Article

High-dimensional separability for one- and few-shot learning

Alexander N. Gorban [1,2,*], Bogdan Grechuk [1], Evgeny M. Mirkes [1,2], Sergey V. Stasenko[2] and Ivan Y. Tyukin [1,2,3]

1 Department of Mathematics, University of Leicester, Leicester, UK
2 Lobachevsky University, Nizhni Novgorod, Russia
3 Department of Geoscience and Petroleum, Norwegian University of Science and Technology
* Correspondence: a.n.gorban@le.ac.uk

Version June 28, 2021 submitted to Journal Not Specified

Abstract: This work is driven by a practical question, corrections of Artificial Intelligence (AI) errors. Systematic re-training of a large AI system is hardly possible. To solve this problem, special external devices, correctors, are developed. They should provide quick and non-iterative system fix without modification of a legacy AI system. A common universal part of the AI corrector is a classifier that should separate undesired and erroneous behavior from normal operation. Training of such classifiers is a grand challenge at the heart of the one- and few-shot learning methods. Effectiveness of one- and few-short methods is based on either significant dimensionality reductions or the blessing of dimensionality effects. Stochastic separability is a blessing of dimensionality phenomenon that allows one-and few-shot error correction: in high-dimensional datasets under broad assumptions each point can be separated from the rest of the set by simple and robust linear discriminant. The hierarchical structure of data universe is introduced where each data cluster has a granular internal structure, etc. New stochastic separation theorems for the data distributions with fine-grained structure are formulated and proved. Separation theorems in infinite-dimensional limits are proven under assumptions of compact embedding of patterns into data space. New multi-correctors of AI systems are presented and illustrated with examples of predicting errors and learning new classes of objects by a deep convolutional neural network.

Keywords: artificial intelligence; blessing of dimensionality; clusters; errors; separability; discriminant; dimensionality reduction

1. Introduction

1.1. One- and few-short learning

Learning new concepts rapidly from small low-sample data is a key challenge in machine learning [1]. Despite of a widespread perception of neural networks as monstrous giant systems, iterative training of which requires a lot of time and resources, one- and few-shot learning appears to be possible. Several modern approaches to organization of this type of learning are based on the preliminary training tasks that are similar but not fully identical to the new task to be learned.. The previous experience of training improves learnability: after massive training the system gains some meta-skills, and new tasks, which are not crucially different from the previous ones, do not require large training sets and training time. This heuristic was utilized in various constructions of one- and few-short learning algorithms [3]. The meta-skills and learnability can be gained in the previous experience of solving various sensible problems or in a specially organized meta-learning [4,5].

One- and few-short learning is based mainly on combinations of reasonable *preparatory learning* that should increase learnability and create meta-skills, and simple learning routines that are used for learning from small number of examples after this propaedeutics. These simple methods act in the system's latent space. This is the feature space created in the course of preparatory learning that can be organized by solving a large number of previous problems or a special set of problems selected for the meta-learning. Typically, a copy of the same pretrained system is used for different one- and few-short learning tasks. Nevertheless, plenty of approaches are applicable to few-shot minor modifications of the features using new tasks.

Effectiveness of one- and few-short simple methods is based on either significant *dimensionality reductions* or the *blessing of dimensionality* effects [6,7].

A significant reduction in dimensionality means that several features have been extracted that are sufficient to solve a large number of previous problems or special set of problems selected for the meta-learning. Thereafter, a well-elaborated library of efficient lower-dimensional statistical learning methods can be applied to a few-shot solving of new problems using the same features. Of course, plenty of approaches are possible to few-shot minor modifications of the features using new tasks.

The blessing of dimensionality is a relatively new idea [8–11]. It means that the simple classical techniques like linear Fisher's discriminant become unexpectedly powerful in high dimensions under some assumptions about regularity of probability distributions [12–14]. These assumptions can be rather mild and typically include absence of extremely dense lumps that are areas with relatively low volume but unexpectedly high probability (for more detail we refer to [15]). These lumps correspond to the narrow but high peaks of probability density.

If a dataset consists of k such lumps then, for moderate values of k, this can be considered as a special case of dimensionality reduction. The centers of clusters are considered as 'principal points' to stress the analogy with principal components [16,17]. Such a clustered structure in system's latent space may appear in the course of preparatory learning: the images of data points in the latent space, *"attract similar and repulse dissimilar"* data points. Similarity could be described by the proximity in the space of the outputs.

The one- and few-short learning can be organized in all three situations described above:

1. If the feature space is effectively reduced, then the challenge of large data set can be mitigated and we can rely on classical linear or non-linear methods of statistical learning.
2. In the situation of 'blessing of dimensionality', with sufficiently regular probability distribution in high dimensions the simple linear (or kernel [18]) one- and few-short methods become effective [7,14,15].
3. If the datapoints in the latent space form dense clusters, then position of new data with respect to these clusters can be utilized for solving new tasks. We can also expect that new data may introduce new clusters, but persistence of the cluster structure seems to be important. The clusters themselves can be distributed in a multidimensional feature space. This is the novel and more general setting we are going to focus below in Sec. 3.

There is a rich set of tools for dimensionality reduction. It includes the classical prototype, principal component analysis (PCA) (see, for example, [17]), and many generalizations, from principal manifolds [20] and kernel PCA [21] to principal graphs [17,22] and autoencoders [23,24].We will briefly describe some of these important tools in the context of data preprocessing (Section 2), but the detailed analysis of dimensionality reduction is out of the main scope of the paper.

In a series of previous works, we focused on the second item [6,11–15,25]. The blessing of dimensionality effects that make the one- and few-shot learning possible for regular distributions of data are based on the stochastic separation theorems. All these theorems have a similar structure: for large dimensions, even in an exponentially large (relatively to the dimension) set of points, each point is separable from the rest by a linear functional, which is given by a simple explicit formula. These blessing of dimensionality phenomena are closely connected to the concentration of measure [26–29]. In particular, the blessing of dimensionality is closely connected to the various versions of the central

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

limit theorem in probability theory [19]. Of course, there remain open questions about sharp estimates for some distribution classes, but the general picture seems to be clear now.

In this work, we focus mainly on the third point and explore the blessings of dimensionality and related methods of one- and few-shot learning for *multidimensional data with rich cluster structure*. Such datasets cannot be described by regular probability densities with a priori bounded Lipschitz constant. Even more general assumptions about absence of sets with relatively small volume but relatively high probability fail. We believe that this option is especially important for applications.

1.2. AI errors and correctors

The main driver of our research is the problem of AI errors and their correction: *All AI systems sometimes make errors and will make errors in the future.* These errors must be detected and corrected immediately and locally in the networks of collaborating systems. If we do not solve this problem, then a new AI winter will come. Recall that the previous AI winters came after the hype peaks of inflated expectations and bold advertising.

Gartner's Hype Cycle is a convenient tool for representation of R&D trends. According to Gartner [30], the data-driven Artificial Intelligence (AI) has already left the Peak of Inflated Expectation and is descending into the Trough of Disillusionment. In more detail, some AI applications are still climbing to the peak, but Machine Learning and Deep Learning are going down. Explainable AI joined them in 2020 [31].

According to Gartner's Hype cycle model, the Trough of Disillusionment will turn into the Slope of Enlightenment that leads to Plateau of Productivity. The modern Peak and Trough is not the first in the history of AI. Surprisingly, previous troughs (AI winters) did not turn into the performance plateaus, but alternated with new hype with new peaks of inflated expectations (Figure 1) [32].

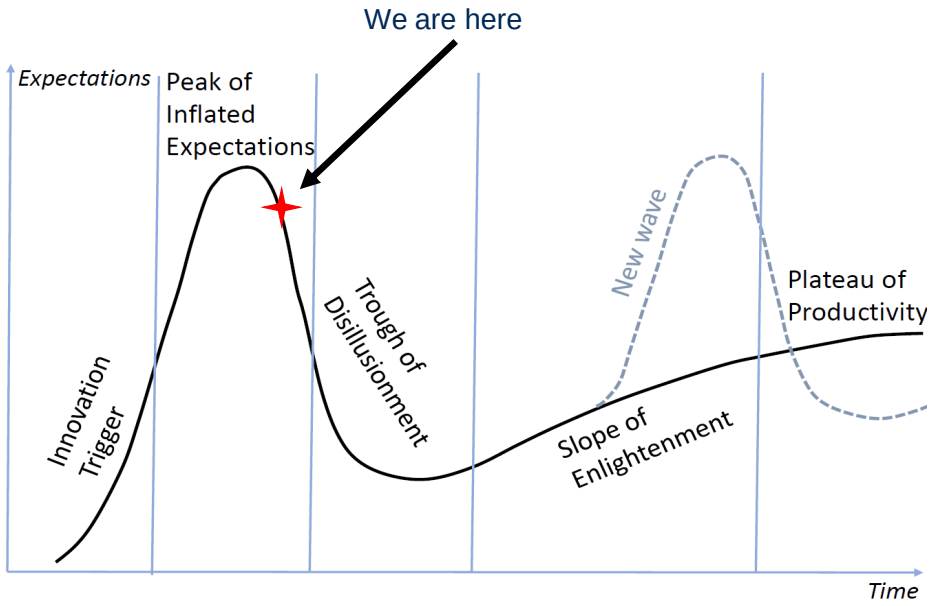

Figure 1. *Gartner Hype Cycle and its phases.* Position of the data-driven AI on the hype cycle is marked by a four-pointed star. A possible new hype peak (new wave) is represented by the dashed line.

What problem is pushing the AI downhill? Is this the same problem that pushed AI down previous slopes decades ago? The main problem for the widespread use of AI around the world are unexpected errors in real-life applications:

- The mistakes can be dangerous;
- Usually, it remains unclear, who is responsible for them;
- The types of errors are numerous and often unpredictable,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

- The real world is not a good i.i.d. sample,
- So we cannot rely on a statistical estimate of the probability of errors in real life.

Errors are unavoidable companions of data driven AI. Fundamental origins of AI errors could be different. Of course, they include software errors, unexpected human behavior and non-intended use as well as many other possible reasons. Nevertheless, the universal cause of errors is uncertainty in training data and in training process. The real world possibilities are not covered by the dataset.

The mistakes should be corrected. The systematic re-training of a big AI system seems to be rarely possible:

- To preserve existing skills we must use the full set of training data.
- This approach requires much recourses for each error.
- However, new errors may appear in the course of re-training.
- The preservation of existing skills is not guaranteed.
- The probability of damage to skills is a priori unknown.

To avoid a recurrence of a detected error, a quick, non-iterative system fix is required. This is the main challenge for the one- and few-shot learning methods.

To provide fast error correction, we must consider developing *correctors*, external devices that complement legacy artificial intelligence systems, diagnose the risk of error and correct errors. The original AI system remains part of the extended "system + corrector" complex. Therefore, the correction is reversible, and the original system can always be extracted from the augmented AI complex. Correctors have two different functions: (i) they should recognize potential errors and (2) provide corrected outputs for situations with potential errors. The idealized scheme of a legacy AI system augmented with an elementary corrector is presented in Fig. 2.

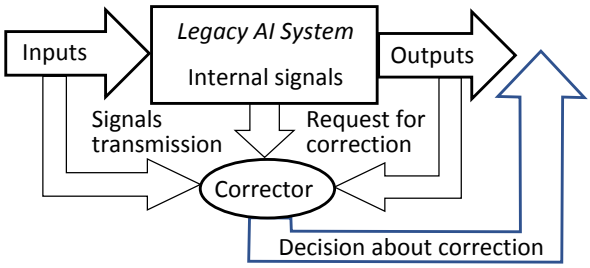

Figure 2. *Scheme of operation of an elementary corrector of legacy AI systems.* The elementary corrector receives the input signals of the legacy AI system, the internal signals generated by this system in the decision-making process, and its output signals and assesses the need for correction. The elementary corrector includes a binary classifier that separates situations with a high risk of error from the normal functioning. A modified decision rule is specified for a corrected AI system in a situation with a high risk. If correction is required, the corrector sends a warning signal and a modified output for further use.

The universal part of the AI corrector is a classifier that should separate situations with erroneous behavior from normal operation. It is a binary classifier for all types of AI. The generalization ability of this classifier is its ability to recognize the errors that it had never seen before. The training set for corrector consists of a collection of situations with normal operation of the legacy AI system (the 'normal' class) and a set of labeled errors. The detection and labeling of errors for training correctors can be performed by various methods, which include human inspection, decisions of other AI systems of their committees, signals of success or failure from the outer world, and other possibilities that are outside the scope of our work.

We can usually expect that a normal class of error-free operations includes many more examples than a set of labeled errors. Moreover, even the situation with one newly labeled error is of considerable

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

interest. All the stochastic separation theorems were invented to develop the one- of few-shot learning rules for the binary error/normal operation classifiers.

A specific component of the AI corrector is the modified decision rule (the 'correction' itself). Of course, the general theory and algorithms are focused on the universal part of the correctors. For many classical families of data distributions, it is proved that the well-known Fisher discriminant is surprisingly a powerful tool for constructing correctors if the dimension of the data space is sufficiently high (most results of this type are collected in [15]). This is proven for a wide class of distributions, including log-concave distributions, their convex combinations, and product distributions with bounded support.

In this article, we refuse the classical hypothesis of the regularity of the data distribution and assume that the data can have a rich fine-grained structure with many clusters and corresponding peaks in the probability density. Moreover, the notion of probability distribution in high dimensions may sometimes create more question than answers. Therefore, after development of new stochastic separation theorems for data with fine-grained clusters, we discuss the possibility to substitute the probabilistic approach to foundations of the theory by more robust functional analysis methods with limit transition to infinite dimension.

The source of the fine-grained structure in the data seems to be very natural and universal: the observable world consists of things. The datapoints represent situations. The qualitative difference between situations is in existence/absence of notable things there. Let us avoid the philosophical discussion about the existence of things in reality or their creation in the course of cognition (like Kant's "thing in itself" and its transformation into "thing for us"). We simply use the existence of things as a fact from our experience, captured in the semantics of languages, and try to find a direct and useful (as we seek to demonstrate) formalization of this experience.

1.3. The structure of the paper

In Sec. 2 we consider the problems of post-classical data analysis in very high-dimensional spaces, where classical statistical methods seem to be useless because in high dimensions we will never have enough data points. We begin with Donoho's definition of post-classical problems, where the number of attributes is greater than the number of data points [9], discuss alternative definitions and preprocessing of such data. In particular, we discuss the following preprocessing operations:

- Correlation transformation;
- PCA;
- Supervised PCA;
- Semi-supervised PCA;
- Transfer Component Analysis (TCA);
- The novel expectation-maximization Domain Adaptation PCA ('DAPCA').

In Sec. 3 the stochastic separation theorems for the data distributions with fine-grained structure are formulated and proved. For these theorems, we model clusters by geometric bodies (balls or ellipsoids) and work with distributions of ellipsoids in high dimensions. The hierarchical structure of data universe is introduced where each data cluster has a granular internal structure, etc. Separation theorems in infinite-dimensional limits are proven under assumptions of compact embedding of patterns into data space.

Sec. 4 represents the structure of correctors for multiple clusters of errors (multi-correctors). For such data sets, several elementary correctors and a dispatcher are required, which distributes situations for analysis to the most appropriate elementary corrector. In multi-corrector, each elementary corrector separates its own area of high-risk error situations and contains an alternative rule for making decisions in situations from this area. The input signals of the correctors are the input, internal and output signals of the AI system to be corrected, as well as any other available attributes of the situation. The system of correctors is controlled by a dispatcher, which is formed on the basis of a cluster analysis of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

errors and distributes the situations specified by the signal vectors between elementary correctors for evaluation and, if necessary, correction.

In Sec. 5 we present a case study that illustrates how "clustered" or "granuled universes" may arise in real data, and show how the structure underpinned by granular representation can be used in challenging machine learning and AI problems. These problems include learning new classes of data in legacy deep learning AI models and predicting AI errors. We present simple algorithms and workflows which can be used to solve these challenging tasks circumventing the needs for computationally expensive re-training. We also illustrate potential technical pitfalls and dichotomies requiring additional attention from the algorithms' users and designers.

Discussion (Sec. 6) aims to explain a main message: the success or failure of many machine learning algorithms, the possibility of meta-learning and gaining experience depend on the world structure. The possibility to represent a real world situations as a collection of things with some features (properties) and relationships between these entities is the fundamental basis of knowledge of both humans and AI. Despite the universality of this ontology, things–features–relations, two possibilities of destruction of such a structure are also well known:

- "Someone's noise is another one's signal!" This 'axiom' of data mining means that after some a priori unknown transformation of data, a new 'things–features–relations' structure will emerge from the noise, and the old structure will become background noise. From the philosophical point of view, this data mining operational situation looks pretty dramatic, with things disappearing and new thing emerging from the chaos.
- The physical world of fields, and then quantum mechanics or quantum field theory contradicts the simplistic 'things–features–relations' ontology. Nevertheless, there is more than enough tasks in data analysis that can be captured in the form of the 'things–features–relations' world.

When working with ordinary universes that are made of things (with features and relations) and allow for meta-learning, we must keep in mind that the situation can be destroyed, and this is just another face of reality.

Appendices A and B include auxiliary mathematical results and technical information.

2. Preprocessing in post-classical data world

2.1. Postclassical data

High-dimensional post-classical world was defined in [9] by the inequality

$$\text{The number of attributes } d \gg \text{ The number of examples } N. \tag{1}$$

This post-classical world is different from the 'classical world', where we can consider infinite growth of the sample size for the given number of attributes. The classical statistical methodology was developed for the classical world based on the assumption of

$$d < N \text{ and } N \to \infty.$$

Thus, the classical statistical learning theory is mostly useless in the multidimensional post-classical world. These results all fail if $d > N$. The $d > N$ case is not anomalous for the modern big data problems. It is the generic case: both the sample size and the number of attributes grow, but in many important cases the number of attributes grows faster than the number of labeled examples [9].

High-dimensional effects of the curse and blessing of dimensionality appear in a much wider area than specified by the inequality (1). A typical example gives the penomenon of quasiorthogonal

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

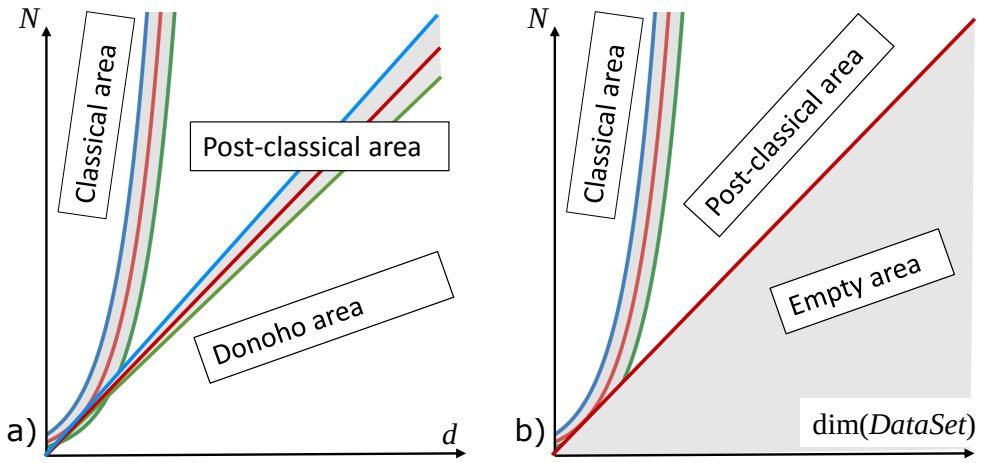

Figure 3. Different zones of data world: a) Separation of Donoho's postclassical data world, where $d > N$ (below the bisector), the classical world, where $d \ll logN$ and the 'postclassical' area below the exponent, $d \gg logN$; b) Classical and postclassical data worlds according to the definition (3) (the area below the bisector is empty). The gray areas around the borders between the different areas symbolize the fuzziness of the borders. Here, d is the number of attributes, N is the number of samples, and $\dim(DataSet)$ is the intrinsic dimensionality of the dataset, $d \geq \dim(DataSet)$ and $N > \dim(DataSet)$.

dimension [33–35]: for a given $\varepsilon > 0$ and $\vartheta > 0$ (assumed small) a random set of N vectors $\boldsymbol{x}_i$ on a high-dimensional unit d-dimensional sphere satisfies the inequality

$$|(\boldsymbol{x}_i, \boldsymbol{x}_j)| < \varepsilon$$

for all $i \neq j$ with probability $p > 1 - \vartheta$ when $N < a\exp(bd)$ and a and b depend on ε and ϑ only. This means that the quasiorthogonal dimension of an Euclidean space grows exponentially with dimension d. Such effects are important in machine learning [35]. Therefore, the Donoho boundary should be modified: the postclassical effects appear in high dimension when

$$d \gg logN. \tag{2}$$

The two different definitions of postclassical area, (1), (2), are illustrated in Fig. 3a.

The definition of the postclassical data world needs one more comment. The inequalities (1), (2) used the number of attributes as the equivalent of the dimension of the data space. Behind this approach is the hypothesis that there is no strong dependency between attributes. In the real situations, the data dimensionality can be much less that the number of attributes, for example, in the case of the strong multicollinearity. If, say, the data are located along a straight line then for most approaches the dimension of the dataset is 1 and the value of d does not matter. Therefore, the definition (2) of the postclassical world needs to be modified further with the dimension of the dataset, $\dim(DataSet)$ instead of d:

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

There are many various definitions of data dimensionality, see a brief review in [36,37]. For all of them, we can assume that $\dim(DataSet) < N$ and $\dim(DataSet) \leq d$ (see Fig. 3b).

