The role of machine learning in mathematical statistics

László (Laci) Györfi[1]

[1]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary

November 27, 2013

e-mail: gyorfi@cs.bme.hu
www.szit.bme.hu/˜gyorfi

Happy birthday

First visit to Leuven: 1987

Happy birthday

First visit to Leuven: 1987

Edward van der Meulen, 19 papers

Happy birthday

First visit to Leuven: 1987
Edward van der Meulen, 19 papers
Jan Beirlant, 7 papers

Happy birthday

First visit to Leuven: 1987
Edward van der Meulen, 19 papers
Jan Beirlant, 7 papers
An Carbonez, 4 papers

Happy birthday

First visit to Leuven: 1987
Edward van der Meulen, 19 papers
Jan Beirlant, 7 papers
An Carbonez, 4 papers
Luc Devroye, 2 books, 11 papers

Happy birthday

First visit to Leuven: 1987
Edward van der Meulen, 19 papers
Jan Beirlant, 7 papers
An Carbonez, 4 papers
Luc Devroye, 2 books, 11 papers
Paul Algoet, 1 paper

Happy birthday

First visit to Leuven: 1987
Edward van der Meulen, 19 papers
Jan Beirlant, 7 papers
An Carbonez, 4 papers
Luc Devroye, 2 books, 11 papers
Paul Algoet, 1 paper
Kris de Brabanter, 1 paper

Portfolio selections

- static portfolio selection (single period)

Portfolio selections

- static portfolio selection (single period)
- constantly rebalanced portfolio

Portfolio selections

- static portfolio selection (single period)
- constantly rebalanced portfolio
- general rebalancing

Portfolio selections

- static portfolio selection (single period)
- constantly rebalanced portfolio
- general rebalancing

- rebalancing

Portfolio selections

- static portfolio selection (single period)
- constantly rebalanced portfolio
- general rebalancing

- rebalancing
- multi-asset, multi-period

Portfolio selections

- static portfolio selection (single period)
- constantly rebalanced portfolio
- general rebalancing

- rebalancing
- multi-asset, multi-period
- empirical (nonparametric statistics, machine learning)

Growth rate

investment in the stock market

d assets

Growth rate

investment in the stock market

d assets

$S_n^{(j)}$ price of asset j at the end of trading period (day) n,

$j = 1, \ldots, d$

Growth rate

investment in the stock market

d assets

$S_n^{(j)}$ price of asset j at the end of trading period (day) n,

$j = 1, \ldots, d$

$S_0^{(j)} = 1$

Growth rate

investment in the stock market

d assets

$S_n^{(j)}$ price of asset j at the end of trading period (day) n,

$j = 1, \ldots, d$

$S_0^{(j)} = 1$

$$S_n^{(j)} = e^{nW_n^{(j)}} \approx e^{nW^{(j)}}$$

Growth rate

investment in the stock market

d assets

$S_n^{(j)}$ price of asset j at the end of trading period (day) n,

$j = 1, \ldots, d$

$S_0^{(j)} = 1$

$$S_n^{(j)} = e^{nW_n^{(j)}} \approx e^{nW^{(j)}}$$

asymptotic growth rate

$$W^{(j)} = \lim_{n\to\infty} W_n^{(j)} = \lim_{n\to\infty} \frac{1}{n} \ln S_n^{(j)}$$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$
static portfolio selection

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$

static portfolio selection

a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$, $b^{(j)} \geq 0$, $\sum_j b^{(j)} = 1$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$

static portfolio selection

a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$, $b^{(j)} \geq 0$, $\sum_j b^{(j)} = 1$

initial capital $S_0 = 1$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$
static portfolio selection
a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$, $b^{(j)} \geq 0$, $\sum_j b^{(j)} = 1$
initial capital $S_0 = 1$

$$S_n = \sum_j b^{(j)} S_n^{(j)}$$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$
static portfolio selection
a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$, $b^{(j)} \geq 0$, $\sum_j b^{(j)} = 1$
initial capital $S_0 = 1$

