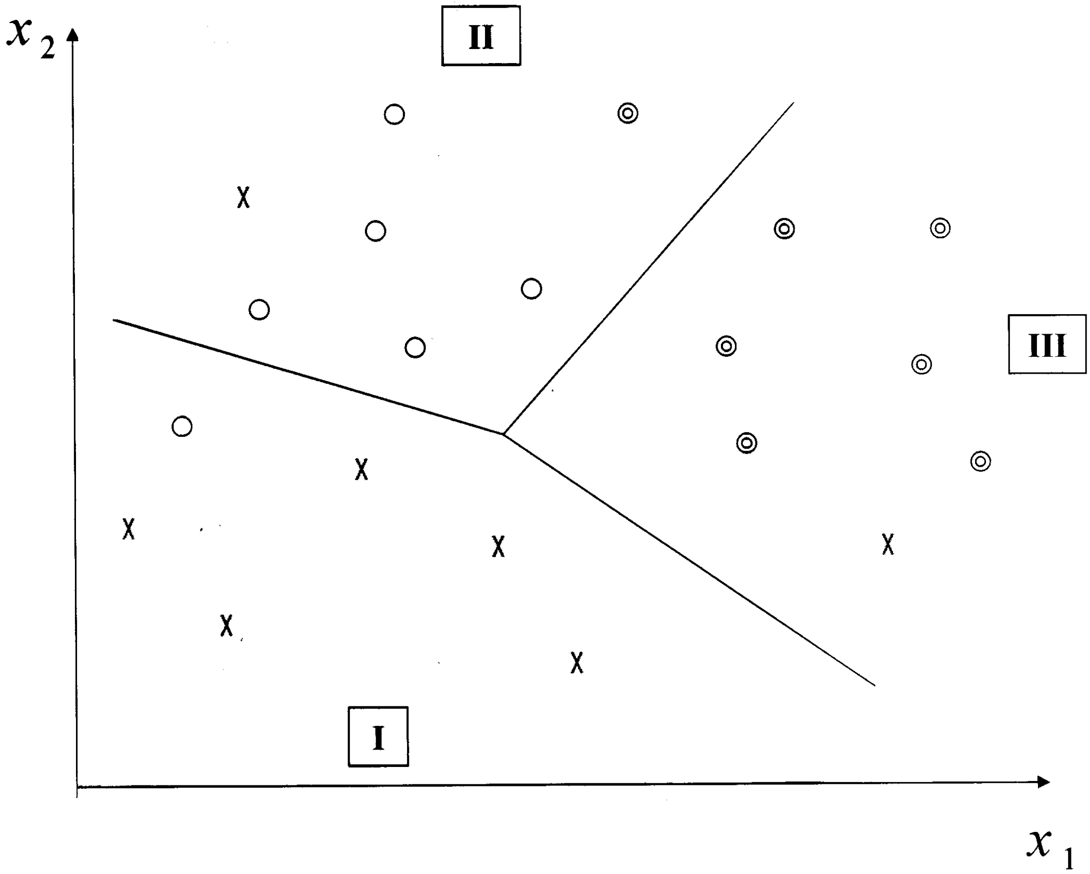
Chervonenkis
Alexey Jakovlevich

Institute of Control Science
Russian Academy of Sciences

Royal Holloway University of London

Entropy Properties of a Decision Rule Class in connection with machine learning abilities