The post-classical world effects include blessing and curse of dimensionality. Both these blessing and curse are based on the concentration of measure phenomena [26–29] and are, in that sense, two sides of the same coin [14,25].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

2.2. Measure examples by examples and reduce the number of attributes to $\dim(DataSet)$

Assume that the number of data points is less than the number of attributes (1). In this situation, we can decrease the dimension of space by many simple transformations. It is possible to apply PCA and delete all the components with vanishing eigenvalues. This could be a non-optimal approach if originally d is very large. It is also possible to restrict the analysis by the space generated by the data vectors. Let the data sample be a set of N vectors $\boldsymbol{x}_i$ in $\mathbb{R}^d$. One way to reduce the description is the following *correlation transformation* that maps the dataspace into cross-correlation space:

1. Centralize data (subtract the mean);
2. Delete coordinates with vanishing variance; (*Caution*: signals with small variance may be important, whereas signals with large variance may be irrelevant for the target task! This standard operation can help but can also impair the results.)
3. Standardize data (normalize to unit standard deviations in coordinates), or use another normalization, if this is more appropriate; (*Caution*: transformation to the dimensionless variables is necessary but selection of the scale (standard deviation) affects the relative importance of the signals and can impair the results.)
4. Normalize the data vectors to unit length: $\boldsymbol{x}_i \mapsto \boldsymbol{x}_i / \|\boldsymbol{x}_i\|$ (*Caution*: this simple normalization is convenient but deletes one attribute, the length. If this attribute is expected to be important than it could be reasonable to use the mean value of $\|\boldsymbol{x}_i\|$ that gives normalization to the unit average length.)
5. Introduce coordinates in the subspace spanned by the dataset, $\mathrm{Span}\{\boldsymbol{x}_i\}$ using projections on $\boldsymbol{x}_i$.
6. Each new datapoint $\boldsymbol{y}$ will be represented by a N-dimensional vector of inner products with coordinates $(\boldsymbol{y}, \boldsymbol{x}_i)$.

After this transformation, the data matrix becomes the Gram matrix $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. For the centralized and normalized data, these inner products can be considered as correlation coefficients. For this dataset, the number of attributes coincides with the number of datapoints. The next step may be PCA or other method of dimensionality reduction. The simple and routine normalization operations can significantly affect the results of data analysis and choosing the right option cannot be made a priori.

In a case study [38], the combination of the correlation transformation with PCA performs well even when the number of attributes was $5 \cdot 10^5$ and the number of samples was about 60. The correlation transformation reduced the dimension to the sample size, and there were five or six major components left after PCA. However, if the data set is truly multidimensional, then the correlation transformation can return a data matrix with strong diagonal dominance. Centralized random vectors will be almost orthogonal due to the phenomenon of quasi-orthogonality. This effect can make the application of PCA after the correlation transformation less efficient.

There is a different approach to dealing with relatively small samples in multidimensional data spaces. In the Donoho area (see (1) and Fig. 3a) we can try to produce a probabilistic generative model and then use it for generation additional data.

The zeroth approximation is the naïve Bayes model. This means assuming that the attributes are independent. The probability distribution is the product of distributions of attributes values. In dimension d, we need to fit the d one-dimensional densities, which is much easier than reconstructing the d-dimensional density in the entire data space. The naïve Bayes approximation can be augmented by accounting strong pair correlations, etc. The resulting approximation may be represented in the form of a Bayesian network [39,40].

There are many methods for generating the probability distribution from data, based on the maximum likelihood estimation married with the network representation of the distribution, like deep latent Gaussian models [41].

The physical interpretation of the log-likelihood as energy (or free energy) gave rise to many popular heuristic approaches like the Boltzmann machine or restricted Boltzmann machine [42] that create approximation of the energy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Extensive experience was accumulated in the use of various generative models of probability distribution. They can be used to leave the Donoho area by augmentation of the dataset with additional samples generated by the model. The statistical status of such augmentation is not always clear because selection of the best model is an intractable problem and we never have enough data and time to solve it. In large dimension, the models are tested on a standard task: accurate imputations of missing data for the samples never seen before. These tests should check if the majority of correlations captured by the model are significant (and not spurious) and may be used to evaluate the False Discovery Rate (FDR).

A good heuristic should provide a reasonable balance between the risk of missing significant correlations and the risk of including spurious correlations. This is a typical multiple testing problem and in the postclassical data world we cannot be always sure that we solved this problem properly. The standard correcting for multiplicity (see, for example, [44]) may result in too many false negative errors (missed correlations) and prematurely throw the baby out with the bathwater. But without such corrections, any findings should be seen as hypothesis generating, not as definitive results [43]. This difficulty can be considered as the fundamental incompleteness of the postclassical datasets.

2.3. Unsupervised, supervised, and semisupervised PCA

PCA remains the standard and very popular tool for dimensionality reduction and unsupervised data preprocessing. It was introduced by K. Pearson in 1900 as a tool for data approximation by straight lines and planes of best fit. Of course, minimization of the mean square distance from the data point to its projection on a plane (i.e. mean square error of the approximation) is equivalent to maximization of the variance of projections (because Pythagorean theorem). This second formulation became the main definition of PCA in textbooks [45]. The third definition of PCA, which we will use below, is more convenient for developing various generalizations. [17].

Let a data sample $\boldsymbol{x}_i \in \mathbb{R}^d$ $(i = 1, \dots, N)$ be given and centralized, and let Π be a projector of $\mathbb{R}^d$ on a q-dimensional plane. The problem is to find the q-dimensional plane that maximizes the scattering of the data projections

$$\frac{1}{2}\sum_{i,j=1}^{n} \|\Pi(\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2. \tag{4}$$

For projection on a straight line (1D subspace) with the normalized basis vector $\boldsymbol{e}$ the scattering (4) is

$$\frac{1}{2}\sum_{i,j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{e})^2 = N\sum_{i=1}^{N} (\boldsymbol{x}_i, \boldsymbol{e})^2 = N(N-1)(\boldsymbol{e}, Q\boldsymbol{e}) \tag{5}$$

where the coefficients of the quadratic form $(\boldsymbol{e}, Q\boldsymbol{e})$ are the sample covariance coefficients $q_{lm} = \frac{1}{N-1}\sum_i x_{il}x_{im}$, and x_{il} $(l = 1, \dots, d)$ are coordinates of the data vector $\boldsymbol{x}_i$.

If $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_q\}$ is an orthonormal basis of the q-dimensional plane in data space, then the maximum scattering of data projections (4) is achieved, when $\boldsymbol{e}_1, \dots, \boldsymbol{e}_q$ are eigenvectors of Q that correspond to the q largest eigenvalues of Q (taking into account possible multiplicity) $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$. This is the standard PCA exactly. A deep problem with using PCA in data analysis is that the major components are not necessarily the most important or even relevant for the target task. Users rarely need to simply explain a certain fraction of variance. Instead, they need to solve a classification, prediction, or other meaningful task. Discarding certain major principal components is a common practice in many applications. First principal components are frequently considered to be associated with technical artifacts in the analysis of omics datasets in bioinformatics, and their removal might improve the downstream analyses [46,47]. Even more than 10 first principal components have to be removed sometimes, in order to increase the signal/noise ratio [48].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

The component ranking can be made more meaningful if we change the form (4) and include additional information about the target problem in the principal component definition. The form (4) allows many useful generalizations. Introduce weight W_{ij} for each pair:

$$H = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\|\Pi(\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2. \tag{6}$$

The weigh W_{ij} may be positive for some pairs (repulsion) or negative for some other pairs (attraction). The weight matrix is symmetric, $W_{ij} = W_{ji}$. Again, the problem of H maximization leads to a diagonalization of a symmetric matrix. Consider projection on a 1D subspace with the normalized basis vector $\boldsymbol{e}$ and define a new quadratic form with coefficients q^W_{lm}:

$$H = \sum_{lm}\left[\sum_i\left(\sum_r W_{ir}\right)x_{il}x_{im} - \sum_{ij} W_{ij}x_{il}x_{jm}\right]e_l e_m = \sum_{lm} q^W_{lm} e_l e_m. \tag{7}$$

Maximum of H (6) on q-dimensional planes is achieved when this plane is spanned by q eigenvectors of the matrix $Q^W = (q^W_{lm})$ (7) that correspond to q largest eigenvalues of Q^W (taking into account possible multiplicity) $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$.

To prove this statement we can mention that the functional H for a q-dimensional plane (6) is the sum of the functionals (7) calculated for vectors from any orthonormal basis of this plane. Let this basis be $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_q\}$. Decompose each $\boldsymbol{e}_i$ in the orthonormal basis of Q^W eigenvectors and follow the classical proof for PCA.

There are several methods for the weights assignment:

- *Classical PCA*, $W_{ij} \equiv 1$;
- *Supervised PCA for classification* tasks [49,50]. The data set is split into several classes, K_v $(v = 1, 2, \ldots, r)$. Follow the strategy "attract similar and repulse dissimilar". If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, then $W_{ij} = -\alpha < 0$ (attraction). If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to different classes, then $W_{ij} = 1$ (repulsion). This preprocessing can substitute several layers of feature extraction deep learning network [51].
- *Supervised PCA for any supervising task*. The data set for supervising tasks is augmented by labels (the desired outputs). There is proximity (or distance, if possible) between these desired outputs. The weight W_{ij} is defined as a function of this proximity. The closer the desired outputs are, the smaller the weights should be. They can change sign (from classical repulsion, $W_{ij} > 0$ to attraction, $W_{ij} < 0$) or simply change the strength of repulsion.
- *Semi-supervised PCA* was defined for a mixture of labeled and unlabeled data [52]. The data are labeled for classification task. For the labeled data, weights are defined as above for supervised PCA. Inside the set of unlabeled data the classical PCA repulsion is used.

All these modifications of PCA are formally very close. They are defined by a maximization of the functional (6) for different distributions of weights. This maximization is transformed into the spectral problem of a symmetric matrix Q^W (see (7) or its simple modification (8)). The dimensionality reduction is achieved by projection of data onto linear span of q eigenvectors of Q^W that correspond to the largest eigenvalues.

How many components to retain, is a non-trivial question even for the classic PCA [53]. The methods based on the evaluation of the fraction of variance unexplained or, what is the same, the relative mean square error of the data approximation by the projection, are popular but we should have in mind that this projection should not only approximate the data, but also be a filter that selects meaningful features. Therefore, the selection of components to keep depends on the problem we aim to solve and heuristic approaches with several trials of different numbers of components may be more useful than an unambiguous formal criterion. Special attention is needed to the cases when some eigenvalues of Q^W become negative. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ but for other eigenvalues

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

$0 \geq \lambda_{r+1} \geq \ldots$. In this case, a further increase in the dimension of the approximating plane above r does not lead to an increase in H but definitely increases the quality of data approximation. The standard practice is not to use eigenvectors that correspond to non-positive eigenvalues [51].

2.4. DAPCA – Domain Adaptation PCA

The classical hypothesis of machine learning is existence of the probability distribution and the same (even unknown) distribution for the training and test sets. The problem of domain adaptation arises when the training set differs from the data that the system should work with under operational conditions. Such situations are typical. The problem is that the new data have no known labels. We have to utilize a known labeled training set (from the "source domain') and a new unlabeled training set (from the "target domain"). The idea is to modify the data and to make the non-labeled data as close to the labeled one as possible. This transformation should erase the difference between the data distributions in two sets and, at the same time, do not destroy the possibility to solve effectively the machine learning problem for the labeled set.

The key question in domain learning is definition of the objective functional: how to measure the difference in distributions between the source domain sample and the target domain sample. The clue to the answer gives the idea [54]:

- Select a family of classifiers in data space;
- Choose the best classifier from this family for separation the source domain samples from the target domain samples;
- The error of this classifier is an objective function for maximization (large classification error means that the samples are indistinguishable by the selected family of classifiers).

Ideally, there are two systems: a classifier that distinguishes the feature vector as either a source or target and a feature generator that learns a combination of tasks: to mimic the discriminator and to ensure the successful learning in the source domain. There are many attempts to implement this idea [55,56]. In particular, an effective neural network realization trains a deep neural network system accurately classify source samples, but decreasing the ability of the associated classifier that uses the same feature set to detect whether each example belongs to the source or target domains [57]. The scattering objective function (6) can combine these two targets for learning of feature generation: success in the learning in the source domain and indistinguishability of the source and target data sets

Transfer Component Analysis (TCA) was proposed to specify attraction between the clouds of projections of labeled and unlabeled data [58]. The distance between the source and target samples was defined as the distance between the projections of their mean points. Attraction between the mean points of the labeled and unlabeled data was postulated. Let $\boldsymbol{\mu}_{\mathrm{L}}$ and $\boldsymbol{\mu}_{\mathrm{U}}$ be these mean points. Their attraction means that a new term should be added to Q^W (7):

$$q^W_{lm} = \sum_i \left(\sum_r W_{ir} \right) x_{il} x_{im} - \sum_{ij} W_{ij} x_{il} x_{jm} - \beta(\mu_{\mathrm{L}l} - \mu_{\mathrm{U}l})(\mu_{\mathrm{L}m} - \mu_{\mathrm{U}m}), \tag{8}$$

where weights W_{ir} are assigned by the same rules as in semisupervised PCA, and $\beta > 0$ is the attraction coefficient between the mean points of the labeled and unlabeled data samples.

Domain Adaptation PCA (DAPCA) also takes advantage of this idea of task mix within a weighted PCA framework (6). The classifier used is the classical kNN (k Nearest Neighbors). Let the source data set (input vectors) be $\mathbf{X}$, the target data set be $\mathbf{Y}$, $\mathbf{X}$ is split into different classes: $\mathbf{X} = K_1 \cup \ldots \cup K_r$. Enumerate points in $\mathbf{Y} \cup \mathbf{X}$ The weights are:

- If $\boldsymbol{x}_i, \boldsymbol{x}_j \in K_v$ then $W_{ij} = -\alpha < 0$ (the source samples from one class, attraction);
- If $\boldsymbol{x}_i \in K_u$ $\boldsymbol{x}_j \in K_v$ $(u \neq v)$ then $W_{ij} = 1$ (the source samples from different classes, repulsion);
- $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbf{Y}$ then $W_{ij} = \beta > 0$ (the target samples, repulsion)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

- For each target sample $x_i \in \mathbf{Y}$ find k closest source samples in $\mathbf{X}$. Denote this set E_i. For each $x_j \in E_i$, $W_{ij} = -\gamma < 0$ (the weight for connections of a target sample and the k closest source samples, attraction).

The weights in this method depend on three non-negative numbers, α, β, and γ, and on the number of nearest neighbors, k. Of course, the values of the constants can vary for different samples and classes, if there is sufficient reason for such a generalization.

kNN classification can be affected by irrelevant features that create difference between the source and target domains and should be erased in the feature selection procedure. This difficulty can be resolved by the *iterative DAPCA*. Use the basic algorithm as the first iteration. It gives the q-dimensional plane of major components (the eigenvectors Q^W) with the orthogonal projector in it Π_1. Find for each target sample k nearest neighbors from the source samples in the projection on this plane (use for definition of k nearest neighbors the seminorm $\|\Pi_1(x) - \Pi_1(y\|)$. Assign new W_{ij} using these nearest neighbors. Find new projector Π_2 and new nearest neighbors. Iterate. The iterations converge in a finite number of steps, because the functional H (7) increases at each step (as in the k-means and similar splitting algorithms). Even if the convergence (in high dimensions) is too long, then the early stop can produce a useful feature set. The iterative DAPCA helps also to resolve the classical distance concentration difficulty: in essentially large dimensional distributions the kNN search may be affected by the distance concentration phenomena: most of the distances are close to the median value [59]. Even use of fractional norms or quasinorms does not save the situation [60], but dimensionality reduction with deleting the irrelevant features may help.

If the target domain is empty then TPA, DAPSA, and iterative DAPCA degenerate to the semi-supervised PCA in the source domain. If there is no source domain then they turn into classical PCA in the target domain.

The described procedures of supervised PCA, semi-supervised PCA, TCA, DAPCA, or iterative DAPCA prepare a relevant feature space. The distribution of data in this space is expectedly far from a regular unimodal distribution. It is assumed that in this space the samples will form dense clumps with a lower data density between them.

3. Stochastic separation for fine-grained distributions

3.1. Fisher separability

Recall that the classical Fisher discriminant between two classes with means μ_1 and μ_2 is separation of the classes by a hyperplane orthogonal to $\mu_1 - \mu_2$ in the inner product

$$\langle x, y\rangle = (x, S^{-1}y), \tag{9}$$

where $(\cdot,\cdot)$ is the standard inner product and S is the average (or the weighted average) of the sample covariance matrix of these two classes.

Let the data set be preprocessed. In particular, we assume that it is *centralized, normalized, and approximately whitened*. In this case, we use in the definition of Fisher's discriminant the standard inner product instead of $\langle\cdot,\cdot\rangle$.

Definition 1. *A point x is* Fisher separable *from a set $\mathbf{Y} \subset \mathbb{R}^n$ with threshold $\alpha \in (0, 1]$, or α-Fisher separable in short, if inequality*

$$\alpha(x, x) \geq (x, y), \tag{10}$$

holds for all $y \in \mathbf{Y}$.

Definition 2. *A finite set $\mathbf{Y} \subset \mathbb{R}^n$ is* Fisher separable *with threshold $\alpha \in (0, 1]$, or α-Fisher separable in short, if inequality (10) holds for all $x, y \in \mathbf{Y}$ such that $x \neq y$.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Separation of points by simple and explicit inner products (10) is, from the practical point of view, more convenient than general linear separability that can be provided by support vector machines, for example. Of course, linear separability is more general that Fisher separability. This is obvious from the everyday low-dimensional experience, but in high dimensions Fisher separability becomes a generic phenomenon [11,12].

Theorem 1 below is a prototype of most stochastic separation theorems.

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

- The probability distribution has no heavy tails;
- The sets of relatively small volume should not have large probability.

These conditions are not necessary and could be relaxed [15].

In the following Theorem 1 [13] the absence of heavy tails is formalized as the tail cut: the support of the distribution is a subset of 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 inequality:

$$\rho(\boldsymbol{x}) < \frac{C}{r^n V_n(\mathbb{B}_n)}, \tag{11}$$

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 $R \leq 1/(2\alpha)$ decays for $n \to \infty$ in a geometric progression with denominator R/r. Condition $1 > r > 1/(2\alpha)$ is possible only if $\alpha > 0.5$, hence, in Theorem 1 we assume $\alpha \in (0.5, 1]$.

Theorem 1. *[13] 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 (11). Then with probability $p > 1 - \delta$ point $\boldsymbol{x}$ is Fisher-separable from Y with threshold α (10).*

Proof. For a given $\boldsymbol{y}$, the set of such $\boldsymbol{x}$ that $\boldsymbol{x}$ is not α-Fisher separable from $\boldsymbol{y}$ by inequality (10) is a ball given by inequality (10)

$$\left\{ z \;\middle|\; \left\| z - \frac{\boldsymbol{y}}{2\alpha} \right\| < \frac{\|\boldsymbol{y}\|}{2\alpha} \right\}. \tag{12}$$

This is the ball of excluded volume. The volume of the ball (12) 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 $p > 1 - \delta$. □

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 $1/(2\alpha) < r < 1$ under assumption that $1/2 < \alpha \leq 1$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Let, for example, $\alpha = 0.8$, $r = 0.9$, $C = 1$, $\delta = 0.01$. Table 1 shows the upper bounds on $|Y|$ given by Theorem 1 in various dimensions n that guarantees α-Fisher separability of a random point $\boldsymbol{x}$ from Y with probability ≥ 0.99 if the ratio $\rho(\boldsymbol{x})/\rho_{\text{uniform}}$ is bounded by the geometric progression $1/r^n$.

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 $\alpha = 0.8$, $r = 0.9$, $C = 1$ in various dimensions.

n	10	25	50	100	150	200
$\lvert Y\rvert \leq$	0.38	91	8.28×10^5	6.85×10^{13}	5.68×10^{21}	4.70×10^{29}
$\rho(\boldsymbol{x})/\rho_{\text{uniform}} \leq$	2.86	13.9	194	3.76×10^4	7.30×10^6	1.41×10^9

For example, for $n = 100$, we see that for any set with $|Y| < 6.85 \times 10^{13}$ points in the unit ball, and any distribution whose density ρ deviates from the uniform one by a factor at most 3.76×10^4, a random point from this distribution is Fisher-separable (2) with $\alpha = 0.8$ from all points in Y with 99% probability.

If we consider Y as a random set in $\mathbb{B}_n$ that satisfies (11) for each point then with high probability Y is α-Fisher separable (each point from the rest of Y) under some constraints of $|Y|$ from above. From Theorem 1 we get the following corollary.

Corollary 1. *If $Y \subset \mathbb{B}_n$ is a random set $Y = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_{|Y|}\}$ 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:*

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

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

Table 2. The upper bound on $|Y|$ that guarantees α-Fisher's separability of Y with probability ≥ 0.99 according to Corollary 1 for $\alpha = 0.8$, $r = 0.9$, $C = 1$ in various dimensions.

n	10	25	50	100	150	200
$\lvert Y\rvert \leq$	0.61	9.5	910	8.28×10^6	7.53×10^{10}	6.85×10^{14}

Multiple generalizations of Theorem 1 are proven with sharp estimates of $|Y|$ for various families of probability distributions. In this section, we derive the stochastic separation theorems for distributions with cluster structure that violates significantly the assumption (11). For this purpose, in the following subsections we introduce models of cluster structures and modify the notion of Fisher separability to separate clusters. The structure of separation functionals remains explicit with a one-shot non-iterative learning but assimilates both information about the entire distribution and about the cluster being separated.

