Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 29 (2019) 98–100

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Comment

Some insights from high-dimensional spheres Comment on "The unreasonable effectiveness of small neural ensembles in high-dimensional brain" by Alexander N. Gorban et al.

Věra Kůrková

Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic

Received 19 March 2019; accepted 25 March 2019
Available online 27 March 2019
Communicated by L. Perlovsky

The title of this article by Gorban et al. refers to Wigner's famous lecture, "The unreasonable effectiveness of mathematics in the natural sciences" [1], delivered 60 years ago in 1959. In the lecture, Wigner emphasized the crucial role of mathematics in developing consistent theories in physics. Similarly, Gorban et al. focus on the role of mathematics in understanding nature, namely the functioning and structure of brains. They utilize mathematics of high-dimensional spaces to explain "how can high-dimensional brain organize reliable and fast learning in high-dimensional world of data by simple tools?".

A number of neurobiological studies have observed the energy efficiency of the brain which seems to exhibit both sparse activity (only a small fraction of neurons have a high rate of firing at any time) and sparse connectivity (each neuron is connected to only a limited number of other neurons) [2]. Gorban et al. suggest that sparse coding of information in the brain can be explained using high-dimensional geometry. They approach the investigation of biological neural networks by exploring a much simpler case of artificial ones.

In recent years, neurocomputing achieved impressive successes [3]. In particular, randomized models and algorithms for neural networks have turned out to be quite efficient for performing high-dimensional tasks. Theoretical analysis complements the experimental evidence of almost deterministic behavior of stochastic algorithms on large networks and/or large data sets. With increase in data dimension and network size, outputs tend to be sharply concentrated around precalculated values. This behavior can be explained by the geometry of high-dimensional spaces, which have many counter-intuitive properties - difficult to visualize for us who live in three-dimensional space. Mathematics alone guides us in these higher dimensions, where senses cannot reach.

Using classical calculus (integration in spherical polar coordinates), one can compute the relative area of the d-dimensional sphere, which is occupied by the polar cap. More precisely, let S^{d-1} denote the unit sphere (the set of vectors of length 1) in the d-dimensional Euclidean space and $C(g,\varepsilon) = \{f \in S^{d-1} \mid \langle u, v\rangle \geq \varepsilon\}$ the polar cap centered at a fixed vector g, which contains all vectors f which have the angular distance from g at most $\alpha = \arccos\varepsilon$ (the inner product $\langle f, g\rangle$ is at least ε), see Fig. 1. For a fixed angle α, with increasing dimension d the normalized surface area μ of such cap decreases exponentially fast to zero as $\mu(C(g,\varepsilon)) \leq e^{-\frac{d\varepsilon^2}{2}}$ (see, e.g., [4]). It is quite surprising to

DOI of original article: https://doi.org/10.1016/j.plrev.2018.09.005.
E-mail address: vera@cs.cas.cz.

https://doi.org/10.1016/j.plrev.2019.03.014

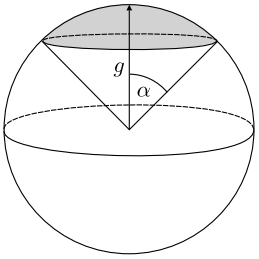

Fig. 1. Polar cap.

realize that most of the area of a high-dimensional sphere is concentrated around the equator (metaphorically in the "tropic" region).

The exponential decrease of sizes of polar caps is the very essence of two properties of high-dimensional spaces called the *curse of dimensionality* [5,6] and the *blessing of dimensionality* [7,8]. As noted by Gorban et al., they are two sides of the same coin. The upper bound $e^{-\frac{d\varepsilon^2}{2}}$ on the relative measure of polar caps implies that one can pack exponentially many disjoint caps in the d-dimensional sphere. For small ε corresponding to the angle $\alpha = \arccos \varepsilon$ close to 90 degrees, this packing number gives the lower bound $e^{\frac{d\varepsilon^2}{2}}$ on the *quasiorthogonal dimension* $\dim_\varepsilon d$. It is defined by replacing the condition on strict orthogonality with merely quasiorthogonality (scalar product of each pair of distinct vectors is smaller than ε) [9]. While there are only d exactly orthogonal unit vectors in the d-dimensional space, the number of ε-orthogonal vectors grows with d exponentially. Even for moderate dimensions of data d, the number of such highly uncorrelated vectors could be unmanageably large.

On the other hand, geometry of high-dimensional spheres implies concentration of values of functions of many variables and possibilities of reduction of dimensionality of data. The upper bound $e^{-\frac{d\varepsilon^2}{2}}$ on the size of a polar cap can be rephrased as follows: inner products of a fixed vector with uniformly randomly chosen vectors concentrate around zero. A generalization obtained by replacing the inner product with a sufficiently smooth (Lipschitz) function on the sphere gives the Lèvy Lemma [10]. It states that almost all values of a Lipschitz function on a high-dimensional sphere are close to their median. This property of high-dimensional spheres is called the *concentration of measure phenomenon*. Similar property was also discovered in probability theory, where it has been studied in terms of bounds on large deviations of sums of random variables by Hoeffding [11], Chernoff [12], and Azuma [13]. It implies that randomized techniques work almost deterministically [14]. Concentration of measure is also the essence of the proof of the Johnson-Lindenstrauss Flattening Lemma. It guarantees a possibility of dimension reduction of d-dimensional data by a random projection to a lower dimension bounded from below by $\frac{8}{\varepsilon} \log d$ such that the projection is a near-isometry (preserves distances within a multiplicative factor $1 \pm \varepsilon$) (see, e.g., [15]).

