Knowledge Transfer Between Artificial Intelligence (AI) systems

Ivan Tyukin, A.N. Gorban

CTS 2017, Saint-Petersburg, 26th October

Content

- AI knowledge Transfer and One-shot learning as Major problems of AI development
- AI Correction: Problem Statement
- Solution: Stochastic Separation Theorems
- Corollary: SKYNET Technology
- Examples
- Conclusion

The 3d Advent of Neural Networks …

1st Advent: **Perceptron** (F. Roseblatt, 1957).

Terminated by (Minsky and Papert, 1969)

2nd Advent:. **Universal Approximation Theorems** (Cybenko, 1989; Barron 1992, Gorban) and Error backpropagation algorithms (Werbos, Gorban).

Stopped naturally (mid 1990s: too expensive computationally)

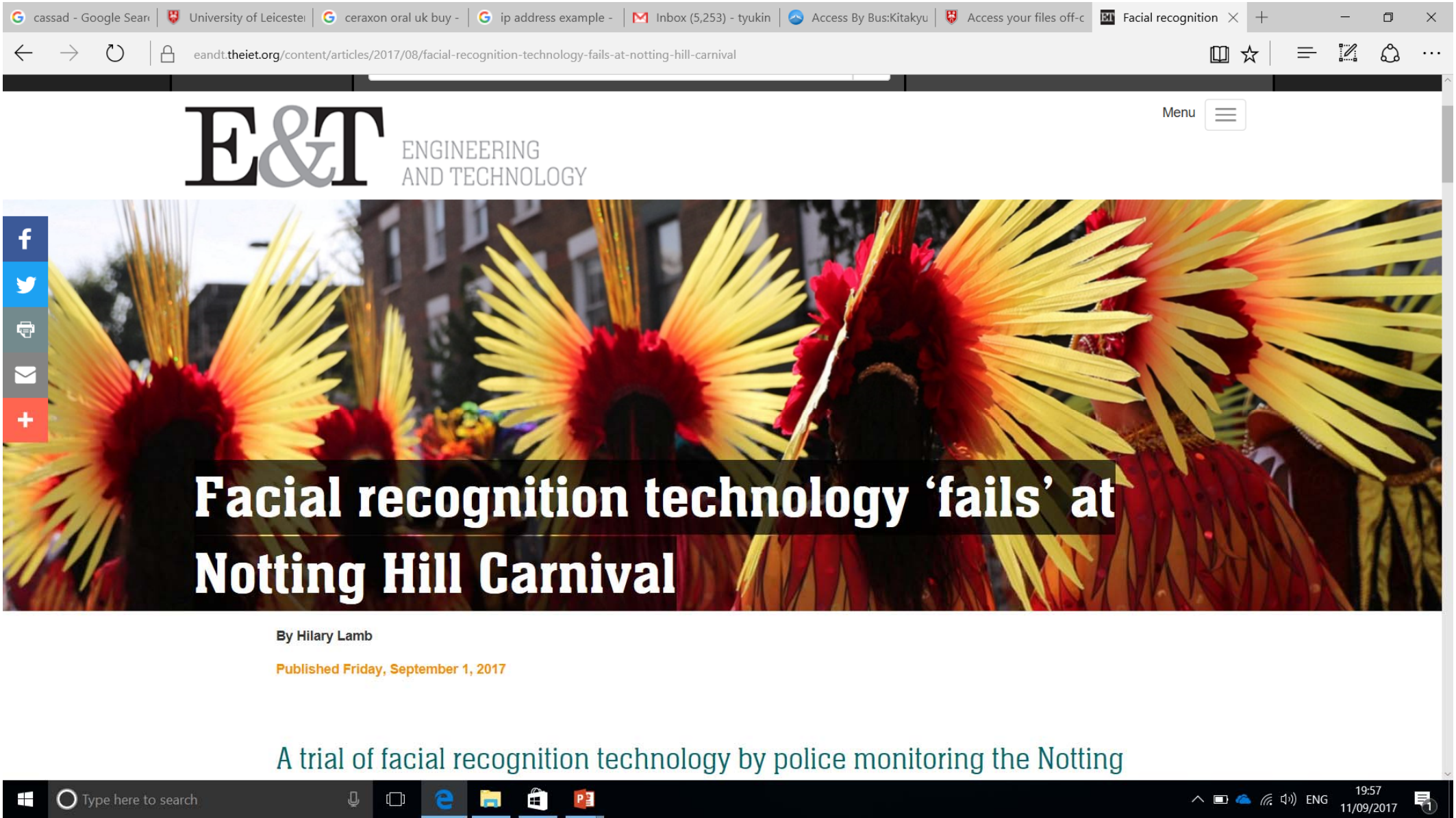
3rd Advent. After 2010: AlexNet and the **Rise of Deep Learning** …. (Until Now)

The World of BIG Legacy AI Systems

- Google (Google Cars, FaceNet, DeepMind)
- Facebook (DeepText, face recognition API)
- Amazon (Recognition, Alexa)
- Apple (Siri, face recognition phone unlock)
- IBM (Watson, face recognition API)
- NEC (face recognition)

Built using massive volumes of data, and have large number of "internal" variables

What stops the AIs: All AI Systems Make mistakes!



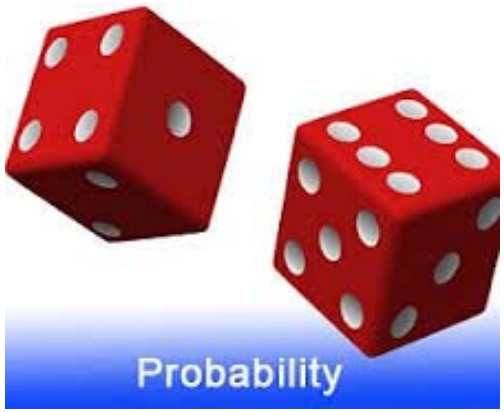
How to Fix Errors and Learn?

Re-training

Random Data Sample

Training

Random Variable



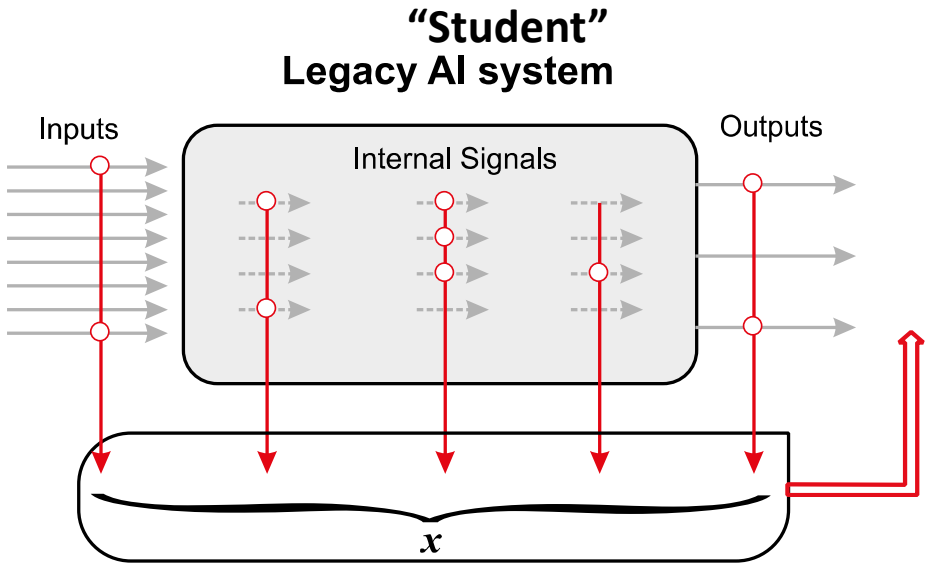
New Mistakes

Does Not Help

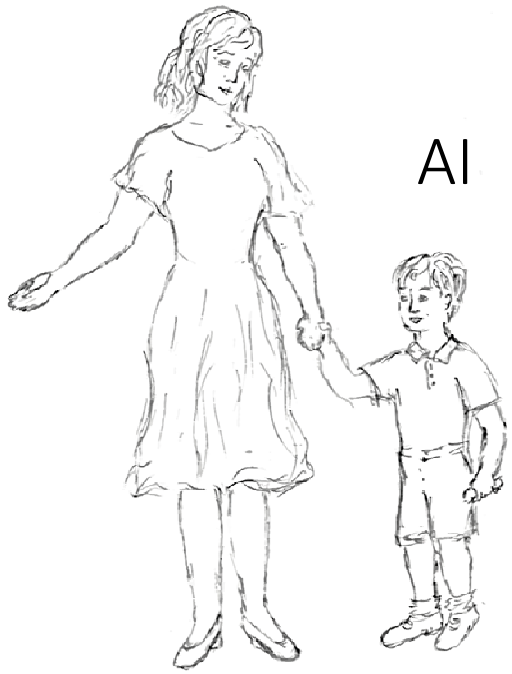
What do we do? Correct Mistakes

With Guaranteed High Probabilities

Setup



Transferred Knowledge



Building Blocks of Correctors (Knowledge Transfer) must

1) be AVAILABLE and reusable in the student AI system;
2) be able to generalize;
3) be efficient computationally;
4) allow for non-iterative learning;

