Bringing the Blessing of Dimensionality to the Edge

1st Ivan Y. Tyukin[1,2,3]
[1] *Department of Mathematics*
University of Leicester
[2] *Lobachevsky University*
[3] *St-Petersburg State Electrotechnical University*
[1] Leicester, United Kingdom
I.Tyukin@le.ac.uk

2nd Alexander N Gorban
University of Leicester
and Lobachevsky University
Leicester, UK, and Nizhni Novgorod, Russia
a.n.gorban@le.ac.uk

3rd Alistair McEwan
Department of Engineering
University of Leicester
Leicester, United Kingdom
aam19@le.ac.uk

4th Sepehr Meshkinfamfard
Department of Mathematics
University of Leicester
Leicester, United Kingdom
sm901@le.ac.uk

***Abstract*—In this work we present a novel approach and algorithms for equipping Artificial Intelligence systems with capabilities to become better over time. A distinctive feature of the approach is that, in the supervised setting, the approaches' computational complexity is sub-linear in the number of training samples. This makes it particularly attractive in applications in which the computational power and memory are limited. The approach is based on the concentration of measure effects and stochastic separation theorems. The algorithms are illustrated with examples.**

***Index Terms*—Stochastic separation theorems, artificial intelligence, machine learning, computer vision**

I. Introduction

Large and growing streams of data are ubiquitous in modern society and technology. They are the backbone of modern healthcare, public safety and security services, sciences, including satellite Earth Observation missions, and world-wide web. Sustained functioning and progress in these essential areas depend on the ability to extract and process information from large and growing data. Since processing overwhelmingly large volumes of data can no longer be accomplished by humans alone, we must rely on Artificial Intelligence (AI) systems built on state-of-the-art machine learning and data analytics technologies.

With the explosive pace of progress in computing such as the emergence and accessibility of Nvidia GPU devices, Google TPUs, and cloud resources to name just a few, current AI systems are now capable of spotting minute patterns in large data sets and can outperform humans in highly complicated tasks like chess, Go, and medical diagnosis [1]. However, the super-human power of modern AIs to learn from massive volumes of data make their conclusions vulnerable to data inconsistencies, poor data quality, and uncertainty inherent to any data. This uncertainty, together with engineering constraints on AI's implementation, lead to inevitable errors in data-driven AIs.

Consequences of AI errors range from mere technical inconveniences to significant public and societal risks: incorrect cancer treatment options by IBM Watson and several Tesla and Uber crashes in 2018 are few examples. Structuring data, improving its quality, and removing uncertainty is generally very resource intensive and thus unsustainable across sectors and industries. More fundamentally, constraints on AIs implementations such as e.g. quantization errors and memory limitations are major limits to AI's performance "at the edge", i.e. in embedded setting.

Significant efforts have been applied to date to address errors in AI systems. Using ensembles [2], [3], [4], augmenting training data [5], [6], [7], enforcing continuity [8], and AI knowledge transfer [9], [10], [11] have been extensively discussed in the literature. These measures, however, do not warrant error-free behaviour as AIs based on empirical data are expected to make mistakes.

Recently, [12], [13], [14], [15], [16], [17] we have shown that spurious errors of AI systems operating in essentially high-dimensional spaces (convolutional and deep learning neural networks are an example of such AI), can be efficiently removed by Fisher discriminants. The advantage of this approach over, for instance, support vector machines [18] is that the computational complexity for constructing Fisher discriminants is at most linear in the number of points in the training set whereas the worst-case complexity for support vector machines scales as a cubic function of the training set's size [19].

The method applies to identified singular spurious errors as well as their moderate-sized clusters. The question, however, is what if the volume of errors which the AI's produce is

The work was supported by Innovate UK Knowledge Transfer Partnership grant KTP010522 and by the grant of the Russian Science Foundation (Project No. 19-19-00566).

similar to that of AI's correct response? The second question is if it is possible for deployed AI's to learn autonomously without supervision? The above questions are fundamentally important for the broad spectrum of AI applications. At the same time, they are particularly acute for embedded and near-edge systems due to constraints imposed by their limited computational capabilities and memory resources.

In this work, we show that stochastic separation theorems, or the blessing of dimensionality [20], [21], stemming from the concentration of measure effects [22], [23], [24], can be adapted and applied to address these important issues too. In particular we present and justify mathematically and in experiments two algorithms that are capable of delivering the required functionality at the computational costs which are compatible with strict computational requirements of embedded deployment.

