Stochastics: models, visualization, theorems

István Fazekas

University of Debrecen, Hungary

Lublin, 2013

1 Neural Networks

Artificial neural networks

1. Motivated by biological neural networks (human brain)
2. Computation mechanism
3. Consists of lot of small units
4. Connections among the units
5. Learning from data

However, the phenomenon is considered as a black box.

The perceptron

(1) Input

$\boldsymbol{x}(n) = (x_1(n), \ldots, x_m(n))^\top$ is a known m-dimensional vector. At any time $n = 1, 2, \ldots, N$ we have an input vector.

(2) Synaptic weights

The true weights $w_1, \ldots, w_m$ are unknown. We have to find them. At time n, $w_1(n), \ldots, w_m(n)$ are known. These approximate the true weights.

$\boldsymbol{w} = (w_1, \ldots, w_m)^\top$ is the true weight vector, $\boldsymbol{w}(n) = (w_1(n), \ldots, w_m(n))^\top$ is its approximation.

(3) Bias

The true bias b is unknown. Its nth approximation is $b(n)$.

(4) Summing junction

$$v(n) = b(n) + \sum_{i=1}^{m} w_i(n) x_i(n). \tag{1.1}$$

(5) Activation function, transfer function φ

(6) Output

For input $\boldsymbol{x}(n)$, the neuron produces the output $y(n) = \varphi(v(n))$.

Training

Steps:

1. design
2. training
3. application

Training set:

$\boldsymbol{x}(n)$ input with $d(n)$ output is known for $n = 1, \ldots, N$

These are the training points. The perceptron produces for input $\boldsymbol{x}(n)$ the output $y(n)$.

The error is

$$\mathcal{E}(n) = (d(n) - y(n))^2.$$

Find the minimum: training

Activation functions

Logistic function

$$\varphi(x) = \frac{1}{1+\exp(-ax)}\,, \quad x \in \mathbb{R}\,,$$

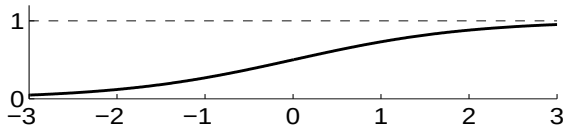
where $a > 0$.

$\varphi(0) = 1/2$, $\varphi(.)$ is increasing, $\varphi(\infty) = 1$, $\varphi(-\infty) = 0$. The derivative

$$\varphi'(x) = \frac{a\exp(-ax)}{(1+\exp(-ax))^2}\,.$$

So $\varphi'(0) = a/4$

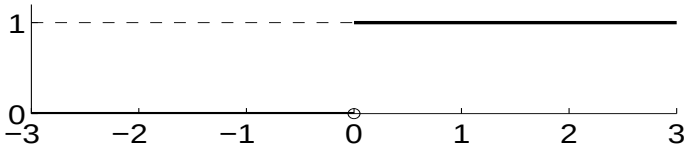
Logistic function



Logistic function with $a = 1$

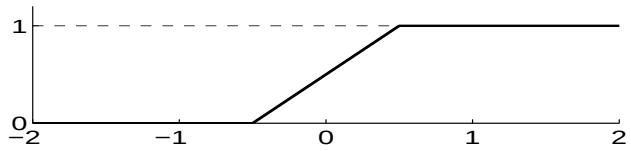
Threshold function, Heaviside function, hard limit

$$\varphi(x) = \begin{cases} 1, & \text{if} \quad x \geq 0\,, \\ 0, & \text{if} \quad x < 0\,. \end{cases}$$



Piecewise linear

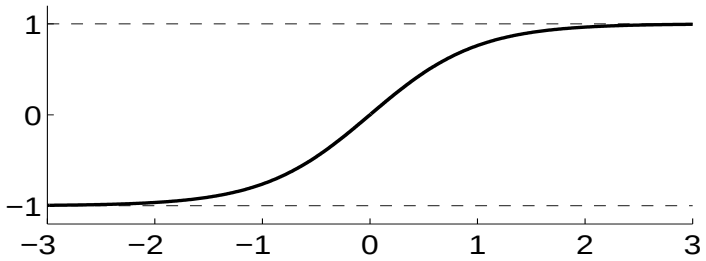
$$\varphi(x) = \begin{cases} 0, & \text{if} \quad x < -0.5\,, \\ x + 0.5, & \text{if} \quad -0.5 \leq x < 0.5\,, \\ 1, & \text{if} \quad x \geq 0.5\,. \end{cases}$$



Uniform on $[-0.5, 0.5]$.

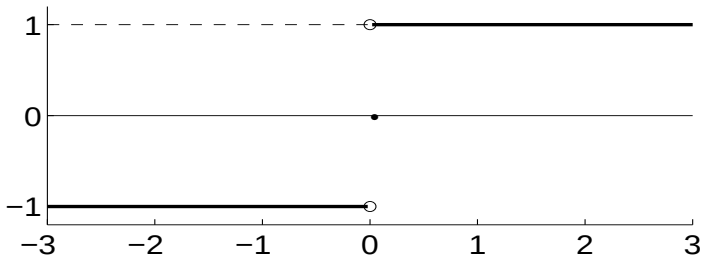
Tangent hyperbolic

$$\varphi(x) = \frac{2}{1 + \exp(-2x)} - 1 = \tanh(x).$$



Sign

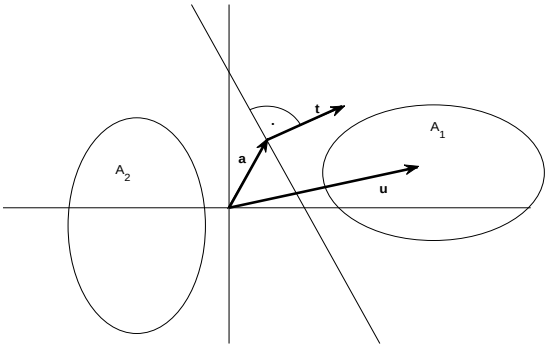
$$\varphi(x) = \begin{cases} -1, & \text{if} \quad x < 0\,, \\ 0, & \text{if} \quad x = 0\,, \\ 1, & \text{if} \quad x > 0\,. \end{cases}$$



Linear

$\varphi(x) = x$, $x \in \mathbb{R}$

Training perceptron



Linearly separable sets

- If the training point is correctly classified, then no correction is made. That is if either $\boldsymbol{w}(n)^\top \boldsymbol{x}(n) > 0$ and $\boldsymbol{x}(n) \in A_1$, or $\boldsymbol{w}(n)^\top \boldsymbol{x}(n) \leq 0$ and $\boldsymbol{x}(n) \in A_2$, then

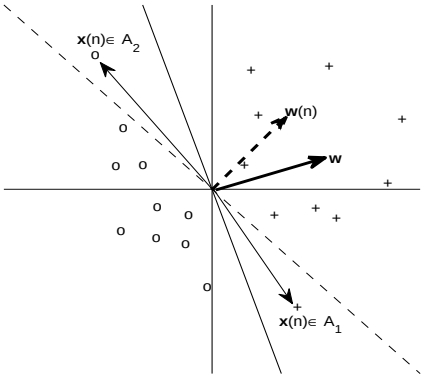
$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n).$$

- If $\boldsymbol{w}(n)^\top \boldsymbol{x}(n) \leq 0$, but $\boldsymbol{x}(n) \in A_1$, then we move $\boldsymbol{w}(n)$ to direction $\boldsymbol{x}(n)$:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + \eta(n)\boldsymbol{x}(n).$$

- If $\boldsymbol{w}(n)^\top \boldsymbol{x}(n) > 0$, but $\boldsymbol{x}(n) \in A_2$, then we move $\boldsymbol{w}(n)$ to the direction opposite to $\boldsymbol{x}(n)$:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta(n)\boldsymbol{x}(n).$$

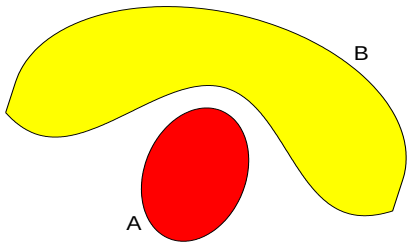


Two types of corrections

Perceptron convergence theorem (Rosenblatt, Novikoff)

Assume that A_1 and A_2 are linearly separable. Assume that $\|\boldsymbol{w}^\top \boldsymbol{x}(n)\| \geq \delta > 0$ and $\|\boldsymbol{x}(n)\|^2 \leq R < \infty$ for each training point. Let the training parameter $\eta > 0$ be fixed. Then the perceptron algorithm terminates at a hyperplane that correctly separates the two sets.

Multi Layer Perceptron, MLP



Linearly non-separable sets

A perceptron separates two linearly separable sets. An MLP separates two non-linearly separable sets.

Parts of an MLP:

1. An input layer
2. Hidden layers
3. An output layer

The input signal comes in at the input end of the network, propagates forward (neuron by neuron) through the network and finally produces the output of the network. The inputs of a neuron i are the outputs of the neurons being just to the left of neuron i. The

outputs of a neuron i are the inputs of the neurons being just to the right of neuron i. Each neuron has its own weights and activation function.

Training MLP

The main steps

- Initialization of the weights
- Give training points
- The input signal propagates through the network without changing the weights
- The output signal is compared with the true output
- The error is propagated backward through the network, the weights are updated

Notation

i, j, k: ith, jth, kth neuron (i, j, k are from the left to the right);
n: nth step of training;
$y_i(n)$: the output of neuron i (at the same time the input of neuron j if the layer of j is just right to the layer of i);
$y_0(n) \equiv 1$;
$w_{j0}(n) = b_j(n)$: the bias of j;
$w_{ji}(n)$: the weight on the edge from i to j;
$v_j(n)$: the value produced by the summing junction of j:

$$v_j(n) = \sum_{i \in \{\text{left neighbour layer of } j\}} w_{ji}(n) y_i(n);$$

$\varphi_j(.)$: transfer function of neuron j;
$y_j(n)$: the output of neuron j: $y_j(n) = \varphi_j(v_j(n))$;
$d_j(n)$: the true output (we compare $y_j(n)$ with it, $d_j(n)$ is known in the output layer)

Denote by C the set of the neurons in the output layer.

Then the error at the nth step is

$$\mathcal{E}(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n) = \frac{1}{2} \sum_{j \in C} (d_j(n) - y_j(n))^2.$$

To find the minimum use gradient method (delta rule):

$$w_{ji}(n+1) - w_{ji}(n) = \Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial w_{ji}(n)}, \tag{1.2}$$

where $\eta > 0$ is the training parameter.

Error back-propagation

Numerical problem: calculate the gradient of $\mathcal{E}$.

Error back-propagation: a recursive calculation of the gradient It is calculated backwards layer by layer.

The local gradient of neuron j can be obtained from the local gradients of the neurons being to the right of j. And the local gradients in the output layer can be calculated directly.

The correction of $w_{ji}(n)$ is

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_i(n),$$

where $\delta_j(n)$ is the local gradient:

$$\delta_j(n) = -\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = e_j(n)\varphi_j'(v_j(n)). \tag{1.3}$$

When j is a neuron in the output layer, then $e_j(n) = d_j(n) - y_j(n)$ is known.

However, if j is a neuron in a hidden layer, then it is not known. However, it can be calculated recursively.

$$\delta_j(n) = \varphi_j'(v_j(n)) \sum_{k \in \{\text{right neighbour layer of } j\}} \delta_k(n) w_{kj}(n). \tag{1.4}$$

It is the most important formula of back-prop.

Versions of error back-propagation

The gradient is calculated by error back-propagation algorithm and applied in other methods.

Conjugate gradient methods: Fletcher-Reeves formula (conjugate gradient back-propagation with Fletcher-Reeves updates), Polak-Ribiére formula (conjugate gradient back-propagation with Polak-Ribiére updates), Powell-Beale formula (conjugate gradient back-propagation with Powell-Beale restarts).

Quasi Newton methods: Broyden-Fletcher-Goldfarb-Shanno formula (BFGS quasi-Newton back-propagation).

Levenberg-Marquardt method (Levenberg-Marquardt back-propagation).

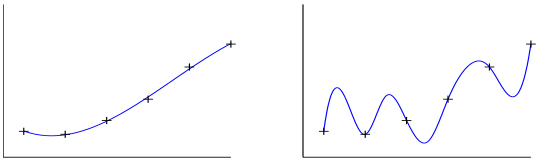
Generalized delta rule or momentum method (gradient descent with momentum back-propagation).

Sequential (one by one), batch (by epochs).

Generalization

A network generalizes well if it produces correct outputs for inputs not being in the training set.

If the shape of the separating curve (or the approximating function) is very 'irregular', then it is probably overfitted.



Left: generalizes well; right: overfitted, overtrained

Radial-Basis Function Networks (RBF)

Let

$$\boldsymbol{x}_i \in \mathbb{R}^{m_0}, \quad i = 1, 2, \ldots, N,$$

be the input vectors and

$$d_i \in \mathbb{R}^1, \quad i = 1, 2, \ldots, N,$$

the corresponding output values.

Find the function $F : \mathbb{R}^{m_0} \to \mathbb{R}$ that approximates well the value d_i at point $\boldsymbol{x}_i$ for each i.

According to Tikhonov's regularization the goodness of fit is measured by two terms.

(i) *Standard error term*:

$$\mathcal{E}_s(F) = \frac{1}{2}\sum_{i=1}^{N}(d_i - y_i)^2 = \frac{1}{2}\sum_{i=1}^{N}[d_i - F(\boldsymbol{x}_i)]^2.$$

(ii) *Regularizing term* $\quad \mathcal{E}_c(F) = \frac{1}{2}\|DF\|^2_{\mathcal{H}}$.

Here D is a linear differential operator. Actually D is defined on a space of functions $F : \mathbb{R}^{m_0} \to \mathbb{R}$ being square integrable. This Hilbert space is denoted by $\mathcal{H}$ and our operator is $D : \mathcal{H} \to \mathcal{H}$. $\|DF\|^2_{\mathcal{H}}$ is the square of the norm of D at space $\mathcal{H}$. The function F is an unknown element of $(\mathcal{H}, \langle \cdot\,, \cdot \rangle_{\mathcal{H}})$. D includes the prior information about the problem. $\mathcal{E}_c(F)$ is the penalty function.

We have to minimize:

$$\mathcal{E}(F) = \mathcal{E}_s(F) + \lambda\mathcal{E}_c(F) = \frac{1}{2}\sum_{i=1}^{N}[d_i - F(\boldsymbol{x}_i)]^2 + \frac{1}{2}\lambda\|DF\|^2_{\mathcal{H}}, \tag{1.5}$$

where $\lambda > 0$ is the regularization parameter.

If λ is small, then the training points determine the solution F_λ. If λ is large then the prior informations specify the solution. $\mathcal{E}(F)$ is called Tikhonov's functional.