$$S_n = \sum_j b^{(j)} S_n^{(j)}$$

$$\max_j b^{(j)} S_n^{(j)} \leq S_n \leq d \max_j b^{(j)} S_n^{(j)}$$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$
static portfolio selection
a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$, $b^{(j)} \geq 0$, $\sum_j b^{(j)} = 1$
initial capital $S_0 = 1$

$$S_n = \sum_j b^{(j)} S_n^{(j)}$$

$$\max_j b^{(j)} S_n^{(j)} \leq S_n \leq d \max_j b^{(j)} S_n^{(j)}$$

If $b^{(j)} > 0$ then

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n = \lim_{n\to\infty} \max_j \frac{1}{n} \ln S_n^{(j)} = \max_j W^{(j)}$$

Static portfolio selection: single period investment

the aim is to achieve $\max_j W^{(j)}$
static portfolio selection
a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$, $b^{(j)} \geq 0$, $\sum_j b^{(j)} = 1$
initial capital $S_0 = 1$

$$S_n = \sum_j b^{(j)} S_n^{(j)}$$

$$\max_j b^{(j)} S_n^{(j)} \leq S_n \leq d \max_j b^{(j)} S_n^{(j)}$$

If $b^{(j)} > 0$ then

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n = \lim_{n\to\infty} \max_j \frac{1}{n} \ln S_n^{(j)} = \max_j W^{(j)}$$

we can do much better, applying dynamic portfolio selection

Relative prices

$$x_i^{(j)} = \frac{S_i^{(j)}}{S_{i-1}^{(j)}}$$

Relative prices

$$x_i^{(j)} = \frac{S_i^{(j)}}{S_{i-1}^{(j)}}$$

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i

Relative prices

$$x_i^{(j)} = \frac{S_i^{(j)}}{S_{i-1}^{(j)}}$$

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i

$x_i^{(j)}$ is the factor by which capital invested in stock j grows during the market period i

Relative prices

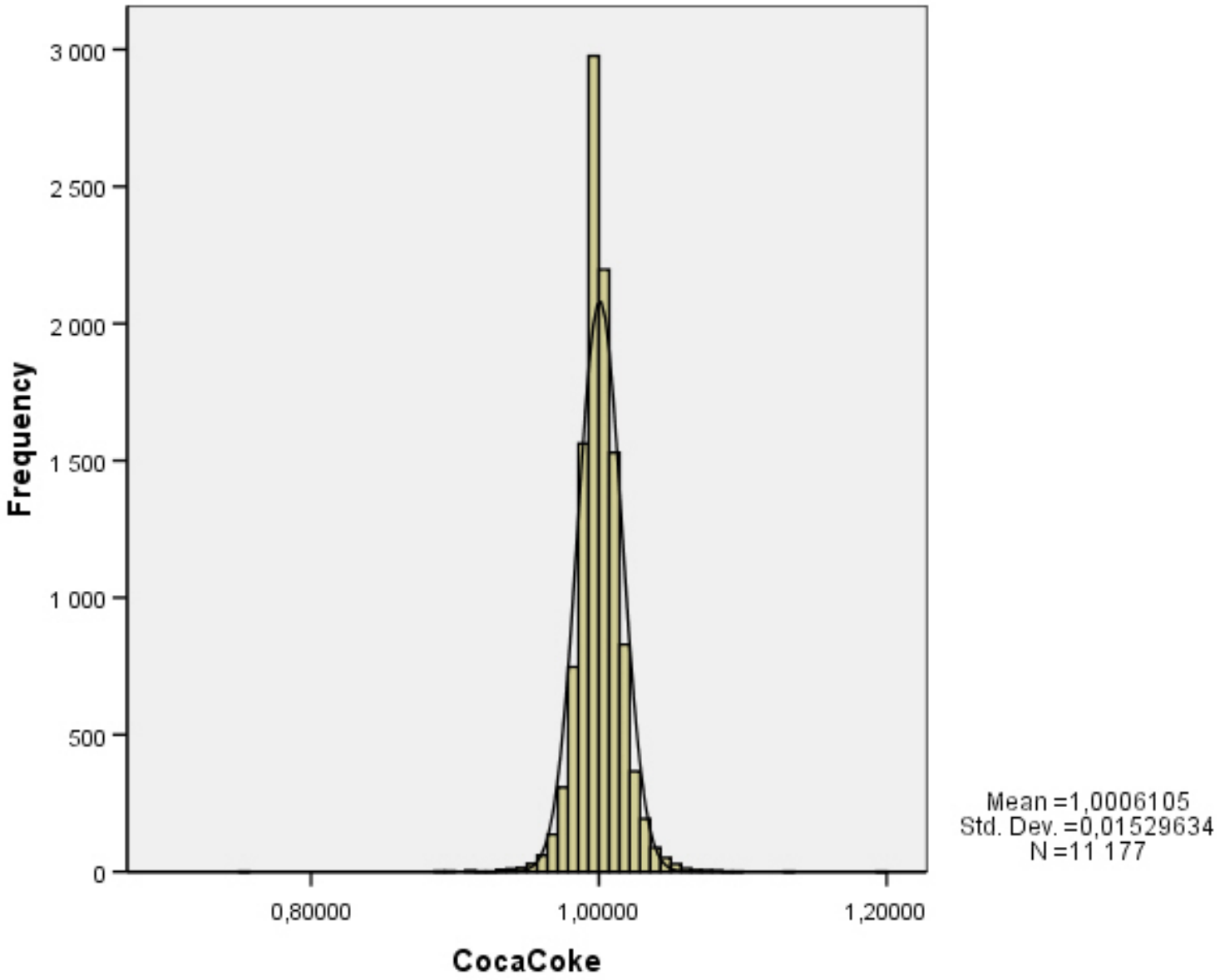
$$x_i^{(j)} = \frac{S_i^{(j)}}{S_{i-1}^{(j)}}$$

