Probabilistic learning model PAC - lecture notes

Hakl, František
2015

Institute of Computer Science
The Czech Academy of Sciences

Probabilistic learning model PAC — lecture notes

František Hakl

Technical report No. 1227

December 2015

Abstrakt:

This report sumarises basic theoretical properties of Valiant's distribution-independent learnability PAC model which determines a probabilistic framework of supervised learning. The form of this text is self-contained and provides a detailed proves of the concerned theorems and lemmas. Essentials of VC-dimension of class sets are also included. The text presented is aimed as a study material for MSc. and PhD. students such as for computer science engineers.

Keywords:
Supervised learning, PAC model, Vapnik-Chervonenkis dimension, learning algorithms

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Probabilistic learning models

František Hakl

lecture notes
Faculty of Nuclear Science and Physical Engineering
Czech Technical University, Prague
December 2015

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

This research was supported by the Czech Science Foundation, Contract №15-18108S .

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Contents

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

List of Figures

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Symbols and notation

The following fonts and symbols, everywhere in the text of this booklet have the meaning described bellow:

$\stackrel{\text{def}}{=}$	definition of newly introducec object, set or number
$\stackrel{2}{\Rightarrow}$	implication from equation, etc.
$\stackrel{4}{=}$	equality follows from 4
$\stackrel{5}{\neq}$	nonequality follows from 5
$\stackrel{xx}{\leq}$	follows from xx
$\stackrel{rs}{\geq}$	follows from rs
$\stackrel{xx}{<}$	less than follows from xx
$\stackrel{xx}{>}$	great than follows from xx
$K = \{1, \ldots, m\}$	set of natural numbers between 1 and m
B_α^N	ball in $\Re^N$ space centered at initial point and with radius α
$\Re^N$	real N-dimensional space
Ω^m	m-tuple Cartesian product of set Ω
$\breve{z}$	element in Cartesian (multiple) product of specified set
$\bar{X}$	general (sub)set of a given set
X	system of subsets of a given set
$\mathcal{X}$	set of systems of sets over a given set
$2^{\bar{X}}$	potential set of $\bar{X}$, $2^{\bar{X}} \stackrel{\text{def}}{=} \{\bar{Z} \,\vert\, \bar{Z} \subset \bar{X}\}$
$[\Omega]_\lambda$	linear hull of set Ω of vectors from linear vector space (=set of all linear combination of finite set of vectors from Ω)
$\bar{B}^\perp$	orthogonal complement of the set $\bar{B}$
$[\bar{B}]_\kappa$	convex hull of set $\bar{B}$
$\mathcal{H}_n^*$	vektor space (with specified parameters)
$C_{\bar{A}}$	set of continuous functions defined on a given set $\bar{A}$
$\mathbf{NP}_k^\alpha$	set of objest with predefined properties, generally set of functions, mappings, etc., context dependent
B^*	general learning algorithm
$\bar{A} \triangle \bar{B}$	symetric difference of two sets $\bar{A}$ and $\bar{B}$
$\bar{A} \dot{-} \bar{B}$	difference of sets $\bar{A}$ and $\bar{B}$
$o\left(\widetilde{f}\right)$	$\widetilde{g} = o\left(\widetilde{f}\right)$ means that absolute value of function $\widetilde{g}$ is on assymptotic neighbourhood of zero majorized by function $\widetilde{f}$
$O\left(\widetilde{f}\right)$	$\widetilde{g} = O\left(\widetilde{f}\right)$ means that absolute value of function $\widetilde{g}$ is on assymptotic neighbourhood of infinity majorised be values of $\widetilde{f}$
$\boldsymbol{A}$	matrix of real numbers, colums or rows will be denoted by notion LINES OF MATRIX
$\vec{\boldsymbol{x}}$	vector from a given vector space (all vectors will be considerd as columns)

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

$\vec{\boldsymbol{x}}_i$	i-th indice of the vector)
$\boldsymbol{A}^{(n)}$	square matrix of order 2^n
$\vec{\boldsymbol{x}}^{(n)}$	vector of dimension 2^n
$\langle \vec{\boldsymbol{x}} \, \vert \vec{\boldsymbol{y}} \rangle$	scalar product of vectors, $\langle \vec{\boldsymbol{x}} \, \vert \vec{\boldsymbol{y}} \rangle = \sum_{i=1}^{n} \vec{\boldsymbol{x}}_i \vec{\boldsymbol{y}}_i$
$\vec{\boldsymbol{x}} \odot \vec{\boldsymbol{y}}$	tensor product of vectors or matrices
$\overset{\wedge}{\text{AND}}$	binary function AND
$\overset{\vee}{\text{OR}}$	binary function OR
$\overset{\sqcup}{\text{XOR}}$	binary function XOR
$\overset{\neg}{\text{NOT}}$	unary function NOT
$\lfloor M \rfloor$	greatest whole number less than M (=whole bellow part)
$\lceil M \rceil$	smallest whole number greater than M (=whole upper part)
$vpm\,(j)$	$vpm\,(j)$ is vector from $\{-1,+1\}^n$, (resp. $\{0,1\}^n$, by context), whose coordinates correspond with binary inscription of natural number j (coordinates of vector $vpm\,(j)$ correspond to pozitions of 1 in binary expansion of j, are equal to 1).
$vbin\,(j)$	$vbin\,(j)$ is vector from 0,1, whose coordinates correspond with binary inscription of natural number j (coordinates of vector $vbin\,(j)$ correspond to pozitions of 1 in binary expansion of j, are equal to 1).
$nint\,(\vec{\boldsymbol{v}})$	whole number, derived from vector $\vec{\boldsymbol{v}} \in \{-1,+1\}^n$, (resp. $\vec{\boldsymbol{v}} \in \{0,1\}^n$, by context), where ones components of vector $nint\,(\vec{\boldsymbol{v}})$ correspond to 1 in binary expansion of number $nint\,(\vec{\boldsymbol{v}})$, the rest components correspond to position of 0 in binary expansion of $nint\,(\vec{\boldsymbol{v}})$
$\binom{a}{i}$	binomical coefficien
$\widetilde{f}$	general mapping between two sets
$\left(\widetilde{f} * \widetilde{g}\right)$	convolution of functions $\widetilde{f}$ a $\widetilde{g}$
$\widetilde{A}\,(b)$	value of function (mapping) $\widetilde{A}$ at point b
$\ln(x)$	natural logarithm of x
$\log_2(x)$	logarithm of radix 2
$x \equiv y \mod r$	$x \equiv y \mod r$ means that there exist whole number k such that $x = yk + r$, where x, y, and r are whole numbers
$Prob_{\widetilde{\Pi}}\left(\bar{A}\right)$	probability of set $\bar{A}$ under probability distribution $\widetilde{\Pi}$
$\left\vert \bar{V} \right\vert$	cardinality of set $\bar{V}$
$\vert\delta\vert$	absolute value of number δ
$\left\Vert \vec{\boldsymbol{z}} \right\Vert_{max}$	maximum norm of vector
$\left\Vert \vec{\boldsymbol{z}} \right\Vert_{E}$	Eucleidian norm of vector
$\left\Vert \vec{\boldsymbol{z}} \right\Vert_{\mathcal{L}_2^{\bar{K}}}$	specified norm of vector
$\bar{A}\big\vert_{\mathcal{P}}$	orthogonal projection of set $\bar{A}$ into subspace $\mathcal{P}$

Newly mentioned object will be emphasized in the text by another font, like NEWLY INTRODUCED NOTION .

Preface

About this booklet. This work grew out of lecture notes for MSc. courses that I taught at the Mathematical department of FNSPE CTU in Prague since 1997. Consequently, this is a textbook covering enough material for two semester courses. The main focus of this textbook is on introduction to PAC learning model and its variants.

The audience. This text is intended for students in mathematics and other fields such as computer science and electrical engineering. Also a significant portion of the presented material is suitable for undergraduates.

Prerequisites. The necessary prerequisites is basic linear algebra and theory of probability. In some places knowledge of some basic analysis, functional analysis and topology is needed.

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Chapter 1

PAC learning model

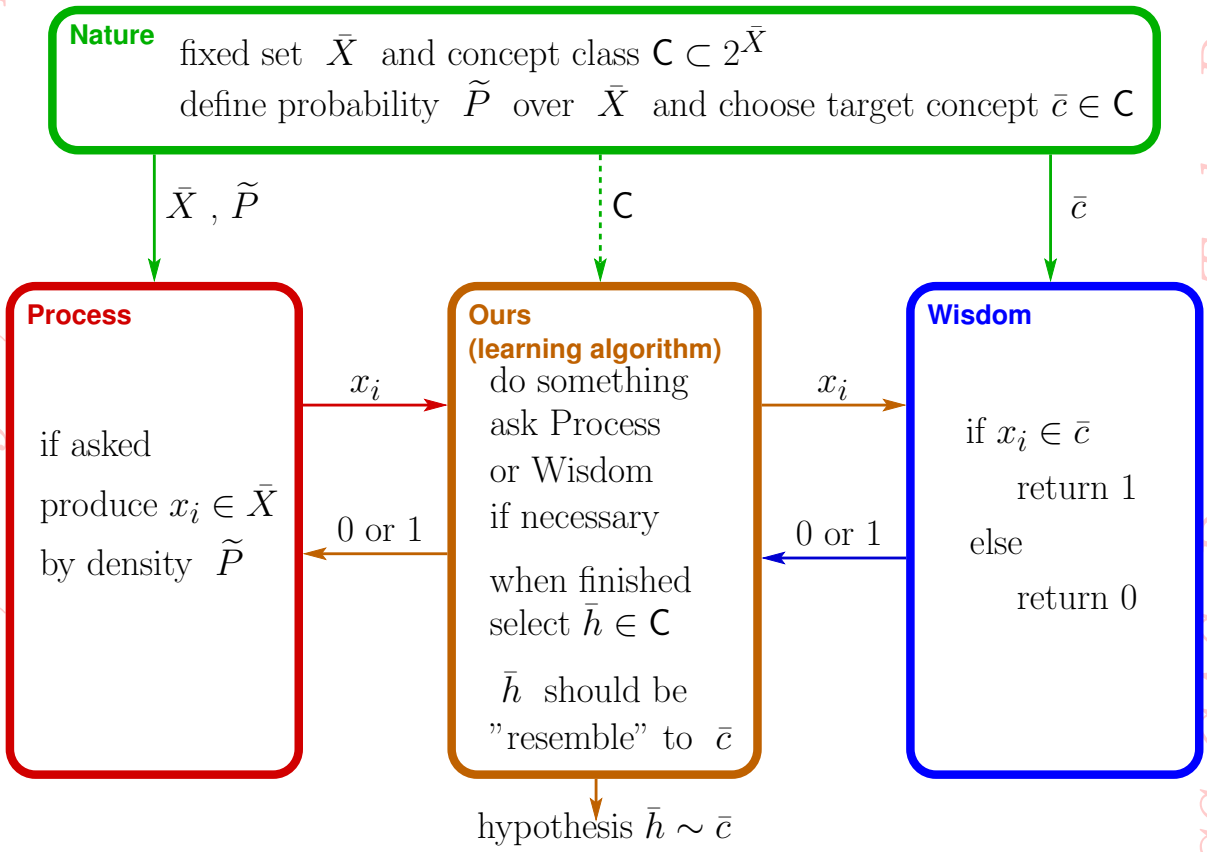


Figure 1.1: Learning scenario.

1.1 Concepts and concept classes

Key subject relating to the term learning algorithm is any subset of firmly given set $\bar{X}$. The terms concept, hypothesis, concept class and hypothesis class are used to designate such systems in the literature relevant to our further explanation. That is also why this terminology will also be used in the following definitions and explanations.

Definition 1.1.1 *Let* $\bar{X}$ *be an arbitrary set. Then we call* $\bar{c} \subset \bar{X}$ CONCEPT *(over set* $\bar{X}$ *). Nonempty collection of sets* $\mathsf{C} \subset 2^{\bar{X}}$ *is* CONCEPT CLASS *(over set* $\bar{X}$ *). Members of this class* $\bar{c} \in \mathsf{C}$ *are concepts of the class* C *.*

One of the methods for describing concept is to set up the concept as a sample of a given set in a representation. Formally, we then define the concept $\mathsf{C}_{\widetilde{f}}$ for Boolean function $\widetilde{f}$.

Definition 1.1.2 *Let* $\widetilde{f}$ *be a mapping from the set* $\bar{X}$ *into two point set* $\{-1,+1\}$ *. Then define concept* $\bar{c}_{\widetilde{f}}$ *as*

$$\bar{c}_{\widetilde{f}} \stackrel{def}{=} \left\{ x \in \bar{X} \,\middle|\, \widetilde{f}\,(x) = 1 \right\}.$$

Obviously, the preceding definition can be used to define concept class corresponding to set of functions over $\bar{X}$.

Definition 1.1.3 *Let us assume that* $\bar{F}$ *is an arbitrary set of function which maps* $\bar{X}$ *into* $\{-1,+1\}$ *. Then we say that* $\bar{F}$ REPRESENT CONCEPT CLASS

$$\mathsf{C}_{\bar{F}} \stackrel{def}{=} \left\{ \bar{A} \subset \bar{X} \,\middle|\, (\exists \widetilde{f} \in \bar{F})(\bar{A} = \bar{c}_{\widetilde{f}}) \right\}.$$

Among others such classes we point out so called $\mathbf{HALFSPACE}_n$ which is most natural concept class in variety applications.

Definition 1.1.4 *Let* n *be natural number and*

$$\bar{F} \stackrel{def}{=} \left\{ \widetilde{f} : \Re^n \to \{-1,+1\} \,\middle|\, \widetilde{f}\,(\vec{\boldsymbol{x}}) = \widetilde{sgn}\,(\langle \vec{\boldsymbol{x}} \,|\vec{\boldsymbol{w}}\rangle - t), \quad t \in \Re, \;\; \vec{\boldsymbol{w}}, \vec{\boldsymbol{x}} \in \Re^n \right\}.$$

Then $\mathbf{HALFSPACE}_n \stackrel{def}{=} \mathsf{C}_{\bar{F}}$.

It is straightforward that $\mathbf{HALFSPACE}_n$ contains all halfspaces of n-dimensional Euclidean space.

Definition 1.1.5 *Let* n *be natural number and*

$$\bar{F} \stackrel{def}{=} \left\{ \widetilde{f} : \Re^n \to \{-1,+1\} \,\middle|\, \widetilde{f}\,(\vec{\boldsymbol{x}}) = \widetilde{sgn}\,(\|\vec{\boldsymbol{x}} - \vec{\boldsymbol{c}}\|_E - r), \quad r \in \Re^+, \;\; \vec{\boldsymbol{c}}, \vec{\boldsymbol{x}} \in \Re^n \right\}.$$

Then $\mathbf{BALL}_n \stackrel{def}{=} \mathsf{C}_{\bar{F}}$.

It is straightforward that $\mathbf{BALL}_n$ contains all balls in n-dimensional Euclidean space.

1.2 Basic Terms of Learning Algorithms

For better comprehension of the functionality of the Probably Approximately Correct learning model we will use a more formal example of the role of learning than the one presented at the beginning of this section. Let us consider a problem involving a specific given set $\bar{A} \in \Re^n$ knowing a priori that $\bar{A}$ is a sphere in geometric terms. Let us assume to have a given sequence of points $\vec{\boldsymbol{x}}_i \in \Re^n$, while each of them is known either to lie within the set $\bar{A}$ or not. Our task is to decide how does the set $\bar{A}$ look like, and - in this specific case - to find the centre and radius of the relevant sphere. It is evident that our decision is highly dependent on the properties of the given sequence $\vec{\boldsymbol{x}}_i$. Let us assume that, based on an algorithm (deterministic or nondeterministic), we have opted for a sphere in $\Re^n$, claiming it is equal to the set $\bar{A}$. It is natural to expect that this particular set should contain all the points $\vec{\boldsymbol{x}}_i \in \bar{A}$ and, on the other hand, should contain no $\vec{\boldsymbol{x}}_i$ that does not lie within $\bar{A}$. This requirement will be called consistency and a set which we assume to be equal to the sphere $\bar{A}$ will be called a hypothesis (the equality of the sphere selected by us and $\bar{A}$ is solely our hypothesis, based on the properties of the algorithm used and on the sequence of points $\vec{\boldsymbol{x}}_i$).

Let us designate our chosen sphere $\bar{K}$. Then, it is evident that the condition of consistency must be met in case, when symmetric difference of the sets $\bar{K}$ and $\bar{A}$ will be an empty set, and it is natural to demand that this should hold for the generated hypothesis or that this should at least hold with a high rate of probability. The terms discussed above are treated in the following definitions.

Definition 1.2.1 *Let* C *be a concept class and a set system* H *satisfies* $\mathsf{C} \subset \mathsf{H} \subset 2^{\bar{X}}$. *Then,* H *is* HYPOTHESIS CLASS *for concept class* C.

Definition 1.2.2 *Let* $\breve{x} \stackrel{def}{=} \{x_1, \ldots, x_m\}$, $x_i \in \bar{X}$, $i \in \{1, \ldots, m\}$, $\vec{\boldsymbol{z}} \in \{-1, +1\}^m$ *and let* $\bar{c} \subset \bar{X}$. *Then the ordered tuple*

$$\left(\breve{x}, \vec{\boldsymbol{z}}\right)$$

is a SAMPLE OF CONCEPT $\bar{c}$ OF LENGHT m *if and only if*

$$(\forall i \in \{1, \ldots, m\}) \left((x_i \in \bar{c}) \Leftrightarrow (\vec{\boldsymbol{z}}_i = 1)\right).$$

For concept class C *define* SAMPLE SPACE OF CONCEPT CLASS *as*

$$\bar{S}_{\mathsf{C}} \stackrel{def}{=} \bigcup_{m \geq 1} \left\{ \bigcup_{\bar{c} \in \mathsf{C}} \left\{ \left(\breve{x}, \vec{\boldsymbol{z}}\right) \middle| \left(\breve{x}, \vec{\boldsymbol{z}}\right) \text{ is a sample of the length } m \text{ of concept } \bar{c} \right\} \right\}.$$

A set $\bar{b} \subset \bar{X}$ *is* CONSISTENT SET *with sample* $\left(\breve{x}, \vec{\boldsymbol{z}}\right)$ *if and only if for all* $i \in \{1, \ldots, m\}$, *holds* $\left(x_i \in \bar{b} \Leftrightarrow \vec{\boldsymbol{z}}_i = 1\right)$.

A rate of difference between two sets has to be available for the purpose of further explanation. We define this rate of difference as follows:

Definition 1.2.3 *Let* $\bar{X}$ *be a probability space with probability distribution* $\widetilde{P}$ *. Furthermore,* C *is concept class over* $\bar{X}$ *and* H *is hypothesis class for* C *. Then, for each* $\bar{c} \in \mathsf{C}$ *and* $\bar{h} \in \mathsf{H}$ *the number*

$$\mathsf{e}_{\widetilde{P}}\left(\bar{c}, \bar{h}\right) \stackrel{def}{=} Prob_{\widetilde{P}}\left(\bar{h} \triangle \bar{c}\right)$$

is ERROR OF HYPOTHESIS $\bar{h}$ REGARDING A CONCEPT $\bar{c}$ *and probability* $\widetilde{P}$ *(symbol* $\bar{h} \triangle \bar{c}$ *denotes symmetric difference between sets* $\bar{h}$ *and* $\bar{c}$ *, e.g.* $\bar{h} \triangle \bar{c} = \left(\bar{c} \dot{-} \bar{h}\right) \cap \left(\bar{h} \dot{-} \bar{c}\right)$*).*

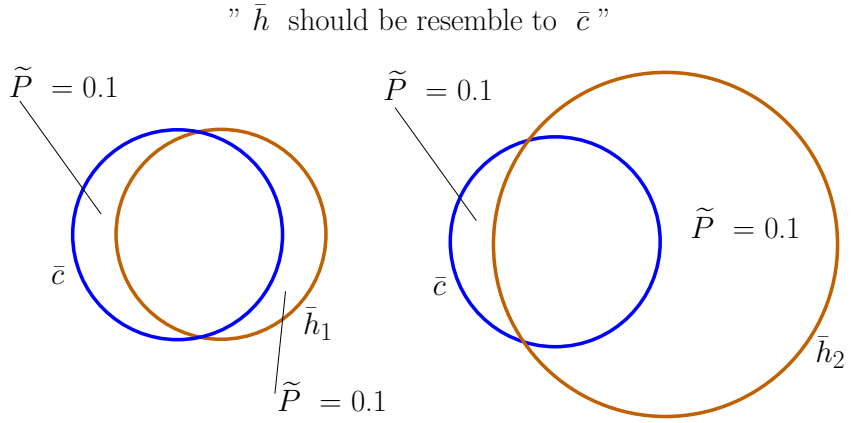


Figure 1.2: Example of two hypotheses with the same error.

Since the set $\bar{A}$ has been described solely by the sequence of its elements (plus a priori information on its shape) our considered learning algorithm is essentially a mapping between the set of all the samples of the set $\bar{A}$ into the set of all spheres in $\Re^n$. We will not require a precise learning algorithm to provide a exact hypothesis for all the possible tasks of this type but we will only require the given learning algorithm to produce, in most cases, a "satisfactory" hypothesis (for instance a hypothesis consistent with the given sample). The following definition represents a quantification of this requirement.

Definition 1.2.4 *Let us have a defined function* $\widetilde{m}(\epsilon, \delta)$ *mapping a set* $(0,1) \times (0,1)$ *into a set of natural numbers, and let* H *be a hypothesis class for the concept class* C*, defined over the set* $\bar{X}$ *. Then* LEARNING ALGORITHM OF COMPLEXITY $\widetilde{m}(\epsilon, \delta)$ *of the concept class* C *is each mapping* $\widetilde{\mathrm{A}^*} : \bar{S}_{\mathsf{C}} \to \mathsf{H}$ *such that for all* $\bar{c} \in \mathsf{C}$*, for all* $0 < \epsilon < 1$*,* $0 < \delta < 1$ *and for any probability* $\widetilde{P}$ *defined on* $\bar{X}$ *is the probability of the set*

$$\left\{ \breve{x} \in \bar{X}^m \,\middle|\, \left(\breve{x}, \vec{\boldsymbol{z}}\right) \text{ is sample } \bar{c} \quad a \quad \mathsf{e}_{\widetilde{P}}\left(\bar{c}, \widetilde{\mathrm{A}^*}\left(\left(\breve{x}, \vec{\boldsymbol{z}}\right)\right)\right) \geq \epsilon \right\}$$

smaller than the number δ*. If such a learning algorithm exists we say that* C IS UNIFORMLY LEARNABLE *according to the hypothesis class* H *. We will call each such learning algorithm the* (ϵ, δ)-LEARNING ALGORITHM *.*

REMARKS:
In the literature, this particular model of computational complexity is designated as PAC (Probably Approximately Correct) learning.

It is evident that each set may be described very precisely by means of a great number of its members. Algorithms using very long samples may probably produce much more veritable hypotheses. If we are satisfied with the given accuracy of generating hypotheses, described by numbers ϵ and δ, it is natural to ask how many a sample must be used for the given concept class to guarantee this accuracy. That is why we will introduce the term sample complexity of learning algorithms as follows:

Definition 1.2.5 *Minimal value* $\widetilde{m}(\epsilon,\delta)$ *for which* A^* *is a learning algorithm is a* SAMPLE COMPLEXITY *of learning algorithm* A^* .

This complexity is defined solely with a view to the length of the sample of the learning algorithm used; the issue of the inner computational complexity of the algorithm itself is not taken into consideration in this case! The roughest estimate of sample complexity is given by the subsequent statement, based solely on probability properties (providing a basic estimate of the number of samples that have to used for the learning algorithm which produces a consistent hypothesis).

Theorem 1.2.1 *Let us assume that* C *is a concept class over finite set* $\bar{X}$ *and* $\mathsf{H}=\mathsf{C}$. *Let learning algorithm* A^* *require at least*

$$\frac{1}{\epsilon}\ln\left(\frac{|\mathsf{C}|}{\delta}\right) \tag{1.1}$$

queries and for any given concept $\bar{c}\in\mathsf{C}$ *and any probability density* $\widetilde{P}$ *defined on* $\bar{X}$, *produce a consistent hypothesis. Than*

$$Prob_{\widetilde{P}}\left(\mathsf{e}_{\widetilde{P}}\left(\bar{c},\widetilde{\mathrm{A}^*}\left(\left(\breve{x},\vec{\boldsymbol{z}}\right)\right)\right)\geq\epsilon\right)<\delta.$$

■ *Proof:*
Let us assume that $\bar{h}$ is a hypothesis consistent with the sample $(\{x_1,\dots,x_m\},\vec{\boldsymbol{z}})$, which is generated by a learning algorithm. Without loos of generality, it is possible to assume that for some $1\leq k\leq m$ is $x_1,\dots,x_k\in\bar{c}$ and $x_{k+1},\dots,x_m\notin\bar{c}$. Hence, it ensues from the consistency of $\bar{h}$ that $x_1,\dots,x_k\in\bar{h}$ a $x_{k+1},\dots,x_m\notin\bar{h}$. Therefore, the following estimate holds

$$\mathsf{e}_{\widetilde{P}}\left(\bar{c},\bar{h}\right)=\sum_{x\in\bar{c}\triangle\bar{h}}Prob_{\widetilde{P}}(x)\leq 1-\sum_{i=1}^{m}Prob_{\widetilde{P}}(x_i).$$

Hence, we obtain from the requirement $\epsilon\leq\mathsf{e}_{\widetilde{P}}\left(\bar{c},\bar{h}\right)$

$$\sum_{i=1}^{m}Prob_{\widetilde{P}}(x_i)\leq 1-\epsilon$$

and from the positiveness of the numbers $Prob_{\widetilde{P}}(x_i)$ it further ensues that

$$(\forall i\in\{1,\dots,m\})(Prob_{\widetilde{P}}(x_i)\leq 1-\epsilon).$$

Hence, the probability of selection of m points $x_i \in \bar{X}$ such that they are consistent with hypothesis $\bar{h}$ for which $\mathsf{e}_{\widetilde{P}}\left(\bar{c}, \bar{h}\right) \geq \epsilon$ meets the inequality

$$Prob_{\widetilde{P}}\left(\left\{x_1, \ldots, x_m \left| \mathsf{e}_{\widetilde{P}}\left(\bar{c}, \bar{h}\right) \geq \epsilon \text{ and } \bar{h} \text{ is consistent } \right.\right\}\right) = \prod_{i=1}^{m} Prob_{\widetilde{P}}\left(x_i\right) \leq (1-\epsilon)^m$$

(selected queries may be repeated; this is an independent selection).

But we are interested to know the probability that such hypothesis really exists. The number of hypotheses is, however, finite and equals the number of concepts in the class C. Therefore, the overall probability of the incidence of the hypothesis with the properties described above is lower than the number

$$|\mathsf{C}| \cdot (1-\epsilon)^m.$$

We demand that probability should, at the most, equal the value of δ, hence that it should hold that $|\mathsf{C}| \cdot (1-\epsilon)^m \leq \delta$. Let us choose

$$m \stackrel{\text{def}}{=} \left\lceil \frac{1}{\epsilon} \ln\left(\frac{|\mathsf{C}|}{\delta}\right) \right\rceil.$$

Then the inequality ($\delta < 1$) holds [1] .

$$|\mathsf{C}| \cdot (1-\epsilon)^m \leq |\mathsf{C}| e^{-\epsilon m} \leq |\mathsf{C}| e^{\ln\left(\frac{\delta}{|\mathsf{C}|}\right)} = \delta.$$

– *q. e. d.* –

Example 1.2.1 *Let* $\bar{X} \stackrel{def}{=} \{1, \cdots, k\}$, $\mathsf{C} \stackrel{def}{=} \left\{\bar{a} \subset \bar{X} \left| \left(\exists i \in \bar{X}\right)\left(\bar{a} = \left\{j \in \bar{X} \left| i \leq j \right.\right\}\right)\right.\right\}$. *Further define learning algorithm* A^* *as*

$$\widetilde{\mathrm{A}^*}\left((x_1, \ldots, x_m, z_1, \ldots, z_m)\right) = \bar{h} = \langle b, k\rangle \cap \bar{X}, \quad \textit{where} \quad b = \min_{i \in \{1,\ldots,m\}} \left\{x_i \left| z_i = 1 \right.\right\}.$$

Obviously the algorithm A^* *produces consistent hypotheses only. At the same time,* $|\mathsf{C}| = k$. *Let* $0 < \epsilon, \delta < 1$ *and*

$$m \geq \frac{1}{\epsilon} \widetilde{ln}\left(\frac{k}{\delta}\right).$$

Then A^* *is* (ϵ, δ)*-learning algorithm.*

Example 1.2.2 *Let* $\bar{X} \stackrel{def}{=} \{1, \cdots, k\}^n$, *let*

$$\mathsf{C} \stackrel{def}{=} \left\{\langle a_1, b_1\rangle \times \langle a_2, b_2\rangle \times \cdots \times \langle a_n, b_n\rangle \left| \left(\forall j \in \{1, \ldots, n\}\right)\left(a_j \leq b_j\right)\right.\right\}.$$

Further define learning algorithm A^* *as*

$$\widetilde{\mathrm{A}^*}\left((\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_m, z_1, \ldots, z_m)\right) = \langle a_1, b_1\rangle \times \langle a_2, b_2\rangle \times \cdots \times \langle a_n, b_n\rangle$$

[1] Let $\widetilde{f}\left(\epsilon\right) \stackrel{\text{def}}{=} e^{-\epsilon} - (1-\epsilon)$. Then $\widetilde{f}\left(0\right) = 0$ and $\widetilde{f}'\left(\epsilon\right) = 1 - e^{-\epsilon} > 0$ on $(0,1)$. Hence $e^{-\epsilon} > 1 - \epsilon$ on $(0,1)$.

where

$$(\forall j \in \{1,\dots,n\}) \left(a_j = \min_{i \in \{1,\dots,m\}} \left\{ (\vec{\boldsymbol{x}}_i)_j \,|\vec{\boldsymbol{z}}_i = 1 \right\} \quad \textit{and} \quad b_j = \max_{i \in \{1,\dots,m\}} \left\{ (\vec{\boldsymbol{x}}_i)_j \,|\vec{\boldsymbol{z}}_i = 1 \right\} \right) .$$

Obviously the algorithm A^* *produces consistent hypotheses only. At the same time,* $|\mathsf{C}| = \binom{k}{2}^n$. *Let* $0 < \epsilon, \delta < 1$ *and*

$$m \geq \frac{1}{\epsilon} \widetilde{ln} \left(\frac{\binom{k}{2}^n}{\delta} \right) .$$

Then A^* *is* (ϵ, δ)*-learning algorithm.*

Let us note that the estimate 1.1 is not dependent on the specific rate of probability $\widetilde{P}$. This can be explained by the fact that the influence of different probabilities of the elements in $\bar{X}$ is compensated by the probability of the selection of this element according to the density of probability $\widetilde{P}$. In other words, elements from $\bar{X}$ with considerable probability, which - in case of belonging to symmetric difference $\bar{c} \triangle \bar{h}$ greatly contribute to the actual size of error, being more frequent in an average sample, which - in view of the fact that algorithm always produces a consistent hypothesis - eliminates their impact. That is why the estimate 1.1 is independent of probability $\widetilde{P}$.

In the subsequent explication, we will demonstrate that this estimate can be markedly improved, notably by applying the term VC-dimension.