The paper is organized as follows. Section II contains necessary theoretical preliminaries and formal statement of the problem. In Section III we present a new algorithm for improving AIs "at the edge", Section IV presents numerical example, and Section V concludes the paper.

NOTATION

The following notational agreements are used throughout the text:

- $\mathbb{R}^n$ stands for the n-dimensional linear real vector space;
- $\mathbb{N}$ denotes the set of natural numbers;
- symbols $\boldsymbol{x} = (x_1, \ldots, x_n)$ will denote elements of $\mathbb{R}^n$;
- $(\boldsymbol{x}, \boldsymbol{y}) = \sum_k x_k y_k$ is the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$ is the standard Euclidean norm in $\mathbb{R}^n$;
- $\mathbb{B}_n$ denotes for the unit ball in $\mathbb{R}^n$ centered at the origin: $\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n | \ (\boldsymbol{x}, \boldsymbol{x}) \leq 1\}$;
- V_n is the n-dimensional Lebesgue measure, and $V_n(\mathbb{B}_n)$ is the volume of unit ball;
- if $\mathcal{Y}$ is a finite set then the number of elements in $\mathcal{Y}$ (cardinality of $\mathcal{Y}$) is denoted by $|\mathcal{Y}|$.

II. PROBLEM FORMULATION AND MATHEMATICAL PRELIMINARIES

A. Problem formulation

Following [14], we consider a generic AI system that processes some *input* signals, produces *internal* representations of the input and returns some *outputs*. We assume that there is a sampling process whereby some relevant information about the input, internal signals, and outputs are combined into a common vector, $\boldsymbol{x}$, representing, but not necessarily defining, the *state* of the AI system.

Depending on the sampling process, the vector $\boldsymbol{x}$ may have various numbers of elements. But generally, the objects $\boldsymbol{x}$ are assumed to be elements of $\mathbb{R}^n$, with n depending on the sampling process. Over a period of activity the AI system generates a set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ of representations $\boldsymbol{x}$. In agreement with standard assumptions in machine learning literature [18], we assume that the set $\mathcal{X}$ is is a random sample drawn from some distribution. The distribution that generates vectors $\boldsymbol{x}$ is supposed to be *unknown*. We will, however, impose the following technical assumption on the generating probability distribution

Assumption 1: The probability density function, p, associated with the probability distribution of the random variable $\boldsymbol{x}$ is compactly supported in the unit ball $\mathbb{B}_n$ and

$$\text{there exists a } C > 0: \ p(\boldsymbol{x}) < \frac{Cr^n}{V_n(\mathbb{B}_n)}, \ r \in (0, 2).$$

for all $\boldsymbol{x} \in \mathbb{B}_n$ and relevant n.

The assumption requires that the random variable $\boldsymbol{x}$ is in $\mathbb{B}_n$ which is consistent with the scope of our applications. Additionally, it states that, as the number of variables in vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ grows, no unexpected concentrations in the probability distributions emerge as a result of this growth (cf. the Smeared Absolute Continuity (SmAC) property in [16], [25]). Awareness of the latter property will be important for the algorithms that follow.

Definition 1: A point $\boldsymbol{x} \in \mathbb{R}^n$ is linearly separable from a set $\mathcal{Y} \subset \mathbb{R}^n$, if there exists a linear functional $l(\cdot)$ such that

$$l(\boldsymbol{x}) > l(\boldsymbol{y})$$

for all $\boldsymbol{y} \in \mathcal{Y}$.

Definition 2: A set $\mathcal{X} \subset \mathbb{R}^n$ is linearly separable from a set $\mathcal{Y} \subset \mathbb{R}^n$, if there exists a linear functional $l(\cdot)$ such that

$$l(\boldsymbol{x}) > l(\boldsymbol{y})$$

for all $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{x} \in \mathcal{X}$.

In addition to these standard notions of linear separability, we adopt the notion of Fisher separability [16], [25].

Definition 3: A point $\boldsymbol{x} \in \mathbb{R}^n$ is Fisher separable from a set $\mathcal{Y} \subset \mathbb{R}^n$, if

$$(\boldsymbol{x}, \boldsymbol{x}) > (\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

for all $\boldsymbol{y} \in \mathcal{Y}$. The point is Fisher separable from the set $\mathcal{Y}$ with a threshold $\kappa \in [0, 1)$ if

$$(\boldsymbol{x}, \boldsymbol{x}) > \kappa(\boldsymbol{x}, \boldsymbol{y}) \tag{2}$$

Having introduced all relevant assumptions and notions, we are now ready to proceed with results underpinning our algorithmic developments.