Denote the minimum point of $\mathcal{E}(F)$ by F_λ. F_λ is a minimum point of $\mathcal{E}(F)$, if

$$\mathcal{E}(F_\lambda) \le \mathcal{E}(F_\lambda + \beta h)$$

for any function h and any scalar β. So for any $h \in \mathcal{H}$ fixed non-zero function

$$\left[\frac{d}{d\beta}\mathcal{E}(F_\lambda + \beta h)\right]_{\beta=0} = 0\,.$$

Let G be the Green function of $\widetilde{D}D$. Then

$$F_\lambda(\boldsymbol{x}) = \frac{1}{\lambda}\sum_{i=1}^{N}[d_i - F_\lambda(\boldsymbol{x}_i)]G(\boldsymbol{x}, \boldsymbol{x}_i)\,. \tag{1.6}$$

That is if the training points $(\boldsymbol{x}_i, d_i)$, $i = 1, \ldots, N$, are given, the Green function G is known, and the regularization parameter λ is given, then the minimum point of the Tikhonov functional $\mathcal{E}$ is F_λ in (1.6).

Let

$$w_i = \frac{1}{\lambda}[d_i - F_\lambda(\boldsymbol{x}_i)], \qquad i = 1, 2, \ldots, N, \tag{1.7}$$

$$\boldsymbol{w} = (w_1, w_2, \ldots, w_N)^\top.$$

$$\boldsymbol{d} = (d_1, d_2, \ldots, d_N)^\top,$$

$$G = \begin{pmatrix} G(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & G(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ G(\boldsymbol{x}_N, \boldsymbol{x}_1) & \ldots & G(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}.$$

Then

$$(G + \lambda I)\boldsymbol{w} = \boldsymbol{d}, \tag{1.8}$$

where I is the $N \times N$ unit matrix.

Solving equation (1.8), we obtain the solution of the regularization problem:

$$\boxed{F_\lambda(\boldsymbol{x}) = \sum_{i=1}^{N} w_i G(\boldsymbol{x}, \boldsymbol{x}_i)}\,. \tag{1.9}$$

It is the regularization network.

However, the Green function is not known. We have to choose it. Generally, it is of the form

$$G(\boldsymbol{x}, \boldsymbol{x}_i) = G(\|\boldsymbol{x} - \boldsymbol{x}_i\|)\,.$$

Therefore

$$\boxed{F_\lambda(\boldsymbol{x}) = \sum_{i=1}^{N} w_i G(\|\boldsymbol{x} - \boldsymbol{x}_i\|)}\,. \tag{1.10}$$

Example.

$$G(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{1}{2\sigma_i^2}\|\boldsymbol{x} - \boldsymbol{x}_i\|^2\right),$$

where $\boldsymbol{x}, \boldsymbol{x}_i \in \mathbb{R}^{m_0}$, $\sigma_i > 0$. It is the density function of $\mathcal{N}_{m_0}(\boldsymbol{x}_i, \sigma_i^2 I)$ (up to a scalar).

The number of neurons in the regularization network is large (N).

We want to find a new solution of the form

$$F^*(\boldsymbol{x}) = \sum_{i=1}^{m_1} w_i G(\|\boldsymbol{x} - \boldsymbol{t}_i\|). \tag{1.11}$$

Find the new weights $\{w_i : i = 1, 2, \ldots, m_1\}$ which minimize the Tikhonov functional $\mathcal{E}(F^*)$:

$$\mathcal{E}(F^*) = \frac{1}{2} \sum_{j=1}^{N} \left(d_j - \sum_{i=1}^{m_1} w_i G(\|\boldsymbol{x}_j - \boldsymbol{t}_i\|) \right)^2 + \frac{\lambda}{2} \|DF^*\|^2_{\mathcal{H}}. \tag{1.12}$$

Let

$$\boldsymbol{d} = (d_1, d_2, \ldots, d_N)^\top, \qquad \boldsymbol{w} = (w_1, w_2, \ldots, w_{m_1})^\top$$

$$G = \begin{pmatrix} G(\boldsymbol{x}_1, \boldsymbol{t}_1) & \ldots & G(\boldsymbol{x}_1, \boldsymbol{t}_{m_1}) \\ \vdots & \ddots & \vdots \\ G(\boldsymbol{x}_N, \boldsymbol{t}_1) & \ldots & G(\boldsymbol{x}_N, \boldsymbol{t}_{m_1}) \end{pmatrix}.$$

$$G_0 = \begin{pmatrix} G(\boldsymbol{t}_1, \boldsymbol{t}_1) & \ldots & G(\boldsymbol{t}_1, \boldsymbol{t}_{m_1}) \\ \vdots & \ddots & \vdots \\ G(\boldsymbol{t}_{m_1}, \boldsymbol{t}_1) & \ldots & G(\boldsymbol{t}_{m_1}, \boldsymbol{t}_{m_1}) \end{pmatrix}.$$

Then $\boldsymbol{w}$ is the solution of

$$(G^\top G + \lambda G_0)\boldsymbol{w} = G^\top \boldsymbol{d}. \tag{1.13}$$

We solve it and then the generalized RBF is

$$\boxed{F^*(\boldsymbol{x}) = \sum_{i=1}^{m_1} w_i G(\|\boldsymbol{x} - \boldsymbol{t}_i\|)}.$$

Finding the centers $\boldsymbol{t}_i$.

- Fixed centers selected at random
- k-means clustering
- Supervised selection of centers

Non-parametric statistical estimators

Let $K : \mathbb{R}^{m_0} \to \mathbb{R}$ be a bounded, continuous, symmetric function with maximum at the origin. Assume $\int_{\mathbb{R}^{m_0}} K(\boldsymbol{x}) d\boldsymbol{x} = 1$. Then K is called kernel function.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be a sample form a population with density function f. The kernel type estimator of f is (Parzen–Rosenblatt):

$$\boxed{\widehat{f}_N(\boldsymbol{x}) = \frac{1}{Nh^{m_0}} \sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h} \right)}, \quad \boldsymbol{x} \in \mathbb{R}^{m_0}. \tag{1.14}$$

Here h is the bandwidth.

We want to estimate the (one-dimensional) Y with a function of X (m_0-dimensional). The optimal choice is the conditional expectation:

$$g(\boldsymbol{x}) = \mathbb{E}\{Y \,|\, X = \boldsymbol{x}\}.$$

g is the regression function. Then

$$g(\boldsymbol{x}) = \int y f(y \,|\, \boldsymbol{x}) dy = \int y \frac{f(y, \boldsymbol{x})}{f(\boldsymbol{x})} dy\,,$$

where $f(y, \boldsymbol{x})$ is the joint density function of Y and X, $f(\boldsymbol{x})$ is the density function of X, $f(y \,|\, \boldsymbol{x})$ is the conditional density of Y given X.

Let $(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)$ be a sample for (Y, X). We estimate f by (1.14). We estimate the joint density by

$$\widehat{f}(y, \boldsymbol{x}) = \frac{1}{N h^{m_0+1}} \sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right) K_0\left(\frac{y - y_i}{h}\right)\,, \tag{1.15}$$

$y \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^{m_0}$. (Here K_0 is a kernel.)

Then

$$\widehat{g}(\boldsymbol{x}) = \int y \frac{\widehat{f}(y, \boldsymbol{x})}{\widehat{f}(\boldsymbol{x})} dy =$$

$$= \frac{\frac{1}{N h^{m_0+1}} \sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right) \int y K_0\left(\frac{y - y_i}{h}\right) dy}{\frac{1}{N h^{m_0}} \sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)} = \frac{\sum_{i=1}^{N} y_i K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)}\,.$$

We obtained

$$\boxed{\widehat{g}(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} y_i K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)}}.$$

It is the kernel type estimator of the regression function g (Nadaraya-Watson).

Constrained optimization

Kuhn–Tucker theorem, Karush–Kuhn–Tucker theorem (Book by Boyd and Vandenberghe).

Let $f_k(\boldsymbol{x})$, $k = 0, \ldots, m$, be real valued functions of d variables. Assume that the intersection of their domain is non-empty: D. The primal optimization problem is:

$$\begin{aligned} \text{minimize} \quad & f_0(\boldsymbol{x}) \\ \text{subject to} \quad & f_k(\boldsymbol{x}) \le 0, \quad k = 1, \ldots, m, \end{aligned}$$

Denote p^* the optimal value, that is the infimum of $f_0(\boldsymbol{x})$ subject to the constraints.

Let

$$L = L(\boldsymbol{x}, \boldsymbol{\lambda}) = f_0(\boldsymbol{x}) + \sum_{k=1}^{m} \lambda_k f_k(\boldsymbol{x}),$$

be the Lagrange function, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ are the Lagrange multipliers.

Let

$$g(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x} \in D} L(\boldsymbol{x}, \boldsymbol{\lambda})$$

be the Lagrange dual function. g is always concave. Moreover, for $\boldsymbol{\lambda} \geq \mathbf{0}$ we have $g(\boldsymbol{\lambda}) \leq p^*$. Using this inequality, we want to find the best lower bound. This leads to the following.

The Lagrange dual problem is:

$$\begin{aligned} &\text{maximize} \quad g(\boldsymbol{\lambda}) \\ &\text{subject to} \quad \lambda_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Denote d^* the optimal value, that is the supremum of $g(\boldsymbol{\lambda})$ subject to the constraints. Then $d^* \leq p^*$. $p^* - d^* \geq 0$ is the optimal duality gap.

Assume that $f_k(\boldsymbol{x})$, $k = 0, \ldots, m$, are differentiable. Then D is open.

Theorem. *Let $\boldsymbol{x}^*$ be the optimal solution of the primal problem, let $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ be the optimal solution of the dual optimization problem with zero duality gap.*

Then the Kuhn–Tucker conditions are satisfied:

$$\begin{aligned} &f_k(\boldsymbol{x}^*) \leq 0, && k = 1, \ldots, m, \\ &\lambda_k^* \geq 0, && k = 1, \ldots, m, \\ &\lambda_k^* f_k(\boldsymbol{x}^*) = 0, && k = 1, \ldots, m, \\ &\frac{df_0(\boldsymbol{x}^*)}{d\boldsymbol{x}} + \sum_{k=1}^{m} \lambda_k^* \frac{df_k(\boldsymbol{x}^*)}{d\boldsymbol{x}} = \mathbf{0}. \end{aligned}$$

For convex functions the above conditions are sufficient.

Theorem. *Let $f_k(\boldsymbol{x})$, $k = 0, \ldots, m$, be convex. Assume that the points $\tilde{\boldsymbol{x}}$ and $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ satisfy the Kuhn–Tucker conditions, that is*

$$f_k(\tilde{\boldsymbol{x}}) \leq 0, \qquad k = 1, \ldots, m, \tag{1.16}$$

$$\tilde{\lambda}_k \geq 0, \qquad k = 1, \ldots, m, \tag{1.17}$$

$$\tilde{\lambda}_k f_k(\tilde{\boldsymbol{x}}) = 0, \qquad k = 1, \ldots, m, \tag{1.18}$$

$$\frac{df_0(\tilde{\boldsymbol{x}})}{d\boldsymbol{x}} + \sum_{k=1}^{m} \tilde{\lambda}_k \frac{df_k(\tilde{\boldsymbol{x}})}{d\boldsymbol{x}} = \mathbf{0}. \tag{1.19}$$

Then $\tilde{\boldsymbol{x}}$ is the optimal solution of the primal optimization problem, $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m)$ is the optimal solution of the dual optimization problem, and the optimal duality gap is zero.

The Slater condition: there exists a point $\underline{\boldsymbol{x}}$ in the interior of D such that $f_i(\underline{\boldsymbol{x}}) < 0$, $i = 1, \ldots, m$ (that is we have strict inequality).

If $f_k(\boldsymbol{x})$, $k = 0, \ldots, m$, are convex, then the Slater condition implies, that the duality gap is zero.

Assume that $f_k(\boldsymbol{x})$, $k = 0, \ldots, m$, are convex and differentiable. Assume that the Slater condition is satisfied. Then there is a solution of (1.16)-(1.19).

The Kuhn–Tucker conditions (1.16)-(1.19) are necessary and sufficient for optimality: that is $\tilde{\boldsymbol{x}}$ is the optimal solution of the primal optimization problem, $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m)$

is the optimal solution of the dual optimization problem, and the optimal duality gap is zero.

The optimality (with zero duality gap) of $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}}$ is equivalent to

$$\min_{\boldsymbol{x}} L(\boldsymbol{x}, \tilde{\boldsymbol{\lambda}}) = L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}}) = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}),$$

that is $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}})$ is a saddle-point.

Support Vector Machines (SVM)

Statistical learning theory, Vapnik

Applied in image processing, a bioinformatics, data mining,...

Linear separation, the optimal hyperplane

Assume that

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N), \quad \boldsymbol{x}_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\} \tag{1.20}$$

is the training set. If $\boldsymbol{x}_i \in \mathbb{R}^d$ belong to class A_1, then $y_i = 1$, if it belongs to class A_2, then $y_i = -1$. A_1 and A_2 are separable by the hyperplane

$$\langle \boldsymbol{x}, \boldsymbol{\varphi} \rangle = c \tag{1.21}$$

if

$$\langle \boldsymbol{x}_i, \boldsymbol{\varphi} \rangle > c, \quad \text{if} \quad y_i = 1, \tag{1.22}$$

$$\langle \boldsymbol{x}_i, \boldsymbol{\varphi} \rangle < c, \quad \text{if} \quad y_i = -1, \tag{1.23}$$

where $\boldsymbol{\varphi} \in \mathbb{R}^d$ is a unit vector, $c \in \mathbb{R}$ and $\langle \boldsymbol{a}, \mathbb{B} \rangle$ denotes the scalar product of $\boldsymbol{a}$ and $\mathbb{B}$.

Optimal hyperplane: the widest zone between the two classes.

Steps: geometric conditions, Kuhn–Tucker theorem, numerical procedure. Assume that the two classes are linearly separable.

For any unit vector $\boldsymbol{\varphi}$ let

$$c_1(\boldsymbol{\varphi}) = \min_{y_i = 1} \langle \boldsymbol{x}_i, \boldsymbol{\varphi} \rangle, \tag{1.24}$$

$$c_2(\boldsymbol{\varphi}) = \max_{y_i = -1} \langle \boldsymbol{x}_i, \boldsymbol{\varphi} \rangle. \tag{1.25}$$

Let the unit vector $\boldsymbol{\varphi}_0$ be the maximum point of

$$\varrho(\boldsymbol{\varphi}) = \frac{c_1(\boldsymbol{\varphi}) - c_2(\boldsymbol{\varphi})}{2} \tag{1.26}$$

given $\|\boldsymbol{\varphi}\| = 1$.

Theorem. *$\boldsymbol{\varphi}_0$ and*

$$c_0 = \frac{c_1(\boldsymbol{\varphi}_0) + c_2(\boldsymbol{\varphi}_0)}{2} \tag{1.27}$$

give the optimal hyperplane $\langle \boldsymbol{x}, \boldsymbol{\varphi}_0 \rangle = c_0$.