Learning to Pattern Recognition

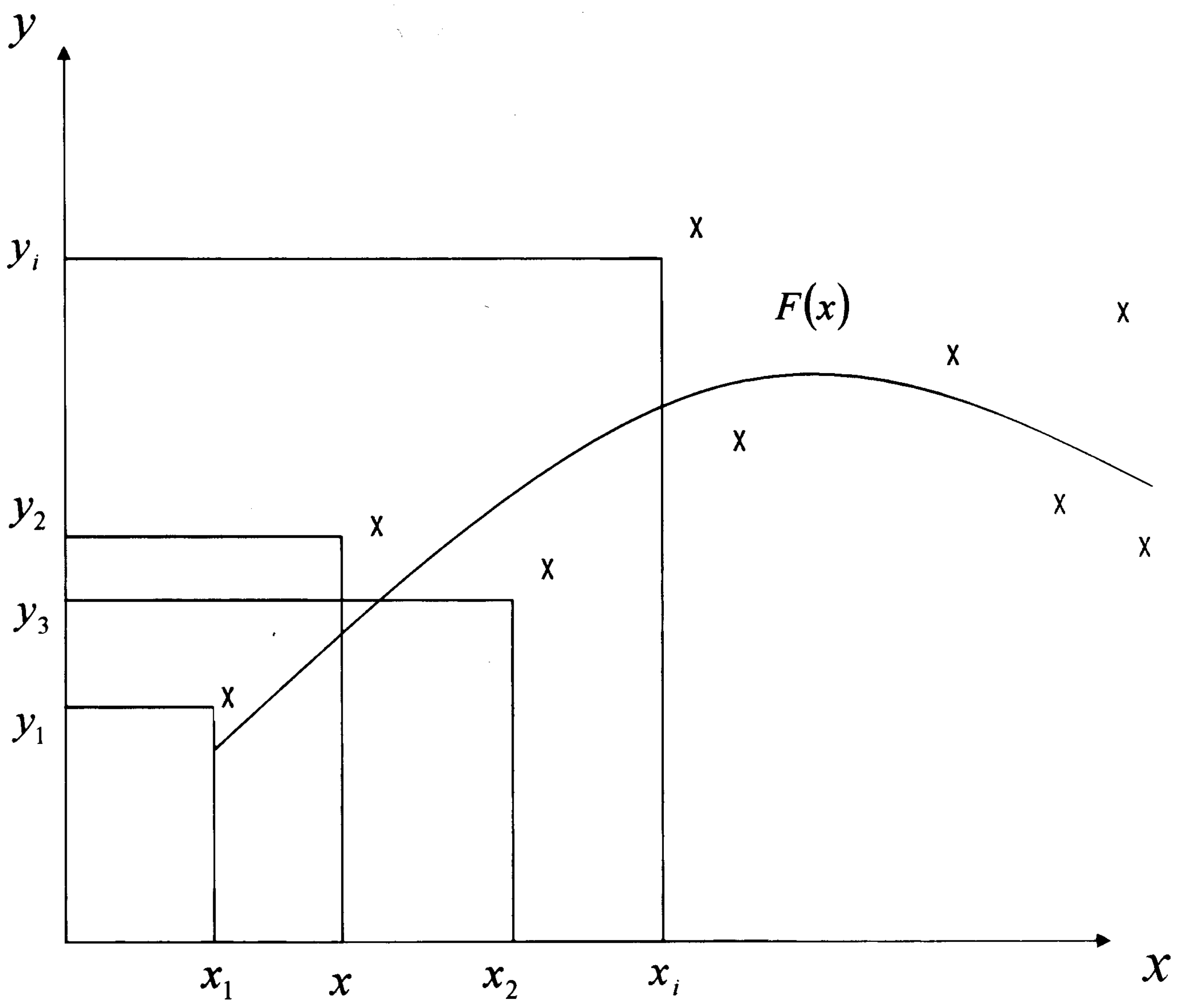


The points represent objects of different classes.

A decision rule is constructed.

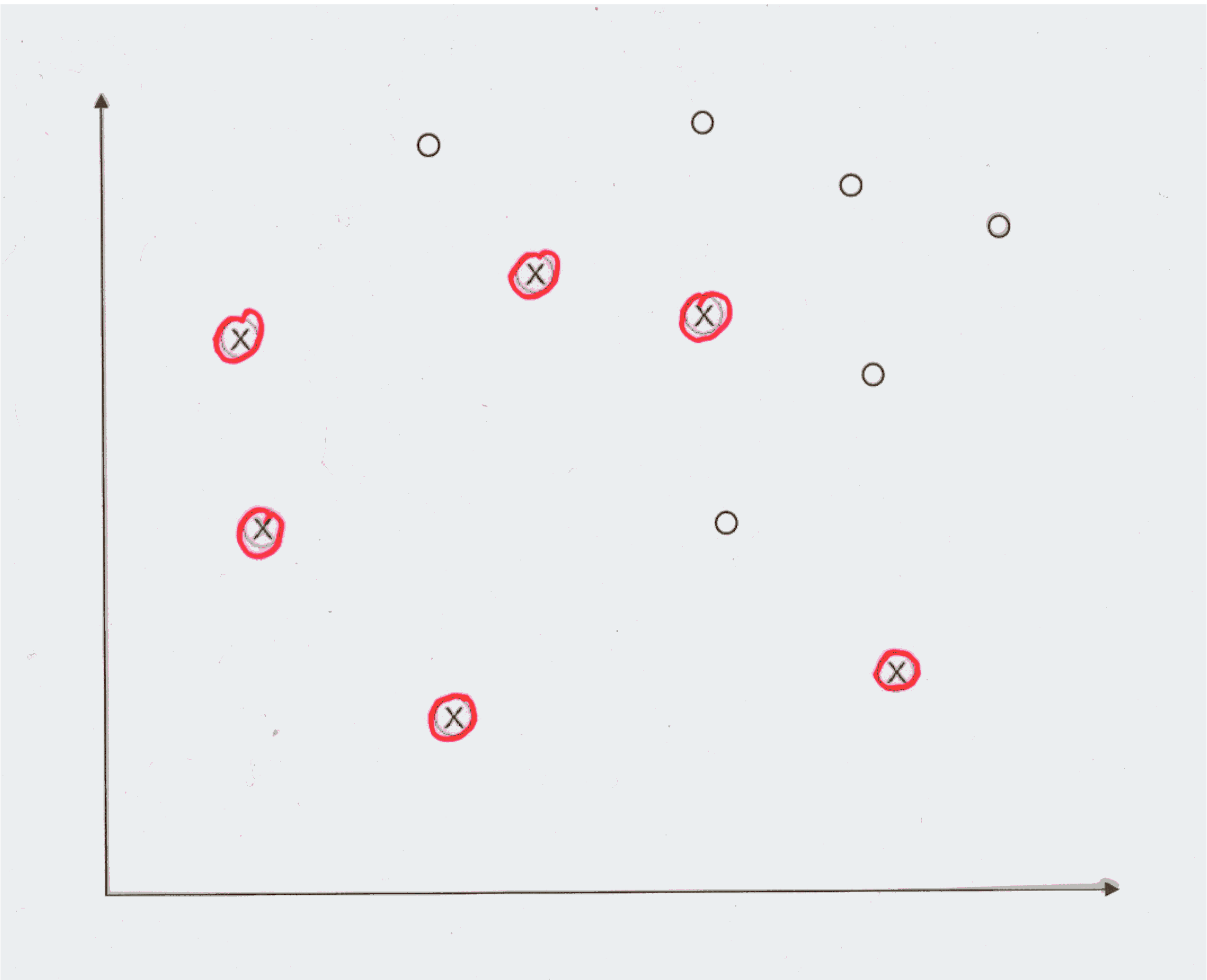
Sometimes we see **errors**.
The goal is to minimize the average number of **errors**.