© F. Hakl, ICS CAS, Prague. Tech. Rep. №1227, Dec 2015

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Chapter 2

Vapnik-Chervonenkis dimension

We now begin to introduce the mathematical theory which is necessary to study PAC learning in cases of infinite concept and hypotesis classes. A main notion discussed in this chapter is Vapnik-Cervonenkis dimension which is a measure of possible set clasification potential of set collection contained in concept class given.

2.1 VC-dimension of concept class

Definition 2.1.1 *Let* $\bar{X}$ *is a given set,* $\bar{Y} \subset \bar{X}$, $\mathsf{C} \subset 2^{\bar{X}}$. *Then* C IS SHATTERED BY $\bar{Y}$ *, iff:*

$$(\forall \bar{z} \subset \bar{Y})(\exists \bar{c} \in \mathsf{C})(\bar{c} \cap \bar{Y} = \bar{z}).$$

Further define VAPNIK-CHERVONENKIS DIMENSION *of the set system* C *as*

$$\mathtt{VC}_{dim}(\mathsf{C}) \stackrel{def}{=} \sup\left\{\left|\bar{Y}\right| \,\middle|\, \mathsf{C} \text{ is shatterd by } \bar{Y}\right\}.$$

Note, that Vapnik-Chervonenkis dimension is equal to the size of maximal set $\bar{A} \subset \mathsf{X}$ satisfying

$$\left\{\bar{B} \,\middle|\, (\exists \bar{Z} \in \mathsf{C})(\bar{B} = \bar{A} \cap \bar{Z})\right\} = 2^{\bar{A}},$$

or to $+\infty$, if an set of any size is shattered by C .

The notion of VC-dimension is so imortant that we illustrate it using following examples (see [AB92]).

Example 2.1.1

1. *Let* $\bar{X} = \Re$ *and* C *are all intervals* $(a, +\infty)$ *, where* $a \in \Re$. *In this case, obviously for an arbitrary* $\{b, c \in \Re \,|b < c\}$ *there does not exists an interval* $(a, +\infty)$ *which contains the point* b *and does not contain the point* c. *Hence* $\mathtt{VC}_{dim}(\mathsf{C}) = 1$.

2. *Let* $\bar{X} = \Re$, s *is a natural number,* $s > 1$ *and* C *contains all union of* s *intervals. Let* $\bar{S} = \{x_1, \ldots, x_{2s} \,|x_i < x_{i+1}\}$. *It is straightforward that* $\bar{S}$ *is shattered by concept class* C . *But for* $\bar{S} = \{x_1, \ldots, x_{2s+1} \,|x_i < x_{i+1}\}$ *each concept containing points* $x_1, x_3, \ldots, x_{2s+1}$ *also contains at least one point from* $x_2, x_4, \ldots, x_{2s}$. *Clearly such* $\bar{S}$ *is not shattered by the concept class* C . *It follows* $\mathtt{VC}_{dim}(\mathsf{C}) = 2s$.

3. *Let $\bar{X} = \Re^n$ and C are all intervals in the space $\Re^n$. Obviously an set of $2n$ points located in the midle of faces of the unit cube is shattered by the concept class C (we can move faces slightly in the direction of coordinates axes). But for arbitrary set $\bar{A}$ of $2n+1$ points, there exists minimal parallelepiped containing whole set $\bar{A}$. Straightforwardly there exists an point $\vec{\boldsymbol{x}} \in \bar{A}$ which is in the interior of mimimal parallelepiped, or lies in the some of its face. Obviously the set $\bar{A} - \{\vec{\boldsymbol{x}}\}$ can not be separate off the set $\{\vec{\boldsymbol{x}}\}$. Hence $\mathtt{VC}_{dim}(\mathsf{C}) = 2n$.*

Example 2.1.2 *For the completness we show an example of the concept class with infinite VC-dimension.*

1. *Put*

$$\mathsf{C} \stackrel{def}{=} \left\{ \bar{A}_\alpha \left| (\exists \alpha \in \Re^n) \left(\bar{A}_\alpha = \left\{ x \in \Re \left| \widetilde{\sin}(\alpha x) \geq 0 \right. \right\} \right) \right. \right\}.$$

Then for arbitrary natural l define sequence $\bar{Z}_l \stackrel{def}{=} \{z_i\}_1^l$, $z_i \stackrel{def}{=} \frac{1}{10^i}$. Further for any subset of the set $\bar{Q} \subset \bar{Z}_l$ the sequence $\delta_1, \ldots, \delta_l$, $\delta_i \in \{0,1\}$ satisfies $\delta_i = 1 \Leftrightarrow z_i \in \bar{Q}$. If we define

$$\alpha \stackrel{def}{=} \pi \left(\sum_{i=1}^{l} (1-\delta_i) 10^i + 1 \right),$$

then equality

$$\alpha z_j = \alpha \frac{1}{10^j} = \pi \left(\sum_{i=1}^{l} (1-\delta_i) 10^{i-j} + \frac{1}{10^j} \right).$$

holds. Hence

$$\alpha z_j = \alpha \frac{1}{10^j} = \pi \left(\sum_{i=1}^{j-1} \frac{1-\delta_i}{10^{j-i}} + \frac{1}{10^j} + (1-\delta_j) + \sum_{i=j+1}^{l} (1-\delta_i) 10^{i-j} \right).$$

Obviously the expression $\sum_{i=j+1}^{l} (1-\delta_i) 10^{i-j}$ is divisible by 2 while the expression $\sum_{i=1}^{j-1} \frac{1-\delta_i}{10^{j-i}} + \frac{1}{10^j}$ is less than to 1.

It follows that $\widetilde{\sin}(\alpha z_j) < 0$ for $\delta_j = 0$ and $\widetilde{\sin}(\alpha z_j) > 0$ for $\delta_j = 1$. Thus the set $\bar{A}_\alpha$ separate each subset $\bar{Z}_l$ defined via sequence $\delta_1, \ldots, \delta_l$. Therefore for arbitrary l there exists the set of the size l which is shattered by the concept class C, So $\mathtt{VC}_{dim}(\mathsf{C}) = +\infty$.

2. *Put*

$$\mathsf{C} \stackrel{def}{=} \left\{ [\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_m]_\kappa \,|m \in N \text{ and } (\forall i \in \{1, \ldots, m\}) (\vec{\boldsymbol{x}} \in \Re^n) \right\},$$

e.g. C is the set of all convex hulls of all finite subsets of $\Re^n$. It is clear, that vertices of a given convex hull over m points can be shattered by C, so $\mathtt{VC}_{dim}(\mathsf{C}) = +\infty$.

2.1.1 General properties of VC-dimension

In the next text we analyze concept classes defined over an finite set $\bar{X}$. We start with the Sauer's lemma ([Sau72]) which provide us upper bound on number of concepts in a concept class with predefined VC-dimension. We feel necessary to point out here that this upper boudnd is genuine combinatorial property of finite sets.

Lemma 2.1.1 (Sauer) *Let* $\bar{X}$ *be a finite set and* $\mathsf{C} \subset 2^{\bar{X}}$. *Then*

$$|\mathsf{C}| \leq \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})} \binom{|\bar{X}|}{i}.$$

Further, there exists $\mathsf{C} \subset 2^{\bar{X}}$ *such that equality holds.*

▪ *Proof:*
Let us define for arbitrary but fixed $y \in \bar{X}$ the following sets

$$\begin{aligned} \mathsf{C}(y) &\overset{\text{def}}{=} \left\{ \bar{A} \dot{-} \{y\} \,\middle|\, \bar{A} \in \mathsf{C} \right\} \\ \mathsf{C}_y &\overset{\text{def}}{=} \left\{ \bar{A} \in \mathsf{C} \,\middle|\, (\exists \bar{B} \in \mathsf{C})(\bar{A} \neq \bar{B} \text{ and } \bar{B} = \bar{A} \cup \{y\}) \right\} \\ \mathsf{C}^y &\overset{\text{def}}{=} \left\{ \bar{A} \in \mathsf{C} \,\middle|\, (\exists \bar{B} \in \mathsf{C}_y)(\bar{A} = \bar{B} \cup \{y\}) \right\}. \end{aligned}$$

We prove the claim of lemma it tree steps.
▪ add 1)
Firstly, we show that for arbitrary $y \in \bar{X}$ is

$$|\mathsf{C}| - |\mathsf{C}(y)| = |\mathsf{C}_y|.$$

Obviously

$$\mathsf{C}(y) = \left\{ \bar{A} \,\middle|\, \bar{A} \in \mathsf{C}, y \notin \bar{A} \right\} \cup \left\{ \bar{A} \dot{-} \{y\} \,\middle|\, \bar{A} \in \mathsf{C}, y \in \bar{A} \right\}.$$

Hence

$$\begin{aligned} |\mathsf{C}(y)| = \left| \left\{ \bar{A} \,\middle|\, \bar{A} \in \mathsf{C}, y \notin \bar{A} \right\} \right| + \left| \left\{ \bar{A} \dot{-} \{y\} \,\middle|\, \bar{A} \in \mathsf{C}, y \in \bar{A} \right\} \right| - \\ \left| \left\{ \bar{A} \,\middle|\, \bar{A} \in \mathsf{C}, y \notin \bar{A} \right\} \cap \left\{ \bar{A} \dot{-} \{y\} \,\middle|\, \bar{A} \in \mathsf{C}, y \in \bar{A} \right\} \right| = |\mathsf{C}| - |\mathsf{C}_y|. \end{aligned}$$

▪ add 2)
Secondly, we prove

$$\mathtt{VC}_{dim}\left(\mathsf{C}_y\right) = n - 1 \Rightarrow \mathtt{VC}_{dim}\left(\mathsf{C}\right) \geq n. \tag{2.1}$$

It follows from the assumption $\mathtt{VC}_{dim}\left(\mathsf{C}_y\right) = n - 1$ that there exists an set $\bar{A} \subset \bar{X} \dot{-} \{y\}$, $\left|\bar{A}\right| = n - 1$ which satisfy

$$\left\{ \bar{B} \,\middle|\, (\exists \bar{Z} \in \mathsf{C}_y)(\bar{B} = \bar{A} \cap \bar{Z}) \right\} = 2^{\bar{A}}. \tag{2.2}$$

To verify it we prove the equality

$$\left\{ \bar{B} \,\middle|\, (\exists \bar{Z} \in (\mathsf{C}_y \cup \mathsf{C}^y))(\bar{B} = (\bar{A} \cup \{y\}) \cap \bar{Z}) \right\} = 2^{\bar{A} \cup \{y\}}.$$

Inclusion $\subset$ is obvious, so let us pay our atention to opposite inclusion $\supset$. Let $\bar{H} \in 2^{\bar{A} \cup \{y\}}$. If $y \notin \bar{H}$ then by 2.2 $\exists \bar{Z} \in \mathsf{C}_y$ such that $\bar{H} = \bar{A} \cap \bar{Z} = (\bar{A} \cup \{y\}) \cap \bar{Z})$. On the contrary, let $y \in \bar{H}$. If we recall 2.2 then for the set $\bar{H} \dot{-} \{y\}$ there exists $\bar{M} \in \mathsf{C}_y$ such that $\bar{H} \dot{-} \{y\} = \bar{A} \cap \bar{M}$. On account of $\bar{M} \in \mathsf{C}_y$ it must be $\bar{M} \cup \{y\} \in \mathsf{C}^y$ and therefore the set equality $\bar{H} = (\bar{A} \cup \{y\}) \cap (\bar{M} \cup \{y\})$ holds. So we get that $\mathtt{VC}_{dim}\left(\mathsf{C}_y \cup \mathsf{C}^y\right) \geq n$ and because $(\mathsf{C}_y \cup \mathsf{C}^y) \subset \mathsf{C}$, the equality $\mathtt{VC}_{dim}\left(\mathsf{C}\right) \geq n$ must be fulfilled.

- add 3)

Now we are ready to conclude the proof of the lemma by induction on the set size $\left|\bar{X}\right|$.

$\left|\bar{X}\right| = 1, 2:$

The claim ot the lemma is straightforward.

$\left|\bar{X}\right| = k \mapsto \left|\bar{X}\right| = k + 1:$

Let us assume that induction assumption is valid fot the case when if $\left|\bar{X}\right| = k$. The definition of the set system C_y follows that the set cintained in it does not contain y, so the whole set system C_y is build over the set $\bar{X} \dot{-} \{y\}$. If we apply induction assumption we obtain

$$|\mathsf{C}_y| \leq \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C}_y)} \binom{\left|\bar{X}\right| - 1}{i} \leq \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})-1} \binom{\left|\bar{X}\right| - 1}{i}$$

(the second inequality holds owning to the fact $\mathtt{VC}_{dim}\left(\mathsf{C}\right) - 1 \geq \mathtt{VC}_{dim}\left(\mathsf{C}_y\right)$ which comes true due to 2.1).

At the same time, in view of induction assumption ($\mathsf{C}(y)$ is defined over the set $\bar{X} \dot{-} \{y\}$), the estimation

$$|\mathsf{C}(y)| \leq \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C}(y))} \binom{\left|\bar{X}\right| - 1}{i} \leq \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})} \binom{\left|\bar{X}\right| - 1}{i},$$

holds, whereas the second inequality immediatelly follows from the definition of the VC-dimension. So we can write (remind $\binom{j}{i} = \binom{j-1}{i} + \binom{j-1}{i-1}$ and $\binom{j}{0} \stackrel{\text{def}}{=} 1$, $\binom{j}{-1} \stackrel{\text{def}}{=} 0$)

$$|\mathsf{C}| = |\mathsf{C}(y)| + |\mathsf{C}_y| \leq \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})} \binom{\left|\bar{X}\right| - 1}{i} + \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})} \binom{\left|\bar{X}\right| - 1}{i-1} = \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})} \binom{\left|\bar{X}\right|}{i}.$$

So we proved the desired inequality. Finally, we have to show that there exists a concept class for which the equality is true. Let us define a concept class

$$\mathsf{C} \stackrel{\text{def}}{=} \left\{\bar{a} \subset \bar{X} \,||\bar{a}| \leq k\right\}.$$

It is easy to verify that $|\mathsf{C}| = \sum_{i=0}^{k} \binom{\left|\bar{X}\right|}{i}$ and that $\mathtt{VC}_{dim}\left(\mathsf{C}\right) = k$, which conclude the proof.

– *q. e. d.* –

The following simple lemma highlights usefulness of the Sauer's lemma.

Lemma 2.1.2 *Let* C *be a finite concept class. Then* $\mathtt{VC}_{dim}\left(\mathsf{C}\right) \leq \log_2\left(|\mathsf{C}|\right)$.

■ *Proof:*
To shatter any set of the size d we need at least 2^d different concepts, hence the set of the size greater than $\log_2(|\mathsf{C}|)$ can not be shattered by the set C. Therefore $\mathtt{VC}_{dim}(\mathsf{C}) \leq \log_2(|\mathsf{C}|)$.

– *q. e. d.* –

The notion introduces in the next definition has very narow relationship to VC-dimension and also to Sauer's lemma.

Definition 2.1.2 *Let C be a nonepty concept class ovet the set $\bar{X}$ and $\bar{S} \subset \bar{X}$. Then we define*

$$\Pi_{\mathsf{C}}(\bar{S}) \stackrel{def}{=} \{\bar{S} \cap \bar{c} \,|\bar{c} \in \mathsf{C}\}.$$

Further for a fixed $m \geq 0$ let us define

$$\Pi_{\mathsf{C}}(m) \stackrel{def}{=} \max\{|\Pi_{\mathsf{C}}(\bar{S})| \,|\bar{S} \subset \bar{X}\ ,\ |\bar{S}| = m\},$$

(maximum is over all sets $\bar{S} \subset \bar{X}$ of the size m).

Apparently $\Pi_{\mathsf{C}}(\bar{S})$ is the system of all subset of the set $\bar{S}$ whose are possible to separate from their complement in $\bar{S}$ using concept from concept class C. The number $\Pi_{\mathsf{C}}(m)$ expresses the cardinality of maximal such system $\Pi_{\mathsf{C}}(\bar{S})$ under condition that the set $\bar{S} \subset \bar{X}$ is finite set of the size m. There exists evident realationship between the number discussed above and the VC-dimension

$$\mathtt{VC}_{dim}(\mathsf{C}) = \sup\{m \,|\Pi_{\mathsf{C}}(m) = 2^m\}.$$

Definition 2.1.3 *For all $d \geq 0$ a $m \geq 0$ put $\Phi_{d,m} \stackrel{def}{=} \sum_{i=0}^{d} \binom{m}{i}$, if $m \geq d$ and $\Phi_{d,m} \stackrel{def}{=} 2^m$ if $m < d$.*

Estimations contained in the successive lemma are usefful in deriving upper bounds on necessary sample size which quarrants that learning algorithm produce an hypothesis with acceptable small error (in the sense of probablity of the symmetric difference between the original concept $\bar{c}$ and produced hypothesis $\bar{h}$).

Lemma 2.1.3 *If C is an arbitrary concept class over $\bar{X}$ and $\mathtt{VC}_{dim}(\mathsf{C}) = d < +\infty$, then*

1. *$\Pi_{\mathsf{C}}(m) \leq \Phi_{d,m}$ for all $d, m \geq 0$.*
2. *$\Phi_{d,m} \leq m^d + 1$ for $d, m \geq 0$ and $\Phi_{d,m} \leq m^d$ for $d \geq 0$ and $m \geq 2$.*
3. *$\Phi_{d,m} \leq 2\frac{m^d}{d!} \leq \left(\frac{em}{d}\right)^d$ for $m \geq d \geq 1$.*

■ *Proof:*
▪ add 1)
Let m be arbitrary and let $\bar{S} \subset \bar{X}$ be arbitrary subset such that $|\bar{S}| = m$ and $|\Pi_{\mathsf{C}}(\bar{S})| = \Pi_{\mathsf{C}}(m)$. Put $\bar{X}^* \stackrel{\text{def}}{=} \bar{S}$ and $\mathsf{C}^* \stackrel{\text{def}}{=} \{\bar{c} \cap \bar{S} \,|\bar{c} \in \mathsf{C}\}$. As $\Pi_{\mathsf{C}^*}(\bar{S}) = \{\bar{S} \cap \bar{c} \,|\bar{c} \in \mathsf{C}^*\}$, we have $|\Pi_{\mathsf{C}^*}(\bar{S})| \leq |\mathsf{C}^*|$. It implies that $\Pi_{\mathsf{C}^*}(m) \leq |\mathsf{C}^*|$. At the same time $\Pi_{\mathsf{C}^*}(\bar{S}) =$

$\left\{\bar{S}\cap\bar{c}\,|\bar{c}\in\mathsf{C}^*\right\}=\left\{\bar{S}\cap\bar{c}\,|\bar{c}\in\mathsf{C}\right\}=\Pi_{\mathsf{C}}\left(\bar{S}\right)$ which follows $\Pi_{\mathsf{C}^*}(m)=\Pi_{\mathsf{C}}(m)$ (note that we assume $\left|\Pi_{\mathsf{C}}\left(\bar{S}\right)\right|=\Pi_{\mathsf{C}}(m)$ and obviously $\Pi_{\mathsf{C}^*}(m)\leq\Pi_{\mathsf{C}}(m)$). In addition $\mathtt{VC}_{dim}(\mathsf{C}^*)\leq\mathtt{VC}_{dim}(\mathsf{C})$. So owning to Sauer's lemma we get the estimation

$$\Pi_{\mathsf{C}}(m)=\Pi_{\mathsf{C}^*}(m)\leq|\mathsf{C}^*|\leq\sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C}^*)}\binom{|\bar{X}^*|}{i}\leq\sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})}\binom{|\bar{X}^*|}{i}=\Phi_{\mathtt{VC}_{dim}(\mathsf{C}),m}\,.$$

Hence the first inequality in the lemma is clearly fulfilled for any m.

- add 2)

This part will be proved by induction on m and d using the relationship $\binom{m}{i}=\binom{m-1}{i}+\binom{m-1}{i-1}$ (recall that $\binom{j}{0}\stackrel{\text{def}}{=}1$, $\binom{j}{-1}\stackrel{\text{def}}{=}0$). So we can write

$$\sum_{i=0}^{d}\binom{m}{i}=\sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{d}\binom{m-1}{i-1}=\sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=1}^{d-1}\binom{m-1}{i}\stackrel{\text{ind.}}{\leq}$$

$$\stackrel{\text{ind.}}{\leq}(m-1)^d+(m-1)^{(d-1)}=m^d\left(\frac{(m-1)^d}{m^d}+\frac{(m-1)^{(d-1)}}{m^d}\right)=$$

$$=m^d\left[\left(1-\frac{1}{m}\right)^d+\frac{1}{m}\left(1-\frac{1}{m}\right)^{(d-1)}\right]=m^d\left(1-\frac{1}{m}\right)^{d-1}\leq m^d.$$

- add 3)

First, we prove the inequality $\Phi_{d,m}\leq 2\frac{m^d}{d!}$. To do it we aplly induction on m and d.

If $d=1$ then $\Phi_{d,m}=m+1\leq 2m$ and inequality holds.

Further, for $m=d>1$ is $\Phi_{d,m}=2^d$. If we recall binomical formulae we can observe that for $d>1$ the expression $2\leq\left(1+\frac{1}{d-1}\right)^{d-1}$ is true. If we apply induction on d we derive

$$2^d\leq\left(\frac{d}{d-1}\right)^{d-1}2^{d-1}\leq\frac{\text{induction}}{\text{assumption}}\leq 2\left(\frac{d}{d-1}\right)^{d-1}\frac{(d-1)^{d-1}}{(d-1)!}=2\frac{d^d}{d!},$$

which concludes the claim for the case $m=d>1$.

Now let us assume that $m>d>1$. Because of $\Phi_{d,m}=\Phi_{d-1,m-1}+\Phi_{d,m-1}$ it is sufficient to verify that

$$2\frac{(m-1)^{d-1}}{(d-1)!}+2\frac{(m-1)^d}{d!}\leq 2\frac{m^d}{d!}.$$

After multiplying by the number $d!$, the previous expression is equivalent to the expressions bellow:

$$\begin{gathered}d(m-1)^{d-1}+(m-1)^d\leq m^d\Leftrightarrow\\(d+m-1)(m-1)^{d-1}\leq m^d\Leftrightarrow\\\tfrac{d+m-1}{m-1}\leq\tfrac{m^d}{(m-1)^d}\Leftrightarrow\\1+\tfrac{d}{m-1}\leq\left(1+\tfrac{1}{m-1}\right)^d,\end{gathered}$$

where the last estimation is based on binomical formula again.

The second estimation $2\frac{m^d}{d!}\leq\left(\frac{em}{d}\right)^d$ is evident for $d=1$. For $d\geq 2$ we use Stirling's formulae

$$n!=n^n\sqrt{2\pi n}\cdot e^{-n+\frac{\widetilde{\delta}(n)}{4}}\,,\quad 0<\widetilde{\delta}(n)<1\,.$$

whose applying yields

$$2\frac{m^d}{d!} < \frac{2m^d}{d^d e^{-d}\sqrt{2\pi d}} = \sqrt{\frac{2}{\pi d}}\left(\frac{em}{d}\right)^d < \left(\frac{em}{d}\right)^d.$$

– *q. e. d.* –

Corollary 2.1.4 *Let* $\bar{X}$ *be a finite set,* $\mathsf{C} \subset 2^{\bar{X}}$ *and* $\mathtt{VC}_{dim}(\mathsf{C}) > 0$. *Then*

$$\mathtt{VC}_{dim}(\mathsf{C}) > \frac{\ln\left(\left|\bar{C}\right|\right)}{1+\ln\left(\left|\bar{X}\right|\right)}.$$

■ *Proof:*
Sauer lemma asserts that $|\mathsf{C}| \le \sum_{i=0}^{\mathtt{VC}_{dim}(\mathsf{C})}\binom{|\bar{X}|}{i}$. Applying last part of 2.1.3, we conclude

$$|\mathsf{C}| \le \left(\frac{e\left|\bar{X}\right|}{\mathtt{VC}_{dim}(\mathsf{C})}\right)^{\mathtt{VC}_{dim}(\mathsf{C})}.$$

In other words,

$$\ln(|\mathsf{C}|) \le \mathtt{VC}_{dim}(\mathsf{C})\left(1+\ln\left(\left|\bar{X}\right|\right)-\ln\left(\mathtt{VC}_{dim}(\mathsf{C})\right)\right) < \mathtt{VC}_{dim}(\mathsf{C})\left(1+\ln\left(\left|\bar{X}\right|\right)\right).$$

– *q. e. d.* –

2.1.2 VC-dimension of union and intersection

Lemma 2.1.5 *Let* Q *concept class over* $\bar{X}$ *and* P *is defined as*

$$\mathsf{P} \stackrel{def}{=} \left\{\bar{a} \subset \bar{X} \,\middle|\, (\exists \bar{b} \in \mathsf{Q})\left(\bar{a} = \bar{X} \dot{-} \bar{b}\right)\right\}.$$

Then $\mathtt{VC}_{dim}(\mathsf{Q}) = \mathtt{VC}_{dim}(\mathsf{P})$.

■ *Proof:*
Let $\bar{z} \subset \bar{X}$ be an arbitrary set shattered by concept class Q, an $\bar{z}_1$, $\bar{z}_2$ be arbitrary dichotomy of $\bar{z}$. Hence, there exists sets $\bar{a} \in \mathsf{Q}$ and $\bar{b} \in \mathsf{Q}$ such that $\bar{z}_1 \subset \bar{a}$, $\bar{z}_1 \cap \bar{b} = \emptyset$ and $\bar{z}_2 \subset \bar{b}$, $\bar{z}_2 \cap \bar{a} = \emptyset$. In other words $\bar{z}_1 \cap \left(\bar{X} \dot{-} \bar{a}\right) = \emptyset$, $\bar{z}_1 \subset \left(\bar{X} \dot{-} \bar{b}\right)$ and $\bar{z}_2 \cap \left(\bar{X} \dot{-} \bar{b}\right) = \emptyset$, $\bar{z}_2 \subset \left(\bar{X} \dot{-} \bar{a}\right)$. It follows that each subset $\bar{z} \subset \bar{X}$ shattered by concept class Q is also shattered by concept class P and vice versa. Therefore $\mathtt{VC}_{dim}(\mathsf{Q}) = \mathtt{VC}_{dim}(\mathsf{P})$.

– *q. e. d.* –

Definition 2.1.4 *Let* $\mathsf{C}_i \subset 2^{\bar{X}}$, $i \in \{1,\dots,k\}$ *are concept classes. Then define the following sets*

$$\mathsf{U}_{\mathsf{C}_1,\dots,\mathsf{C}_k} \stackrel{def}{=} \left\{ \bigcup_{i=1}^{k} \bar{c}_i \,|(\forall i \in \{1,\dots,k\})\,(\bar{c}_i \in \mathsf{C}_i) \right\}$$

and

$$\mathsf{I}_{\mathsf{C}_1,\dots,\mathsf{C}_k} \stackrel{def}{=} \left\{ \bigcap_{i=1}^{k} \bar{c}_i \,|(\forall i \in \{1,\dots,k\})\,(\bar{c}_i \in \mathsf{C}_i) \right\}.$$

Finally denote U_{C} *as the set of all finite unions of subsets* C *and* I_{C} *as the set of all finite intersections of subsets* C*. Further, for each natural number* k *define the sets*

$$\mathsf{U}_{k,\mathsf{C}} \stackrel{def}{=} \left\{ \bigcup_{i=1}^{k} \bar{c}_i \,|(\forall i \in \{1,\dots,k\})\,(\bar{c}_i \in \mathsf{C}) \right\}$$

and

$$\mathsf{I}_{k,\mathsf{C}} \stackrel{def}{=} \left\{ \bigcap_{i=1}^{k} \bar{c}_i \,|(\forall i \in \{1,\dots,k\})\,(\bar{c}_i \in \mathsf{C}) \right\}.$$

Lemma 2.1.6 *Let* C *be a concept class over* $\bar{X}$ *and* $(\forall \bar{a} \in \mathsf{C})\ \left(\bar{X} \dot{-} \bar{a} \in \mathsf{C}\right)$. *Then the equality rovnost*

$$\mathtt{VC}_{dim}\left(\mathsf{U}_{k,\mathsf{C}}\right) = \mathtt{VC}_{dim}\left(\mathsf{I}_{k,\mathsf{C}}\right)$$

holds.

■ *Proof:*
Remember that for all sets the following de Morgan's formulas holds

$$\bar{X} \dot{-} \bigcup_{i=1}^{k} \bar{a}_i = \bigcap_{i=1}^{k} \left(\bar{X} \dot{-} \bar{a}_i\right) \quad \text{and} \quad \bar{X} \dot{-} \bigcap_{i=1}^{k} \bar{a}_i = \bigcup_{i=1}^{k} \left(\bar{X} \dot{-} \bar{a}_i\right).$$

Using these formulas it is easy to prove that $(\forall \bar{a} \in \mathsf{U}_{k,\mathsf{C}})\ (\exists \bar{b} \in \mathsf{I}_{k,\mathsf{C}})\ \left(\bar{a} = \bar{X} \dot{-} \bar{b}\right)$. Obviously the claim of lemma follows from lemma 2.1.5.

– *q. e. d.* –

Main properties of VC-dimension of concept classes union provides next theorems.