Theorem. *The optimal hyperplane is unique.*

An alternative form of the task is the following. Find the vector $\boldsymbol{\psi}_0$ and scalar b_0 which satisfy inequalities

$$\langle \boldsymbol{x}_i, \boldsymbol{\psi} \rangle + b \geq 1, \quad \text{if} \quad y_i = 1, \tag{1.28}$$

$$\langle \boldsymbol{x}_i, \boldsymbol{\psi} \rangle + b \leq -1, \quad \text{if} \quad y_i = -1 \tag{1.29}$$

and give the minimum of

$$\|\boldsymbol{\psi}\|^2 = \langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle. \tag{1.30}$$

Theorem. *The normalized form of the vector $\boldsymbol{\psi}_0$ that minimizes the quadratic function* (1.30) *under the linear constraints* (1.28)–(1.29) *gives the normal vector $\boldsymbol{\varphi}_0$ of the optimal hyperplane:*

$$\boldsymbol{\varphi}_0 = \frac{\boldsymbol{\psi}_0}{\|\boldsymbol{\psi}_0\|}. \tag{1.31}$$

Furthermore, the margin between the optimal hyperplane and separated vectors is:

$$\varrho(\boldsymbol{\varphi}_0) = \sup_{\|\boldsymbol{\varphi}\|=1} \frac{1}{2} \left(\min_{y_i=1} \langle \boldsymbol{x}_i, \boldsymbol{\varphi} \rangle - \max_{y_i=-1} \langle \boldsymbol{x}_i, \boldsymbol{\varphi} \rangle \right) = \frac{1}{\|\boldsymbol{\psi}_0\|}. \tag{1.32}$$

Under constraint

$$y_i(\langle \boldsymbol{x}_i, \boldsymbol{\psi} \rangle + b) \geq 1, \quad i = 1, \ldots, N, \tag{1.33}$$

find the minimum of

$$\frac{1}{2}\|\boldsymbol{\psi}\|^2.$$

It is a convex optimization problem.

Apply the Kuhn–Tucker theorem. The Lagrange function is

$$L(\boldsymbol{\psi}, b, \boldsymbol{\alpha}) = \frac{1}{2}\langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle - \sum_{i=1}^{N} \alpha_i \left(y_i [\langle \boldsymbol{x}_i, \boldsymbol{\psi} \rangle + b] - 1 \right) = \tag{1.34}$$

$$= \frac{1}{2}\langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle - \left\langle \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i, \boldsymbol{\psi} \right\rangle - b \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i \tag{1.35}$$

where $\alpha_i \geq 0$, $i = 1, \ldots, N$, are the Lagrange multipliers. We should find the saddle point of L.

Find the minimum according to $\boldsymbol{\psi}$ and b maximum according to $\boldsymbol{\alpha}$. We can find the minimum with respect to $\boldsymbol{\psi}$ and b analytically. We obtain the following.

Find the maximum of

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{1.36}$$

under constraints $\alpha_i \geq 0$ and

$$\sum_{i=1}^{N} \alpha_i y_i = 0. \tag{1.37}$$

So we find the vector defining the optimal hyperplane:

$$\boldsymbol{\psi}_0 = \sum_{i=1}^{N} y_i \alpha_i^0 \boldsymbol{x}_i.$$

b_0 can be obtained from the following equations.

For non-zero values of α_i^0 the following should be satisfied:

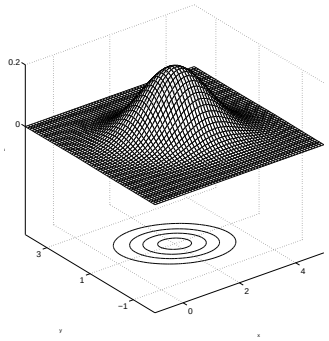
$$y_i(\langle \boldsymbol{x}_i, \boldsymbol{\psi}_0 \rangle + b_0) = 1. \tag{1.38}$$

These vectors $\boldsymbol{x}_i$ are called support vectors.

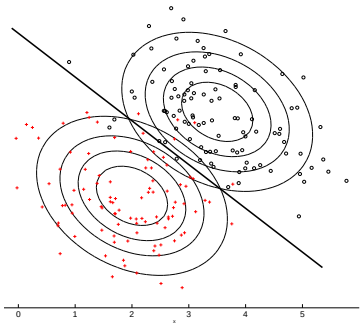
Finally, the optimal hyperplane is

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \alpha_i^0 \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b_0 = 0.$$

Linearly non-separable sets



Normal density



Two samples from normal distributions

Soft margin

Vapnik's proposal: Introduce the non-negative variables $\{\xi_i\}_{i=1}^N$ and claim only

$$y_i(\langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N. \tag{1.39}$$

One should minimize

$$\sum_{i=1}^{N} \xi_i$$

The target function is

$$\Phi(\boldsymbol{\psi}, \boldsymbol{\xi}) = \frac{1}{2}\langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle + C \sum_{i=1}^{N} \xi_i. \tag{1.40}$$

Minimize (1.40) under constraints (1.39) and $\xi_i \geq 0$, $i = 1, 2, \ldots, N$. It is called soft margin.

Apply the Kuhn–Tucker theorem to find the minimum of (1.40) under constraints (1.39) and $\xi_i \geq 0$, $i = 1, 2, \ldots, N$.

The Lagrange function is

$$L(\boldsymbol{\psi}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left(y_i(\langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle + b) - 1 + \xi_i \right) - \sum_{i=1}^{N} \beta_i \xi_i$$

where $\alpha_i \geq 0$, $\beta_i \geq 0$ are multipliers. We want to find saddle point.

Minimum with respect to $\boldsymbol{\psi}$, b and ξ_i, maximum with respect to α_i, β_i.

We obtain

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \alpha_i \alpha_j. \tag{1.41}$$

We have to maximize (1.41) with respect to $\alpha_1, \ldots, \alpha_N$ under constraints

$$\sum_{i=1}^{N} y_i \alpha_i = 0, \tag{1.42}$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N. \tag{1.43}$$

Denote the solution by $\alpha_1^0, \ldots, \alpha_N^0$. Then

$$\boldsymbol{\psi}_0 = \sum_{i=1}^{N} y_i \alpha_i^0 \boldsymbol{x}_i.$$

Moreover, if $0 < \alpha_i^0 < C$, then

$$y_i(\langle \boldsymbol{x}_i, \boldsymbol{\psi}_0 \rangle + b_0) = 1. \tag{1.44}$$

It gives b_0.

Finally the optimal hyperplane is:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \alpha_i^0 \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b_0 = 0.$$

Non-linear separation

If the two sets are not separable linearly, then we transform to a higher dimensional space and try to find linear separation in that space.

Let $\boldsymbol{\varphi}\colon \mathbb{R}^d \to \mathbb{R}^q, \quad d < q$, be a non-linear function. The image space is called feature space. We know, that to find the optimal hyperplane, we need only the inner product of the vectors.

The inner product:

$$\langle \boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{y}) \rangle = \sum_i \mu_i \varphi_i(\boldsymbol{x}) \varphi_i(\boldsymbol{y}), \quad \mu_i > 0.$$

Integral operator:

$$(A_K f)(x) = \int_a^b K(x,y) f(y) dy.$$

$\varphi : [a,b] \to \mathbb{R}$ is eigenfunction, μ is eigenvalue if $A_K \varphi = \mu \varphi$.

Mercer's theorem Let $K : [a,b] \times [a,b] \to \mathbb{R}$ continuous, symmetric, positive semi-definite function.

Then the expansion

$$K(x,y) = \sum_{i=1}^{\infty} \mu_i \varphi_i(x) \varphi_i(y) \tag{1.45}$$

is uniformly convergent.

Kernel functions:

1. Polynomial:
$$K(\boldsymbol{x}, \boldsymbol{x}_i) = (\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + 1)^p.$$
We should choose p.

2. Radial Basis Function:
$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{1}{2\sigma^2} \parallel \boldsymbol{x} - \boldsymbol{x}_i \parallel^2 \right).$$

3. Two-layer perceptron:
$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \tanh(\beta_0 \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + \beta_1).$$

The training set is

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N).$$

Instead of them we should separate

$$(\boldsymbol{\varphi}(\boldsymbol{x}_1), y_1), \ldots, (\boldsymbol{\varphi}(\boldsymbol{x}_N), y_N).$$

However, the optimal hyperplane has the form

$$\langle \boldsymbol{\psi}_0, \boldsymbol{x} \rangle + b_0 = 0$$

where

$$\boldsymbol{\psi}_0 = \sum_{i=1}^{N} y_i \alpha_i^0 \boldsymbol{x}_i,$$

and in the optimization problem only the inner products of the vectors are included, so we replace every inner product by a kernel function.

Replace $\langle\ ,\ \rangle$ by $K(\,.,\,.\,)$.

Then the separating surface is

$$f(\boldsymbol{x}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b = 0. \tag{1.46}$$

To find α_i and b we should maximize

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{1.47}$$

subject to

$$\sum_{i=1}^{N} y_i \alpha_i = 0, \tag{1.48}$$

$$0 \le \alpha_i \le C, \quad i = 1, \ldots, N. \tag{1.49}$$

Vectors $\boldsymbol{x}_i$ corresponding to non-zero α_i^0 are called support vectors.

We classify a point $\boldsymbol{z}$ according to

$$d(\boldsymbol{z}, \boldsymbol{\alpha}) = \operatorname{sgn}\left(\sum_{\boldsymbol{x}_i \text{ is a support vector}} y_i \alpha_i^0 K(\boldsymbol{z}, \boldsymbol{x}_i) + b_0 \right).$$

Sequential Minimal Optimization (SMO)

How to find the maximum of W?

Platt, Cristianini and Shawe-Taylor: SMO

SVM regression

Loss functions

Least squares:

$$(y - f(\boldsymbol{x}, \boldsymbol{\alpha}))^2,$$

where $\boldsymbol{x}$ is the input and y is the output.

Robust statistics (Huber): apply other loss functions.

ε-insensitive loss function:

$$L_\varepsilon(y - f(\boldsymbol{x}, \boldsymbol{\alpha})) = |y - f(\boldsymbol{x}, \boldsymbol{\alpha})|_\varepsilon$$

where

$$|y - f(\boldsymbol{x}, \boldsymbol{\alpha})|_\varepsilon = \begin{cases} 0, & \text{if} \quad |f(\boldsymbol{x}, \boldsymbol{\alpha}) - y| \le \varepsilon, \\ |f(\boldsymbol{x}, \boldsymbol{\alpha}) - y| - \varepsilon, & \text{if} \quad |f(\boldsymbol{x}, \boldsymbol{\alpha}) - y| > \varepsilon. \end{cases}$$

SVM for linear regression

We want to fit a function

$$f(\boldsymbol{x}, \boldsymbol{\psi}) = \sum_{i=1}^{d} \psi_i x^i + b,$$

where $\boldsymbol{x} = (x^1, \ldots, x^d)^\top \in \mathbb{R}^d$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)^\top \in \mathbb{R}^d$.

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ be the sample points (training set), where $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1, \ldots, N$.

Then the empirical risk is

$$R_{\text{emp}}(\boldsymbol{\psi}, b) = \frac{1}{N} \sum_{i=1}^{N} L_{\varepsilon_i}(y_i - \langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle - b).$$

We should minimize it.

Introduce the non-negative variables $\{\xi_i\}_{i=1}^N$ and $\{\xi_i^*\}_{i=1}^N$.

Consider the conditions

$$\begin{aligned} y_i - \langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle - b \le \varepsilon_i + \xi_i, &\quad i = 1, \ldots, N, \\ \langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle + b - y_i \le \varepsilon_i + \xi_i^*, &\quad i = 1, \ldots, N, \\ \xi_i \ge 0, &\quad i = 1, \ldots, N, \\ \xi_i^* \ge 0, &\quad i = 1, \ldots, N. \end{aligned} \tag{1.50}$$

The minimum of $R_{\text{emp}}(\boldsymbol{\psi}, b)$ in $\boldsymbol{\psi}$ and b is attained at the same point as the minimum of $\sum_{i=1}^N (\xi_i + \xi_i^*)$ with respect to ξ_i, ξ_i^*, $\boldsymbol{\psi}$, b and subject to the constrains (1.50).

Vapnik's proposal. Minimize

$$\Phi(\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\xi}^*) = C \left(\sum_{i=1}^{N} (\xi_i + \xi_i^*) \right) + \frac{1}{2} \langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle$$

subject to the constraints (1.50). Apply Kuhn–Tucker theorem. Introduce the non-negative multipliers α_i, α_i^*, β_i, β_i^*.

The Lagrange function:

$$\begin{aligned} L(\boldsymbol{\psi}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}, \boldsymbol{\beta}^*) = C \sum_i (\xi_i + \xi_i^*) + \frac{1}{2} \langle \boldsymbol{\psi}, \boldsymbol{\psi} \rangle + \sum_i \alpha_i (y_i - \langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle - b - \varepsilon_i - \xi_i) \\ + \sum_i \alpha_i^* (-y_i + \langle \boldsymbol{\psi}, \boldsymbol{x}_i \rangle + b - \varepsilon_i - \xi_i^*) - \sum_i \beta_i \xi_i - \sum_i \beta_i^* \xi_i^*. \end{aligned}$$

We have to find its saddle-point, minimize with respect to $\boldsymbol{\psi}$, b, ξ_i, ξ_i^* , maximize with respect to α_i, α_i^*, β_i, β_i^* .

We should maximize

$$W(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \sum_{i=1}^{N} y_i (\alpha_i - \alpha_i^*) - \sum_{i=1}^{N} \varepsilon_i (\alpha_i + \alpha_i^*) - \frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{1.51}$$

subject to the constraints

$$\begin{aligned} &\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ &\qquad 0 \le \alpha_i \le C, \quad i = 1, \ldots, N, \\ &\qquad 0 \le \alpha_i^* \le C, \quad i = 1, \ldots, N. \end{aligned} \tag{1.52}$$

Solving this quadratic programming problem, we find the optimal $\boldsymbol{\alpha}^0$ and $\boldsymbol{\alpha}^{*0}$. Then

$$\boldsymbol{\psi}^0 = \sum_{i=1}^{N}(\alpha_i^0 - \alpha_i^{*0})\boldsymbol{x}_i.$$

The optimal b is obtained by minimizing $R_{\text{emp}}(\boldsymbol{\psi}^0, b)$. The regression hyperplane is

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(\alpha_i^0 - \alpha_i^{*0})\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b^0.$$

SVM for non-linear regression

Apply kernel functions. Maximize

$$W(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) - \sum_{i=1}^{N} \varepsilon_i(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

subject to the constrains (1.52). Denote the maximum point by α_i^0, α_i^{*0}, $i = 1, 2, \ldots, N$. Finally, the approximating function is

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(\alpha_i^0 - \alpha_i^{*0})K(\boldsymbol{x}, \boldsymbol{x}_i) + b^0.$$

References

TEXTBOOKS

Haykin, Simon: Neural networks. A comprehensive foundation, 1999, Prentice Hall, London

Vapnik, V. N.: Statistical Learning Theory, 1998, John Wiley and Sons Inc., New York

Cristianini, Nello and Shawe-Taylor, John: An Introduction to Support Vector Machines and Other Kernel-based Methods, 2000, Cambridge, Cambridge University Press

Luc, Devroye and László, Györfi and Gábor, Lugosi: A probabilistic theory of pattern recognition, 1996, New York, Springer

T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Prediction, Inference and Data Mining, Second Edition, Springer Verlag, 2009

Fletcher, R. Practical methods of optimization. Second edition. A Wiley-Interscience Publication. John Wiley and Sons, Ltd., Chichester, 1987.