It should be emphasized that the Lèvy Lemma assumes uniform probability distribution. Also in learning theory, typically it is assumed that data are independent and identically distributed [16,17]. Under these assumptions, Gorban et al. derived stochastic separation theorems for classification [18,19]. But independence of random variables is a strong assumption. The hypothesis that a probability distribution can be expressed as a product probability is called the "naive Bayes assumption" [20]. Often in real tasks, neither uniform probability nor independence of data is satisfied. Nevertheless, some versions of the concentration of measure phenomenon hold even for more general distributions and in dependent settings. Recently, we derived stochastic theorems on network complexity and sparsity holding also for non-uniform probability distributions [21].

Mathematical theory supports and inspires experimental research in neurocomputing. Theory provides some insights to why artificial neural networks can perform high-dimensional tasks efficiently. However, care should be exercised in attempting to extrapolate from artificial neural networks to biological ones. Artificial networks only outperform humans in tasks where they can employ their own strength: in comparison with humans, machines have enormously larger memories and are blindingly fast at calculating. Thus a computer can play (with itself) more games of Chess or even Go (which is significantly harder) in a couple of days than a person could in a lifetime [22]. Thus it is not so surprising that the AlphaGo Zero program achieved victory over the best world Go player. Automatic game playing is a complex, yet artificial man-made task within a limited ambient space with formally defined rules and so has been a natural target for AI. I can imagine that in the future, robots might be capable to learn games like golf -

although it is questionable why someone would invest money into building such robots. Typically, robots are used for tasks which are dangerous or boring for man and neither one is the case of the golf.

Biological neural networks are much more sophisticated than artificial ones. Lessons learned from artificial neural networks are useful but may not be sufficient for understanding brain. Can human brain ever fully understand itself or are there inherent limitations to our knowledge?

Acknowledgements

V.K. acknowledges support from the Czech Grant Foundation grant GA18-23827S and institutional support of the Institute of Computer Science RVO 67985807.

References

[1] Wigner EP. The unreasonable effectiveness of mathematics in the natural sciences. Commun Pure Appl Math 1960;13(1):1–14.
[2] Laughlin SB, Sejnowski TJ. Communication in neural networks. Science 2003;301:1870–4.
[3] LeCunn Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[4] Ball K. An elementary introduction to modern convex geometry. In: Levy S, editor. Flavors of geometry. Cambridge University Press; 1997. p. 1–58.
[5] Bellman R. Dynamic programming. Princeton University Press; 1957.
[6] Kainen PC, Kůrková V, Sanguineti M. Dependence of computational models on input dimension: tractability of approximation and optimization tasks. IEEE Trans Inf Theory 2012;58:1203–14.
[7] Kainen PC. Utilizing geometric anomalies of high dimension: when complexity makes computation easier. In: Warwick K, Kárný M, editors. Computer-intensive methods in control and signal processing: the curse of dimensionality. Boston, MA: Birkhäuser; 1997. p. 283–94.
[8] Donoho D. High-dimensional data analysis: the curses and blessings of dimensionality. In: AMS math challenge lecture; 2000. p. 1–33.
[9] Kainen PC, Kůrková V. Quasiorthogonal dimension of Euclidean spaces. Appl Math Lett 1993;6:7–10.
[10] Lévy P. Problèmes concrets d'analyse fonctionelle. Paris: Gauthier Villards; 1951.
[11] Hoeffding W. Probability inequalities for sums of bounded random variables. J Am Stat Assoc 1963;58:13–30.
[12] Chernoff H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Stat 1952;23:493–507.
[13] Azuma K. Weighted sums of certain dependent random variables. Tohoku Math J 1967;19:357–67.
[14] Dubhashi D, Panconesi A. Concentration of measure for the analysis of randomized algorithms. Cambridge University Press; 2009.
[15] Matoušek J. Lectures on discrete geometry. New York: Springer; 2002.
[16] Cucker F, Smale S. On the mathematical foundations of learning. Bull Am Math Soc 2002;39:1–49.
[17] Vapnik V. The nature of statistical learning theory. New York: Springer; 1997.
[18] Gorban A, Tyukin I. Stochastic separation theorems. Neural Netw 2017;94:255–9.
[19] Gorban A, Tyukin I, Prokhorov D, Sofeikov K. Approximation with random bases: pro et contra. Inf Sci 2016;364–365:129–45.
[20] Rennie J, Shih L, Teevan J, Karger D. Tackling the poor assumptions of naive Bayes classifiers. In: Fawcett T, Mishra N, editors. Proc. 20th int. conf. on machine learning; 2003. p. 616–23.
[21] Kůrková V, Sanguineti M. Probabilistic bounds for binary classification of large data sets. In: Oneto L, Navarin N, Sperduti A, Anguita D, editors. Recent advances in big data and deep learning. Springer; 2019.
[22] Mastering the game of Go without human knowledge. Nature 2017;550(7676):354–9.