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i

$x_i^{(j)}$ is the factor by which capital invested in stock j grows during the market period i

For NYSE data

$$0.7 \leq x_i^{(j)} \leq 1.2$$


3 000
2 500
2 000
1 500
1 000
500
0
Frequency
0,80000
1,00000
1,20000
CocaCoke
Mean =1,0006105
Std. Dev. =0,01529634
N =11 177

Dynamic portfolio selection: multi-period investment

multi-period investment

Dynamic portfolio selection: multi-period investment

multi-period investment
Constantly Re-balanced Portfolio (CRP)

a portfolio vector $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$

Dynamic portfolio selection: multi-period investment

multi-period investment
Constantly Re-balanced Portfolio (CRP)

a portfolio vector $\mathbf{b} = (b^{(1)}, \ldots b^{(d)})$
$b^{(j)}$ gives the proportion of the investor's capital invested in stock j

Dynamic portfolio selection: multi-period investment

multi-period investment
Constantly Re-balanced Portfolio (CRP)

a portfolio vector $\mathbf{b} = (b^{(1)}, \ldots b^{(d)})$
$b^{(j)}$ gives the proportion of the investor's capital invested in stock j
$\mathbf{b}$ is the constant portfolio vector for each trading day

Evolution

for the first day S_0 denotes the initial capital

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle$$

Evolution

for the first day S_0 denotes the initial capital

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 \langle \mathbf{b}\,, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital

$$S_2 = S_1 \cdot \langle \mathbf{b}\,, \mathbf{x}_2 \rangle = S_0 \cdot \langle \mathbf{b}\,, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}\,, \mathbf{x}_2 \rangle\,.$$

Evolution

for the first day S_0 denotes the initial capital

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital

$$S_2 = S_1 \cdot \langle \mathbf{b}\,,\, \mathbf{x}_2 \rangle = S_0 \cdot \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}\,,\, \mathbf{x}_2 \rangle\,.$$

for the nth day:

$$S_n = S_{n-1} \langle \mathbf{b}\,,\, \mathbf{x}_n \rangle = S_0 \prod_{i=1}^{n} \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle$$