Prakasa Rao, B. L. S. Nonparametric functional estimation. Probability and Mathematical Statistics. Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers), New York, 1983

Boyd, Stephen; Vandenberghe, Lieven: Convex optimization. Cambridge University Press, Cambridge, 2004

Martin T. Hagan, Howard B. Demuth, Mark H. Beale: Neural Network Design. Boston, MA: PWS Publishing, 1996

RESEARCH PAPERS AND BOOKS

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, vol 5 issue 4:115 - 133.

Hebb, D. O. (1949). The Organization of Behavior: A neuropsychological theory. New York: Wiley.

Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review, v65, No. 6, pp. 386408.

Novikoff, A. B. (1962). On convergence proofs on perceptrons. Proceedings of Symposium on the Mathematical Theory of Automata, vol 12, 615-622. Polytechnic Institute of Brooklyn.

Widrow, B., and Hoff, M. E., Jr., 1960, Adaptive switching circuits, in 1960 IRE WESCON Convention Record, Part 4, New York: IRE, pp. 96104

Marvin Minsky and Seymour Papert, 1972 (2nd edition with corrections, first edition 1969) Perceptrons: An Introduction to Computational Geometry, The MIT Press, Cambridge MA,

David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams: Learning representations by back-propagating errors. Nature 323, 533 - 536 (09 October 1986)

Platt, John C.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998, Microsoft Research, Technical Report MSR-TR-98-14, `http://research.microsoft.com/en-us/um/people/jplatt/smoTR.pdf`

Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". Biological Cybernetics 43 (1): 5969.

V. Vapnik and A. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities." Theory of Probability and its Applications, 16(2):264280, 1971.

Tikhonov, A.N. Solution of incorrectly formulated problems and the regularization method. (English. Russian original) Sov. Math., Dokl. 5, 1035-1038 (1963); translation from Dokl. Akad. Nauk SSSR 151, 501-504 (1963).

Tikhonov, A. N.; Arsenin, V. Ya. Metody resheniya nekorrektnykh zadach. (Russian) [Methods for the solution of ill-posed problems] Izdat. "Nauka", Moscow, 1974.

Kuhn, H. W.; Tucker, A. W. Nonlinear programming. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950, pp. 481492. University of California Press, Berkeley and Los Angeles, 1951.

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

MATLAB The Language of Technical Computing. MathWorks,
`http://www.mathworks.com/products/matlab/`

2 Strong laws of large numbers

Topics

1. *The phenomenon*

random walks with normalizations.

2. *The laws*

laws of large numbers (LLN),
the law of the iterated logarithm (LIL),
the central limit theorem.

3. *Significance*

relative frequency and probability,
the fundamental theorem of mathematical statistics,
consistency of estimators,
Monte Carlo methods.

4. *Weak laws and strong laws*

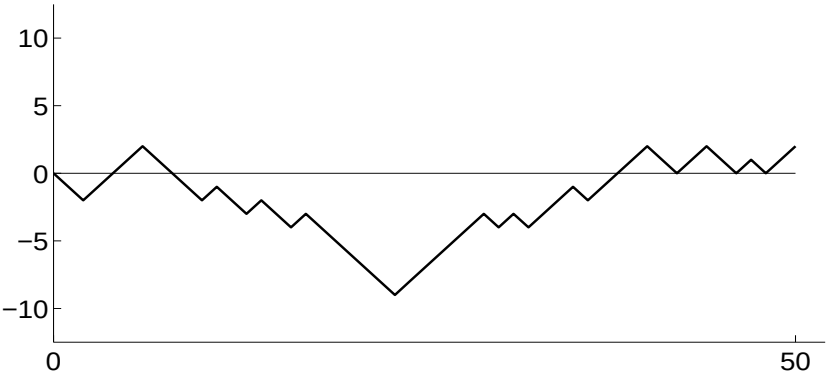
stochastic convergence, almost sure convergence.

5. *Pairwise independence, complete independence*

Etemadi's theorem, Kolmogorov's theorem.

1. The phenomenon: the random walk

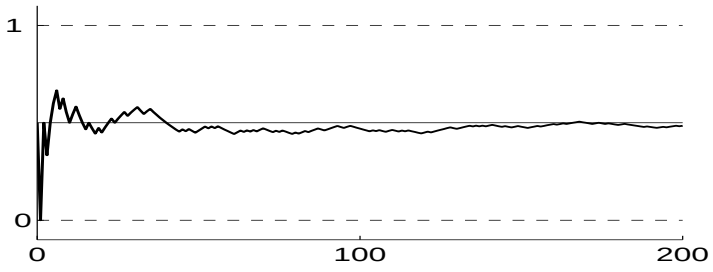
$\xi_i = \pm 1$ in the coin tossing experiment, $S_n = \sum_{i=1}^{n} \xi_i$



A trajectory of the symmetric random walk (S_n)

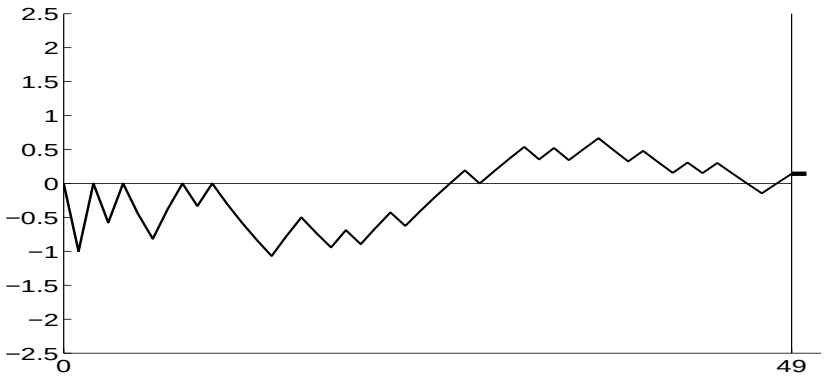
The law of large numbers

$S_n/n \to 0$



The normalized trajectory of the symmetric random walk (S_n/n)

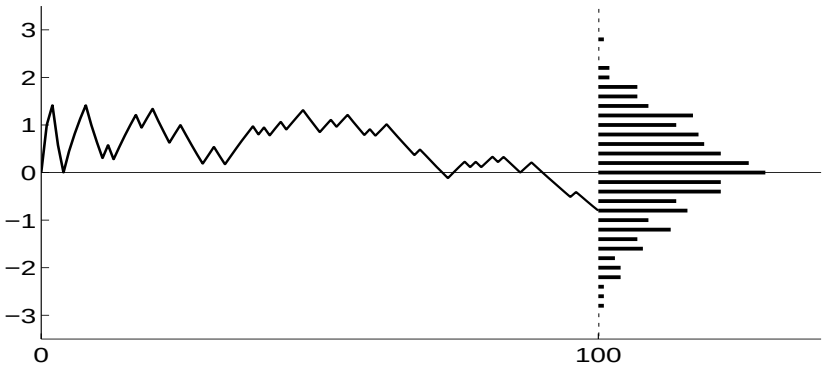
Standardization, another phenomenon



The standardized random walk $(S_n/\sqrt{n})$

The central limit theorem

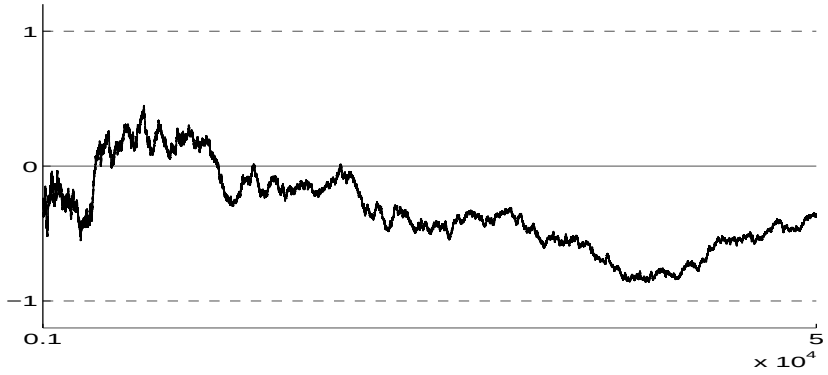
$S_n/\sqrt{n} \approx \mathcal{N}(0,1)$



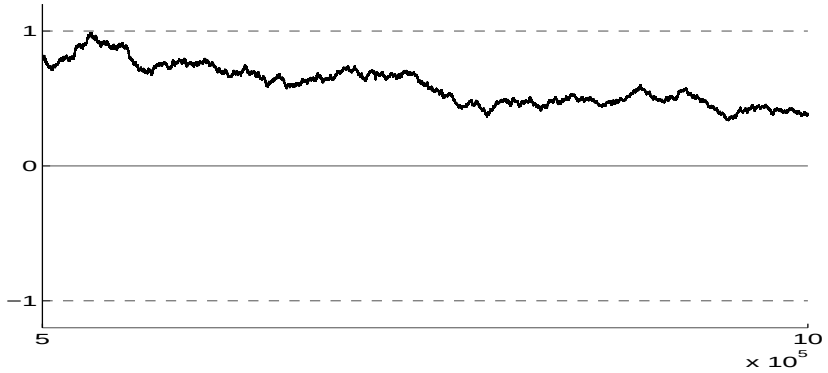
300 repetitions of the first 100 steps of the random walk;
the histogram of $S_{100}/\sqrt{100}$ is bell shaped

Between the two phenomena: LIL

$$P\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1\right) = 1.$$



n: 1 000 – 50 000



LIL; n: 500 000 – 1 000 000

2. The main features of limit theorems

Definition 2.1. We say that the sequence of random variables $\eta_1, \eta_2, \dots$ **converges in probability** to a random variable η, if

$$\forall\, \varepsilon > 0 \qquad \text{we have} \qquad \lim_{n\to\infty} P\big(|\eta_n - \eta| > \varepsilon\big) = 0.$$

In other words: stochastic convergence, convergence in measure.

Definition 2.2. We say that the sequence of random variables $\eta_1, \eta_2, \dots$ **converges almost surely (a.s.)** to a random variable η, if there exists an event N with $P(N) = 0$ and

$$\lim_{n\to\infty} \eta_n(\omega) = \eta(\omega), \quad \text{for} \quad \omega \in \Omega \setminus N.$$

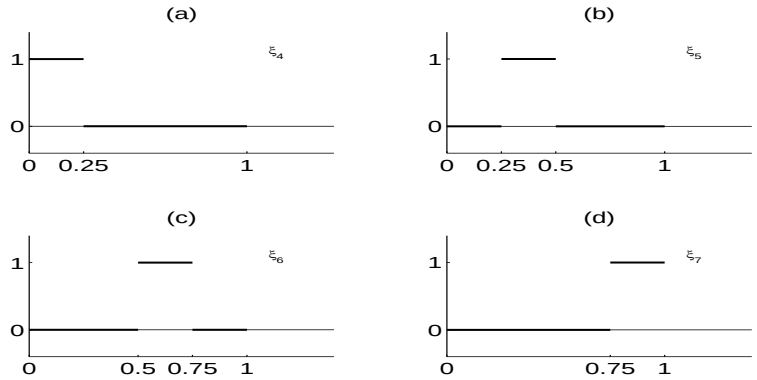
In other words: convergence with probability 1, convergence almost everywhere.

Almost sure convergence is point-wise convergence except a set of zero probability. Almost sure convergence implies stochastic convergence. Stochastic convergence does not imply almost sure convergence. Therefore in the strong laws the convergence is stronger than in the weak laws.

Example 2.1. We construct a sequence converging in probability but not converging almost surely.
Let $(\Omega, \mathcal{A}, P)$ be the interval $[0, 1)$ endowed with the σ-algebra of the Borel sets and with

the Lebesgue measure.
Let $\xi_1(\omega) = 1,\ \omega \in [0,1)$;
$\xi_2(\omega) = 1$, if $\omega \in [0,1/2)$ and 0 otherwise;
$\xi_3(\omega) = 1$, if $\omega \in [1/2,1)$ and 0 otherwise;
$\xi_4(\omega) = 1$, if $\omega \in [0,1/4)$ and 0 otherwise;
.
.
.
$\xi_8(\omega) = 1$, if $\omega \in [0,1/8)$ and 0 otherwise;
...

As the length of the intervals where $\xi_n \neq 0$, converges to 0, so $\xi_n \to 0$ in probability. However the interval, where $\xi_n = 1$, 'returns infinitely many times' above any point, so in the sequence $\xi_n(\omega)$ there are infinitely many zeros, and infinitely many ones.
So $\xi_n(\omega)$ in not convergent, $\omega \in [0,1)$.



We say that $\xi_1, \xi_2, \ldots$ is fundamental (Cauchy) with probability 1 if $\xi_1(\omega), \xi_2(\omega), \ldots$ is fundamental (Cauchy) for almost all $\omega \in \Omega$.

Theorem 2.1. $\lim_{n\to\infty} \xi_n = \xi$ *almost surely iff*

$$\mathbb{P}\left\{\sup_{k\geq n} |\xi_k - \xi| \geq \varepsilon\right\} \to 0, \tag{2.1}$$

as $n \to \infty$, *for every* $\varepsilon > 0$.

Theorem 2.2. *Let*

$$\sum_{k=1}^{\infty} \mathbb{P}\left\{|\xi_k - \xi| \geq \varepsilon\right\} < \infty, \tag{2.2}$$

for every $\varepsilon > 0$.

Then $\lim_{n\to\infty} \xi_n = \xi$ *almost surely.*

Theorem 2.3. *The sequence* $\xi_1, \xi_2, \ldots$ *is fundamental with probability 1 iff*

$$\mathbb{P}\left\{\sup_{k\geq 0} |\xi_{n+k} - \xi_n| \geq \varepsilon\right\} \to 0, \tag{2.3}$$

as $n \to \infty$, *for every* $\varepsilon > 0$.

Let $\xi_1, \xi_2, \dots$ be independent identically distributed random variables;
let $S_n = \sum_{i=1}^n \xi_i$.

The Marcinkiewicz SLLN
Let $\mathbb{E}|\xi_i|^r < \infty$, where $0 < r < 2$, and let $m = \mathbb{E}\xi_i$, if $r \geq 1$, and $m = 0$ otherwise. Then

$$\lim_{n\to\infty} \frac{S_n - nm}{n^{1/r}} = 0 \qquad \text{almost surely.}$$

Kolmogorov's SLLN: when $r = 1$.