Theorem 2.1.7 *Let* $\bar{X}$ *be arbitrary set and* P, Q *are concept classes over* $\bar{X}$*. Then*

$$\mathtt{VC}_{dim}\left(\mathsf{U}_{\mathsf{P},\mathsf{Q}}\right) \leq \mathtt{VC}_{dim}\left(\mathsf{P}\right) + \mathtt{VC}_{dim}\left(\mathsf{Q}\right) + 1.$$

■ *Proof:*
We show this theorem by contradiction. Let $\mathtt{VC}_{dim}\left(\mathsf{P}\right) = m$, $\mathtt{VC}_{dim}\left(\mathsf{Q}\right) = n$, and $\mathtt{VC}_{dim}\left(\mathsf{U}_{\mathsf{P},\mathsf{Q}}\right) = n+m+2$. It follows that there exists a points set $\{x_1,\dots,x_{m+n+2}\} \subset \bar{X}$ which is shattered by concept class $\mathsf{U}_{\mathsf{P},\mathsf{Q}}$, where $x_i \neq x_j$, $i,j \in \{1,\dots,m+n+2\}$. Because of $\mathtt{VC}_{dim}\left(\mathsf{P}\right) = m$

a $\mathtt{VC}_{dim}(\mathsf{Q}) = n$, there exist sets $\bar{a} \subset \{x_1, \ldots, x_{m+1}\}$ and $\bar{b} \subset \{x_{m+2}, \ldots, x_{m+n+2}\}$ such that

$$(\forall \bar{p} \in \mathsf{P})\,(\bar{a} \neq \{x_1, \ldots, x_{m+1}\} \cap \bar{p}) \text{ and } (\forall \bar{q} \in \mathsf{Q})\left(\bar{b} \neq \{x_{m+2}, \ldots, x_{m+n+2}\} \cap \bar{q}\right) .$$

Because the set $\{x_1, \ldots, x_{m+n+2}\}$ is shattered by $\mathsf{U}_{\mathsf{P},\mathsf{Q}}$ there exist sets $\bar{z} \in \mathsf{U}_{\mathsf{P},\mathsf{Q}}$, $\bar{z_{\mathsf{P}}} \in \mathsf{P}$, $\bar{z_{\mathsf{Q}}} \in \mathsf{Q}$, such that

$$\bar{z} = \bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}} \quad \text{and} \quad (\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \{x_1, \ldots, x_{m+n+2}\} = \bar{a} \cup \bar{b} .$$

For the sake of clarity, let us define sets $\bar{A} \stackrel{\text{def}}{=} \{x_1, \ldots, x_{m+1}\}$ and $\bar{B} \stackrel{\text{def}}{=} \{x_{m+2}, \ldots, x_{m+n+2}\}$. As were mentioned before we can write $\bar{A} \cap \bar{B} = \emptyset$, $\bar{a} \subset \bar{A}$, and $\bar{b} \subset \bar{B}$. So

$$\begin{aligned}(\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \{x_1, \ldots, x_{m+n+2}\} &= (\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \left(\bar{A} \cup \bar{B}\right) = \\ &= \left((\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \bar{A}\right) \cup \left((\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \bar{B}\right) = \bar{a} \cup \bar{b} .\end{aligned}$$

Owning to that $\left((\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \bar{B}\right) \dot{-} \bar{B} = \emptyset$ and $\left(\bar{a} \cup \bar{b}\right) \dot{-} \bar{B} = \bar{a}$ we get

$$\left((\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \bar{A}\right) = \bar{a} \quad \text{and using the same argumentation,} \quad \left((\bar{z_{\mathsf{P}}} \cup \bar{z_{\mathsf{Q}}}) \cap \bar{B}\right) = \bar{b} ,$$

or equivalently

$$\left(\bar{z_{\mathsf{P}}} \cap \bar{A}\right) \cup \left(\bar{z_{\mathsf{Q}}} \cap \bar{A}\right) = \bar{a} \quad \text{and} \quad \left(\bar{z_{\mathsf{P}}} \cap \bar{B}\right) \cup \left(\bar{z_{\mathsf{Q}}} \cap \bar{B}\right) = \bar{b} .$$

– *q. e. d.* –

Theorem 2.1.8 *Let* P , Q *be concept classes over an set* $\bar{X}$ *and let* $(\forall \bar{a} \in \mathsf{P})\left(\bar{X} \dot{-} \bar{a} \in \mathsf{P}\right)$. *Then*

$$\mathtt{VC}_{dim}(\mathsf{U}_{\mathsf{P},\mathsf{Q}}) \leq \mathtt{VC}_{dim}(\mathsf{P}) + \mathtt{VC}_{dim}(\mathsf{Q}) .$$

▪ *Proof:*
Assume that P fulfills assumptions of the theorem, e.g. $\bar{p} \in \mathsf{P} \Leftrightarrow \bar{X} \dot{-} \bar{p} \in \mathsf{P}$. Further let $\mathtt{VC}_{dim}(\mathsf{P}) = n$ and $\mathtt{VC}_{dim}(\mathsf{Q}) = m$.

To prove the theorem by contradiction let us assume that there exist mutually different points $\{x_1, \ldots, x_{m+n+1}\}$ that are shattered by concept class $\mathsf{U}_{\mathsf{P},\mathsf{Q}}$. The fact that $\mathtt{VC}_{dim}(\mathsf{Q}) = m$ implies that there exists an set $\bar{a} \subset \{x_1, \ldots, x_{m+1}\}$ which can not be separated from $(\{x_1, \ldots, x_{m+1}\} \dot{-} \bar{a}$ by any concept in Q . Consequently for any set $\bar{Y} \subset \{x_{m+2}, \ldots, x_{m+n+1}\}$ there exists an concept $\bar{p} \in \mathsf{P}$, which separates the set $\bar{a} \cup \bar{Y}$ from the set $\{x_1, \ldots, x_{m+n+1}\} \dot{-} \left(a \cup \bar{Y}\right)$. Now let us take into mind two different cases:
▪ add a)
Let $\bar{a} \neq \emptyset$. Jelikož výše uvedená množina je neprázdná, můžeme z ní vybrat libovolný bod x_0. Potom ovšem pro libovolnou množinu $\bar{Y} \subset \{x_{m+1}, \ldots, x_{m+n+1}\}$ existuje koncept $\bar{p} \in \mathsf{P}$ oddělující $x_0 \cup \bar{Y}$ od $\{x_1, \ldots, x_{m+n+1}\} \dot{-} \left(x_0 \cup \bar{Y}\right)$. Jelikož P obsahuje s každým svým konceptem i jeho doplněk, zajistíme tímto konceptem i opačné oddělení. Jelikož toto platí pro libovolnou množinu $\bar{Y} \subset \{x_{m+2}, \ldots, x_{m+n+1}\}$, je množina $\{x_0, x_{m+1}, \ldots, x_{m+n+1}\}$ (jejíž mohutnost je $n+1$) třídou konceptů P rozdělena, což je v rozporu s tím, že $\mathtt{VC}_{dim}(\mathsf{P}) = n$

▪ add b)
Let $\bar{a} = \emptyset$. Každá množina $\bar{Y} \subset \{x_{m+2}, \ldots, x_{m+n+1}\}$ je odseparovatelná od množiny $\{x_1, \ldots, x_{m+n+1}\} - \bar{Y}$ nějakým konceptem z P. Vezmeme libovolný bod $x_0 \in \{x_1, \ldots, x_{m+1}\}$. Potom pro libovolnou množinu $\bar{Y} \subset \{x_{m+2}, \ldots, x_{m+n+1}\}$ existuje koncept $\bar{p} \in \mathsf{P}$, který tuto množinu odseparuje od množiny $x_0 \cup \{x_{m+1}, \ldots, x_{m+n+1}\} - \bar{Y}$. Dále koncept $\bar{X} - \bar{p}$ zajistí opačnou separaci. So the set $\{x_0, x_{m+2}, \ldots, x_{m+n+1}\}$ is shattered by concept class P, which leads to contradiction.

– *q. e. d.* –

Classes of VC-dimension one

We turn now attention to most simpliest, but still applicable, concept classes of the $\mathtt{VC}_{dim}(\mathsf{C}) = 1$. To sake of completness let us characterize concept classes with $\mathtt{VC}_{dim}(\mathsf{C}) \leq 1$.

Theorem 2.1.9 *Let* C *be nonempty concept class over* $\bar{X}$. *Then*

1. $\mathtt{VC}_{dim}(\mathsf{C}) = 0$ *if and only if* C *contains exactly one set.*
2. *Let one of the following conditions is true:*
   (a) C *is linearly ordered by inclusion, or*
   (b) *any two sets in* C *are disjoint.*

   Then $\mathtt{VC}_{dim}(\mathsf{C}) = 1$

▪ *Proof:*
▪ add 1: $\Rightarrow$)
Let C contains a least two different sets $\bar{a}$, $\bar{b}$. Without loss of generality we can assume that there exists an point $z \in \bar{a}$ such that $z \notin \bar{b}$. In this case $\bar{a} \cap \{z\} = z$ and $\bar{b} \cap \{z\} = \emptyset$. So we have $\mathtt{VC}_{dim}(\mathsf{C}) = 1$ which is contradiction.
▪ add 1: $\Leftarrow$)
In opposite, let be $\mathsf{C} \stackrel{\text{def}}{=} \{\{\bar{a}\}\}$. Let there exists $y \in \bar{X}$ which is shattered by C. If $y \in \bar{a}$, then $y \cap \bar{a} \neq \emptyset$, and $\{y\}$ is not shattered by C. If $y \notin \bar{a}$, then $y \cap \bar{a} \neq \{y\}$, and resembly $\{y\}$ is not shattered by C. Hence no one-point subset of $\bar{X}$ can be shattered by C.
▪ add 2: a)
Let $\{a, b\} \subset \bar{X}$, $a \neq b$, be shattered by C. So there exist sets $\bar{c}_1, \bar{c}_2 \in \mathsf{C}$ such that $\bar{c}_1 \cap \{a, b\} = \{b\}$ and $\bar{c}_2 \cap \{a, b\} = \{a\}$. But C is linearly ordered by inclusion, hence without loss of generality, $\bar{c}_1 \subset \bar{c}_2$. It follows $\bar{c}_2 \cap \{a, b\} = \{a, b\}$ and we have a contradiction.
▪ add 2: b)
Let $\{a, b\} \subset \bar{X}$, $a \neq b$, be shattered by C. So there exist sets $\bar{c}_1, \bar{c}_2 \in \mathsf{C}$ such that $\bar{c}_1 \cap \{a, b\} = \{b\}$ and $\bar{c}_2 \cap \{a, b\} = \{a, b\}$. Hence $\bar{c}_1 \cap \bar{c}_2 = \{b\}$ which contradicts assumptions b) of the second part of the theorem.

– *q. e. d.* –

Theorem 2.1.10 *Let* $\bar{X}$ *be an arbitrary set,* $\mathsf{C} \subset 2^X$, *and* $\mathtt{VC}_{dim}(\mathsf{C}) = 1$. *Then* $\mathtt{VC}_{dim}(\mathsf{U}_{n,\mathsf{C}}) \leq n$.

■ *Proof:*
Let $\mathtt{VC}_{dim}(\mathsf{U}_{n,\mathsf{C}}) = n+1$. Hence there exists an set $\bar{A} \subset \bar{X}$, $\left|\bar{A}\right| = n+1$, which is shaterred by the concept class $\mathsf{U}_{n,\mathsf{C}}$. So there exists an $\bar{c} \in \mathsf{U}_{n,\mathsf{C}}$ in such a way that $\bar{c} \cap \bar{A} = \bar{A}$. Because of the set $\bar{c}$ is union of n sets, $\bar{c} = \bar{c}_1 \cup \cdots \cup \bar{c}_n$, and the size of the set $\bar{A}$ is $n+1$, there exists an index j, such that $\left|\bar{c}_j \cap \bar{A}\right| \geq 2$. Therefore there exists $a, b \in \bar{c}_j$ such that $a \neq b$, $a, b \in \bar{A}$. In accordance with the fact that $\bar{A}$ is shattered by $\mathsf{U}_{n,\mathsf{C}}$, the set $\{a, b\}$ is shattered by concept class $\mathsf{U}_{n,\mathsf{C}}$ too. Because of $\mathtt{VC}_{dim}(\mathsf{C}) = 1$ the set $\{a, b\}$ is not shattered by concept class C, so

$$\left(\exists \bar{g} \in 2^{\{a,b\}}\right)(\forall \bar{c} \in \mathsf{C})\left(\bar{c} \cap \{a, b\} \neq \bar{g}\right) . \tag{2.3}$$

In addition $\bar{g} \neq \{a, b\}$ due to $a, b \in \bar{c}_j$. Hence $\bar{g} \in \{\emptyset, \{a\}, \{b\}\}$. Let $\bar{c}_1 \cup \cdots \cup \bar{c}_n \in \mathsf{U}_{n,\mathsf{C}}$ be arbitrary. So we have (accordingly to 2.3)

$$(\bar{c}_1 \cup \cdots \cup \bar{c}_n) \cap \{a, b\} = (\bar{c}_1 \cap \{a, b\}) \cup \cdots \cup (\bar{c}_n \cap \{a, b\}) \neq \bar{g}$$

because $\bar{g}$ is empty set or one-point set. This contradicts that $\{a, b\}$ is shattered by $\mathsf{U}_{n,\mathsf{C}}$.

– *q. e. d.* –

Theorem 2.1.11 *Let* $\bar{X}$ *be an arbitrary set,* $\mathsf{C}_i \subset 2^X$, $i \in \{1, \ldots, n\}$, *and set systems* C_i *are linearly ordered by inclusion. Then* $\mathtt{VC}_{dim}(\mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n}) \leq n$.

■ *Proof:*
Let $\mathtt{VC}_{dim}(\mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n}) = n+1$. Hence there exists an set $\bar{A} \subset \bar{X}$, $\left|\bar{A}\right| = n+1$, which is shaterred by the concept class $\mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n}$. Further, let us define the following sets:

$$\mathsf{H} = \left\{\bar{h} \,\middle|\, (\exists a \in \bar{A})\left(\bar{h} = \bar{A} \dot{-} \{a\}\right)\right\} \quad , \quad \mathsf{G}_i \stackrel{\text{def}}{=} \left\{\bar{A} \cap \bar{c}_j \,|\bar{c}_j \in \mathsf{C}_i,\ j \geq 1\right\} \quad ,$$

$$\mathsf{B} \stackrel{\text{def}}{=} \left\{\bar{b} \,|(\exists \bar{c} \in \mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n})\left(\bar{b} = \bar{A} \cap \bar{c} \quad \text{and} \quad \left|\bar{b}\right| = n\right)\right\} .$$

Straightforwardly, we have

1. The set H contains all subsets $\bar{h}$ of the set $\bar{A}$ with $\left|\bar{h}\right| = n$ and $|\mathsf{H}| = n+1$.
2. Each G_i, $i \in \{1, \ldots, n\}$, is a finite set system linearly ordered by inclusion.
3. For any $i, j \in \{1, \ldots, n\}$ there exist only one set $\bar{e} \in \mathsf{G}_i$ such that $|e| = j$.
4. Because $\bar{A}$ is shattered by $\mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n}$ the equality of collection sets $\mathsf{H} = \mathsf{B}$ must be true. Hence $|\mathsf{B}| = n+1$.

Now let $\bar{b} \in \mathsf{B}$ be arbitrary. It follows, that $\bar{b} = \bar{A} \cap \bar{c}_1 \cap \cdots \cap \bar{c}_n$, where $\bar{c}_i \in \mathsf{C}_i$, $i \in \{1, \ldots, n\}$. Obviously $(\forall i \in \{1, \ldots, n\})\left(\bar{b} \subset \bar{c}_i\right)$. Let us assume, for a while, that $(\forall i \in \{1, \ldots, n\})\left(\bar{A} \dot{-} \bar{b} \in \bar{c}_i\right)$ (note, that $\bar{A} \dot{-} \bar{b}$ is one point set). But this imply that $\bar{A} = \bar{A} \cap \bar{c}_1 \cap \cdots \cap \bar{c}_n = \bar{b}$, where $\bar{c}_i \in \mathsf{C}_i$, which is contradiction $(\bar{b} \neq \bar{A})$. So there must exists $j \in \{1, \ldots, n\}$ such that $\bar{A} \cap \bar{c}_j = \bar{b}$, where $\bar{c}_j \in \mathsf{C}_j$, e.g. $\bar{b} \in \mathsf{G}_j$. As we mentioned in the item 3) on the property list above, such $\bar{b}$ with cardinality n is unique in set system G_j. Because we have only n set systems G_j available, we get that $|\mathsf{B}| = n$, which contradicts item 4) of list mentioned. Confessedly the set $\bar{B}$ can not be shattered by the system $\mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n}$ and $\mathtt{VC}_{dim}\left(\mathsf{I}_{\mathsf{C}_1,\ldots,\mathsf{C}_n}\right) \leq n$.

– *q. e. d.* –

Lemma 2.1.12 *Let $\mathsf{C} \subset 2^{\bar{X}}$ be a concept class, let $\mathtt{VC}_{dim}(\mathsf{C}) = d \geq 1$ be finite, and $k \geq 1$. Then $\mathtt{VC}_{dim}\left(\mathsf{U}_{k,\mathsf{C}}\right) \leq 2dk\log_2(3k)$ and $\mathtt{VC}_{dim}\left(\mathsf{I}_{k,\mathsf{C}}\right) \leq 2dk\log_2(3k)$.*

■ *Proof:*
We prove the statement of the lemma for union only, the case of intersection is resemble. The proof for the case $k = 1$ is trivial so let us assume that $k \geq 2$ and let $\bar{T} \subset \mathsf{X}$ be a finite set, $\left|\bar{T}\right| = m \geq 1$. Accordance with lemma 2.1.3, $\Pi_{\mathsf{C}}\left(\left|\bar{T}\right|\right) \leq \Phi_{d,m}$. Further, every set in the system $\Pi_{\mathsf{U}_{\mathsf{k},\mathsf{C}}}\left(\bar{T}\right)$ is of the form $\bigcup_{i=1}^{k} \bar{a}_i$, where $\bar{a}_i \in \Pi_{\mathsf{C}}\left(\bar{T}\right)$, $1 \leq i \leq k$. It follows that

$$\left|\Pi_{\mathsf{U}_{\mathsf{k},\mathsf{C}}}\left(\bar{T}\right)\right| \leq \left|\Pi_{\mathsf{C}}\left(\bar{T}\right)\right|^k \leq \left(\Phi_{d,m}\right)^k.$$

If $\left(\Phi_{d,m}\right)^k < 2^m$, then $\bar{T}$ can not be shattered by $\mathsf{U}_{k,\mathsf{C}}$ and $\mathtt{VC}_{dim}\left(\mathsf{U}_{k,\mathsf{C}}\right)$ is less than m. So, using statement of the 2.1.3 it suffices to show that $\left(\frac{em}{d}\right)^{dk} < 2^m$ for $m = 2dk\log_2(3k)$, which is equivalent to the expression $\log_2(3k) < \frac{9k}{2e}$. This inequality is clearly satisfied for the value $k = 2$ and it is obvious that is satisfied for any k greater.

– *q. e. d.* –

2.2 VC-dimension of linear concepts

The very useful tool which can be used to construct an upper bound estimation of VC-dimension of the concept of halfspaces in Euclidean space $\Re^n$ is the Radon's lemma (see. [Lej85]) which proof is based on the notion of affine independent [1] vector set.

Lemma 2.2.2 (Radon) *Let $\bar{S} \stackrel{def}{=} \{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_k\} \subset \Re^n$, $k \geq n+2$, $\vec{\boldsymbol{x}}_i$ are mutually different. Then there exists sets $\bar{S}_1$ and $\bar{S}_2$ such that $\bar{S}_1 \cup \bar{S}_2 = \bar{S}$, $\bar{S}_1 \cap \bar{S}_2 = \emptyset$ and*

$$[\bar{S}_1]_\kappa \cap [\bar{S}_2]_\kappa \neq \emptyset,$$

[1] Let $\bar{Y}$ be a vector space of the dimension d. Then a linear combination $\sum_{i=1}^{n} \alpha_i \vec{\boldsymbol{x}}_i$ of arbitrary finite number of vectors $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_n \in \bar{Y}$ is called affine iff $\sum_{i=1}^{n} \alpha_i = 1$. The set $\bar{A}$ of all affine combinations of the vectors $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_n$ forms affine hull of vectors $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_n$. Affinne hull is in fact affine subspace of $\bar{Y}$. Let $\vec{\boldsymbol{v}} \in \bar{A}$. Then the set $\bar{P} \stackrel{\text{def}}{=} \left\{\vec{\boldsymbol{x}} \in \bar{Y} \,\middle|\, (\exists \vec{\boldsymbol{a}} \in \bar{A})\,(\vec{\boldsymbol{x}} = \vec{\boldsymbol{a}} - \vec{\boldsymbol{v}})\right\}$ is a linear subspace of the space $\bar{Y}$. The dimension of this subspace is called affine rank of the system $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_n$. Obviously, the affine rank of any vector set in d-dimensional vector space can be at most d. The system $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_n$ is called affine independent iff its affine rank is equal to the number $n-1$. So, any $(d+2)$-tuple of vectors in d dimensional space is affine dependent. The following theorem holds:

(the symbol $[\bar{S}]_\kappa$ denotes a convex hull of the set $\bar{S}$).

■ *Proof:*
The maximal number of afine independent vectors in linear space of dimension n is equal to $n+1$. Hence, any set of $n+2$ vectors is afine dependent. Therefore there exist numbers $\alpha_1, \ldots, \alpha_{n+2}$ such that

$$\sum_{i=1}^{n+2} \alpha_i \cdot \vec{\boldsymbol{x}}_i = \vec{\mathbf{0}}, \quad \text{and} \quad \sum_{i=1}^{n+2} \alpha_i = 0.$$

In addition, there exists $j \in \{1, \ldots, n+2\}$ satisfying $\alpha_j \neq 0$. Without loos of generality we can assume that $\alpha_j > 0$. Further let us define a dichotomy of the set $\bar{S}$ as

$$\bar{S}_1 \stackrel{\text{def}}{=} \left\{ \vec{\boldsymbol{x}}_i \in \bar{S} \,|\alpha_i > 0 \right\} \quad \text{and} \quad \bar{S}_2 \stackrel{\text{def}}{=} \left\{ \vec{\boldsymbol{x}}_i \in \bar{S} \,|\alpha_i \leq 0 \right\},$$

($\sum_{i=1}^{n+2} \alpha_i = 0$ and $\alpha_j > 0$ implies that both of the sets $\bar{S}_1, \bar{S}_2$ are nonempty). Now we define

$$\omega \stackrel{\text{def}}{=} \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_1\}} \alpha_i \,, \quad \vec{\boldsymbol{z}}_1 \stackrel{\text{def}}{=} \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_1\}} \frac{\alpha_i}{\omega} \cdot \vec{\boldsymbol{x}}_i \,, \quad \vec{\boldsymbol{z}}_2 \stackrel{\text{def}}{=} \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} \frac{-\alpha_i}{\omega} \cdot \vec{\boldsymbol{x}}_i \,.$$

The fact that

$$0 = \sum_{i=1}^{n+2} \alpha_i = \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_1\}} \alpha_i + \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} \alpha_i = \omega + \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} \alpha_i$$

follows $\omega = -\sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} \alpha_i$. Hence

$$1 = \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_1\}} \frac{\alpha_i}{\omega} = \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} \frac{-\alpha_i}{\omega}.$$

So, it imply that $\vec{\boldsymbol{z}}_1 \in [\bar{S}_1]_\kappa$ and $\vec{\boldsymbol{z}}_2 \in [\bar{S}_2]_\kappa$. Further

$$\sum_{i=1}^{n+2} \alpha_i \cdot \vec{\boldsymbol{x}}_i = \vec{\mathbf{0}} \;\Rightarrow\; \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_1\}} \alpha_i \cdot \vec{\boldsymbol{x}}_i = \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} -\alpha_i \cdot \vec{\boldsymbol{x}}_i \;\Rightarrow\; \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_1\}} \frac{\alpha_i}{\omega} \cdot \vec{\boldsymbol{x}}_i = \sum_{\{i|\vec{\boldsymbol{x}}_i \in \bar{S}_2\}} \frac{-\alpha_i}{\omega} \cdot \vec{\boldsymbol{x}}_i \,.$$

So finally $\vec{\boldsymbol{z}}_1 = \vec{\boldsymbol{z}}_2$ and $[\bar{S}_1]_\kappa \cap [\bar{S}_2]_\kappa \neq \emptyset$.

– *q. e. d.* –

We are going to use Radon lemma to prove the following important fact.

Theorem 2.2.3 $\mathtt{VC}_{dim}\left(\mathbf{HALFSPACE}_n\right) = \mathtt{VC}_{dim}\left(\mathbf{BALL}_n\right) = n+1.$

Theorem 2.2.1 *Vectors $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_n$ are affine independent iff*

$$(\forall \alpha_1, \ldots, \alpha_n) \left[\left(\sum_{i=1}^{n} \alpha_i \vec{\boldsymbol{x}}_i = \vec{\mathbf{0}} \quad a \quad \sum_{i=1}^{n} \alpha_i = 0 \right) \quad \textit{if and only if} \quad \alpha_1 = \alpha_2 = \ldots = \alpha_n = 0 \right].$$

■ *Proof:*

▪ add HALFSPACE)

First of all let $\{\vec{\boldsymbol{x}}_1,\dots,\vec{\boldsymbol{x}}_k\}\subset\Re^n$, $\vec{\boldsymbol{a}}\in\Re^n$, $t\in\Re$. Let a vector $\vec{\boldsymbol{y}}$ is defined as $\vec{\boldsymbol{y}}\stackrel{\text{def}}{=}\sum_{i=1}^k\alpha_i\vec{\boldsymbol{x}}_i$, where $\sum_{i=1}^k\alpha_i=1$ and $(\forall i\in\{1,\dots,k\})\,(\alpha_i\geq 1)$. Let in addition $\langle\vec{\boldsymbol{x}}_i\,|\vec{\boldsymbol{a}}\rangle<t$, $i\in\{1,\dots,k\}$. Then

$$\langle\vec{\boldsymbol{y}}\,|\vec{\boldsymbol{a}}\rangle=\sum_{i=1}^k\alpha_i\langle\vec{\boldsymbol{x}}_i\,|\vec{\boldsymbol{a}}\rangle<\sum_{i=1}^k\alpha_i\cdot t=t\,. \tag{2.4}$$

Now we show that $\mathtt{VC}_{dim}\left(\mathbf{HALFSPACE}_n\right)\leq n+1$. Assume contradiction, so put $\mathtt{VC}_{dim}\left(\mathbf{HALFSPACE}_n\right)>n+1$. In other words, $\mathbf{HALFSPACE}_n$ shatters an set $\bar{S}$ containing at least $n+2$ points. Claim of Radon lemma implies that there exists two disjoint sets $\bar{S}_1$ a $\bar{S}_2$ such that $\bar{S}_1\cup\bar{S}_2=\bar{S}$, a $[\bar{S}_1]_\kappa\cap[\bar{S}_2]_\kappa\neq\emptyset$. At the same time, $\bar{S}$ is shattered by $\mathbf{HALFSPACE}_n$, hence there exists a hyperplane separating sets $\bar{S}_1$ and $\bar{S}_2$. But such hyperplane must separate whole convex hulls of $\bar{S}_1$ and $\bar{S}_2$ (see 2.4) which contradicts Radons lemma.

We finish the proof by construction of a set of $n+1$ points, which is shattered by $\mathbf{HALFSPACE}_n$. Let $\bar{S}\stackrel{\text{def}}{=}\left\{\vec{\mathbf{0}},\vec{\boldsymbol{e}}_i,\dots,\vec{\boldsymbol{e}}_n\right\}$, where $\vec{\mathbf{0}}$ is zero vector and $\vec{\boldsymbol{e}}_i$ je i-th vector of standard basis. Let $\bar{S}_1$ denote arbitrary subset of the set $\bar{S}$. Now we can define vector $\vec{\boldsymbol{\alpha}}\in\Re^n$ and number t as:

$$\vec{\boldsymbol{\alpha}}_i\stackrel{\text{def}}{=}\begin{cases}1 & \text{if}\quad\vec{\boldsymbol{x}}_i\in\bar{S}_1\\ -1 & \text{if}\quad\vec{\boldsymbol{x}}_i\notin\bar{S}_1\end{cases}\quad\text{a}\quad t\stackrel{\text{def}}{=}\begin{cases}\frac{1}{2} & \text{if}\quad\vec{\mathbf{0}}\in\bar{S}_1\\ -\frac{1}{2} & \text{if}\quad\vec{\mathbf{0}}\notin\bar{S}_1.\end{cases}$$

Apparently a hyperplane $\{\vec{\boldsymbol{x}}\,|\langle\vec{\boldsymbol{x}}\,|\vec{\boldsymbol{\alpha}}\rangle-t=0\}$ in the space $\Re^n$ separate sets $\bar{S}\dot{-}\bar{S}_1$ and $\bar{S}_1$. Since the set $\bar{S}_1$ was chosen arbitrarily, we get

$$\mathtt{VC}_{dim}\left(\mathbf{HALFSPACE}_n\right)=n+1.$$

▪ add BALL)

Let us assume a contradiction, so put $\mathtt{VC}_{dim}\left(\mathbf{BALL}_n\right)>n+1$. In this case there exists $\bar{S}=\{\vec{\boldsymbol{x}}_1,\dots,\vec{\boldsymbol{x}}_{n+2}\}\subset\Re^n$ which is shattered by the concept class $\mathbf{BALL}_n$. As in the previous case, the Radon lemma implies that there exists two disjoint sets $\bar{S}_1$ a $\bar{S}_2$ such that $\bar{S}_1\cup\bar{S}_2=\bar{S}$, a $[\bar{S}_1]_\kappa\cap[\bar{S}_2]_\kappa\neq\emptyset$. Because $\bar{S}$ is shattered by $\mathbf{BALL}_n$, $\left(\exists\bar{B}_1,\bar{B}_2\in\mathbf{BALL}_n\right)\left(\bar{S}_1=\bar{B}_1\cap\bar{S}\text{ and }\bar{S}_2=\bar{B}_2\cap\bar{S}\right)$. Let us assume that there exists $j\in\{1,\dots,n+2\}$ such that $\vec{\boldsymbol{x}}_j\in\bar{B}_1\cap\bar{B}_2$. Than $\vec{\boldsymbol{x}}_j\in\bar{S}_1$ and at the same time $\vec{\boldsymbol{x}}_j\in\bar{S}_2$, which contradicts the fact that $\bar{S}_1\cap\bar{S}_2=\emptyset$. So $\bar{S}\cap\left(\bar{B}_1\cap\bar{B}_2\right)=\emptyset$. Let

$$\bar{B}_1=\{\vec{\boldsymbol{x}}\in\Re^n\,|\,\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{c}}_1\|_E\leq r_1\}\quad\text{and}\quad\bar{B}_2=\{\vec{\boldsymbol{x}}\in\Re^n\,|\,\|\vec{\boldsymbol{x}}-\vec{\boldsymbol{c}}_2\|_E\leq r_2\}\,,$$

where $\vec{\boldsymbol{c}}_1,\vec{\boldsymbol{c}}_2\in\Re^n$ and $r_1,r_2\in\Re^+$ are fixed. Further let us express the intersection $\bar{I}$ of surfaces of the sets $\bar{B}_1,\bar{B}_2$ as

$$\bar{I}\stackrel{\text{def}}{=}\left\{\vec{\boldsymbol{y}}\in\Re^n\,\middle|\,\sum_{i=1}^n\vec{\boldsymbol{y}}_i^2-2\vec{\boldsymbol{y}}_i\left(\vec{\boldsymbol{c}}_1\right)_i+\left(\vec{\boldsymbol{c}}_1\right)_i^2=r_1\quad\text{and}\quad\sum_{i=1}^n\vec{\boldsymbol{y}}_i^2-2\vec{\boldsymbol{y}}_i\left(\vec{\boldsymbol{c}}_2\right)_i+\left(\vec{\boldsymbol{c}}_2\right)_i^2=r_2\right\}.$$

Now let us take into mind two possible situations,

1. $\bar{I} \neq \emptyset$:
   In this case each vektor $\vec{\boldsymbol{y}} \in \bar{I}$ satisfy

$$\sum_{i=1}^{n} 2\vec{\boldsymbol{y}}_i \left((\vec{\boldsymbol{c}}_2)_i - (\vec{\boldsymbol{c}}_1)_i \right) = r_1 - r_2 - \sum_{i=1}^{n} \left((\vec{\boldsymbol{c}}_1)_i^2 - (\vec{\boldsymbol{c}}_2)_i^2 \right) \ .$$

   Obviously it imply that

$$\bar{I} \subset \left\{ \vec{\boldsymbol{x}} \in \Re^n \,\middle|\, \langle \vec{\boldsymbol{x}} \,|\vec{\boldsymbol{c}}_2 - \vec{\boldsymbol{c}}_1 \rangle = \frac{r_1 - r_2 - \sum_{i=1}^{n} \left((\vec{\boldsymbol{c}}_1)_i^2 - (\vec{\boldsymbol{c}}_2)_i^2 \right)}{2} \right\} \stackrel{\text{def}}{=} \bar{H} \ .$$

   The set $\bar{H}$ is a hyperplane in the space $\Re^n$ and this hyperplane separate sets $\bar{B}_1 \dot{-} \bar{B}_2$ and $\bar{B}_2 \dot{-} \bar{B}_1$, therefore it separates also the sets $\bar{S}_1$ and $\bar{S}_2$ which contradicts Radon's lemma. So we have proved that $\mathtt{VC}_{dim}\left(\mathbf{BALL}_n\right) \leq n+1$.