3.2. Granular models of clusters

The simplest model of a fine-grained distribution of data assumes that the data are grouped into dense clusters and each cluster is located inside a relatively small body (a granule) with random position. Under these conditions, the distributions of data inside the small granules do not matter and may be put out of consideration. What is important, is the geometric characteristics of the granules and their distribution. This is a simple one-level version of the granular data representation [61,62]. The possibility to replace points by compacts in neural network learning was considered by Kainen [63]. He developed the idea that "compacta can replace points". In discussion, we will touch also a promising multilevel hierarchical granular representation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Spherical granules allows a simple straightforward generalization of Theorem 1. Consider spherical granules G_z of radius R with centers $z \in \mathbb{B}_n$:

$$G_z = \{z' \mid \|z' - z\| \leq R\}.$$

Let G_x and G_y be two such granules. Let us reformulate the Fisher separation condition with threshold α for granules:

$$\alpha(x, x') \geq (x, y') \text{ fo all } x' \in G_x,\ y' \in G_y. \tag{13}$$

Elementary geometric reasoning gives that the separability condition (13) holds if x (the center of G_x) does not belong to the ball with radius $\frac{1}{2\alpha}\|y\| + R(1 + \frac{1}{\alpha})$ centered at $\frac{1}{2\alpha}y$:

$$x \notin \left\{ z \;\middle|\; \left\| z - \frac{y}{2\alpha} \right\| < \frac{\|y\|}{2\alpha} + R\left(1 + \frac{1}{\alpha}\right) \right\}. \tag{14}$$

This is analogous to the ball of excluded volume (12) for spherical granules. The difference from (12) is that both z and y are inflated into balls of radius R.

Let $\mathbf{B}$ be the closure of the ball defined in (12):

$$\mathbf{B} = \left\{ z \;\middle|\; \left\| z - \frac{y}{2\alpha} \right\| \leq \frac{\|y\|}{2\alpha} \right\}.$$

Condition (14) implies that the distance between x and $\mathbf{B}$ is at least $R(1 + \frac{1}{\alpha})$. In particular, $\|x - \beta x\| \geq R(1 + \frac{1}{\alpha})$, where β is the largest real number such that $\beta x \in \mathbf{B}$. Then βx belongs to the boundary of $\mathbf{B}$, hence (10) holds as an equality for βx:

$$\alpha(\beta x, \beta x) = (\beta x, y),$$

or, equivalently, $\alpha\beta\|x\|^2 = (x, y)$. Then

$$\alpha(x, x) = \alpha\|x\| \cdot \|x - \beta x\| + \alpha\beta\|x\|^2 \geq \alpha\|x\| \cdot R\left(1 + \frac{1}{\alpha}\right) + (x, y) = (1 + \alpha)R\|x\| + (x, y).$$

Thus, if x satisfies (14) then

$$\alpha(x, x) \geq (1 + \alpha)R\|x\| + (x, y) \text{ that is } \alpha((x, x) - R\|x\|) \geq (x, y) + R\|x\|. \tag{15}$$

Let $x' \in G_x,\ y' \in G_y$. The Cauchy–Schwarz inequality gives $|(x' - x, x)| \leq \|x' - x\|\|x\| \leq R\|x\|$ and $|(y' - y, x)| \leq \|y' - y\|\|x\| \leq R\|x\|$. Therefore, $(x, x') \geq (x, x) - R\|x\|$ and $(x, y) + R\|x\| \geq (x, y')$. Combination of two last inequalities with (15) gives separability (13).

If the point y belongs to the unit ball $\mathbb{B}_n$ then the radius of the ball of excluded volume (14) does not exceed

$$\xi = \frac{1}{2\alpha} + R\left(1 + \frac{1}{\alpha}\right). \tag{16}$$

Further on, the assumption $\xi < 1$ is used.

Theorem 2. *Consider a finite set of spherical granules G_y with radius R and set of centers Y in $\mathbb{B}_n$. Let G_x be a granule with radius R and a randomly chosen center x from a distribution in the unit ball with the bounded probability density $\rho(x)$. Assume that $\rho(x)$ satisfies inequality (11) and the upper estimate of the radius of excluded ball (16) $\xi < 1$. Let $1 > r > \xi$ and*

$$|Y| < \delta \frac{1}{C}\left(\frac{r}{\xi}\right)^n. \tag{17}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Then the separability condition (13) holds for G_x and all G_y ($\boldsymbol{y} \in Y$) with probability $p > 1 - \delta$.

Proof. The separability condition (13) holds for the granule G_x and all G_y ($\boldsymbol{y} \in Y$) if $\boldsymbol{x}$ does not belong to the excluded ball (14) for all $\boldsymbol{y} \in Y$. The volume of the excluded ball is $V = \xi^n V_n(\mathbb{B}_n)$ for each $\boldsymbol{y} \in Y$. The probability that point $\boldsymbol{x}$ belongs to such a ball does not exceed $C\left(\frac{\xi}{r}\right)^n$ in accordance with the boundedness condition (11). Therefore, the probability that $\boldsymbol{x}$ belongs to the union of such balls does not exceed $|Y|C\left(\frac{\xi}{r}\right)^n$. This probability is less than δ if $|Y| < \delta\frac{1}{C}\left(\frac{r}{\xi}\right)^n$. □

Table 3 shows how the number $|Y|$ that guarantees separability (13) of a random granule G_x from an arbitrarily selected set of $|Y|$ granules with probability 0.99 grows with dimension for $\alpha = 0.9$, $r = 0.9$, $C = 1$ and $R = 0.1$.

Table 3. The upper bound on $|Y|$ that guarantees separation of granules G_x and all G_y ($\boldsymbol{y} \in Y$) (13) with probability 0.99 according to Theorem 2 for $\alpha = 0.9$, $r = 0.9$, $C = 1$ and $R = 0.1$ in various dimensions.

n	25	50	100	150	200
$\lvert Y\rvert \leq$	0.55	30	9.26×10^4	2.81×10^8	8.58×10^{11}

The separability condition (13) can be considered as Fisher separability (10) with inflation points to granules. From this point of view, Theorem 2 is a version of Theorem 1 with inflated points. An inflated version of Corollary 1 also exists.

Corollary 2. *Let $Y \subset \mathbb{B}_n$ be a random set $Y = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_{|Y|}\}$. Assume that for each j the density of conditional distribution of vector $\boldsymbol{y}_j$ for any given positions of the other $\boldsymbol{y}_k$ in $\mathbb{B}_n$ exists and satisfies inequality (11). Consider a finite set of spherical granules G_y with radius R and centers $\boldsymbol{y} \in Y$ in $\mathbb{B}_n$. For the radius of the excluded ball (16) assume $\xi < r$, where $r < 1$ is defined in (11). Then, with probability*

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

for every two $\boldsymbol{x}, \boldsymbol{y} \in Y$ ($\boldsymbol{x} \neq \boldsymbol{y}$) the separability condition (13) holds. Equivalently, it holds with probability $p > 1 - \delta$ ($\delta > 0$) if

$$|Y| < \sqrt{\frac{\delta}{C}}\left(\frac{r}{\xi}\right)^{n/2}.$$

This upper border of $|Y|$ grows with n in geometric progression.

The idea of spherical granules implies that, in relation to the entire data set, the granules are more or less uniformly compressed in all directions and their diameter is relatively small (or, equivalently, the granules are inflated points, and this inflation is limited isotropically). Looking around, we can hypothesize quite different properties: in some directions, the granules can have large variety, it can be as large, as variety of the whole set, but the dispersion decays in the sequence of the granule's principal components while the entire set is assumed to be whitened. Large diameter of granules is not an obstacle to the stochastic separation theorems. The following Proposition gives simple but instructive example.

Proposition 1. *Let $1 \geq \alpha > 1/2$, $1 > r > 1/(2\alpha)$, $1 > \delta > 0$. Consider an arbitrary set of N intervals $I_j = [\boldsymbol{u_j}, \boldsymbol{v_j}] \in \mathbb{B}_n$ ($j = 1, \dots, N$). Let $\boldsymbol{x}$ 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 (11) and $N < \frac{\delta}{2C}(2r\alpha)^n$. Then with probability $p > 1 - \delta$ point $\boldsymbol{x}$ is Fisher-separable from any $\boldsymbol{y} \in \cup_j I_j$ with threshold α (10).*

Proof. For given $\boldsymbol{x}$ and α, the Fisher's separability inequality defines a half-space for $\boldsymbol{y}$ (10). An interval $I = [u, v]$ belongs to this half-space if and only if its ends, u and v, belong to it, that is, $\boldsymbol{x}$ is α-Fisher

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

separable from u and v. Therefore, we can apply Theorem 1 to prove α-Fisher separability of $\boldsymbol{x}$ from the set $Y = \{u_j\} \cup \{v_j\}$, $|Y| = 2N$. □

The same statements are true for separation of a point from a set of simplexes of various dimension. For such estimates, only the number of vertices matters.

Consider granules in the form of ellipsoids with decaying sequence of length of the principal axes. Let $d_1 > d_2 > ....$ ($d_i > 0$) be an infinite sequence of the upper bounds for semi-axes. Each ellipsoid granule in $\mathbb{R}^n$ has a center, $\boldsymbol{z}$, an orthonormal basis of principal axes $E = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$, and a sequence of semi-axes, $A = \{a_1 \geq a_2 \geq \ldots \geq a_n\}$ ($d_i \geq a_i > 0$). This ellipsoid is given by the inequality:

$$\mathbf{S}_{z,E,A} = \left\{ \boldsymbol{z}' \left| \sum_{j=1}^{n} \frac{1}{a_j^2} (\boldsymbol{z}' - \boldsymbol{z}, \boldsymbol{e}_j)^2 \leq 1 \right. \right\}. \tag{18}$$

Let the sequence $d_1 > d_2 > ....$ ($d_i > 0$, $d_i \to 0$) be given.

Theorem 3. *Consider a set of N elliptic granules (18) with centers $\boldsymbol{z} \in \mathbb{B}_n$ and $a_i \leq d_i$. Let D be the union of all these granules. Assume that $\boldsymbol{x} \in \mathbb{B}_n$ is a random point from a distribution in the unit ball with the bounded probability density $\rho(\boldsymbol{x}) \leq \rho_{\max}$. Then for positive ε, ς*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') < \varepsilon \text{ for all } \boldsymbol{z}' \in D, \ \ \& \ \ (\boldsymbol{x}, \boldsymbol{x}) > 1 - \varsigma) > 1 - N\rho_{\max} V_n(\mathbb{B}_n) a \exp(-bn), \tag{19}$$

where a and b do not depend on the dimensionality.

In proof of Theorem 3 we construct explicit estimates of probability in (19). This construction (eq. (26) below) is an important part of Theorem 3. It is based on the following lemmas about quasiorthogonality of random vectors.

Lemma 1. *Let $\boldsymbol{e} \in \mathbb{R}^n$ be any normalized vector, $\|\boldsymbol{e}\| = 1$. Assume that $\boldsymbol{x} \in \mathbb{B}_n$ is a random point from a distribution in $\mathbb{B}_n$ with the bounded probability density $\rho(\boldsymbol{x}) \leq \rho_{\max}$. Then, for any $\varepsilon > 0$ the probability*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{e}) \geq \varepsilon) \leq \frac{1}{2} \rho_{\max} V_n(\mathbb{B}_n) (\sqrt{1 - \varepsilon^2})^n. \tag{20}$$

Proof. The inequality $(\boldsymbol{x}, \boldsymbol{e}) \geq \varepsilon$ defines a spherical cap. This spherical cap can be estimated from above by the volume of a hemisphere of radius $\sqrt{1 - \varepsilon^2}$ (Fig. 4). The volume W of this hemisphere is

$$W = \frac{1}{2} V_n(\mathbb{B}_n) (\sqrt{1 - \varepsilon^2})^n$$

The probability that $\boldsymbol{x}$ belongs to this cap is bounded from above by the value $\rho_{max} W$, which gives the estimate (20). □

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

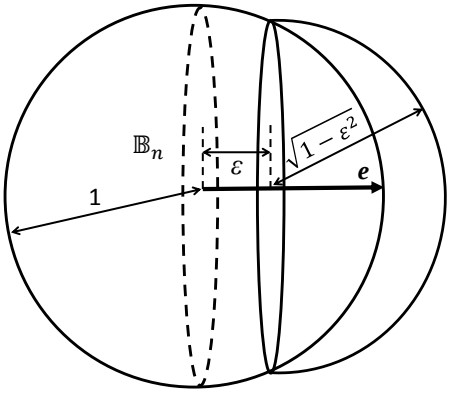

Figure 4. *Approximation of a spherical cap by a hemisphere.* A spherical cap is portion of $\mathbb{B}_n$ cut off by a plane on distance ε from the center. It is approximated from above by a hemisphere of radius $\sqrt{1-\varepsilon^2}$. The vector $\boldsymbol{x}$ should belong to this spherical cap to ensure the inequality $(\boldsymbol{x},\boldsymbol{e}) \geq \varepsilon$.

Lemma 2. *Let $\boldsymbol{e}_1,\dots,\boldsymbol{e}_N \in \mathbb{R}^n$ be normalized vectors, $\|\boldsymbol{e}_i\| = 1$. Assume that $\boldsymbol{x} \in \mathbb{B}_n$ is a random point from a distribution in $\mathbb{B}_n$ with the bounded probability density $\rho(\boldsymbol{x}) \leq \rho_{\max}$. Then, for any $\varepsilon > 0$ the probability*

$$\mathbf{P}((\boldsymbol{x},\boldsymbol{e}_i) \leq \varepsilon \text{ for all } i = 1,\dots,N) \geq 1 - \frac{1}{2} N \rho_{\max} V_n(\mathbb{B}_n)(\sqrt{1-\varepsilon^2})^n \tag{21}$$

Proof. Notice that

$$\mathbf{P}((\boldsymbol{x},\boldsymbol{e}_i) \leq \varepsilon \text{ for all } i = 1,\dots,N) \geq 1 - \sum_i \mathbf{P}((\boldsymbol{x},\boldsymbol{e}_i) \geq \varepsilon).$$

According to Lemma 1, each term in the last sum is estimated from above by the expression $\frac{1}{2}\rho_{\max} V_n(\mathbb{B}_n)(\sqrt{1-\varepsilon^2})^n$ (20). □

It is worth to mention that the term $(\sqrt{1-\varepsilon^2})^n$ decays exponentially when n increases.

Let $\mathbf{S}_{z,E,A}$ be an ellipsoid (18). Decompose a vector $\boldsymbol{x} \in \mathbb{R}^n$ in an orthonormal basis $E = \{\boldsymbol{e}_1,\dots,\boldsymbol{e}_n\}$: $\boldsymbol{x} = \sum_i (\boldsymbol{x},\boldsymbol{e}_i)\boldsymbol{e}_i = \|x\| \sum_i \boldsymbol{e}_i \cos\alpha_i$, where $\cos\alpha_i = (\boldsymbol{x},\boldsymbol{e}_i)/\|\boldsymbol{x}\|$. Notice that $\sum_i \cos^2\alpha_i = 1$ (the n-dimensional Pythagoras theorem).

Lemma 3. *For a given $\boldsymbol{x} \in \mathbb{R}^n$. maximization of a linear functional $(\boldsymbol{x},\boldsymbol{z}')$ on an ellipsoid (18) gives*

$$\max_{\boldsymbol{z}' \in \mathbf{S}_{z,E,A}} (\boldsymbol{x},\boldsymbol{z}') = (\boldsymbol{x},\boldsymbol{z}) + \|\boldsymbol{x}\| \sqrt{\sum_i a_i^2 cos\alpha_i^2}, \tag{22}$$

and the maximizer has the following coordinates in the principal axes:

$$z_i' = z_i + \frac{a_i^2 \cos\alpha_i}{\sqrt{\sum_i a_i^2 \cos\alpha_i^2}}, \tag{23}$$

where $z_i' = (\boldsymbol{z}',\boldsymbol{e}_i)$, and $z_i = (\boldsymbol{z},\boldsymbol{e}_i)$ are coordinates of the vectors $\boldsymbol{z}'$, $\boldsymbol{z}$ in the basis E.

Proof. Introduce coordinates in the ellipsoid $\mathbf{S}_{z,E,A}$ (18): $\Delta_i = z_i' - z_i$. In these coordinates, the objective function is

$$(\boldsymbol{x},\boldsymbol{z}') = (\boldsymbol{x},\boldsymbol{z}) + \|\boldsymbol{x}\| \sum_i \Delta_i \cos\alpha_i.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

For given x, z we have to maximize $\sum_i \Delta_i \cos\alpha_i$ under the equality constraints:

$$F(\Delta_1,\dots,\Delta_n) = \frac{1}{2}\sum_i \frac{\Delta_i^2}{a_i^2} = \frac{1}{2},$$

because the maximizer of a linear functional on a convex compact set belongs to the border of this compact.

The method of Lagrange multipliers gives:

$$\cos\alpha_i = \lambda\frac{\partial F}{\partial \Delta_i} = \lambda\frac{\Delta_i}{a_i^2},\quad \Delta_i = \frac{1}{\lambda}a_i^2\cos\alpha_i.$$

To find the Lagrange multiplier λ, we use the equality constrain again and get

$$\frac{1}{\lambda^2}\sum_i a_i^2\cos^2\alpha_i = 1,\quad \lambda = \pm\sqrt{\sum_i a_i^2\cos^2\alpha_i},$$

where the '+' sign corresponds to the maximum and the '–' sign corresponds to the minimum of the objective function. Therefore, the required maximizer has the form (23) and the corresponding maximal value is given by (22). □

Proof of Theorem 3 . The proof is organized as follows. Select sufficiently small $R > 0$ and find such k that $d_{k+1} < R$. For each elliptic granule select the first k vectors of its principal axes. There will be N vectors of the first axes, N vectors of the second axes, etc. Denote these families of vectors $\mathcal{E}_1$, $\mathcal{E}_2$, ..., $\mathcal{E}_k$: $\mathcal{E}_i$ is a set of vectors of the ith principal axis for granules. Let $\mathcal{E}_0$ be the set of the centers of granules. Select a small $\vartheta > 0$. Use Lemma 2 and find the probability that for all $e \in \mathcal{E}_i$ and for all $i = 1,\dots,k$ the following quasiorthogonality condition holds: $|(x,e)| \le \frac{\vartheta}{\sqrt{k}d_i}$. Under this condition, evaluate the value of the separation functionals (22) in all granules as

$$(x,z') \le (x,z) + \|x\|\sqrt{\sum_i a_i^2 cos\alpha_i^2} \le (x,z) + \sqrt{\vartheta^2 + R^2}, \tag{24}$$

where z is the center of the granule. Indeed,

$$\|x\|^2\sum_i a_i^2 cos\alpha_i^2 \le \sum_{i=1}^{k} d_i^2(x,e_i)^2 + \sum_{i=k+1}^{n}\|x\|^2R^2 cos\alpha_i^2.$$

The quasiorthogonality condition gives that the first sum does not exceed ϑ. Recall that $\|x\| \le 1$ and $\sum_i cos\alpha_i^2 = 1$. Therefore, the second sum does not exceed R^2. This gives us the required estimate (24).

The first term, (x,z) is also small with high probability. This quasiorthogonality of x and N vectors of the centers of granules follows from Lemma 2. It should be noted that the requirement of qusiorthogonality of x to several families of vectors (N centers and kN principal axes) increases the pre-exponential factor in the negative term in (21). This increase can be compensated by a slight increase in the dimensionality because of the exponential factor there.

Let us construct the explicit estimates for given $\varepsilon > 0$, $\varsigma > 0$. Take

$$\vartheta = R = \varepsilon/(1+\sqrt{2}). \tag{25}$$

Under conditions of Theorem 3 several explicit exponential estimates of probabilities hold:

1. Volume of a ball with radius $1-\varsigma$ is $V_n(\mathbb{B}_n)(1-\varsigma)^n$. therefore for probability of x belong to this ball, we have
$$\mathbf{P}((x,x) \le 1-\varsigma) \le \rho_{\max}V_n(\mathbb{B}_n)(1-\varsigma)^n;$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

2. For every $z \in \mathcal{E}_0$,
$$\mathbf{P}((x,z) \geq \vartheta) \leq \rho_{\max}\frac{1}{2}V_n(\mathbb{B}_n)(\sqrt{1-\vartheta^2})^n;$$
3. For every $e \in \mathcal{E}_i$
$$\mathbf{P}\left(|(x,e)| \geq \frac{\vartheta}{\sqrt{k}d_i}\right) \leq \rho_{\max}V_n(\mathbb{B}_n)\left(\sqrt{1-\left(\frac{\vartheta}{\sqrt{k}d_i}\right)^2}\right)^n.$$

Thus, the probability

$$\begin{aligned}&\mathbf{P}\left((x,x) \geq 1-\varsigma \;\&\; (x,z) \leq \vartheta \text{ for all } z \in \mathcal{E}_0 \;\&\; |(x,e)| \leq \frac{\vartheta}{\sqrt{k}d_i} \text{ for all } e \in \mathcal{E}_i,\, i=1,\dots,k\right)\\ &\geq 1-\rho_{\max}V_n(\mathbb{B}_n)\left[(1-\varsigma)^n + \frac{1}{2}N(\sqrt{1-\vartheta^2})^n + N\sum_{i=1}^{k}\left(\sqrt{1-\left(\frac{\vartheta}{\sqrt{k}d_i}\right)^2}\right)^n\right].\end{aligned} \tag{26}$$

If $(x,z) \leq \vartheta$ for all $z \in \mathcal{E}_0$ and $|(x,e)| \leq \frac{\vartheta}{\sqrt{k}d_i}$ for all $e \in \mathcal{E}_i$, $i=1,\dots,k$ then, according to the choice of ϑ (25) and inequality (24), $(x,z') \leq \varepsilon$ for all points from the granules $z' \in D$. Therefore, (26) proves Theorem 3 with explicit estimate of the probability.

If, in addition, $(x,x) \geq 1-\varsigma$, $0 < \alpha \leq 1$ and $\alpha(1-\varsigma) > \varepsilon$ then

$$\alpha(x,x) > (x,z') \text{ for all } z' \in D$$

for all points from the granules $z' \in D$. This is the analogue of α-Fisher separability of point x from elliptic granules. □

Theorem 3 describes stochastic separation of a random point in n-dimensional dataspace from a set of N elliptic granules. For given N probability of α-Fisher separability exponentially approaches 1 with dimensionality growth. Equivalently, for a given probability, the upper bound on the number of granules that guarantees such a separation with this probability grows exponentially with the dimension of the data. We require two properties of the probability distribution: compact support and the existence of a probability density bounded from above. The interplay between the dependence of the maximal density on the dimension (similarly to (11)) and the exponents in the probability estimates (26) determines the estimate of the separation probability.