Reconstruction of numerical dependencies.



The Goal is to minimize the Mean Square Error $(y - F(x))^2$.

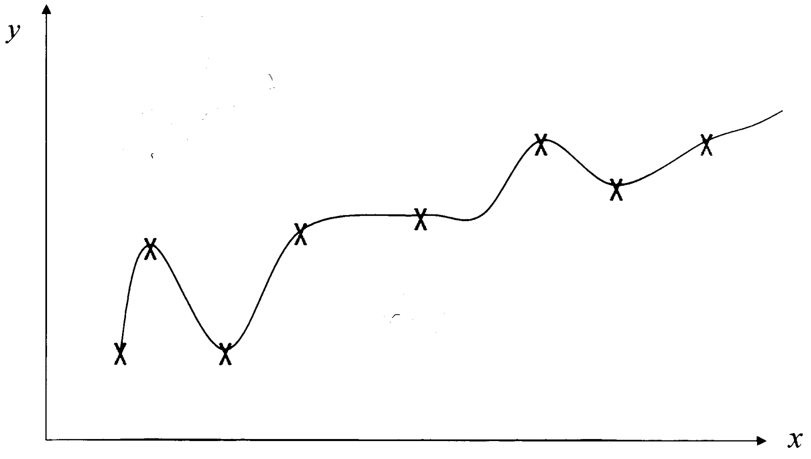
Over fitting in pattern recognition problem.

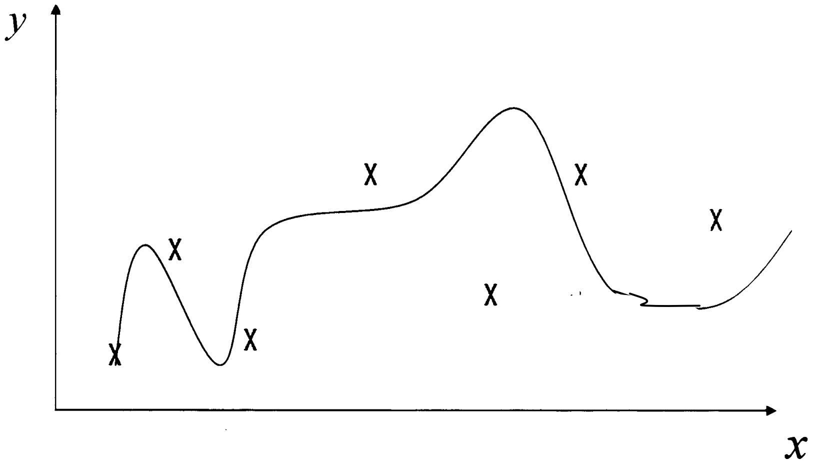


Having a training set one ca assume that all points within **red** circles belong to the class 1.
All the rest belong to the class 2.

We see **no errors** on the **training** set.
But there will be **a lot of errors** on **new** data.

Over fitting in polynomial regression reconstruction





Choosing sufficiently **large degree** of a polynomial one can get an approximation delivering **zero error** on the training set.

But on the new data the **errors** will be **large**.

Formal definition.

Penalty function for an error

$$Q(y, y^*),$$

where **y** is a true value, y* is a predicted value.

True average risk:

$$R_{true} = \int Q(y, F(x))\, dP_{xy}.$$

Empirical risk

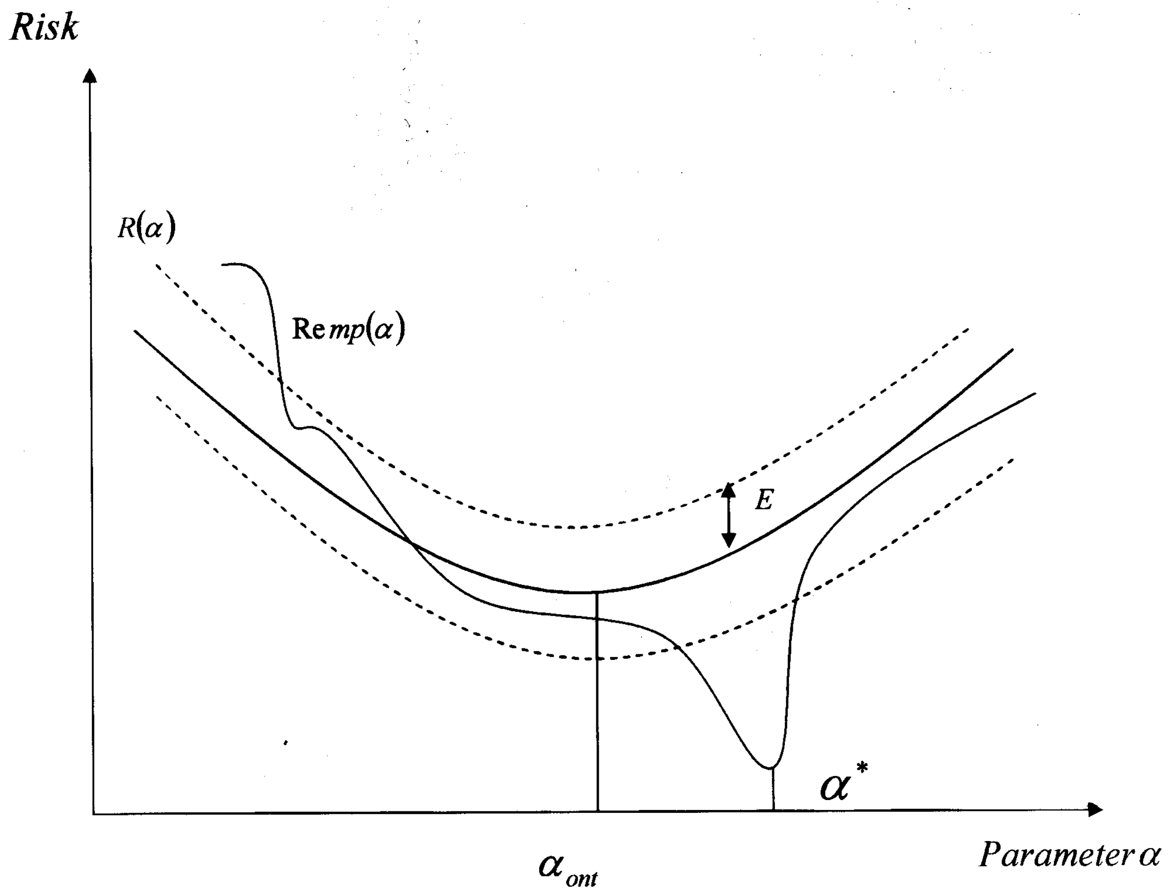
$$R_{emp} = 1/l \sum Q(y_i, F(x_i)).$$

The goal is to minimize true risk.

The mean is to minimize empirical risk.

Approval: according to the large number law

$$R_{emp} \xrightarrow{P} R_{true}$$



The **red** line shows the dependence of the empirical risk on parameters.

The black line shows the dependence of the true risk on the parameters.

The point, delivering minimum to empirical risk is far from the point delivering minimum to the true risk.

It would be not so if the dependencies R_{true} and R_{emp} are uniformly close.

Uniform convergence of frequencies to probabilities.

A є S - a system of random events.

P(A) – probability of an event A,

$x_1, x_2, \ldots .x_l$ - a random sample sequence

$\nu(A)$ – frequency of the eventчастота A on the sample sequence.

Bernully Theorem :

$$\nu(A) \rightarrow P(A)$$

Uniform convergence:

Sup $|\nu(A) - P(A)| \rightarrow 0$ A є S

Random functions F(x,α) $\alpha \in \Lambda$

M(α) = E F(x,α) -expectation

R(α) = $1/l \, \Sigma \, F(x_i,\alpha)$ – average value

The large number law:

$$R(\alpha) \rightarrow M(\alpha)$$

Uniform convergence:

Sup |R(α) - M(α)| $\rightarrow$ 0 $\alpha \in \Lambda$

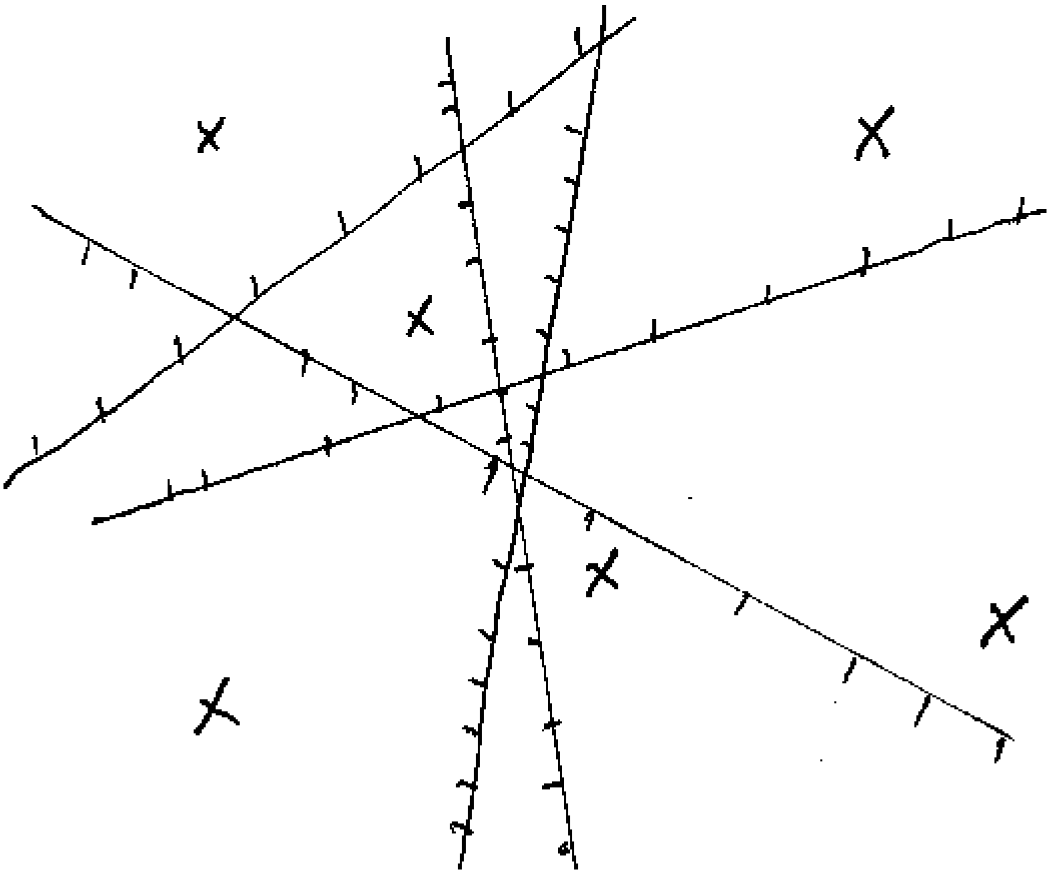
Conditions of the uniform convergence of frequencies to probabilities