Evolution

for the first day S_0 denotes the initial capital

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital

$$S_2 = S_1 \cdot \langle \mathbf{b}\,,\, \mathbf{x}_2 \rangle = S_0 \cdot \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}\,,\, \mathbf{x}_2 \rangle\,.$$

for the nth day:

$$S_n = S_{n-1} \langle \mathbf{b}\,,\, \mathbf{x}_n \rangle = S_0 \prod_{i=1}^{n} \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle = S_0 e^{n W_n(\mathbf{b})}$$

Evolution

for the first day S_0 denotes the initial capital

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 \langle \mathbf{b}\,, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital

$$S_2 = S_1 \cdot \langle \mathbf{b}\,, \mathbf{x}_2 \rangle = S_0 \cdot \langle \mathbf{b}\,, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}\,, \mathbf{x}_2 \rangle\,.$$

for the nth day:

$$S_n = S_{n-1} \langle \mathbf{b}\,, \mathbf{x}_n \rangle = S_0 \prod_{i=1}^{n} \langle \mathbf{b}\,, \mathbf{x}_i \rangle = S_0 e^{n W_n(\mathbf{b})}$$

with the average growth rate

$$W_n(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ln \langle \mathbf{b}\,, \mathbf{x}_i \rangle\,.$$

log-optimum portfolio

Special market process: $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is independent and identically distributed (i.i.d.)

log-optimum portfolio

Special market process: $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is independent and identically distributed (i.i.d.)

log-optimum portfolio $\mathbf{b}^*$

$$\mathbf{E}\{\ln \langle \mathbf{b}^* , \mathbf{X}_1 \rangle\} = \max_{\mathbf{b}} \mathbf{E}\{\ln \langle \mathbf{b} , \mathbf{X}_1 \rangle\}$$

log-optimum portfolio

Special market process: $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is independent and identically distributed (i.i.d.)

log-optimum portfolio $\mathbf{b}^*$

$$\mathbf{E}\{\ln \langle \mathbf{b}^* , \mathbf{X}_1 \rangle\} = \max_{\mathbf{b}} \mathbf{E}\{\ln \langle \mathbf{b} , \mathbf{X}_1 \rangle\}$$

Best Constantly Re-balanced Portfolio (BCRP)

log-optimum portfolio

Special market process: $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is independent and identically distributed (i.i.d.)

log-optimum portfolio $\mathbf{b}^*$

$$\mathbf{E}\{\ln \langle \mathbf{b}^* , \mathbf{X}_1 \rangle\} = \max_{\mathbf{b}} \mathbf{E}\{\ln \langle \mathbf{b} , \mathbf{X}_1 \rangle\}$$

Best Constantly Re-balanced Portfolio (BCRP)

for dependent market process we can do even better

Dynamic portfolio selection: general case

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i

Dynamic portfolio selection: general case

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i

$\mathbf{b} = \mathbf{b}_1$ is the portfolio vector for the first day

Dynamic portfolio selection: general case

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i
$\mathbf{b} = \mathbf{b}_1$ is the portfolio vector for the first day
initial capital S_0

$$S_1 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle$$

Dynamic portfolio selection: general case

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i
$\mathbf{b} = \mathbf{b}_1$ is the portfolio vector for the first day
initial capital S_0

$$S_1 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital, the portfolio vector $\mathbf{b}_2 = \mathbf{b}(\mathbf{x}_1)$

$$S_2 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}(\mathbf{x}_1) \, , \, \mathbf{x}_2 \rangle \, .$$