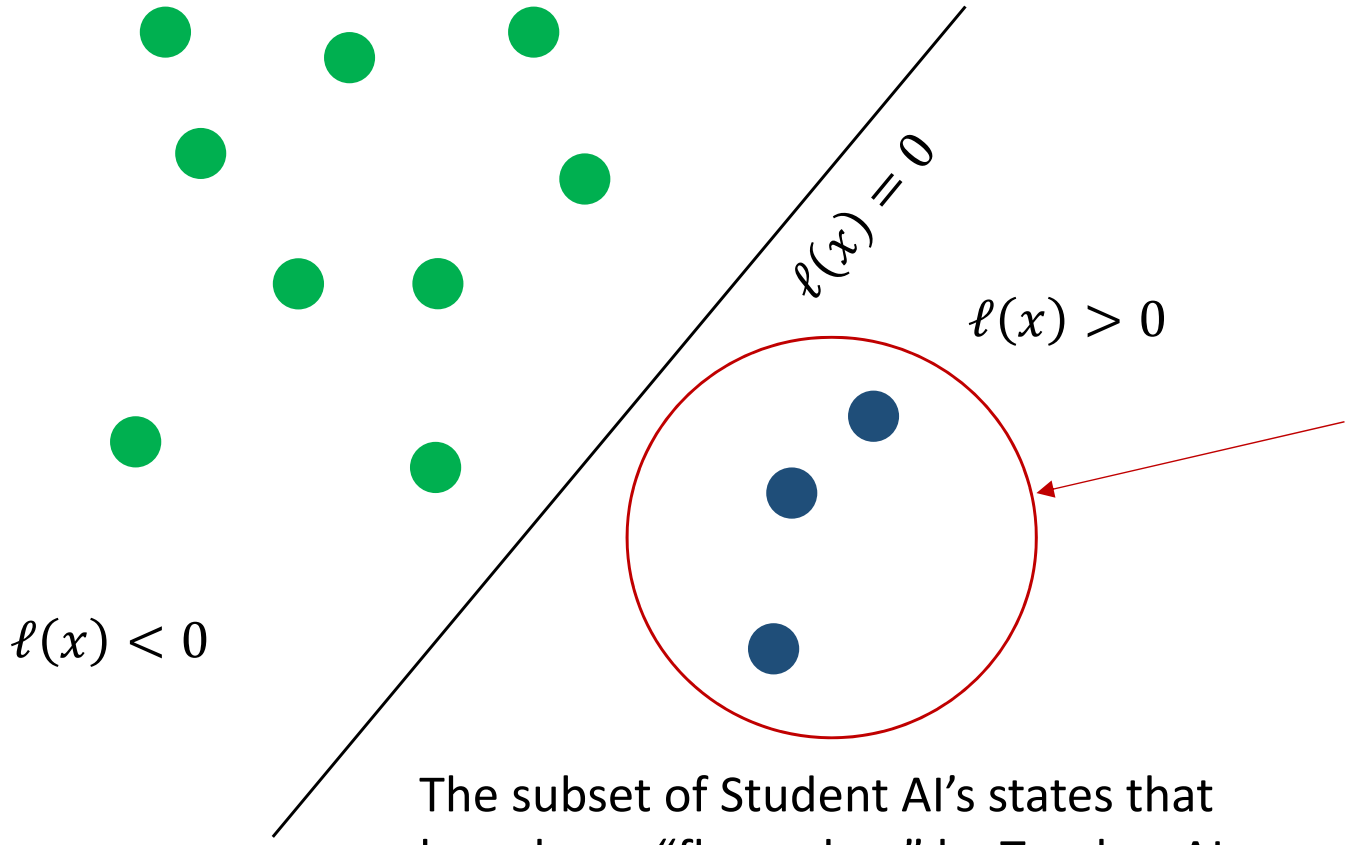
Knowledge Transfer building blocks: linear functionals

$$l(x) = \langle x, w \rangle + b, \;\; w \in \mathbb{R}^n, \;\; b \in \mathbb{R}$$

A simple idea:

Incorporate KT units into Student AI decision-making

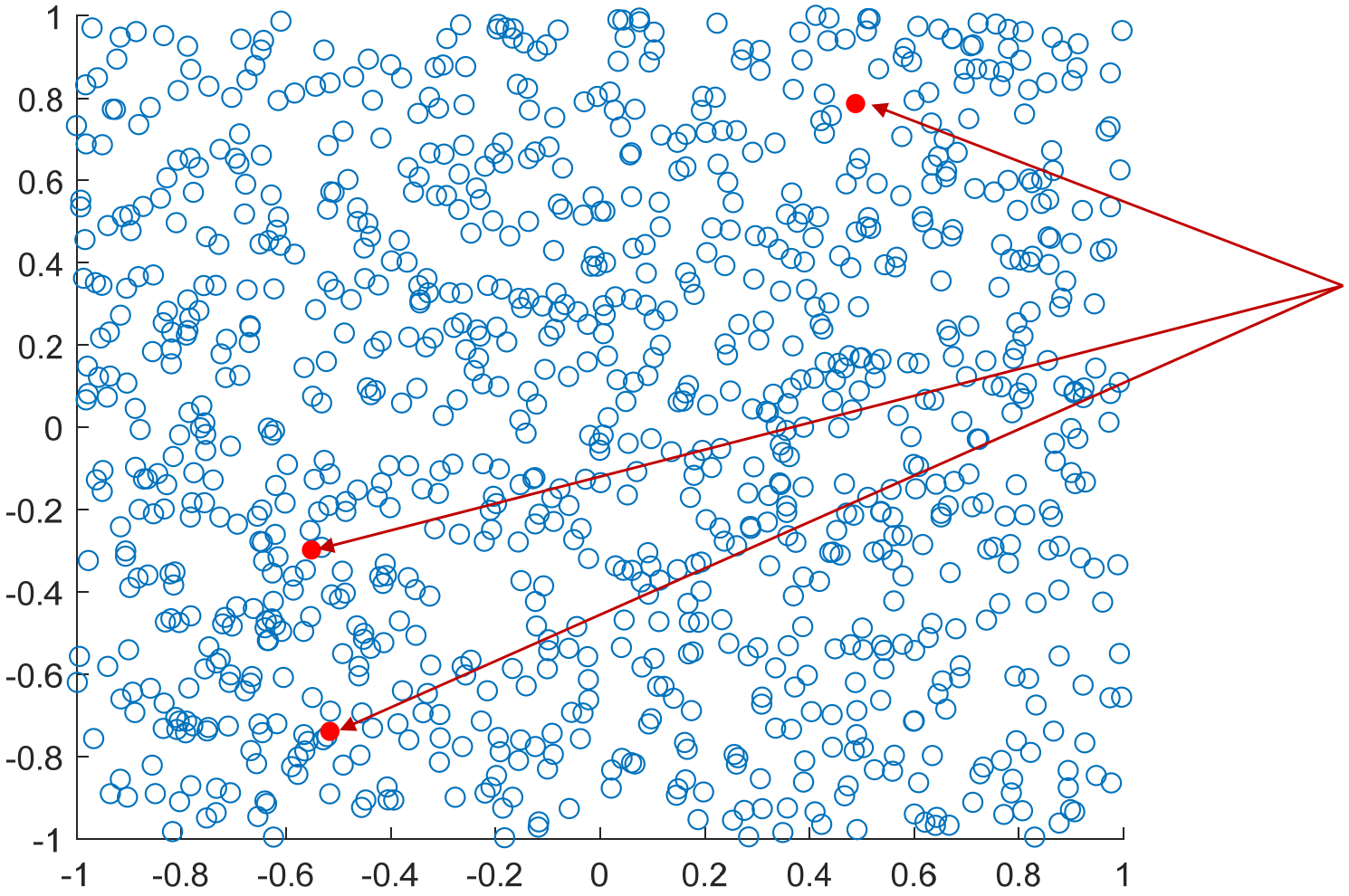
The subset of Student AI's states that have not been "flagged up" by Teacher AI



New knowledge "parcelled" by the AI Knowledge Transfer Unit, $\ell(x)$

The subset of Student AI's states that have been "flagged up" by Teacher AI

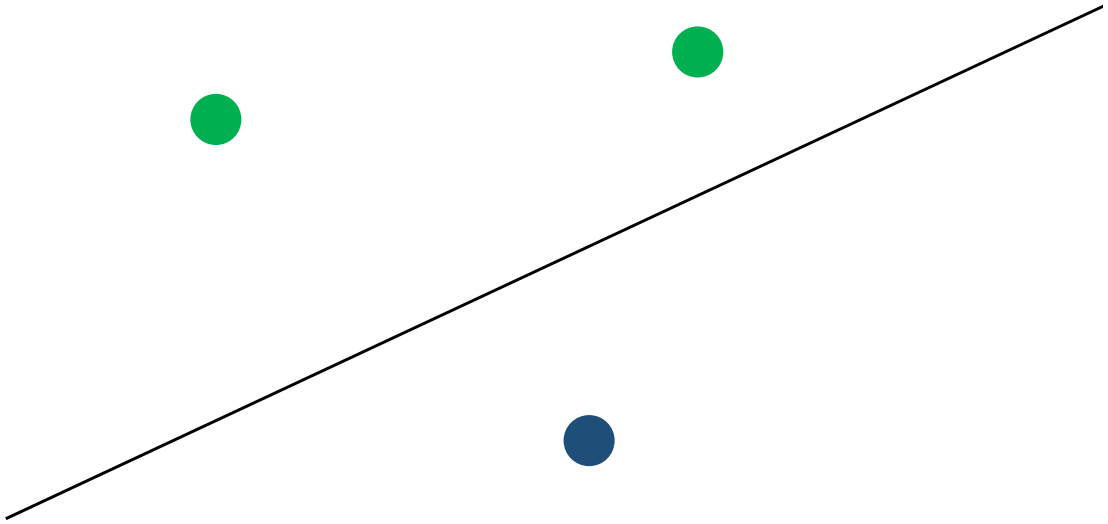
Q1: How is this possible? An example in 2d …