2. $\bar{I} = \emptyset$:
   Obviously any hyperplane separate the sets $\bar{B}_1, \bar{B}_2$ and therefore especially the sets $\bar{S}_1, \bar{S}_2$ which contradicts Radon's lemma as in previous case.

Now let the set $\bar{S} = \{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_{n+1}\} \subset \Re^n$ be any set shattered by concept class $\mathbf{HALFSPACE}_n$. Let $\bar{S}_1, \bar{S}_2$ be any dichotomy of the set $\bar{S}$ and let the hyperplane $\{\vec{\boldsymbol{x}} \in \Re^n \,|\langle \vec{\boldsymbol{x}} \,|\vec{\boldsymbol{w}} \rangle = t\}$ separates the sets $\bar{S}_1, \bar{S}_2$. Put $\vec{\boldsymbol{z}} \stackrel{\text{def}}{=} t \cdot \frac{\vec{\boldsymbol{w}}}{\|\vec{\boldsymbol{w}}\|_E^{\,2}}$. Clearly $\vec{\boldsymbol{z}}$ lies on hyperplane and the sets $\bar{B}_1, \bar{B}_2 \in \mathbf{BALL}_n$ defined as

$$\bar{B}_1 = \left\{ \vec{\boldsymbol{x}} \in \Re^n \,\middle|\, \left\| \vec{\boldsymbol{x}} - \left(\vec{\boldsymbol{z}} - r \frac{\vec{\boldsymbol{w}}}{\|\vec{\boldsymbol{w}}\|_E} \right) \right\|_E \leq r \right\}$$

and

$$\bar{B}_2 = \left\{ \vec{\boldsymbol{x}} \in \Re^n \,\middle|\, \left\| \vec{\boldsymbol{x}} - \left(\vec{\boldsymbol{z}} + r \frac{\vec{\boldsymbol{w}}}{\|\vec{\boldsymbol{w}}\|_E} \right) \right\|_E \leq r \right\}$$

separates the sets $\bar{S}_1, \bar{S}_2$ too for sufficiently large value of radius $r > 0$ because the set $\bar{S}$ is finite and therefore bounded.

– q. e. d. –

2.2.1 Application of Cover's lemma

Now we turn our attention to Cover's lemma. This lemma enumerate the number of dichotomies whose can be obtained by intersection of homogeneous halfspaces with a finite sets of points. To point out the main idea of its proof we start with the following two lemmas.

Definition 2.2.1 *Let $\bar{A}, \bar{B} \subset \Re^N$ be given. Then a vector $\vec{\boldsymbol{w}} \in \Re^N$ is* HOMOGENOUS LINEAR SEPARATOR *of sets $\bar{A}$ and $\bar{B}$ iff $\left(\forall \vec{\boldsymbol{a}} \in \bar{A}\right) \left(\langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{a}} \rangle > 0\right)$ and at the same time $\left(\forall \vec{\boldsymbol{b}} \in \bar{B}\right) \left(\left\langle \vec{\boldsymbol{w}} \,\middle|\, \vec{\boldsymbol{b}} \right\rangle < 0\right)$. In this case, the sets $\bar{A}$ and $\bar{B}$ are called* HOMOGENOUS LINEARLY SEPARABLE SETS . *The set of all homogenously linearly separable tuples of sets $\left(\bar{A}, \bar{B}\right)$ will be denoted by the symbol $\mathbf{HLS}_{sets}$.*

Lemma 2.2.4 ([Cov]) *Let $\bar{A}, \bar{B} \subset \Re^N$ be finite sets, $\vec{y} \in \Re^N$, $\vec{y} \neq \vec{0}$. Than pair of sets $\left\{\bar{A}, \bar{B} \cup \{\vec{y}\}\right\}$ and $\left\{\bar{A} \cup \{\vec{y}\}, \bar{B}\right\}$ are simultaneously homogeneously linearly separable iff there exists hyperplane, containing zero vector and vector $\vec{y}$, which separate sets $\bar{A}$ and $\bar{B}$.*

∎ *Proof:*
Let us denote $\bar{W} \stackrel{\text{def}}{=} \left\{\vec{w} \in \Re^n \middle| \left(\forall \vec{a} \in \bar{A}\right) \left(\langle \vec{w} \,|\vec{a}\rangle < 0\right) \text{ and } \left(\forall \vec{b} \in \bar{B}\right) \left(\left\langle \vec{w} \middle| \vec{b} \right\rangle > 0\right)\right\}$. Then the sets $\bar{A}$ a $\{\vec{y}\} \cup \bar{B}$ are homogenously linearly separable if and only if there exists a vector $\vec{w}_a \in \bar{W}$ such that $\langle \vec{w}_a \,|\vec{y}\rangle < 0$ and the sets $\{\vec{y}\} \cup \bar{A}$ and $\bar{B}$ are homogenously linearly separable if and only iff there exists $\vec{w}_b \in \bar{W}$ such that $\langle \vec{w}_b \,|\vec{y}\rangle > 0$. Hence if we define a vector $\vec{w}_{ab} \stackrel{\text{def}}{=} \langle \vec{w}_a \,|\vec{y}\rangle \vec{w}_b - \langle \vec{w}_b \,|\vec{y}\rangle \vec{w}_a$ than it is straightforward that the sets $\bar{A}$ and $\bar{B}$ are homogenously linearly separable by hyperplane $\{\vec{x} \in \Re^n \,|\langle \vec{x} \,|\vec{w}_{ab}\rangle = 0\}$.

Now prove the opposite implication. Let the sets $\bar{A}$ and $\bar{B}$ be homogenously linearly separate by hyperplanes containing a vector $\vec{y}$. Then there exists $\vec{w}_{\vec{y}} \in \bar{W}$ in such a way that $\left\langle \vec{w}_{\vec{y}} \middle| \vec{y} \right\rangle = 0$. By reason that $\bar{W}$ is an open set, there exists a number $\epsilon > 0$ such that the vectors $\vec{w}_{\vec{y}} + \epsilon\vec{y}$ and $\vec{w}_{\vec{y}} - \epsilon\vec{y}$ belong to the set $\bar{W}$. Hence the tuple of sets $\left\{\bar{A}, \bar{B} \cup \{\vec{y}\}\right\}$ and $\left\{\bar{A} \cup \{\vec{y}\}, \bar{B}\right\}$ are both simultaneously homogenous linearly separable by hyperplanes passed by vectors $\vec{w}_{\vec{y}} + \epsilon\vec{y}$ and $\vec{w}_{\vec{y}} - \epsilon\vec{y}$.

– *q. e. d.* –

The following lemma will be useful in explanation of main step of Cover's lemma proof.

Lemma 2.2.5 *Let $\{\vec{x}_1, \ldots, \vec{x}_k\}$ is an set of linearly independent vectors of $\Re^n$ and let following vectors are defined as:*

$$\vec{y}_i \stackrel{def}{=} \vec{x}_i|_{\vec{x}_k^{\perp}}, \ \forall i \in \{1, \ldots, k-1\},$$
$$\vec{y}_k \stackrel{def}{=} \vec{x}_k.$$

Then vectors $\{\vec{y}_1, \ldots, \vec{y}_k\}$ are linearly independent.

∎ *Proof:*
We prove this lemma via contradiction. Let $\{\vec{y}_1, \ldots, \vec{y}_k\}$ be linearly dependent vectors. As for all $i \in \{1, \ldots, k-1\}$ holds $\vec{y}_k \perp \vec{y}_i$ vectors $\left\{\vec{y}_1, \ldots, \vec{y}_{k-1}\right\}$ must be linearly dependent. Hence there exist nontrivial linear combination

$$\sum_{i=1}^{k-1} \alpha_i \vec{y}_i = \vec{0}.$$

Therefore vectors $\vec{y}_i$ are defined as orthogonal projection along vector $\vec{x}_k$ it holds that

$$\vec{y}_i = \vec{x}_i - \beta_i \vec{x}_k.$$

It implies that

$$\sum_{i=1}^{k-1} \alpha_i . \vec{x}_i - \sum_{i=1}^{k-1} \alpha_i \beta_i \vec{x}_k = \vec{0}.$$

But $\vec{\boldsymbol{y}}_i \neq \vec{\mathbf{0}}$, $i \in \{1, \ldots, k-1\}$, hence there exist at least two nonzero α_i which implies contradiction with linear independency of vectors $\{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_k\}$.

– *q. e. d.* –

Theorem 2.2.6 (Cover, [Cov]) *Let* $\bar{X} \stackrel{def}{=} \{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_d\} \subset \Re^N$ *are linearly independent vectors. Than there exists*

$$2\sum_{k=0}^{d-1} \binom{N-1}{k}$$

mutually different disjoint splittings of the set $\bar{X}$ *into sets* $\bar{A}$ *and* $\bar{B}$ *whose are homogeneously linearly separable (i.e. they can be separated via hyperplane which contains zero vector).*

■ *Proof:*
We prove this theorem by induction on the number of vectors d and dimension N. Let us denote the number of mutually different homogenously separable disjoint splittings of the set $\bar{X}$ by the symbol $\widetilde{C}(d, N)$. Furthe let us assume that the set $\bar{Y} \stackrel{\text{def}}{=} \bar{X} \cup \{\vec{\boldsymbol{x}}_{d+1}\}$ fulfills the assumption of the theorem proved and that $\bar{A}$, $\bar{B}$ be arbitrary disjoint splitting of the set $\bar{Y}$. Because of the set $\bar{X}$ is finite and linear homogenous separator of the set is defined via nonstrict inequality, at least one ot the tuples $\left(\bar{A} \cup \{\vec{\boldsymbol{x}}_{d+1}\}, \bar{B}\right)$ and $\left(\bar{A}, \bar{B} \cup \{\vec{\boldsymbol{x}}_{d+1}\}\right)$ are homogenouly linearly separable sets. In addition, it can occur that both of these set tuples are simultaneously homogenously linearly separable, therefore it is reasonable to define the following set:

$$\bar{S} \stackrel{\text{def}}{=} \left\{ \bar{A} \subset X \,\middle|\, \bar{A} = \bar{X} \dot{-} \bar{B} \;\wedge\; \left(\bar{A} \cup \{\vec{\boldsymbol{x}}_{d+1}\}, \bar{B}\right) \in \mathbf{HLS}_{sets} \;\wedge\; \left(\bar{A}, \bar{B} \cup \{\vec{\boldsymbol{x}}_{d+1}\}\right) \in \mathbf{HLS}_{sets} \right\}.$$

So if we recapitulate the previous explanation, we see that the number $\widetilde{C}(d+1, N)$ can be expressed as the sum of $\widetilde{C}(d, N)$ and the number of simultaneously linearly separable dichotomies $\left\{\bar{A} \cup \{\vec{\boldsymbol{x}}_{d+1}\}, \bar{B}\right\}$ and $\left\{\bar{A}, \bar{B} \cup \{\vec{\boldsymbol{x}}_{d+1}\}\right\}$ of the set $\bar{X} \cup \{\vec{\boldsymbol{x}}_{d+1}\}$, e.g. $\left|\bar{S}\right|$. Applying the statement of the lemma 2.2.4 we deduce that the number of such set tuples $\bar{A}$, $\bar{B}$ is equal to the number of tuples $\bar{A}$, $\bar{B}$, that are homogenously linearly separable by hyperplane containing the vector $\vec{\boldsymbol{x}}_{d+1}$. Apparently the sets $\bar{A}$, $\bar{B}$ are separated by such a hyperplane if and only if the sets $\bar{A}'$, $\bar{B}'$ are linearly homogenously separated, where $\bar{A}'$ and $\bar{B}'$ are orthogonal projection of the sets $\bar{A}$ and $\bar{B}$ into the orthogonal complement of the vector $\vec{\boldsymbol{x}}_{d+1}$.

Owning to the lemma 2.2.5 the sets $\bar{A}'$, $\bar{B}'$ contains linear independent vectors (projections of vectors $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_d$ into orthogonal complement of the vector $\vec{\boldsymbol{x}}_{d+1}$). So we can understand these projections as $(N-1)$ dimensional vectors (in orthogonal complement of the vector $\vec{\boldsymbol{x}}_{d+1}$) and thus we can apply on them the induction assupmtion. Hence we get the following recursive formula

$$\widetilde{C}(d+1, N) = \widetilde{C}(d, N) + \widetilde{C}(d, N-1). \tag{2.5}$$

On the base of this recursive formula we derive the number of mutually different disjoint splittings of the set $\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_d$ in two homogenously linearly separable parts.

Evidently $\widetilde{C}(1,N)=2$ for all $N\geq 1$, because one-point set $\{\vec{\boldsymbol{x}}\}$ can be split up in two homogenously linearly separable tuples of sets $\{\emptyset,\{\vec{\boldsymbol{x}}\}\}$ and $\{\{\vec{\boldsymbol{x}}\},\emptyset\}$ only.

To illustrate recursive formula for $\widetilde{C}(d,N)$ let us write down these values for initial d and N in the following scheme (the each number in this scheme in a fixed position is equal to the sum of the number on the left and the number in the position which is immediately north-west):

	$d=1$	2	3	4	5	6	7	8
$N=1$	2							
2	2	4	8					
3	2	4	8					
4	2	4	8	16				
5	2	4	8	16	32			
6	2	4	8	16	32	64		
7	2	4	8	16	32	64	128	

$\ldots$

We can deduce from the above scheme of values $\widetilde{C}(d,N)$ that the number $\widetilde{C}(d,N)$ is equal to double of sum of combinatorial numbers. This observation can be easily verified using equality 2.5 (recall that $\binom{j}{i}=\binom{j-1}{i}\binom{j-1}{i-1}$):

$$\sum_{k=0}^{N-1}\binom{d}{k}=\sum_{k=0}^{N-1}\binom{d-1}{k}+\sum_{k=0}^{N-2}\binom{d-1}{k}=\sum_{k=0}^{N-1}\binom{d-1}{k}+\sum_{k=1}^{N-1}\binom{d-1}{k-1}=\sum_{k=0}^{N-1}\binom{d}{k}.$$

– *q. e. d.* –

2.3 VC-dimension of composed mapping

For an arbitrary system of (elementar) functions we can define the VC-dimension of such functions as the VC-dimension of all the sets, which are inverse image of the intervals $(-\infty, \alpha\rangle$. Next, we will show how to use the knowledge of the VC-dimension of such class concepts to determine upper estimate of VC-dimension of function sets contain a composition of those elementar functions. First of all, let us define the term composed mapping, though.

Definition 2.3.1 *Let* $\mathbf{D}$ *be a connected directed acyclic graph (DAG) and* $\bar{I}$ *denote the set of vertices* $\mathbf{D}$*. For a vertex* $j \in \bar{I}$ *let the numbers* d_j^+ *and* d_j^- *denote the number of edges leading into the vertex* j *and leading from the vertex* j*, respectively. Let the following properties hold:*

1. *There exists exactly one vertex with* $d_j^- = 0$*. We call this vertex an* OUTPUT VERTEX .
2. *Let there exist at least one vertex with* $d_j^+ = 0$*. We call such vertex an* INPUT VERTEX .
3. *For each vertex* j *with* $d_j^+ > 0$ *let there exist a mapping* $\widetilde{Z_j} : \left\{ \Re^{d_j^+} \times W_j \right\} \to \Re$ *, where* $\bar{W}_j$ *is a given parametric space of the mapping* $\widetilde{Z_j}$ *.*
4. *Let for each noninput vertex* j *the value* v_j *of the vertex* j *satisfy the following:*

$$v_j = \widetilde{Z_j}\left(\left(v_{i_1} \cdots v_{i_{d_j^+}},\right), \vec{\boldsymbol{w}}_j\right),$$

where $v_{i_1} \cdots v_{i_{d_j^+}}$ *are values of vertices from which an edge leads to the vertex* j *and* $\vec{\boldsymbol{w}}_j \in \bar{W}_j$*.*

Then, $\left(\mathbf{D}, \left(\widetilde{Z_j}, \bar{W}_j\right), j \in \bar{I}\right)$ *is called* COMPOSED MAPPING *. Vertex* j *is an* INTERIOR VERTEX *if* j *is not an input vertex or the output vertex. The number of input vertices is* DIMENSION OF COMPOSED MAPPING *. Let* z *be the number of noninput vertices. Each point in* $\bar{W}_1 \times \bar{W}_2 \cdots \times \bar{W}_z$*, is called* PARAMETRIZATION OF COMPOSED MAPPING *.*

Example 2.3.1 *For example, let* $\widetilde{s} : \Re^3 \times \bar{W}_s \times \bar{W}_f \times \bar{W}_g \times \bar{W}_h \to \Re$ *, where* $\bar{W}_s, \bar{W}_f, \bar{W}_g, \bar{W}_h$ *are parameter spaces of mappings* $\widetilde{s}$ *,* $\widetilde{f}$ *,* $\widetilde{g}$ *,* $\widetilde{h}$ *, respectively, and*

$$\widetilde{s}\left(\widetilde{f}\,(x, y, z, \vec{\boldsymbol{w}}_f), \widetilde{g}\left(y, \widetilde{h}\,(z, \vec{\boldsymbol{w}}_h), \vec{\boldsymbol{w}}_g\right), \vec{\boldsymbol{w}}_s\right).$$

The corresponding graph is sketched in Figure 2.1.

In the following text, we will derive basic upper estimate of the VC-dimension of composed mapping using the properties of the functions $\widetilde{Z_j}$. For this purpose, we will have to operate with parts of composed mapping that will be defined by means of what is called proper numbering of vertices of composed mapping.

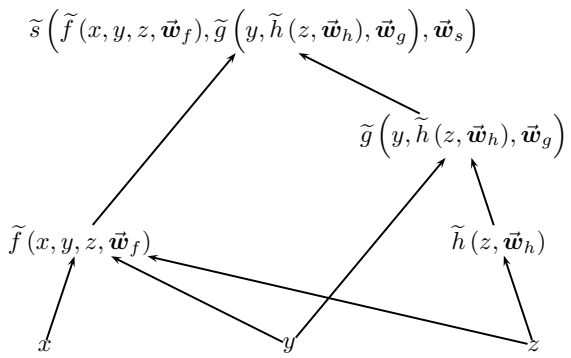


Figure 2.1: Corresponding graph for composed mapping.

Definition 2.3.2 *Let G be an acyclic connected directed graph, and noninput vertices be numbered by natural numbers $i \in \{1,\ldots,z\}$. Then, this numbering is* PROPER NUMBERING *iff each edge leads from a vertex with a smaller number.*

Proper numbering of a connected directed acyclic graph will be attained for instance in such a way whereby, after removing all the nodes having no input edge, we will first gradually number in a graph thus originated all the nodes having no input edge (this is possible since the graph is acyclic); these nodes will be removed from the graph, and this particular procedure will be repeated on the resultant graph, while gradually numbering, using subsequent numbers.

Definition 2.3.3 *Let $\left(\mathbf{D},\left(\widetilde{Z_j},\bar{W}_j\right),j\in\bar{I}\right)$ be a composed mapping with dimension n, let $\mathbf{D}$ has proper numbering with $(z-1)$ inner vertices, and let the output vertex have number z. Then, let us define for each $j\in\{1,\ldots,z\}$ the following concept class*

$$\mathsf{C}_j^{loc} \stackrel{def}{=} \left\{\bar{c}\subset\Re^{d_j^+}\,\middle|\,(\exists\vec{\boldsymbol{w}}\in W_j)\left(\bar{c}=\left\{\vec{\boldsymbol{s}}\in\Re^{d_j^+}\,\middle|\,\widetilde{Z_j}(\vec{\boldsymbol{w}},\vec{\boldsymbol{s}})\le 0\right\}\right)\right\}.$$

We will call each concept class C_j^{loc} a LOCAL CONCEPT CLASS .

Furthermore, let us denote the value of noninput vertex $j\in\{1,\ldots,z\}$ in dependence of $\omega\in\bar{W}_1\times\bar{W}_2\cdots\times\bar{W}_j$ and input values $\vec{\boldsymbol{x}}\in\Re^n$ of composed mapping as $v_{j,\omega,\vec{\boldsymbol{x}}}$ and let us define PARTIAL CONCEPT CLASS OF COMPOSED MAPPING *as the following system of sets*

$$\mathsf{C}_j^{par} \stackrel{def}{=} \left\{\bar{c}\subset\Re^{n}\,\middle|\,(\exists\omega\in\bar{W}_1\times\bar{W}_2\cdots\times\bar{W}_j)\left(\bar{c}=\left\{\vec{\boldsymbol{x}}\in\Re^{n}\,\middle|\,v_{j,\omega,\vec{\boldsymbol{x}}}\le 0\right\}\right)\right\}.$$

Than we will call the number $\mathtt{VC}_{dim}(\mathsf{C}_z^{par})$ as the VC-DIMENSION OF COMPOSED MAPPING .

In other words, for a given j and C_j^{loc}j is a system of all subsets $\bar{c}\in\Re^{d_j^+}$ such that there exist parameters in $\bar{W}_j$ in such a way that the set $\bar{c}$ is an inverse image of the interval $(-\infty,0\rangle$ in mapping $\widetilde{Z_j}(\vec{\boldsymbol{w}},\vec{\boldsymbol{s}})$. Concept classes C_z^{par} are defined in a similar fashion.

Definition 2.3.4 *Let* $\bar{V} \stackrel{def}{=} \{\vec{\boldsymbol{x}}_i, \ldots, \vec{\boldsymbol{x}}_m\} \subset \Re^n$ *and let* $\left(\mathbf{D}, \left(\widetilde{Z_j}, \bar{W}_j\right), j \in \bar{I}\right)$ *be a composed mapping of the dimension* n *with* $(n+z)$ *vertices and with proper numbering. Further, let* $l \in \{1, \ldots, z\}$ *and for each* $\omega \in \bar{W}_1 \times \bar{W}_2 \cdots \times \bar{W}_z$ *let the matrix* $\boldsymbol{R}^{\omega,\bar{V},l} \in \Re^{l \times m}$ *be defined as*

$$\boldsymbol{R}^{\omega,\bar{V},l}_{i,j} \stackrel{def}{=} \widetilde{sgn}\left(v_{j,\omega,\vec{\boldsymbol{x}}_i}\right), \quad j \in \{1, \ldots, l\},\ i \in \{1, \ldots, m\}\ .$$

Then, we say that two parametrizations ω_1 *and* ω_2 *are* $l, \bar{V}$-EQUIVALENT *if and only if*

$$\boldsymbol{R}^{\omega_1,\bar{V},l} = \boldsymbol{R}^{\omega_2,\bar{V},l}\ .$$

Lemma 2.3.1 *The relation* $l, \bar{V}$*-equivalency defined in 2.3.4 is equivalence relation on the set* $\bar{W}_1 \times \bar{W}_2 \cdots \times \bar{W}_l$.

■ *Proof:*
Reflexivity and symmetry is obvious. Transitivity follows from the fact that if $\widetilde{sgn}(a) = \widetilde{sgn}(b)$ and, at the same time, $\widetilde{sgn}(b) = \widetilde{sgn}(c)$, then also $\widetilde{sgn}(a) = \widetilde{sgn}(c)$.

– *q. e. d.* –

Theorem 2.3.2 *Let the symbol* $\mathtt{S}_{l,\bar{V}}$ *denote a number of equivalence classes of the* $l, \bar{V}$*-equivalence, let* m *be an arbitrary positive integer and* $\bar{V} \stackrel{def}{=} \{\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_m\} \subset \Re^n$. *Then, for all* $k \in \{1, \ldots, z\}$ *the following estimation*

$$\Pi_{\mathsf{C}^{\mathsf{par}}_{\mathsf{k}}}(m) \leq \mathtt{S}_{k,\bar{V}} \leq \Pi_{\mathsf{C}^{\mathsf{loc}}_{1}}(m) \cdot \Pi_{\mathsf{C}^{\mathsf{loc}}_{2}}(m) \cdot \cdots \cdot \Pi_{\mathsf{C}^{\mathsf{loc}}_{\mathsf{k}}}(m)$$

holds.

■ *Proof:*
For a fixed $k \in \{1, \ldots, z\}$ let us define the following sets

$$\Gamma \stackrel{\text{def}}{=} \left\{ \boldsymbol{R}^{\omega,\bar{V},l} \,\middle|\, \omega \in \bar{W}_1 \times \bar{W}_2 \cdots \times \bar{W}_k \right\}$$

and

$$\Gamma_j \stackrel{\text{def}}{=} \left\{ \vec{\boldsymbol{z}} \in \Re^m \,\middle|\, \left(\exists \omega \in \bar{W}_1 \times \bar{W}_2 \cdots \times \bar{W}_z\right) \left(\forall i \in \{1, \ldots, m\}\right) \left(\vec{\boldsymbol{z}}_i = \boldsymbol{R}^{\omega,\bar{V},k}_{i,j}\right) \right\},\ j \in \{1, \ldots, k\}\ .$$

Obviously, the number of equivalence classes is equal to $\left|\bar{\Gamma}\right|$ and $\left|\bar{\Gamma}\right| \leq \left|\bar{\Gamma}_1\right| \cdot \left|\bar{\Gamma}_2\right| \cdot \ldots \cdot \left|\bar{\Gamma}_k\right|$ ($\bar{\Gamma}_j$ contains all possible j-th columns of the matrices $\boldsymbol{R}^{\omega,\bar{V},k}$). Further, it is obvious that each $\vec{\boldsymbol{z}} \in \Gamma_j$ define disjoint splitting $\bar{V}_-$, $\bar{V}_+$ of the set $\bar{V}$, where

$$\bar{V}_- = \left\{ \vec{\boldsymbol{x}}_i \in \bar{V} \,|\, \vec{\boldsymbol{z}}_i \leq 0 \right\} \quad \text{and} \quad \bar{V}_+ = \left\{ \vec{\boldsymbol{x}}_i \in \bar{V} \,|\, \vec{\boldsymbol{z}}_i > 0 \right\}\ .$$

But the number of such disjoint splittings is less or equal to $\left|\Pi_{\mathsf{C}^{\mathsf{loc}}_{\mathsf{j}}}\left(\bar{V}\right)\right| \leq \Pi_{\mathsf{C}^{\mathsf{loc}}_{\mathsf{j}}}(m)$, so $\left|\bar{\Gamma}_j\right| \leq \Pi_{\mathsf{C}^{\mathsf{loc}}_{\mathsf{j}}}(m)$, $j \in \{1, \ldots, k\}$. Hence, $\mathtt{S}_{k,\bar{V}} \leq \Pi_{\mathsf{C}^{\mathsf{loc}}_{1}}(m) \cdot \Pi_{\mathsf{C}^{\mathsf{loc}}_{2}}(m) \cdot \cdots \cdot \Pi_{\mathsf{C}^{\mathsf{loc}}_{\mathsf{k}}}(m)$.

The inequality $\left|\Pi_{\mathsf{C}_k^{\mathsf{par}}}\left(\bar{V}\right)\right| \leq \mathtt{S}_{k,\bar{V}}$ follows directly from the definition of the number $\mathtt{S}_{k,\bar{V}}$ (last column of the matrix $\boldsymbol{R}^{\omega,\bar{V},k}$ defines all possible disjoint splittings of the set $\bar{V}$). But this restriction holds for all $\bar{V}$ of the size m, which follows $\Pi_{\mathsf{C}_k^{\mathsf{par}}}(m) \leq \mathtt{S}_{k,\bar{V}}$.

$-$ *q. e. d.* $-$

Up to this point in our explication we have not used anywhere the properties of partial mappings $\widetilde{Z}_i$ of a composed mapping; so far we have used only the fact that the graph of composed mapping is solely acyclic and directed. Theorem 2.3.2 provides instruction on how to proceed in deriving upper estimate of VC-dimension of more general composed mappings since it makes it possible to reduce the process of finding such an estimate into estimates of the VC-dimension of mapping corresponding to the individual nodes of the graph onlyy. We will illustrate this procedure on a simple example of the composition mappings. We start with the following lemma.

Lemma 2.3.3 *Let* $\{\alpha_i\}_1^z$ *be positive numbers and* $\sum_{i=1}^z \alpha_i = 1$. *Then,*

$$-\sum_{i=1}^{z} \alpha_i \ln(\alpha_i) \leq \ln(z). \tag{2.6}$$

■ *Proof:*
Firstly, let $\alpha_i = \frac{1}{z}$, $i \in \{1,\ldots,z\}$. Then,

$$-\sum_{i=1}^{z} \frac{1}{z} \ln\left(\frac{1}{z}\right) = \sum_{i=1}^{z} \ln\left(z^{\frac{1}{z}}\right) = \ln(z) \ ,$$

e.g. for $\alpha_i = \frac{1}{z}$, $i \in \{1,\ldots,z\}$, equality holds.
Further, let $\vec{\boldsymbol{s}} \in \left((1,1,\cdots,1)^T\right)^{\perp}$. Hence, $\sum_{i=1}^z \vec{\boldsymbol{s}}_i = 0$ and $(\forall t \in \Re)\left(\sum_{j=1}^z \frac{1}{z} + t\vec{\boldsymbol{s}}_i = 1\right)$.
Let $t_{min}, t_{max} \in \Re$ be numbers such that

$$(\forall t \in (t_{min}, t_{max}))\,(\forall i \in \{1,\ldots,z\})\left(\frac{1}{z} + t\vec{\boldsymbol{s}}_i > 0\right) \ .$$

Let us define function $\widetilde{f} : (t_{min}, t_{max}) \to \Re$ as

$$\widetilde{f}(t) \stackrel{\text{def}}{=} \sum_{i=1}^{z} \left(\frac{1}{z} + t\vec{\boldsymbol{s}}_i\right) \ln\left(\frac{1}{z} + t\vec{\boldsymbol{s}}_i\right).$$

Then (recall $\sum_{i=1}^z \vec{\boldsymbol{s}}_i = 0$)

$$\frac{d\widetilde{f}}{dt} = \sum_{i=1}^{z} \vec{\boldsymbol{s}}_i + \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z} + t\vec{\boldsymbol{s}}_i\right) = \sum_{i=1}^{z} \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z} + t\vec{\boldsymbol{s}}_i\right) \ .$$

Obviously for $t = 0$ the following holds:

$$\sum_{i=1}^{z} \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z} + t\vec{\boldsymbol{s}}_i\right) = 0 \quad \text{e.g.} \quad \frac{d\widetilde{f}}{dt}(0) = 0 \ .$$

Because $\ln(\cdot)$ is monotonously increasing function, we can write

$$\begin{array}{rrcl} t>0 & s>0 & \Rightarrow & \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right) > \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}\right) \\ & s<0 & \Rightarrow & \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right) > \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}\right) \\ t<0 & s>0 & \Rightarrow & \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right) < \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}\right) \\ & s<0 & \Rightarrow & \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right) < \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}\right) \end{array} .$$

If we take into account the fact $\sum_{i=1}^{z} \vec{\boldsymbol{s}}_i = 0$ it is straightforward that

$$t>0 \Rightarrow \sum_{i=1}^{z} \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right) > 0 \quad \text{and} \quad t<0 \Rightarrow \sum_{i=1}^{z} \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right) < 0.$$

In other words

$$(\forall t \in (t_{min}, t_{max})) \left(t \neq 0 \Rightarrow \frac{d\widetilde{f}}{dt}(t) \neq 0 \right) .$$

Finally, let us show that the second derivative of the $\widetilde{f}$ in $t=0$ is positive:

$$\frac{d}{dt}\left(\frac{d\widetilde{f}}{dt}\right) = \frac{d}{dt}\left(\sum_{i=1}^{z} \vec{\boldsymbol{s}}_i \ln\left(\frac{1}{z}+t\vec{\boldsymbol{s}}_i\right)\right) = \sum_{i=1}^{z} \frac{\vec{\boldsymbol{s}}_i^2}{\frac{1}{z}+t\vec{\boldsymbol{s}}_i} .$$

$$\widetilde{f}''(0) = \sum_{i=1}^{z} z \cdot \vec{\boldsymbol{s}}_i^2 > 0 .$$

Hence, the function $\widetilde{f}$ has maximum in $t=0$ for arbitrary vector $\vec{\boldsymbol{s}} \in \left((1,1,\cdots,1)^T\right)^{\perp}$.