Dynamic portfolio selection: general case

$\mathbf{x}_i = (x_i^{(1)}, \ldots x_i^{(d)})$ the return vector on day i
$\mathbf{b} = \mathbf{b}_1$ is the portfolio vector for the first day
initial capital S_0

$$S_1 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital, the portfolio vector $\mathbf{b}_2 = \mathbf{b}(\mathbf{x}_1)$

$$S_2 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}(\mathbf{x}_1) \, , \, \mathbf{x}_2 \rangle \, .$$

nth day a portfolio strategy $\mathbf{b}_n = \mathbf{b}(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) = \mathbf{b}(\mathbf{x}_1^{n-1})$

$$S_n = S_0 \prod_{i=1}^{n} \left\langle \mathbf{b}(\mathbf{x}_1^{i-1}) \, , \, \mathbf{x}_i \right\rangle$$

Dynamic portfolio selection: general case

$\mathbf{x}_i = (x_i^{(1)}, \dots x_i^{(d)})$ the return vector on day i
$\mathbf{b} = \mathbf{b}_1$ is the portfolio vector for the first day
initial capital S_0

$$S_1 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle$$

for the second day, S_1 new initial capital, the portfolio vector $\mathbf{b}_2 = \mathbf{b}(\mathbf{x}_1)$

$$S_2 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}(\mathbf{x}_1) \, , \, \mathbf{x}_2 \rangle \, .$$

nth day a portfolio strategy $\mathbf{b}_n = \mathbf{b}(\mathbf{x}_1, \dots, \mathbf{x}_{n-1}) = \mathbf{b}(\mathbf{x}_1^{n-1})$

$$S_n = S_0 \prod_{i=1}^{n} \left\langle \mathbf{b}(\mathbf{x}_1^{i-1}) \, , \, \mathbf{x}_i \right\rangle = S_0 e^{n W_n(\mathbf{B})}$$

with the average growth rate

$$W_n(\mathbf{B}) = \frac{1}{n} \sum_{i=1}^{n} \ln \left\langle \mathbf{b}(\mathbf{x}_1^{i-1}) \, , \, \mathbf{x}_i \right\rangle .$$

log-optimum portfolio

$\mathbf{X}_1, \mathbf{X}_2, \ldots$ drawn from the vector valued stationary and ergodic process

log-optimum portfolio

$\mathbf{X}_1, \mathbf{X}_2, \ldots$ drawn from the vector valued stationary and ergodic process

log-optimum portfolio $\mathbf{B}^* = \{\mathbf{b}^*(\cdot)\}$

$$\mathbf{E}\{\ln \left\langle \mathbf{b}^*(\mathbf{X}_1^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_1^{n-1}\} = \max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_1^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_1^{n-1}\}$$

Optimality

Algoet and Cover (1988): If $S_n^* = S_n(\mathbf{B}^*)$ denotes the capital after day n achieved by a log-optimum portfolio strategy $\mathbf{B}^*$, then for any portfolio strategy $\mathbf{B}$ with capital $S_n = S_n(\mathbf{B})$ and for any stationary ergodic process $\{\mathbf{X}_n\}_{-\infty}^{\infty}$,

$$\limsup_{n\to\infty}\left(\frac{1}{n}\ln S_n - \frac{1}{n}\ln S_n^*\right) \le 0 \quad \text{almost surely}$$

Optimality

Algoet and Cover (1988): If $S_n^* = S_n(\mathbf{B}^*)$ denotes the capital after day n achieved by a log-optimum portfolio strategy $\mathbf{B}^*$, then for any portfolio strategy $\mathbf{B}$ with capital $S_n = S_n(\mathbf{B})$ and for any stationary ergodic process $\{\mathbf{X}_n\}_{-\infty}^{\infty}$,

$$\limsup_{n\to\infty}\left(\frac{1}{n}\ln S_n - \frac{1}{n}\ln S_n^*\right) \le 0 \quad \text{almost surely}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\ln S_n^* = W^* \quad \text{almost surely,}$$

where

$$W^* = \mathbf{E}\left\{\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_{-\infty}^{-1}), \mathbf{X}_0\right\rangle \mid \mathbf{X}_{-\infty}^{-1}\}\right\}$$

is the maximal growth rate of any portfolio.

