.

Example 20. Let $\alpha = 1$ and $\delta = 0.01$. Table 10 shows the upper bounds on M in Corollary 14 in various dimensions n. For example, for $n = 100$, we see that over a billion points from the uniform distribution in the unit cube are 1-Fisher-separable with probability greater than 99%.

7. Fisher separability for dependent data from product distribution

The key assumption in Section 6 is that all points in set F are chosen independently. This section establishes a sufficient condition for Fisher separability with high probability in a datasets with dependent data points, as soon as the corresponding conditional distributions are product distributions in the unit cube $U_n = [0,1]^n$.

Formally, we assume the following.

(*) For any $\boldsymbol{x} \in F$ and $\boldsymbol{y} \in F$, and any $\boldsymbol{y}_0 \in U_n$, the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y} = \boldsymbol{y}_0$ is a product distribution with support in U_n.

For every $\boldsymbol{x} \in F, \boldsymbol{y} \in F, \boldsymbol{y}_0 \in U_n$ and index $i \in 1, 2, \ldots, n$, let $\sigma_i^2(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}_0)$ be the variance of the conditional distribution of the ith component of $\boldsymbol{x}$ given $\boldsymbol{y} = \boldsymbol{y}_0$. Let

$$\sigma_0^2 = \min_{\boldsymbol{x}\in F, \boldsymbol{y}\in F, \boldsymbol{y}_0\in U_n}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}_0)\right)$$

be the minimal value of average variance of the components of such conditional distribution. Also, let $\boldsymbol{c}^* = (1/2, \ldots, 1/2)$ be the center of U_n.

Theorem 22. *Assume that (*) holds. For any $\delta > 0$, $\alpha \in (0, 1]$, if*

$$\sigma_0^2 > \frac{1}{16\alpha^2}$$

and

$$M < \sqrt{\delta}\exp\left(\frac{256\alpha^4}{(1+2\alpha)^4}\left(\sigma_0^2 - \frac{1}{16\alpha^2}\right)^2 n\right)$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is $(\alpha, \boldsymbol{c}^*)$-Fisher separable with probability greater than $1-\delta$.

Proof. By (16), the statement of the theorem follows from (13) with $\boldsymbol{c} = \boldsymbol{c}^*$ and

$$f(n,\alpha) = \exp\left(-2\left(\frac{4\alpha}{2\alpha+1}\right)^4\left(\sigma_0^2 - \frac{1}{16\alpha^2}\right)^2 n\right).$$

We will show that for any $\boldsymbol{x} \in F, \boldsymbol{y} \in F$, and $\boldsymbol{y}_0 \in U_n$

$$\mathbb{P}[\alpha(\boldsymbol{x}-\boldsymbol{c}^*, \boldsymbol{x}-\boldsymbol{c}^*) \leq (\boldsymbol{x}-\boldsymbol{c}^*, \boldsymbol{y}-\boldsymbol{c}^*) \mid \boldsymbol{y} = \boldsymbol{y}_0] \leq f(n,\alpha), \tag{86}$$

which would imply (13) and finish the proof. The set of all $\boldsymbol{x} \in \mathbb{R}^n$ which does not satisfy the inequality $\alpha(\boldsymbol{x}-\boldsymbol{c}^*, \boldsymbol{x}-\boldsymbol{c}^*) \le (\boldsymbol{x}-\boldsymbol{c}^*, \boldsymbol{y}-\boldsymbol{c}^*)$ is the ball with center $\boldsymbol{c} = \boldsymbol{c}^* + (\boldsymbol{y}_0 - \boldsymbol{c}^*)/2\alpha$ and radius $r = \|\boldsymbol{y}_0 - \boldsymbol{c}^*\|/2\alpha$, see Gorban et al. (2018). Because $\boldsymbol{y}_0 \in U_n$, $r^2 = \|\boldsymbol{y}_0 - \boldsymbol{c}^*\|^2/4\alpha^2 \le n(1/2)^2/4\alpha^2 = n/16\alpha^2$, and (86) would follow from

$$\mathbb{P}\left[(\boldsymbol{x}-\boldsymbol{c}, \boldsymbol{x}-\boldsymbol{c}) \le \frac{n}{16\alpha^2} \mid \boldsymbol{y} = \boldsymbol{y}_0\right] \le f(n, \alpha). \tag{87}$$

Let x_i be the random variable whose distribution is the conditional distribution of the ith component of $\boldsymbol{x}$ given $\boldsymbol{y} = \boldsymbol{y}_0$. Let $z_i = -(x_i - c_i)^2$, where c_i is the ith component of $\boldsymbol{c}$. Then

$$E[z_i] = -E[(x_i - c_i)^2] \le -E[(x_i - E[x_i])^2] = -\sigma_i^2(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}_0),$$

and

$$\sum_{i=1}^{n} E[z_i] \le -\sum_{i=1}^{n} \sigma_i^2(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}_0) \le -n\sigma_0^2.$$