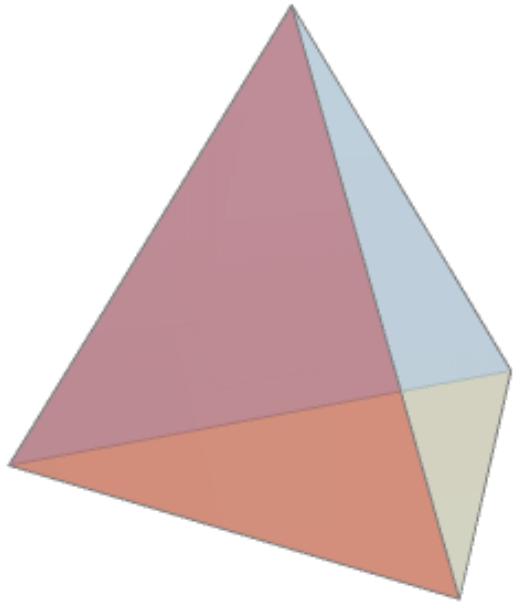


Separate these from the rest by a linear functional ?

We will consider the case when n, the dimension, is large

Q2: Triviality ?



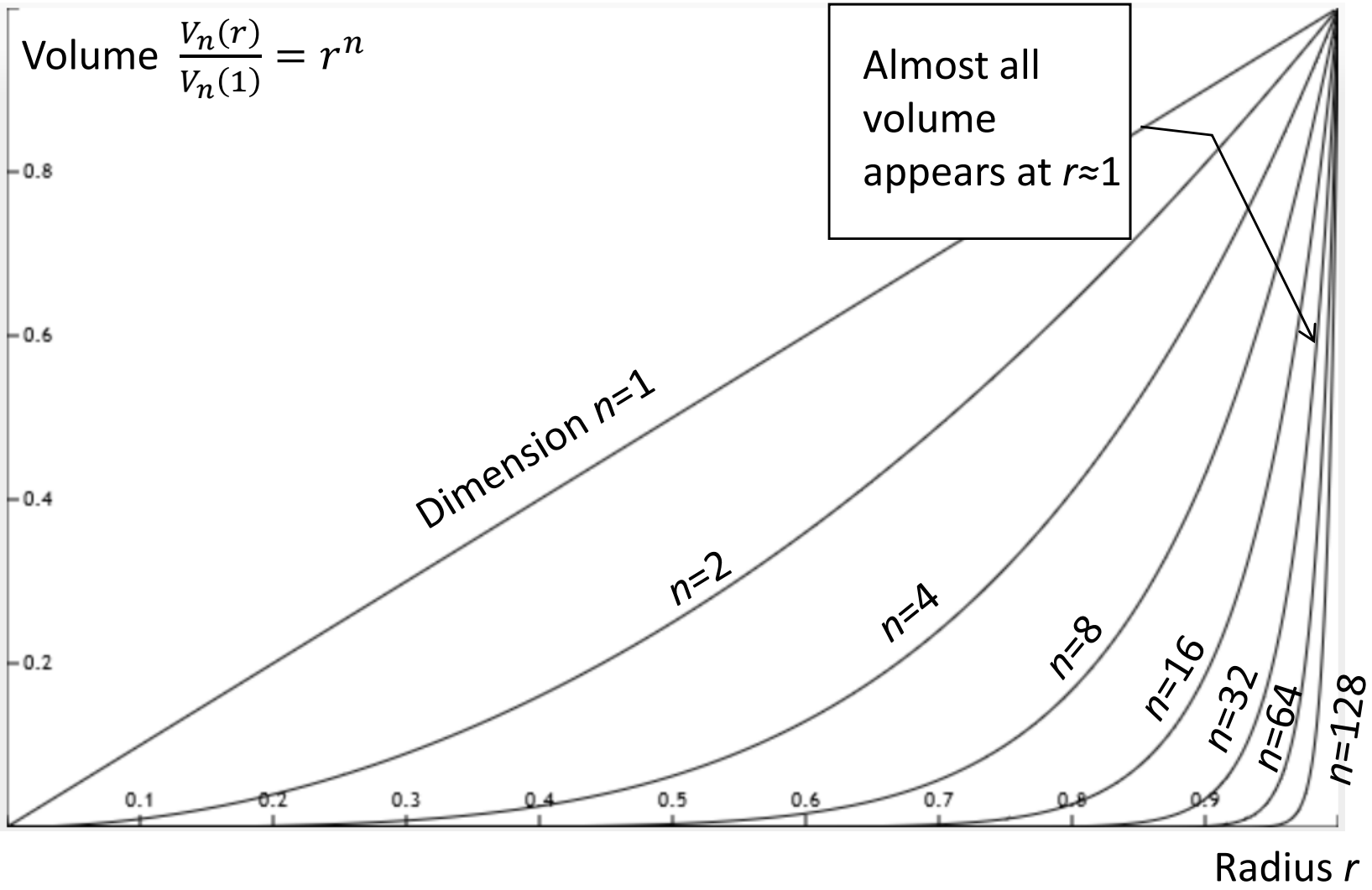


With probability 1, $M = n + 1$ random points are linearly separable in R^n

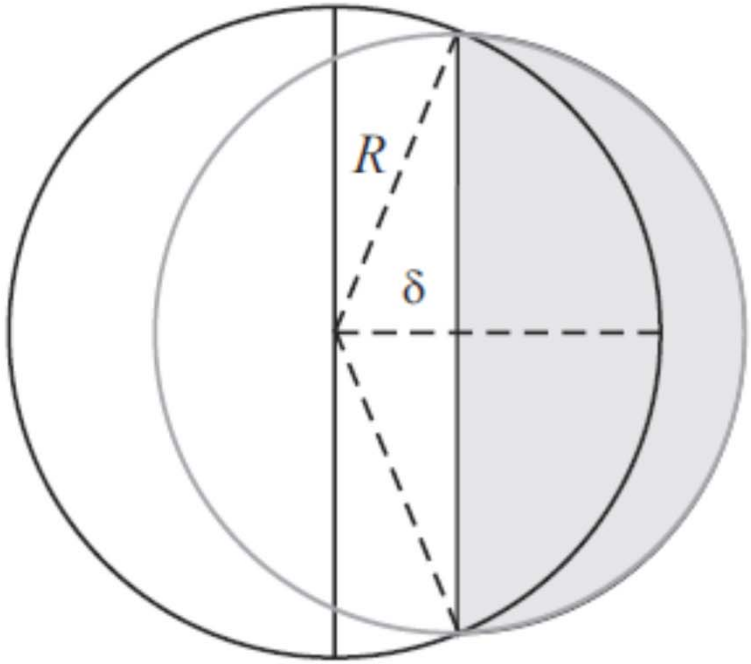
We are interested in the case when $M \gg n$

Geometrical preliminaries:
peculiar properties of high-dimension

Concentration of the volume



Volume of high-dimensional ball
is concentrated near its border (sphere)

