Accepted Manuscript

Stochastic separation theorems

A.N. Gorban, I.Y. Tyukin

PII: S0893-6080(17)30177-6
DOI: http://dx.doi.org/10.1016/j.neunet.2017.07.014
Reference: NN 3795

To appear in: *Neural Networks*

Received date : 1 April 2017
Revised date : 16 July 2017
Accepted date : 21 July 2017

Please cite this article as: Gorban, A.N., Tyukin, I.Y., Stochastic separation theorems. *Neural Networks* (2017), http://dx.doi.org/10.1016/j.neunet.2017.07.014

Stochastic Separation Theorems

A.N. Gorban[a,*], I.Y. Tyukin[a]

[a]Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK

Abstract

The problem of non-iterative one-shot and non-destructive correction of unavoidable mistakes arises in all Artificiall Intelligence applications in the real world. Its solution requires robust separation of samples with errors from samples where the system works properly. We demonstrate that in (moderately) high dimension this separation could be achieved with probability close to one by linear discriminants. Based on fundamental properties of measure concentration, we show that for $M < a\exp(bn)$ random M-element sets in $\mathbb{R}^n$ are linearly separable with probability p, $p > 1-\vartheta$, where $1 > \vartheta > 0$ is a given small constant. Exact values of $a, b > 0$ depend on the probability distribution that determines how the random M-element sets are drawn, and on the constant ϑ. These *stochastic separation theorems* provide a new instrument for the development, analysis, and assessment of machine learning methods and algorithms in high dimension. Theoretical statements are illustrated with numerical examples.

Keywords: Fisher's discriminant, random set, measure concentration, linear separability, machine learning, extreme point

1. Introduction

Artificial Intelligence (AI) systems make errors. They should be corrected without damage of existing skills. The *problem of non-destructive correction* arises in many areas of research and development, from AI to mathematical neuroscience, where the reverse engineering of the brain ability to learn on-the-fly remains a great challenge. It is very desirable that the corrector of errors is *non-iterative* (one-shot) because iterative re-training of a large system requires much time and resource and cannot be done immediately without impeding activity.

The non-desrructive correction requires separation of the situations (samples) with errors from the samples corresponding to correct behavior by a simple and robust classifier. Linear discriminants introduced by Fisher (1936) are simple, robust, require just the inverse covariance matrix of data, and may be easily modified for assimilation of new data. Rosenblatt (1962) revived the common interest in linear classifiers. His works sparked intensive scientific debate (Minsky and Papert, 1969) and gave rise to development of numerous crucial concepts such as e.g. Vapnik-Chervonenkis theory (Vapnik and Chervonenkis, 1971), learnability (Natarajan, 1989), and generalization capabilities of neural networks (Vapnik, 2000), (Bousquet and Elisseeff, 2002). Linear functionals (adaptive summators) are basic building blocks of significantly more sophisticated AI systems such as e.g. multilayer perceptrons, (Rumelhart et al., 1986), Convolutional Neural Networks (Le Cun and Bengio, 1995), (LeCun et al., 2015) and their derivatives. Much is known about linear functionals as "stand-alone" learning machines, including their generalization margins (Freund and Schapire, 1999), (Vapnik, 2000) and numerous methods for their construction: linear discriminants and regression, perceptron learning, and Support Vector Machines (Vapnik, 1982) among others.

In this work, we demonstrate that in high dimensions and even for exponentially large samples, linear classifiers in their classical Fisher's form are powerful enough to separate errors from correct responses with high probability and to provide efficient solution to the non-destructive corrector problem. We prove that linear functionals, as learning machines, have surprising and, as far as we are concerned, new peculiar extremal properties: *in high dimension, with probability $p > 1-\vartheta$ and for $M < a\exp(bn)$ with $a, b > 0$ every point in random i.i.d. drawn M-element sets in $\mathbb{R}^n$ is linearly separable from the rest.* Moreover, the separating linear functional can be found explicitly, without iterations. This property holds for a broad set of relevant distributions, including products of probability measures with bounded support and equidistribution in a unit ball, providing mathematical foundations for one-trial correction of legacy AI systems (cf. (Gorban et al., 2016a)).

A problem of data fusion in multiagent systems has clear similarity to the problem of non-destructive correction. According to Forney et al. (2017), data collected by different agents may not be naively combined due to changes in the context, and special procedures for their assimilation without damage of gained skills are needed. The proven stochastic separation effects can be used to approach this problem. They also shed light on the possible origins of remarkable selectivity to stimuli observed in-vivo in the real brain (Quian Quiroga et al., 2005).

2. Preliminaries

2.1. Notation

Throughout the text, $\mathbb{R}^n$ is the n-dimensional linear real vector space. Unless stated otherwise, symbols $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,n})$

*Corresponding author
Email addresses: ag153@le.ac.uk (A.N. Gorban), it37@le.ac.uk (I.Y. Tyukin)