In Theorem 3 we analyzed separation of a random point from a set of granules but it seems to be much more practical to consider separation of a random granule from a set of granules. For analysis of random granules a joint distributions of the position of the center and the basis of principal axes is needed. Existence of strong dependencies between the position of the center and the directions of principal axes may in special cases destroy the separability phenomenon. For example, if the first principal axis has length 1 or more and is parallel to the vector of the center (i.e. $e_1 = x/\|x\|$) then this granule is not separated even from the origin. On the other hand, independence of these distributions guarantees stochastic separability, as follows from Theorem 4 below. Independence by itself is not needed. The essential condition is that for each orientation of the granule, the position of its center remains rather uncertain.

Theorem 4. *Consider a set of N elliptic granules (18) with centers $z \in \mathbb{B}_n$ and $a_i \leq d_i$. Let D be the union of all these granules. Assume that $x \in \mathbb{B}_n$ is a random point from a distribution in the unit ball with the bounded probability density $\rho(x) \leq \rho_{\max}$. Let x be a center of a random elliptic granule $\mathbf{S}_x = \mathbf{S}_{x,E_x,A_x}$ (18). Assume*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

that for any basis of principal axes E and sequence of semi-axes $A = \{a_i\}$ $(a_i \le d_i)$ the conditional distribution of the centers of granules $\boldsymbol{x}$ given $E_x = E$, $A_x = A$ has a density in $\mathbb{B}_n$ uniformly bounded from above:

$$\rho(\boldsymbol{x} \mid E_x = E,\ A_x = A) \le \rho_{\max}$$

and $\rho_{\max}$ does not depend on E_x, A_x Then for positive ε, ς

$$\begin{aligned}\mathbf{P}((\boldsymbol{x},\boldsymbol{z}') \le \varepsilon \text{ for all } \boldsymbol{z}' \in D \ \ \&\ \ (\boldsymbol{x},\boldsymbol{x}') \ge (\boldsymbol{x},\boldsymbol{x}) - \varepsilon \text{ for all } \boldsymbol{x}' \in \mathbf{S}_x \ \ \&\ \ (x,x) \ge 1-\varsigma \\ \ge 1 - N\rho_{\max} V_n(\mathbb{B}_n) a \exp(-bn),\end{aligned} \tag{27}$$

where a and b do not depend on the dimensionality.

In the proof of Theorem 4 we estimate the probability (27) by a sum of decaying exponentials, which give explicit formulas for a and b as was done for Theorem 3 in (26).

Proof. We will prove (27) for an elipsoid $\mathbf{S}_x$ (18) with given (not random) basis E and semiaxes $a_i \le d_i$, and with a random center $\boldsymbol{x} \in \mathbb{B}_n$ assuming that the distribution density of $\boldsymbol{x}$ is bounded from above by $\rho_{\max}$.

Select sufficiently small $R > 0$ and find such k that $d_{k+1} < R$. For each granule, including $\mathbf{S}_x$ with the center $\boldsymbol{x}$ select the first k vectors of its principal axes. There will be $N+1$ vectors of the first axes, $N+1$ vectors of the second axes, etc. Denote these families of vectors $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_k$: $\mathcal{E}_i$ is a set of vectors of the ith principal axis for all granules, $\mathbf{S}_x$. Let $\mathcal{E}_0$ be the set of of the centers of granules (excluding the centre $\boldsymbol{x}$ of the granule $\mathbf{S}_x$.)

For a given $\vartheta > 0$ the following estimate of probability holds (analogously to (26)).

$$\begin{aligned}&\mathbf{P}\left((\boldsymbol{x},\boldsymbol{x}) \ge 1-\vartheta \ \&\ (\boldsymbol{x},\boldsymbol{z}) \le \vartheta \text{ for all } \boldsymbol{z} \in \mathcal{E}_0 \ \&\ |(\boldsymbol{x},\boldsymbol{e})| \le \frac{\vartheta}{\sqrt{k}d_i} \text{ for all } \boldsymbol{e} \in \mathcal{E}_i,\ i = 1,\dots,k\right) \\ &\ge 1 - \rho_{\max} V_n(\mathbb{B}_n)\left[(1-\vartheta)^n + \frac{1}{2}N(\sqrt{1-\vartheta^2})^n + (N+1)\sum_{i=1}^{k}\left(\sqrt{1-\left(\frac{\vartheta}{\sqrt{k}d_i}\right)^2}\right)^n\right].\end{aligned} \tag{28}$$

If $(\boldsymbol{x},\boldsymbol{x}) \ge 1-\vartheta$ and $(\boldsymbol{x},\boldsymbol{z}) \le \vartheta$ for all $\boldsymbol{z} \in \mathcal{E}_0$, and $|(\boldsymbol{x},\boldsymbol{e})| \le \frac{\vartheta}{\sqrt{k}d_i}$ for all $\boldsymbol{e} \in \mathcal{E}_i$, $i = 1,\dots,k$, then by (24)

$$(\boldsymbol{x},\boldsymbol{z}') \le \vartheta + \sqrt{\vartheta^2 + R^2} \ \ \&\ \ (\boldsymbol{x},\boldsymbol{x}') \ge 1 - \vartheta - \sqrt{\vartheta^2 + R^2} \text{ for all } \boldsymbol{z}' \in D, \boldsymbol{x}' \in \mathbf{S}_x.$$

Therefore, if we select $R = \frac{\varepsilon}{1+\sqrt{2}}$ and $\vartheta = \min\left\{\varsigma, \frac{\varepsilon}{1+\sqrt{2}}\right\}$, then the estimate (28) proves Theorem 4. Additionally, for this choice, $(\boldsymbol{x},\boldsymbol{x}') \ge 1-\varepsilon$ for all $\boldsymbol{x}' \in \mathbf{S}_x$. Therefore, if $\varepsilon < \frac{\alpha}{1+\alpha}$, then $\alpha(\boldsymbol{x},\boldsymbol{x}') > (\boldsymbol{x},\boldsymbol{z}')$ for all $\boldsymbol{z}' \in D$ and $\boldsymbol{x}' \in \mathbf{S}_x$ with probability estimated in (28). This result can be considered as α-Fisher separability of elliptic granules in high dimensions with high probability. □

Note that the the proof does not actually use that $d_i \to 0$. All that we use that $\limsup_{i\to\infty} d_i < R$ for $R = \frac{\epsilon}{1+\sqrt{2}}$, where $\epsilon < \frac{\alpha}{1+\alpha}$. Hence the proof remains valid whenever $\limsup_{i\to\infty} d_i < \frac{\alpha}{(1+\sqrt{2})(1+\alpha)}$.

It may be useful to formulate a version of Theorem 4 when $\mathbf{S}_{\boldsymbol{x}}$ is the granule of an arbitrary (non-random) shape but with a random center as a separate Proposition.

Proposition 2. *Let D be the union of N elliptic granules (18) with centers in $\mathbb{B}_n$ with $a_i \le d_i$. Let $\mathbf{S}_{z,E,A}$ be one more such granule. Let $\boldsymbol{x} \in \mathbb{B}_n$ be a random point from a distribution in the unit ball with the bounded probability density $\rho(x) \le \rho_{max}$. Let $\mathbf{S}_{\boldsymbol{x}} = \mathbf{S}_{z,E,A} + (\boldsymbol{x} - \boldsymbol{z})$ be the granule $\mathbf{S}_{z,E,A}$ shifted such that its center becomes $\boldsymbol{x}$. Then Theorem 4 is true for $\mathbf{S}_{\boldsymbol{x}}$.*

The proof is the same as the proof of Theorem 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

The estimates (26) and (28) are far from being sharp. Detailed analysis for various classes of distributions may give better estimates as it was done for separation of finite sets [15]. This work needs to be done for separation of granules as well.

3.3. Superstatistic presentation of "granules"

The alternative approach to the granular structure of the distributions are *soft clusters*. They can be studies in the frame of *superstatistical* approach with representation of data distribution by a random mixture of distributions of points in individual clusters. We start with the following remark.

Notice that Proposition 2 has the following easy corollary.

Corollary 3. *Let $\mathbf{S}_x$ and D be as in Proposition 2. Let $\boldsymbol{x}'$ and $\boldsymbol{z}'$ be the points selected uniformly at random from $\mathbf{S}_x$ and D, correspondingly. Then for positive ϵ, ζ*

$$\mathbf{P}((\boldsymbol{x},\boldsymbol{z}') \leq \epsilon \ \&\ (\boldsymbol{x},\boldsymbol{x}') \geq (\boldsymbol{x},\boldsymbol{x}) - \epsilon \ \&\ (\boldsymbol{x},\boldsymbol{x}) \geq 1-\zeta) \geq 1 - N\rho_{max}V_n(\mathbb{B}_n)a\exp(-bn),$$

where the constants a, b are the same as in Theorem 4.

Proof. Let $f(n) = N\rho_{\max}V_n(\mathbb{B}_n)a\exp(-bn)$. Let $A \subset \mathbb{B}_n$ be the set of $\boldsymbol{x}$ such that (27) holds. Proposition 2 states that $\mathbf{P}(\boldsymbol{x} \in A) \geq 1 - f(n)$. Let E be the event that $(\boldsymbol{x},\boldsymbol{z}') \leq \epsilon$ & $(\boldsymbol{x},\boldsymbol{x}') \geq (\boldsymbol{x},\boldsymbol{x}) - \epsilon$ & $(\boldsymbol{x},\boldsymbol{x}) \geq 1-\zeta$. By the law of total probability,

$$\begin{aligned}\mathbf{P}(E) &= \mathbf{P}(E|\boldsymbol{x}\in A)\mathbf{P}(\boldsymbol{x}\in A) + \mathbf{P}(E|\boldsymbol{x}\notin A)\mathbf{P}(\boldsymbol{x}\notin A)\\ &\geq \mathbf{P}(E|\boldsymbol{x}\in A)\mathbf{P}(\boldsymbol{x}\in A) = 1\cdot\mathbf{P}(\boldsymbol{x}\in A) \geq 1 - f(n).\end{aligned}$$

□

Corollary 3 is weaker than Proposition 2. While Proposition 2 states that, with probability at least $1 - f(n)$, *the whole* granule $\mathbf{S}_x$ can be separated from *all* points in D, Corollary 3 allows for the possibility that there could be a small portions of $\mathbf{S}_x$ and D which are not separated from each other. As we will see below, this weakening allows us to prove the result in much greater generality, where the uniform distribution in granules is replaced by much more general log-concave 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 $K = \{z \in \mathbb{R}^n \,|\, \rho(z) > 0\}$ is convex and $g(z) = -\log(\rho(z))$ is a convex function on K. For example, the uniform distribution in any full-dimensional subset of $\mathbb{R}^n$ (and in particular uniform distribution in granules (18)) has a log-concave density.

We say that ρ is whitened, or *isotropic*, if $\mathbb{E}[\boldsymbol{x}] = \mathbf{0}$, and

$$\mathbb{E}[(\boldsymbol{x},\theta)^2)] = 1 \qquad \forall\theta \in \mathbb{S}^{n-1}, \tag{29}$$

where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. Equation (29) 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, hence every log-concave random vector $\boldsymbol{x}$ can be represented as

$$\boldsymbol{x} = \Sigma\boldsymbol{y} + \boldsymbol{x}_0, \tag{30}$$

where $\boldsymbol{x}_0 = \mathbb{E}[\boldsymbol{x}]$, Σ is (non-random) matrix and $\boldsymbol{y}$ is some isotropic log-concave random vector.

An example of standard normal distribution shows that the support of isotropic log-concave distribution may be the whole $\mathbb{R}^n$. However, such distributions are known to be concentrated in a ball of radius $\sqrt{n}(1+\delta)$ with high probability.

Specifically, [64, Theorem 1.1] implies that for any $\delta \in (0,1)$ and any isotropic log-concave random vector in $\mathbb{R}^n$,

$$\mathbf{P}(\|\boldsymbol{x}\| \leq (1+\delta)\sqrt{n}) \geq 1 - c\exp(-c'\delta^3\sqrt{n}) \tag{31}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

where $c, c' > 0$ are some absolute constants. Note that we have $\sqrt{n}$ but not n in the exponent, and this cannot be improved without requiring extra conditions on the distribution. 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 K. [64, Theorem 1.1] also implies that

$$\mathbf{P}(\|\boldsymbol{x}\| \leq (1+\delta)\sqrt{n}) \geq 1 - c\exp(-c'\delta^4 n) \tag{32}$$

for any $\delta \in (0,1)$, and any isotropic strongly log-concave random vector $\boldsymbol{x}$ in $\mathbb{R}^n$.

Fix some $\delta > 0$ and infinite sequence $d = (d_1 > d_2 > \dots)$ with each $d_i > 0$ and $d_i \to 0$. Let us call log-concave random vector $\boldsymbol{x}$ (δ, d)-admissible if set $\Sigma \cdot B(\mathbf{0}, (1+\delta)\sqrt{n}) + \boldsymbol{x}_0$ is a subset of some ellipsoid $\mathbf{S}_{\boldsymbol{x}_0, E, A}$ (18), where Σ and $\boldsymbol{x}_0$ are defined in (30) and $B(\mathbf{0}, (1+\delta)\sqrt{n})$ is the ball with center $\mathbf{0}$ and radius $(1+\delta)\sqrt{n}$. Then (31) and (32) imply that $\boldsymbol{x} \in \mathbf{S}_{\boldsymbol{x}_0, E, A}$ with high probability. In combination with Proposition 2, this implies the following results.

Proposition 3. *Let $\delta > 0$ and infinite sequence $d = (d_1 > d_2 > \dots)$ with each $d_i > 0$ and $d_i \to 0$ be fixed. Let $\boldsymbol{x} \in \mathbb{B}_n$ be a random point from a distribution in the unit ball with the bounded probability density $\rho(x) \leq \rho_{max}$. Let $\boldsymbol{x}''$ be a point selected from some (δ, d)-admissible log-concave distribution, and let $\boldsymbol{x}' = \boldsymbol{x}'' - \mathbb{E}[\boldsymbol{x}''] + \boldsymbol{x}$. Let $\boldsymbol{z}'$ be the point selected from a mixture of N (δ, d)-admissible log-concave distributions with centers in $\mathbb{B}_n$. Then for positive ϵ, ζ*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon \;\&\; (\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \;\&\; (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta) \geq 1 - N\rho_{max} V_n(\mathbb{B}_n) a \exp(-bn) - 2c\exp(-c'\delta^3\sqrt{n}),$$

for some constants a, b, c, c' that do not depend on the dimensionality.

Proof. If follows from (31) and (δ, d)-admissibility of the distribution from which $\boldsymbol{x}''$ has been selected that

$$\mathbf{P}(\boldsymbol{x}' \notin \mathbf{S}_0) \leq c\exp(-c'\delta^3\sqrt{n})$$

for some ellipsoid $\mathbf{S}_0$ (18). Similarly, since $\boldsymbol{z}'$ is selected from a mixture of N (δ, d)-admissible log-concave distributions, we have

$$\mathbf{P}\left(\boldsymbol{z}' \notin \bigcup_{i=1}^{N} \mathbf{S}_i\right) \leq c\exp(-c'\delta^3\sqrt{n})$$

for some ellipsoids $\mathbf{S}_1, \dots, \mathbf{S}_N$ (18). Let E be the event that $(\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon \;\&\; (\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \;\&\; (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta$. If E does not happen than either (i) $\boldsymbol{x}' \notin \mathbf{S}_0$, or (ii) $\boldsymbol{z}' \notin \bigcup_{i=1}^{N} \mathbf{S}_i$, or (iii) $\boldsymbol{x}' \in \mathbf{S}_0$ and $\boldsymbol{z}' \in \bigcup_{i=1}^{N} \mathbf{S}_i$, but E still does not happen. The probabilities of (i) and (ii) are at most $c\exp(-c'\delta^3\sqrt{n})$, while the probability of (iii) is at most $N\rho_{\max} V_n(\mathbb{B}_n) a \exp(-bn)$ by Proposition 2. □

Exactly the same proof in combination with (32) imply the following version for strongly log-concave distributions.

Proposition 4. *Let $\delta, \gamma > 0$ and infinite sequence $d = (d_1 > d_2 > \dots)$ with each $d_i > 0$ and $d_i \to 0$ be fixed. Let $\boldsymbol{x} \in \mathbb{B}_n$ be a random point from a distribution in the unit ball with the bounded probability density $\rho(x) \leq \rho_{max}$. Let $\boldsymbol{x}''$ be a point selected from some (δ, d)-admissible γ-SLC distribution, and let $\boldsymbol{x}' = \boldsymbol{x}'' - \mathbb{E}[\boldsymbol{x}''] + \boldsymbol{x}$. Let $\boldsymbol{z}'$ be the point selected from a mixture of N (δ, d)-admissible γ-SLC distributions with centers in $\mathbb{B}_n$. Then for positive ϵ, ζ*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon \;\&\; (\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \;\&\; (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta) \geq 1 - N\rho_{max} V_n(\mathbb{B}_n) a \exp(-bn) - 2c\exp(-c'\delta^4 n),$$

for some constants a, b, c, c' that do not depend on the dimensionality.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

3.4. The superstatistic form of the prototype stochastic separation theorem

Theorem 1 evaluates the probability that a random point $\boldsymbol{x} \in \mathbb{B}_n$ with bounded probability density is α-Fisher separable from an eponentially large finite set Y and demonstrates that under some natural conditions this probability tends to zero when dimension n tends to ∞. This phenomenon has a simple explanation: for any $\boldsymbol{y} \in \mathbb{B}_n$ the set of such $\boldsymbol{x} \in \mathbb{B}_n$ that $\boldsymbol{x}$ is not α-Fisher separable from $\boldsymbol{y}$ is a ball with radius $\|\boldsymbol{y}\|/(2\alpha) < 1$ and the fraction of this volume in $\mathbb{B}_n$ decays as

$$\left(\frac{\|\boldsymbol{y}\|}{2\alpha}\right)^n.$$

These arguments can be generalized with some efforts for the situation when we consider an elliptic granule instead of a random point $\boldsymbol{x}$ and an *arbitrary* probability distribution instead of a finite set Y. Instead of the estimate of the probability of a point $\boldsymbol{x}$ falling into a the ball of excluded volume (12) we use the following Proposition for separability of a random point $\boldsymbol{x}'$ of a granule $\mathbf{S}_{\boldsymbol{x}}$ with a random center $\boldsymbol{x}$ from an arbitrary point $\boldsymbol{z}' \in \mathbb{B}_n$.

Proposition 5. *Let $\mathbf{S}_{\boldsymbol{x}}$ be the granule defined in Proposition 2. Let $\boldsymbol{x}'$ be the point selected uniformly at random from $\mathbf{S}_{\boldsymbol{x}}$. Let $\boldsymbol{z}' \in \mathbb{B}_n$ be an arbitrary (non-random) point. Then for positive ϵ, ζ*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta) \geq 1 - \rho_{max} V_n(\mathbb{B}_n) a \exp(-bn),$$

where the constants a, b do not depend on the dimensionality.

Proof. The fact that

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta) \geq 1 - \rho_{\max} V_n(\mathbb{B}_n) a \exp(-bn)$$

is proved in Theorem 4, while the fact that

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon) \geq 1 - \rho_{\max} V_n(\mathbb{B}_n) a \exp(-bn)$$

follows from Lemma 1. □

Propositions 3 and 4 can be straightforwardly generalized in the same way

Proposition 6. *Let $\delta > 0$ and infinite sequence $d = (d_1 > d_2 > \dots)$ with each $d_i > 0$ and $d_i \to 0$ be fixed. Let $\boldsymbol{x} \in \mathbb{B}_n$ be a random point from a distribution in the unit ball with the bounded probability density $\rho(x) \leq \rho_{max}$. Let $\boldsymbol{x}''$ be a point selected from some (δ, d)-admissible log-concave distribution, and let $\boldsymbol{x}' = \boldsymbol{x}'' - \mathbb{E}[\boldsymbol{x}''] + \boldsymbol{x}$. Let $\boldsymbol{z}' \in \mathbb{B}_n$ be an arbitrary (non-random) point. Then for positive ϵ, ζ*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta) \geq 1 - \rho_{max} V_n(\mathbb{B}_n) a \exp(-bn) - c \exp(-c' \delta^3 \sqrt{n}),$$

for some constants a, b, c, c' that do not depend on the dimensionality.

Proposition 7. *Let $\delta, \gamma > 0$ and infinite sequence $d = (d_1 > d_2 > \dots)$ with each $d_i > 0$ and $d_i \to 0$ be fixed. Let $\boldsymbol{x} \in \mathbb{B}_n$ be a random point from a distribution in the unit ball with the bounded probability density $\rho(x) \leq \rho_{max}$. Let $\boldsymbol{x}''$ be a point selected from some (δ, d)-admissible γ-SLC distribution, and let $\boldsymbol{x}' = \boldsymbol{x}'' - \mathbb{E}[\boldsymbol{x}''] + \boldsymbol{x}$. Let $\boldsymbol{z}' \in \mathbb{B}_n$ be an arbitrary (non-random) point. Then for positive ϵ, ζ*

$$\mathbf{P}((\boldsymbol{x}, \boldsymbol{z}') \leq \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}') \geq (\boldsymbol{x}, \boldsymbol{x}) - \epsilon \ \&\ (\boldsymbol{x}, \boldsymbol{x}) \geq 1 - \zeta) \geq 1 - \rho_{max} V_n(\mathbb{B}_n) a \exp(-bn) - c \exp(-c' \delta^4 n),$$

for some constants a, b, c, c' that do not depend on the dimensionality.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

We remark that because Propositions 5, 6 and 7 hold for an arbitrary (non-random) point $z' \in \mathbb{B}_n$, they also hold for point selected from *any* probability distribution within $\mathbb{B}_n$, and in particular if point z' selected uniformly at random from *any* set $D \subset \mathbb{B}_n$.