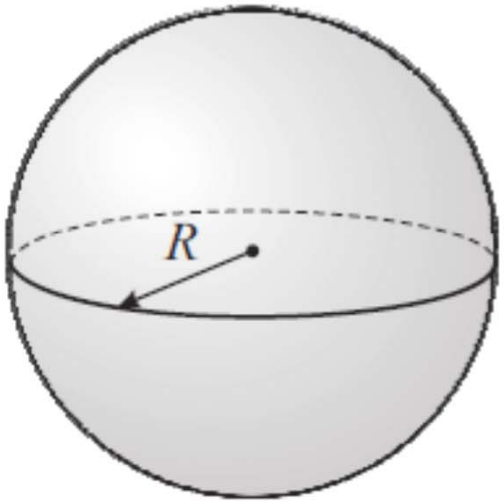


(a) Covering of an n-ball

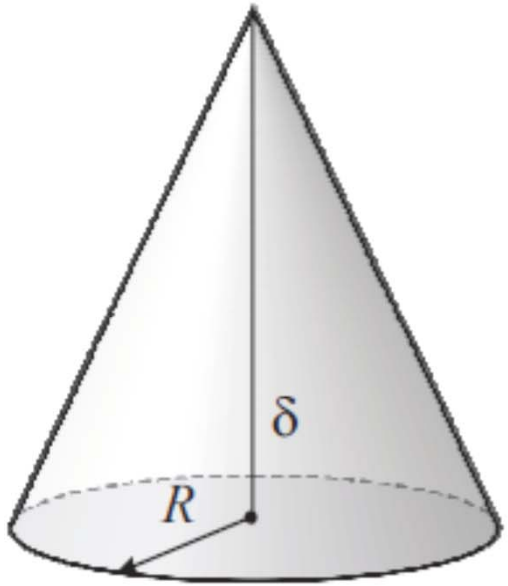
$$\frac{\mathcal{V}(\mathcal{B}^n(R)) - \mathcal{V}(\mathcal{B}^n(\sqrt{R^2 - \delta^2}))}{\mathcal{V}(\mathcal{B}^n(R))}$$

$$1 - \left(1 - \frac{\delta^2}{R^2}\right)^{\frac{n}{2}}$$

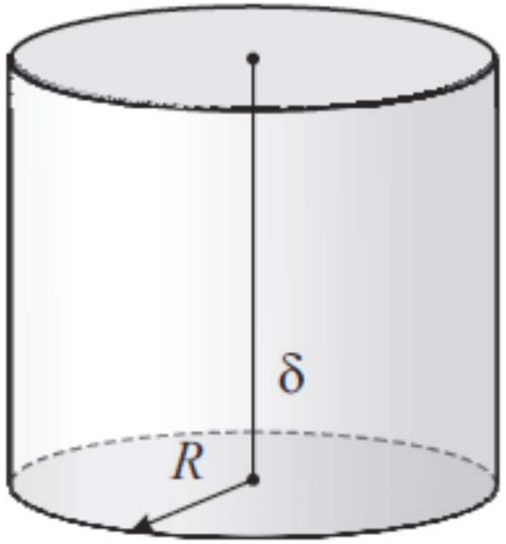
$$1 - \left(1 - \frac{\delta^2}{R^2}\right)^{\frac{n}{2}} > 1 - e^{-\frac{n\delta^2}{2R^2}}$$



(b) $\mathcal{V} \sim \sqrt{\frac{2\pi}{n+1}} R V_{n-1}(R)$

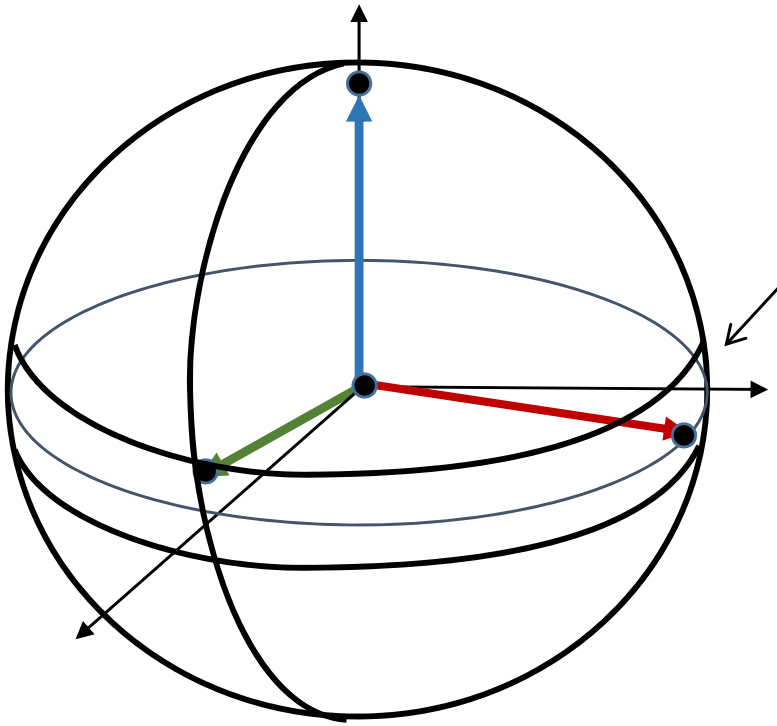


(c) $\mathcal{V} = \frac{1}{n+1} \delta V_{n-1}(R)$



(d) $\mathcal{V} = \delta V_{n-1}(R)$

Waist concentration



In high dimension, the volume of the ball is concentrated near its equator (each equator)

Corollary: *Two random unit vectors are almost orthogonal almost for sure (the angle between them is close to π/2 with high probability).*

Almost orthonormal random bases theorem

Theorem (Gorban, Tyukin, 2016)**.** *Let $h_1,\dots h_N$ be i.i.d. random vectors from a n-dimensional unit sphere (equidistribution), $0<\vartheta<1, 0<\varepsilon<1$. Then for*

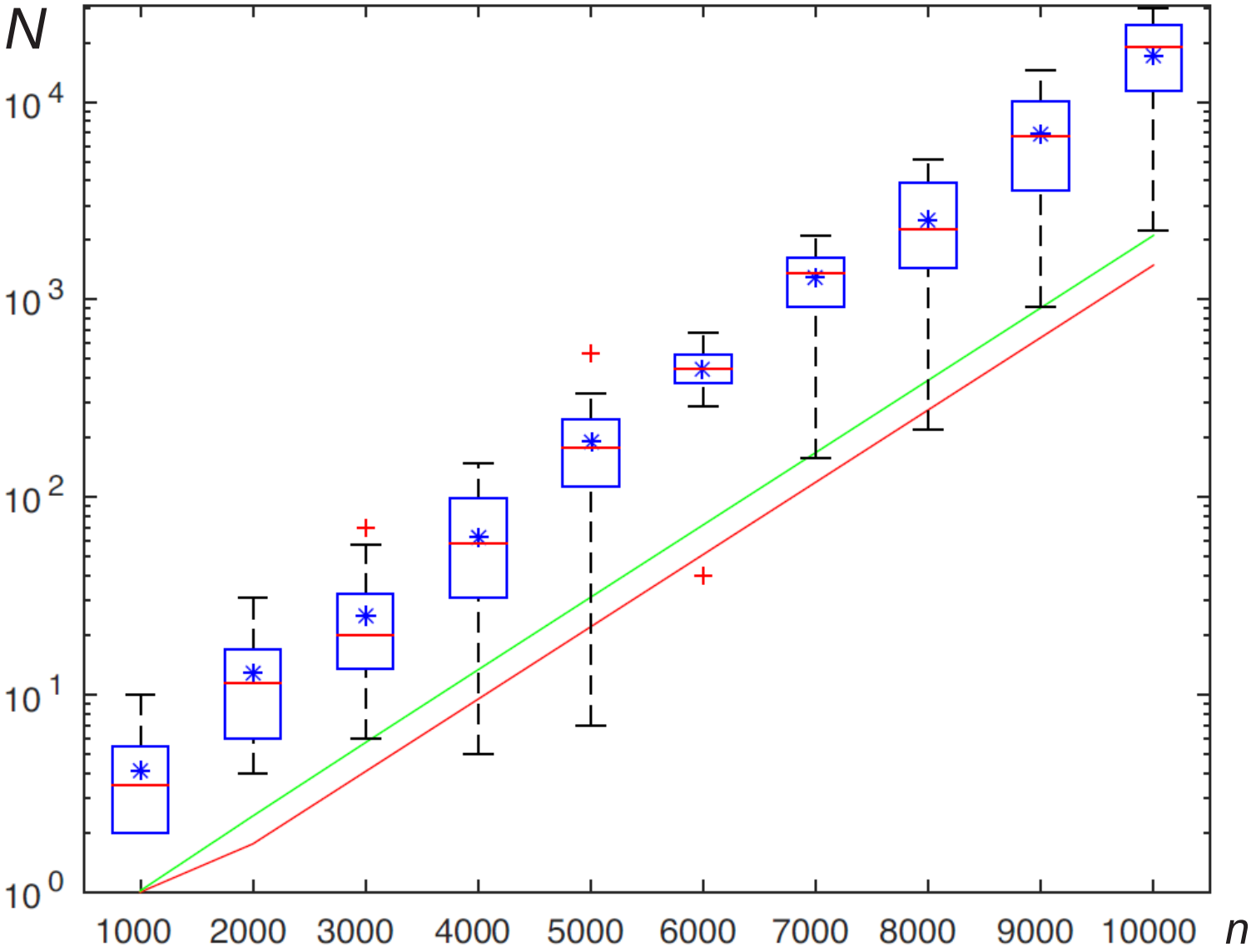
$$N \le e^{\frac{\epsilon^2 n}{4}} \left[\log\left(\frac{1}{1-\vartheta}\right)\right]^{\frac{1}{2}}$$

with probability $P>1-\vartheta$ all vectors h_i are pairwise ε-orthogonal: $|\langle h_i, h_j\rangle| < \varepsilon$ for $i\neq j$.

Refined estimate

$$N \le \sqrt{\frac{\log^2(1-\vartheta)}{4} + 2\log\frac{1}{1-\vartheta} e^{\frac{n\epsilon^2}{2}}} + \frac{\log(1-\vartheta)}{2}$$

Numerical evaluation/example



Lengths N of pairwise almost orthogonal chains of vectors that are independently randomly sampled from $[-1,1]^n$ as a function of dimension n.
Red line – the basic estimate, green line – the refined estimate.
$\varepsilon=0.06$, $\theta=0.1$.