Index $\Delta^{S}(x_1,\ldots,\ x_l)$ of the event class S is defined as the total number of all possible splitting of a sequence by the sets A є S.

Growth function is defined as

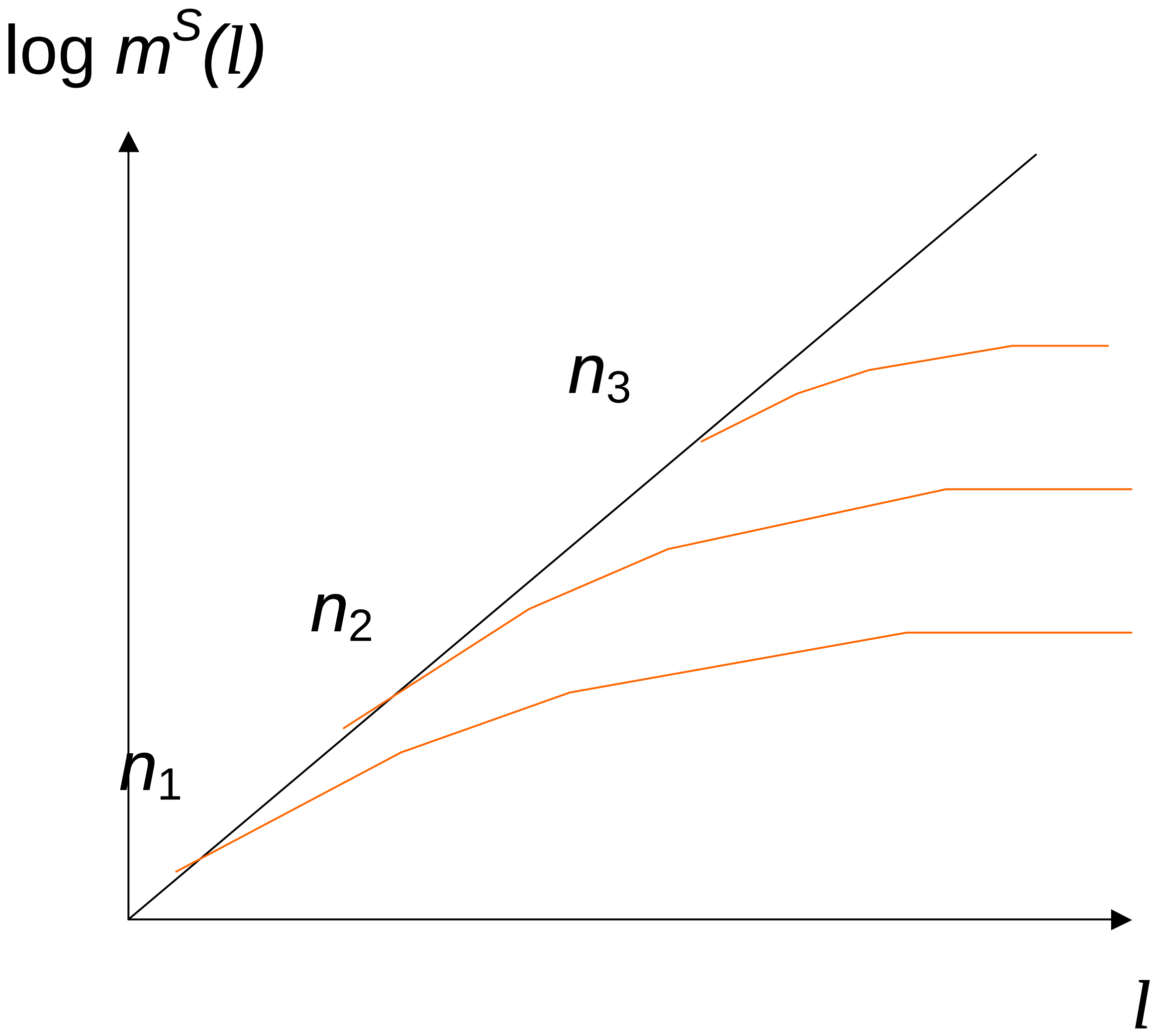
$$M^{\Lambda}(l) = \max \Delta^{S}(x_1,\ldots,\ x_l),$$

Where maximum is searched over all possible sample sequences of length l.