Hence,

$$\mathbb{P}\left[(\boldsymbol{x}-\boldsymbol{c}, \boldsymbol{x}-\boldsymbol{c}) \le \frac{n}{16\alpha^2} \mid \boldsymbol{y} = \boldsymbol{y}_0\right] = \mathbb{P}\left[\sum_{i=1}^{n} z_i \ge -\frac{n}{16\alpha^2}\right]$$

$$\le \mathbb{P}\left[\sum_{i=1}^{n} (z_i - E[z_i]) \ge -\frac{n}{16\alpha^2} + n\sigma_0^2\right].$$

In fact, $c_i = 1/2 + (y_i^0 - 1/2)/2\alpha$, where y_i^0 is the ith component of $\boldsymbol{y}_0$. Because $0 \le y_i^0 \le 1$, we get $1/2 - 1/4\alpha \le c_i \le 1/2 + 1/4\alpha$, hence $-(1/2 + 1/4\alpha)^2 \le z_i \le 0$. By Hoeffding's inequality (Boucheron et al., 2013, Theorem 2.8),

$$\mathbb{P}\left[\sum_{i=1}^{n} (z_i - E[z_i]) \ge t\right] \le \exp\left(-\frac{2t^2}{n(1/2 + 1/4\alpha)^4}\right).$$

With $t = -\frac{n}{16\alpha^2} + n\sigma_0^2$, this proves (87). □

We remark that because $\sigma_0 \le 0.5$, the bound $\sigma_0^2 > \frac{1}{16\alpha^2}$ in Theorem 22 may hold only if $\alpha > 1/2$.

For $\alpha = 1$, we have the following corollary.

Corollary 15. *Assume that (*) holds. For any $\delta > 0$, if*

$$\sigma_0^2 > \frac{1}{16}$$

and

$$M < \sqrt{\delta} \exp\left(\frac{256}{81}\left(\sigma_0^2 - \frac{1}{16}\right)^2 n\right) \tag{88}$$

then set $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is $(1, \boldsymbol{c}^*)$-Fisher separable with probability greater than $1 - \delta$.

Example 21. Let $\alpha = 1$ and $\delta = 0.01$. Table 11 shows the upper bounds on M in Corollary 15 for $\sigma_0 = 0.4$, 0.45, and 0.5 in various dimensions n.

For example, for $n = 500$ and $\sigma_0 = 0.4$, we see that over 300,000 points are Fisher-separable with probability greater than 99%.

Corollary 15 is not applicable if $\sigma_0^2 \le 1/16$. However, this is unavoidable. Indeed, let set F contain points $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $\boldsymbol{y}$ is uniformly distributed among the vertices of the unit cube, and $\boldsymbol{x}$ is uniformly distributed among the vertices of the (twice smaller) cube with main diagonal connecting $\boldsymbol{c}^*$ and $\boldsymbol{y}$. Then the variance of the components of $\boldsymbol{x}$ is 1/16, but $\boldsymbol{x}$ and $\boldsymbol{y}$ are not $(1, \boldsymbol{c}^*)$-Fisher separable with probability 1.

Table 11
The upper bounds on M in Corollary 15 for $\alpha = 1$ and $\delta = 0.01$, various σ_0 and dimensions.

	$\sigma_0 = 0.4$	$\sigma_0 = 0.45$	$\sigma_0 = 0.5$
$n = 10$	0.13	0.18	0.3
$n = 50$	0.44	2.21	25
$n = 100$	2	49	6691
$n = 200$	40	24,017	$4.4 \cdot 10^8$
$n = 500$	334,248	$2.8 \cdot 10^{12}$	$1.3 \cdot 10^{23}$
$n = 1000$	$1.1 \cdot 10^{12}$	$8 \cdot 10^{25}$	$1.7 \cdot 10^{47}$

8. Summary: a short guide on proven theorems

We established new stochastic separation theorems for a broad class of log-concave and product distributions. All the theorems state that if the number of points M does not exceed some bound M_0, then the points are Fisher separable with high probability. In all theorems, the bound M_0 grows exponentially in dimension n. The exact rate of growth of M_0 depends on the distribution assumptions we impose. If we make stronger assumptions, we can prove theorems with faster-growing upper bound M_0, and can ensure separation of more points.

We can get separation theorems with the fastest-growing bound M_0 if we assume that the data are i.i.d. and are taken from a fixed given distribution such as the standard normal distribution (Theorems 12 and 13), uniform distributions in a ball (Theorem 15) or in a unit cube (Corollary 14), or multivariate exponential distribution (Theorem 16).