B. Mathematical Preliminaries

Our first result is provided in Theorem 1 (cf. [26], [16])

Theorem 1: Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be given, $\boldsymbol{x}_i \in \mathbb{B}_n$, and let $\boldsymbol{x}$ be drawn from a distribution satisfying Assumption 1. Then $\boldsymbol{x}$ is Fisher separable from the set $\mathcal{X}$ with probability

$$P \geq 1 - MC\left(\frac{r}{2}\right)^n, \ r \in (0, 2). \tag{3}$$

Proof of Theorem 1. Consider events

$$A_i: \ \boldsymbol{x} \text{ is Fisher separable from } \boldsymbol{y}_i$$

It is clear that

$$P(\text{not } A_i) = \int_{(\boldsymbol{x}, \boldsymbol{x}) - (\boldsymbol{x}, \boldsymbol{x}_i) \leq 0} p(\boldsymbol{x}) d\boldsymbol{x}.$$

Recall that

$$P(A_1 \& A_2 \& \cdots \& A_M) \geq 1 - \sum_{i=1}^{M} P(\text{not } A_i).$$

Combining the last two observations we can conclude that the probability that $\boldsymbol{x}$ is separable from all $\boldsymbol{x}_i$ is bounded from below by the expression in (3). □

Consider two random sets $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ and $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\}$. Let there be a process (e.g. a learning algorithm) which, for the given $\mathcal{X}$, $\mathcal{Y}$ or their subsets, produces a classifier

$$f(\cdot) = \sum_{i=1}^{d} \alpha_i(\boldsymbol{z}_i, \cdot), \ \alpha_j \in \mathbb{R}.$$

The vectors $\boldsymbol{z}_i$, $i = 1, \ldots, d$ are supposed to be known. Furthermore, we suppose that the function f is such that

$$f(\boldsymbol{y}_j) > \sum_{m,k=1}^{d} \alpha_m \alpha_k (\boldsymbol{z}_m, \boldsymbol{z}_k) \tag{4}$$

for all $\boldsymbol{y}_j \in \mathcal{Y}$. In other words, if we denote $\boldsymbol{w} = \sum_{i=1}^{d} \alpha_i \boldsymbol{z}_i$, the following holds true:

$$(\boldsymbol{w}, \boldsymbol{w}) < (\boldsymbol{w}, \boldsymbol{y}_i) \text{ for all } i = 1, \ldots, K. \tag{5}$$

Note that since the $\mathcal{Y}, \mathcal{X}$ are random, it is natural to expect that the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ is also random. The following statement can now be formulated:

Theorem 2: Consider sets $\mathcal{X}$ and $\mathcal{Y}$. Let $p_\alpha(\boldsymbol{\alpha})$ be the probability density function associated with the random vector $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}$ satisfies condition (4) with probability 1. Then the set $\mathcal{X}$ is separable from the set $\mathcal{Y}$ with probability

$$P \geq 1 - \sum_{i=1}^{M} \int_{H(\boldsymbol{\alpha}, \boldsymbol{x}_i) \leq 0} p_\alpha(\boldsymbol{\alpha}) d\boldsymbol{\alpha}, \tag{6}$$

where

$$H(\boldsymbol{\alpha}, \boldsymbol{x}_i) = \sum_{k,m=1}^{d} \alpha_k \alpha_m (\boldsymbol{z}_k, \boldsymbol{z}_m) - \sum_{m=1}^{d} \alpha_m (\boldsymbol{z}_m, \boldsymbol{x}_i).$$

Proof of Theorem 2. Consider events

$$A_i : \ (\boldsymbol{w}, \boldsymbol{w}) > (\boldsymbol{w}, \boldsymbol{x}_i).$$