Mathematical foundations for
AI knowledge transfer and one-shot Learning

Results

For n large enough, with probability close to one,

1) all points in large but finite i.i.d. samples are *linearly separable* from the rest …

2) (sufficiently small) subsets of large but finite i.i.d. samples are *linearly separable* from the rest …

Separation of two random sets.

for n large enough, with probability close to 1, β_2 is bounded from below by a small number **(Theorem 1)**

Special case: positive correlation in Y

Theorem 2 (Tyukin, Gorban, Sofeikov, Romanenko 2017). *Consider an equidistribution in a unit ball* $B_n(1)$ *in* R^n, *and let* M *and* Y *be i.i.d. samples (defined above) from this distribution. Let* $Y_c = \{x_{M+r_1}, \dots, x_{M+r_m}\}$ *be a subset of* Y *such that*

$$\beta_2(m-1) \le \sum_{r_j, r_j \neq r_i} \langle x_{r_i}, x_{r_j} \rangle \le \beta_1(m-1) \quad \forall\, i = 1, \dots, m.$$

Then

$$P_1(M, Y_c) \ge \max_{\varepsilon \in (0,1)} (1-(1-\varepsilon)^n)^m \left(1 - \frac{\Delta(\varepsilon, m)^n}{2}\right)^M$$

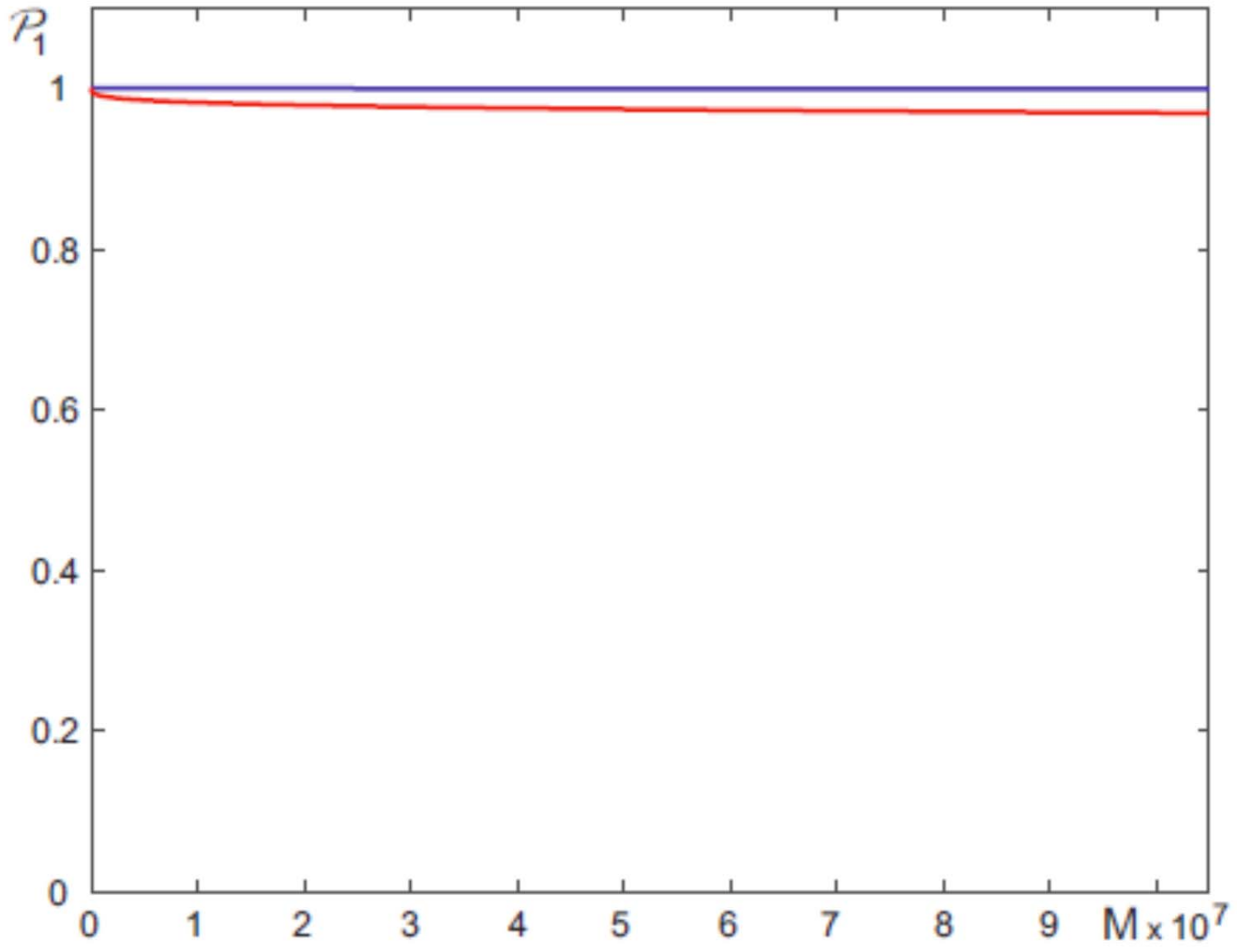
$$\Delta(\varepsilon, m) = \left(1 - \frac{1}{m}\left(\frac{(1-\varepsilon)^2 + \beta_2(m-1)}{\sqrt{1+(m-1)\beta_1}}\right)^2\right)^{\frac{1}{2}}$$

s.t. $(1-\varepsilon)^2 + \beta_2(m-1) > 0,\ 1 + (m-1)\beta_1 > 0.$

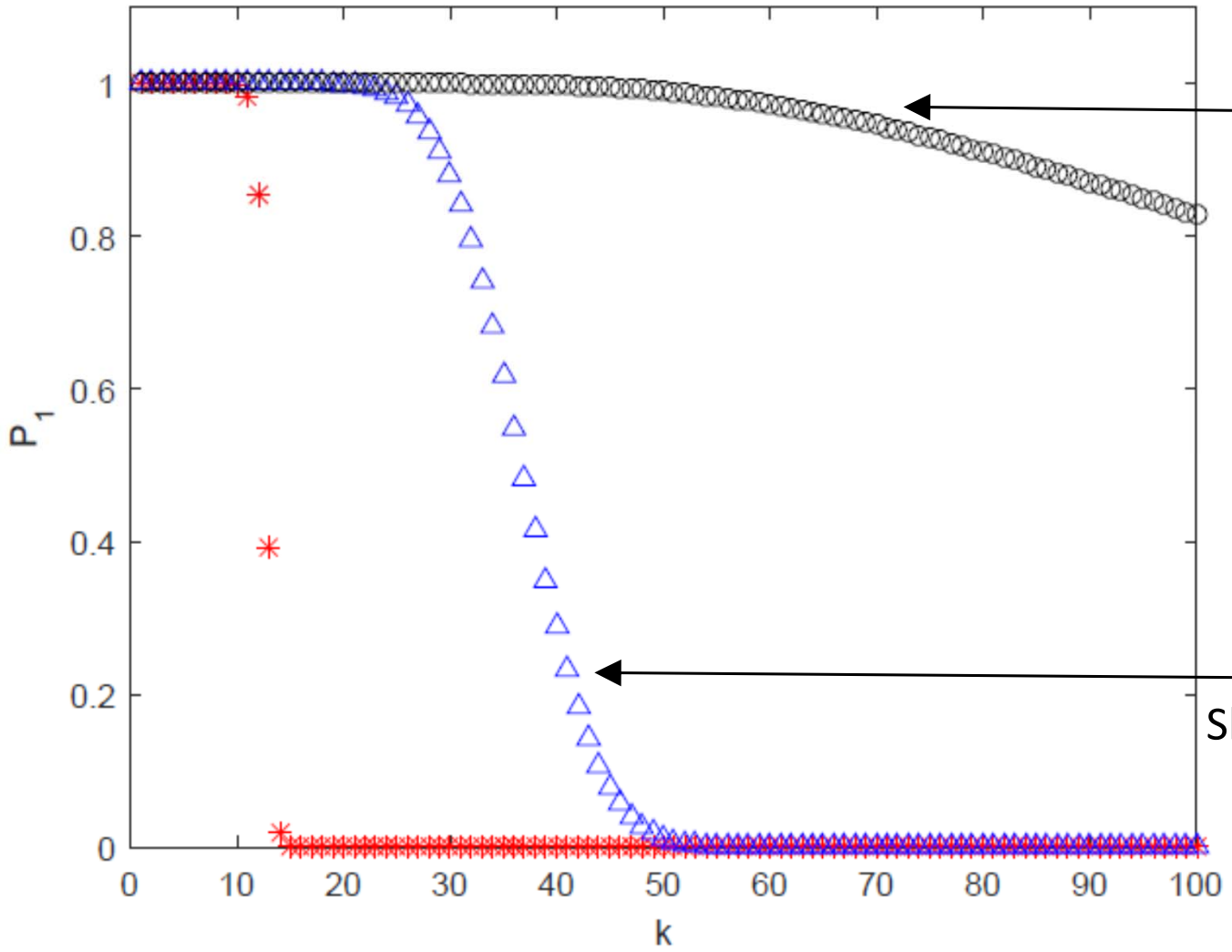
How large is the probability of separating one point?