More generally, we have established new separation theorems for i.i.d. data from any fixed given distribution f, assuming that f is either rotation invariant (Theorem 14) or a product distribution (Theorem 21 and Corollary 13).

In the Theorems listed above, the distribution f is assumed to be known and the bound M_0 explicitly depend on f. More generally, we may assume that distribution f is unknown but is known to belong to some family $\mathcal{F}$ of distributions. In this case, the bound M_0 should depend on $\mathcal{F}$ but not on f. We have proved such separation theorems for i.i.d. data from (unknown) product distribution (Theorems 19 and 20), rotation invariant distribution (Theorems 17 and 18), isotropic strongly log-concave distribution (Theorems 8 and 9), and, more generally, any mixture of strongly log-concave distributions (Theorem 11). This last theorem is very general, because any distribution with exponentially decaying tails may be approximated by a mixture of log-concave ones.

Finally, we have Theorems with i.i.d. assumption relaxed. In particular, in Theorem 1 the probability of separability of a random point from a finite set was estimated without any assumption about the randomness and distributions of this finite set. Theorem 10 treats the case when the data are independent but not identically distributed, and their distributions are strongly log-concave but not isotropic. Theorem 22 treats the case when the data may be dependent but the conditional distributions are product distributions.

The results are illustrated in Figs. 3–8.

9. Conclusion: what are these estimates for?

The theorems presented in the paper have, roughly speaking, the following structure: for a given class of distributions, a random set of M vectors in $\mathbb{R}^n$ is α-Fisher separable with probability $\ge p$ if $M \le M_0$, where M_0 depends on n, p, and α and this dependence is specific for the selected class of probability distributions. For the distributions without heavy tails and "clumps" (sets with relatively low volume but high probability) M_0 grows fast with n: exponentially for strictly log-concave distributions (tails that decay as $\exp(-a\|\boldsymbol{x}\|^2)$ or faster) and as exponent of $\sqrt{n}$ for general

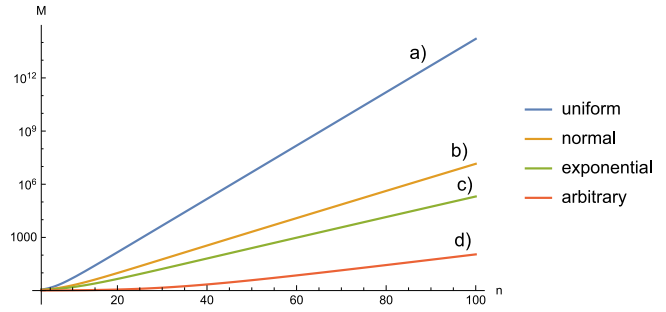

Fig. 3. The number M of points which are guaranteed to be 1-Fisher separable with probability 99% as a function of dimension n for (a) the uniform distribution in a ball (Corollary 4), (b) the standard normal distribution (Theorem 13), (c) multivariate exponential distribution (Theorem 16), and (d) lower bound for M which works for an arbitrary log-concave rotation-invariant distribution (Theorem 18).

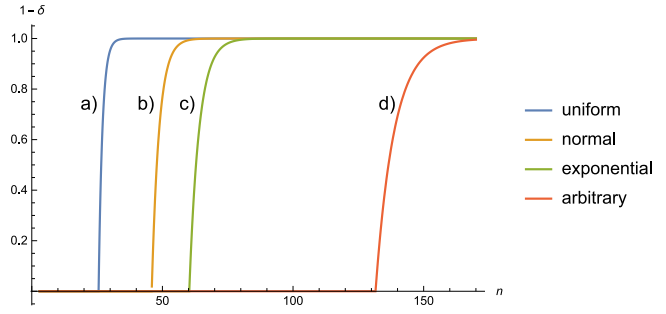

Fig. 4. The lower bound $1-\delta$ for the probability that the set of $M = 10{,}000$ points is 1-Fisher separable as a function of dimension n for (a) the uniform distribution in a ball (Corollary 4), (b) the standard normal distribution (Theorem 13), (c) multivariate exponential distribution (Theorem 16), and (d) an arbitrary log-concave rotation-invariant distribution (Theorem 18).

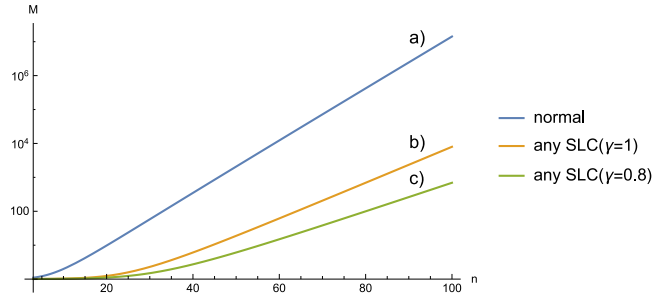

Fig. 5. The number M of points which are guaranteed to be 1-Fisher separable with probability 99% as a function of dimension n for (a) the standard normal distribution (Theorem 13), (b) an arbitrary strictly log-concave distribution with $\gamma = 1$ (Theorem 9), and (c) an arbitrary strictly log-concave distribution with $\gamma = 0.8$ (Theorem 9). Recall that the standard normal distribution is strictly log-concave with $\gamma = 1$.