– *q. e. d.* –

Theorem 2.3.4 *Let* $\left(\mathbf{D}, \left(\widetilde{Z_j}, \bar{W}_j\right), j \in \bar{I}\right)$ *be a composed mapping with dimension* n*,let* w *be the number of edges,* z *be the number of noninput vertices and* $q \stackrel{def}{=} w+z$*. Further let for all noninput vertices* k *is* $\mathtt{VC}_{dim}\left(\mathsf{C}_k^{loc}\right) = d_k^+ + 1$ *. Then, for any* $m > \max\left\{d_i^+ \,|i \in \{1,\ldots,z\}\right\}$ *is*

$$\Pi_{\mathsf{C}_z^{\mathsf{par}}}(m) \leq \left(\frac{ezm}{q}\right)^q \tag{2.7}$$

and further the estimation

$$\mathtt{VC}_{dim}\left(\mathsf{C}_z^{par}\right) < 2q\log_2(ez) \tag{2.8}$$

holds.

▪ *Proof:*

▪ add Proof 2.7)

Let us assume that proper numbering is defined in the graph $\mathbf{D}$ and let d_i^+ denote the number of edges leading to the node i. According to the claim of the lemma 2.1.3 the following estimate holds for each considered i and $m \geq d_i^+ + 1$

$$\Pi_{\mathsf{C}_i}(m) \leq \left(\frac{em}{d_i^+ + 1}\right)^{d_i+^+1} ,$$

because by assuption is $\mathtt{VC}_{dim}\left(\mathsf{C}_k^{loc}\right) = d_k^+ + 1$. But then

$$\Pi_{\mathsf{C}_\mathsf{L}}(m) \leq \Pi_{\mathsf{C}_1}(m) \cdot \Pi_{\mathsf{C}_2}(m) \cdot \cdots \Pi_{\mathsf{C}_z}(m) \leq \prod_{i=1}^{z} \left(\frac{em}{d_i^+ + 1}\right)^{d_i^+ + 1} . \tag{2.9}$$

Obviously it holds

$$q = \sum_{i=1}^{z} \left(d_i^+ + 1\right) \tag{2.10}$$

and let us set

$$\alpha_i \stackrel{\text{def}}{=} \frac{d_i^+ + 1}{q}.$$

By substituting into the estimate 2.6, we come to the expression

$$\sum_{i=1}^{z} \frac{d_i^+ + 1}{q} \ln\left(\frac{q}{d_i^+ + 1}\right) \leq \ln(z).$$

Hence (multiplying by q and using the equality $\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$)

$$\sum_{i=1}^{z} \left(d_i^+ + 1\right) \ln\left(\frac{1}{d_i^+ + 1}\right) \leq q\ln(z) - \left(\sum_{i=1}^{z} \left(d_i^+ + 1\right)\right) \ln(q) = q\ln(z) - q\ln(q).$$

Thus, finally, by delogaritmizing we obtain that

$$\prod_{i=1}^{z} \left(\frac{1}{d_i^+ + 1}\right)^{d_i^+ + 1} \leq \left(\frac{z}{q}\right)^q.$$

Using the equation 2.10, we can arrange the previous expression into the following form

$$\prod_{i=1}^{z} \left(\frac{em}{d_i^+ + 1}\right)^{d_i^+ + 1} \leq \left(\frac{z}{q}\right)^q \cdot \prod_{i=1}^{z} (em)^{d_i^+ + 1} = \left(\frac{z}{q}\right)^q \cdot (em)^q$$

from which the first proven estimate ensues by substituting into 2.9.

- add Proof 2.8)

First, we will prove the inequality for $z = 1$. Hence, it obviously holds that $\mathtt{VC}_{dim}\left(\mathsf{C}_1^{par}\right) = w + 1$ and it holds that

$$\mathtt{VC}_{dim}\left(\mathsf{C}_1^{par}\right) = w + 1 < 2(1 + w)\log_2(ez) .$$

Further, we will prove the inequality 2.8 or the case $z \geq 2$. For each real number $a > 4$, the following estimates $(q > 1)$ hold

$$2\log_2(a) < a \Leftrightarrow 2a\log_2(a) < a^2 \Leftrightarrow 2a\log_2(a) < 2^{2\log_2(a)} \Leftrightarrow (2a\log_2(a))^q < 2^{2q\log_2(a)}.$$

If we substitute $a \stackrel{\text{def}}{=} ez$, $z \geq 2$, then $a > 2.71 \cdot z > 4$ and we obtain that the inequality given below always holds

$$\left(\frac{ez2q\log_2(ez)}{q}\right)^q < 2^{2q\log_2(ez)}. \tag{2.11}$$

Let $\mathtt{VC}_{dim}\left(\mathsf{C}_z^{par}\right) = m$. Then, from the definition of the VC-dimension, m is a maximum number such that the equality $2^m = \Pi_{\mathsf{C}_z^{\mathsf{par}}}(m)$ holds; furthermore, if this inequality holds for some m, it holds for all natural m which are smaller. But according to the estimate 2.7 and 2.11 it must also hold that

$$\Pi_{\mathsf{C}_\mathsf{L}}\left(2q\log_2(ez)\right) \le \left(\frac{ez2q\log_2(ez)}{q}\right)^q < 2^{2q\log_2(ez)}$$

and, therefore, an set of the size $2q\log_2(ez)$ cannot be shattered by the class C_z^{par}, hence $\mathtt{VC}_{dim}\left(\mathsf{C}_z^{par}\right) < 2q\log_2(ez)$.

– *q. e. d.* –

2.4 VC-dimension of symmetric difference

Lemma 2.4.1 *Let $\bar{S}, \bar{h}_1, \bar{h}_2 \subset \bar{X}$. Then,*

$$\bar{h}_1 \cap \bar{S} = \bar{h}_2 \cap \bar{S} \quad \Leftrightarrow \quad \left(\bar{X} \dot{-} \bar{h}_1\right) \cap \bar{S} = \left(\bar{X} \dot{-} \bar{h}_2\right) \cap \bar{S}$$

■ *Proof:*
Let us assume that $\bar{h}_1 \cap \bar{S} \neq \bar{h}_2 \cap \bar{S}$. Without loss of generality, let $y \in \bar{h}_1 \cap \bar{S}$ and $y \notin \bar{h}_2 \cap \bar{S}$. It follows $y \in \bar{h}_1$, $y \in \bar{S}$ and $y \notin \bar{h}_2$. Hence, $y \notin \left(\bar{X} \dot{-} \bar{h}_1\right) \cap \bar{S}$ and $y \in \left(\bar{X} \dot{-} \bar{h}_2\right) \cap \bar{S}$. Therefore, $\bar{h}_1 \cap \bar{S} \neq \bar{h}_2 \cap \bar{S}$ implies $\left(\bar{X} \dot{-} \bar{h}_1\right) \cap \bar{S} \neq \left(\bar{X} \dot{-} \bar{h}_2\right) \cap \bar{S}$, which completes the proof.

– *q. e. d.* –

Lemma 2.4.2 *Let $\bar{S}, \bar{c}, \bar{h}_1, \bar{h}_2 \subset \bar{X}$. Further, let*

$$\left(\bar{X} \dot{-} \bar{c}\right) \cap \bar{h}_1 \cap \bar{S} = \left(\bar{X} \dot{-} \bar{c}\right) \cap \bar{h}_2 \cap \bar{S} \quad \textit{and} \quad \bar{c} \cap \left(\bar{X} \dot{-} \bar{h}_1\right) \cap \bar{S} = \bar{c} \cap \left(\bar{X} \dot{-} \bar{h}_2\right) \cap \bar{S}\ . \tag{2.12}$$

Then, $\bar{h}_1 \cap \bar{S} = \bar{h}_2 \cap \bar{S}$.

■ *Proof:*
We prove the claim by contradiction. Let us assume that $\bar{h}_1 \cap \bar{S} \neq \bar{h}_2 \cap \bar{S}$. There is clearly no loss of generality in assuming $y \in \bar{h}_1 \cap \bar{S}$ and $y \notin \bar{h}_2 \cap \bar{S}$. It follows $y \in \bar{h}_1$, $y \in \bar{S}$ and $y \notin \bar{h}_2$. Hence, $y \notin \left(\bar{X} \dot{-} \bar{c}\right) \cap \bar{h}_2 \cap \bar{S}$. So the left part of 2.12 implies that $y \notin \left(\bar{X} \dot{-} \bar{c}\right) \cap \bar{h}_1 \cap \bar{S}$. At the same time, $y \in \bar{h}_1 \cap \bar{S}$ which follows $y \notin \left(\bar{X} \dot{-} \bar{c}\right)$. So $y \in \bar{c}$. Recalling previous argumentation, we have

$$y \in \bar{c}, \quad y \in \bar{S}, \quad y \notin \bar{X} \dot{-} \bar{h}_1, \quad y \in \bar{X} \dot{-} \bar{h}_2\ .$$

Put together, these equations leads to

$$y \notin \bar{c} \cap \left(\bar{X} \dot{-} \bar{h}_1\right) \cap \bar{S} \quad \text{and} \quad y \in \bar{c} \cap \left(\bar{X} \dot{-} \bar{h}_2\right) \cap \bar{S}$$

which contradicts the right part of 2.12.

– *q. e. d.* –

Lemma 2.4.3 *Let $\bar{S}, \bar{c}, \bar{h}_1, \bar{h}_2 \subset \bar{X}$. Further, let*

$$\left[\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S}\right] \cup \left[\bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_1\right) \cap \bar{S}\right] = \left[\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S}\right] \cup \left[\bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_2\right) \cap \bar{S}\right]. \tag{2.13}$$

Then,

$$\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S} = \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S} \quad \textit{and} \quad \bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_1\right) \cap \bar{S} = \bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_2\right) \cap \bar{S}\,.$$

■ *Proof:*
Let us assume that $\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S} \neq \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S}$. For example, let there exist y such that $y \in \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S}$ and $y \notin \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S}$. In this case, $y \in \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S}$ implies $y \in \bar{X}\dot{-}\bar{c}$, e.g. $y \notin \bar{c}$, $y \in \bar{h}_1$ and $y \in \bar{S}$. Hence, $y \notin \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S}$ implies that $y \notin \bar{h}_2$. It means that

$$y \in \left[\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S}\right] \cup \left[\bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_1\right) \cap \bar{S}\right]$$

and

$$y \notin \left[\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S}\right] \cup \left[\bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_2\right) \cap \bar{S}\right]$$

which conclude the proof of the lemma.

– *q. e. d.* –

Theorem 2.4.4 *Let $m \geq 1$, $\bar{c} \subset \bar{X}$ and $\mathsf{H} \subset 2^{\bar{X}}$, $\mathsf{R} \stackrel{def}{=} \left\{\bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathsf{H}\right\}$. Then,*

1. $\Pi_{\mathsf{H}}(m) = \Pi_{\mathsf{R}}(m)$ *and*
2. $\mathtt{VC}_{dim}(\mathsf{H}) = \mathtt{VC}_{dim}(\mathsf{R})$. *(Note that* R *depends on the concept* $\bar{c}$ *.)*

■ *Proof:*
▪ add 1)
Let $\bar{h}_1, \bar{h}_2 \in \mathsf{H}$ and $\bar{S}, \bar{c} \subset \bar{X}$ be arbitrary but fixed. Then, by the lemma 2.4.1 is

$$\bar{h}_1 \cap \bar{S} = \bar{h}_2 \cap \bar{S} \quad \Leftrightarrow \quad \left(\bar{X}\dot{-}\bar{h}_1\right) \cap \bar{S} = \left(\bar{X}\dot{-}\bar{h}_2\right) \cap \bar{S}\,. \tag{2.14}$$

From 2.14 it is apparent that

$$\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S} = \left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S} \quad \text{and} \quad \bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_1\right) \cap \bar{S} = \bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_2\right) \cap \bar{S}\,. \tag{2.15}$$

Owing to the lemma 2.4.2 (see 2.12) we are going to see that simultaneous validity of equalities 2.15 implies both equalities 2.14.
Further, it is clear and straightforward that 2.15 implies the following set equality

$$\left[\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_1 \cap \bar{S}\right] \cup \left[\bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_1\right) \cap \bar{S}\right] = \left[\left(\bar{X}\dot{-}\bar{c}\right) \cap \bar{h}_2 \cap \bar{S}\right] \cup \left[\bar{c} \cap \left(\bar{X}\dot{-}\bar{h}_2\right) \cap \bar{S}\right]. \tag{2.16}$$

Similarly as in the previous text, owning to the lemma 2.4.3 (see 2.13), we can see that the opposite implication is also true, e.q. that 2.16 implies 2.15. So 2.16 is equivalent to $\bar{h}_1 \cap \bar{S} = \bar{h}_2 \cap \bar{S}$.
Now recall definition of the symmetric difference of the sets $\bar{c}$ and $\bar{h}$,

$$\bar{h} \triangle \bar{c} = \left(\bar{h} \dot{-} \bar{c}\right) \cup \left(\bar{c} \dot{-} \bar{h}\right) = \left(\bar{h} \cap \left(\bar{X} \dot{-} \bar{c}\right)\right) \cup \left(\bar{c} \cap \left(\bar{X} \dot{-} \bar{h}\right)\right) .$$

It follows (use $\left(\bar{A} \cup \bar{B}\right) \cap \bar{S} = \left(\bar{A} \cap \bar{S}\right) \cup \left(\bar{B} \cap \bar{S}\right)$)

$$\left(\bar{h} \triangle \bar{c}\right) \cap \bar{S} = \left[\left(\bar{h} \cap \left(\bar{X} \dot{-} \bar{c}\right)\right) \cap \bar{S}\right] \cup \left[\left(\bar{c} \cap \left(\bar{X} \dot{-} \bar{h}\right)\right) \cap \bar{S}\right] .$$

So, finally, using the last expression, we can rewrite the equation 2.16 to a more legible form

$$\left(\bar{h}_1 \triangle \bar{c}\right) \cap \bar{S} = \left(\bar{h}_2 \triangle \bar{c}\right) \cap \bar{S} .$$

Thus, we have proved that for arbitrary $\bar{h}_1, \bar{h}_2 \in \mathsf{H}$ and $\bar{S}, \bar{c} \subset \bar{X}$ the equivalence

$$\bar{h}_1 \cap \bar{S} = \bar{h}_2 \cap \bar{S} \quad \Leftrightarrow \quad \left(\bar{h}_1 \triangle \bar{c}\right) \cap \bar{S} = \left(\bar{h}_2 \triangle \bar{c}\right) \cap \bar{S} \tag{2.17}$$

is fulfiled.
The rest of the proof goes as follows. For the fixed set $\bar{S}$ we show that $\left|\Pi_{\mathsf{H}}\left(\bar{S}\right)\right| = \left|\Pi_{\mathsf{R}}\left(\bar{S}\right)\right|$. To do this let us define a mapping $\widetilde{Z} : \Pi_{\mathsf{H}}\left(\bar{S}\right) \to \Pi_{\mathsf{R}}\left(\bar{S}\right)$ as

$$\widetilde{Z}\left(\bar{u}\right) \stackrel{\text{def}}{=} \left(\bar{h}_{\bar{u}} \triangle \bar{c}\right) \cap \bar{S} , \quad \text{where} \quad \bar{h}_{\bar{u}} \in \left\{\bar{h} \in \mathsf{H} \,\middle|\, \bar{h} \cap \bar{S} = \bar{u}\right\} .$$

First, note that the mapping $\widetilde{Z}$ is well defined. Let $\bar{u} \in \Pi_{\mathsf{R}}\left(\bar{S}\right)$ and $\bar{h}'_{\bar{u}}, \bar{h}''_{\bar{u}} \in \left\{\bar{h} \in \mathsf{H} \,\middle|\, \bar{h} \cap \bar{S} = \bar{u}\right\}$. Then, owning to 2.17, we get

$$\widetilde{Z}\left(\bar{u}\right) = \left(\bar{h}'_{\bar{u}} \triangle \bar{c}\right) \cap \bar{S} = \left(\bar{h}''_{\bar{u}} \triangle \bar{c}\right) \cap \bar{S} .$$

Therefore, the image of the set $\bar{u}$ is independent on the choice of the set $\bar{h}_{\bar{u}}$. Also it is straightforward that equivalence 2.17 implies injectivity of the mapping $\widetilde{Z}$. Finally, let $\bar{q} \in \Pi_{\mathsf{R}}\left(\bar{S}\right)$. Clearly, $\bar{q}$ is rewritable in the form $\bar{q} = \left(\bar{h}_{\bar{q}} \triangle \bar{c}\right) \cap \bar{S}$ and, hence, $\widetilde{Z}\left(\bar{h}_{\bar{q}}\right) = \bar{q}$. In other words, the mapping $\widetilde{Z}$ is one-to-one which follows $\left|\Pi_{\mathsf{H}}\left(\bar{S}\right)\right| = \left|\Pi_{\mathsf{R}}\left(\bar{S}\right)\right|$ immediately. The set $\bar{S}$ was chosen arbitrarily, so the definition of the $\Pi_{\mathsf{H}}\left(\bar{S}\right)$ follows that for any $m \geq 1$ is

$$\Pi_{\mathsf{H}}\left(m\right) = \Pi_{\mathsf{R}}\left(m\right) .$$

▪ add 2)
Let us take the alternative definition $\mathtt{VC}_{dim}\left(\mathsf{H}\right) \stackrel{\text{def}}{=} \sup\left\{m \in N \,|2^m = \Pi_{\mathsf{H}}\left(m\right)\right\}$. Since the first part of the lemma is proved for arbitrary value of $m \in N$, it is straightforward that for the fixed set $\bar{c}$ (note that R depends on the set $\bar{c}$) is

$$\sup\left\{m \in N \,|2^m = \Pi_{\mathsf{H}}\left(m\right)\right\} = \sup\left\{m \in N \,|2^m = \Pi_{\mathsf{R}}\left(m\right)\right\} .$$

– q. e. d. –

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Chapter 3

Sample Complexity and VC–dimension

In this chapter, we will deal with the bottom and upper estimates of the number of examples necessary for learning algorithms of the complexity $\widetilde{m}(\epsilon, \delta)$ as has been introduced in [BEHW89].

One such estimate for a sufficient quantity of enquiries is given in Theorem 1.2.1. In this chapter, we will show that a dominating role in these estimates is played primarily by the concept VC-dimension studied above. Proceeding from that particular concept, we can formulate better estimates of the necessary length of the sample used for generating a hypothesis approximating the given concept.

First, we will discuss the above-mentioned estimates for a fixed chosen system of concepts and hypotheses. In the subsequent text, we will consider cases when the given concepts are distinguished according to the dimension of space whose subsets they are, and we will analyze sample complexity of learning algorithms with a view to the dimension of vector spaces containing the given concepts as their subsets. A natural requirement is to postulate a criterion guaranteeing that the length of the sample $\widetilde{m}(\epsilon, \delta)$ of learning algorithm will be upper bound by polynom in the dimension of space.

In conclusion, we will carry out a similar analysis concerning the length of words that describe the concepts from the given concept system and which, in their essence, express through their length descriptive complexity of concepts.

3.1 Estimate of the Number of Samples

In this part, we will prove a series of lemmas and theorems, which will be necessary for the proof of Theorem 3.1.12, giving the lower and upper estimate of sample complexity for (ϵ, δ)-algorithms.

Throughout this chapter and in the subsequent ones we will assume that each considered concept class, just as each considered hypothesis class, would be made up solely of Borelian sets. For the sake of comprehensiveness, let us recall the term Borelian sets ([Jar55]). Let us assume that Ω s a non-empty system of sets, which is σ-additive, i.e. for any arbitrary sequence of sets $\bar{A}_n \in \Omega$ it holds that $\bigcup_{n=1}^{\infty} \bar{A}_n \in \Omega$. Furthermore, if it holds that for any arbitrary $\bar{A}, \bar{B} \in \Omega$ there is also $\bar{A} \dot{-} \bar{B} \in \Omega$, we will call the system of sets Ω

σ-additive circle. It can be illustrated that for each non-empty system of sets Ω there exists just one smallest σ-additive circle containing Ω (this is precisely the intersection of all the σ-additive circles comprising the system Ω). We will call this minimal σ-additive circle the Borelian circle over the system Ω. Let $\mathcal{M}$ be an arbitrary metric space and let Θ be a Borelian circle over a system of all open sets in $\mathcal{M}$. Then, we will call the elements of this circle BORELIAN SETS of the space $\mathcal{M}$.

Apart from the requirement regarding the Borelian character of sets which form the individual concepts it will be necessary for other proofs to characterize in greater detail the concepts and hypotheses under consideration. This necessary characteristic can be suitably captured by the term ϵ-transversal.

Definition 3.1.1 *For any arbitrary* $\mathsf{R} \subset 2^{\bar{X}}$ *and for any arbitrary probability density* $\widetilde{P}$ *defined on* $\bar{X}$ *and for an arbitrary* $\epsilon > 0$ *let us define* $\mathsf{R}_{\widetilde{P},\epsilon} \stackrel{def}{=} \left\{ \bar{r} \in \mathsf{R} \,\middle|\, Prob_{\widetilde{P}}(\bar{r}) > \epsilon \right\}$. *Then, we will call* $\bar{T}_{\widetilde{P},\epsilon} \subset \bar{X}$ ϵ-TRANSVERSAL R *just when*

$$\left(\forall \bar{r} \in \mathsf{R}_{\widetilde{P},\epsilon}\right)\left(\bar{r} \cap \bar{T}_{\widetilde{P},\epsilon} \neq \emptyset\right)$$

Example 3.1.1 *Let* $\bar{a} \stackrel{def}{=} \langle 0, 1\rangle$ *and* R *be a system of all closed intervals on* $\bar{a}$ *and let* $\widetilde{P}$ *be an uniform probability density on* $\bar{a}$. *Then, the set of all the points* ϵk, $1 \leq k \leq \frac{1}{\epsilon}$ *forms the* ϵ*-transversal* R *for any arbitrary value* $\epsilon > 0$.

Example 3.1.2 *If* R *is formed by all the open subsets of the interval* $\bar{a}$ *then - for uniform probability density - there exists no finite* ϵ*-transversal for any* ϵ.

We will examine the probability of the selection of the ϵ-transversal of the system R on the basis of arbitrary random selection of points from $\bar{X}$. Specifically, we will be interested in the probability of the event described in the following definition.

Definition 3.1.2 *For each* $m \geq 1$ *and* $\epsilon > 0$ *let* $\bar{Q}_{m,\epsilon}$ *denote a set of all* $\breve{x} \in \bar{X}^m$ *such that the different elements* $\breve{x}$ *do not form* ϵ*-transversal for* R, *so*

$$\bar{Q}_{m,\epsilon} \stackrel{def}{=} \left\{ \breve{x} \in \bar{X}^m \,\middle|\, \left(\exists \bar{r} \in \mathsf{R}_{\widetilde{P},\epsilon}\right)\left(\breve{x} \cap \bar{r} = \emptyset\right) \right\}.$$

Further define the set

$$\bar{J}_{\epsilon}^{2m} \stackrel{def}{=} \left\{ \left(\breve{x}, \breve{y}\right) \in \bar{X}^m \times \bar{X}^m \,\middle|\, \left(\exists \bar{r} \in \mathsf{R}_{\widetilde{P},\epsilon}\right)\left(\breve{x} \cap \bar{r} = \emptyset \quad \text{and} \quad \left|\breve{y} \cap \bar{r}\right| \geq \frac{\epsilon m}{2}\right) \right\}.$$

In the proofs of the probability properties of learning algorithms we will be interested primarily in the property of the system of sets $\mathsf{R} \stackrel{\text{def}}{=} \left\{ \bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathsf{H} \right\}$, where $\bar{c}$ is some fixed concept $\bar{c} \subset \bar{X}$ and H is the hypothesis class for this particular concept. **The importance of the term ϵ-transversal is in that if the sample of the concept $\bar{c}$ is simultaneously ϵ-transversal for R, then it contains an counter-example for each hypothesis whose error – as seen in terms of the target concept $\bar{c}$ is greater than ϵ.** To be able to examine the properties of the system R, we will take into consideration hypothesis classes meeting the following criterion given in the definition below.

Definition 3.1.3 *The hypothesis class* H *is* WELL-BEHAVED *if the sets* $\bar{Q}_{m,\epsilon}$ *and* $\bar{J}_{\epsilon}^{2m}$ *are measurable for any arbitrary probability density* $\widetilde{P}$ *, any arbitrary* $m \geq 1$*,* $\epsilon > 0$ *and any arbitrary system of sets* $\mathsf{R} \stackrel{def}{=} \left\{ \bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathsf{H} \right\}$*, where* $\bar{c}$ *is an arbitrary Borelian sets.*

An overwhelming majority of concept classes and hypothesis classes commonly used is well-behaved. One of the possibilities of verification of this particular property is to verify universal separability.

Definition 3.1.4 *The hypothesis class* $\mathsf{H} \subset 2^{\bar{X}}$ *is* UNIVERSALLY SEPARABLE *, if there exists a countable subset* T *of the class* H *such that for all* $\bar{h} \in \mathsf{H}$ *there exists a sequence* $\{\bar{h}_i\}_1^{\infty}$ *of sets from* T *such that*

$$\left(\forall x \in \bar{X}\right) \left(\exists n \geq 1\right) \left(\left(\forall i \geq n\right) \left(x \in \bar{h}_i \text{ if and only if } x \in \bar{h} \right) \right).$$

The implication formulated in the following theorem holds.

Theorem 3.1.1 *If* H *is universally separable, then* H *is well-behaved.*

▪ *Proof:*
Let $\bar{c} \subset \bar{X}$ be a Borelian set and $\mathsf{R} \stackrel{\text{def}}{=} \left\{ \bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathsf{H} \right\}$. We will show that the sets $\bar{Q}_{m,\epsilon}$ and $\bar{J}_{\epsilon}^{2m}$ are Borelian. Proof will be given only for $\bar{Q}_{m,\epsilon}$, being similar for $\bar{J}_{\epsilon}^{2m}$.

Since H is universally separable, it ensues from the definition that R is universally separable as well. Let $\mathsf{T} \subset \mathsf{R}$ be a set from the definition of universal separability. Let $\{\gamma_i\}_1^{\infty}$ be a descending sequence of positive numbers converging to zero, $\{\epsilon_i\}_1^{\infty}$ be a descending sequence of positive numbers converging to $\epsilon > 0$. Let us define for each $i, j \geq 1$

$$\mathsf{T}_{i,j} \stackrel{\text{def}}{=} \left\{ \bar{t} \in \mathsf{T} \,\middle|\, \left(\exists \bar{r} \in \mathsf{R}\right) \left(Prob_{\widetilde{P}}\left(\bar{r}\right) \geq \epsilon_i \quad \text{and} \quad Prob_{\widetilde{P}}\left(\bar{t} \triangle \bar{r}\right) \leq \gamma_j \right) \right\}.$$

We will show that the following equality holds:

$$\bar{Q}_{m,\epsilon} = \bigcup_{i=1}^{\infty} \bigcap_{j=1}^{\infty} \bigcap_{\bar{t} \in \mathsf{T}_{i,j}} \left\{ \breve{x} \in \bar{X}^m \,\middle|\, \breve{x} \cap \bar{t} = \emptyset \right\}.$$

▪ add inclusion $\supset$:)
for $\breve{x}$ from the set on the right there exists such $\bar{t} \in \mathsf{T}_{i,j}$, where $\breve{x} \cap \bar{t} = \emptyset$ furthermore, $\epsilon_i - \gamma_j > \epsilon$. For each such $\bar{t}$ there is $\bar{t} \in \mathsf{R}$ and $Prob_{\widetilde{P}}\left(\bar{t}\right) > \epsilon$, therefore $\breve{x} \in \bar{Q}_{m,\epsilon}$.
▪ add inclusion $\subset$:)
For each $\breve{x} \in \bar{Q}_{m,\epsilon}$ there exists $i \geq 1$ in such a way that for some $\bar{r} \in \mathsf{R}$ there is $\breve{x} \cap \bar{r} = \emptyset$ and $Prob_{\widetilde{P}}\left(\bar{r}\right) > \epsilon_i$. Since R is universally separable there exists in T a sequence of sets converging, in terms of points, to $\bar{r}$ (in the sense of the definition of universal separability), hence for each $j \geq 1$ there exists $\bar{t} \in \mathsf{T}$, for which $Prob_{\widetilde{P}}\left(\bar{t} \triangle \bar{r}\right) \leq \gamma_j$ a $\breve{x} \cap \bar{t} = \breve{x} \cap \bar{r} = \emptyset$. Therefore, $\bar{x}$ is in the set on the right.
Hence $\bar{Q}_{m,\epsilon}$ is a Borelian set and the theorem is valid.

– *q. e. d.* –

We will now prove the following technical lemmas providing elementary inequalities necessary for estimates of the model complexity of learning algorithms (let us establish a convention that $\widetilde{P}^m$ is probability density defined on the cartesian product $\bar{X}^m$, derived from probability density $\widetilde{P}$ defined on $\bar{X}$). Chebyshev's inequality will be used for the proof of the immediately following lemma [1]

Lemma 3.1.2 *Let $\bar{X}$ be a probability space, $m \in N$, $\bar{r} \subset \bar{X}$, such that $Prob(\bar{r}) \geq \epsilon \stackrel{def}{=} \frac{2}{m}$. Then, the probability that an m–member independent selection of points from $\bar{X}$ will contain at least one point from $\bar{r}$ is greater than $\frac{1}{2}$.*

■ *Proof:*
The proof is based on the application of Chebyshev's inequality. Let $\widetilde{A}$ be a random variable equal to the number of elements belonging to $\bar{r}$ during implementation of m independent random selections from $\bar{X}$. This random variable $\widetilde{A}$ has a binomic distribution. [2] .

For the sake of simplicity let us assume that the set $\bar{r}$ has probability ϵ (this can be assumed because if probability of the set $\bar{r}$ is greater, the theorem being proved holds all the more). Then, the random variable $\widetilde{A}$ has mean value $\mu = m\epsilon$ and variance $\sigma^2 = m\epsilon(1-\epsilon)$. Having substituted into Chebyshev's inequality, we get

$$Prob\left(\left|\widetilde{A} - m\epsilon\right| \geq \lambda\right) \leq \frac{m\epsilon(1-\epsilon)}{\lambda^2} < \frac{m\epsilon}{\lambda^2}.$$

If we substitute $m = \frac{2}{\epsilon}$ a $\lambda \stackrel{\text{def}}{=} 2$, we get

$$Prob\left(\left|\widetilde{A} - 2\right| \geq 2\right) \stackrel{\text{def}}{=} \gamma < \frac{1}{2}.$$

[1] Chebyshev's theorem, (see e.g. [And85])
Let $\widetilde{V}$ be an arbitrary random variable with a mean value μ and variance σ^2. Then, it holds that

$$(\forall \lambda > 0)\left(Prob\left(\left|\widetilde{V} - \mu\right| \geq \lambda\right) \leq \frac{\sigma^2}{\lambda^2}\right). \tag{3.1}$$

[2] Let event $\bar{A}$ occur with probability p, then supplementary event $\bar{X} - \bar{A}$ will occur with probability $1-p$. Let $\widetilde{B_{k,n}}$ be a phenomenon whereby phenomenon $\bar{A}$ occurred precisely k–times out of n independent attempts. Then, the random variable $\widetilde{B_{k,n}}$ has a binomic distribution

$$Prob\left(\widetilde{B_{k,n}}\right) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Mean value of this distribution is equal to

$$\widetilde{E}\left(\widetilde{B_{k,n}}\right) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np$$

and variance is equal to

$$\widetilde{D^2}\left(\widetilde{B_{k,n}}\right) = \widetilde{E}\left(\left(\widetilde{B_{k,n}} - \widetilde{E}\left(\widetilde{B_{k,n}}\right)\right)^2\right) = np(1-p).$$

Therefore, the following formula obviously holds for opposite event

$$Prob\left(\left|\widetilde{A}-2\right|<2\right)=1-\gamma>\frac{1}{2}.$$

Since $\widetilde{A}$ assumes solely positive integer values, it ensues from the inequality $\left|\widetilde{A}-2\right|$ that že $\widetilde{A}$ assumes one of the values $1,2,3$ with a probability greater than $\frac{1}{2}$.