Rate of convergence: Hsu, Robbins, Erdős, Baum, Katz,...
Non-identically distributed r.v.'s
Not independent r.v.'s
Banach space valued case
Multiindex case
The law of the iterated logarithm
If $0 < \sigma^2 = \mathbb{E}\xi_1^2 < \infty$, $\mathbb{E}\xi_1 = 0$, then

$$P\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = 1\right) = 1,$$

$$P\left(\liminf_{n\to\infty} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = -1\right) = 1.$$

The central limit theorem
Assume $\sigma^2 = \mathbb{D}^2\xi_1$ is finite and positive, let $m = \mathbb{E}\xi_1$. Then

$$\lim_{n\to\infty} P\left(\frac{S_n - nm}{\sqrt{n\sigma}} < x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt\,, \qquad \forall x \in \mathbb{R}\,.$$

Rate of convergence: Berry-Esseen
Local CLT
Non-identically distributed r.v.'s: Lyapunov, Lindeberg
Not independent r.v.'s
Functional CLT: Donsker's theorem

An interplay between SLLN and CLT
Almost sure central limit theorems
Let $X_1, X_2, \dots,$ be i.i.d. real r.v.'s with mean 0 and variance 1.
Let $S_k = X_1 + \dots + X_k$.
Then

$$\frac{1}{\log n} \sum_{k=1}^{n} \frac{1}{k}\, \delta_{\frac{S_k(\omega)}{\sqrt{k}}} \;\Rightarrow\; \mathcal{N}(0,1)\,, \qquad \text{for almost every } \omega \in \Omega\,, \tag{2.4}$$

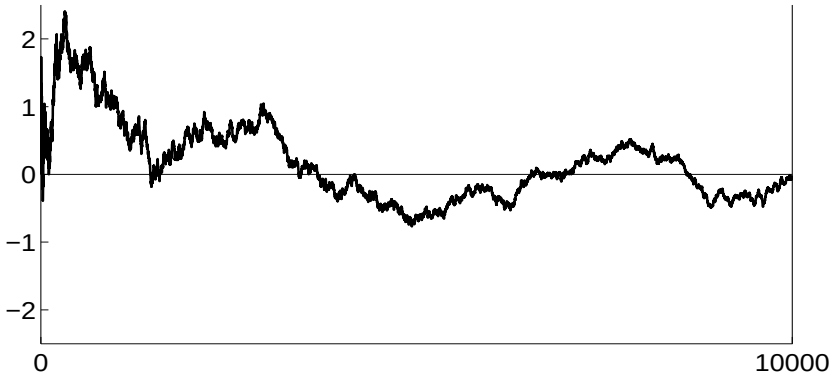
if $n \to \infty$, where δ_x is the unit mass at point x, $\Rightarrow \mathcal{N}(0,1)$ denotes weak convergence to the standard normal law.

See Brosamler (1988), Schatte (1988), Lacey–Philipp (1990), Rychlik (1994),...

A.s. CLT

A trajectory of a standardized random walk

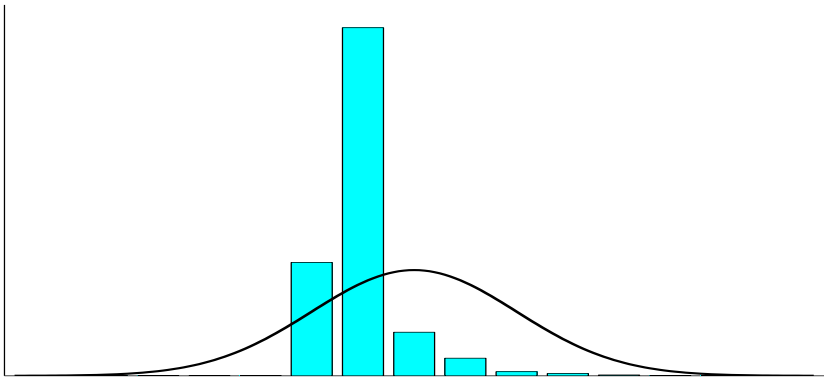
$S_k/\sqrt{k}$, $k = 1, \ldots, 10000$



10 000 steps of a random walk

The histogram of the discrete distribution

$$P\left(\eta = \frac{S_k(\omega)}{\sqrt{k}}\right) \approx \frac{1}{\log n}\frac{1}{k}, \quad k = 1, \ldots, 10000.$$



3. Applications of the LLN: relative frequency and probability

Consider an experiment $\mathcal{K}$ and an event A with $P(A) = p$.

Repeat $\mathcal{K}$ n times under fixed conditions and independently.

Denote k_A/n the relative frequency of A.

Then $k_A/n = S_n/n$, if $S_n = \xi_1 + \cdots + \xi_n$, where ξ_i is the indicator of A during the i-th performance of the experiment.

Then $\xi_1, \ldots, \xi_n$ are independent with Bernoulli distribution, i.e. $P(\xi_i = 1) = p$, $P(\xi_i = 0) = 1 - p$, $m = \mathbb{E}\xi_i = p = P(A)$.

Therefore

$$\lim_{n\to\infty} \frac{k_A}{n} = \lim_{n\to\infty} \frac{S_n}{n} = m = P(A).$$

It is the Bernoulli LLN.

Theorem in probability theory: The relative frequency converges to the probability.

Empirical fact: in a real life experiment the relative frequency 'converges' to the probability.

So our model is in accordance with the empirical facts.

Applications in mathematical analysis

The Weierstrass approximation theorem; Bernstein polynomials

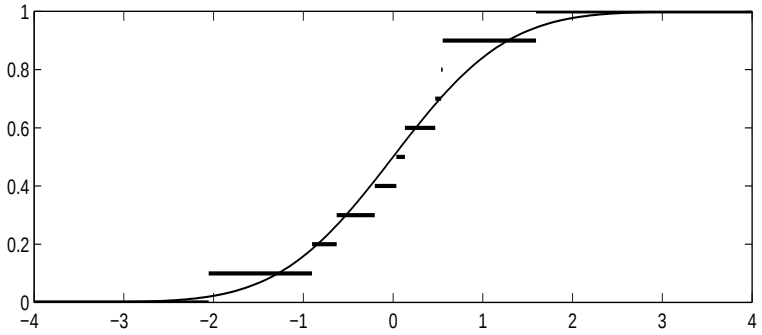
Applications in statistics

Sample in statistics: independent identically distributed random variables $\xi_1, \dots, \xi_n$.

(1) The empirical mean (sample mean, average) is a consistent estimator of the theoretical mean. That is $\bar{\xi} \to m$, where $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ is the empirical mean and $m = \mathbb{E}\xi$ is the expectation.

(2) The fundamental theorem of mathematical statistics.

The following is true with probability 1: the empirical distribution function F_n^* converges to the theoretical distribution function F uniformly on the whole real line.



Monte Carlo methods (stochastic simulation)

Calculating integrals. Let $f : [0,1] \longrightarrow [0,1]$.

$$\int_0^1 f(x)\,dx = ?$$

Let $\xi_1, \eta_1, \xi_2, \eta_2, \dots$ be independent, uniformly distributed on $[0,1]$. Then (ξ_i, η_i), $i = 1, 2, \dots$, are independent and uniformly distributed on the unit square.

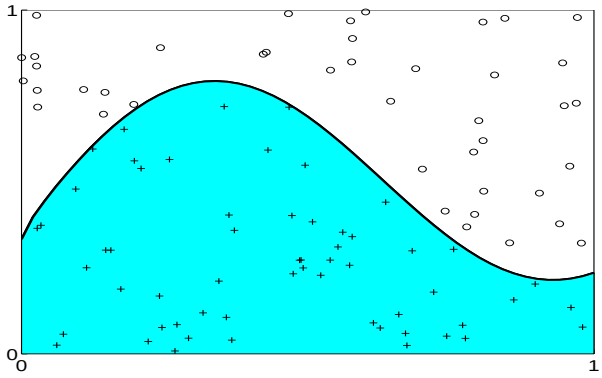
Let

$$\varrho_i = \begin{cases} 1, & \text{if} \quad f(\xi_i) > \eta_i \\ 0, & \text{if} \quad f(\xi_i) \le \eta_i \end{cases} .$$

Then $\varrho_1, \dots, \xi_n$ are independent and identically distributed. $\mathbb{E}\varrho_i = P\left(f(\xi_i) > \eta_i\right) = \int_0^1 f(x)\,dx$. By the SLLN

$$\lim_{n\to\infty} \frac{1}{n} \sum\nolimits_{i=1}^{n} \varrho_i = \int_0^1 f(x)\,dx \qquad \text{almost surely.}$$

So, observing the sequence (ξ_i, η_i), $i = 1, 2, \dots, n$, we can calculate $\int_0^1 f(x)\,dx$. An approximation of the integral is the number of points (ξ_i, η_i) falling below the curve divided by the total number of points (ξ_i, η_i).



4. Weak law: stochastic convergence; strong law: almost sure convergence

Kolmogorov's zero-one law

Let $\mathcal{A}_1, \mathcal{A}_2, \dots$ be σ-algebras.

Definition 2.3. Let

$$\mathcal{A}_k^\infty = \sigma\{\mathcal{A}_k, \mathcal{A}_{k+1}, \dots\}, \qquad \mathcal{T} = \cap_{n=1}^\infty \mathcal{A}_n^\infty$$

$\mathcal{T}$ is called the tail σ-algebra.

Theorem 2.4 (Kolmogorov's zero-one law)**.** *Let $\mathcal{A}_1, \mathcal{A}_2, \dots$ be independent σ-algebras. If $A \in \mathcal{T}$, then $\mathbb{P}(A)$ is either zero or one.*

Kolmogorov's Inequality

Let $\xi_1, \xi_2, \dots, \xi_n$ be random variables and let $S_k = \sum_{i=1}^k \xi_i$ be the partial sum.

Theorem 2.5 (Kolmogorov's Inequality)**.** *Let $\xi_1, \xi_2, \dots, \xi_n$ be independent random variables with $\mathbb{E}\xi_i = 0$, $\mathbb{E}\xi_i^2 < \infty$, $i \le n$.*

(a) Then for every $\varepsilon > 0$

$$\mathbb{P}\left\{\max_{1\le k\le n} |S_k| \ge \varepsilon\right\} \le \frac{\mathbb{E}S_n^2}{\varepsilon^2}. \tag{2.5}$$

(b) If also $\mathbb{P}(|\xi_i| \le c) = 1$, $i \le n$, then

$$\mathbb{P}\left\{\max_{1\le k\le n} |S_k| \ge \varepsilon\right\} \ge 1 - \frac{(c+\varepsilon)^2}{\mathbb{E}S_n^2}. \tag{2.6}$$

Proof. First hitting time of level ε:

$$A_k = \{|S_i| < \varepsilon,\ i = 1, 2, \dots, k-1,\ |S_k| \ge \varepsilon\}.$$

One-Series Theorem

Theorem 2.6 (Kolmogorov and Khinchin). *Let $\xi_1, \xi_2, \dots$ be independent random variables with $\mathbb{E}\xi_i = 0$.*

(a) Then if

$$\sum \mathbb{E}\xi_n^2 < \infty, \tag{2.7}$$

the series $\sum \xi_n$ converges with probability 1.

(b) If the random variables ξ_n, $n \geq 1$, are uniformly bounded (i.e., $\mathbb{P}(|\xi_n| \leq c) = 1$, $c < \infty$), the converse is true: the convergence of $\sum \xi_n$ with probability 1 implies (2.7).

Proof. (a) Use Theorem 2.3 and part (a) of Kolmogorov's inequality.
(b) Use Theorem 2.3 and part (b) of Kolmogorov's inequality.

Two-Series Theorem

Theorem 2.7 (Two-Series Theorem). *(a) A sufficient condition for the convergence of the series $\sum \xi_n$ of independent random variables, with probability 1, is that both series $\sum \mathbb{E}\xi_n$ and $\sum \mathbb{D}^2\xi_n$ converge.*

(b) If $\mathbb{P}(|\xi_n| \leq c) = 1$, the condition is also necessary.

Proof. (a) Use One-Series Theorem.
(b) Symmetrization and use One-Series Theorem.

Notation:

$$\xi^c = \xi, \quad \text{if} \quad |\xi| \leq c, \quad \text{and} \quad \xi^c = 0, \quad \text{if} \quad |\xi| > c.$$

Kolmogorov's Three-Series Theorem

Theorem 2.8 (Kolmogorov's Three-Series Theorem). *Let $\xi_1, \xi_2, \dots$ be a sequence of independent random variables.*

(a) A sufficient condition for the convergence of $\sum \xi_n$ with probability 1 is that the series

$$\sum \mathbb{E}\xi_n^c, \quad \sum \mathbb{D}^2\xi_n^c, \quad \sum \mathbb{P}(|\xi_n| \geq c)$$

converge for some $c > 0$;

(b) a necessary condition is that these series converge for every $c > 0$.

Proof. Use Two-Series Theorem.

The Borel-Cantelli lemma

a) If $\sum_{i=1}^{\infty} P(A_i) < \infty$*, then the probability that infinitely many of the events* A_i *occur is 0.*

b) Let the events $A_1, A_2, \ldots$ *be independent. If* $\sum_{n=1}^{\infty} P(A_n) = \infty$*, then the probability that infinitely many of the events* A_i *occur is 1.*

Remark that Erdős and Rényi proved, that part b) of the Borel-Cantelli lemma is true for pairwise independent events, too.

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

means that A_n occurs for infinitely many n.

Proof of the Borel-Cantelli lemma

a) We should prove that $\sum_{i=1}^{\infty} P(A_i) < \infty$ implies $P(\limsup A_i) = 0$.

$P(\limsup A_i) = P(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k) \le P(\bigcup_{k=n}^{\infty} A_k)$ for every n.

But

$$P(\bigcup_{k=n}^{\infty} A_k) \le \sum_{k=n}^{\infty} P(A_k) \to 0, \quad \text{as } n \to \infty,$$

as it is the tail of a convergent series.

b) We should prove that $\sum_{i=1}^{\infty} P(A_n) = \infty$ implies $P(\limsup A_n) = 1$.

By the continuity of the probability,

$$P(\limsup A_n) = P(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k) = \lim_{n\to\infty} P(\bigcup_{k=n}^{\infty} A_k) = \lim_{n\to\infty} \lim_{m\to\infty} P(\bigcup_{k=n}^{m} A_k). \tag{2.8}$$

Using de Morgan's law and the independence,

$$P(\overline{\bigcup_{k=n}^{m} A_k}) = P(\bigcap_{k=n}^{m} \overline{A}_k) = \prod_{k=n}^{m} P(\overline{A}_k).$$

Applying $P(\overline{A}_k) = 1 - P(A_k) \le \exp\{-P(A_k)\}$, the previous expression has the following upper bound

$$\prod_{k=n}^{m} \exp\{-P(A_k)\} = \exp\{-\sum_{k=n}^{m} P(A_k)\}.$$

By our assumption, this expression converges to 0, as $m \to \infty$. So $\lim_{m\to\infty} P(\bigcup_{k=n}^{m} A_k) = 1$. Therefore, by (2.8), $P(\limsup A_n) = 1$.

Lemma 2.1 (Chebyshev's inequality)**.** *Assume that the variance of* ξ *is finite. Then for any* $\varepsilon > 0$ *we have*

$$P\big(|\xi - \mathbb{E}\xi| \ge \varepsilon\big) \le \mathbb{D}^2(\xi)/\varepsilon^2.$$

Cantelli's SLLN

Theorem 2.9 (Cantelli)**.** *Let $\xi_1, \xi_2, \dots$ be independent random variables with finite fourth moments and let*

$$\mathbb{E}|\xi_n - \mathbb{E}\xi_n|^4 \le C, \quad n \ge 1,$$

for some constant C.