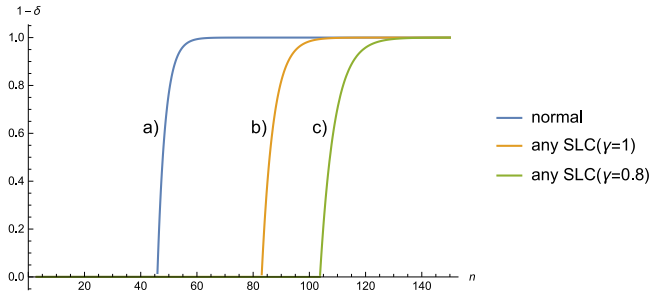

Fig. 6. The lower bound $1-\delta$ for the probability that the set of $M = 10{,}000$ points is 1-Fisher separable as a function of dimension n for (a) the standard normal distribution (Theorem 9), (b) arbitrary strictly log-concave distribution with $\gamma = 1$ (Theorem 9), and (c) arbitrary strictly log-concave distribution with $\gamma = 0.8$ (Theorem 9).

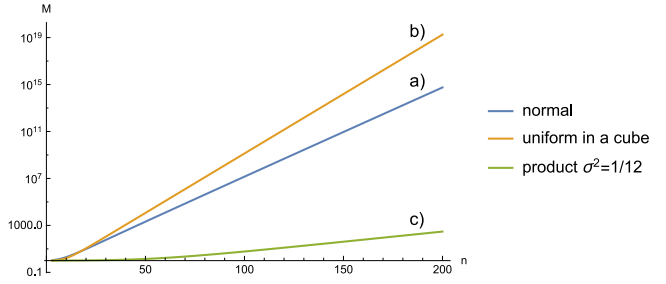

Fig. 7. The number M of points which are guaranteed to be 1-Fisher separable with probability 99% as a function of dimension n for (a) the standard normal distribution (Theorem 9), (b) the uniform distribution in a cube (Corollary 14) and (c) an arbitrary mean-centered product distribution with $\sigma^2 = \frac{1}{12}$ (Corollary 12). Recall that the uniform distribution in a cube has $\sigma^2 = \frac{1}{12}$.

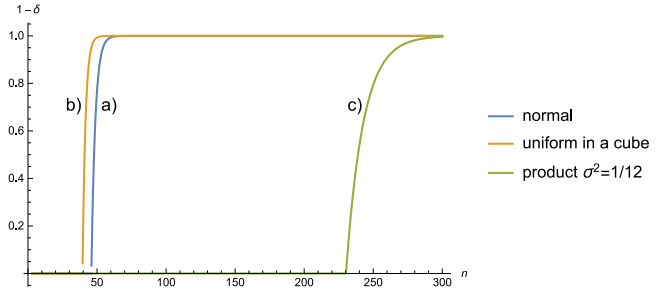

Fig. 8. The lower bound $1-\delta$ for the probability that the set of $M = 10{,}000$ points is 1-Fisher separable as a function of dimension n for (a) the standard normal distribution (Theorem 9), (b) the uniform distribution in a cube (Corollary 14) and (c) an arbitrary product distribution with $\sigma^2 = \frac{1}{12}$ (Corollary 12).

log-concave distributions (which may have exponential tails that decay as $\exp(-a\|\boldsymbol{x}\|)$). The main problem solved in the work was to find the best (optimal and explicit) estimates.

Stochastic separation theorems form a relatively new chapter of the measure concentration theory (for the collection of the classical results about concentration of measure we refer to Giannopoulos and Milman (2000), Ledoux (2001) and Vershynin (2018)). Concentration of random sets in thin shells is well-known: equivalence of microcanonical and canonical ensembles in statistical physics due to concentration near the level sets of energy (Gibbs, 1960), concentration of the volume of a ball near its border, the sphere, and concentration of the sphere near its equators (Ball, 1997; Lévy, 1951) (and general 'waist concentration' (Gromov, 2003)), etc. Stochastic separation theorems describe the fine structure of this thin layer.