Events A_i are equivalent to that $H(\boldsymbol{\alpha}, \boldsymbol{x}_i) > 0$. Eq. (6) provides a lower bound for the probability that all these events hold. Recall that vectors $\boldsymbol{\alpha}$ satisfy (5), and hence

$$\sum_{m=1}^{d} \alpha_m (\boldsymbol{z}_m, \boldsymbol{x}_i) = (\boldsymbol{w}, \boldsymbol{x}_i) < (\boldsymbol{w}, \boldsymbol{y}_j) = \sum_{m=1}^{d} \alpha_m (\boldsymbol{z}_m, \boldsymbol{y}_j)$$

for all $\boldsymbol{x}_i \in \mathcal{X}$ and $\boldsymbol{y}_j \in \mathcal{Y}$ with probability at least (6). The statement now follows immediately from Definition 2. □

Theorem 2 generalizes earlier k-tuple separation theorems [15] to a very general class of practically relevant distributions. No independence assumptions are imposed on the components of vectors $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$. We do, however, require that some information about distribution of the classifier parameters, $\boldsymbol{\alpha}$, is available.

Observe, for example, that if there exist $L > 0$, $\lambda \in (0, 1)$ and a function $\beta : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ such that

$$\int_{H(\boldsymbol{\alpha}, \boldsymbol{y}) \leq 0} p_\alpha(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \leq L \lambda^{\beta(d,n)}$$

for any $\boldsymbol{y} \in \mathbb{R}^n$ then (6) becomes

$$P \geq 1 - ML\lambda^{\beta(d,n)}.$$

If $d = n$ and elements of the set $\mathcal{Y}$ are sufficiently strongly correlated, then Theorem 1 provides a good approximation of the separability probability bound for a simple separating function in which $\boldsymbol{w}$ is just a scaled centroid of $\mathcal{Y}$.

III. Fast removal of AI errors

According to theoretical constructions presented in the previous section, the following is an advantage for successful and efficient separation of random sets in high dimension: one of the sets should be sufficiently concentrated (spatially localized and have an exponentially smaller volume relative to the other [Theorems 1, 2]). Moreover, successful separability of a single point is dependent on absence of unexpected concentrations in the probability distributions. Below we present an algorithm for fast and efficient error correction of AI systems which is motivated by these observations and intuition stemming from our theoretical results.

Consider two finite sets, the set $\mathcal{X} \subset \mathbb{R}^n$, and $\mathcal{Y} \subset \mathbb{R}^n$. The task is to construct a linear functional separating the set $\mathcal{X}$ from $\mathcal{Y}$.

Algorithm 1:

1) Let $\bar{\boldsymbol{x}}$ be the centroid of $\mathcal{X}$. 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) Select $m \leq n$ Principal Components, $h_1, \ldots, h_m$, 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.
4) For the sets $\mathcal{X}_r$ and $\mathcal{Y}_{*,r}$, construct the Fisher discriminant separating $\mathcal{X}_r$ from $\mathcal{Y}_{*,r}$. Let $\boldsymbol{w}_r$ be a vector such that if $(\boldsymbol{w}_r, \boldsymbol{z}) > (\boldsymbol{w}_r, \boldsymbol{w}_r)$ then we associate element $\boldsymbol{z}$ with the set $\mathcal{Y}_{r,*}$; if $(\boldsymbol{w}_r, \boldsymbol{z}) \leq (\boldsymbol{w}_r, \boldsymbol{w}_r)$ then we associate $\boldsymbol{z}$ with the set $\mathcal{X}_r$.
5) Calculate $\boldsymbol{w} = [h_1 \vdots \cdots \vdots h_m]\boldsymbol{w}_r$ and return $l(\cdot) = (\boldsymbol{w}, \cdot)$ as the desired linear functional for $\mathcal{X}_c$ and $\mathcal{Y}_*$.

IV. Numerical Example

To illustrate the efficiency of the approach, we tested the algorithm in the object detection task in which the primary object detector was an OpenCV implementation of the Haar face detector. The detector has been applied to a video footage capturing traffic and pedestrians walking on the streets of Montreal. For the purposes of testing and validation, we used