3.5. Compact embedding of patterns and hierarchical universe

Stochastic separation theorems tell us that in large dimensions, randomly selected data points (or clusters of data) can be separated by simple and explicit functionals from an existing data set with high probability, as long as the data set is not too large (or the number of data clusters is not too large). The number of data points (or clusters) allowed in conditions of these theorems is bounded from above by an exponential function of dimension. Such theorems for data points (see, for example, Teorem 1 and [15]) or clusters (Theorems 2 –4) are valid for broad families of probability distributions. Explicit estimations of probability to violate the separability property were found.

There is a circumstance that can devalue this (and many other) probabilistic results in high dimension. We almost never know the probability of a multivariate data distribution beyond strong simplification assumptions. In the postclassical world, observations cannot really help because we never have enough data to restore the probability density (again, strong simplification like independence assumption or dimensionality reduction can help, but this is not a general multidimensional case). A radical point of view is possible, according to which there is no such thing as a general multivariate probability distribution, since it is unobservable.

In the infinite-dimensional limit the situation can look simpler: instead of finite but small probabilities that decrease and tend to zero with increasing dimension (like in (26) and (28)) some statements become generic and hold "almost always". Such limits for concentrations on spheres and their equators were discussed by Lévy [65] as an important part of the measure concentration effects. In physics, this limit corresponds to the so-called thermodynamic limit of statistical mechanics [66,67]. In the infinite-dimensional limit many statements about high or low probabilities transform into 0-1 laws: something happens almost always or almost newer. The original Kolmogorov 0-1 law states, roughly speaking, that an event that depends on an infinite collection of independent random variables but is independent of any finite subset of these variables, has probability zero or one (for precise formulation we refer to the monograph [68]). The infinite-dimensional 0-1 asymptotic might bring more light and be more transparent than the probabilistic formulas.

From the infinite-dimensional point of view, the 'elliptic granule' (18) with decaying sequence of diameters $d_1 > d_2 > ....$ ($d_i > 0$, $d_i \to 0$) is a compact. The specific elliptic shape used in Theorem 3 is not much important and many generalization are possible for the granules with decaying sequence of diameters. The main idea, from this point of view, is compact embedding of specific patterns into general population of data. This point of view was influenced by the hierarchy of Sobolev Embedding Theorems where the balls of embedded spaces appear to be compact in the image space.

The finite-dimensional hypothesis about granular structure of the data sets can be transformed into the infinite-dimensional view about compact embedding: the patterns correspond to the compact subsets of the dataspace. Moreover, this hypothesis can be extended to the hypothesis about hierarchical structure (Fig. 5): The data that correspond to a pattern also have the intrernal granular structure. To reveal this structure, we can apply centralization and whitening to a granule. After that, the granule will transform into a new unit ball, the external set (the former 'Universe') will typically become infinitely far ('invisible') and the internal structure can be seeking in the form of collection of compact granules in new topology.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

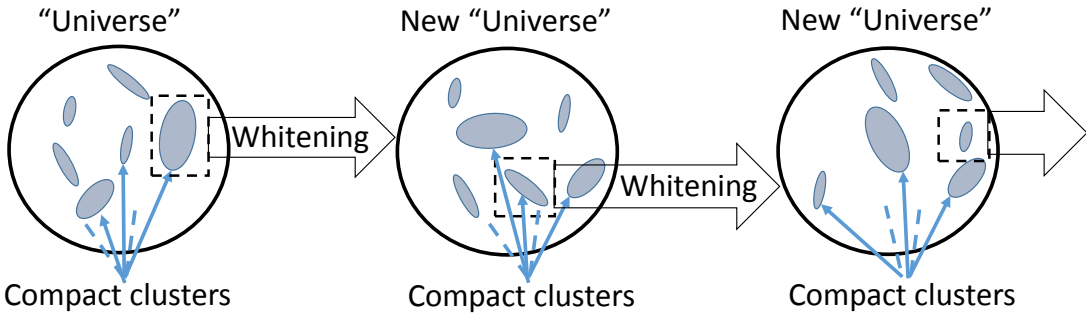

Figure 5. *Hierarchical Universe.* Each pattern is represented by a compact set embedded in the data universe. When we select this compact and apply whitening, it becomes a new universe and we see a set of compact patterns inside, etc.

It should be stressed that this vision is not a theorem. It is proposed instead of typical dominance of smooth or even uniform distributions that populate theoretical studies in machine learning. On another hand, hierarchical structure was observed in various data analytics exercises: if there exists a natural semantic structure then we expect that data have the corresponding cluster structure. Moreover, various preprocessing operations make this structure more visible (see, for example, discussion of preprocessing in Section 2).

The compact embedding idea was recently explicitly used in data analysis (see, for example, [69–71]).

The infinite-dimensional representation and compact embedding hypothesis brings light to the very popular phenomenon of vulnerability of AI decisions in high-dimension world. According to recent research, such vulnerability seems to be a generic property of various formalizations of learning and attack processes in high-dimensional systems [72,73].

Let Q be an infinite-dimensional Banach space. The patterns, or representations a pattern, or their images in an observer systems, etc. are modeled below by compact subsets of Q.

Theorem 5 (Theorem of high-dimensional vulnerability)**.** *Consider two compact sets, $K_{0,1} \subset Q$. For* almost every $\boldsymbol{y} \in Q$ *there exists such continuous linear functional $\boldsymbol{l}$ on Q, $\boldsymbol{l} \in Q^*$, that*

$$\boldsymbol{l}(\boldsymbol{x_1} - \boldsymbol{x_0}) > 0 \ \textit{for all} \ \ \boldsymbol{x_0} \in K_0, \boldsymbol{x_1} \in (K_1 + \boldsymbol{y}). \tag{33}$$

In particular, for every $\varepsilon > 0$ there exist such $\boldsymbol{y} \in Q$ and continuous linear functional $\boldsymbol{l}$ on Q, $\boldsymbol{l} \in Q^*$, that $\|\boldsymbol{y}\| < \varepsilon$ and (33) holds. If (33) holds, then $K_0 \cap (K_1 + \boldsymbol{y}) = \emptyset$. The perturbation $\boldsymbol{y}$ takes K_1 out of the intersection with K_0. Moreover, linear separation of K_0 and perturbed K_1 (i.e. $(K_1 + \boldsymbol{y})$) is possible for almost always (33) (for almost any perturbation).

The definition of "almost always" is clarified in detail in Appendix A. The set of exclusions, i.e. the perturbations that do not satisfy (33) in Theorem 5, is completely thin in the following sense, according to Definition A1. A set $Y \subset Q$ is completely thin, if for any compact space K the set of continuous maps $\Psi : K \to Q$ with non-empty intersection $\Psi(K) \cap Y \neq \emptyset$ is set of first Bair category in the Banach space $C(K, Q)$ of continuous maps $K \to Q$ equipped by the maximum norm.

Proof of Theorem 5 . Let $\overline{\text{co}}(V)$ be a closed convex hull of a set $V \subset Q$. The following sets are convex compacts in Q: $\overline{\text{co}}(K_0)$, $\overline{\text{co}}(K_1)$, and $\overline{\text{co}}(K_0) - \overline{\text{co}}(K_1)$. Let

$$\boldsymbol{y} \notin (\overline{\text{co}}(K_0) - \overline{\text{co}}(K_1)). \tag{34}$$

Then the set $\overline{\text{co}}(K_1) + \boldsymbol{y} - \overline{\text{co}}(K_0)$ does not contain zero. It is a convex compact set. According to the Hahn–Banach separation theorem [74], there exists a continuous linear separating functional $\boldsymbol{l} \in Q^*$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

that separates the convex compact $\overline{\text{co}}(K_1) + \boldsymbol{y} - \overline{\text{co}}(K_0)$ from 0. The same functional separates its subset, $K_1 + \boldsymbol{y} - K_0$ from zero, as required.

The set of exclusions, $\overline{\text{co}}(K_0) - \overline{\text{co}}(K_1)$ (see (34)) is a compact convex set in Q. According to Riesz's theorem, it is nowhere dense in Q [74]. Moreover, for any compact space K the set of continuous maps $\Psi : K \to Q$ with non-empty intersection $\Psi(K) \cap Y \neq \emptyset$ is a nowhere dense subset of Banach space $C(K, Q)$ of continuous maps $K \to Q$ equipped by the maximum norm.

Indeed, let $\Psi(K) \cap Y \neq \emptyset$. The set $\Psi(K)$ is compact. Therefore, as it is proven, an arbitrary small perturbation $\boldsymbol{y}$ exists that takes $\Psi(K)$ out of the intersection with Y: $(\Psi(K) + \boldsymbol{y}) \cap Y = \emptyset$. The minimal value

$$\min_{\boldsymbol{x_1} \in (\Psi(K)+\boldsymbol{y}),\, \boldsymbol{x_2} \in Y} \|\boldsymbol{x_1} - \boldsymbol{x_2}\| = \delta > 0$$

exists and is positive because compactness $(\Psi(K) + \boldsymbol{y})$ and Y.

Therefore, $\Psi'(K) \cap Y = \emptyset$ for all Ψ' from a ball of maps in $C(K, Q)$

$$\left\{ \Psi' \;\middle|\; \|\Psi' - (\Psi + \boldsymbol{y})\| < \frac{\delta}{2} \right\}$$

This proofs that the set of continuous maps $\Psi : K \to Q$ with non-empty intersection $\Psi(K) \cap Y$ is a nowhere dense subset of $C(K, Q)$. Thus, the set of exclusions is completely thin. □

The following Corollary is simple but it may seem counterintuitive:

Corollary 4. *A compact set $K_0 \subset Q$ can be separated from a countable set of compacts $K_i \subset Q$ by a single and arbitrary small perturbation $\boldsymbol{y}$ ($\boldsymbol{y} < \varepsilon$ for an arbitrary $\varepsilon > 0$):*

$$(K_0 + \boldsymbol{y}) \cap K_i = \emptyset.$$

Almost all perturbations $\boldsymbol{y} \in Q$ provide this separation and the set of exclusions is completely thin.

Proof. First, refer to Theorem 5 (for separability of K_0 from one K_i). Then mention that countable union of completely thin set of exclusions is completely thin, whereas the whole Q is not (according to the Bair theorem, Q is not a set of first category). □

Separability theorems for compactly embedded patterns might explain why the vulnerability to adversarial perturbations and stealth attacks is typical for high-dimensional AI systems based on data [72]. Two properties are important simultaneously: high dimensionality and compactness of patterns.

4. Multi-correctors of AI systems

In this section, we present the construction of error correctors for multidimensional AI systems operating in a multidimensional world. It combines a set of elementary correctors (Fig. 2) and a dispatcher that distributes the tasks between them. The population of possible errors is presented as a collection of clusters. Each elementary corrector works with its own cluster of situations with a high risk of error. It includes a binary classifier that separates that cluster from the rest of situations. Dispatcher is based on an unsupervised classifier that performs cluster-analysis of errors, selects the most appropriate cluster for each operating situation, transmits the signals for analysis to the corresponding elementary corrector and requests the correction decision from it (Fig. 6).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

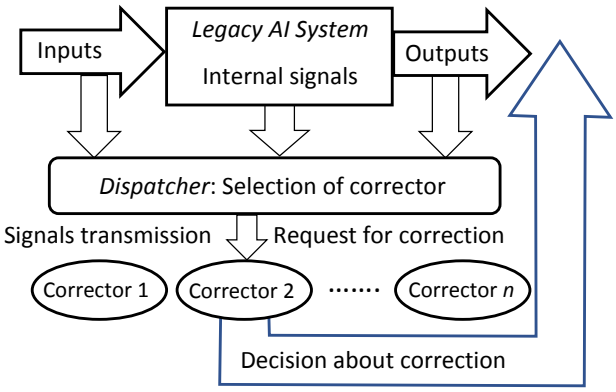

Figure 6. *Multi-corrector – a system of elementary correctors, controlled by the dispatcher, for reversible correction of legacy AI systems.* The dispatcher receives signals from the AI system to be corrected (input signals of the AI system, internal signals generated in the decision-making process, and output signals), and selects from the elementary correctors the one that most corresponds to the situation and will process this situation to resolve the issue of correction. The decision rule, on the basis of which the dispatcher distributes situations between elementary correctors, is formed as a result of a cluster analysis of situations with diagnosed errors. Each elementary corrector processes situations from one cluster. When new errors are detected, the dispatcher modifies the definition of clusters. Cluster models are prepared and modified using the data stream online algorithms.

In brief, operation of multi-correctors (Fig. 6) can be described as follows:

- The correction system is organized as a set of elementary correctors, controlled by the dispatcher;
- Each elementary corrector 'owns' a certain class of errors and includes a binary classifier that separates situations with a high risk of these errors, which it owns, from other situations;
- For each elementary corrector, a modified rule is set for operating of the corrected AI system in a situation with a high risk of error diagnosed by the classifier of this corrector;
- The input to the corrector is a complete vector of signals, consisting of the input, internal and output signals of the corrected artificial intelligence system, (as well as, if available, any other available attributes of the situation);
- The dispatcher distributes situations between elementary correctors;
- The decision rule, based on which the dispatcher distributes situations between elementary correctors, is formed as a result of cluster analysis of situations with diagnosed errors;
- Cluster analysis of situations with diagnosed errors is performed using an online algorithm;
- Each elementary corrector owns situations with errors from a single cluster;
- After receiving of a signal about the detection of new errors, the dispatcher modifies the definition of clusters according to the selected online algorithm and accordingly modifies the decision rule, on the basis of which situations are distributed between elementary correctors;
- After receiving a signal about detection of new errors, the dispatcher chooses an elementary corrector, which must process the situation, and the classifier of this corrector learns according to a non-iterative explicit rule.

Flowcharts of these operations are presented in Appendix B. Multi-correctors satisfy the following requirements:

1. Simplicity of construction;
2. Correction should not damage the existing skills of the system;
3. Speed (fast non-iterative learning);
4. Correction of new errors without destroying previous corrections.

For implementation of this structure, the construction of classifiers for elementary correctors and the online algorithms for clustering should be specified. For elementary correctors many choices are possible, for example:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

- Fisher's linear discriminant is simple, robust, and is proven to be applicable in high-dimensional data analysis [13,15];
- Kernel versions of non-iterative linear discriminants extend the area of application of the proposed systems, their separability properties were quantified and tested [18];
- Decision trees of mentioned elementary discriminants with bounded depth. These algorithms require small (bounded) number of iterations.

The population of clustering algorithms is huge [75]. The first choice for testing of multi-correctors [76] was partition around centroids by k means algorithm. The closest candidates for future development are multi-centroid algorithms that present clusters by networks if centroids (see, for example, [77]. This approach to clustering meets the idea of compact embedding, when the network of centers corresponds to the ε-net approximating the compact.

5. Multicorrectors in clustered universe: a case study

5.1. General setup and tasks at hand

5.1.1. Data sets

In what follows our use-cases will evolve around a standard problem of supervised multi-class classification. In order to be specific and to ensure reproducibility of our observations and results, we will work with a well-known and widely available CIFAR-10 dataset [79], [78]. The CIFAR-10 dataset is a collection of 32×32 colour images that are split across 10 classes:

'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', 'truck'

with 'airplane' being a label of Class 1, and 'truck' being a label of Class 10. The original CIFAR-10 dataset is further split into two subsets: a *training* set containing $5,000$ images per class (total number of images in the training set is $50,000$), and a *testing* set with $1,000$ images per class (total number of images in the testing set is $10,000$).

5.1.2. Tasks and approach

We focus on two fundamental tasks: for a given *legacy classifier*,

- (Task 1) devise an algorithm to *learn a new class* without catastrophic forgetting and re-training, and
- (Task 2) develop an algorithm to *predict* classification errors in the legacy classifier.

Let us now specify these tasks in more detail.

As a *legacy classifier* we have used a deep convolutional neural network whose structure is shown in Table 4. The network's training set comprised $45,000$ images corresponding to Class 1 - 9 ($5,000$ images per class), and the test set comprised $9,000$ images from the CIFAR-10 *testing* set ($1,000$ images per class). No data augmentation was invoked as a part of the training process. The network by stochastic gradient descent with the momentum parameter set to 0.9 and mini-batches of size 128. Overall, we trained the network over 70 epochs executed in 7 training episodes of 10-epoch training, and the learning rate was equal to $0.1/(1 + 0.001k)$, where k is the index of a training instance (a mini-batch) within a training episode.

The network's accuracy, expressed as the percentage of correct classifications, was 0.84 and 0.73 on the training and testing sets, respectively (rounded to the second decimal point). Each 10-epoch training episode took approximately 1.5 hours to complete on an HP Zbook 15 G3 laptop with a Core i7-6820HQ CPU, 16 Gb of RAM and Nvidia Quadro 1000M GPU.

Task 1 (learning a new class). Our first task was to equip the trained network with a capability to learn a new class without expensive retraining. In order to achieve this aim we adopted an approach

 doi:10.20944/preprints202106.0718.v1

Table 4. Architecture of the *legacy classifier*

Layer number	Type	Size
1	Input	$32 \times 32 \times 3$
2	Conv2d	$4 \times 4 \times 64$
3	ReLU	
4	Batch normalization	
5	Dropout 0.25	
6	Conv2d	$2 \times 2 \times 64$
7	ReLU	
8	Batch normalization	
9	Dropout 0.25	
10	Conv2d	$3 \times 3 \times 32$
11	ReLU	
12	Batch normalization	
13	Dropout 0.25	
14	Conv2d	$3 \times 3 \times 32$
15	ReLU	
16	Batch normalization	
17	Maxpool	pool size 2×2, stride 2×2
18	Dropout 0.25	
19	Fully connected	128
20	ReLU	
21	Dropout 0.25	
22	Fully connected	128
23	ReLU	
24	Dropout 0.25	
25	Fully connected	9
26	Softmax	9

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Table 5. Latent representation of an image

Attributes	$x_1, \ldots, x_9$	$x_{10}, \ldots, x_{137}$	$x_{138}, \ldots, x_{265}$	$x_{266}, \ldots, x_{393}$
Layers	26 (Softmax)	19 (Fully connected)	22 (Fully connected)	23 (ReLU)

and algorithms presented in [6], [76]. According to this approach, for every input image $\boldsymbol{u}$ we generated its latent representation $\boldsymbol{x}$ of which the composition is shown in Table 5.

Using these latent representations of images, we formed two sets: $\mathcal{X}$ and $\mathcal{Y}$. The set $\mathcal{X}$ contained latent representations of the new class (Class 10 - 'trucks') from the CIFAR-10 training set (5,000 images), and the set $\mathcal{Y}$ contained latent representations of all other images in CIFAR-10 training set (45,000 images). These set have then been used to construct a multi-corrector in accordance with the following algorithm presented in [76].

Algorithm 1. *(Few-shot AI corrector [76]: 1NN version. Training). Input: sets $\mathcal{X}$, $\mathcal{Y}$, the number of clusters, k, threshold, θ (or thresholds $\theta_1, \ldots, \theta_k$).*

1. *Determining the centroid $\bar{\boldsymbol{x}}$ of the $\mathcal{X}$. Generate two sets, $\mathcal{X}_c$, the centralized set $\mathcal{X}$, and $\mathcal{Y}^*$, the set obtained from $\mathcal{Y}$ by subtracting $\bar{\boldsymbol{x}}$ from each of its elements.*
2. *Construct Principal Components for the centralized set $\mathcal{X}_c$.*
3. *Using Kaiser, broken stick, conditioning rule, or otherwise, select $m \leq n$ Principal Components, $h_1, \ldots, h_m$, corresponding to the first largest eivenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0$ of the covariance matrix of the set $\mathcal{X}_c$, and project the centralized set $\mathcal{X}_c$ as well as $\mathcal{Y}^*$ onto these vectors. The operation returns sets $\mathcal{X}_r$ and $\mathcal{Y}_r^*$, respectively:*

$$\mathcal{X}_r = \{\boldsymbol{x} | \boldsymbol{x} = H\boldsymbol{z},\ \boldsymbol{z} \in \mathcal{X}_c\}$$

$$\mathcal{Y}_r^* = \{\boldsymbol{y} | \boldsymbol{y} = H\boldsymbol{z},\ \boldsymbol{z} \in \mathcal{Y}^*\},\ H = \begin{pmatrix} h_1^T \\ \vdots \\ h_m^T \end{pmatrix}.$$

4. *Construct matrix W*

$$W = \operatorname{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_m}}\right)$$

corresponding to the whitening transformation for the set $\mathcal{X}_r$. Apply the whitening transformation to sets $\mathcal{X}_r$ and $\mathcal{Y}_r^*$. This returns sets $\mathcal{X}_w$ and $\mathcal{Y}_w^*$:

$$\mathcal{X}_w = \{\boldsymbol{x} | \boldsymbol{x} = W\boldsymbol{z},\ \boldsymbol{z} \in \mathcal{X}_r\}$$
$$\mathcal{Y}_w^* = \{\boldsymbol{y} | \boldsymbol{y} = W\boldsymbol{z},\ \boldsymbol{z} \in \mathcal{Y}_r^*\}.$$

5. *Cluster the set $\mathcal{Y}_w^*$ into k clusters $\mathcal{Y}_{w,1}^*, \ldots, \mathcal{Y}_{w,k}^*$ (using e.g. the k-means algorithm or otherwise). Let $\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_k$ be their corresponding centroids.*
6. *For each pair $(\mathcal{X}_w, \mathcal{Y}_{w,i}^*)$, $i = 1, \ldots, k$, construct (normalized) Fisher discriminants $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$:*

$$\boldsymbol{w}_i = \frac{(\operatorname{Cov}(\mathcal{X}_w) + \operatorname{Cov}(\mathcal{Y}_{w,i}^*))^{-1}\bar{\boldsymbol{y}}_i}{\|(\operatorname{Cov}(\mathcal{X}_w) + \operatorname{Cov}(\mathcal{Y}_{w,i}^*))^{-1}\bar{\boldsymbol{y}}_i\|}.$$

An element $\boldsymbol{z}$ is associated with the set $\mathcal{Y}_{w,i}^*$ if $(\boldsymbol{w}_i, \boldsymbol{z}) > \theta$ and with the set $\mathcal{X}_w$ if $(\boldsymbol{w}_i, \boldsymbol{z}) \leq \theta$.