Red line – estimate (from below) of the probability of linear separation

Dimension n=30, **m=1 (single-point separation)**



How does k-tuple separability depend on k (or m) ?

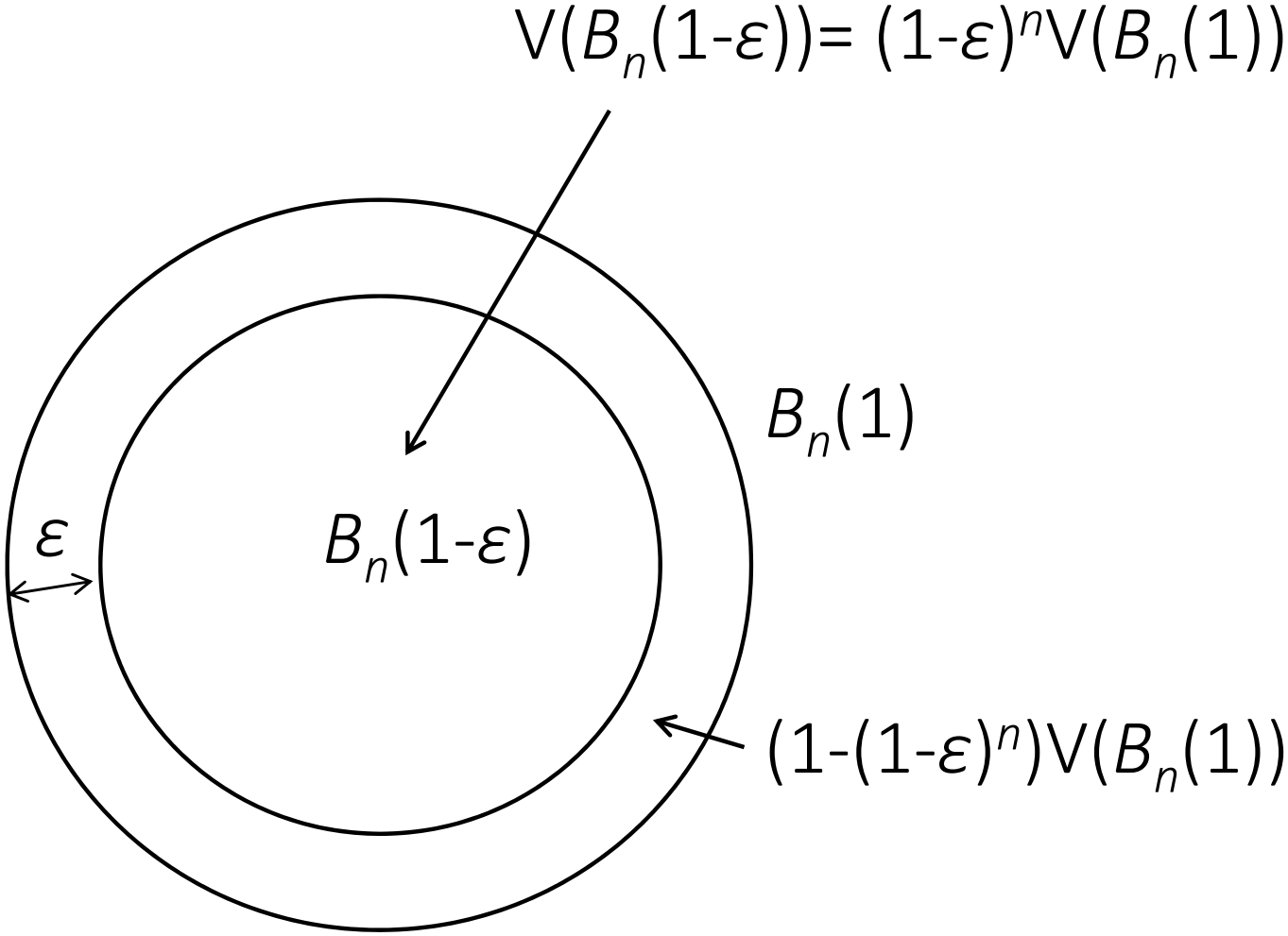


More correlated
$\beta_1 = 0.5, \beta_2 = 0.07$

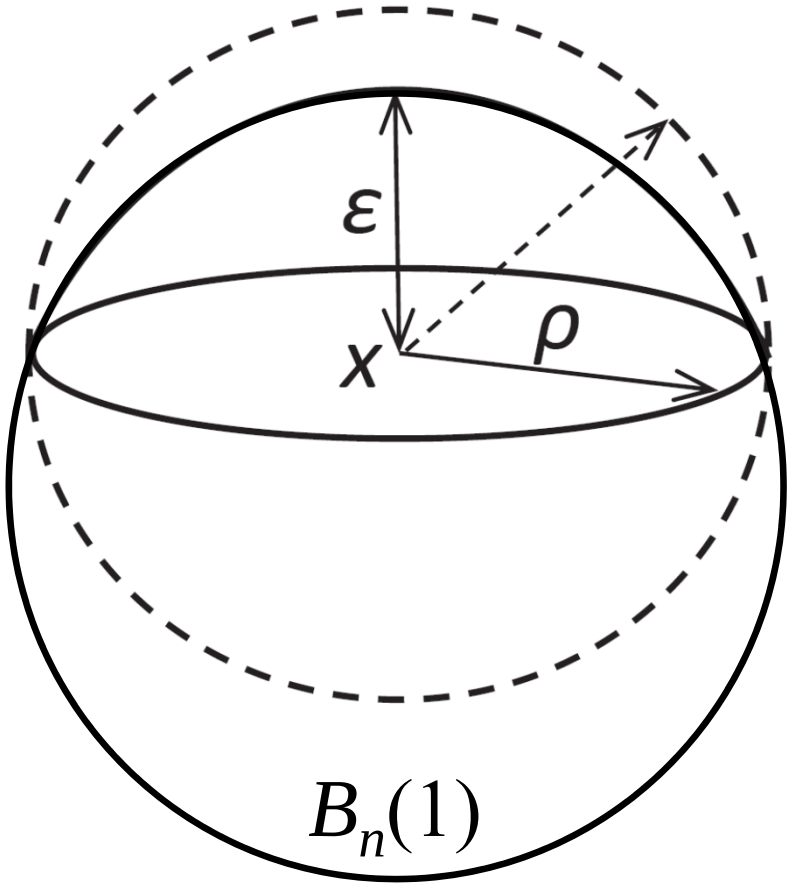
n = 2000
M=100 000

Slightly positively correlated
$\beta_1 = 0.5, \beta_2 = 0.05$

Proof (m=1)


$V(B_n(1-\varepsilon))= (1-\varepsilon)^n V(B_n(1))$
$B_n(1)$
ε
$B_n(1-\varepsilon)$
$(1-(1-\varepsilon)^n)V(B_n(1))$

Proof (m=1)

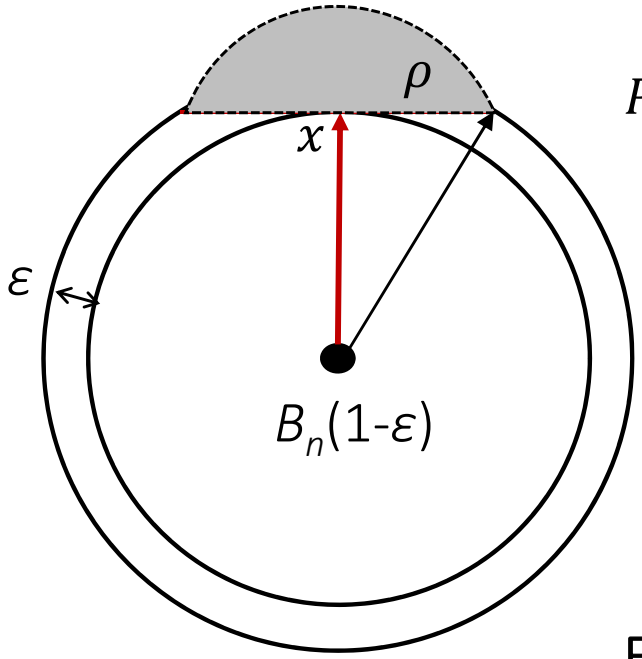


A test point *x* and a spherical cap on distance ε from the surface of unit sphere. Escribed sphere of radius $\rho(\varepsilon)$ is showed by dashed line. The volume of this escribed sphere is much less than the volume of $B_n(1)$. Their ratio is $(\rho(\varepsilon))^n$.

$$\rho(\varepsilon) = (1 - (1 - \varepsilon)^2)^{\frac{1}{2}}$$

Proof (m=1)

The probability that a random point belongs to a ε-layer near the border sphere of $B_n(1)$.



$$P_1(M, Y_c) \geq \max_{\varepsilon \in (0,1)} (1 - (1-\varepsilon)^n) \left(1 - \frac{\rho(\varepsilon)^n}{2}\right)^M$$

The probability that M random points from $B_n(1)$ are outside a hemisphere (half-ball) of radius $\rho(\varepsilon)$