The picture shows all possible splitting of a sequence by the sets corresponding to half planes.

Growth function is either trivially equal to 2^l, is bounded from above by a polynomial.



Polynomial growth forms sufficient condition for the uniform convergence of frequencies to probabilities over the event class S.

This condition appears to be sufficient but not necessary. There exist examples, when the growth function is 2^l, but still the uniform convergence holds.

In particular, it is so if the space *X* is countable, and the system S consists of all subsets of the space *X*.

If instead of maximum of index

$$\Delta^S(x_1, , x_l)$$

we take expectation of its log

$$H^s(l) = \mathrm{E} \log \Delta^S(x_1, , x_l),$$

we can get necessary and sufficient conditions for the uniform convergence of frequencies to probabilities.

They are:

$$1/l \quad H^s(l) \to 0.$$

This condition may be interpreted as follows.

Entropy per symbol must go to zero while the sample sequence length goes to infinity.

Conditions for the Uniform Convergence of means to expectations

(For uniformly limited classes of functions).

Given random functions F(x,α) $\alpha \in \Lambda$

$M(\alpha) = E\, F(x,\alpha)$ -expectation

$R(\alpha) = 1/l \sum F(x_i,\alpha)$ – mean value

$\text{Sup}\, |R(\alpha) - M(\alpha)| \to 0 \qquad \alpha \in \Lambda$

It is possible to reduce the problem to the previous one.
It is enough to construct a set **S** of events defined as

$$A = \{ x: F(x_i,\alpha) > C\}$$

for all possible α and C values, and to apply the conditions (and estimates) of the uniform convergence to this set of events.

But then we get only sufficient conditions.

To deduce necessary and sufficient conditions we propose the following construction.

Given a sample sequence

$$x_1, x_2, \ldots .x_l$$

we construct in l - dimensional Euclidian space a setв T, consisting of the points with coordinates

$F(x_1,\alpha), F(x_2,\alpha) \ldots .F(x_l,\alpha)$

for all possible values of $\alpha \in \Lambda$.

Then we define ε– extension of this set as a unification of all cubes with edge length ε and centers in the points of the set T, and its volume

$$V^{\varepsilon}(x_1, x_2, \ldots x_l).$$

ε-entropy of a function class on the samples of length *l* we call expectation of this volume log over all sample sequences of length *l* :

$$H^{\varepsilon}(l) = \mathrm{E}\ \log V^{\varepsilon}(x_1, x_2, \ldots x_l).$$

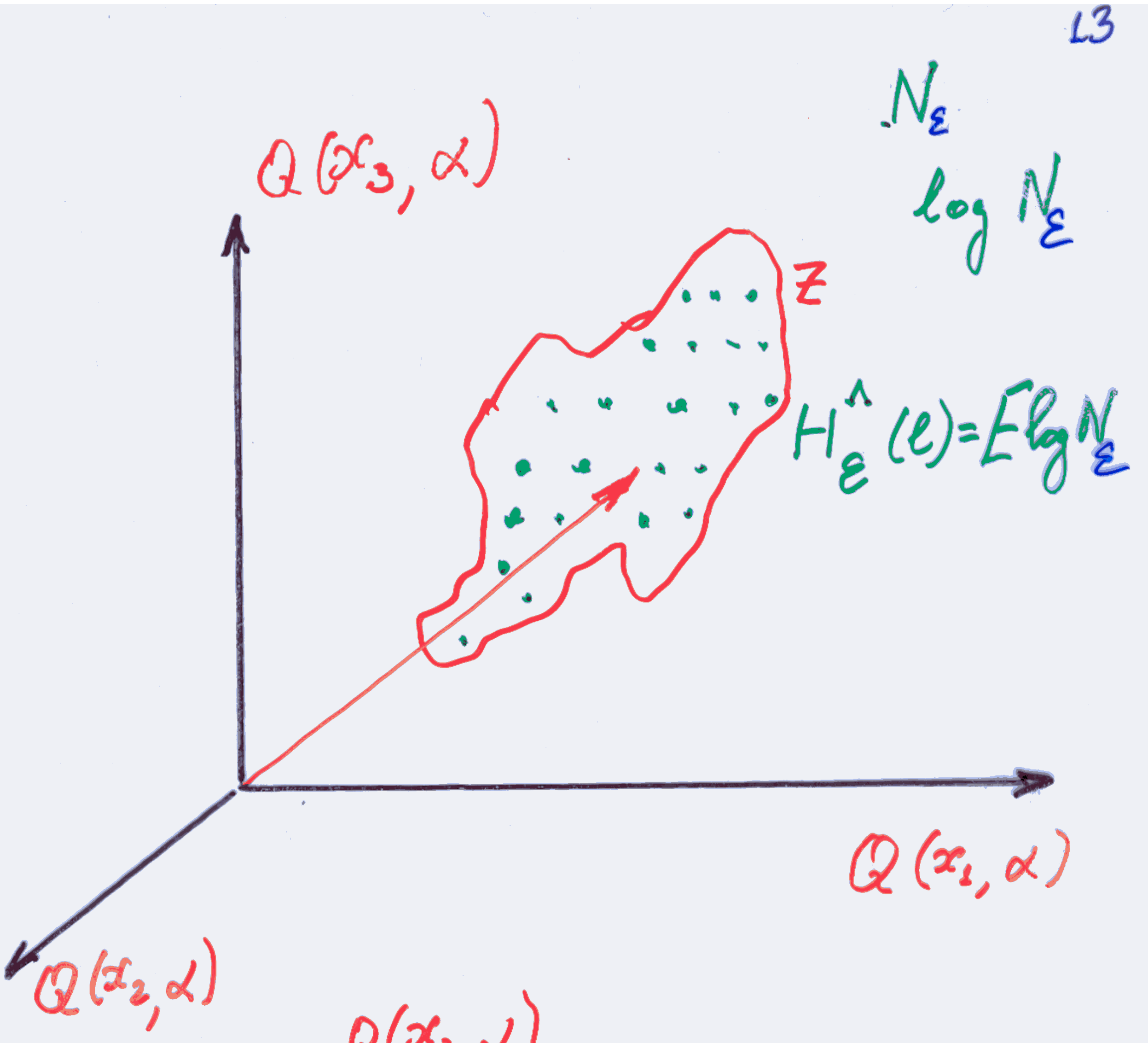
Then the necessary and sufficient condition for the uniform convergence of means to expectations is the following:

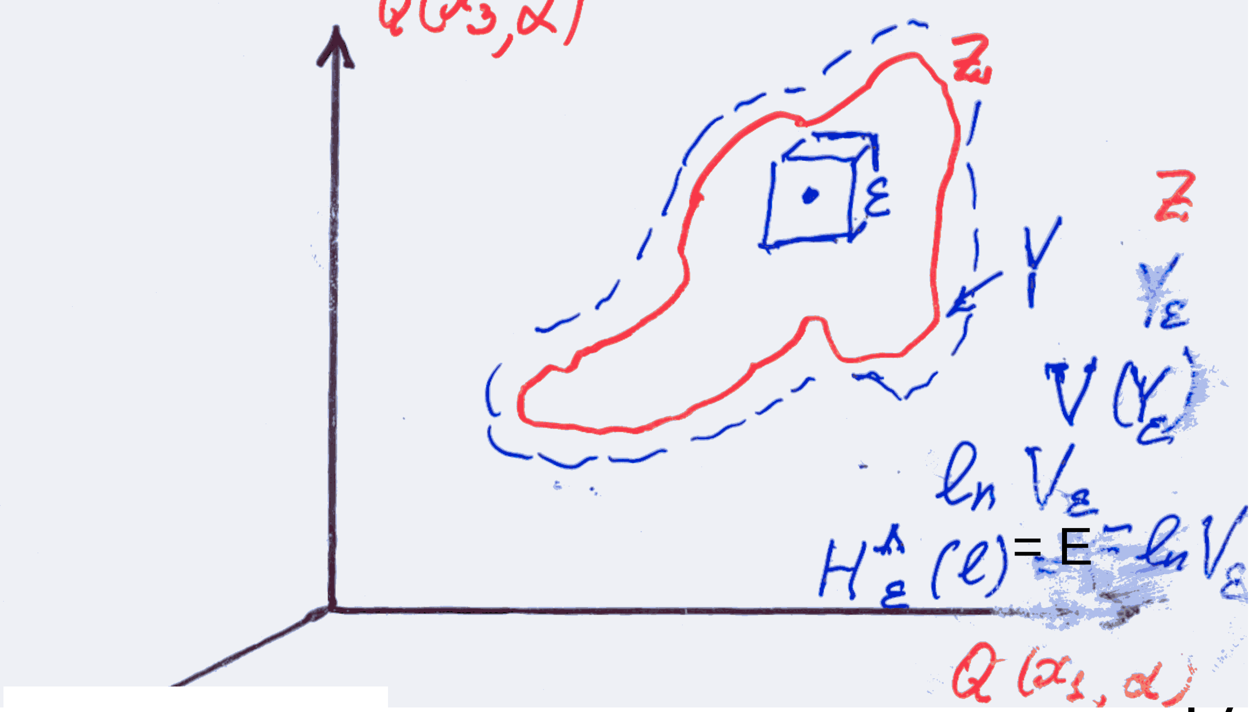
For all ε > 0

$$1/\,l\ \ H^{\varepsilon}(l) \to \log \varepsilon.$$

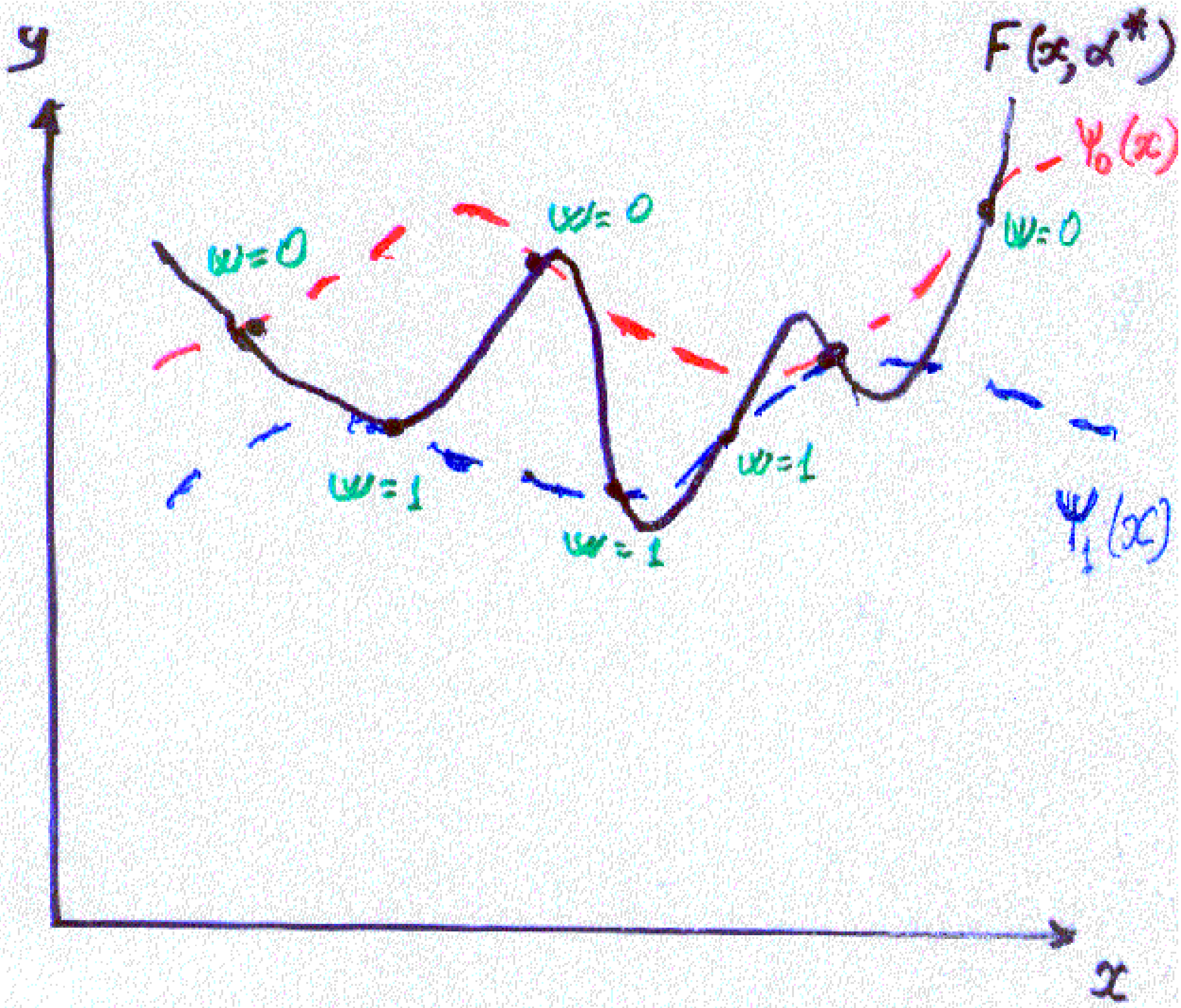
That means that ε– extension of our set T behaves asymptotically as a single cube with edge length equal to ε.

Really it appears that, if this condition holds for some ε > 0, then it holds for all ε > 0.


L3
$Q(x_3, \alpha)$
N_ε
$\log N_\varepsilon$
Z
$H^{\Lambda}_{\varepsilon}(\ell) = E \log N_\varepsilon$
$Q(x_1, \alpha)$
$Q(x_2, \alpha)$


$Q(x_3, \alpha)$
Z
ε
Z
Y
Y_ε
$V(Y_\varepsilon)$
$\ln V_\varepsilon$
$H^{\Lambda}_{\varepsilon}(\ell) = E \ln V_\varepsilon$
$Q(x_1, \alpha)$

$Q(x_2,\alpha)$



It appears that if

$$1/l \; H^{\varepsilon}(l) \;\rightarrow \log \varepsilon + \eta \quad (\eta \;> 0),$$

then there exist two functions,
the upper one $\varphi_0(x)$
and the lower one нижняя $\varphi_1(x)$

$$(\varphi_0(x) >= \varphi_1(x))$$

And the average distance between them is non-zero

$\int (\varphi_0(x) - \varphi_1(x))\, dPx \geq \varepsilon (e^{\eta} - 1)$.

The functions and the average distance may be found independently on the sample size.

Now if one has an almost arbitrary sample sequence,

$$x_1, x_2, \ldots . x_l$$

and assigns arbitrarily in what points x_i a function should be close to the upper function ($\omega_i = 0$), and in what points it should be close to the lower function ($\omega_i = 1$),

then there exists such a value α^*, that the function $F(x_i, \alpha^*)$ has arbitrarily close values to the upper or to the lower function depending on our assignment.

It is true for any sample sequence length.

Similar property is true for the case, when the uniform convergence of frequencies to probabilities does not hold.

In this case the system of events S may be characterized by a system of binary indicator functions

$$F(\mathrm{x}, a) \qquad (a \in L)$$

In this case there is no need to make ε–extension, and the result seems more clear and precise.

I remind that the uniform convergence of frequencies to probabilities does not hold if and only if

$$1/l \quad H^{s}(l) \rightarrow C > 0.$$

Then there exists a set $S \subseteq X$, such that almost for any sample sequence

$$x_1, x_2, \ldots .x_l \quad (x_i \in S)$$

and any binary sequence

$$w_1, w_2, \ldots .w_l \quad (w_i = 0,1)$$

there is such a value α*, that

$$F(x_i, \alpha^*) = w_i.$$