If multiple thresholds are given then an element $\boldsymbol{z}$ is associated with the set $\mathcal{Y}_{w,i}^*$ if $(\boldsymbol{w}_i, \boldsymbol{z}) > \theta_i$ and with the set $\mathcal{X}_w$ if $(\boldsymbol{w}_i, \boldsymbol{z}) \leq \theta_i$.

Output: vectors $\boldsymbol{w}_i$, $i = 1, \ldots, k$, matrices H and W.

Integration logic of the multi-corrector into the final system was as follows [76]:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Algorithm 2. *(Few-shot AI corrector [76]: 1NN version. Deployment). Input: a data vector* $\boldsymbol{x}$*, the set's* $\mathcal{X}$ *centroid vector* $\bar{\boldsymbol{x}}$*, matrices* H*,* W*, the number of clusters,* k*, cluster centroids* $\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_k$*, threshold,* θ *(or thresholds* $\theta_1, \ldots, \theta_k$*), discriminant vectors,* $\boldsymbol{w}_i$*,* $i = 1, \ldots, k$*.*

1. *Compute*
$$\boldsymbol{x}_w = WH(\boldsymbol{x} - \bar{\boldsymbol{x}})$$
2. *Determine*
$$\ell = \arg\min_i \|\boldsymbol{x}_w - \bar{\boldsymbol{y}}_i\|.$$
3. *Associate the vector* $\boldsymbol{x}$ *with the set* $\mathcal{Y}$ *if* $(\boldsymbol{w}_\ell, \boldsymbol{x}_w) > \theta$ *and with the set* $\mathcal{X}$ *otherwise.*

   If multiple thresholds are given then associate the vector $\boldsymbol{x}$ *with the set* $\mathcal{Y}$ *if* $(\boldsymbol{w}_\ell, \boldsymbol{x}_w) > \theta_\ell$ *and with the set* $\mathcal{X}$ *otherwise.*

Output: a label attributed to the vector $\boldsymbol{x}$*.*

Remark 1. Note that, since the set $\mathcal{Y}$ corresponds to data samples from previously learned classes, a positive response in the multi-corrector (condition $(\boldsymbol{w}_\ell, \boldsymbol{x}_w) > \theta$ holds) "flags" that this data point is to be associated with classes that have already been learned (Classes 1 – 9). Absence of a positive response indicates that the data point is to be associated with the new class (Class 10).

Task 2 (predicting errors of a trained legacy classifier). In addition to learning a new class without retraining, we considered the problem of predicting correct performance of a trained legacy classifier. In this setting, the set $\mathcal{X}$ of vectors corresponding to *incorrect* classifications on CIFAR-10 *training* set, and the set $\mathcal{Y}$ contained latent representations of images form CIFAR-10 training set that have been correctly classified. Similar to the previous task, predictor of the classifier's error was constructed in accordance with Algorithms 1, 2.

Testing protocols. Performance of the algorithms was assessed on CIFAR-10 *testing* set. For Task 1, we tested how well our new system - the legacy network shown in Table 4 combined with the multi-corrector constructed by Algorithms 1, 2 - performs on images from CIFAR-10 *testing* set. For Task 2, we assessed how well the multi-corrector, trained on CIFAR-10 *training* set, predicts errors of the legacy network for images of 9 classes (Class 1 - 9) taken from CIFAR-10 *testing* set.

5.2. Results

Task 1 (learning a new class). Performance of the multi-corrector in the task of learning a new class is illustrated in Figure 7. In these experiments, we projected onto the first 20 principal components. The rationale for choosing these 20 principal components was that for these components the ratio of the largest eigenvalue to the eigenvalue that is associated with the principal component is always smaller than 10. The figure shows ROC curves in which True positives are images from the new class and identified as a new class, and False positives are defined as images from already learned classes (Classes 1 - 9) but identified as a new class (Class 10) by the combined system. As we can see from Figure 7, performance of the system saturates at about 10 clusters which indicates a peculiar granular structure of the data universe in this example: clusters are apparently not equal in terms of their impact on the overall performance, and the benefit of using more clusters decays rapidly as the number of clusters grows.

We note that the system performance and generalisation depends on both, ambient dimension (the number of principal components used) and the number of clusters. This phenomenon is illustrated in in Figure 8. When the number of dimensions increases (top row in Figure 8) the gap between a single-cluster corrector and a multi-cluster corrector narrows. Yet, as can be observed from this experiment, the system generalises well.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

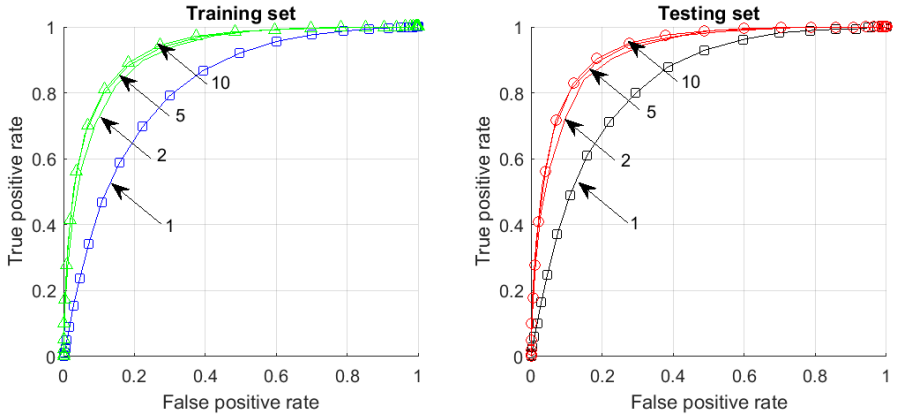

Figure 7. Clustered universe in learning a new class. Arrows and numbers show the number of clusters in the multi-corrector for which that specific ROC curve was constructed.

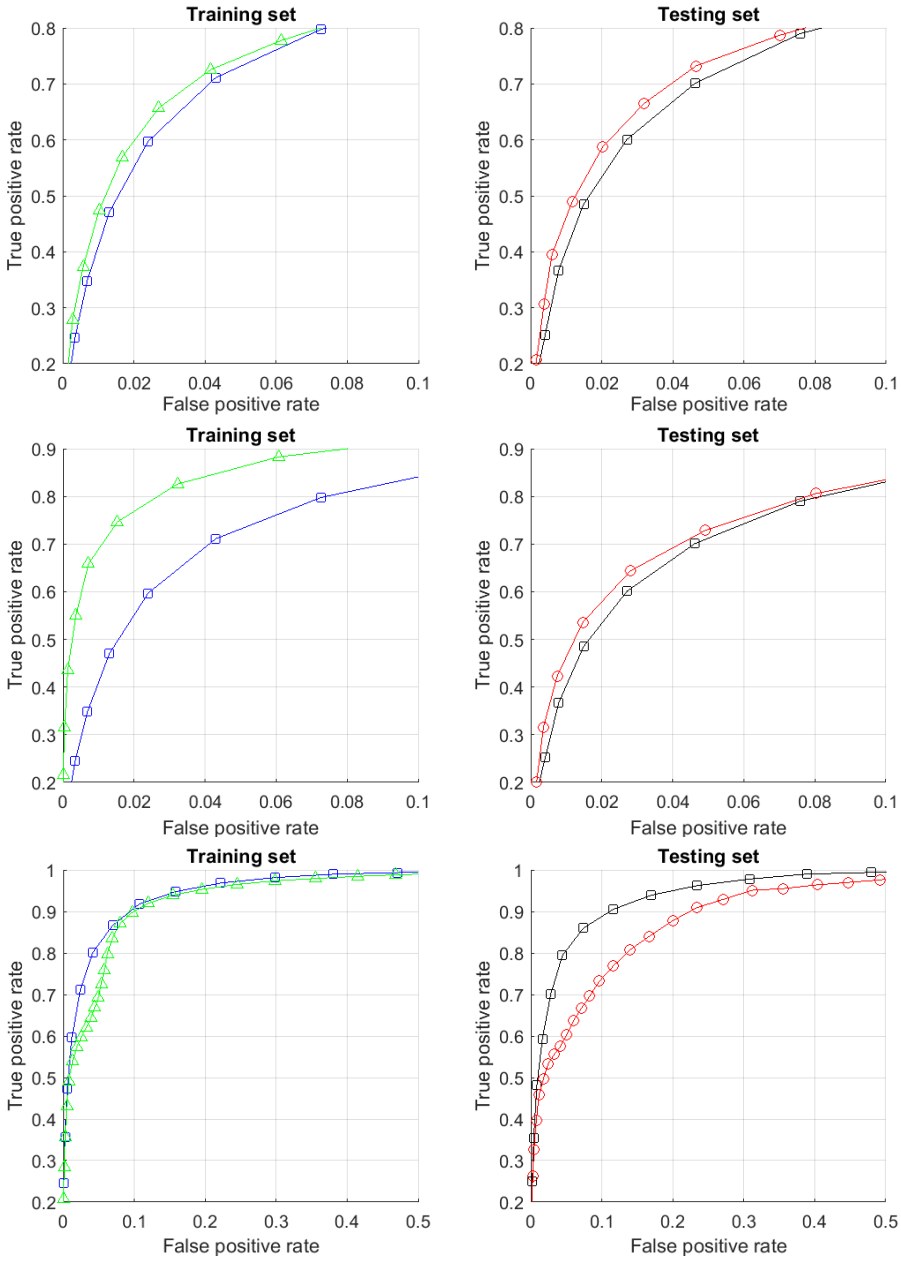

Figure 8. Clustered universe in learning a new class - the impact of dimension of the ambient space. Curves marked with squares correspond to corrector with a single cluster, curves marked by triangles and circles correspond to correctors with multiple clusters. Top panel: the application of Algorithm 1 and 2 to the same data but with retained first 100 principal components instead of the first 20 components (see Figure 7). Middle panel: projecting onto the first 100 principal components and using 300 clusters. Bottom panel: projecting onto the first 300 principal components and using 300 clusters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

When the number of clusters increases from 10 to 300 the system overfits. This is not surprising as given the size of our training set (50,000 images to learn from) splitting the data into 300 clusters implies that each 100-dimensional discriminant in Algorithm 1 is constructed, on average, from mere 170 samples. The lack of data to learn from and "diffusion" and shattering of clusters in high dimension could be contributor to the instability. Nevertheless, as the right plot shows, the system still generalises at the level that is similar to the 10-cluster scenario.

When the ambient dimension increases further we observe a dramatic performance collapse for the multi-corrector constructed by Algorithms 1, 2. Now 300-dimensional vectors are built from, on average 170 points. The procedure is inherently unstable and in this sense such results are expected in this limit.

Task 2 (predicting errors). A very similar picture occurs in the task of predicting errors of legacy classifiers. For our specific case, performance of 10-cluster multi-corrector with projection onto 20 principal components in shown in Figure 9. In this task, True positives are *errors* of the original classifier which have been correctly identified as errors by the corrector. False positives are data correctly classified by the original deep neural network but which nevertheless have been labeled as errors by the corrector. According to Figure 9, the multi-corrector model generalises well and delivers circa 70% specificity and sensitivity on the test set.

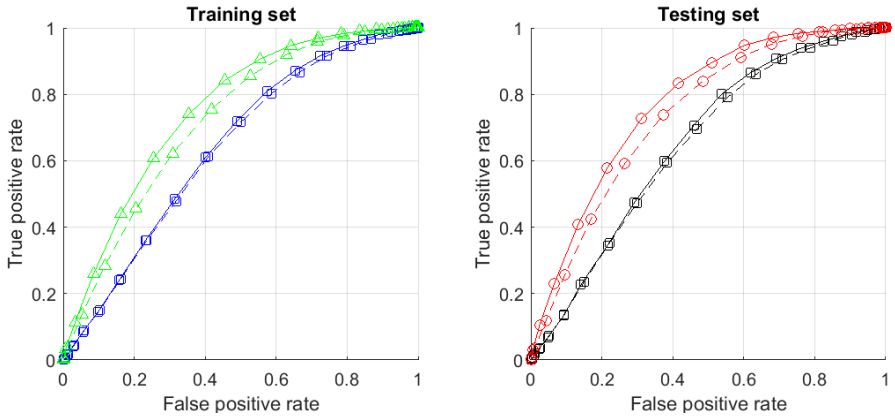

Figure 9. Prediction of errors. Curves marked by triangles and circles correspond to 10-cluster multi-corrector. Curves marked by squares are produced by a single-cluster classifier. Dashed lines show performance of the same system but constructed on data sets in the reduced feature space formed by attributes $1-137$ (see Table 5).

Another interesting phenomenon illustrated by Fig. 9 is the apparent importance of how the information from the legacy AI model is aggregated into correcting cascades. Dashed lines in Fig. 9 show what happens if latent representations are formed by signals taken from layers 26 and 19 only. In this case the impact of clustering becomes less pronounced suggesting the importance of feature selection for optimal performance.

Computational efficiency. Computational costs of constructing multi-correctors is remarkably small. For example, learning a new class with a 10-cluster multi-corrector and 20 principal components took 1.32 seconds on the same hardware used to train the original legacy classifier. When the number of clusters and dimension increases to 300 and 300, respectively, the amount of time needed to construct the multi-corrector was 37.7 seconds. These figures show that not only clustered universes and multi-correctors are feasible in applications but also they are extremely efficient computationally. We do not wish to suggest that they are a replacement of deeper retraining. Yet, as we see from these experiments, they can be particularly efficient in the tasks of incremental learning - learning an additional class in a multi-class problem - if implemented appropriately.

 doi:10.20944/preprints202106.0718.v1

6. Discussion

> "If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic *hypothesis* (or the atomic *fact*, or whatever you wish to call it) that *all things are made of atoms—little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another*. In that one sentence, you will see, there is an *enormous* amount of information about the world, if just a little imagination and thinking are applied." Richard Feynman [86].

Observing the successes and failures of data driven AI we should go further than Richard Feynman. Perhaps the most influential and non-trivial ontological *hypothesis* (or *fact*) is that the *world consists of things*. This miracle of things makes our language and cognition possible. Even in description of flows of substance (continua) we are most successful when we identify discrete things: vortexes and various structures. Philosophers can discuss if things exist or are created by us in the course of cognition. But the practical point of view is undoubted: the hypothesis of things is very useful and there is an *enormous* amount of information in it about the world. (In some sense, the atomic hypothesis can be read as an extension of the hypothesis of things to the microscale.)

We can go even further and apply this hypothesis of things to features of objects and to relations between them, like L. Boltzmann advised to M. Planck to apply the atomic hypothesis to emission of electromagnetic radiation (and we remember the great success that followed this advice). We will find that features and relations are also combined into distinct discrete entities. The main purpose of big data preprocessing is to uncover all of these entities hidden in the data. After that, various pattern recognition tasks become viable.

The preprocessing in the postclassical data world (Sec. 2, Fig. 3) is a challenging task because no classical statistical methods are applicable when the sample size is much less than data dimensionality (the Donoho area (Sec. 2.1, 1) [9]). The correlation transformation (Sec. 2.2) moves data out of the Donoho area, but the specific non-classical effects persist while the sample size remains much less than the exponential of the data dimensionality (2). Dimensionality reduction methods should combine two set of goals: sensible grouping and extraction of relevant features. For these purposes, combination of supervised and unsupervised learning techniques is necessary. Data labels from supervised approaches add sense to analysis of unlabeled data. The simple geometric methods like supervised PCA, semisupervised PCA (Sec. 2.3), and Domain Adaptation PCA (DAPCA) (Sec. 2.3) serve as prototypes of more complex and less controllable approaches. They can also be used to simplify large deep learning systems [51].

Data in postlclassical world are rarefied. At the same time, values of regular functionals on data are concentrated near their median values [26,29]. Combinations of these properties produce 'blessing of dimensionality' [8,9,65]. The most important manifestation of these effects for applied data analysis beyond the central limit theorem are quasiorthogonality [33–35] and stochastic separation theorems [12,15]. These results give the theoretical backgrounds for construction of new type of intellectual devices, correctors of AI systems. In this paper, we presented a new family of stochastic separation theorems for fine-grained data distributions with different geometry of clusters (Sec. 3). These results allowed us to develop the multi-correctors of multidimensional AI with a granular distribution of errors. On real data, such correctors showed better performance than simple correctors.

Various versions of multi-correctors that provide fast and reversible correction of AI errors should be supplemented by the special operation of interiorization of corrections. Accumulation of many corrections will. step by step, spend the blessing of dimensionality resource: after implementing elementary corrections, the probability of success of new correctors decreases, which can be considered as accumulation of technical debt. In psychology, interiorization is the process of making skills, attitudes, thoughts, and knowledge an integrated part of one's own being. For large legacy AI systems,

 doi:10.20944/preprints202106.0718.v1

interiorization of corrections means the supervising re-training of the system, where the complex "legacy system+multi-corrector" acts as a supervisor and labels the data, while the system itself learns by assimilating the fast flow of generated data.

The infinite-dimensional version of theorems about separation of compact clusters and families of such clusters demonstrates the importance of hypothesis about compact embedding of data clusters (Sec. 3.5). The idealized concept of granular Hierarchical Universe (Fig. 5) is intended to replace the ideal picture of a smooth unimodal distribution popular in statistical science.

In the future, the concept of granular Hierarchical Universe will be extended to feature space and to relationships between different entities (in data space, feature space, etc.), which is natural in the deep learning framework.

Author Contributions: conceptualization and methodology, ANG and IYT; writing–original draft preparation, ANG, IYT and BG; writing–review and editing, all authors; software and validation, IYT and EMM.

Funding: IYT was funded by UKRI (Alan Turing AI Acceleration Fellowship EP/V025295/1). ANG, EMM, and IYT were founded by the Ministry of Science and Higher Education of the Russian Federation (Project No. 075-15-2020-808).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
PCA	Principal Component Analysis
TCA	Transfer Component Analysis
DAPCA	Domain Adaptation PCA

Appendix A "Almost Always" in infinite-dimensional spaces

As it was mentioned in Sec. 3.5, in the infinite-dimensional limit many statements about high or low probabilities transform into 0-1 laws: something happens almost always or almost newer. Such limits for concentrations on spheres and their equators were discussed by Lévy [65] as an important part of the measure concentration effects. In physics, this limit corresponds to the so-called thermodynamic limit of statistical mechanics [66,67]. The original Kolmogorov 0-1 law states, roughly speaking, that an event that depends on an infinite collection of independent random variables but is independent of any finite subset of these variables, has probability zero or one (for precise formulation we refer to the monograph [68]). The infinite-dimensional 0-1 asymptotic might bring more light and be more transparent than the probabilistic formulas.

This may be surprising, but the problem is what "almost always" means. Formally, various definitions of genericity are constructed as follows. All systems (or cases, or situations and so on) under consideration are somehow parameterized – by sets of vectors, functions, matrices etc. Thus, the "space of systems" Q can be described. Then the *"meagre (or thin) sets"* are introduced into Q, i.e. the sets, which we shall later neglect. The union of a finite or countable number of meager sets, as well as the intersection of any number of them should be meager set again, while the whole Q is not thin. There are two traditional ways to determine thinness.

1. The sets of *measure zero* are negligible.
2. The sets of *Baire first category* are negligible.

The first definition requires existence of a special measure such that all relevant distributions are expected to be absolute continuous with respect to it. In Theorem 1, for example, we assumed that the probability distribution (yet unknown) has density and is absolutely continuous with respect

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

Lebesgue measure. Moreover, we used a version of the "Smeared (or Smoothed) Absolute Continuity" (SmAC) condition (11) [13,32], which means that the sets of relatively small volume cannot have high probability, whereas absolute continuity means that sets of zero volume have probability zero. Unfortunately, in the infinite-dimensional spaces we usually do not have such a sensible measure. It is very easy to understand if we look on the volumes of balls in Hilbert space with orthonormal basis $\{e_i\}$. If the measure of a ball is function of its radius and the measure of a ball of radius R is finite, then the balls of radius $R/4$ have zero measure (because infinitely many such balls with the centers at points $Re_i/2$ can be packed in the ball of radius $R/4$), and, therefore, the ball of radius R has zero measure because it can be covered by a countable set of balls of radius $R/4$. Hence, all balls have either zero or infinite measure.

The second definition is widely accepted when we deal with the functional parameters. The construction begins with nowhere dense sets. The set Y is nowhere dense in Q, if in any nonempty open set $V \subset Q$ (for example, in a ball) there exists a nonempty open subset $W \subset V$ (for example, a ball), which does not intersect with Y: $W \cap Y = \emptyset$. Roughly speaking, Y is "full of holes" – in any neighborhood of any point of the set Y there is an open hole. Countable union of nowhere dense sets is called the set of first category. The second usual way is to define thin sets as the *sets of first category*. A residual set (a "thick" set) is the complement of a set of the first category. If a set is not meagre it is said to be of the second category. The Baire classification is non-trivial in the so-called Baire spaces, where every intersection of a countable collection of open dense sets is also dense. Complete metric spaces and, in particular, Banach spaces are Baire spaces. Therefore, for Banach spaces of functions, the common definition of negligible set is "set of first Baire category". Such famous results as transversality theorem in differential topology [80] or Pugh closing lemma [81] and Kupka-Smale theorem [82] in differential dynamics.

Despite of these great successes, it is also widely recognized that the Bair category approach to generic properties requires at least great care. Here are some examples of correct but useless statements about "generic" properties of function: Almost every continuous function is not differentiable; Almost every C^1-function is not convex. Their meaning for applications is most probably this: the genericity used above for continuous functions or for C^1-function is irrelevant to the subject.

Contradictions between the measure-based and category-based definitions of negligible sets are well-known even in dimension one: even the real line R can be divided into two sets, one of which has zero measure, the other is of first category [83]. Genericity in the sense of measure and genericity in the sense of category differ significantly in the applications where both concepts can be used.

The conflict between the two main views on genericity and negligibility stimulated efforts to invent new and stronger approaches. The formal requirements to new definitions are:

- A union of countable family of thin sets should be thin.
- Any subset of a thin set should be thin.
- The whole space is not thin.