Then as $n \to \infty$

$$\frac{S_n - \mathbb{E}S_n}{n} \to 0 \quad (P\text{-a.s.}). \tag{2.9}$$

Proof. Apply Chebysev's inequality and the Borel-Cantelli lemma.

Weak laws of large numbers

Let $\xi_1, \xi_2, \dots$ be a sequence of r.v.'s, and $S_n = \xi_1 + \cdots + \xi_n$, $n = 1, 2, \dots$, be the sequence of their partial sums.
The weak law of large numbers states the stochastic convergence of S_n/n.

Theorem 2.10. *Let $\xi_1, \xi_2, \dots$ be pairwise independent identically distributed r.v.'s. Assume $\mathbb{E}\xi_i^2 < \infty$. Denote $m = \mathbb{E}\xi_i$ the common expectation. Then*

$$\lim_{n\to\infty} \frac{S_n}{n} = m \qquad \textit{in probability.}$$

So the average of the the observations converges to the theoretical mean.
'Weak' means that the convergence is stochastic. That is for large n's S_n/n is close to m with high probability. Stochastic convergence is a metric convergence. Almost sure convergence is not a metric convergence.

Khintchine proved that the above theorem remains valid if instead of $\mathbb{E}\xi_i^2 < \infty$ we assume only $\mathbb{E}|\xi_i| < \infty$.

Proof of the weak law

Proof. By Chebyshev's inequality, for any $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) = P\left(\left|\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\right| > \varepsilon\right) \le$$

$$\le \frac{1}{\varepsilon^2}\mathbb{D}^2\left(\frac{S_n}{n}\right) = \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n}\mathbb{D}^2\xi_i = \frac{1}{\varepsilon^2 n}\mathbb{D}^2\xi_1 \to 0,$$

as $n \to \infty$. (We used that the variance is additive for pairwise independent r.v.'s.)

5. Independence

Khintchine (1929): pairwise independence ⇒ WLLN
Kolmogorov (1933): (complete) independence ⇒ SLLN
Etemadi (1981): pairwise independence ⇒ SLLN

The Toeplitz Lemma

Lemma 2.1 (Toeplitz). *Let $\{a_n\}$ be a sequence of nonnegative numbers, $b_n = \sum_{i=1}^{n} a_i$, $b_n > 0$ for $n \geq 1$, and $b_n \uparrow \infty$, $n \to \infty$.*

Let $\{x_n\}$ be a sequence of numbers converging to x.

Then

$$\frac{1}{b_n} \sum_{j=1}^{n} a_j x_j \to x. \tag{2.10}$$

In particular, if $a_n = 1$ then

$$\frac{x_1 + \cdots + x_n}{n} \to x. \tag{2.11}$$

The Kronecker Lemma

Lemma 2.2 (Kronecker). *Let $\{b_n\}$ be an increasing sequence of positive numbers, $b_n \uparrow \infty$, $n \to \infty$,*
and let $\{x_n\}$ be a sequence of numbers such that $\sum x_n$ converges.

Then

$$\frac{1}{b_n} \sum_{j=1}^{n} b_j x_j \to 0, \quad n \to \infty. \tag{2.12}$$

In particular, if $b_n = n$, $x_n = \dfrac{y_n}{n}$ and $\sum \left(\dfrac{y_n}{n}\right)$ converges, then

$$\frac{y_1 + \cdots + y_n}{n} \to 0, \quad n \to \infty. \tag{2.13}$$

Proof. Apply the Toeplitz Lemma.

A Kolmogorov's SLLN

Theorem 2.11 (Kolmogorov). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite second moments, and let there be positive numbers b_n such that $b_n \uparrow \infty$ and*

$$\sum \frac{\mathbb{D}^2 \xi_n}{b_n^2} < \infty. \tag{5}$$

Then

$$\frac{S_n - \mathbb{E}S_n}{b_n} \to 0 \quad (P\text{-a.s.}). \tag{2.14}$$

In particular, if

$$\sum \frac{\mathbb{D}^2 \xi_n}{n^2} < \infty \tag{2.15}$$

then

$$\frac{S_n - \mathbb{E}S_n}{n} \to 0 \quad (P\text{-a.s.}). \tag{2.16}$$

Proof. Apply the two-series theorem and Kronecker's Lemma.

A Lemma for Moments

Lemma 2.3. *Let ξ be a nonnegative random variable. Then*

$$\sum_{n=1}^{\infty} \mathbb{P}(\xi \geq n) \leq \mathbb{E}\xi \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}(\xi \geq n). \tag{2.17}$$

Kolmogorov's SLLN

Theorem 2.12 (Kolmogorov)**.** *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables. If $\mathbb{E}|\xi_1| < \infty$, then*

$$\frac{S_n}{n} \to m \quad (P\text{-a.s.}) \tag{2.18}$$

where $m = \mathbb{E}\xi_1$.

If

$$\frac{S_n}{n} \to C \quad (P\text{-a.s.}) \tag{2.19}$$

where C is a finite constant, then $\mathbb{E}|\xi_1| < \infty$.

Proof. Use truncation, Lemma 2.3, Theorem 2.11.

Etemadi's SLLN

It turned out that pairwise independence implies the strong law. Etemadi's result implies both Khintchine's WLLN, and Kolmogorov's SLLN:

Theorem 2.13 (Etemadi's SLLN)**.** *Let $\xi_1, \xi_2, \ldots$ be pairwise independent identically distributed r.v.'s.*

Let $S_n = \xi_1 + \cdots + \xi_n$. Assume $\mathbb{E}|\xi_i| < \infty$, $m = \mathbb{E}\xi_i$. Then

$$\lim_{n\to\infty} \frac{S_n}{n} = m \qquad \textit{almost surely.} \tag{2.20}$$

Lemma 2.2. *Let ξ be a non-negative r.v. Then*

$$\mathbb{E}\xi < \infty \quad \textit{if and only if} \quad \sum\nolimits_{n=1}^{\infty} P(\xi \geq n) < \infty.$$

Proof.

$$\sum_{n=1}^{\infty} P(\xi \geq n) = \sum_{n=1}^{\infty}\sum_{k=n}^{\infty} P(k \leq \xi < k+1) =$$

$$= \sum_{k=1}^{\infty} kP(k \leq \xi < k+1) \leq \sum_{k=0}^{\infty} \mathbb{E}\xi\chi_{\{k\leq\xi<k+1\}} = \mathbb{E}\xi = \cdots \leq$$

$$\leq \sum_{k=0}^{\infty} (k+1)P(k \leq \xi < k+1) = \cdots = 1 + \sum_{n=1}^{\infty} P(\xi \geq n).$$

Proof of Etemadi's SLLN

Let

$$\xi_i^+ = \xi_i \chi_{\{\xi_i \geq 0\}}, \qquad \xi_i^- = -\xi_i \chi_{\{\xi_i \leq 0\}},$$

the positive and the negative parts of ξ_i, $i = 1, 2, \ldots$. Then both $\xi_1^+, \xi_2^+, \ldots$, and $\xi_1^-, \xi_2^-, \ldots$ are sequences of identically distributed, pairwise independent r.v.'s. Moreover

$$\xi_i = \xi_i^+ - \xi_i^-.$$

Therefore, if the theorem is valid for the positive parts and the negative parts, the it valid for ξ_i. So it is enough to prove for non-negative r.v.'s.

Assume $\xi_i \geq 0$. Truncate ξ_i at level i:

$$\eta_i = \xi_i \chi_{\{\xi_i \leq i\}}, \quad i = 1, 2, \ldots; \quad S_n^* = \eta_1 + \cdots + \eta_n, \quad n = 1, 2, \ldots.$$

Let $\alpha > 1$ be fixed and let $k_n = [\alpha^n]$ (integer part).

First we prove that a subsequence of S_n^*/n is convergent. More precisely, we shall prove that

$$\lim_{n\to\infty} \frac{S_{k_n}^*}{k_n} = m, \quad \text{almost surely.} \tag{2.21}$$

To this end, first we prove, that

$$\lim_{n\to\infty} \mathbb{E} S_{k_n}^*/k_n = m. \tag{2.22}$$

By the monotone convergence theorem,

$$m = \mathbb{E}\xi_1 = \lim_{n\to\infty} \mathbb{E}\xi_1 \chi_{\{\xi_1 \leq n\}} = \lim_{n\to\infty} \mathbb{E}\xi_n \chi_{\{\xi_n \leq n\}} = \lim_{n\to\infty} \mathbb{E}\eta_n.$$

However, $\mathbb{E}S_{k_n}^*/k_n$ is the arithmetical mean of the numbers $\mathbb{E}\eta_1, \ldots, \mathbb{E}\eta_{k_n}$. As the sequence of the arithmetical means of a convergent sequence has the same limit as the limit of the sequence itself, therefore (2.22) follows.

Later we shall show that for any $\varepsilon > 0$

$$\Delta = \sum_{n=1}^{\infty} P\left(\left| \frac{S_{k_n}^* - \mathbb{E}S_{k_n}^*}{k_n} \right| > \varepsilon \right) < \infty. \tag{2.23}$$

But (2.23) and (2.22) imply (2.21).

To see it, we remark that by Borel-Cantelli's lemma and (2.23), for any $\varepsilon > 0$ the following event has probability 1: $|S_{k_n}^* - \mathbb{E}S_{k_n}^*|/k_n$ can be greater than ε only for finitely many indices k_n. Applying this property for a null sequence of ε's, we obtain

$$(S_{k_n}^* - \mathbb{E}S_{k_n}^*)/k_n \to 0, \qquad \text{almost surely.}$$

This and (2.22) imply (2.21).

Now we prove that (2.21) is valid for the whole sequence S_n/n, that is

$$\lim_{n\to\infty} \frac{S_{k_n}}{k_n} = m, \quad \text{almost surely.} \tag{2.24}$$

We have

$$\sum_{n=1}^{\infty} P(\eta_n \neq \xi_n) = \sum_{n=1}^{\infty} P(\xi_n > n) = \sum_{n=1}^{\infty} P(\xi_1 > n) < \infty$$

because $\mathbb{E}|\xi_1| < \infty$. So, by the Borel-Cantelli lemma, with probability 1: $\xi_n \neq \eta_n$ can happen only for finitely many n. Therefore (2.21) implies (2.24).

Now we turn to the indices not belonging to the subsequence k_n. Let $k_n < i \leq k_{n+1}$. Then, by the non-negativity of ξ_i,

$$\frac{S_{k_n}}{k_{n+1}} \leq \frac{S_i}{i} \leq \frac{S_{k_{n+1}}}{k_n}\,.$$

Therefore

$$\frac{k_n}{k_{n+1}}\frac{S_{k_n}}{k_n} \leq \frac{S_i}{i} \leq \frac{k_{n+1}}{k_n}\frac{S_{k_{n+1}}}{k_{n+1}}.$$

So, using (2.24),

$$\frac{1}{\alpha}m \leq \liminf_{n\to\infty} \frac{S_n}{n} \leq \limsup_{n\to\infty} \frac{S_n}{n} \leq \alpha m$$

almost surely. It is valid for any $\alpha > 1$, so (2.20) is true. □

To complete the proof we have to prove (2.23)**,** that is

$$\Delta = \sum_{n=1}^{\infty} P\left(\left|\frac{S^*_{k_n} - \mathbb{E}S^*_{k_n}}{k_n}\right| > \varepsilon\right) < \infty.$$

Using Chebyshev's inequality, the additivity of the variance for pairwise independent r.v.'s, and the fact that $\mathbb{D}^2(X) \leq \mathbb{E}(X^2)$, we obtain

$$\Delta \leq c\sum_{n=1}^{\infty} \frac{\mathbb{D}^2(S^*_{k_n})}{k_n^2} = c\sum_{n=1}^{\infty} \frac{1}{k_n^2}\sum_{i=1}^{k_n} \mathbb{D}^2(\eta_i) \leq c\sum_{n=1}^{\infty} \frac{1}{k_n^2}\sum_{i=1}^{k_n} \mathbb{E}\eta_i^2 = c\sum_{i=1}^{\infty} \mathbb{E}\eta_i^2 \sum_{k_n \geq i} \frac{1}{k_n^2},$$

where at the last step we interchanged the order of summations.
(Here and in what follows c is a constant which may depend on the formula.)

By the definition of k_n, we obtain

$$\sum_{k_n \geq i} 1/k_n^2 \leq ci^{-2}(1 + \alpha^{-2} + \alpha^{-4} + \cdots) \leq ci^{-2}.$$

By the definition of η_i, using the additivity of the expectation, then interchanging summations,

$$\Delta \leq c\sum_{i=1}^{\infty} \mathbb{E}\eta_i^2 \sum_{k_n \geq i} \frac{1}{k_n^2} \leq c\sum_{i=1}^{\infty} \frac{\mathbb{E}\eta_i^2}{i^2} = c\sum_{i=1}^{\infty} \frac{\mathbb{E}(\xi_1^2 \chi_{\{\xi_1 \leq i\}})}{i^2} =$$

$$= c\sum_{i=1}^{\infty} \frac{1}{i^2}\sum_{k=0}^{i-1} \mathbb{E}(\xi_1^2 \chi_{\{k<\xi_1\leq k+1\}}) = c\sum_{k=0}^{\infty} \mathbb{E}(\xi_1^2 \chi_{\{k<\xi_1\leq k+1\}}) \sum_{i=k+1}^{\infty} \frac{1}{i^2} \leq$$

$$\leq c\sum_{k=0}^{\infty} \frac{1}{k+1}\mathbb{E}(\xi_1^2 \chi_{\{k<\xi_1\leq k+1\}}) \leq c\sum_{k=0}^{\infty} \mathbb{E}(\xi_1 \chi_{\{k<\xi_1\leq k+1\}}) = c\mathbb{E}\xi_1 < \infty.$$

In formula being last but one, we applied that $\sum_{i=k+1}^{\infty} i^{-2} \leq c(k+1)^{-1}$
which is the consequence of approximating a sum by an integral. □

References

N. Etemadi (1981); An elementary proof of the strong law of large numbers. Z. Wahrsch. Verw. Gebiete, 55(1):119–122, 1981.

Khinchin, A. (1929); "Sur la loi des grands nombres." Comptes rendus de l'Acadmie des Sciences 189, 477-479, 1929.

Kolmogorov, Andrey (1933); Grundbegriffe der Wahrscheinlichkeitsrechnung (in German). Berlin: Julius Springer

Ash, Robert B. (1972); Real analysis and probability. Academic Press, New York-London, 1972.

Bauer, Heinz (1996); Probability theory. Walter de Gruyter and Co., Berlin, 1996.

Gut, Allan (2005); Probability: a graduate course. Springer Texts in Statistics. Springer, New York, 2005.

Loève, Michel (1963); Probability theory. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1963.

Rényi, A. (1970); Probability theory. North-Holland Publishing Co., Amsterdam-London, 1970.

Shiryayev, A. N. (1984); Probability. Graduate Texts in Mathematics, 95. Springer-Verlag, New York, 1984.