The first theorems of this class were considered as the manifestation of the blessing of dimensionality (Gorban & Tyukin, 2018; Gorban, Tyukin and Romanenko, 2016). Indeed, the fast and non-iterative correction of the AI errors is based on the phenomenon of stochastic separation in high dimensions. The legacy AI systems are supplemented by *correctors*. These simple smart devices separate recognized errors and their surroundings from situations with correct functioning and replace the legacy AI solution with the corrected one. One of the possible structures of correcting system is presented in Fig. 9. The correcting system receives a vector of signals that represents the situation in maximal detail. It consists of input vectors of

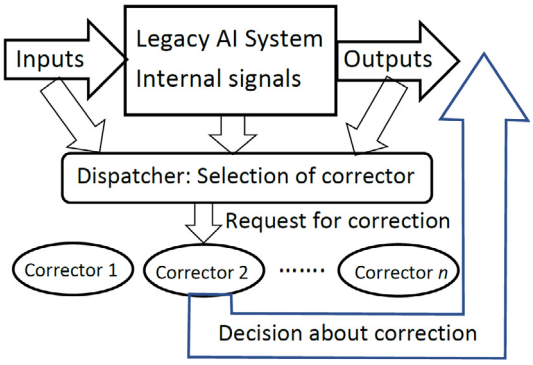

Fig. 9. Corrector of a legacy AI system.

the legacy AI system, vector of internal signals of that and the output vector (Fig. 9). There are several elementary correctors (Corrector 1, Corrector 2, …Corrector *n* in Fig. 9). Each elementary corrector includes a classifier, which separates a cluster of recognized errors from all other situations, and keeps the modified decision rule for this cluster. Dispatcher selects for each situation the closest cluster and sends the vector that represents the situation to the corresponding elementary corrector for further decision. The elementary corrector takes the decision "an error or not an error" and acts according to this decision. Stochastic separation theorems are necessary to evaluate the probability of accurate work of such a system. Of course, its accuracy increases with dimensionality of data. Correctors can be used for solution of the classical problem of sensitivity and specificity improvement (removing false-positive and false-negative results of classification), for knowledge transfer between artificial intelligence systems (Tyukin et al., 2018), for training of multiagent systems and other purposes.

If the AI system works for a long time, then errors and their correctors accumulate. The 'technical debt' increases, and flexibility drops down (Sculley et al., 2015). In this situation, the Interiorization of the accumulated knowledge is necessary. This is incorporation of knowledge into system's inner structure. Interiorization can be organized as supervised learning that uses the system with correctors as the supervisor. The AI system, equipped with correctors ('teacher'), labels randomly generated examples (proposes the answers or actions) and the AI system without correctors ('student') learns to give the proper answer. At the beginning, the student is the same legacy AI system, as the teacher, but without correctors. During the learning process, the student's skills change. The random generation of examples can be improved by selection of the more realistic examples and by elements of adversarial learning (selection of the examples with higher probability of errors). This play of the system with itself is a realization of the famous selfplay technology of DeepMind (for discussion of the selfplay principle and DeepMind Alpha Go Zero technology we refer to Holcomb et al. (2018)).

Stochastic separation theorems have three critical applications. One of them is one-shot correction of errors in intellectual systems. Recently, it was realized that the possibility to correct an AI system opens also the possibility to attack it. The dimensionality of the AI's decision-making space is a major contributor to the AI's vulnerability (Tyukin et al., 2020). So, the stochastic separation theorems demonstrate also the new version of the curse of dimensionality. As we said, the blessing and curse of dimensionality are two sides of the same coin. Thus, the second application is vulnerability analysis of high-dimensional AI systems in high-dimensional world.

The third application is to explain the "unreasonable effectiveness" of small neural ensembles in the multidimensional brain and the emergence of static and associative memories in the ensembles of single neurons (Gorban et al., 2019). A simple enough functional neuronal model is capable of explaining: i) the extreme selectivity of single neurons to the information content of high-dimensional data, ii) simultaneous separation of several uncorrelated informational items from a large set of stimuli, and (iii) dynamic learning of new items by associating them with already "known" ones (Tyukin et al., 2019). These results constitute a basis for organization of complex memories in ensembles of single neurons. The stochastic separation theorems give the theoretical background of existence and efficiency of 'concept cells' and sparse coding in a brain Gorban et al. (2019), Quian Quiroga (2019) and Tapia et al. (2020). (These 'hardware components of thought and memory' are presented in detail by Quian Quiroga et al. (2013, 2005) and Viskontas et al. (2009).)

There are also many technical applications of stochastic separation theorems with optimal bounds in various areas of data analysis and machine learning, for example, for estimation of dimensionality of data. The estimated dimension depends linearly on the exponents from these bounds for the methods based on the data separability properties (Bac & Zinovyev, 2020; Mirkes et al., 2020). Therefore, if we use bound with exponent twice far from the optimal one, then we misestimate the data dimension twice.

In recent review by Bac and Zinovyev (2020) the typology of these methods is proposed and a new family of methods based on the data separability properties is presented.