– *q. e. d.* –

Before formulating the lemma, which gives an upper bound combinatorial estimate of the probability of the set $\bar{J}_{\epsilon}^{2m}$, we still have to show special properties of the integral of multiple-variable functions.

Definition 3.1.5 *Let π be permutation of the set $\{1,2,\cdots,n\}$ and $(\vec{\boldsymbol{x}}_1,\vec{\boldsymbol{x}}_2,\cdots,\vec{\boldsymbol{x}}_n)\in\Re^n$. Then define*

$$\vec{\boldsymbol{x}}_\pi\stackrel{def}{=}\left(\vec{\boldsymbol{x}}_{\widetilde{\pi}(1)},\vec{\boldsymbol{x}}_{\widetilde{\pi}(2)},\cdots,\vec{\boldsymbol{x}}_{\widetilde{\pi}(n)}\right)\ .$$

To show special properties of positive functions with permuted variables we need the following lemma

Lemma 3.1.3 *Let $n\geq 1$ be a natural number and*

$$\bar{K}\stackrel{def}{=}\left\{\vec{\boldsymbol{a}}\in\Re^n\,\middle|\,\vec{\boldsymbol{a}}_i\geq 0, i\in\{1,\ldots,n\},\ and\ \sum_{i=1}^{n}\vec{\boldsymbol{a}}_i\leq 1\right\}.$$

Then

$$\vec{\boldsymbol{a}}\in\bar{K}\Rightarrow\sum_{i=1}^{n}i\cdot\vec{\boldsymbol{a}}_i\leq n\ .$$

■ *Proof:*

Obviously, the lemma is true for all vectors $\vec{\boldsymbol{a}}\in\bar{K}$ of the form $\vec{\boldsymbol{a}}=(0,0,\cdots,0,\vec{\boldsymbol{a}}_n)^T$, where $0\leq\vec{\boldsymbol{a}}_n\leq 1$. Further let us assume that there exists an index $j\in\{1,\ldots,n-1\}$ such that $0<\vec{\boldsymbol{a}}_j$. In such a case, $\vec{\boldsymbol{a}}_n\leq 1-\vec{\boldsymbol{a}}_j<1$. Now let us define vector $\vec{\boldsymbol{b}}\in\Re^n$ as

$$\vec{\boldsymbol{b}}_i\stackrel{\text{def}}{=}\begin{cases}\vec{\boldsymbol{a}}_i & i\notin\{j,n\}\\ 0 \quad for & i=j\\ \vec{\boldsymbol{a}}_n+\vec{\boldsymbol{a}}_j & i=n\ .\end{cases}$$

Obviously, $\vec{\boldsymbol{b}}\in\bar{K}$ and

$$\sum_{i=1}^{n}i\cdot\vec{\boldsymbol{b}}_i-\sum_{i=1}^{n}i\cdot\vec{\boldsymbol{a}}_i=(n-j)\,\vec{\boldsymbol{a}}_j>0\ .$$

So the maximum of the sum $\sum_{i=1}^{n}i\cdot\vec{\boldsymbol{a}}_i$ can not be reached at any vector $\vec{\boldsymbol{a}}\in\bar{K}$ except the vector $(0,0,\cdots,0,1)^T$. It is straightforward that this maximal value is equal to n.

– *q. e. d.* –

Lemma 3.1.4 *Let* $\widetilde{f} : \Re^n \to \{0,1\}$ *and* $\widetilde{p} : \Re^n \to \Re$ *be a nonnegative function,* $\int_{\Re^n} \widetilde{p}\,(\vec{\boldsymbol{x}})dx = 1$ *and* $\int_{\Re^n} \widetilde{f}\,(\vec{\boldsymbol{x}})dx$ *exists. Let* $\bar{\Omega}$ *be the set of all permutations of the* $\{1,2,\cdots,n\}$*. Then,*

$$\int_{\Re^n} \sum_{\widetilde{\pi}\in\bar{\Omega}} \left(\widetilde{f}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\widetilde{p}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\right) d\vec{\boldsymbol{x}} \le \max_{\vec{\boldsymbol{x}}\in\Re^n} \left\{\sum_{\widetilde{\pi}\in\bar{\Omega}} \widetilde{f}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\right\} . \tag{3.2}$$

■ *Proof:*
Let us define index set

$$\bar{M} \stackrel{\text{def}}{=} \left\{0,1,2,\cdots, \max_{\vec{\boldsymbol{x}}\in\Re^n} \left\{\sum_{\widetilde{\pi}\in\bar{\Omega}} \widetilde{f}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\right\}\right\}$$

and sets

$$\bar{D}_i \stackrel{\text{def}}{=} \left\{\vec{\boldsymbol{x}} \in \Re^n \,\middle|\, \sum_{\widetilde{\pi}\in\bar{\Omega}} \widetilde{f}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}}) = i\right\} \quad \text{and} \quad \bar{\Pi}_{\vec{\boldsymbol{x}}} \stackrel{\text{def}}{=} \left\{\vec{\boldsymbol{y}} \in \Re^n \,\middle|\, \left(\exists\widetilde{\pi}\in\bar{\Omega}\right)\left(\vec{\boldsymbol{y}} = \vec{\boldsymbol{x}}_{\widetilde{\pi}}\right)\right\} .$$

for any $i \in \bar{M}$ and $\vec{\boldsymbol{x}} \in \Re^n$.
Obviously, $\bar{D}_i \cap \bar{D}_j = \emptyset$, $i \neq j$, $i,j \in \bar{M}$, and $\cup_{i\in\bar{M}} \bar{D}_i = \Re^n$. Further because $\bar{\Omega}$ is the set of all permutations for any $\vec{\boldsymbol{y}} \in \bar{\Pi}_{\vec{\boldsymbol{x}}}$ is $\bar{\Pi}_{\vec{\boldsymbol{y}}} = \bar{\Pi}_{\vec{\boldsymbol{x}}}$. So if $\vec{\boldsymbol{x}} \in \bar{D}_i$ then $\bar{\Pi}_{\vec{\boldsymbol{x}}} \subset \bar{D}_i$.
Further define relation $\mathcal{R}$ as

$$\vec{\boldsymbol{x}}\mathcal{R}\vec{\boldsymbol{y}} \;\Leftrightarrow\; \vec{\boldsymbol{y}} \in \bar{\Pi}_{\vec{\boldsymbol{x}}} ,$$

where $\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \in \Re^n$. Let $\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}, \vec{\boldsymbol{z}} \in \Re^n$. It is straightforward that $\vec{\boldsymbol{x}}\mathcal{R}\vec{\boldsymbol{x}}$ (identity permutation), $\vec{\boldsymbol{x}}\mathcal{R}\vec{\boldsymbol{y}} \Rightarrow \vec{\boldsymbol{y}}\mathcal{R}\vec{\boldsymbol{x}}$ (inverse permutation) and $\vec{\boldsymbol{x}}\mathcal{R}\vec{\boldsymbol{y}}, \vec{\boldsymbol{y}}\mathcal{R}\vec{\boldsymbol{z}} \Rightarrow \vec{\boldsymbol{x}}\mathcal{R}\vec{\boldsymbol{z}}$ (composition of corresponding permutations). In other words, $\mathcal{R}$ is equivalence relation on $\Re^n$ and $\bar{\Pi}_{\vec{\boldsymbol{x}}}$ are equivalence classes of the equivalence $\mathcal{R}$. Further, let

$$\bar{Z}_i \stackrel{\text{def}}{=} \left\{\bar{\Pi}_{\vec{\boldsymbol{x}}} \,\middle|\, \vec{\boldsymbol{x}} \in \bar{D}_i\right\} , \; i \in \bar{M} .$$

Because $\bar{\Pi}_{\vec{\boldsymbol{x}}}$ are equivalence classes, all sets in $\bar{Z}_i$ form disjoint splitting of the set $\bar{D}_i$, e.g.

$$\bar{D}_i = \cup_{\bar{t}\in\bar{Z}_i} \bar{t} , \; i \in \bar{M} . \tag{3.3}$$

Now, for any equivalence class $\bar{t}$ define

$$\bar{\Upsilon}_{\bar{t}} \stackrel{\text{def}}{=} \left\{\vec{\boldsymbol{x}} \in \bar{t} \,\middle|\, \widetilde{f}\,(\vec{\boldsymbol{x}}) = 1\right\} . \tag{3.4}$$

Clearly, if $\vec{\boldsymbol{x}} \in \bar{D}_i$, then size of the set $\bar{\Upsilon}_{\bar{\Pi}_{\vec{\boldsymbol{x}}}}$ is i. Hence

$$\int_{\bar{D}_i} \left[\sum_{\widetilde{\pi}\in\bar{\Omega}} \widetilde{f}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\widetilde{p}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\right] d\vec{\boldsymbol{x}} \stackrel{3.3}{=} \int_{\cup_{\bar{t}\in\bar{Z}_i}\bar{t}} \left[\sum_{\widetilde{\pi}\in\bar{\Omega}} \widetilde{f}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\widetilde{p}\,(\vec{\boldsymbol{x}}_{\widetilde{\pi}})\right] d\vec{\boldsymbol{x}} \stackrel{3.4}{=}$$

$$\stackrel{3.4}{=} \int_{\cup_{\bar{t}\in\bar{Z}_i}\bar{t}} \left[\sum_{\vec{\boldsymbol{x}}\in\bar{\Upsilon}_{\bar{t}}} \widetilde{p}\,(\vec{\boldsymbol{x}})\right] d\vec{\boldsymbol{x}} \stackrel{3.5}{\le} i \cdot \int_{\bar{A}_i} \widetilde{p}\,(\vec{\boldsymbol{x}})d\vec{\boldsymbol{x}} \stackrel{3.5}{\le} i \cdot \int_{\bar{D}_i} \widetilde{p}\,(\vec{\boldsymbol{x}})d\vec{\boldsymbol{x}} ,$$

where the sets $\bar{A}_i$ are defined as

$$\bar{A}_i \stackrel{\text{def}}{=} \left\{ \vec{\boldsymbol{x}} \in \bar{D}_i \,\middle|\, \vec{\boldsymbol{x}} = \max_{\vec{\boldsymbol{y}} \in \bar{\Upsilon}_{\bar{\Pi}_{\vec{\boldsymbol{x}}}}} \{\widetilde{p}\,(\vec{\boldsymbol{y}})\} \right\} \subset \bar{D}_i \ , \ i \in \bar{M} \ . \tag{3.5}$$

Finally, for all $i \in \bar{M} \dot{-} \{0\}$ (note that $\int_{\bar{D}_0} \left(\sum_{\tilde{\pi} \in \bar{\Omega}} \widetilde{f}\,(\vec{\boldsymbol{x}}_{\tilde{\pi}}) \widetilde{p}\,(\vec{\boldsymbol{x}}_{\tilde{\pi}}) d\vec{\boldsymbol{x}} \right)$ is zero) put

$$\alpha_i \stackrel{\text{def}}{=} \int_{\bar{D}_i} \widetilde{p}\,(\vec{\boldsymbol{x}}) d\vec{\boldsymbol{x}} \ .$$

Due to the fact that $\bar{D}_i$ form disjoint splitting of $\Re^n$ and $\int_{\Re^n} \widetilde{p}\,(\vec{\boldsymbol{x}}) d\vec{\boldsymbol{x}} = 1$, we have that $\alpha_i \geq 0$, $i \in \bar{M} \dot{-} \{0\}$, and $\sum_{i=1}^{|\bar{M}|} \alpha_i \leq 1$. If we apply lemma 3.1.3, we get 3.2.

– *q. e. d.* –

Lemma 3.1.5 *Let* R *be a non-empty class concept on* $\bar{X}$ *, and* $\widetilde{P}$ *probability density on* $\bar{X}$ *for which* $\bar{Q}_{m,\epsilon}$ *and* $\bar{J}_{\epsilon}^{2m}$ *are measurable for arbitrary* $m \geq 1$ *and* $\epsilon > 0$*. Then:*

1. *it holds for each* $\epsilon > 0$ *and* $m \geq \frac{2}{\epsilon}$ *that*

$$Prob_{\widetilde{P}^m}\left(\bar{Q}_{m,\epsilon}\right) < 2 Prob_{\widetilde{P}^{2m}}\left(\bar{J}_{\epsilon}^{2m}\right),$$

2. *it holds for each* $\epsilon > 0$ *and* $m \geq 1$ *that*

$$Prob_{\widetilde{P}^{2m}}\left(\bar{J}_{\epsilon}^{2m}\right) \leq \Pi_{\mathsf{R}}\left(2m\right) 2^{-\frac{\epsilon m}{2}}.$$

■ *Proof:*
▪ add 1)
We will show $Prob_{\widetilde{P}^{2m}}\left(\bar{J}_{\epsilon}^{2m}\right) > \frac{1}{2} Prob_{\widetilde{P}^m}\left(\bar{Q}_{m,\epsilon}\right)$. Let $\widetilde{\chi}_{\bar{B}}$ denote a characteristic function of the set $\bar{B}$. Then, according to Fubini's theorem

$$Prob_{\widetilde{P}^{2m}}\left(\bar{J}_{\epsilon}^{2m}\right) = \int_{\bar{X}^{2m}} \widetilde{\chi_{\bar{J}_{\epsilon}^{2m}}}\left(\breve{x}^{2m}\right) d\widetilde{P}^{2m}\left(\breve{x}^{2m}\right) =$$

$$\int_{\breve{x} \in \bar{X}^m} \left(\int_{\breve{y} \in \bar{X}^m} \widetilde{\chi_{\bar{J}_{\epsilon}^{2m}}}\left(\breve{x}, \breve{y}\right) d\widetilde{P}^m\left(\breve{y}\right) \right) d\widetilde{P}^m\left(\breve{x}\right) \geq$$

$$\geq \int_{\breve{x} \in \bar{Q}_{m,\epsilon}} \left(\int_{\breve{y} \in \bar{X}^m} \widetilde{\chi_{\bar{J}_{\epsilon}^{2m}}}\left(\breve{x}, \breve{y}\right) d\widetilde{P}^m\left(\breve{y}\right) \right) d\widetilde{P}^m\left(\breve{x}\right).$$

Let $\breve{x} \in \bar{Q}_{m,\epsilon}$. Then definition of $\bar{Q}_{m,\epsilon}$ follows that there exists an $\bar{r}_{\breve{x}} \in \mathsf{R}_{\widetilde{P},\epsilon}$ such that $\breve{x} \cap \bar{r}_{\breve{x}} = \emptyset$. Further let the set $\bar{K}_{2m,\epsilon}^{\bar{r}_{\breve{x}}}$ be defined as

$$\bar{K}_{2m,\epsilon}^{\bar{r}_{\breve{x}}} \stackrel{\text{def}}{=} \left\{ \left(\breve{x}, \breve{y}\right) \in \bar{X}^m \times \bar{X}^m \,\middle|\, \left(\breve{x} \cap \bar{r}_{\breve{x}} = \emptyset \quad \text{and} \quad \left| \breve{y} \cap \bar{r}_{\breve{x}} \right| \geq \frac{\epsilon m}{2} \right) \right\} .$$

Obviously $\bar{K}^{\bar{r}\breve{\bar{x}}}_{2m,\epsilon} \subset \bar{J}^{2m}_{\epsilon}$, and therefore

$$Prob_{\tilde{P}^{2m}}\left(\bar{J}^{2m}_{\epsilon}\right) \geq \int_{\breve{\bar{x}}\in\bar{Q}_{m,\epsilon}} \left(\int_{\breve{\bar{y}}\in\bar{X}^m} \widetilde{\chi_{\bar{K}^{\bar{r}\breve{\bar{x}}}_{2m,\epsilon}}}\left(\breve{\bar{x}},\breve{\bar{y}}\right) d\tilde{P}^m\left(\breve{\bar{y}}\right)\right) d\tilde{P}^m\left(\breve{\bar{x}}\right).$$

For each fixed $\breve{\bar{x}} \in \bar{Q}_{m,\epsilon}$ the inner integral represents the probability that in m independent selections of bionomically distributed variable there occurs an event with probability ϵ at least $\frac{m\epsilon}{2}$ times. According to lemma 3.1.2, for an arbitrary $m \geq \frac{2}{\epsilon}$, this probability is greater than $\frac{1}{2}$. Hence, we get

$$Prob_{\tilde{P}^{2m}}\left(\bar{J}^{2m}_{\epsilon}\right) > \int_{\breve{\bar{x}}\in\bar{Q}_{m,\epsilon}} \frac{1}{2} d\tilde{P}^m\left(\breve{\bar{x}}\right) = \frac{1}{2} Prob_{\tilde{P}^m}\left(\bar{Q}^m_{\epsilon}\right).$$

- add 2)

Let $\bar{\Omega} \overset{\text{def}}{=} \{\sigma_1, \ldots, \sigma_{2m!}\}$ be the set of all permutation of the set $\{1, \ldots, 2m\}$. Evidently

$$Prob_{\tilde{P}^{2m}}\left(\bar{J}^{2m}_{\epsilon}\right) = \int_{\bar{X}^{2m}} \widetilde{\chi_{\bar{J}^{2m}_{\epsilon}}}\left(\breve{\bar{x}}^{2m}\right) d\tilde{P}^{2m}\left(\breve{\bar{x}}^{2m}\right) = \int_{\bar{X}^{2m}} \widetilde{\chi_{\bar{J}^{2m}_{\epsilon}}}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right) d\tilde{P}^{2m}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right)$$

for all the permutations σ_j, $j \in \{1, \ldots, 2m!\}$, where the symbol $\breve{\bar{x}}^{2m}_{\sigma_j}$ denotes $\breve{\bar{x}}^{2m}$ with a permutated order of elements, and $\widetilde{\chi}(\bar{a})$ is a characteristic function of the set $\bar{a}$. Hence, it ensues that

$$Prob_{\tilde{P}^{2m}}\left(\bar{J}^{2m}_{\epsilon}\right) = \int_{\bar{X}^{2m}} \frac{1}{(2m)!} \sum_{\sigma_j\in\bar{\Omega}} \widetilde{\chi_{\bar{J}^{2m}_{\epsilon}}}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right) d\tilde{P}^{2m}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right).$$

Using the theorem contained in the formula 3.2 and the fact that $\int_{\bar{X}^{2m}} d\tilde{P}^{2m}\left(\breve{\bar{x}}\right) = 1$, we get

$$\int_{\bar{X}^{2m}} \frac{1}{(2m)!} \sum_{\sigma_j\in\bar{\Omega}} \widetilde{\chi_{\bar{J}^{2m}_{\epsilon}}}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right) d\tilde{P}^{2m}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right) \leq \frac{1}{2m!} \max_{\breve{\bar{x}}\in\bar{X}^{2m}} \left\{ \sum_{\sigma_j\in\bar{\Omega}} \widetilde{\chi_{\bar{J}^{2m}_{\epsilon}}}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right) \right\}. \tag{3.6}$$

It is, therefore, sufficient for proof if we demonstrate the validity of the estimate

$$\frac{1}{2m!} \sum_{\sigma_j\in\bar{\Omega}} \widetilde{\chi_{\bar{J}^{2m}_{\epsilon}}}\left(\breve{\bar{x}}^{2m}_{\sigma_j}\right) \leq \Pi_{\mathsf{R}}(2m)\, 2^{-\frac{\epsilon m}{2}}$$

for any arbitrary $\breve{\bar{x}}^{2m} \in \bar{X}^{2m}$.
Let us take fixed $\breve{\bar{x}} \in \bar{X}^{2m}$ and define the set

$$\bar{\Psi}_{\breve{\bar{x}}} \overset{\text{def}}{=} \left\{ \bar{r} \in \mathsf{R}_{\tilde{P},\epsilon} \,\middle|\, (\exists\sigma \in \bar{\Omega}) \left(\text{if } \breve{\bar{x}}^{2m}_{\sigma} \overset{\text{def}}{=} \left(\breve{\bar{z}},\breve{\bar{y}}\right), \text{then} \quad \bar{r}\cap\breve{\bar{z}} = \emptyset \quad \text{and} \quad \left|\bar{r}\cap\breve{\bar{y}}\right| \geq \frac{m\epsilon}{2} \right) \right\}.$$

Further, let us define for each $\bar{r} \in \bar{\Psi}_{\breve{\bar{x}}}$ the set

$$\bar{\Theta}_{\breve{\bar{x}},\bar{r}} \overset{\text{def}}{=} \left\{ \sigma \in \bar{\Omega} \,\middle|\, \text{if } \breve{\bar{x}}^{2m}_{\sigma} \overset{\text{def}}{=} \left(\breve{\bar{z}},\breve{\bar{y}}\right), \text{then} \quad \bar{r}\cap\breve{\bar{z}} = \emptyset \quad \text{and} \quad \left|\bar{r}\cap\breve{\bar{y}}\right| \geq \frac{m\epsilon}{2} \right\}. \tag{3.7}$$

Let us establish cardinality of the set $\bar{\Theta}_{\breve{x},\bar{r}}$. Let us denote $l \stackrel{\text{def}}{=} \left|\bar{r} \cap \breve{x}\right| = \left|\bar{r} \cap \breve{y}\right|$. Let us now discuss the question in how many ways it is possible to permutate elements in $\breve{x}$ in a way to meet the conditions of 3.7 (for an $\bar{r}$ given fixed). The number of these possibilities is evidently equal to the number of ways in which l points may be placed in m positions in $\breve{y}$ multiplied by the number of permutations $(2m-l)$ of points not belonging to $\bar{r}$, multiplied by the number of permutations of l points that belong to $\bar{r}$ and which are already in selected positions $\breve{y}$. Hence,

$$\left|\bar{\Theta}_{\breve{x},\bar{r}}\right| = \binom{m}{l} \cdot (2m-l)! \cdot l! \ . \tag{3.8}$$

Further, the number of different points in $\breve{x}$ at most equals the number $2m$, and it follows that with the help of all sets in $\mathsf{R}_{\widetilde{P},\epsilon}$ we can create at best $\Pi_{\mathsf{R}}(2m)$ of different sets in the form of $\bar{r} \cap \breve{x}$. Hence, the estimate holds (in fact it is $\left\{\bar{v} \left| \left(\exists \bar{r} \in \bar{\Psi}_{\breve{x}}\right)\left(\bar{v} = \bar{r} \cap \breve{x}\right)\right.\right\} \subset \Pi_{\mathsf{R}}\left(\breve{x}\right)$)

$$\left|\bar{\Psi}_{\breve{x}}\right| \leq \Pi_{\mathsf{R}}(2m) \, .$$

We are concerned, as our target, with the number of permutations σ such for which $\breve{x}_{\sigma}^{2m} \in \bar{J}_{\epsilon}^{2m}$. However, this number equals the cardinality of the set $\bigcup_{\bar{r} \in \bar{\Psi}_{\breve{x}}} \bar{\Theta}_{\breve{x},\bar{r}}$, so we can write

$$\sum_{\sigma_j \in \bar{\Omega}} \widetilde{\chi_{\bar{J}_{\epsilon}^{2m}}}\left(\breve{x}_{\sigma_j}^{2m}\right) \leq \left|\bigcup_{\bar{r} \in \bar{\Psi}_{\breve{x}}} \bar{\Theta}_{\breve{x},\bar{r}}\right| \leq \left|\bar{\Theta}_{\breve{x},\bar{r}}\right|\left|\bar{\Psi}_{\breve{x}}\right| \, .$$

If we recall the definition of the set $\bar{J}_{\epsilon}^{2m}$ (see 3.1.2), we can see that $l = \left|\bar{r} \cap \breve{y}\right| \geq \frac{m\epsilon}{2}$. So we get

$$\frac{1}{(2m)!}\left|\bar{\Theta}_{\breve{x},\bar{r}}\right|\left|\bar{\Psi}_{\breve{x}}\right| \stackrel{3.8}{=} \frac{\binom{m}{l}}{\binom{2m}{l}}\left|\bar{\Psi}_{\breve{x}}\right| = \frac{m(m-1)\cdots(m-l+1)}{2m(2m-1)\cdots(2m-l+1)}\left|\bar{\Psi}_{\breve{x}}\right| \leq \frac{1}{2^l}\left|\bar{\Psi}_{\breve{x}}\right| \leq \frac{1}{2^{\frac{m\epsilon}{2}}}\left|\bar{\Psi}_{\breve{x}}\right|$$

which directly implies that

$$\frac{1}{(2m)!}\sum_{\sigma_j \in \bar{\Omega}} \widetilde{\chi_{\bar{J}_{\epsilon}^{2m}}}\left(\breve{x}_{\sigma_j}^{2m}\right) \leq \Pi_{\mathsf{R}}(2m)\, 2^{-\frac{m\epsilon}{2}} .$$

Because $\breve{x}$ was chosen arbitrarily,

$$\frac{1}{2m!}\max_{\breve{x} \in \bar{X}^{2m}}\left\{\sum_{\sigma_j \in \bar{\Omega}} \widetilde{\chi_{\bar{J}_{\epsilon}^{2m}}}\left(\breve{x}_{\sigma_j}^{2m}\right)\right\} \leq \Pi_{\mathsf{R}}(2m)\, 2^{-\frac{m\epsilon}{2}} \ ,$$

which together with 3.6 completes the proof.

– *q. e. d.* –

The previous lemma makes it possible to obtain upper estimate of probability of an event when, based on a random sample, we do not obtain ϵ-transversal for the sets in error sets R .

It is evident, on the basis of lemma 2.1.3, that if the VC-dimension of the hypothesis class H is finite, then the number $\Pi_{\mathsf{H}}(m)$ is polynomially bounded in view of m. That is why the asymptomatic behavior of the expression $\Pi_{\mathsf{H}}(2m)\,2^{-\frac{\epsilon m}{2}}$ in the neighborhood of infinity is given by the exponential member, and obviously this value approaches zero for large values of m very rapidly. Now we will estimate the size m (hence length of the sample) so that the expression $2\Pi_{\mathsf{H}}(2m)\,2^{-\frac{\epsilon m}{2}}$ is smaller than the number δ.

Lemma 3.1.6 *Let α, β, m be positive numbers, $m \geq \alpha\widetilde{\log_2}(e)$ [3] and $m \geq \alpha\widetilde{\log_2}(\beta m)$. Then,*

$$(\forall t > 0)\left(m + t > \alpha\widetilde{\log_2}(\beta(m+t))\right) \quad . \tag{3.9}$$

■ *Proof:*
Let us define a function $\widetilde{f}(m) \stackrel{\text{def}}{=} m - \alpha\widetilde{\log_2}(\beta m)$. Then,

$$\widetilde{f}'(m) = 1 - \frac{\alpha\widetilde{\log_2}(e)}{m}.$$

Hence

$$\widetilde{f}'(m) = 0 \quad \Rightarrow \quad m = \alpha\widetilde{\log_2}(e)$$

and obviously for all $m > \alpha\widetilde{\log_2}(e)$ the function $\widetilde{f}(m)$ is monotone increasing. At the same time, $\widetilde{f}(m)$ is positive.

– *q. e. d.* –

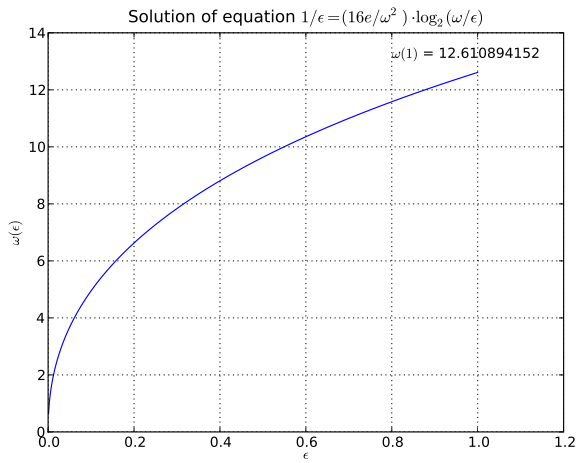
Lemma 3.1.7 *Let α, β, ϵ be positive numbers, $\frac{1}{\epsilon} \geq \alpha\widetilde{\log_2}(e)$ and $\frac{1}{\epsilon} \geq \alpha\widetilde{\log_2}\left(\frac{\beta}{\epsilon}\right)$. Then,*

$$(\forall t \in (0, \epsilon))\left(\frac{1}{\epsilon - t} > \alpha\widetilde{\log_2}\left(\frac{\beta}{\epsilon - t}\right)\right) \quad . \tag{3.10}$$

■ *Proof:*
It is obvious that $\frac{1}{\epsilon - t} = \frac{1}{\epsilon} + \frac{t}{(\epsilon - t)\epsilon}$. Let $m = \frac{1}{\epsilon}$. For such m prepositions of the lemma 3.1.6 are satisfied, which follows the proof.

– *q. e. d.* –

[3] $\widetilde{\log_2}(e) = \frac{1}{\widetilde{\ln}(2)} = \simeq 1.44269504088896$



	+ 0.0	+0.01	+0.02	+0.03	+0.04
0.00	–	1.81	2.46	2.94	3.33
0.05	3.67	3.97	4.25	4.50	4.73
0.10	4.95	5.15	5.35	5.53	5.71
0.15	5.88	6.04	6.19	6.34	6.49
0.20	6.63	6.76	6.90	7.02	7.15
0.25	7.27	7.39	7.51	7.62	7.73
0.30	7.84	7.94	8.05	8.15	8.25
0.35	8.35	8.44	8.54	8.63	8.72
0.40	8.81	8.90	8.98	9.07	9.15
0.45	9.24	9.32	9.40	9.48	9.56

Figure 3.1: Graph of the solution of equation $\frac{1}{\epsilon} = \frac{16e}{\omega^2(\epsilon)}\log_2\left(\frac{\omega(\epsilon)}{\epsilon}\right)$. The corresponding source code in python language is listed in Appendix 4.1.1.

Lemma 3.1.8 *Let m be a natural number, $0 < \epsilon < 1$ and*

$$m \geq \max\left\{\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right\}. \tag{3.11}$$

Then, $2\Phi_{d,2m}2^{-\frac{\epsilon m}{2}} \leq \delta$.