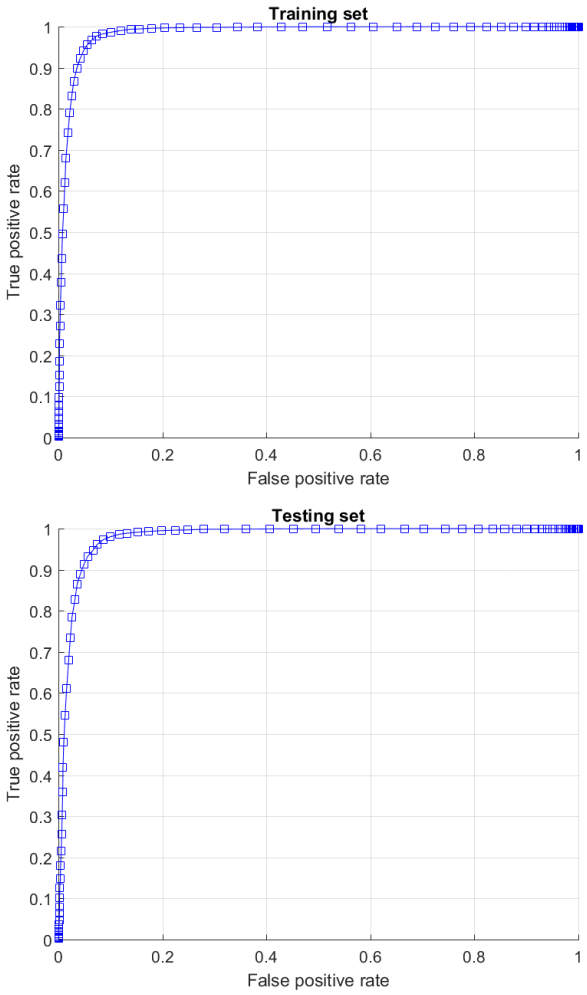

Fig. 1. ROC curves after the application of the AI error correcting algorithm

the MTCNN face detector as a vehicle to generate ground truth data. All the data as well as the code generating true positive and false positive images can be provided by request.

For this particular dataset, the total number of true positives was 21896, and the total number of false positives was 9372. All the detects have been resized to 64 by 64 crops (in RGB encoding). Each crop produces a 12288-dimensional vector. From this dataset, we generated a training set containing 50 percent of positive and false positives, and passed this training set to Algorithm 1. In the algorithm, true positives have been associated with the set $\mathcal{X}$, and false positives were associated with the set $\mathcal{Y}_*$. The number of Principal Components was limited to 200.

Training took 183.87 seconds on a Core i7 laptop, and the outcomes of the process as well as performance on the testing set are shown in Fig. 1. As we can see from this figure, direct implementation of Algorithm 1 allows one to filter 90 percent of all errors at the cost of missing circa 5 percent of true positives. Most importantly, implementation of the correcting functional on an ARM Cortex-A53 processor took less than 1 millisecond per each 12288-dimensional vector implying significant capacity of the approach for embedded near-edge applications.

V. Conclusion

In this work we presented a novel approach for equipping edge-based or near-edge devices with capabilities to remove spurious as well as a rather overwhelming number of errors. The approach is based on stochastic separation theorems [12], [13], [14], [15], [16], [17] and the concentration of measure phenomena. Our results demonstrate that the new capability can be delivered to the edge and deployed in a fully automated way, whereby a more sophisticated AI system, e.g. MTCNN face detector, monitors performance of a less powerful counterpart. The experiment extends our earlier work [15] and directly responds to the fundamental challenge of removing AI errors in industrial applications at minimal computational costs.

Acknowledgment

The work was supported by Innovate UK Knowledge Transfer Partnership grant KTP010522 and by the grant of the Russian Science Foundation (Project No. 19-19-00566).

References

[1] H. Liang, B. Tsui, H. Ni, C. Valentim, S. Baxter, G. Liu, W. Cai, D. Kermany, K. Sun, J. Chen, L. He, J. Zhu, P. Tian, H. Shao, L. Zheng, R. Hou, S. Hewett, G. Li, P. Liang, X. Zang, Z. Zhang, L. Pan, H. Cai, R. Ling, S. Li, Y. Cui, S. Tang, H. Ye, X. Huang, W. He, W. Liang, Q. Zhang, J. Jiang, W. Yu, J. Gao, W. Ou, Y. Deng, Q. Hou, B. Wang, C. Yao, Y. Liang, S. Zhang, Y. Duan, R. Zhang, S. Gibson, C. Zhang, O. Li, E. Zhang, G. Karin, N. Nguyen, X. Wu, C. Wen, J. Xu, W. Xu, B. Wang, W. Wang, J. Li, B. Pizzato, C. Bao, D. Xiang, W. He, S. He, Y. Zhou, W. Haw, M. Goldbaum, A. Tremoulet, C.-N. Hsu, H. Carter, L. Zhu, K. Zhang, and H. Xia, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature Medicine*, vol. 25, pp. 433–438, 2019.