Stochastic separation theorems shed light on the fundamental *problem of learning from few examples in high dimensions*. This problem is central for understanding when and why modern large-scale systems can learn from post-classic data and generalize so well in practice. Classical generalization bounds stemming from the Vapnik–Chervonenkis theory (Vapnik, 1999) alone are too conservative to explain these successes. It has been demonstrated in Zhang et al. (2016) that absolutely identical deep neural networks are capable to exhibit both sides of the learning spectrum: to successfully generalize from meaningful training data and, at the same time, 'memorize' random assignments of labels without any generalization. Few-shot learning schemes such as matching (Vinyals et al., 2016) and prototypical networks (Snell et al., 2017), and success of stochastic configuration networks in practice (Wang & Li, 2017) are another manifestations of the same phenomenon.

These results suggest that neural networks' generalization capabilities are intrinsically linked with internal regularities in the data sets and also with representations of these regularities in the networks' latent spaces. Stochastic separation theorems reveal an important characteristic of this important regularity: if an object has a 'compact' representation in the network's latent space then such object can be learned from just few or even single example. The notion of 'compactness' here should be specified. For various classes of problems it can be thought of as covering of data by bounded number of balls with limited radii for some bounds, depending on the dimension and variability of the data, or as a sufficiently fast decay of a sequence of dataset diameters. Absence of such compact representations may require exponentially large training samples to learn from. In this respect, the theorems suggest that a successful learning process in modern networks with large VC dimension must include building an adequate data representation in the network's latent space.

The extreme rarefaction of data in the post-classical multidimensional world leads to many unexpected phenomena: applicability of simple discriminants to apparently complex problem of correcting AI, the possibility of stealth attacks on AI systems and the apparent simplicity of the concept cells and sparse coding in the brain. Kreinovich (2019) characterized this bunch of phenomena as "unheard-of simplicity", following Pasternak's famous verses. Stochastic separation theorems with optimal bounds provide a tool for dealing with these problems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work was supported by the University of Leicester, the Ministry of Science and Higher Education of the Russian Federation (Project No. 14.Y26.31.0022) and UKRI Alan Turing AI Acceleration Fellowship EP/V025295/1.

References

Bac, J., & Zinovyev, A. (2020). Lizard brain: tackling locally low-dimensional yet globally complex organization of multi-dimensional datasets. *Frontiers in Neurorobotics*, *13*(110), http://dx.doi.org/10.3389/fnbot.2019.00110.

Ball, K. (1997). An elementary introduction to Modern Convex Geometry. In S. Levy (Ed.), *Flavors of geometry* (pp. 1–58). Cambridge, UK: Cambridge University Press, http://library.msri.org/books/Book31/contents.html.

Bárány, I., & Füredi, Z. (1988). On the shape of the convex hull of random points. *Probability Theory and Related Fields*, *77*, 231–240. http://dx.doi.org/10.1007/BF00334039.

Bobkov, G. G. (2010). Gaussian concentration for a class of spherically invariant measures. *Journal of Mathematical Sciences*, *167*(3), 326–339. http://dx.doi.org/10.1007/s10958-010-9922-0.

Boucheron, G., Lugosi, G., & P., Massart. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Camastra, F. (2003). Data dimensionality estimation methods: a survey. *Pattern Recognition*, *36*(12), 2945–2954. http://dx.doi.org/10.1016/S0031-3203(03)00176-6.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and Blessings of Dimensionality. In *Invited lecture at mathematical challenges of the 21st century, AMS national meeting, Los Angeles, CA, USA, August 6-12, 2000*. CiteSeerX, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.329.3392.

Donoho, D., & Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society, Series A*, *367*, 4273–4293. http://dx.doi.org/10.1098/rsta.2009.0152.

Giannopoulos, A. A., & Milman, V. D. (2000). Concentration property on probability spaces. *Advances in Mathematics*, *156*, 77–106. http://dx.doi.org/10.1006/aima.2000.1949.

Gibbs, J. W. (1960). *Elementary principles in statistical mechanics, developed with especial reference to the rational foundation of thermodynamics*. New York, NY, USA: Dover Publications.

Gorban, A. N., Golubkov, A., Grechuk, B., Mirkes, E. M., & Tyukin, I. Y. (2018). Correction of AI systems by linear discriminants: Probabilistic foundations. *Information Sciences*, *466*, 303–322. http://dx.doi.org/10.1016/j.ins.2018.07.040.

Gorban, A. N., Kégl, B., Wunsch, D., & Zinovyev, A. (Eds.), (2008). *Principal manifolds for data visualisation and dimension reduction*. Berlin/Heidelberg, Germany: Springer, http://dx.doi.org/10.1007/978-3-540-73750-6.

Gorban, A. N., Makarov, V. A., & Tyukin, I. Y. (2019). The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Physics of Life Reviews*, *29*, 55–88. http://dx.doi.org/10.1016/j.plrev.2018.09.005.