■ *Proof:*
The lemma 2.1.3 postulates the inequality $\Phi_{d,2m} \leq \left(\frac{2em}{d}\right)^d$, thus it suffices to show that $2\left(\frac{2em}{d}\right)^d \leq \delta 2^{\frac{m\epsilon}{2}}$. This is equivalent to the inequality

$$\frac{m\epsilon}{2} \geq d\log_2\left(\frac{2me}{d}\right) + \log_2\left(\frac{2}{\delta}\right).$$

The first condition on m in 3.11 implies $\frac{m\epsilon}{4} \geq \log_2\left(\frac{2}{\delta}\right)$, so the rest of the proof is reduced to showing

$$\frac{m\epsilon}{4} \geq d\log_2\left(\frac{2me}{d}\right). \tag{3.12}$$

To do so, let us define (see graph on Figure 3.1)

$$\alpha_\epsilon \stackrel{\text{def}}{=} \frac{16e}{12.611^2}, \qquad \beta_\epsilon \stackrel{\text{def}}{=} 12.611\ .$$

It is obvious that

$$(\forall\epsilon \in (0,1\rangle)\left(\frac{1}{\epsilon} \geq 1 > \frac{16e}{12.611^2}\log_2(e) = \alpha_\epsilon\log_2(e) \simeq 0.394539011\right). \tag{3.13}$$

At the same time numerical calculation shows (again see graph on Figure 3.1) that for $\epsilon = 1$ is

$$\frac{1}{\epsilon} \geq \frac{16e}{12.611^2}\log_2\left(\frac{12.611}{\epsilon}\right). \tag{3.14}$$

Inequalities 3.13 and 3.14 follows that both assumptions of the lemma 3.1.7 are true and hence the inequality 3.14 is valid for any $\epsilon \in (0,1\rangle$.
Further let us define

$$\alpha_m \stackrel{\text{def}}{=} \frac{4d}{\epsilon}, \qquad \beta_m \stackrel{\text{def}}{=} \frac{2e}{d}, \qquad m = \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right).$$

Hence we get for any $\epsilon \in (0,1\rangle$

$$\frac{1}{\epsilon} \geq \frac{16e}{12.611^2}\log_2\left(\frac{12.611}{\epsilon}\right) \Leftrightarrow 2\log_2\left(\frac{12.611}{\epsilon}\right) \geq \log_2\left(\frac{16e}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right) \Leftrightarrow$$

$$\Leftrightarrow \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right) \geq \frac{4d}{\epsilon}\log_2\left(\frac{16e}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right) \Leftrightarrow m \geq \alpha_m \log_2(\beta_m m). \quad (3.15)$$

At the same time, for any $0 < \epsilon < 1$ it is straightforward that

$$m = \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right) \geq \frac{4d}{\epsilon}\log_2(e) = \alpha_m \log_2(e). \quad (3.16)$$

Resembly as in the previous case inequalities 3.15 and 3.16 follows that both assumptions of the lemma 3.1.6 are true and hence the last inequality 3.15 is valid for any

$$m \geq \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right).$$

But

$$m \geq \alpha_m \log_2(\beta_m m) \Leftrightarrow \frac{m\epsilon}{4} \geq d\log_2\left(\frac{2me}{d}\right),$$

and finally

$$\frac{m\epsilon}{2} = \frac{m\epsilon}{4} + \frac{m\epsilon}{4} \geq d\log_2\left(\frac{2me}{d}\right) + \log_2\left(\frac{2}{\delta}\right).$$

– *q. e. d.* –

Lemma 3.1.9 *Let the matrix* $\boldsymbol{Z}$ *be a real matrix of the type* $m \times n$ *and let it hold that* $(\forall i \in \{1,\ldots,m\})\ \left(\sum_{j=1}^{n} \frac{\boldsymbol{Z}_{i,j}}{n} > \alpha\right)$. *Then, there exists* $j_0 \in \{1,\ldots,n\}$ *so that*

$$\sum_{i=1}^{m} \frac{\boldsymbol{Z}_{i,j_0}}{m} > \alpha.$$

■ *Proof:*
Proof will be made by means of contradiction. It ensues from the assumption of the lemma that $\sum_{i=1}^{m}\sum_{j=1}^{n} \boldsymbol{Z}_{i,j} > \alpha nm$. If the lemma's theorem did not hold, then – on the contrary – $\sum_{j=1}^{n}\sum_{i=1}^{m} \boldsymbol{Z}_{i,j} < \alpha mn$ would have to hold, which is a contradiction.

– *q. e. d.* –

Lemma 3.1.10 *Let* $\bar{X} \stackrel{def}{=} \{x_1, \ldots, x_d\}$ *be a finite set,* $k \in \{1, \ldots, d\}$ *, and* $0 \le \omega \le 1$ *Further let*

$$\mathsf{P} \stackrel{def}{=} \left\{ \bar{a} \subset \bar{X} \,||\bar{a}| \ge k \right\} \quad \textit{and} \quad \mathsf{Q} \stackrel{def}{=} \left\{ \bar{b} \subset \bar{X} \,\middle|\, 1 \le \left|\bar{b}\right| < k \right\} .$$

Let us randomly select a set $\bar{s} \subset \bar{X}$ *and let the probability of getting* $\bar{s} \in \mathsf{P}$ *be equal to* ω*. Finally, let us denote the average size of selected sets as* ρ*. Then,*

$$\rho \le k \cdot (1 - \omega) + d \cdot \omega \, . \tag{3.17}$$

Further, let on the set $\bar{X}$ *be defined uniform probability* $\widetilde{P}$ *and let* ϑ *denotes average probability of the selected sets. Than*

$$\vartheta \le \frac{k}{d} \cdot (1 - \omega) + \omega \, . \tag{3.18}$$

■ *Proof:*
Denote selected sets as $\mathsf{C} \stackrel{\text{def}}{=} \{\bar{c}_1, \ldots, \bar{c}_p\}$, where p is the number of trials. Obviously

$$\rho = \frac{\sum_{i=1}^{p} |\bar{c}_i|}{p} = \frac{\sum_{\bar{c}_i \in \mathsf{Q}} |\bar{c}_i|}{p} + \frac{\sum_{\bar{c}_i \in \mathsf{P}} |\bar{c}_i|}{p} \le$$

$$\le \frac{k \cdot |\{i \,|\bar{c}_i \in \mathsf{Q}\}|}{p} + \frac{d \cdot |\{i \,|\bar{c}_i \in \mathsf{P}\}|}{p} = k \cdot (1 - \omega) + d \cdot \omega \, .$$

The proof of the second estimation is straightforward.

– *q. e. d.* –

Lemma 3.1.11 *Let for any* $\vec{\boldsymbol{z}} \in \{0, 1\}^n$ *the symbol* $n_+(\vec{\boldsymbol{z}})$ *denotes the number of ones in* $\vec{\boldsymbol{z}}$ *. Than*

$$\frac{1}{2^n} \sum_{\vec{\boldsymbol{z}} \in \{0,1\}^n} n_+(\vec{\boldsymbol{z}}) = \frac{n}{2} \, .$$

■ *Proof:*
Let $\vec{\boldsymbol{o}} \in \{0, 1\}^n$ is defined as $\vec{\boldsymbol{o}} \stackrel{\text{def}}{=} (1, 1, \cdots, 1)$ and let $\bar{A}, \bar{B}$ be sets such that for all $\vec{\boldsymbol{a}} \in \bar{A}$ is the vector $-(\vec{\boldsymbol{a}} - \vec{\boldsymbol{o}}) \in \bar{B}$ and for all $\vec{\boldsymbol{b}} \in \bar{B}$ is the vector $-\left(\vec{\boldsymbol{b}} - \vec{\boldsymbol{o}}\right) \in \bar{A}$. Hence the sets $\bar{A}$, $\bar{B}$ form disjoint splitting of the set $\{0, 1\}^n$ and in addition $\left|\bar{A}\right| = \left|\bar{B}\right|$. So we have

$$\frac{1}{2^n} \sum_{\vec{\boldsymbol{z}} \in \{0,1\}^n} n_+(\vec{\boldsymbol{z}}) = \frac{1}{2^n} \sum_{\vec{\boldsymbol{a}} \in \bar{A}} n_+(\vec{\boldsymbol{a}}) - n_+(\vec{\boldsymbol{a}} - \vec{\boldsymbol{o}}) = \frac{1}{2^n} \sum_{\vec{\boldsymbol{a}} \in \bar{A}} n = \frac{1}{2^n} \cdot \frac{2^n}{2} \cdot n = \frac{n}{2} \, .$$

– *q. e. d.* –

The lemmas and theorems postulated above make it possible directly to prove the fundamental theorem of this part. However, let us omit in our considerations the trivial cases described in the following definition.

Definition 3.1.6 *Concept class* C *defined over the set* $\bar{X}$ *is called* NONTRIVIAL CONCEPT CLASS *iff*

$$(\exists \bar{c}_1, \bar{c}_2 \in \mathsf{C}) \quad \textit{such that} \quad \bar{c}_1 \neq \bar{c}_2 \quad \textit{and} \quad \left(\bar{c}_1 \cap \bar{c}_2 \neq \emptyset \quad \textit{or} \quad \bar{c}_1 \cup \bar{c}_2 \neq \bar{X}\right) .$$

Concept class is called TRIVIAL CONCEPT CLASS *in other cases.*

So what is a trivial concept class? If we negate definition condition, we get

$$(\forall \bar{c}_1, \bar{c}_2 \in \mathsf{C}) \left(\bar{c}_1 = \bar{c}_2 \quad \text{or} \quad \left(\bar{c}_1 \cap \bar{c}_2 = \emptyset \quad \text{and} \quad \bar{c}_1 \cup \bar{c}_2 = \bar{X}\right)\right) .$$

Clearly, we can reformulate the definition 3.1.6 in the form that the concept class is trivial if and only if it consists of one concept only, or has two concepts whose form disjoint splitting of the set $\bar{X}$.

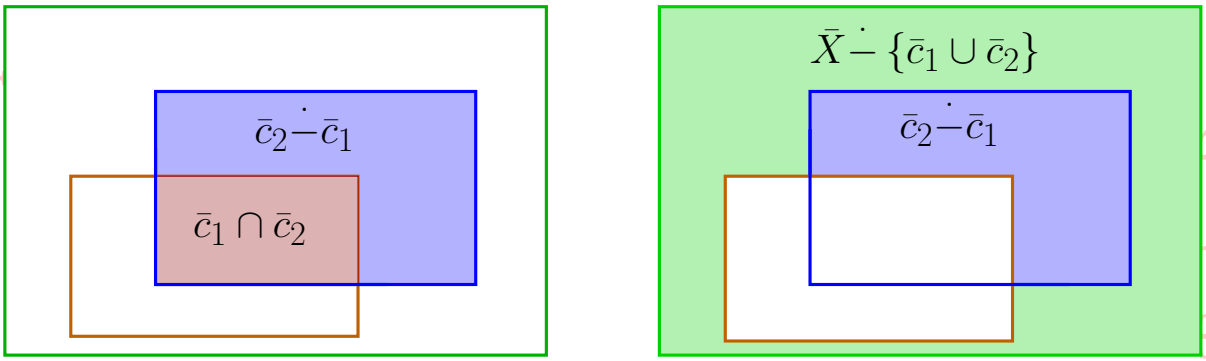


Figure 3.2: Two necessary minimal contents of nontrivial concept classes (stained areas must be nonempty sets).

It can be expected that we focus on nontrivial classes. The next theorem postulates upper bound on necessary queries that have to be provided by an environment (we mean some kind of supervision) to learning algoritm to be an (ϵ, δ)-algorithm, which is the main result of the standard PAC learning theory.

Theorem 3.1.12 *Let* C *be a nontrivial, well-behaved concept class. Then, the following holds:*

1. *If* $\mathtt{VC}_{dim}(\mathsf{C}) = d$ *and* $d < \infty$ *then*

   (a) *for any* $0 < \epsilon < \frac{1}{2}$ *there is no* (ϵ, δ)*-learning algorithm which exploits less than*

   $$\max\left(\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right), d\left(1 - 2\left(\epsilon\left(1-\delta\right) + \delta\right)\right)\right) \tag{3.19}$$

   queries.

   (b) *for arbitrary* $0 < \epsilon < 1$*, any learning algorithm using at least*

   $$\max\left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right) \tag{3.20}$$

   queries and returning a consistent hypothesis only is an (ϵ, δ)*-learning algorithm.*

2. C *is uniformly learnable if and only if* $\mathtt{VC}_{dim}(\mathsf{C}) < \infty$.

■ *Proof:*

Let C be a nontrivial concept class and let $\mathtt{VC}_{dim}(\mathsf{C}) = d$ be finite.

▪ add Proof of 3.19)

Let $0 < \epsilon < \frac{1}{2}$ and $m \stackrel{\text{def}}{=} \max\left(\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right), d\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)\right)$. We have to prove that any arbitrary learning algorithm for C must exploit at least m queries. Proof will be divided into the following steps:

Proof for the first member of the estimate: Proof of the estimate given by the first member in 3.19 will be made for the following two cases:

Case 1: C contains at least two different concepts $\bar{c}_1$ and $\bar{c}_2$ which have a non-empty intersection. Let A^* be a learning algorithm for C and let us denote $\bar{a} \stackrel{\text{def}}{=} \bar{c}_1 \cap \bar{c}_2$ and $\bar{b} \stackrel{\text{def}}{=} \bar{c}_2 - \bar{c}_1$. Let $\widetilde{P}$ be such a probability density on $\bar{X}$ that $Prob_{\widetilde{P}}\left(\bar{b}\right) \stackrel{\text{def}}{=} \epsilon$, $Prob_{\widetilde{P}}\left(\bar{a}\right) = 1-\epsilon$ and let the probability of all the other elements from $\bar{X}$ be zero. In view of the probability density thus defined, we can assume in further text – without loss of generality – that $\bar{X} \stackrel{\text{def}}{=} \bar{a} \cup \bar{b}$, $\mathsf{C} \stackrel{\text{def}}{=} \left\{\bar{a}, \bar{X}\right\}$ and $\mathsf{H} \stackrel{\text{def}}{=} \left\{\emptyset, \bar{a}, \bar{b}, \bar{X}\right\}$.

According to the first bound of the number of selected samples, $m \leq \frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right)$. Using the inequality [4]

$$\frac{1-\epsilon}{\epsilon} < \frac{-1}{\ln\left(1-\epsilon\right)},$$

we get

$$m < \frac{-1}{\ln\left(1-\epsilon\right)}\ln\left(\frac{1}{\delta}\right),$$

$$m\ln\left(1-\epsilon\right) > \ln\left(\delta\right),$$

$$\left(1-\epsilon\right)^m > \delta.$$

The last inequality says that the probability of selecting all the samples from the set $\bar{a}$ is greater than δ.

All the possible learning algorithms for the concept class C can obviously be divided into two groups:

[4] Power series of the function $\ln\left(1-\epsilon\right)$ is

$$\ln\left(1-\epsilon\right) = -\epsilon - \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} - \frac{\epsilon^4}{4}\ldots, \quad \text{for} \quad -1 < \epsilon < 1.$$

Taking into mind that this power series is absolute convergent (we can multiply it by any polynomial and apply arbitrary bracketing), we have

$$\left(1-\epsilon\right)\ln\left(1-\epsilon\right) = -\epsilon + \left(\epsilon^2 - \frac{\epsilon^2}{2}\right) + \left(\frac{\epsilon^3}{2} - \frac{\epsilon^3}{3}\right) + \left(\frac{\epsilon^4}{3} - \frac{\epsilon^4}{4}\right) + \ldots > -\epsilon,$$

because $0 < \epsilon < 1$, and hence all the numbers in brackets are positive. So $\left(1-\epsilon\right)\ln\left(1-\epsilon\right) > -\epsilon$, which can be expressed as

$$\frac{1-\epsilon}{\epsilon} < \frac{-1}{\ln\left(1-\epsilon\right)}.$$

1. algorithms which, in case of selecting precisely m samples from $\bar{a}$, generate hypothesis $\bar{a}$. Such algorithms produce for the target concept $\bar{X}$ a hypothesis with an error equal to $Prob_{\widetilde{P}}\left(\bar{b}\right) = \epsilon$.
2. algorithms, which – in this instance – generate hypothesis $\bar{b}$, $\bar{X}$ or $\emptyset$. However, since – according to the assumption of the theorem $\epsilon < \frac{1}{2}$ – such algorithms produce, in case of the target concept $\bar{a}$, an error at least equal to ϵ.

We have, therefore, shown that there exists the concept class C such that for any learning algorithm A^* there exists the target concept ($\bar{x}$ in the group 1.) and $\bar{a}$ in the group 2.)) for which we get the hypothesis with error greater than ϵ with probability greater than δ. Thus, there cannot exist an (ϵ, δ)-learning algorithm for C, requiring less than $\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right)$ queries.

Case 2: C contains at least two different concepts $\bar{c}_1$ and $\bar{c}_2$ for which $\bar{c}_1 \cup \bar{c}_2 \neq \bar{X}$. Let us put $\bar{a} \stackrel{\text{def}}{=} \bar{X} \dot{-} (\bar{c}_1 \cap \bar{c}_2)$ and $\bar{b} \stackrel{\text{def}}{=} \bar{c}_1$. Let us define probability density $\widetilde{P}$ just as in the previous case, e.g. $Prob_{\widetilde{P}}\left(\bar{b}\right) \stackrel{\text{def}}{=} \epsilon$, $Prob_{\widetilde{P}}\left(\bar{a}\right) = 1 - \epsilon$ and the probability of all the other elements from $\bar{X}$ is zero. Put $\bar{X} \stackrel{\text{def}}{=} \bar{a} \cup \bar{b}$, $\mathsf{C} \stackrel{\text{def}}{=} \left\{\bar{b}, \emptyset\right\}$, $\mathsf{H} \stackrel{\text{def}}{=} \left\{\emptyset, \bar{a}, \bar{b}, \bar{X}\right\}$. Further, let us take into consideration random selection of m negative samples which are all from the set $\bar{a}$. The rest of the proof is in the same manner as in the preceding case.

Proof for the second member of the estimate: Now what remains to be done is to prove the validity of the second estimate in the expression 3.19. Since C is a nontrivial class concept, $\mathtt{VC}_{dim}(\mathsf{C}) = d$ is equal at least to 1. Therefore, there exists a d-element subset $\bar{\Gamma} \subset \bar{X}$, which is shattered by the class concept C. Let probability density $\widetilde{P}$ be uniform on the set $\bar{\Gamma}$ and zero on its complement to the set $\bar{X}$ (thanks to this selection the entire proof is based solely on the combinatorial properties of finite sets). In view of this probability density we can again put – without loss of to generality – $\bar{X} \stackrel{\text{def}}{=} \bar{\Gamma}$ and $\mathsf{C} \stackrel{\text{def}}{=} 2^{\bar{\Gamma}}$. Let us assume that some learning algorithm A^* has obtained a sample of the length of m elements from the set $\bar{X}$, and let us denote this sample $\left(\breve{x}, \vec{\boldsymbol{z}}\right)$, $\breve{x} \in \bar{X}^m$, while the number of mutually different elements x_i in the sample $\breve{x}$ is l. Let us define the set system of all concepts consistent with the sample $\left(\breve{x}, \vec{\boldsymbol{z}}\right)$

$$\mathsf{B}_{\left(\breve{x}, \vec{\boldsymbol{z}}\right)} \stackrel{\text{def}}{=} \left\{\bar{c} \in \mathsf{C} \,\middle|\, \bar{c} \quad \text{is consistent with} \quad \left(\breve{x}, \vec{\boldsymbol{z}}\right)\right\}.$$

If we subtract $\breve{x}$ from $\bar{X}$, there remains a set of the size $(d - l)$, which has 2^{d-l} subsets. Each of these subsets, unified with $\breve{x}$, is a consistent hypothesis for the sample $\breve{x}$. Thus, $\left|\mathsf{B}_{\left(\breve{x}, \vec{\boldsymbol{z}}\right)}\right| = 2^{d-l}$. Let us consider that the algorithm A^* always produce consistent hypotheses. Therefore

$$\left(\exists \bar{h} \in \mathsf{B}_{\left(\breve{x}, \vec{\boldsymbol{z}}\right)}\right)\left(\forall \bar{c} \in \mathsf{B}_{\left(\breve{x}, \vec{\boldsymbol{z}}\right)}\right)\left(\widetilde{\mathrm{A}^*}\left(\left(\breve{x}, \vec{\boldsymbol{z}}\right) = \bar{h}\right)\right),$$

in the other words, for the sample $\left(\breve{x}, \vec{\boldsymbol{z}}\right)$ the algorithm A^* must generate a hypothesis from $\mathsf{B}_{\left(\breve{x}, \vec{\boldsymbol{z}}\right)}$ identical one for all the target concepts from $\mathsf{B}_{\left(\breve{x}, \vec{\boldsymbol{z}}\right)}$ (the

result of the algorithm depends solely on its definition and on the sample used). The number of elements in which this hypothesis can differ from the target concept equals $d-l$. To explain this fact in a detail we can denote those elements as $\{y_1,\dots,y_{d-l}\} = \bar{X} - \breve{x}$. Obviously, all points y_i can be placed into sets $\bar{h}\cap\bar{c}$, $\bar{X}-\left(\bar{h}\cup\bar{c}\right)$ and $\bar{h}\triangle\bar{c}$ arbitrarily. Therefore all subset of $\{y_1,\dots,y_{d-l}\}$ are contained inside such as outside of sets $\bar{h}\triangle\bar{c}$ for $\bar{c}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}$. Because the probability of each element y_i is equal to the value $\frac{1}{d}$ we can estimate mean error of the generated hypothesis over all the target concepts in $\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}$ as (see lemma 3.1.11)

$$\frac{1}{2^{d-l}}\sum_{\bar{c}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}}\frac{\left|\bar{h}\triangle\bar{c}\right|}{d}=\frac{1}{2^{d-l}}\sum_{\vec{\boldsymbol{z}}\in\{0,1\}^{d-l}}\frac{n_+\left(\vec{\boldsymbol{z}}\right)}{d}\overset{3.1.11}{=}\frac{d-l}{2d}\geq\frac{d-m}{2d}\,.$$

For this moment let us define for each target concept $\bar{c}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}$ and each hypothesis $\bar{h}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}$ the number

$$\boldsymbol{Z}_{\bar{c},\bar{h}}\overset{\text{def}}{=}\mathsf{e}_{\widetilde{P}}\left(\bar{c},\bar{h}\right).$$

As we explained above, obviously it holds that

$$\left(\forall\bar{h}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}\right)\left(\sum_{\bar{c}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}}\frac{\boldsymbol{Z}_{\bar{c},\bar{h}}}{\left|\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}\right|}>\frac{d-m}{2d}\right).$$

According to (lemma 3.1.9) there exists target concept $\bar{b}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}$ with average error at least $\frac{d-m}{2d}$ (in the sense that for random choice of $\breve{x}\in\bar{X}^m$ the average error of generated hypotheses is at least $\frac{d-m}{2d}$. In other words, there exists a concept in relation to which all the consistent hypotheses have mean error ρ

$$\rho\geq\frac{d-m}{2d}\,. \tag{3.21}$$

Finally let us define

$$\mathsf{E}\overset{\text{def}}{=}\left\{\bar{h}\in\mathsf{B}_{\left(\breve{x},\vec{\boldsymbol{z}}\right)}\,\middle|\,\mathsf{e}_{\widetilde{P}}\left(\bar{b},\bar{h}\right)\geq\epsilon\right\}\;,\;\bar{S}\overset{\text{def}}{=}\left\{\left(\breve{x},\vec{\boldsymbol{z}}\right)\text{ is a sample of }\bar{b}\text{ and }\breve{x}\in X^m\right\},$$

and

$$\mathsf{P}\overset{\text{def}}{=}\left\{\left(\breve{x},\vec{\boldsymbol{z}}\right)\in\bar{S}\,\middle|\,\widetilde{\mathrm{A}^*}\left(\left(\breve{x},\vec{\boldsymbol{z}}\right)\right)\in\mathsf{E}\right\}\quad,\quad\mathsf{Q}\overset{\text{def}}{=}\left\{\left(\breve{x},\vec{\boldsymbol{z}}\right)\in\bar{S}\,\middle|\,\widetilde{\mathrm{A}^*}\left(\left(\breve{x},\vec{\boldsymbol{z}}\right)\right)\notin\mathsf{E}\right\}.$$

Let us assume, that the probability of the set P be equal to $\delta\in(0,1)$ and recall that all hypotheses in the set system E are produced by the learning algorithm A^* as a response to the samples from P. So $Prob_{\widetilde{P}}(\mathsf{P})=Prob_{\widetilde{P}}(\mathsf{E})$. Now let us assume that $Prob_{\widetilde{P}}(\mathsf{E})=\delta-\gamma$, where $\gamma\in(0,\delta)$. Hence we can estimate average probability of the produced hypothesis over all samples from the set $\bar{S}$ using the lemma 3.1.10 (note that the probability on $\bar{X}$ is assumed to be uniform) and the 3.21 as

$$\epsilon(1-\delta)+\delta>\epsilon(1-(\delta-\gamma))+(\delta-\gamma)\overset{3.18}{\geq}\rho\overset{3.21}{\geq}\frac{d-m}{2d}\,.$$

It follows that

$$m > d(1 - 2(\epsilon(1-\delta)+\delta)\,,$$

which contradicts the second part of upperbound in the assumption 3.19. So $Prob_{\tilde{P}}(\mathsf{E}) \geq \delta$.

▪ add Proof of the estimate 3.20)

Let $\bar{c} \in \mathsf{C}$ be fixed and let

$$m \geq \max\left\{\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right\}$$

and

$$\bar{Q}_{m,\epsilon} = \left\{\breve{x} \in \bar{X}^m \,\middle|\, \left(\exists \bar{r} \in \mathsf{R}_{\tilde{P},\epsilon}\right)\left(\breve{x} \cap \bar{r} = \emptyset\right)\right\}.$$

If we recall the definiton 3.1.2, the lemma 3.1.5, the theorem 2.4.4 and the lemma 3.1.8 we get

$$Prob_{\tilde{P}}\left(\bar{Q}_{m,\epsilon}\right) \overset{3.1.5}{\leq} 2\Pi_{\mathsf{R}}(2m)\,2^{-\frac{\epsilon m}{2}} \overset{2.4.4}{=} 2\Pi_{\mathsf{H}}(2m)\,2^{-\frac{\epsilon m}{2}} \overset{3.1.8}{\leq} \delta\,,$$

where $\mathsf{R} \overset{\text{def}}{=} \left\{\bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathsf{H}\right\}$. Due to the fact that the algorithm A^* produces consistent hypotheses only we can write

$$\bar{h} = \widetilde{\mathrm{A}^*}\left(\left(\breve{x}, \vec{\boldsymbol{z}}\right)\right) \quad \text{and} \quad \mathsf{e}_{\tilde{P}}\left(\bar{c}, \bar{h}\right) > \epsilon \qquad \Rightarrow \qquad \breve{x} \in \bar{Q}_{m,\epsilon}\,.$$

But the probability of the set $\bar{Q}_{m,\epsilon}$ is less than ϵ which concludes the proof of this part.

▪ add Proof of the part 2)

Let us assume that $\mathtt{VC}_{dim}(\mathsf{C}) \overset{\text{def}}{=} +\infty$. Obviously, in this case it is possible for every natural d to perform a sequence of steps in the proof of the second part of the estimate 3.19. However, if we select the values ϵ and δ such that the expression $(1 - 2(\epsilon(1-\delta)+\delta))$ is positive, we get for this particular selection of ϵ, δ that there exists no (ϵ,δ)-learning algorithm for any fixed chosen length of samples.

Conversely, on the basis of the validity of the estimate 3.20 it is possible to construct an (ϵ,δ)-learning algorithm in such a way that we first well-order, in whatever fashion, the hypotheses class H (each set can be well-ordered, see the Zermelo's theorem, e.g. [PBKN90], p. 44, and let us further recall that in case of well-ordering, there exist no incomparable elements in the set). Then, we assign to each sample from $\bar{S}_{\mathsf{C}}$ the first concept (in view of the actual ordering) which is consistent with it. That this sample exists ensues from the inclusion $\mathsf{C} \subset \mathsf{H}$. According to 3.20, this algorithm is an (ϵ,δ)-learning algorithm.

– *q. e. d.* –

Obviously, this theorem is much stronger than the theorem 1.2.1 since the value $|\mathsf{C}|$ in the upper estimate has now been replaced by the value $\mathtt{VC}_{dim}(\mathsf{C})$ which may be substantially smaller. And in addition, this theorem is apllicable in the case of infinite set $\bar{X}$ and infinite concept class C.

Example 3.1.3 *Let $\bar{X}$, C and A^* be defined in the same manner as in the example1.2.1. Hence $\mathtt{VC}_{dim}(\mathsf{C}) = 1$ and if*

$$m \geq \left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right)$$

then the learning algorithm A^* *is* (ϵ,δ)*-learning algorithm.*

Example 3.1.4 *Let* $\bar{X}$, C *and* A^* *be defined in the same manner as in the example1.2.2. Hence* $\mathtt{VC}_{dim}(\mathsf{C}) = 2n$ *and if*

$$m \geq \left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{16n}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right)$$

then the learning algorithm A^* *is* (ϵ,δ)*-learning algorithm.*

Note that in the both cases discussed above the bellow bound on the m is independent on the number k (e.g. on the size of the space $\bar{X}$).

3.1.1 Delta rule learning algorithm

As an example of consistent learning algorithm, let us now study the so-called delta rule algorithm which is able to find a separation hyperplane for a given tuple of linearly separable sets of Euclidean space.

Definition 3.1.7 *Let* $(\vec{\boldsymbol{x}}_1, y_1), \ldots, (\vec{\boldsymbol{x}}_m, y_m)$ *be a given sequence of tuples in* $\Re^n \times \{-1,+1\}$, $t \geq 1$. *Further, let vector's sequence* $\{\vec{\boldsymbol{w}}_i\}_1^\infty$ *satisfy the following recursive formulas*

1. *put* $\vec{\boldsymbol{w}}_1 \overset{def}{=} \vec{\boldsymbol{0}}$, $k = 1$
2. *let* $k = k+1$ *and* $\bar{J} \overset{def}{=} \{j \in \{1,\ldots,m\} \,|\, \widetilde{sgn}\left(\langle \vec{\boldsymbol{w}}_k \,|\, \vec{\boldsymbol{x}}_j \rangle\right) \neq y_j\}$
   (a) *if* $\bar{J} = \emptyset$ *put* $\vec{\boldsymbol{w}}_{k+1} = \vec{\boldsymbol{w}}_k$ *and STOP,*
   (b) *else let* $j_k \in \bar{J}$ *be arbitrary. Then put*

$$\vec{\boldsymbol{w}}_{k+1} \overset{def}{=} \vec{\boldsymbol{w}}_k + y_{j_k}\vec{\boldsymbol{x}}_{j_k}$$

   and REPEAT step 2).

Then we say that this sequence arose by application of DELTA RULE *on* $(\vec{\boldsymbol{x}}_1, y_1), \ldots, (\vec{\boldsymbol{x}}_m, y_m)$.