[2] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. on Patt. Analyis and Mach. Intel.*, vol. 12, no. 10, pp. 993–1001, 1990.

[3] T. K. Ho, "Random decision forests," in *Proc. of the 3rd International Conference on Document Analysis and Recognition*, 1995, pp. 993–1001.

[4] ——, "The random subspace method for constructing decision forests," *IEEE Trans. on Patt. Analyis and Mach. Intel.*, vol. 20, no. 8, pp. 832–844, 1998.

[5] A. Kuznetsova, S. Hwang, B. Rosenhahn, and L. Sigal, "Expanding object detectors horizon: Incremental learning framework for object detection in videos," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 28–36.

[6] I. Misra, A. Shrivastava, and M. Hebert, "Semi-supervised learning for object detectors from video," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3594–3602.

[7] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3282–3289.

[8] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, https://arxiv.org/abs/1604.04326.

[9] L. Pratt, "Discriminability-based transfer between neural networks," *Advances in Neural Information Processing*, no. 5, pp. 204–211, 1992.

[10] T. Chen, I. Goodfellow, and J. Shlens, "Net2net: Accelerating learning via knowledge transfer," *ICLR 2016*, 2015.

[11] V. Vapnik and R. Izmailov, "Knowledge transfer in svm and neural networks," *Annals of Mathematics and Artificial Intelligence*, pp. 1–17, 2017.

[12] A. Gorban and I. Tyukin, "Stochastic separation theorems," *Neural Networks*, vol. 94, pp. 255–259, 2017.
[13] A. Gorban, I. Tyukin, and I. Romanenko, "The blessing of dimensionality: Separation theorems in the thermodynamic limit," 09 2016, a talk given at TFMST 2016, 2nd IFAC Workshop on Thermodynamic Foundations of Mathematical Systems Theory. September 28-30, 2016, Vigo, Spain. [Online]. Available: https://arxiv.org/abs/1610.00494v1
[14] A. Gorban, R. Burton, I. Romanenko, and I. Tyukin, "One-trial correction of legacy AI systems and stochastic separation theorems," *Information Sciences*, vol. 484, pp. 237–254, 2019.
[15] I. Y. Tyukin, A. N. Gorban, K. Sofeikov, and I. Romanenko, "Knowledge transfer between artificial intelligence systems," *Frontiers of Neurorobotics*, vol. 12, Article 49, 2018.
[16] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of ai systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303–322, 2018.
[17] I. Tyukin, A. Gorban, S. Green, and D. Prokhorov, "Fast construction of correcting ensembles for legacy artificial intelligence systems: Algorithms and a case study," *Information Sciences*, vol. 485, pp. 230–247, 2019.
[18] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 2000.
[19] O. Chapelle, "Training a support vector machine in the primal," *Neural computation*, vol. 19, no. 5, pp. 1155–1178, 2007.
[20] A. Gorban and I. Tyukin, "Blessing of dimensionality: mathematical foundations of the statistical physics of data," *Philosophical Transactions of the Royal Society A*, vol. 376, p. 20170237, 2018.
[21] P. C. Kainen, "Utilizing geometric anomalies of high dimension: When complexity makes computation easier," in *Computer Intensive Methods in Control and Signal Processing*. Springer, 1997, pp. 283–294.
[22] A. Gorban, I. Tyukin, D. Prokhorov, and K. Sofeikov, "Approximation with random bases: Pro et contra," *Information Sciences*, vol. 364–365, pp. 129–145, 2016.
[23] P. Kainen and V. Kurkova, "Quasiorthogonal dimension of euclidian spaces," *Appl. Math. Lett.*, vol. 6, no. 3, pp. 7–10, 1993.
[24] M. Gromov, "Isoperimetry of waists and concentration of maps," *GAFA, Geomteric and Functional Analysis*, vol. 13, pp. 178–215, 2003.
[25] A. N. Gorban, B. Grechuk, and I. Y. Tyukin, "Augmented artificial intelligence," *arXiv preprint arXiv:1802.02172*, 2018.
[26] A. N. Gorban, V. A. Makarov, and I. Y. Tyukin, "The unreasonable effectiveness of small neural ensembles in high-dimensional brain," *Physics of Life Reviews*, 2018.