3 General approach to strong laws of large numbers

Introduction

Hájek and Rényi (1955) proved the following inequality.

Let $X_1, \dots, X_n$ be independent random variables with zero mean values and finite variances $\mathbb{E}X_k^2 = \sigma_k^2$, $k = 1, \dots, n$. Denote by $S_k = X_1 + \cdots + X_k$, $k = 1, \dots, n$, the partial sums. Let $\beta_1, \dots, \beta_n$ be a non-decreasing sequence of positive numbers. Then for any $\varepsilon > 0$ and for any m with $1 \le m \le n$ we have

$$\mathbb{P}\left(\max_{m\le l\le n}\left|\frac{S_l}{\beta_l}\right| \ge \varepsilon\right) \le \frac{1}{\varepsilon^2}\left[\frac{1}{\beta_m^2}\sum_{l=1}^{m}\sigma_l^2 + \sum_{l=m+1}^{n}\frac{\sigma_l^2}{\beta_l^2}\right]. \tag{3.1}$$

In Hájek-Rényi (1955) this inequality was used to obtain strong laws of large numbers (SLLN).

In Fazekas and Klesov (2000) it was shown that a Hájek-Rényi type maximal inequality for moments is always a consequence of an appropriate Kolmogorov type maximal inequality. Moreover, the Hájek-Rényi type maximal inequality automatically implies the SLLN. The most important is that no restriction is assumed on the dependence structure of the random variables.

We list some abstract versions of the Hájek-Rényi inequality. Then we show that several Hájek-Rényi type inequalities and SLLN's can be inserted into the framework of our general theory. We shall mention results e.g. on martingales, mixingales, ϱ-mixing sequences, associated sequences, negatively associated sequences, and demimartingales.

Notation

$X_1, X_2, \dots,$ will denote a sequence of random variables. The partial sums will be $S_n = \sum_{i=1}^{n} X_i$ for $n \ge 1$ and $S_0 = 0$. A sequence $\{b_n\}$ will be called non-decreasing if $b_i \le b_{i+1}$ for $i \ge 1$.

Hájek-Rényi type maximal inequalities for moments

Definition 3.1. We say that the random variables $X_1, \dots, X_n$ satisfy the **first Kolmogorov type maximal inequality for moments**, if for each m with $1 \le m \le n$

$$\mathbb{E}\Big[\max_{1\le l\le m}|S_l|\Big]^r \le K\sum\nolimits_{l=1}^{m}\alpha_l \tag{3.2}$$

where $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $K > 0$.

Definition 3.2. We say that the random variables $X_1, \dots, X_n$ satisfy the **first Hájek-Rényi type maximal inequality for moments**, if

$$\mathbb{E}\left[\max_{1\le l\le n}\left|\frac{S_l}{\beta_l}\right|\right]^r \le C\sum_{l=1}^{n}\frac{\alpha_l}{\beta_l^r} \tag{3.3}$$

where $\beta_1, \dots, \beta_n$ is a non-decreasing sequence of positive numbers, $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $C > 0$.

Theorem 3.1. *(Fazekas-Klesov (2000), Theorem 1.1.) Let the random variables $X_1, \dots, X_n$ be fixed. If the first Kolmogorov type maximal inequality* (3.2) *for moments is satisfied, then the first Hájek-Rényi type maximal inequality* (3.3) *for moments is satisfied with $C = 4K$.*

Here and in what follows we mean that the Hájek-Rényi type maximal inequality is valid with the same parameters n, $\alpha_1, \dots, \alpha_n$, $r > 0$ as the appropriate Kolmogorov's inequality and for an arbitrary non-decreasing sequence of positive numbers $\beta_1, \dots, \beta_n$. Moreover, C may depend on K and r only.

The above result allows us to obtain an abstract form of the **SLLN**.

Theorem 3.2. *(Theorem 2.1 in Fazekas-Klesov (2000).) Let $X_1, X_2, \dots$ be a sequence of random variables. Let the non-negative numbers $\alpha_1, \alpha_2, \dots$, $r > 0$, and $K > 0$ be fixed. Assume that for each $m \geq 1$ the first Kolmogorov type maximal inequality* (3.2) *for moments is satisfied. Let $b_1, b_2, \dots$ be a non-decreasing unbounded sequence of positive numbers. If*

$$\sum\nolimits_{l=1}^{\infty} \frac{\alpha_l}{b_l^r} < \infty \tag{3.4}$$

then

$$\lim_{n\to\infty} \frac{S_n}{b_n} = 0 \quad a.s. \tag{3.5}$$

Proof. The proof in Fazekas-Klesov (2000) is based on a theorem of Dini (see Lemma 3.2 below). Actually it suffices to apply Lemma 3.1 below which is a simple consequence of Dini's theorem. Let $\{\beta_n\}$ be a sequence satisfying the properties given in Lemma 3.1. Then apply Theorem 3.1. □

Lemma 3.1. *Let $\{b_k\}$ be a non-decreasing unbounded sequence of positive numbers. Let $\{\alpha_k\}$ be a sequence of non-negative numbers, with $\sum_{k=1}^{\infty} \frac{\alpha_k}{b_k^r} < \infty$, where $r > 0$.*
Then there exists a non-decreasing unbounded sequence $\{\beta_k\}$ of positive numbers such that $\sum_{k=1}^{\infty} \frac{\alpha_k}{\beta_k^r} < \infty$ and $\lim_{k\to\infty} \frac{\beta_k}{b_k} = 0$.

In Fazekas-Klesov (1998) a direct proof of Lemma 3.1 is given. A more general Hájek-Rényi type inequality.

Definition 3.3. We say that the random variables $X_1, \dots, X_n$ satisfy the **second Kolmogorov type maximal inequality for moments**, if for each k, m with $1 \leq k < m \leq n$

$$\mathbb{E}\Big[\max_{k\leq l\leq m} |X_k + \dots + X_l|\Big]^r \leq K \sum\nolimits_{l=k}^{m} \alpha_l \tag{3.6}$$

where $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $K > 0$.

Definition 3.4. We say that the random variables $X_1, \dots, X_n$ satisfy the **second Hájek-Rényi type maximal inequality for moments**, if for each m with $1 \leq m \leq n$

$$\mathbb{E}\left[\max_{m\leq l\leq n} \left|\frac{S_l}{\beta_l}\right|\right]^r \leq C\left[\frac{1}{\beta_m^r}\sum_{l=1}^{m} \alpha_l + \sum_{l=m+1}^{n} \frac{\alpha_l}{\beta_l^r}\right] \tag{3.7}$$

where $\beta_1, \dots, \beta_n$ is a non-decreasing sequence of positive numbers, $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $C > 0$.

Theorem 3.3. *Let the random variables $X_1, \dots, X_n$ be fixed. If the second Kolmogorov type maximal inequality* (3.6) *for moments is satisfied, then the second Hájek-Rényi type maximal inequality* (3.7) *for moments is satisfied with $C = 4D_rK$, where $D_r = 1$ for $0 < r \le 1$, and $D_r = 2^{r-1}$ for $r \ge 1$.*

A popular way to obtain the SLLN is the application of the second Hájek-Rényi type maximal inequality.

Second proof of Theorem 3.2 if the second Kolmogorov type maximal inequality (3.6) *for moments is satisfied. By Theorem 3.3,*

$$\mathbb{E}\left[\sup_{k\ge m}\left|\frac{S_k}{b_k}\right|\right]^r = \lim_{n\to\infty}\mathbb{E}\left[\max_{m\le k\le n}\left|\frac{S_k}{b_k}\right|\right]^r \le C\left[\frac{1}{b_m^r}\sum_{l=1}^{m}\alpha_l + \sum_{l=m+1}^{\infty}\frac{\alpha_l}{b_l^r}\right].$$

As $m \to \infty$, the above expression converges to 0. □

Applications of the moment inequalities

Some basic inequalities

A Kolmogorov type inequality serves the starting point of each application of our main theorem.

Doob's inequalities for martingales are well-known.

Qi-Man Shao (2000) proved a comparison theorem for moment inequalities between negatively associated and independent random variables. That general theorem implies the following Kolmogorov type inequality for negatively associated random variables $X_1, \dots, X_n$.

$$\mathbb{E}\Big[\max_{1\le k\le n}\big|\sum\nolimits_{i=1}^{k} X_i\big|\Big]^p \le 2^{3-p}\sum\nolimits_{i=1}^{n}\mathbb{E}|X_i|^p \quad \text{for} \quad 1 < p \le 2. \tag{3.8}$$

It was obtained by Matuła (1992) for $p = 2$.

Applications of the moment inequalities

Proofs of SLLN's

It is worth to fix explicitly the conditions like we did in Theorem 3.2 because it helps to handle some difficult particular cases.

In Fazekas-Klesov (2000) the general SLLN was applied among others to prove Brunk-Prokhorov type SLLN for martingales, to extend Shao's (1995) Marcinkiewicz-Zygmund type SLLN for ϱ-mixing sequences, and to extend Hansen's (1991) SLLN for mixingales.

Using the general Theorem 3.2, **Kuczmaszewska (2005)** proved an SLLN for negatively associated sequences. (She obtained the Kolmogorov type maximal inequality which serves the base of the proof.) Moreover, she presented a general Marcinkiewicz-Zygmund type SLLN for ϱ-mixing sequences. (Both Kuczmaszewska (2005) and Fazekas-Klesov (2000) applied the Kolmogorov type maximal inequality for ϱ-mixing sequences given in Shao (1995).)

Rate of convergence in the SLLN

Theorem 3.4 (Shuhe-Ming). *(Lemma 1.2 in Shuhe-Ming (2006).) Let $X_1, X_2, \ldots$ be a sequence of random variables. Let the non-negative numbers $\alpha_1, \alpha_2, \ldots$, $r > 0$, and $K > 0$ be fixed. Assume that for each $m \geq 1$ the first Kolmogorov type maximal inequality* (3.2) *for moments is satisfied. Let $b_1, b_2, \ldots$ be a non-decreasing unbounded sequence of positive numbers. If* (3.4) *is satisfied, i.e. $\sum_{l=1}^{\infty} \frac{\alpha_l}{b_l^r} < \infty$, then $\lim_{n\to\infty} \frac{S_n}{b_n} = 0$ a.s., moreover*

$$\frac{S_n}{b_n} = \mathrm{O}\left(\frac{\beta_n}{b_n}\right) \quad a.s.$$

where

$$\beta_n = \max_{1\leq k\leq n} b_k \nu_k^{\delta/r}, \quad \nu_k = \sum\nolimits_{l=k}^{\infty} \frac{\alpha_l}{b_l^r},$$

δ is an arbitrary number with $0 < \delta < 1$.

We remark that $\lim_{n\to\infty} \beta_n/b_n = 0$.

Shuhe and Ming (2006) followed the approach described in F-K (2000) but they utilized the full strength of Dini's theorem.

Lemma 3.2 (Dini). *(***Dini's theorem***, see Fikhtengolts (1969), sect. 375.5.) Let $c_1, c_2, \ldots$ be non-negative numbers, $\nu_n = \sum_{k=n}^{\infty} c_k$. If $0 < \nu_n < \infty$ for all $n = 1, 2, \ldots$, then for any $0 < \delta < 1$ we have $\sum_{n=1}^{\infty} c_n/\nu_n^{\delta} < \infty$.*

Theorem 3.4 was used to obtain convergence rates in SLLN's for random variables satisfying certain dependence conditions. We list those cases where the SLLN was obtained in Fazekas-Klesov (2000) while the rate of convergence in Shuhe-Ming (2006). Sequences with superadditive moment function. The proofs are based on an inequality in Móricz (1976). Marcinkiewicz-Zygmund SLLN for sequences with superadditive moment function. SLLN using Petrov's natural characteristics of the order of growth of sums.

The simple versions of the so called almost sure limit theorems are based on an SLLN with logarithmic normalizing factors. In Fazekas-Klesov (2000), Theorem 8.1, a short proof is given for the next SLLN. Shuhe and Ming (2006) gave the rate of convergence.

Theorem 3.5 (Shuhe-Ming). *(Shuhe-Ming (2006), Theorem 2.5.) Let for some $\beta > 0$ and $C > 0$*

$$|\mathrm{cov}(X_k, X_l)| \leq C\left(\frac{l}{k}\right)^{\beta}, \qquad 1 \leq l \leq k. \tag{3.9}$$

Then

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^{n} \frac{X_k - \mathbb{E}X_k}{k} = 0 \qquad a.s. \tag{3.10}$$

Moreover, for any $0 < \delta < 1/2$

$$\frac{1}{\log n} \sum_{k=1}^{n} \frac{X_k - \mathbb{E}X_k}{k} = \mathrm{O}\left(\frac{1}{(\log n)^{\delta}}\right) \qquad a.s. \tag{3.11}$$

Hájek-Rényi type inequalities for the probability

Definition 3.5. We say that the random variables $X_1, \dots, X_n$ satisfy the **first Kolmogorov type maximal inequality for the probability**, if for each m with $1 \le m \le n$

$$\mathbb{P}\Big(\max_{1\le l\le m} |S_l| \ge \varepsilon \Big) \le \frac{K}{\varepsilon^r} \sum_{l=1}^{m} \alpha_l \quad \text{for any} \quad \varepsilon > 0 \tag{3.12}$$

where $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $K > 0$.

Definition 3.6. We say that the random variables $X_1, \dots, X_n$ satisfy the **first Hájek-Rényi type maximal inequality for the probability**, if

$$\mathbb{P}\left(\max_{1\le l\le n} \left| \frac{S_l}{\beta_l} \right| \ge \varepsilon \right) \le \frac{C}{\varepsilon^r} \sum_{l=1}^{n} \frac{\alpha_l}{\beta_l^r} \quad \text{for any} \quad \varepsilon > 0 \tag{3.13}$$

where $\beta_1, \dots, \beta_n$ is a non-decreasing sequence of positive numbers, $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $C > 0$.

Theorem 3.6. *(Tómács-Líbor (2006).) Let the r.v.'s $X_1, \dots, X_n$ be fixed. If the first Kolmogorov type maximal inequality* (3.12) *for the probability is satisfied, then the first Hájek-Rényi type maximal inequality* (3.13) *for the probability is satisfied with $C = 4K$.*

Theorem 3.7. *(Tómács-Líbor (2006).)* **SLLN** *Let $X_1, X_2, \dots$ be a sequence of random variables. Let the non-negative numbers $\alpha_1, \alpha_2, \dots$, $r > 0$, and $K > 0$ be fixed. Assume that for each $m \ge 1$ the first Kolmogorov type maximal inequality* (3.12) *for the probability is satisfied. Let $b_1, b_2, \dots$ be a non-decreasing unbounded sequence of positive numbers. If*

$$\sum\nolimits_{l=1}^{\infty} \frac{\alpha_l}{b_l^r} < \infty \tag{3.14}$$

then

$$\lim_{n\to\infty} \frac{S_n}{b_n} = 0 \quad a.s. \tag{3.15}$$

A more general Hájek-Rényi type inequality for the probability

Definition 3.7. We say that the random variables $X_1, \dots, X_n$ satisfy the **second Kolmogorov type maximal inequality for the probability** if for each k, m with $1 \le k < m \le n$

$$\mathbb{P}\Big(\max_{k\le l\le m} |X_k + \dots + X_l| \ge \varepsilon \Big) \le \frac{K}{\varepsilon^r} \sum_{l=k}^{m} \alpha_l \quad \text{for each} \quad \varepsilon > 0 \tag{3.16}$$

where $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $K > 0$.