Gorban, A. N., & Tyukin, I. Y. (2017). Stochastic separation theorems. *Neural Networks*, *94*, 255–259. http://dx.doi.org/10.1016/j.neunet.2017.07.014.

Gorban, A. N., & Tyukin, I. Y. (2018). Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society, Series A*, *376*, Article 20170237. http://dx.doi.org/10.1098/rsta.2017.0237.

Gorban, A. N., Tyukin, I., Prokhorov, D., & Sofeikov, K. (2016). Approximation with random bases: Pro et contra. *Information Sciences*, *364–365*, 129–145. http://dx.doi.org/10.1016/j.ins.2015.09.021.

Gorban, A. N., Tyukin, I. Y., & Romanenko, I. (2016). The blessing of dimensionality: Separation theorems in the thermodynamic limit. *IFAC-PapersOnLine*, *49*(24), 64–69. http://dx.doi.org/10.1016/j.ifacol.2016.10.755.

Gorban, A. N., & Zinovyev, A. (2010). Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems*, *20*, 219–232. http://dx.doi.org/10.1142/S0129065710002383.

Grechuk, B. (2019). Practical stochastic separation theorems for product distributions. In *Proc. 2019 international joint conference on neural networks (IJCNN)* (pp. 1–8). Budapest, Hungary: IEEE Press, http://dx.doi.org/10.1109/IJCNN.2019.8851817.

Gromov, M. (2003). Isoperimetry of waists and concentration of maps. *Geometric and Functional Analysis*, *13*, 178–215. http://dx.doi.org/10.1007/s00039-009-0703-1.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*, 13–30. http://dx.doi.org/10.1080/01621459.1963.10500830.

Holcomb, S. D., Porter, W. K., Ault, S. V., Mao, G., & Wang, J. (2018). Overview on DeepMind and its AlphaGo Zero AI. In *Proceedings of ICBDE '18, the 2018 international conference on big data and education* (pp. 67–71). New York: Association for Computing Machinery, http://dx.doi.org/10.1145/3206157.3206174.

Jolliffe, I. (1993). *Principal component analysis*. Berlin/Heidelberg, Germany: Springer.

Kainen, P. C. (1997). Utilizing geometric anomalies of high dimension: when complexity makes computation easier. In K. Warwick, & M. M. Kárný (Eds.), *Computer-intensive methods in control and signal processing: The curse of dimensionality* (pp. 283–294). New York, NY, USA: Springer, http://dx.doi.org/10.1007/978-1-4612-1996-5_18.

Kainen, P., & Kůrková, V. (1993). Quasiorthogonal dimension of Euclidian spaces. *Applied Mathematics Letters*, *6*, 7–10. http://dx.doi.org/10.1016/0893-9659(93)90023-G.

Kainen, P., & Kůrková, V. (2020). Quasiorthogonal dimension. In O. Kosheleva, S. P. Shary, G. Xiang, & R. Zapatrin (Eds.), *Beyond traditional probabilistic data processing techniques: Interval, fuzzy etc. Methods and their applications* (pp. 615–629). Cham: Springer, http://dx.doi.org/10.1007/978-3-030-31041-7_35.

Kreinovich, V. (2019). The heresy of unheard-of simplicity: Comment on The unreasonable effectiveness of small neural ensembles in high-dimensional brain by AN Gorban, VA Makarov, and IY Tyukin. *Physics of Life Reviews*, *29*, 93–95. http://dx.doi.org/10.1016/j.plrev.2019.04.006.

Kůrková, V. (2019). Some insights from high-dimensional spheres: Comment on "The unreasonable effectiveness of small neural ensembles in high-dimensional brain" by Alexander N. Gorban et al. *Physics of Life Reviews*, *29*, 98–100. http://dx.doi.org/10.1016/j.plrev.2019.03.014.

Kůrková, V., & Sanguineti, M. (2019). Probabilistic bounds for binary classification of large data sets. In L. Oneto, N. Navarin, A. Sperduti, & D. Anguita (Eds.), *Proceedings of the international neural networks society, Genova, Italy, 16–18 2019, Volume 1* (pp. 309–319). Berlin/Heidelberg, Germany: Springer, http://dx.doi.org/10.1007/978-3-030-16841-4_32.

Ledoux, M. (2001). *Mathematical surveys & monographs*: *No. 89*, *The concentration of measure phenomenon*. AMS.

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

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, *4*(1), 66–70. http://dx.doi.org/10.3923/ajms.2011.66.70.

Lopez, J. L., & Sesma, J. (1999). Asymptotic expansion of the incomplete beta function for large values of the first parameter. *Integral Transforms and Special Functions*, *8*(3–4), 233–236. http://dx.doi.org/10.1080/10652469908819230.