Of course, if we take care not to throw the baby out with the bath water then in $\mathbb{R}^n$, where both classical definition are applicable, we expect that thin sets should be of first category and have zero measure, both. It was not clear a priori whether such a theory is possible with proof nontrivial and important generic properties. It turned out that it is possible. To substantiate the effectiveness of evolutionary optimization, a theory of completely negligible sets in Banach spaces was developed. [84,85].

Let Q be a real Banach space. Consider compact subsets in Q parametrized by points of a compact space K. It can be presented as a Banach space $C(K, Q)$ of continuous maps $K \to Q$ in the maximum norm.

Definition A1. *A set $Y \subset Q$ is completely thin, if for any compact space K the set of continuous maps $\Psi : K \to Q$ with non-empty intersection $\Psi(K) \cap Y \neq \emptyset$ is set of first Bair category.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

The union of a finite or countable number of completely thin sets is completely thin. Any subset of a completely thin point is completely thin, while the whole Q is not. A set Y in the Banach space Q is completely thin, if for any compact set K in Q and arbitrary positive $\varepsilon > 0$ there exists a vector $q \in Q$, such that $\|q\| < \varepsilon$ and $K + q$ does not intersect Y: $(K + q) \cap Y = \emptyset$. All compact sets in infinite-dimensional Banach spaces and closed linear subspaces with infinite codimension are completely thin.

Only empty set is completely thin in a finite-dimensional space $\mathbb{R}^n$.

Examples below demonstrate that almost all continuous functions have very natural properties: the set of zeros is nowhere dense, and the (global) maximizer is unique. Below the wording "almost always" means: the set of exclusions is completely thin.

Proposition A1 ([84,85]). *Let X have no isolated points. Then*

- *Almost always a function $f \in C(X)$ has nowhere dense set of zeros $\{x \in X \,|\, f(x) = 0\}$ (the set of exclusions is completely thin in $C(X)$).*
- *Almost always a function $f \in C(X)$ has only one point of global maximum.*

The following proposition is a tool for proof that some typical properties of functions hold almost always for all functions from a generic compact set.

Proposition A2 ([84,85]). *If a set Y in the Banach space Q is completely thin, then for any compact metric space K the set of continuous maps $\Psi : K \to Q$ with non-empty intersection $\Psi(K) \cap Y \neq \emptyset$ is completely thin in the Banach space $C(K, Q)$.*

Proposition A3 ([84,85]). *Let X have no isolated points. Then for any compact space K and almost every continuous map $\Psi : K \to C(X)$ all functions $f \in \Psi(K)$ have nowhere dense sets of zeros (the set of exclusions is completely thin in $C(K, C(X))$).*

In other words, in almost every compact family of continuous functions all the functions have nowhere dense sets of zeros.

Qualitatively, the concept of a completely thin set was introduced as a tool for identifying typical properties of infinite-dimensional objects, the violation of which is unlikely ('improbable') in any reasonable sense.

Appendix B Flowchart of multi-corrector operation

In Sec. 4, we introduced multi-corrector of AI systems. The basic scheme of this device is presented in Fig. 6. It includes several elementary correctors (see Fig. 2) and a dispatcher. A cluster of errors is owned by each elementary corrector. An elementary corrector evaluates the risk of errors from its own cluster for an arbitrary operation situation and takes the decision: to correct or not to correct the legacy AI decision for this situation. For any situation, the dispatcher selects the most appropriate elementary corrector to make a decision about correction. To find a suitable corrector, it uses a cluster error model. When new errors are found, the cluster model changes. More detailed presentation of multi-corrector operation is given by the following flowcharts.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

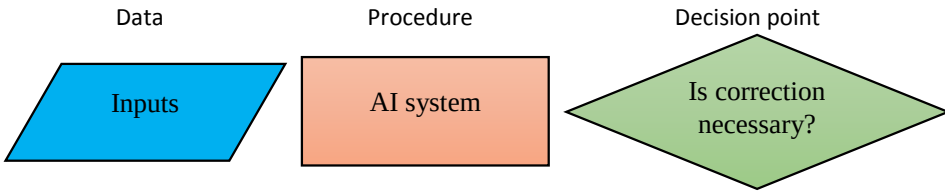

Figure A1. *Notations used in the flowcharts.* All flowcharts use a unified set of blocks: blocks in the form of parallelograms display data, rectangular blocks display procedures, and blocks in the form of rhombuses display the branching points of processes (algorithms) or decision points. The arrows reflect the transfer of data and control.

Flowcharts and blocks are numbered. The flowchart number is mentioned at the top of the drawings. If a block is present in different flowcharts, then it carries the number assigned to it in the top-level flowchart. The relations between different flowcharts are presented in Fig. A2.

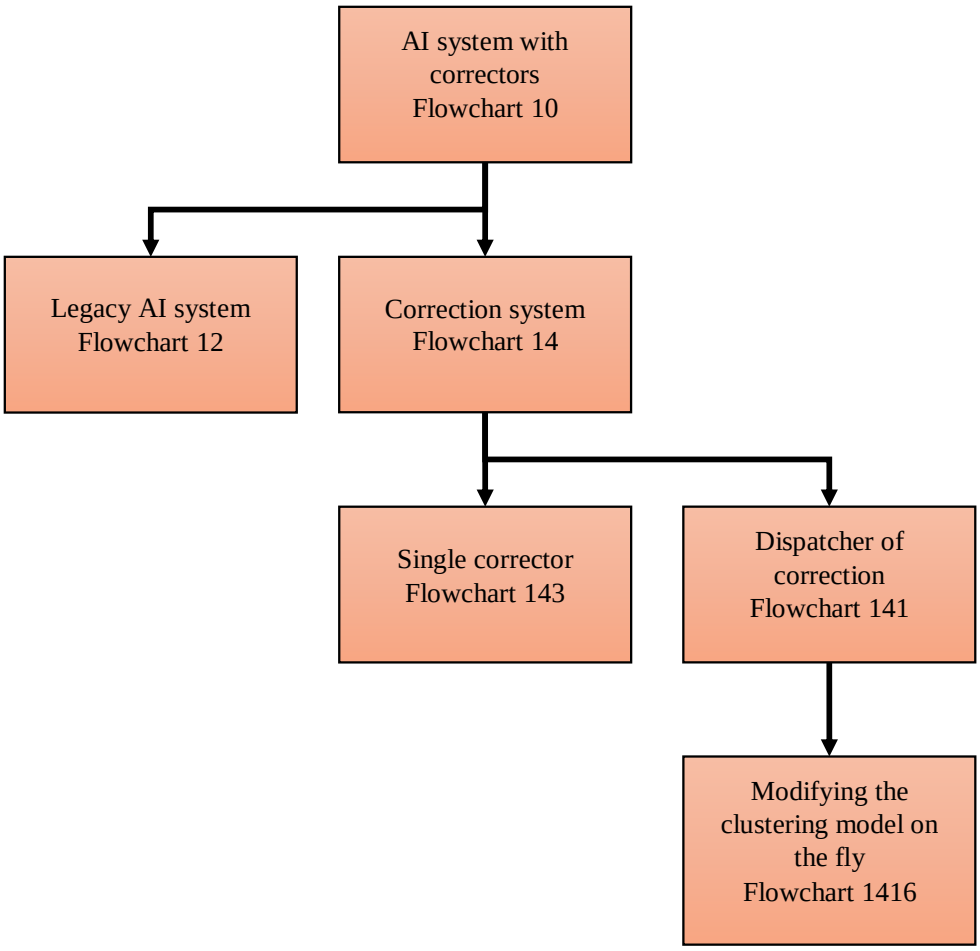

Figure A2. *The tree of flowcharts*: 10 – Operation of the modified AI system (Fig. A3); 12 – Operation of the legacy AI system (Fig. A4), 14 – Operation of the correction system (Fig. A5); 143 – Single corrector operation (Fig. A6); 141 – The work of the dispatcher (Fig. A7); 1415 – Online modification of the cluster model (Fig. A8);

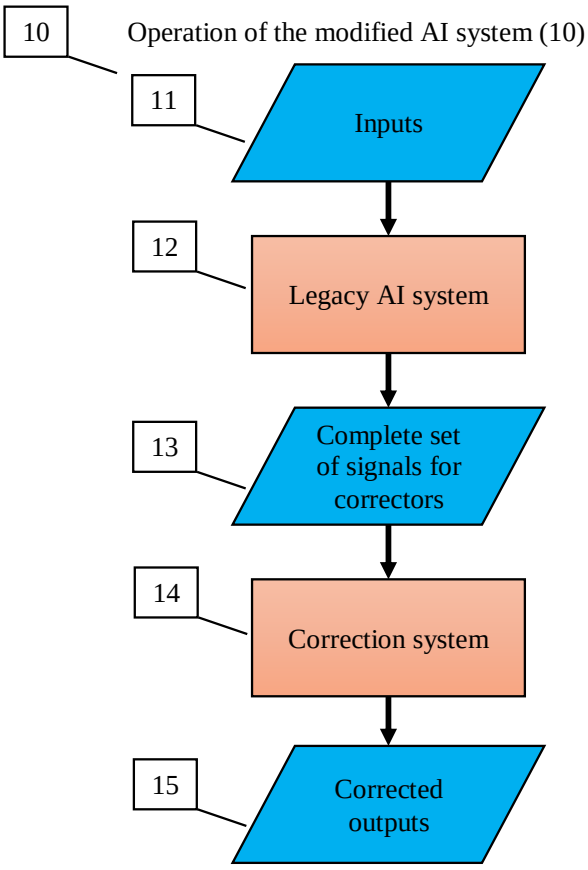

Figure A3. *Operation of the modified AI system (10).* Input signals (11) are fed to the input of the AI system (12), which at the output gives out the complete vector of the signal (13) that can be used for correction. The complete signal vector (13) is fed to the input of the correction system (14). The correction system (14) calculates the correction of the output signals (15).

 doi:10.20944/preprints202106.0718.v1

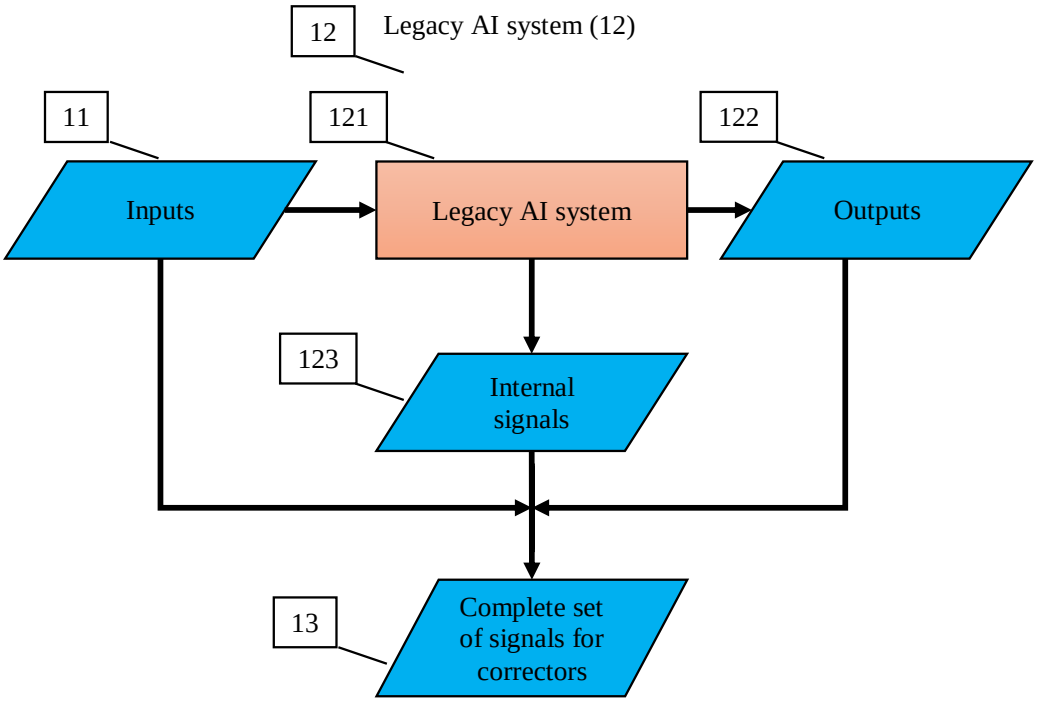

Figure A4. *Operation of the legacy AI system (12).* Input signals (11) are fed to the input of the AI system. The AI system generates vectors of internal signals (123) and output signals (122). Input signals (11), internal signals (123), and output signals (122) form the complete signal vector (13).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

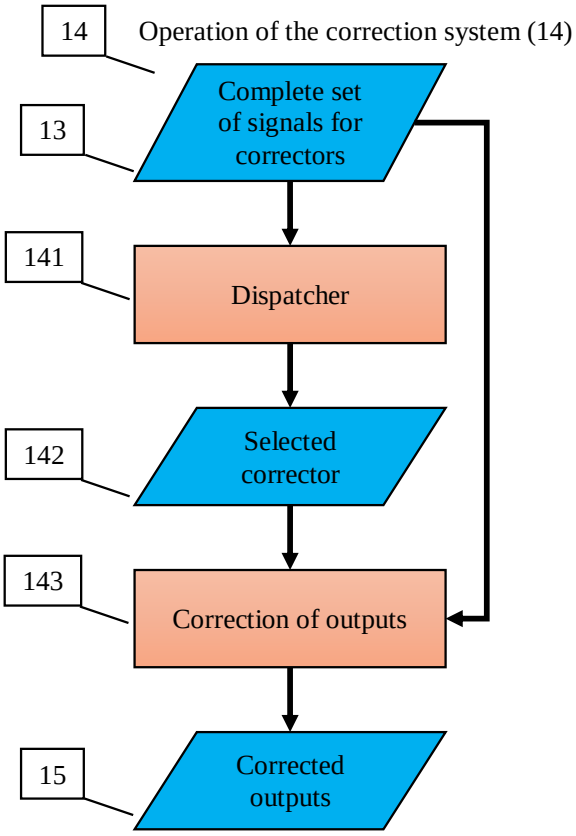

Figure A5. *Operation of the correction system (14).* The complete vector of signals (13) is fed to the dispatcher input (141). The dispatcher (141) selects from the correctors the one that most closely matches the situation (142). The selected corrector (142) and the complete signal vector (13) are used to correct the signals (13). The computed corrected outputs (15) are returned.

 doi:10.20944/preprints202106.0718.v1

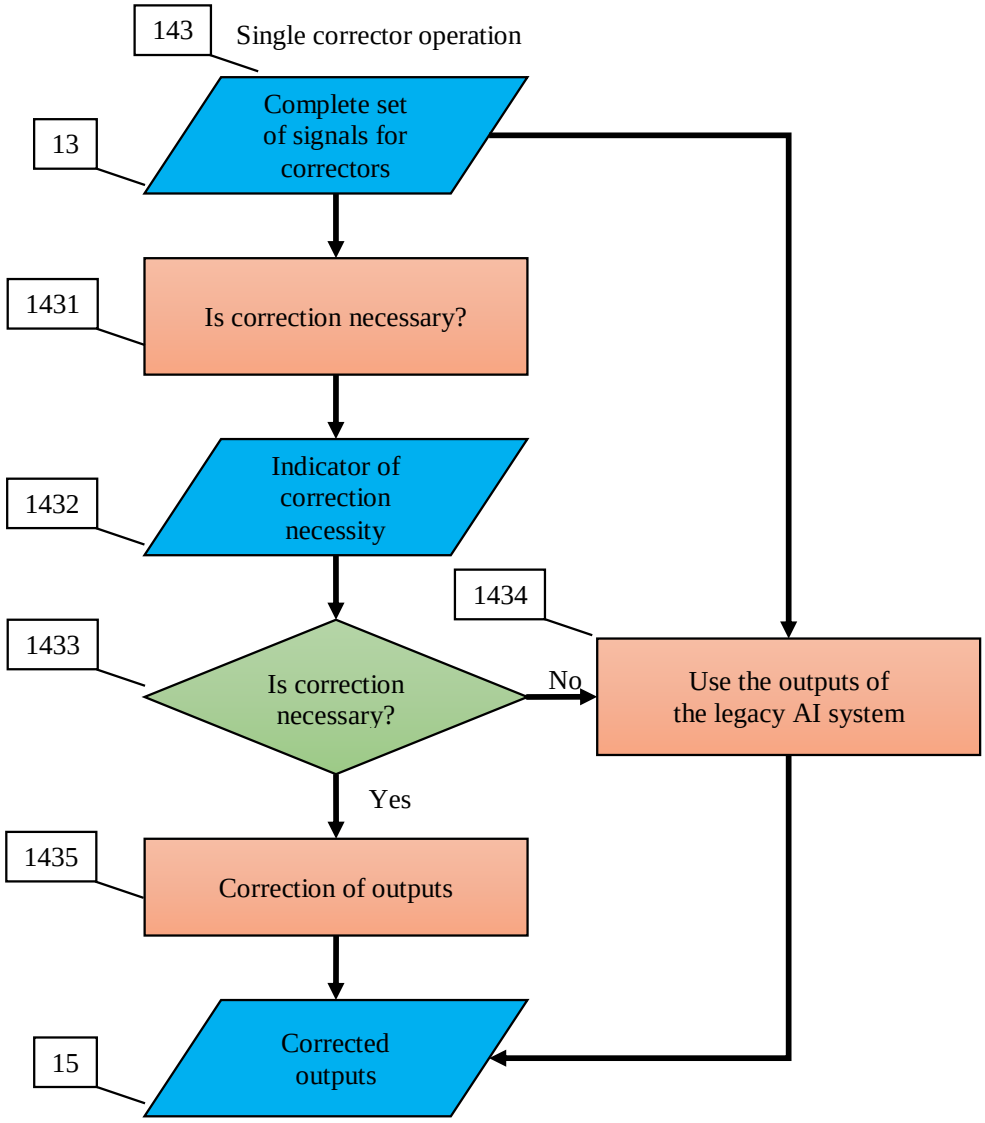

Figure A6. *Single corrector operation (143).* The complete vector of signals (13) is used to decide whether a correction is needed (1431). If it is necessary, then correction (1435) is performed, and the resulting vector of output signals (15) is sent to the output. If there is no need for correction, then the vector of output signals is extracted (1434) from the complete vector of signals (13), and the resulting vector of output signals (15) is transmitted to the output.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2021 doi:10.20944/preprints202106.0718.v1

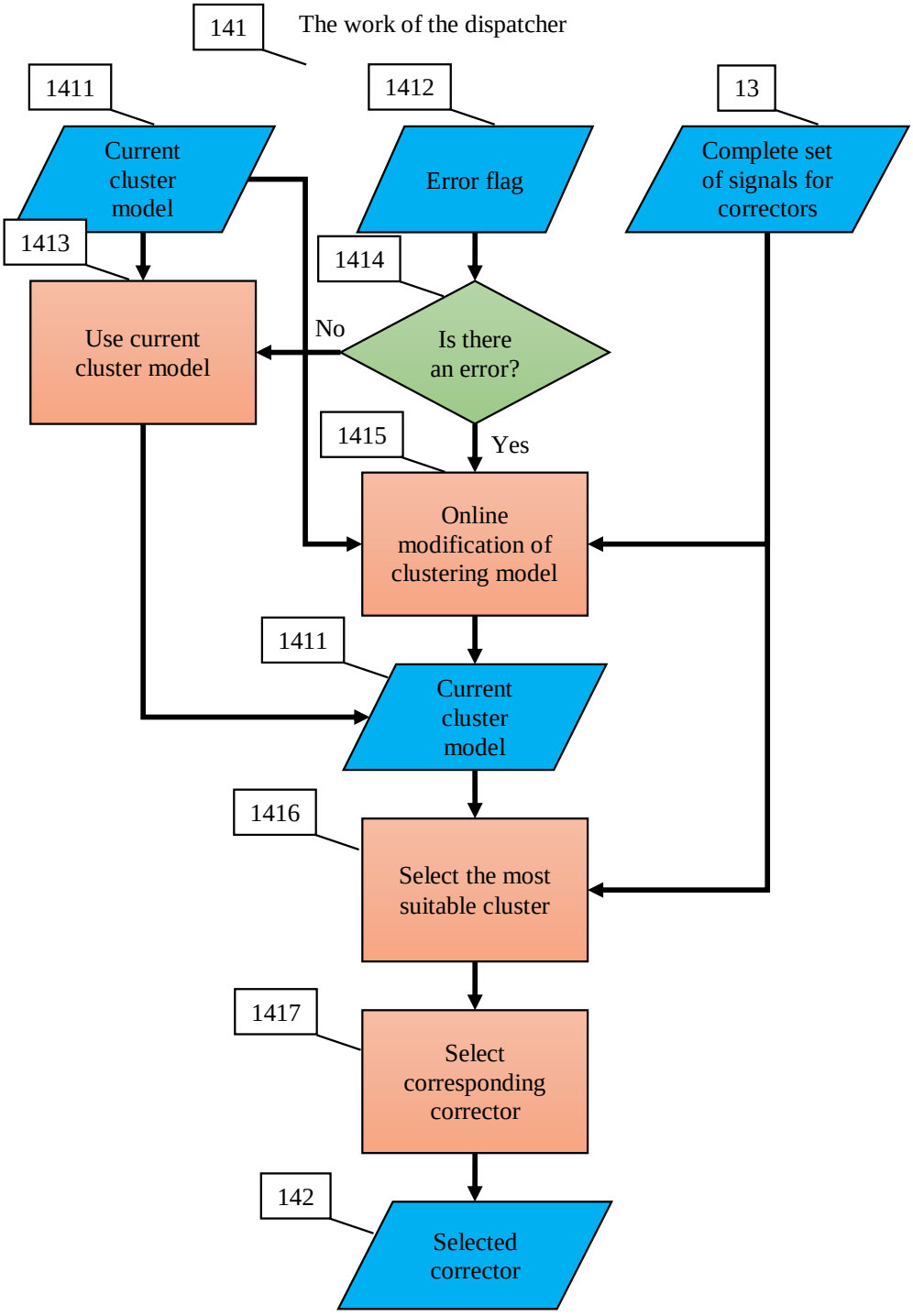

Figure A7. *The work of the dispatcher (141).* If the error flag (1412) is detected (1414), then the current cluster model (1411) and the complete signal vector (13) are used to modify the cluster model (1415) online. The modified cluster model becomes the current one (1411). If the error flag (1412) is not detected (1414), then the current cluster model (1411) is selected (1413) for use (1411). Based on the cluster model (1411) and the complete signal vector (13), the most suitable cluster (1416) is selected. Then, the corrector (142) corresponding to this cluster is selected (1417).