A basic properties of the delta rule algorithm sumarizes the next theorem:

Theorem 3.1.13 *Assume that sequence* $\{\vec{\boldsymbol{w}}_i\}_1^\infty$ *arose by application of delta rule and let there exists a vector* $\widehat{\vec{\boldsymbol{w}}}$ *such that for all indexes* $i \in \{1,\ldots,m\}$ *holds* $\widetilde{sgn}\left(\left\langle \widehat{\vec{\boldsymbol{w}}} \,|\, \vec{\boldsymbol{x}}_i \right\rangle\right) = y_i$. *Further let*

$$\alpha \overset{def}{=} \max_{i\in\{1,\ldots,m\}} \left\{\|\vec{\boldsymbol{x}}_i\|^2\right\} \quad \text{and} \quad \beta \overset{def}{=} \min_{i\in\{1,\ldots,m\}} \left\{\left|\left\langle \widehat{\vec{\boldsymbol{w}}} \,|\, \vec{\boldsymbol{x}}_i \right\rangle\right|\right\} > 0\,.$$

Then there exists an natural number $z > 0$ *satisfying* $\vec{\boldsymbol{w}}_{z+1} = \vec{\boldsymbol{w}}_z$ *and* z *can be estimated as*

$$z \leq \frac{\alpha\left\|\widehat{\vec{\boldsymbol{w}}}\right\|^2}{\beta^2} + 1\,.$$

■ *Proof:*
By definition of delta rule it is obvious that for arbitrary $k > 1$ for which $\vec{\boldsymbol{w}}_k \neq \vec{\boldsymbol{w}}_{k-1}$ holds

$$\vec{\boldsymbol{w}}_k = \sum_{p=1}^{k-1} \ddot{\vec{\boldsymbol{x}}}_{j_p}, \quad \text{where} \quad \ddot{\vec{\boldsymbol{x}}}_{j_p} \stackrel{\text{def}}{=} \begin{cases} \vec{\boldsymbol{x}}_{j_p} & \text{pro} \quad y_{j_p} = 1 \\ -\vec{\boldsymbol{x}}_{j_p} & \text{pro} \quad y_{j_p} = -1 \end{cases}.$$

Further let us assume that $k > t$. By the fact that $\widetilde{sgn}\left(\left\langle \widehat{\vec{\boldsymbol{w}}} \left| \vec{\boldsymbol{x}}_{j_p} \right. \right\rangle\right) \neq y_{j_p}$ and by definition of $\ddot{\vec{\boldsymbol{x}}}_{j_p}$ the inequality $\left\langle \widehat{\vec{\boldsymbol{w}}} \left| \ddot{\vec{\boldsymbol{x}}}_{j_p} \right. \right\rangle > 0$ is true for all vectors $\ddot{\vec{\boldsymbol{x}}}_{j_p}$. Thus we can write the estimation

$$\left\langle \widehat{\vec{\boldsymbol{w}}} \left| \vec{\boldsymbol{w}}_k \right. \right\rangle = \sum_{p=1}^{k-1} \left\langle \widehat{\vec{\boldsymbol{w}}} \left| \ddot{\vec{\boldsymbol{x}}}_{j_p} \right. \right\rangle \geq (k-1)\,\beta > 0.$$

Hence, we get $\left|\left\langle \widehat{\vec{\boldsymbol{w}}} \left| \vec{\boldsymbol{w}}_k \right. \right\rangle\right|^2 \geq (k-1)^2\,\beta^2$. Further, using Schwartz inequality $\left|\left\langle \vec{\boldsymbol{a}} \left| \vec{\boldsymbol{b}} \right. \right\rangle\right|^2 \leq \|a\|^2 \|b\|^2$ we finally show

$$\frac{\beta^2\,(k-1)^2}{\left\|\widehat{\vec{\boldsymbol{w}}}\right\|^2} \leq \|\vec{\boldsymbol{w}}_k\|^2. \tag{3.22}$$

But, at the same time, for any k is

$$\vec{\boldsymbol{w}}_k = \vec{\boldsymbol{w}}_{k-1} + \ddot{\vec{\boldsymbol{x}}}_{j_{k-1}},$$

and after square

$$\|\vec{\boldsymbol{w}}_k\|^2 = \|\vec{\boldsymbol{w}}_{k-1}\|^2 + 2\left\langle \vec{\boldsymbol{w}}_{k-1} \left| \ddot{\vec{\boldsymbol{x}}}_{j_{k-1}} \right. \right\rangle + \left\|\ddot{\vec{\boldsymbol{x}}}_{j_{k-1}}\right\|^2.$$

By properties of delta rule algorithm must be $\widetilde{sgn}\left(\left\langle \vec{\boldsymbol{w}}_{k-1} \left| \vec{\boldsymbol{x}}_{j_{k-1}} \right. \right\rangle\right) \neq y_{j_{k-1}}$ and therefore $\left\langle \vec{\boldsymbol{w}}_{k-1} \left| \ddot{\vec{\boldsymbol{x}}}_{j_{k-1}} \right. \right\rangle \leq 0$. Hence, the previous inequality one could rewrite to the form

$$\|\vec{\boldsymbol{w}}_k\|^2 - \|\vec{\boldsymbol{w}}_{k-1}\|^2 \leq \left\|\ddot{\vec{\boldsymbol{x}}}_{j_{k-1}}\right\|^2. \tag{3.23}$$

Now, let as sum up equations 3.23 over all $j \in \{1, \ldots, k\}$ (remember that $\vec{\boldsymbol{w}}_1 = \vec{\boldsymbol{0}}$). We get

$$\|\vec{\boldsymbol{w}}_k\|^2 \leq \sum_{p=1}^{k-1} \left\|\ddot{\vec{\boldsymbol{x}}}_{j_p}\right\|^2 \leq (k-1)\,\alpha.$$

Blended together the last inequality and inequality 3.22 we can coclude

$$\frac{\beta^2\,(k-1)^2}{\left\|\widehat{\vec{\boldsymbol{w}}}\right\|^2} \leq (k-1)\,\alpha$$

and finally

$$k - 1 \leq \frac{\alpha \left\|\widehat{\vec{\boldsymbol{w}}}\right\|^2}{\beta^2}.$$

There is a constant independent on the value of k on the right side of the last expression. It follows that an $z > 0$ mentioned in the statement of the theorem must exists.

– *q. e. d.* –

Our chiev interest is to illustrate by examples the PAC learning model. Reconsider now the definition of delta rule learning algorithm. Let us change step *2.(b)* in definition 3.1.7 by *"else let* $j_k = \min\left\{\bar{J}\right\}$. *Then put"* . To do it so, we get deterministic version of delta rule algorithm and we can apply bellow bound to sufficient number of examples used by (ϵ, δ)-algorithm. Obviously, each vector $\vec{\boldsymbol{w}}$ produced by delta rule algorithm is homogenous linear separator of positive and negative patterns in the sequence $\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, \cdots, \vec{\boldsymbol{x}}_m$. Hence corresponding concept class C is the set of all halfspaces with zero vector in its border and $\mathsf{C} = \mathsf{H}$. As we already shown, $\mathtt{VC}_{dim}(\mathsf{C}) = n$, which implies that we need at least

$$m \geq \max\left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8n}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right)$$

to keep the delta rule process to be (ϵ, δ)-algorithm.

3.1.2 Lover bound for maximal steps of delta rule algorithm.

From another point of view the statement of the Theorem 3.1.13 only says that the delta rule algorithm is finite process without any posibility to estimate the number of necessary iterations. Unfortunately, as we explain in the following text the number of delta rule iterations can not be upper bounded by polynomial in n.

Definition 3.1.8 *Let* $\bar{A}, \bar{B} \subset \Re^n$. *Then the tuple* $(\vec{\boldsymbol{w}}, t) \in \Re^n \times \Re$ *is* LINEAR SEPARATOR *of sets* $\bar{A}$, $\bar{B}$ *iff*

$$\left(\forall \vec{\boldsymbol{a}} \in \bar{A}\right)\left(\langle \vec{\boldsymbol{a}} \,|\vec{\boldsymbol{w}}\rangle < t\right) \quad \textit{and} \quad \left(\forall \vec{\boldsymbol{b}} \in \bar{B}\right)\left(\left\langle \vec{\boldsymbol{b}} \,|\vec{\boldsymbol{w}}\right\rangle > t\right) .$$

Lemma 3.1.14 *Let* $\bar{A} \subset \{-1,+1\}^n$, $\bar{B} \stackrel{def}{=} \{-1,+1\}^n \dot{-} \bar{A}$ *and* $(\vec{\boldsymbol{w}}, t)$ *be a linear separator of sets* $\bar{A}$ *and* $\bar{B}$. *Then there exists a linear separator* $(\vec{\boldsymbol{w}}^*, t)$ *of sets* $\bar{A}$, $\bar{B}$ *such that*

$$\left(\forall \vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \in \{-1,+1\}^n\right)\left(\vec{\boldsymbol{x}} \neq \vec{\boldsymbol{y}} \Rightarrow \langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{x}}\rangle \neq \langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{y}}\rangle\right) . \tag{3.24}$$

■ *Proof:*

Let $(\vec{\boldsymbol{w}}, t)$ be an arbitrary linear separator of sets $\bar{A}$ and $\bar{B}$. Further, let there exists $\vec{\boldsymbol{y}}, \vec{\boldsymbol{z}} \in \{-1,+1\}^n$ such that $\vec{\boldsymbol{x}} \neq \vec{\boldsymbol{y}}$ and $\langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{x}}\rangle = \langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{y}}\rangle$.
Now let us define a positive number β as

$$0 < 2\beta < \min_{\vec{\boldsymbol{r}}, \vec{\boldsymbol{s}} \in \{-1,+1\}^n} \left\{\left|\langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{r}}\rangle - \langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{s}}\rangle\right| > 0\right\} .$$

Because $\vec{\boldsymbol{x}} \neq \vec{\boldsymbol{y}}$, there exists an index k for which $\vec{\boldsymbol{x}}_k \neq \vec{\boldsymbol{y}}_k$. Put $\vec{\boldsymbol{w}^*} = \vec{\boldsymbol{w}} + \beta\vec{\boldsymbol{e}}_k$, where $\vec{\boldsymbol{e}}_k$ has all coordinates zero except k-th, which is 1. Due to definition of number β the tuples $(\vec{\boldsymbol{w}}^*, t)$ is linear separator of sets $\bar{A}$ and $\bar{B}$ and at the same time

$$\left(\forall \vec{\boldsymbol{r}}, \vec{\boldsymbol{s}} \in \{-1,+1\}^n\right)\left(\langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{r}}\rangle \neq \langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{s}}\rangle \Rightarrow \langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{r}}\rangle \neq \langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{s}}\rangle\right)$$

and

$$|\langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{x}}\rangle - \langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{y}}\rangle| \geq 2\beta \;,$$

so the number of tuples $\vec{\boldsymbol{y}}, \vec{\boldsymbol{z}}$ for which $\langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{x}}\rangle = \langle \vec{\boldsymbol{w}} \,|\vec{\boldsymbol{y}}\rangle$ is decreased at least by 1. Obviously we can repeat this construction till all such tuples $\vec{\boldsymbol{y}}, \vec{\boldsymbol{z}}$ are exhausted.

– *q. e. d.* –

Based on the previous lemma we can define recursive algorithm which allows us to postulate exponential lower bound on the number of hypercube vertices dichotomies. Thist recursive algorithm is illustrated on the Fig. 3.3.

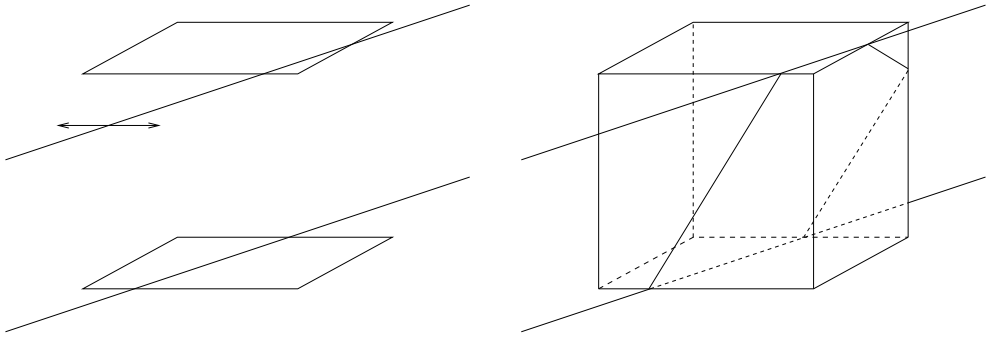


Figure 3.3: Construction of $2^k{+}1$ dichotomies derived from a fixed dichotomy of $\{-1,+1\}^k$ (for $k=2$).

Theorem 3.1.15 *The number of linearly separable dichotomies of the cube* $\{-1,+1\}^n$ *is larger than* $2^{\frac{n(n-1)}{2}}$.

■ *Proof:*
We proceed the proof by induction on n. The case $n=1$ is clear because there exist 4 linearly separable dichotomies of $\{-1,+1\}$ exactly. So we suppose that $n \geq 1$ and that statement holds for this n. Let $\left(\bar{A}, \bar{B}\right)$ be a fixed linearly separable dichotomy of the cube $\{-1,+1\}^n$. As follows from the lemma 3.1.14 we can assume that corresponding linear separator $(\vec{\boldsymbol{w}}^*, t)$ fullfils the condition 3.24. So there exists 2^n mutually different real numbers ω_i, $i \in \{1,\ldots,2^n\}$, such that $\omega_i < \omega_{i+1}$, $i \in \{1,\ldots,2^n-1\}$, and

$$\{\omega_1, \omega_2, \cdots, \omega_{2^n}\} = \{\omega \,|\omega = \langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{x}}\rangle,\ \vec{\boldsymbol{x}} \in \{-1,+1\}^n\}$$

Let us define

$$t_0 \stackrel{\text{def}}{=} \omega_1 - 1, \qquad t_i \stackrel{\text{def}}{=} \frac{\omega_i + \omega_{i+1}}{2},\ i \in \{1,\ldots,2^n-1\}, \qquad t_{2^n} \stackrel{\text{def}}{=} \omega_{2^n} + 1,$$

$$\bar{A}^- \stackrel{\text{def}}{=} \left\{(\vec{\boldsymbol{a}}, -1) \in \{-1,+1\}^{n+1} \,\middle|\, \vec{\boldsymbol{a}} \in \bar{A}\right\}, \qquad \bar{B}^- \stackrel{\text{def}}{=} \left\{\left(\vec{\boldsymbol{b}}, -1\right) \in \{-1,+1\}^{n+1} \,\middle|\, \vec{\boldsymbol{b}} \in \bar{B}\right\},$$

and let for all $i \in \{0,\ldots,2^n\}$ is

$$\bar{A}^+_{t_i} \stackrel{\text{def}}{=} \left\{(\vec{\boldsymbol{x}}, +1) \in \{-1,+1\}^{n+1} \,|\langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{x}}\rangle < t_i\right\}, \quad \bar{B}^+_{t_i} \stackrel{\text{def}}{=} \left\{(\vec{\boldsymbol{x}}, +1) \in \{-1,+1\}^{n+1} \,|\langle \vec{\boldsymbol{w}}^* \,|\vec{\boldsymbol{x}}\rangle > t_i\right\}$$

and

$$\bar{C}_i^A = \bar{A}^- \cup \bar{A}_{t_i}^+ \quad , \quad \bar{C}_i^B = \bar{B}^- \cup \bar{B}_{t_i}^+.$$

Obviously, a tuple $\left(\bar{C}_i^A, \bar{C}_i^B\right)$ is a dichotomy of the cube $\{-1,+1\}^{n+1}$. Further we show that

$$\left(\left(\vec{\boldsymbol{w}}^*, \frac{t-t_i}{2}\right), \frac{t+t_i}{2}\right). \tag{3.25}$$

is a linear separator of $\left(\bar{C}_i^A, \bar{C}_i^B\right)$.

Let $\vec{\boldsymbol{x}} \in \bar{A}^-$. Then $\left(\exists \vec{\boldsymbol{a}} \in \bar{A}\right)$ $\left(\vec{\boldsymbol{x}} = (\vec{\boldsymbol{a}}, -1)\right)$. Hence

$$\left\langle \left(\vec{\boldsymbol{w}}^*, \frac{t-t_i}{2}\right) | (\vec{\boldsymbol{a}}, -1) \right\rangle = \langle \vec{\boldsymbol{w}}^* | \vec{\boldsymbol{a}} \rangle - \frac{t-t_i}{2} < t - \frac{t-t_i}{2} = \frac{t+t_i}{2}.$$

Let $\vec{\boldsymbol{x}} \in \bar{A}_{t_i}^+$. Then $\left(\exists \vec{\boldsymbol{y}} \in \{-1,+1\}^n\right)$ $\left(\vec{\boldsymbol{x}} = (\vec{\boldsymbol{y}}, 1) \text{ and } \langle \vec{\boldsymbol{w}}^* | \vec{\boldsymbol{y}} \rangle < t_i\right)$. Hence

$$\left\langle \left(\vec{\boldsymbol{w}}^*, \frac{t-t_i}{2}\right) | (\vec{\boldsymbol{y}}, 1) \right\rangle = \langle \vec{\boldsymbol{w}}^* | \vec{\boldsymbol{y}} \rangle + \frac{t-t_i}{2} < t_i + \frac{t-t_i}{2} = \frac{t+t_i}{2}.$$

Let $\vec{\boldsymbol{x}} \in \bar{B}^-$. Then $\left(\exists \vec{\boldsymbol{b}} \in \bar{B}\right)$ $\left(\vec{\boldsymbol{x}} = \left(\vec{\boldsymbol{b}}, -1\right)\right)$. Hence

$$\left\langle \left(\vec{\boldsymbol{w}}^*, \frac{t-t_i}{2}\right) \middle| \left(\vec{\boldsymbol{b}}, -1\right) \right\rangle = \left\langle \vec{\boldsymbol{w}}^* \middle| \vec{\boldsymbol{b}} \right\rangle - \frac{t-t_i}{2} > t - \frac{t-t_i}{2} = \frac{t+t_i}{2}.$$

Let $\vec{\boldsymbol{x}} \in \bar{B}_{t_i}^+$. Then $\left(\exists \vec{\boldsymbol{y}} \in \{-1,+1\}^n\right)$ $\left(\vec{\boldsymbol{x}} = (\vec{\boldsymbol{y}}, 1) \text{ and } \langle \vec{\boldsymbol{w}}^* | \vec{\boldsymbol{y}} \rangle > t_i\right)$. Hence

$$\left\langle \left(\vec{\boldsymbol{w}}^*, \frac{t-t_i}{2}\right) | (\vec{\boldsymbol{y}}, 1) \right\rangle = \langle \vec{\boldsymbol{w}}^* | \vec{\boldsymbol{y}} \rangle + \frac{t-t_i}{2} > t_i + \frac{t-t_i}{2} = \frac{t+t_i}{2}.$$

The previous four inequalities follows, that the tuple 3.25 is a linear separator of the $\left(\bar{C}_i^A, \bar{C}_i^B\right)$. Thus for a given dichotomy $\left(\bar{A}, \bar{B}\right)$ of $\{-1,+1\}^n$ we obtained 2^n+1 mutually different dichotomies of the $\{-1,+1\}^{n+1}$. In addition, it is obvious that for different dichotomies $\left(\bar{A}', \bar{B}'\right)$ of $\{-1,+1\}^n$ derived dichotomies of $\{-1,+1\}^{n+1}$ differs in corresponding sets $\bar{A}^-$ and $\bar{B}^-$.

Finally, let K_n denotes the number of all linearly separable dichotomies of the set $\{-1,+1\}^n$. Hence we proved recursive formula $K_{n+1} \geq (2^n+1)K_n$. As we mentioned $K_1 = 4$, so

$$K_n \geq \prod_{i=0}^{n-1} \left(2^i + 1\right) > \prod_{i=0}^{n-1} 2^i = 2^{\frac{n(n-1)}{2}}.$$

– *q. e. d.* –

The previous theorem follows the fact that there exists linearly separable splitting of the cube $\{-1,+1\}^n$ with integer linear separator whose indices can not be upper bounded by any polynomial in n.

Theorem 3.1.16 *There exists a linearly separable dichotomy of the* $\{-1,+1\}^n$ *such that any integer linear separator* $(\vec{\boldsymbol{w}}, t)$ *of this dichotomy satisfies estimation*

$$2^{\frac{n-2}{2}} \leq \sum_{k=1}^{n} |\vec{\boldsymbol{w}}_k| + |t| \,.$$

■ *Proof:*
Let us prove bellow bound. Denote

$$\bar{P}_n \stackrel{\text{def}}{=} \left\{ \left(\bar{A}, \bar{B}\right) \middle| \left(\bar{A}, \bar{B}\right) \text{ is linearly separable splitting of } \{-1,+1\}^n \right\}.$$

Further let for all $\left(\bar{A}, \bar{B}\right) \in \bar{P}_n$ is

$$\bar{W}_{\left(\bar{A},\bar{B}\right)} \stackrel{\text{def}}{=} \left\{ (\vec{\boldsymbol{w}}, t) \in \Re^n \times \Re \middle| (\vec{\boldsymbol{w}}, t) \text{ is integer linear separator of } \left(\bar{A}, \bar{B}\right) \right\}.$$

Finally define integer function $\widetilde{\zeta}(n)$ as

$$\widetilde{\zeta}(n) \stackrel{\text{def}}{=} \max_{\left(\bar{A},\bar{B}\right)\in\bar{P}_n} \left\{ \min_{(\vec{\boldsymbol{w}},t)\in\bar{W}_{\left(\bar{A},\bar{B}\right)}} \left\{ \left\lceil \log_2 \left(\sum_{k=1}^{n} |\vec{\boldsymbol{w}}_k| + |t| \right) \right\rceil \right\} \right\}.$$

Definition of the function $\widetilde{\zeta}(n)$ follows that for any dichotomy $\left(\bar{A}, \bar{B}\right) \in \bar{P}_n$ there exists an integer linear separator $(\vec{\boldsymbol{w}}, t) \in \bar{W}_{\left(\bar{A},\bar{B}\right)}$ which has all indices bounded to the interval $\left(-2^{\widetilde{\zeta}(n)}, 2^{\widetilde{\zeta}(n)}\right)$.

Using binary notation we need at most $\widetilde{\zeta}(n) + 1$ bits (including sign) to store each indices of the linear separator. Hence, to store whole linear separator, e.g. $n+1$ indices, $(n+1)(\widetilde{\zeta}(n)+1)$ bits is sufficient.

This amount of bits allow to store at most $2^{(n+1)(\widetilde{\zeta}(n)+1)}$ different linear separators. At the same time, by the theorem 3.1.15, there exists more than $2^{\frac{n(n-1)}{2}}$ different linearly separable dichotomies. Hence we get inequality

$$2^{(n+1)(\widetilde{\zeta}(n)+1)} \geq 2^{\frac{n(n-1)}{2}}.$$

Therefore $(n+1)(\widetilde{\zeta}(n)+1) \geq \frac{n(n-1)}{2}$ and

$$\widetilde{\zeta}(n) + 1 \geq \frac{n(n-1)}{2(n+1)} = \frac{n-2}{2} + \frac{1}{n+1} > \frac{n-2}{2}.$$

.

– *q. e. d.* –

Example 3.1.5 *Let* $\bar{X} \stackrel{def}{=} \{-1,+1\}^n$ *and*

$$\mathsf{C} = \left\{ \bar{A} \subset \{-1,+1\}^n \middle| \bar{A} \text{ and } \left(\{-1,+1\}^n \dot{-} \bar{A} \right) \text{ are linearly separable} \right\}.$$

Let $\bar{c} \in \mathsf{C}$ *be such a set, that any integer linear separator* $(\vec{\boldsymbol{w}}, t)$ *of the sets* $\bar{c}$ *and* $\{-1,+1\}^n \dot{-} \bar{c}$ *satisfies* $2^{\frac{n-2}{2}} \leq \sum_{k=1}^{n} |\vec{\boldsymbol{w}}_k| + |t|$. *Let us use delta rule as a learning algorithm to find* $(\vec{\boldsymbol{w}}, t)$. *The algorithm stops when return a consistent hypothesis and this happens after finite number of steps. But in each step of delta rule algorithm absolute values of the weight vector* $\vec{\boldsymbol{w}}$ *can increase by one only, so we need at least* $\frac{2^{\frac{n-2}{n}}}{n+1}$ *steps of delta rule algorithm.*

3.1.3 Linear separation and linear programming

Definition 3.1.9 (Mangasarian LP) *Let* $\bar{A} \stackrel{def}{=} \{\vec{a}_1, \ldots, \vec{a}_i\}$ *and* $\bar{B} \stackrel{def}{=} \left\{\vec{b}_1, \ldots, \vec{b}_j\right\}$ *be a finite subsets of the* $\Re^n$ *. Then* MANGASARIAN LINEAR PROBLEM *is defined as the problem find vectors* $\vec{y} \in \Re^i$, $\vec{z} \in \Re^j$, $\vec{w} \in \Re^n$ *and* $t \in \Re$ *that minimizes*

$$\sum_{\alpha=1}^{i} \vec{y}_\alpha + \sum_{\beta=1}^{j} \vec{z}_\beta$$

subject to

$$\begin{array}{lcll} \vec{y}_\alpha + \langle \vec{w} \,|\vec{a}_\alpha \rangle - t & \geq & 1 \quad for \quad \alpha \in \{1, \ldots, i\} \\ \vec{z}_\beta - \left\langle \vec{w} \left| \vec{b}_\beta \right. \right\rangle + t & \geq & 1 \quad for \quad \beta \in \{1, \ldots, j\} \\ \vec{y}_\alpha & \geq & 0 \quad for \quad \alpha \in \{1, \ldots, i\} \\ \vec{z}_\beta & \geq & 0 \quad for \quad \beta \in \{1, \ldots, j\} \; . \end{array}$$

Theorem 3.1.17 *Let* $\bar{A} \stackrel{def}{=} \{\vec{a}_1, \ldots, \vec{a}_i\}$ *and* $\bar{B} \stackrel{def}{=} \left\{\vec{b}_1, \ldots, \vec{b}_j\right\}$ *be a finite subsets of the* $\Re^n$ *. Then*

1. *There exists a linear separator of the sets* $\bar{A}$ *and* $\bar{B}$ *if and only if the optimal value of the corresponding Mangasarian LP is zero.*
2. *If the optimal value of the corresponding Mangasarian LP is zero and* $(\vec{y}^*, \vec{z}^*, \vec{w}^*, t^*)$ *is optimal solution, than* $(\vec{w}^*, t^*)$ *is linear separator of the sets* $\bar{A}$ *and* $\bar{B}$ *.*

■ *Proof:*
▪ add 1 $\Rightarrow$)
Let $(\vec{w}, t)$ be linear separator of the sets $\bar{A}$ and $\bar{B}$. It means that $\langle \vec{w} \,|\vec{a}_\alpha \rangle - t > 0$, $\alpha \in \{1, \ldots, i\}$, and $\left\langle \vec{w} \left| \vec{b}_\beta \right. \right\rangle - t < 0$, $\beta \in \{1, \ldots, j\}$. Let

$$\omega \stackrel{\text{def}}{=} \min \left\{ \min \left\{ \langle \vec{w} \,|\vec{a}_\alpha \rangle - t \,| \alpha \in \{1, \ldots, i\} \right\}, \min \left\{ -\left\langle \vec{w} \left| \vec{b}_\beta \right. \right\rangle + t \,| \beta \in \{1, \ldots, j\} \right\} \right\} \; .$$

Clearly, $\omega > 0$ and $\langle \vec{w} \,|\vec{a}_\alpha \rangle - t \geq \omega$, $\alpha \in \{1, \ldots, i\}$, and $\left\langle \vec{w} \left| \vec{b}_\beta \right. \right\rangle - t \leq \omega$. So $\left\langle \frac{\vec{w}}{\omega} \,|\vec{a}_\alpha \right\rangle - \frac{t}{\omega} \geq 1$, $\alpha \in \{1, \ldots, i\}$, $\left\langle \frac{\vec{w}}{\omega} \left| \vec{b}_\beta \right. \right\rangle - \frac{t}{\omega} \leq 1$. So $\left(\vec{0}, \vec{0}, \frac{\vec{w}}{\omega}, \frac{t}{\omega}\right)$ is the solution of the corresponding Mangasarian LP with optimal value 0.
▪ add 1 $\Leftarrow$ and 2)
Let $(\vec{y}^*, \vec{z}^*, \vec{w}^*, t^*)$ is an optimal solution with zero optimal value. It follows that $\vec{y} = \vec{0}$ and $\vec{z} = \vec{0}$. Hence a tuple $(\vec{w}, t)$ is a linear separator of the sets $\bar{A}$ and $\bar{B}$.

– *q. e. d.* –

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Chapter 4

Appendices

4.1 Source codes

4.1.1 Python code for $\omega(\epsilon)$ solver

The python code listed bellow were used to compute table and graph on the Fig. 3.1.

```
#!/usr/bin/python

import string, os, os.path, time, re, random, shutil, sys
import matplotlib.pyplot as plt
import numpy as np
import math

omega = []
epsilon = []

eps = 0.001
step = 0.001
e = 2.71828182845905
log2e = 1.442695040888963387

result = 1.0

while eps <= 1.0 + step :
   epsilon.append(eps)
   iter = result
   residum = 1000000000
   while residum > 0.00000000001 : # Newton iterations method $x_{n+1}=x_n-f(x)/f'(x)
      value = 1/eps - log2e*math.log(iter/eps)*16*e/iter/iter
      derivative = (2*math.log(iter/eps) - 1.0)*log2e*16*e/iter/iter/iter
      iter = iter - value/derivative

      residum = abs(1/eps - log2e*math.log(iter/eps)*16*e/iter/iter)

   result = iter
   omega.append(result)
   eps += step
```


© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

```
#########  COMPUTE TABLE  #########
def get_index(r,c) : # r row , c column
   value = c*0.01 + r*0.05
   index = int(value*1000)
   return index-1

text_row = u'0.00  '
col = 0
while col < 5 :
   text_row = text_row + " & %8s"%(col*0.01)
   col += 1
print "%s \\\\ \\hline"%(text_row)

text_row = u'0.00  '
row = 0
col = 0
while row < 10 :
   while col < 5 :
      index = get_index(row,col)
      res = unicode("%.2f"%(omega[index]))
      if col == 0 and row == 0 : res = u'--'
      text_row = text_row + " & %8s"%(res)
      col += 1
   col = 0
   print "%s \\\\"%(text_row)
   text_row = "%.2f  "%((float(row)+1)/20)
   row += 1
print "\\hline"


#########  MAKE PLOT  #########
plt.title(u'Solution of equation $1/\epsilon=(16e/\omega^2)\cdot\
 \log_2(\omega/\epsilon)$')

plt.xlabel(u'$\epsilon$')
plt.ylabel(u'$\omega(\epsilon)$')
plt.text(0.8,13.0,u'$\omega(1)$ = 12.610894152')
plt.grid(True)
plt.plot(epsilon,omega)

plt.savefig(u'omega-as-function-epsilon.pdf',
    dpi=75, facecolor='w', edgecolor='w', orientation='portrait',
    papertype='a4',format=None,transparent=False)

plt.show()
```


© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

© F. Hakl, ICS CAS, Prague, Tech. Rep. №1227, Dec 2015

Bibliography

[AB92] Martin Anthony and Norman Biggs. *Computational Learning Theory*. Press Syndicate of the University of Cambridge, 1992.

[And85] Jiří Anděl. *Matematická statistika*. SNTL - ALFA Bratislava, 1985.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the Association for Computing Machinery*, 36:929–965, oct 1989.

[Cov] Thomas M. Cover. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transaction on Electronic Computers*, pages 326–334.

[Jar55] V. Jarník. *Integrální počet II*. NČAV Praha, 1955.

[Lej85] K. Lejchtvějs. *Vypuklije množestva*. Moskva Nauka, 1985.

[PBKN90] Ladislav Procházka, Ladislav Bican, Tomáš Kepka, and Petr Němec. *Algebra*. ACADEMIA Praha, 1990.

[Sau72] N. Sauer. On the Density of Families of Sets. *Journal of the Association for Computing Machinery*, 13:145–147, feb 1972.