□

For m>1 the separating functional is:

$$l_0(\boldsymbol{x}) = \left\langle \frac{\bar{\boldsymbol{y}}}{\|\bar{\boldsymbol{y}}\|}, \boldsymbol{x} \right\rangle - \frac{1}{\sqrt{m}} \left(\frac{(1-\varepsilon)^2 + \beta_2(m-1)}{\sqrt{1+(m-1)\beta_1}} \right) \qquad \bar{\boldsymbol{y}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_{M+r_i}$$

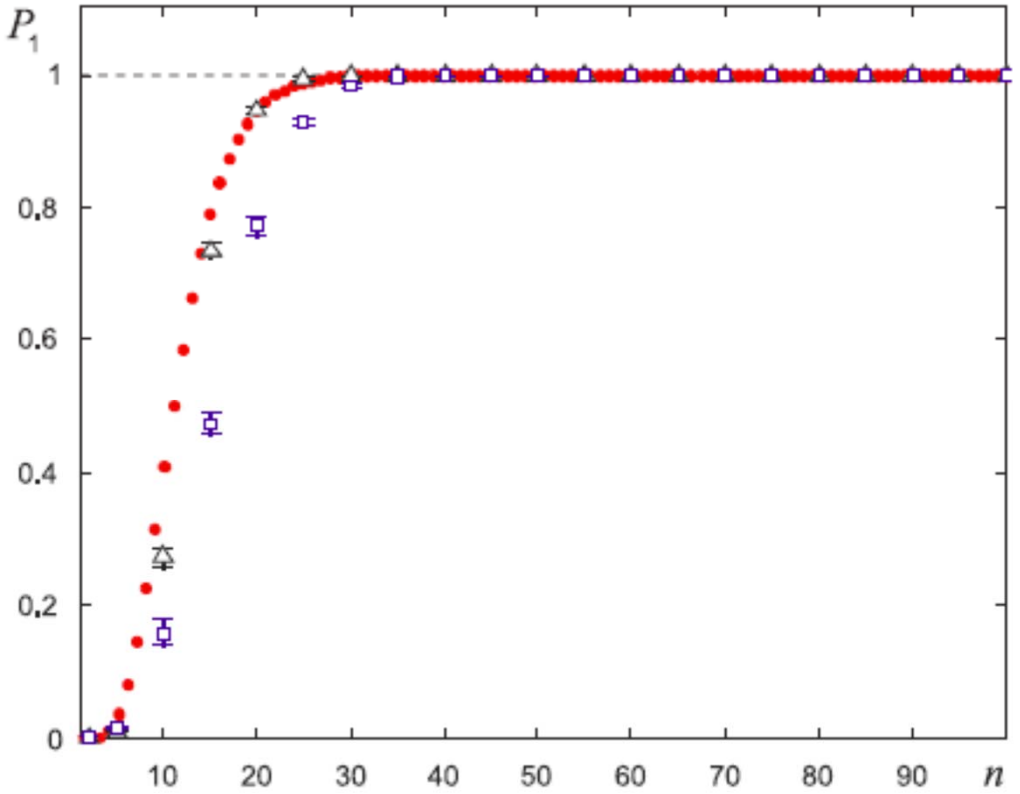
The results can be generalized to product measures (cube, Gaussian)

A.N. Gorban, I. Y. Tyukin. Stochastic Separation Theorems. *Neural Networks 94*, 255-259, 2017. Preprint: https://arxiv.org/abs/1703.01203

Hoeffding's Inequality, 1963 (for PM with compact support)

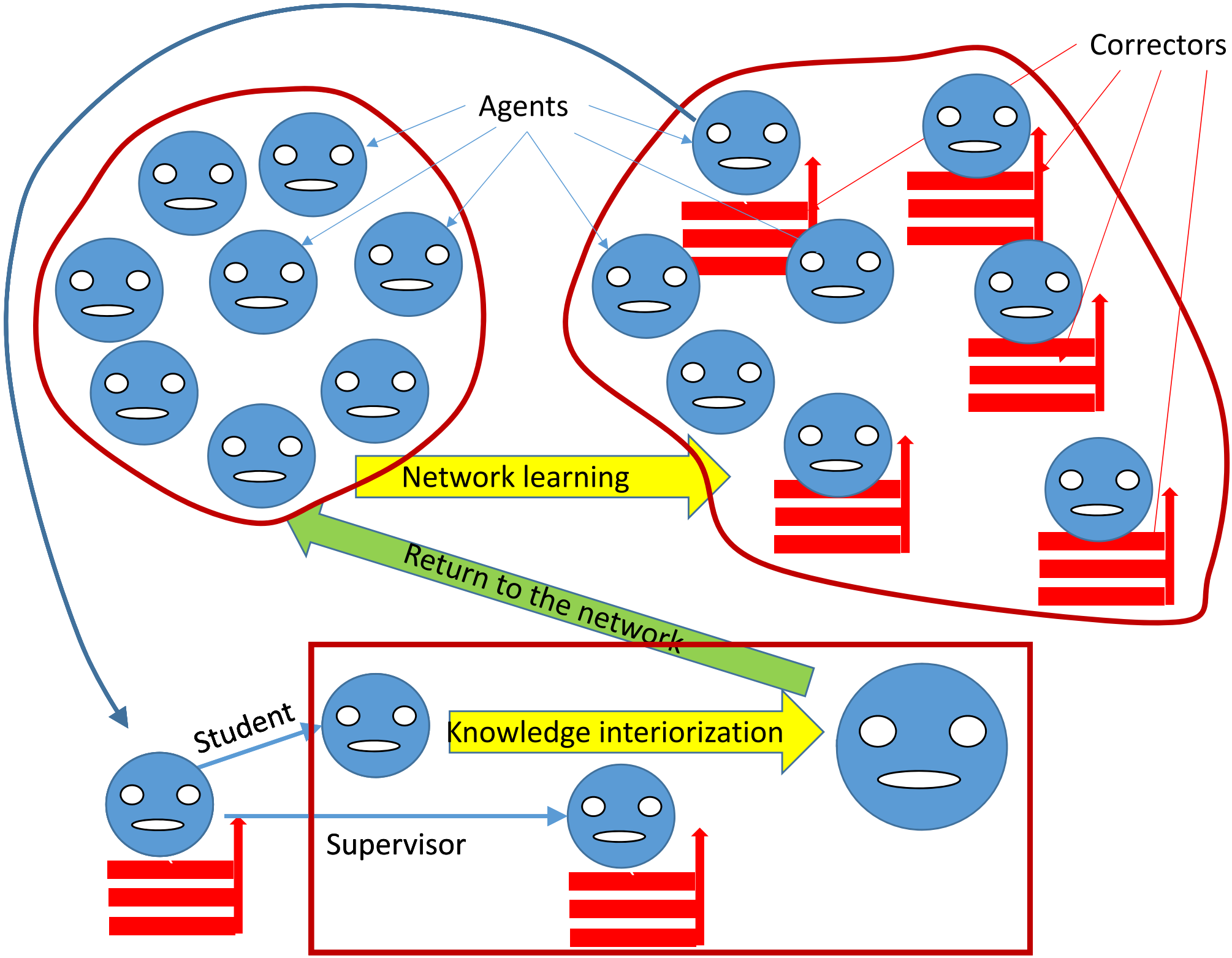
Other distributions

Dimension, n	2	5	10	20	30
Ball (theoretical estimate)	$3.7 \cdot 10^{-5}$	0.0364	0.4096	0.9455	0.9975
Cube					
min:	$4 \cdot 10^{-4}$	0.0089	0.2580	0.9408	0.9986
median:	$4 \cdot 10^{-4}$	0.0110	0.2737	0.9469	0.9992
max:	$5 \cdot 10^{-4}$	0.0137	0.2847	0.9511	1.0
Gaussian					
min:	$5 \cdot 10^{-4}$	0.0122	0.1403	0.7559	0.9792
median:	$10 \cdot 10^{-4}$	0.0153	0.1568	0.7698	0.9817
max:	0.0014	0.0183	0.1778	0.7836	0.9848



Next steps:

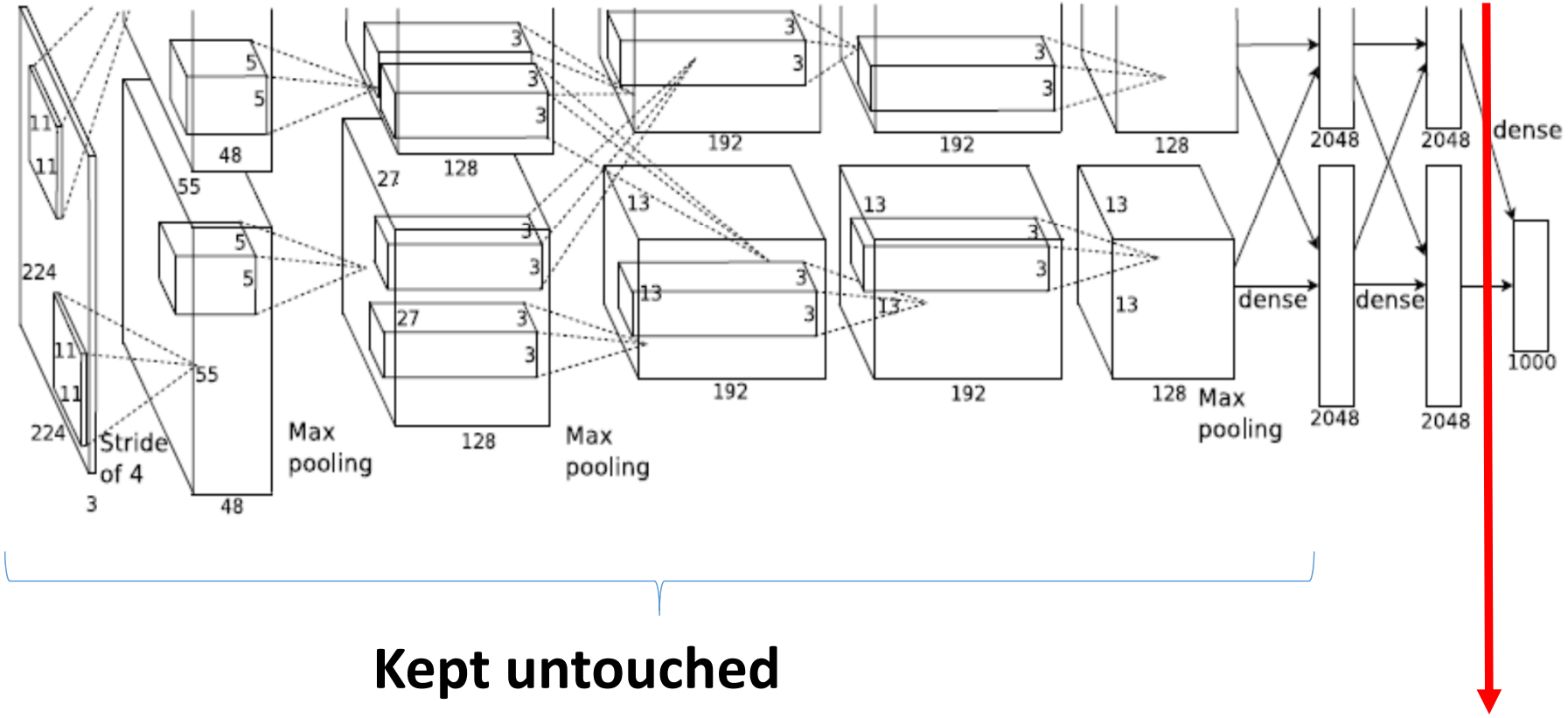
These Theorems enable
“SKY NET” Technology


Agents
Correctors
Network learning
Return to the network
Student
Knowledge interiorization
Supervisor

Examples

Pedestrian detection

KT for Convolutional Nets (Human to CNN)



Kept untouched

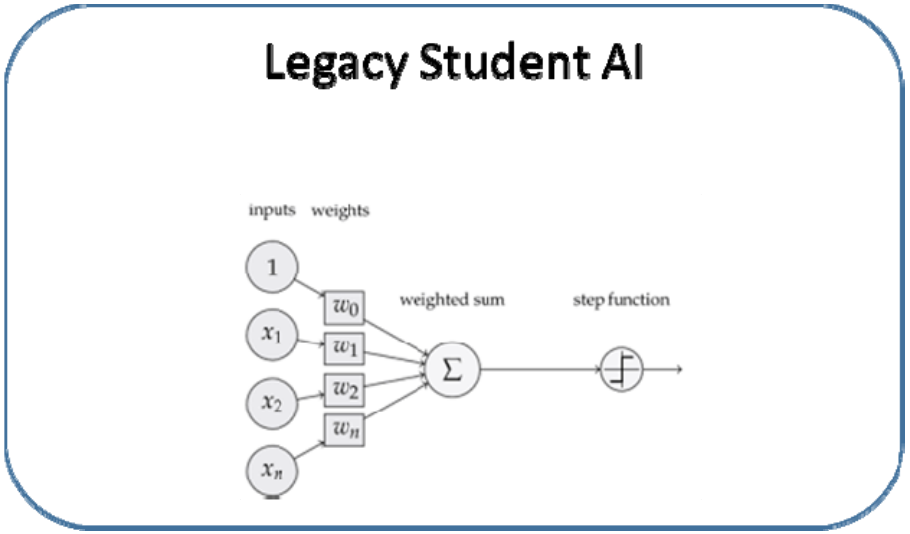
Internal representations $x_1, x_2, \ldots, x_M$

AI Knowledge transfer Unit

$$l(x) = \langle x, w \rangle + b,\ w \in \mathbb{R}^n,\ b \in \mathbb{R}$$

SPORTS
#student
SPORTS
#student

One-shot learning HOG-SVM (Theorem 2, m=1!)



Conclusion

- In AI systems whose state can be represented by elements of topological vector space with relatively high dimensions, errors or new knowledge, can be ``parcelled'' (packaged) by linear functionals and ``attached'' to an existing Legacy AI system → **AI Knowledge Transfer**

- This AI Knowledge Transfer is completely reversible; It is non-iterative and fast (and hence one-shot Learning)

- In the core of these results are various measure concentration phenomena

- This gives rise to new filed and phenomena: "**Statistical Physics of Big Data and Machine learning in High Dimensions**"

- This is very practical!

References

- J. W. Gibbs, Elementary Principles in Statistical Mechanics, developed with especial reference to the rational foundation of thermodynamics, (New York: Dover Publications, 1960 [1902])
- Lévy P, Problèmes concrets d'analyse fonctionnelle. Paris: Gauthier-Villars; 1951. (Second edition.)
- F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, 1962
- M. Gromov, Isoperimetry of waists and concentration of maps, GAFA, Geom. Funct. Anal. 13 (2003) 178-215
- P.C.Kainen, V.Kurkova, Quasiorthogonal dimension of Euclidian spaces, Appl. Math. Lett. 6(3) (1993) 7–10.
- S.Artstein, Proportional concentration phenomena of the sphere, Israel J. Math. 132 (2002) 337–358.

Further reading

- I.Y. Tyukin, A.N. Gorban, K. Sofeikov, I. Romanenko. Knowledge Transfer Between Artificial Intelligence Systems. https://arxiv.org/abs/1709.01547

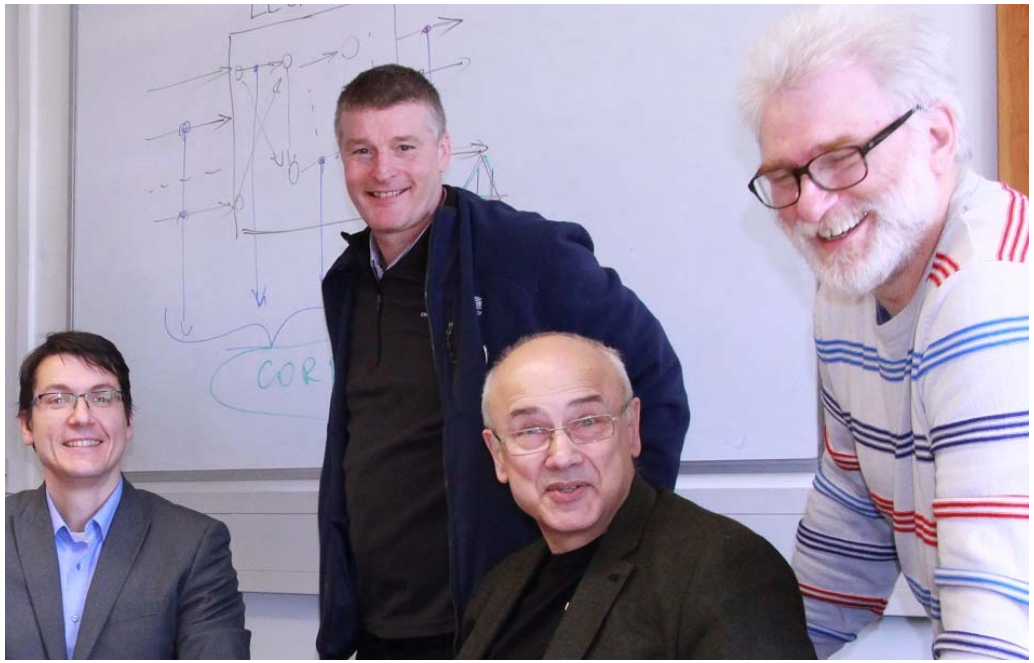
- A.N. Gorban, I.Y. Tyukin, D.V. Prokhorov, K.I. Sofeikov, Approximation with random bases: Pro et Contra, *Information Sciences* 364–365, 129–145 (2016),

- A.N. Gorban, I. Y. Tyukin. Stochastic Separation Theorems. *Neural Networks 94*, 255-259, 2017.

- A.N. Gorban, R. Burton, I. Romanenko, I.Y. Tyukin. One-trial Correction of Legacy AI Systems and Stochastic Separation Theorems. https://arxiv.org/abs/1610.00494

Co-authors and collaborators

- Alexander N. Gorban
- Ilya Romanenko
- Konstantin Sofeikov
- Richard Burton
- Danil Prokhorov
- Jeremy Levesley
- Evgeny Mirkes



Spinout

UnivAI Ltd

www.univai.co