Definition 3.8. We say that the random variables $X_1, \dots, X_n$ satisfy the **second Hájek-Rényi type maximal inequality for the probability**, if for each m with $1 \le m \le n$

$$\mathbb{P}\left(\max_{m\le l\le n} \left| \frac{S_l}{\beta_l} \right| \ge \varepsilon \right) \le \frac{C}{\varepsilon^r} \left[\frac{1}{\beta_m^r} \sum_{l=1}^{m} \alpha_l + \sum_{l=m+1}^{n} \frac{\alpha_l}{\beta_l^r} \right] \quad \text{for any} \quad \varepsilon > 0 \tag{3.17}$$

where $\beta_1, \dots, \beta_n$ is a non-decreasing sequence of positive numbers, $\alpha_1, \dots, \alpha_n$ are non-negative numbers, $r > 0$, and $C > 0$.

Theorem 3.8. *Let the random variables $X_1, \dots, X_n$ be fixed. If the second Kolmogorov type maximal inequality* (3.16) *for the probability is satisfied, then the second Hájek-Rényi type maximal inequality* (3.17) *for the probability is satisfied with $C = (1 + \sqrt[r]{4})^r K$.*

We show that by applying the second Hájek-Rényi type maximal inequality we can prove the SLLN.

Second proof of Theorem 3.7 if the second Kolmogorov type maximal inequality for the probability is satisfied. By Theorem 3.8,

$$\mathbb{P}\left(\sup_{k\geq m}\left|\frac{S_k}{\beta_k}\right|>\varepsilon\right)\leq C\left[\frac{1}{b_m^r}\sum_{l=1}^{m}\alpha_l+\sum_{l=m+1}^{\infty}\frac{\alpha_l}{b_l^r}\right].$$

As $m \to \infty$, the above expression converges to 0. □

Applications of the probability inequalities

Alternative proofs for the Hájek-Rényi inequalities and SLLN's

Most classical Hájek-Rényi type inequalities were proved by direct methods. Below we list some known results and point out how can we insert them into our general framework. However, by our general method we can reproduce the Hájek-Rényi type inequalities up to an absolute constant multiplier.

First we remark that in **Hájek-Rényi (1955)** a direct proof is given for inequality (3.1) if the random variables are independent. Now we see that the original Hájek-Rényi inequality is a consequence of the original Kolmogorov inequality (up to a constant).

Another classical Hájek-Rényi type inequality was proved by **Chow (1960)** for submartingales. We can see that it is a consequence of Doob's inequality (up to a constant).

Kounias and Weng (1969) proved the following Hájek-Rényi type inequality without assuming any dependence condition. Let $X_1, \dots, X_n$ be random variables. Let $r > 0$ be fixed. Let the moments $v_i = \mathbb{E}|X_i|^r$ be finite for each i. Let $s = 1$ if $0 < r \leq 1$ and $s = r$ if $r > 1$. Then

$$\mathbb{P}\left(\max_{1\leq l\leq n}\left|\frac{S_l}{\beta_l}\right|\geq\varepsilon\right)\leq\frac{1}{\varepsilon^r}\left(\sum_{l=1}^{n}\left(\frac{v_l}{\beta_l^r}\right)^{1/s}\right)^s \quad \text{for any} \quad \varepsilon>0 \tag{3.18}$$

where $\beta_1, \dots, \beta_n$ is an arbitrary non-decreasing sequence of positive numbers.

This inequality is of type (3.13) for $r \leq 1$. But for $r > 1$ it seems to be different of type (3.13). However, we shall see that it can be inserted into our framework for $r > 1$, too. The appropriate Kolmogorov type inequality is of the form (3.12) with $\alpha_k = \left(\sum_{i=1}^{k} a_i\right)^r - \left(\sum_{i=1}^{k-1} a_i\right)^r$, $a_i = (\mathbb{E}|X_i|^r)^{1/r}$. We can prove that $\sum_{l=1}^{n} \frac{\alpha_l}{\beta_l^r} \leq \left(\sum_{l=1}^{n} \frac{a_l}{\beta_l}\right)^r$. So in this case (3.13) implies (3.18) up to a multiplier 4.

Szynal (1973) obtained the following Hájek-Rényi type inequality without assuming independence and moment conditions. Let $X_1, \dots, X_n$ be arbitrary random variables. Let $r > 0$ be fixed. Let $s = 1$ if $0 < r \leq 1$ and $s = r$ if $r > 1$. Then

$$\mathbb{P}\left[\max_{m\leq l\leq n}\left|\frac{S_l}{\beta_l}\right|\geq 3\varepsilon\right]\leq 2\left[\sum_{l=1}^{m}\mathbb{E}^{1/s}\Big(\frac{|X_l|^r}{(\beta_m\varepsilon)^r+|X_l|^r}\Big)+\sum_{l=m+1}^{n}\mathbb{E}^{1/s}\Big(\frac{|X_l|^r}{(\beta_l\varepsilon)^r+|X_l|^r}\Big)\right]^s \tag{3.19}$$

for any $\varepsilon > 0$ where $\beta_1, \dots, \beta_n$ is an arbitrary non-decreasing sequence of positive numbers. It seems that (3.19) can not be inserted into the framework of our theorem.

Bickel (1970) obtained a Hájek-Rényi type generalization of Lévy's inequality. Chandra and Ghosal (1996) obtained a Kolmogorov type inequality and a Marcinkiewicz-Zygmund type SLLN for **asymptotically almost negatively associated (AANA)** random variables.

The sequence $X_1, X_2, \dots$ is called AANA if there exists a non-negative sequence $q_n \to 0$ such that

$$\operatorname{cov}\left(f(X_m), g(X_{m+1}, \dots, X_{m+k})\right) \le q_m \left(\operatorname{var}\big(f(X_m)\big)\operatorname{var}\big(g(X_{m+1}, \dots, X_{m+k})\big)\right)^{1/2} \quad (3.20)$$

for all $m, k \ge 1$ and for all coordinatewise increasing continuous functions f and g whenever the right hand side of the above inequality is finite.

The negatively associated (in particular the independent) sequences are AANA. The Kolmogorov type inequality for AANA sequences is the following.

Theorem 3.9. *(Chandra-Ghosal (1996), Theorem 1.) Let $X_1, \dots, X_n$ be zero mean square integrable r.v.'s such that* (3.20) *holds for every $1 \le m < m + k \le n$. Let $A_n = \sum_{l=1}^{n-1} q_l^2$. Then*

$$\mathbb{P}\Big(\max_{1\le l\le n} |S_l| \ge \varepsilon\Big) \le \frac{2}{\varepsilon^2}\left(A_n + \sqrt{1 + A_n^2}\right)^2 \sum_{l=1}^{n} \mathbb{E}X_l^2 \quad \textit{for any} \quad \varepsilon > 0. \quad (3.21)$$

Using (3.21), Kim, Ko and Lee (2004) obtained Hájek-Rényi inequalities of the form (3.13) and (3.17), moreover, if $A = \sum_{l=1}^{\infty} q_l^2 < \infty$, an SLLN of the form Theorem 3.7. Now we can see that these results are immediate consequences of our general theorems.

Rate of convergence in the SLLN's

Combining the method of Theorem 3.7 and the ideas of Shuhe and Ming (2006) we can obtain further results concerning the rate of convergence in the SLLN.

Theorem 3.10. *Let $X_1, X_2, \dots$ be a sequence of random variables. Let the non-negative numbers $\alpha_1, \alpha_2, \dots$, $r > 0$, and $K > 0$ be fixed. Assume that for each $m \ge 1$ the first Kolmogorov type maximal inequality* (3.12) *for the probability is satisfied. Let $b_1, b_2, \dots$ be a non-decreasing unbounded sequence of positive numbers. If* (3.14) *is satisfied, i.e. $\sum_{l=1}^{\infty} \frac{\alpha_l}{b_l^r} < \infty$, then $\lim_{n\to\infty} \frac{S_n}{b_n} = 0$ a.s., moreover*

$$\frac{S_n}{b_n} = \mathrm{O}\left(\frac{\beta_n}{b_n}\right) \quad a.s.$$

where β_n is defined in Theorem 3.4.

This theorem can be used to obtain convergence rates in the SLLN for random variables satisfying certain dependence conditions. Using that inequality, an SLLN can be obtained for **demimartingales** (Christofides (2000)). Now we can see that our general theorems immediately imply the SLLN from the Kolmogorov type inequality, moreover the rate of convergence in the SLLN. The sequence of partial sums of zero mean **associated** random variables is a demimartingale. We mention that for associated random variables

Kolmogorov type inequalities were obtained by Newman and Wright (1981) and by Matuła (1996). From those inequalities Matuła (1996) and Shuhe-Ming (2006) obtained the SLLN and the rate of convergence in the SLLN, respectively. Now we se that those theorems are covered by our general method.

We see that for **AANA** sequences, using the Kolmogorov type inequality of Chandra-Ghosal (1996), our method gives the SLLN and the rate of convergence in it. We mention that negatively associated sequences are AANA. In the special case of negatively associated sequences, the Kolmogorov inequality was obtained by Matuła (1992), the Hájek-Rényi type inequality and the SLLN by Liu, Gan and Chen (1999), while the rate of convergence in the SLLN by Shuhe and Ming (2006). Those results are covered by our method.

References

D. W. K. Andrews, (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, **4**, 458–467.

Bickel, P. J., (1970). A Hájek-Rényi extension of Lévy's inequality and some applications. *Acta Math. Acad. Sci. Hungar.* **21**, 199–206.

H. D. Brunk, (1948). The strong law of large numbers. *Duke Math. J.* **15**, 181–195.

D. L. Burkholder, (1773). Distribution function inequalities for martingales. *Ann. Probab.* **1**(1), 19–42.

Chandra, T. K. and Ghosal, S., (1996). Extensions of the strong law of large numbers of Marcinkiewicz and Zygmund for dependent variables. *Acta Math. Hungar.* **71**(4), 327–336.

Chow, Y. S., (1960). A martingale inequality and the law of large numbers. *Proc. Amer. Math. Soc.* **11**, 107–111.

Christofides, T. C., (2000). Maximal inequalities for demimartingales and a strong law of large numbers. *Statist. Probab. Lett.* **50**(4), 357–363.

Csörgő, M. (1968). On the strong law of large numbers and the central limit theorem for martingales. *Trans. Amer. Math. Soc.* **131**, 259–275.

V. A. Egorov, (1970). On the strong law of large numbers and the law of the iterated logarithm for a sequence of independent random variables. *Probab. Theory Appl.* **15**(3), 520–527.

Fazekas, I., Klesov, O. I., (1998). A general approach to the strong laws of large numbers. *Technical Report*, No. **4**, (1998), Kossuth Lajos University, Hungary.

Fazekas, I., Klesov, O. I., (2000). A general approach to the strong laws of large numbers. *Teor. Veroyatnost. i Primenen.* **45**(3), 568–583; *Theory Probab. Appl.*, **45**(3) (2002), 436–449.

Fikhtengolts, G. M., (1969). *A Course of Differential and Integral Calculus.* Vol. 2. (Russian) Nauka, Moscow.

J. Hájek and A. Rényi, (1955). Generalization of an inequality of Kolmogorov. *Acta Math. Acad. Sci. Hungar.* **6**(3-4), 281–283.

P. Hall and C. C. Heyde, (1980). *Martingale Limit Theory and its Application.* Academic Press, New York.

B. E. Hansen, (1991). Strong laws for dependent heterogeneous processes. *Econometric Theory*, **7**, 213–221. Erratum. *Econometric Theory*, **8** (1982), 421–422.

Kim, Tae-Sung; Ko, Mi-Hwa; Lee, Il-Hyun; (2004). On the strong law for asymptotically almost negatively associated random variables. *Rocky Mountain J. Math.* **34**(3), 979–989.

Kounias, E. G. and Weng, T-S, (1969). An inequality and almost sure convergence. *Ann. Math. Statist.* **40**(3), 1091–1093.

Kuczmaszewska, A., (2005). The strong law of large numbers for dependent random variables. *Statist. Probab. Lett.* **73**(3), 305–314.

Liu, Jingjun; Gan, Shixin; and Chen, Pingyan; (1999). The Hájeck-Rényi inequality for the NA random variables and its application. *Statist. Probab. Lett.* **43**(1), 99–105.

M. Longnecker and R. J. Serfling, (1977). General moment and probability inequalities for the maximum partial sum. *Acta Math. Sci. Hungar.* **30**(1-2), 129–133.

Matuła, P., (1992). A note on the almost sure convergence of sums of negatively dependent random variables. *Statist. Probab. Lett.* **15**(3), 209–213.

Matuła, P., (1996). Convergence of weighted averages of associated random variables. *Probab. Math. Statist.* **16**(2), 337–343.

D. L. McLeish, (1975). A maximal inequality and dependent strong laws. *Ann. Probab.* **3**(5), 829–839.

T. F. Móri, (1993). On the strong law of large numbers for logarithmically weighted sums. *Annales Univ. Sci. Budapest*, **36**, 35–46.

F. Móricz, (1976). Moment inequalities and the strong laws of large numbers. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **35**, 299–314.

Newman, C. M. and Wright, A. L., (1982). Associated random variables and martingale inequalities. *Z. Wahrsch. Verw. Gebiete*, **59**(3), 361–371.

V. V. Petrov, (1975). *Sums of Independent Random Variables*. Springer–Verlag, New York.

Yu. V. Prokhorov, (1950). On the strong law of large numbers. (Russian.) *Izv. AN SSSR, ser. matem.* **14**(6), 523–536.

Qi-Man Shao, (1995). Maximal inequalities for partial sums of ϱ-mixing sequences. *Ann. Probab.* **23**(2), 948–965.

Qi-Man Shao, (2000). A comparison theorem on moment inequalities between negatively associated and independent random variables. *J. Theoret. Probab.* **13**(2), 343–356.

Hu Shuhe, Hu Ming, (2006). A general approach rate to the strong law of large numbers. *Statist. Probab. Lett.* **76**(8), 843–851.

Szynal, Dominik, (1973). An extension of the Hájek-Rényi inequality for one maximum of partial sums. *Ann. Statist.* **1**, 740–744.

H. Teicher, (1968). Some new conditions for the strong law. *Proc. Nat. Acad. Sci. U.S.A.* **59**(3), 705–707.

Tómács, T. and Líbor, Zs., (2006). A Hájek-Rényi type inequality and its applications. *Ann. Math. Inf.* **33**, 141–149.

Sung, Soo Hak; Hu, Tien-Chung; Volodin, Andrei (2008). A note on the growth rate in the Fazekas-Klesov general law of large numbers and on the weak law of large numbers for tail series. *Publ. Math. Debrecen* **73**(1-2), 110.