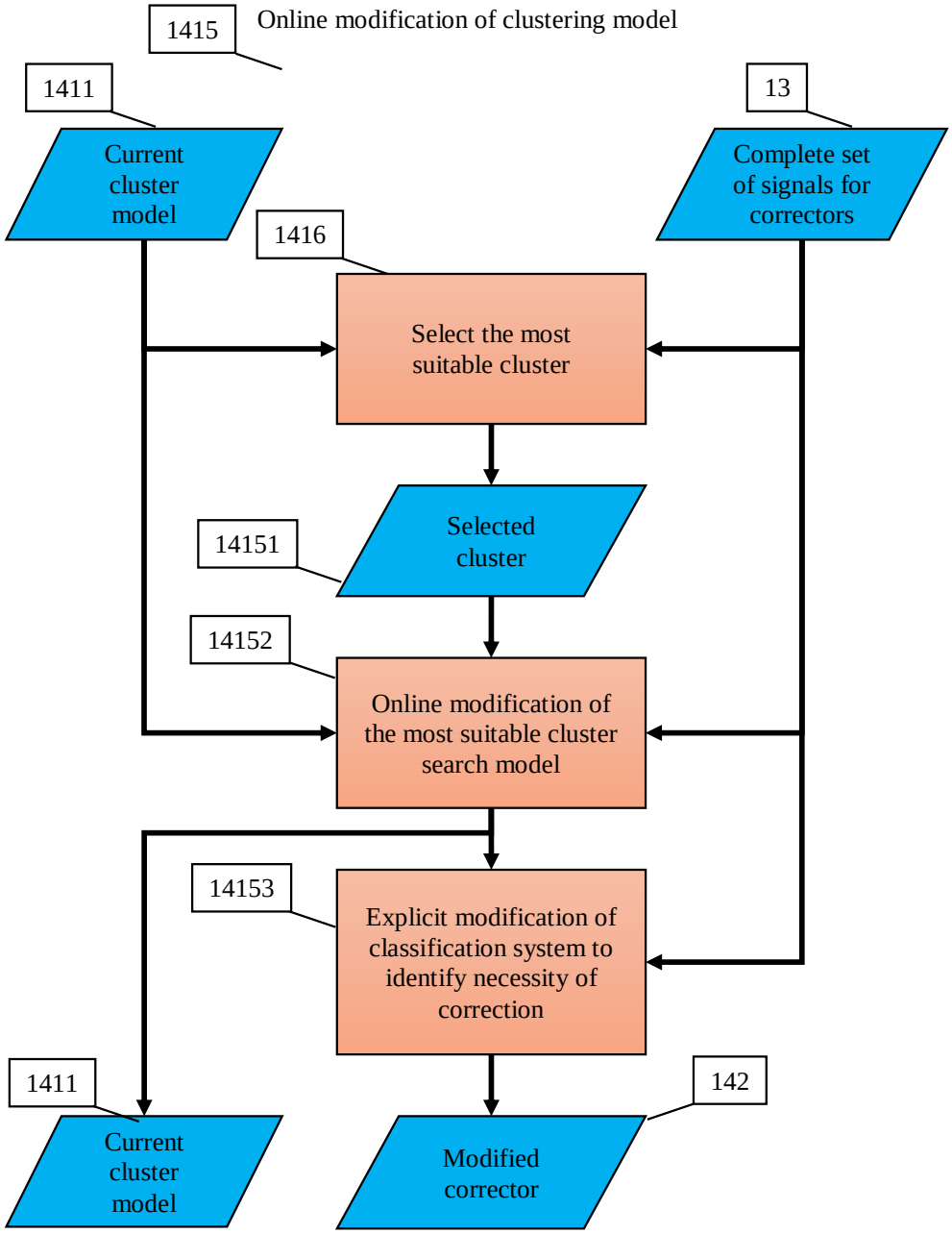

Figure A8. *Online modification of the cluster model (1415).* Based on the current cluster model (1411) and the complete signal vector (13), the most suitable cluster (14151) is selected (1416). Online modification of the rule for determining the most suitable cluster (14152) is performed. After setting up the new cluster model (1411), the classifier for this corrector to make a decision about the need for correction is explicitly modified (14153). The modified corrector (142) together with the new cluster model (1411) forms an updated version of the correction system (14).

References

1. Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching networks for one shot learning. NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, December 2016, Pages 3637–3645. https://dl.acm.org/doi/abs/10.5555/3157382.3157504
2. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. Proceedings of the International Conference on Learning Representations, 2017. https://openreview.net/pdf?id=rJY0-Kcll

3. Wang, Y.; Yao, Q.; Kwok, J. T.; Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* **2020**, *53*(3), 1–34. https://doi.org/10.1145/3386252
4. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 4080–4090. https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf
5. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 1199–1208). https://openaccess.thecvf.com/content_cvpr_2018/html/Sung_Learning_to_Compare_CVPR_2018_paper.html
6. Gorban, A.N.; Tyukin, I.Y. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Phil. Trans. R. Soc. A* **2018**, *376*, 20170237, https://doi.org/10.1098/rsta.2017.0237.
7. Tyukin, I.Y.; Gorban, A.N.; Alkhudaydi, M.H.; Zhou, Q. Demystification of few-shot and one-shot learning. arXiv preprint arXiv:2104.12174 **2021**. https://arxiv.org/abs/2104.12174
8. Kainen, P.C. Utilizing geometric anomalies of high dimension: when complexity makes computation easier. In *Computer-Intensive Methods in Control and Signal Processing: The Curse of Dimensionality*; Warwick, K., Kárný, M., Eds.; Springer: New York, NY, USA, 1997; 283–294, https://doi.org/10.1007/978-1-4612-1996-5_18.
9. Donoho, D.L. *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*; *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.
10. Anderson, J.; Belkin, M.; Goyal, N.; Rademacher, L.; Voss, J. The More, the Merrier: the Blessing of Dimensionality for Learning Large Gaussian Mixtures. In Proceedings of The 27th Conference on Learning Theory, Barcelona, Spain, 13–15 June 2014; Balcan, M.F.; Feldman, V.; Szepesvári, C., Eds.; PMLR: Barcelona, Spain, 2014; Volume 35, 1135–1164. http://proceedings.mlr.press/v35/anderson14.pdf
11. Gorban, A.N.; Tyukin, I.Y.; Romanenko, I. The blessing of dimensionality: Separation theorems in the thermodynamic limit. *IFAC-PapersOnLine* **2016**, *49*, 64–69, https://doi.org/10.1016/j.ifacol.2016.10.755.
12. Gorban, A.N.; Tyukin, I.Y. Stochastic separation theorems. *Neural Networks* **2017**, *94*, 255–259. https://doi.org/10.1016/j.neunet.2017.07.014
13. Gorban, A.N.; Golubkov, A.; Grechuk, B.; Mirkes, E.M.; Tyukin I.Y. Correction of AI systems by linear discriminants: Probabilistic foundations. *Information Sciences* **2018**, *466*, 303–322. https://doi.org/10.1016/j.ins.2018.07.040
14. Gorban, A.N.; Makarov, V.A.; Tyukin, I.Y. The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Phys. Life Rev.* **2019**, *29*, 55–88, https://doi.org/10.1016/j.plrev.2018.09.005.
15. Grechuk, B.; Gorban, A.N.; Tyukin, I.Y. General stochastic separation theorems with optimal bounds. *Neural Networks* **2021**, *138*, 33–56. https://doi.org/10.1016/j.neunet.2021.01.034
16. Flury B. Principal points. *Biometrika* **1990**, *77*, 33–41. https://doi.org/10.1093/biomet/77.1.33
17. Gorban, A.N.; Zinovyev, A.Y. Principal graphs and manifolds. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, Olivas, E.S.; Guererro, J.D.M.; Sober, M.M.; Benedito, J.R.M.; Lopes, A.J.S., Eds. 2010, 28–59. IGI Global. https://doi.org/10.4018/978-1-60566-766-9.ch002
18. Tyukin, I.Y.; Gorban, A.N.; Grechuk, B.; Green, S. Kernel Stochastic Separation Theorems and Separability Characterizations of Kernel Classifiers. In 2019 International Joint Conference on Neural Networks (IJCNN) 2019, 1–6, IEEE. https://doi.org/10.1109/IJCNN.2019.8852278
19. Kreinovich, V.; Kosheleva, O. Limit Theorems as Blessing of Dimensionality: Neural-Oriented Overview. *Entropy* **2021**, *23*(5), 501. https://doi.org/10.3390/e23050501
20. Gorban, A.N.; Kégl, B.; Wunsch, D.; Zinovyev, A. (Eds.) *Principal Manifolds for Data Visualisation and Dimension Reduction*; Springer: Berlin/Heidelberg, Germany, 2008. https://doi.org/10.1007/978-3-540-73750-6.
21. Schölkopf, B. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **1998**, *10*(5), 1299–1319. https://doi.org/10.1162/089976698300017467.
22. Gorban, A.N.; Zinovyev, A. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems* **2010**, *20*(03), 219–232. https://doi.org/10.1142/S0129065710002383
23. Kramer, M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* **1991***37*(2), 233–243. https://doi.org/10.1002/aic.690370209

24. Hinton G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *28*313(5786), 504–507. https://doi.org/10.1126/science.1127647
25. Gorban, A.N.; Makarov, V.A.; Tyukin, I.Y. High-Dimensional Brain in a High-Dimensional World: Blessing of Dimensionality. *Entropy* **2020**, 22, 82, https://doi.org/10.3390/e22010082.
26. Giannopoulos, A.A.; Milman, V.D. Concentration property on probability spaces. *Adv. Math.* **2000**, *156*, 77–106, https://doi.org/10.1006/aima.2000.1949.
27. Gromov, M. Isoperimetry of waists and concentration of maps. *Geom. Funct. Anal.* **2003**, *13*, 178–215, https://doi.org/10.1007/s00039-009-0703-1.
28. Ledoux, M. *The Concentration of Measure Phenomenon*; Number 89 in Mathematical Surveys & Monographs; AMS: Providence, RI, USA, 2005.
29. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2018.
30. Gartner Hype Cycle for Artificial Intelligence, 2019. https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/.
31. Gartner Hype Cycle for Emerging Technologies, 2020. https://www.gartner.com/en/newsroom/press-releases/2020-08-18-gartner-identifies-five-emerging-trends-that-will-drive-technology-innovation-for-the-next-decade.
32. Gorban, A.N.; Grechuk, B.; Tyukin, I.Y. Augmented Artificial Intelligence: a Conceptual Framework, arXiv preprint **2018**. https://arxiv.org/abs/1802.02172.
33. Kainen, P.; Kůrková, V. Quasiorthogonal dimension of Euclidian spaces. *Appl. Math. Lett.* **1993**, *6*, 7–10. https://doi.org/10.1016/0893-9659(93)90023-G.
34. Kainen, P.; Kůrková, V. Quasiorthogonal dimension. In Kosheleva, O., Shary, S.P., Xiang, G., Zapatrin, R. (Eds.). *Beyond Traditional Probabilistic Data Processing Techniques: Interval, Fuzzy etc. Methods and Their Applications*. Springer, Cham, 2020, 615–629. https://doi.org/10.1007/978-3-030-31041-7_35.
35. Gorban, A.N.; Tyukin, I.; Prokhorov, D.; Sofeikov, K. Approximation with random bases: Pro et contra. *Information Sciences* **2016** *364–365*, 129–145. https://doi.org/10.1016/j.ins.2015.09.021
36. Camastra F. Data dimensionality estimation methods: a survey. *Pattern recognition* **2003**, *36*(12), 2945–2954. https://doi.org/10.1016/S0031-3203(03)00176-6
37. Bac J.; Zinovyev A. Lizard brain: Tackling locally low-dimensional yet globally complex organization of multi-dimensional datasets. *Frontiers in neurorobotics* **2020** *13*, 110. https://doi.org/10.3389/fnbot.2019.00110
38. Moczko, E.; Mirkes, E.M.; Cáceres, C.; Gorban, A.N.; Piletsky, S. Fluorescence-based assay as a new screening tool for toxic chemicals. *Scientific reports* **2016**, *6*(1), 33922. https://doi.org/10.1038/srep33922
39. Chen, S. H.; Pollino, C. A. Good practice in Bayesian network modelling. *Environmental Modelling & Software* **2012**, *37*, 134–145. https://doi.org/10.1016/j.envsoft.2012.03.012
40. Cobb, B.R.; Rumí, R.; Salmerón, A. Bayesian Network Models with Discrete and Continuous Variables. In: Lucas P., Gámez J.A., Salmerón A. (eds) *Advances in Probabilistic Graphical Models. Studies in Fuzziness and Soft Computing*, vol 213. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-68996-6_4
41. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of the 31st International Conference on Machine Learning 2014 Jun 22-24 Eds. E.P. Xing and T. Jebara (pp. 1278–1286). PMLR. http://proceedings.mlr.press/v32/rezende14.html
42. Hinton G.E. A Practical Guide to Training Restricted Boltzmann Machines. In: Montavon G., Orr G.B., Müller KR. (eds) Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 7700. Springer, Berlin, Heidelberg, 2012, 599–619. https://doi.org/10.1007/978-3-642-35289-8_32
43. Streiner, D.L.; Norman, G.R. Correction for multiple testing: is there a resolution?. *Chest* **2011***140*(1), 16–18. https://doi.org/10.1378/chest.11-0523
44. Noble, W.S. How does multiple testing correction work?. *Nat Biotechnol* **2009**, *27*, 1135–1137. https://doi.org/10.1038/nbt1209-1135
45. Jolliffe, I. *Principal Component Analysis*; Springer: Berlin/Heidelberg, 1993.
46. Sompairac, N.; Nazarov, P.V.; Czerwinska, U.; Cantini, L.; Biton, A.; Molkenov, A.; Zhumadilov, Z.; Barillot, E.; Radvanyi, F.; Gorban, A.; Kairov, U. Independent component analysis for unraveling the complexity of cancer omics datasets. *International Journal of molecular sciences* **2019**, *20*(18), 4414. http://dx.doi.org/10.3390/ijms20184414

47. Hicks, S.C.; Townes, F.W.; Teng, M.; Irizarry, R.A.. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **2018**, *19*(4), 562–578. https://doi.org/10.1093/biostatistics/kxx053
48. Krumm, N.; Sudmant, P.H.; Ko, A.; O'Roak, B.J.; Malig, M.; Coe, B.P.; Quinlan, A.R.; Nickerson, D.A.; Eichler, E.E., Copy number variation detection and genotyping from exome sequence data. *Genome research* **22**(8), 1525–1532. https://doi.org/10.1101/gr.138115.112
49. Koren, Y.; Carmel, L. Robust linear dimensionality reduction. *IEEE Trans Visual Comput Graph.* **2004** *10*(4), 459–470. https://doi.org/10.1109/TVCG.2004.17
50. Mirkes, E.M., Gorban, A.N., Zinoviev, A. Supervised PCA. https://github.com/Mirkes/SupervisedPCA. 2016.
51. Gorban, A.N.; Mirkes, E.M.; Tukin I.Y. How deep should be the depth of convolutional neural networks: a backyard dog case study. *Cognitive Computation* **2020**, *12*, 388–397. https://doi.org/10.1007/s12559-019-09667-7
52. Song, Y.; Nie, F.; Zhang, C.; Xiang, S. A unified framework for semi-supervised dimensionality reduction. *Pattern Recognition* **2008**, *41*(9), 2789–2799. https://doi.org/10.1016/j.patcog.2008.01.001.
53. Cangelosi, R.; Goriely, A. Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* **2007**, *2*(1), 2. doi:10.1186/1745-6150-2-2
54. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. https://doi.org/10.1007/s10994-009-5152-4
55. Sun, S.: Shi, H.; Wu, Y. A survey of multi-source domain adaptation. *Information Fusion* **2015**, *24*, 84–92. https://doi.org/10.1016/j.inffus.2014.12.003
56. Saito, K.; Watanabe, K., Ushiku, Y.; Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, 3723–3732. https://doi.org/10.1109/CVPR.2018.00392
57. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* **2016**, *17*(59), 2096–2030. http://jmlr.org/papers/v17/15-239.html
58. Matasci, G.; Volpi, M.; Tuia, D.; Kanevski, M. Transfer component analysis for domain adaptation in image classification. In Image and Signal Processing for Remote Sensing XVII (Vol. 8180, p. 81800F). International Society for Optics and Photonics, 2011. https://doi.org/10.1117/12.898229
59. Pestov, V. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Comput. Math. Appl.* **2013**, *65*, 1427–1437. https://doi.org/10.1016/j.camwa.2012.09.011
60. Mirkes, E.M.; Allohibi, J.; Gorban, A.N. Fractional Norms and Quasinorms Do Not Help to Overcome the Curse of Dimensionality. *Entropy* **2020**, *22*, 1105. https://doi.org/10.3390/e22101105.
61. Zadeh, L.A. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.* **1997**, *19*, 111–127. https://doi.org/10.1016/S0165-0114(97)00077-8
62. Pedrycz, W.; Skowron, A.; Kreinovich, V. (Eds) Handbook of granular computing. John Wiley & Sons; 2008.
63. Kainen, P.C. Replacing points by compacta in neural network approximation. *Journal of the Franklin Institute* **2004**, *341*(4), 391–399. https://doi.org/10.1016/j.jfranklin.2004.03.001
64. Guédon, O.; Milman, E. Interpolating thin-shell and sharp large-deviation estimates for lsotropic log-concave measures. *Geometric and Functional Analysis* **2011**, *21*(5), 1043–1068. https://doi.org/10.1007/s00039-011-0136-5
65. Lévy, P. *Problèmes Concrets D'analyse Fonctionnelle*. Gauthier-Villars: Paris, France, 1951.
66. Khinchin, A.Y. *Mathematical Foundations of Statistical Mechanics.* New York: Courier Corporation, 1949. (English translation from the Russian edition, Moscow – Leningrad, 1943.)
67. Thompson, C.J. *Mathematical statistical mechanics*. Princeton University Press; 2015.
68. Kolmogorov, A.N., *Foundations of the theory of probability.* Second English Edition. Courier Dover Publications; 2018.
69. Liu, L.; Shao, L.; Li, X. Evolutionary compact embedding for large-scale image classification. *Information Sciences* **2015**, *316*, 567–581. https://doi.org/10.1016/j.ins.2014.06.030
70. Vemulapalli, R.; Agarwala, A. A compact embedding for facial expression similarity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, 5683–5692. https://openaccess.thecvf.com/content_CVPR_2019/html/Vemulapalli_A_Compact_Embedding_for_Facial_Expression_Similarity_CVPR_2019_paper.html

71. Bhattarai, B.; Liu, H.; Huang, H.H. Ceci: Compact embedding cluster index for scalable subgraph matching. In Proceedings of the 2019 International Conference on Management of Data 2019 Jun 25 (pp. 1447-1462). https://doi.org/10.1145/3299869.3300086
72. Tyukin, I.Y.; Higham, D.J.; Gorban, A. N. 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, 2020. https://doi.org/10.1109/IJCNN48605.2020.9207472.
73. Colbrook, M.J.; Antun. V.; Hansen, A.C. Can stable and accurate neural networks be computed?–On the barriers of deep learning and Smale's 18th problem. arXiv preprint arXiv:2101.08286 2021.
74. Rudin, W. *Functional Analysis*. New York, NY: McGraw-Hill Science/Engineering/Math, 1991.
75. Xu, R; Wunsch, D. *Clustering*. John Wiley & Sons; 2008.
76. Tyukin, I.Y., Gorban, A.N., McEwan, A.A., Meshkinfamfard, S.; Tang, L. Blessing of dimensionality at the edge and geometry of few-shot learning. *Information Sciences* **2021**, *564*, 124–143. https://doi.org/10.1016/j.ins.2021.01.022
77. Tao, C.W. Unsupervised fuzzy clustering with multi-center clusters. *Fuzzy Sets and Systems* **2002**, *128*(3), 305–322. https://doi.org/10.1016/S0165-0114(01)00191-9
78. Krizhevsky, A. Learning multiple layers of features from tiny images. Technical Report, University of Toronto, 2009; **CiteSeerX** https://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.222.9220
79. Krizhevsky, A. CIFAR 10 Dataset, University of Toronto, 2009. https://www.cs.toronto.edu/~kriz/cifar.html
80. Golubitsky, M., Guillemin, V. *Stable Mappings and Their Singularities*. Springer-Verlag. 1974.
81. Pugh, C. The closing lemma. *Amer. J. Math.* **1967**, *89*, 956–1009. https://doi.org/10.2307/2373413
82. Palis, J.; de Melo W. The Kupka-Smale Theorem. In: Geometric Theory of Dynamical Systems. Springer, New York, NY, 1982. https://doi.org/10.1007/978-1-4612-5703-5_3
83. Oxtoby, J.C. *Measure and category: A survey of the analogies between topological and measure spaces*. Springer; NY, 2013.
84. Gorban, A.N. *Equilibrium encircling. Equations of chemical kinetics and their thermodynamic analysis.* Nauka, Novosibirsk, 1984.
85. Gorban, A.N. Selection Theorem for Systems with Inheritance, *Math. Model. Nat. Phenom.* **2007**, *2*(4), 1–45. https://doi.org/10.1051/mmnp:2008024
86. Feynman, R.P.; Leighton, R.B.; Sands, M. *The Feynman Lectures on Physics*, Addison Wesley, 2005, Volume I, Chapter 1, Introduction.