Mirkes, E. M., Allohibi, J., & Gorban, A. N. (2020). Fractional norms and quasinorms do not help to overcome the curse of dimensionality. *Entropy*, *22*(1105), http://dx.doi.org/10.3390/e22101105.

Moczko, E., Mirkes, E. M., Cáceres, C., Gorban, A. N., & Piletsky, S. (2016). Fluorescence-based assay as a new screening tool for toxic chemicals. *Scientific Reports*, *6*(33922), http://dx.doi.org/10.1038/srep33922.

Petrov, V. (1965). On the probabilities of large deviations for sums of independent random variables. *Theory of Probability and its Applications*, *10*(2), 287–298. http://dx.doi.org/10.1137/1110033.

Pham, H. (2007). Some applications and methods of large deviations in finance and insurance. In *Lecture notes in mathematics*: *vol. 1919*, *Paris-princeton lectures on mathematical finance 2004* (pp. 191–244). Berlin, Heidelberg: Springer, http://dx.doi.org/10.1007/978-3-540-73327-0_5.

Quian Quiroga, R. (2019). Akakhievitch revisited Comment on "The unreasonable effectiveness of small neural ensembles in high-dimensional brain" by Alexander N. Gorban et al. *Physics of Life Reviews*, *29*, 111–114. http://dx.doi.org/10.1016/j.plrev.2019.02.014.

Quian Quiroga, R., Fried, I., & Koch, C. (2013). Brain cells for grandmother. *Scientific American*, *308*(2), 30–35, http://www.jstor.org/stable/26017950.

Quian Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. http://dx.doi.org/10.1038/nature03687.

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

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Proc. of 28th international conference on neural information processing systems (NIPS 2015)*, *Advances in neural information processing systems 28* (pp. 2503–2511). N.Y: Curran Associates, Inc., http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.

Sidorov, S., & Zolotykh, N. (2020). Linear and Fisher separability of random points in the d-dimensional spherical layer and inside the d-dimensional cube. *Entropy*, *22*(11), Article 1281. http://dx.doi.org/10.3390/e22111281.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proc. of 30th international conference on neural information processing systems (NIPS 2017)*, *Advances in neural information processing systems 30* (pp. 4077–4087). N.Y: Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Tapia, C. C., Tyukin, I., & Makarov, V. A. (2020). Universal principles justify the existence of concept cells. *Scientific Reports*, *10*(1), 1–9. http://dx.doi.org/10.1038/s41598-020-64466-7.

Tyukin, I., Gorban, A. N., Calvo, C., Makarova, J., & Makarov, V. A. (2019). High-dimensional brain: A tool for encoding and rapid learning of memories by single neurons. *Bulletin of Mathematical Biology*, *81*(11), 4856–4888. http://dx.doi.org/10.1007/s11538-018-0415-5.

Tyukin, I. Y., Gorban, A. N., Sofeikov, K., & Romanenko, I. (2018). Knowledge transfer between artificial intelligence systems. *Frontiers in Neurorobotics*, *12*(49), http://dx.doi.org/10.3389/fnbot.2018.00049.

Tyukin, I. Y., Higham, D. J., & Gorban, A. N. (2020). On adversarial examples and stealth attacks in artificial intelligence systems. In *Proc. 2020 international joint conference on neural networks (IJCNN), Glasgow, United Kingdom, 2020* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/IJCNN48605.2020.9207472.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. http://dx.doi.org/10.1109/72.788640.

Vershynin, R. (2018). *Cambridge series in statistical and probabilistic mathematics*, *High-dimensional probability: An introduction with applications in data science*. Cambridge, UK: Cambridge University Press.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. In D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, & I. Guyon (Eds.), *Proc. of 30th annual conference on neural information processing systems, Barcelona, Spain (NIPS 2016)*, *Advances in neural information processing systems 29* (pp. 3637–3646). N.Y: Curran Associates, Inc., http://papers.neurips.cc/paper/6385-matching-networks-for-one-shot-learning.

Viskontas, I. V., Quian Quiroga, R., & Fried, I. (2009). Human medial temporal lobe neurons respond preferentially to personally relevant images. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21329–21334. http://dx.doi.org/10.1073/pnas.0902319106.

Wang, D., & Li, M. (2017). Stochastic configuration networks: Fundamentals and algorithms. *IEEE Transactions on Cybernetics*, *47*(10), 3466–3479. http://dx.doi.org/10.1109/TCYB.2017.2734043.

Wendel, J. G. (1948). Note on the gamma function. *American Mathematical Monthly*, *55*(9), 563–564, https://www.jstor.org/stable/i314786.

Wong, R. (2001). *Asymptotic approximations of integrals*. Philadelphia, PA: SIAM.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530. https://arxiv.org/abs/1611.03530.