Universally consistent portfolio

These limit relations give rise to the following definition:

Definition

A portfolio strategy $\mathbf{B}$ is called **universally consistent with respect to a class $\mathcal{C}$ of stationary and ergodic processes $\{\mathbf{X}_n\}_{-\infty}^{\infty}$,** if for each process in the class,

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n(\mathbf{B}) = W^* \quad \text{almost surely.}$$

Empirical portfolio selection

$$\mathbf{E}\{\ln \left\langle \mathbf{b}^*(\mathbf{X}_1^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_1^{n-1}\} = \max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_1^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_1^{n-1}\}$$

Empirical portfolio selection

$$\mathbf{E}\{\ln\left\langle \mathbf{b}^*(\mathbf{X}_1^{n-1}), \mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\} = \max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_1^{n-1}), \mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\}$$

fixed integer $k > 0$

$$\mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_1^{n-1}), \mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\} \approx \mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}), \mathbf{X}_n\right\rangle \mid \mathbf{X}_{n-k}^{n-1}\}$$

Empirical portfolio selection

$$\mathbf{E}\{\ln\left\langle \mathbf{b}^*(\mathbf{X}_1^{n-1}),\mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\} = \max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_1^{n-1}),\mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\}$$

fixed integer $k > 0$

$$\mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_1^{n-1}),\mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\} \approx \mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}),\mathbf{X}_n\right\rangle \mid \mathbf{X}_{n-k}^{n-1}\}$$

and

$$\mathbf{b}^*(\mathbf{X}_1^{n-1}) \approx \mathbf{b}_k(\mathbf{X}_{n-k}^{n-1}) = \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}),\mathbf{X}_n\right\rangle \mid \mathbf{X}_{n-k}^{n-1}\}$$

because of stationarity

$$\mathbf{b}_k(\mathbf{x}_1^k) \quad = \quad \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\}$$

because of stationarity

$$\begin{aligned}
\mathbf{b}_k(\mathbf{x}_1^k) &= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\}
\end{aligned}$$

because of stationarity

$$
\begin{aligned}
\mathbf{b}_k(\mathbf{x}_1^k) &= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_{k+1} \right\rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\}
\end{aligned}
$$

because of stationarity

$$
\begin{aligned}
\mathbf{b}_k(\mathbf{x}_1^k) &= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_{k+1} \right\rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}} \mathbf{E}\{\ln \langle \mathbf{b}, \mathbf{X}_{k+1} \rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\},
\end{aligned}
$$

because of stationarity

$$
\begin{aligned}
\mathbf{b}_k(\mathbf{x}_1^k) &= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1}), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_n \right\rangle \mid \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}(\cdot)} \mathbf{E}\{\ln \left\langle \mathbf{b}(\mathbf{x}_1^k), \mathbf{X}_{k+1} \right\rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\} \\
&= \arg\max_{\mathbf{b}} \mathbf{E}\{\ln \langle \mathbf{b}, \mathbf{X}_{k+1} \rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\},
\end{aligned}
$$

which is the maximization of the regression function

$$
m_{\mathbf{b}}(\mathbf{x}_1^k) = \mathbf{E}\{\ln \langle \mathbf{b}, \mathbf{X}_{k+1} \rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\}
$$

Regression function

Y real valued

X observation vector

Regression function

Y real valued

X observation vector

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

i.i.d. data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function

Y real valued

X observation vector

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

i.i.d. data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Regression function

Y real valued

X observation vector

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

i.i.d. data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Kernel regression estimate with window kernel

Regression function

Y real valued

X observation vector

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

i.i.d. data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Kernel regression estimate with window kernel

Bandwidth $r > 0$

Regression function

Y real valued

X observation vector

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

i.i.d. data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Kernel regression estimate with window kernel

Bandwidth $r > 0$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{\|x - X_i\| \le r\}}}{\sum_{i=1}^n I_{\{\|x - X_i\| \le r\}}}$$

Regression function

Y real valued

X observation vector

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

i.i.d. data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Kernel regression estimate with window kernel

Bandwidth $r > 0$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{\|x - X_i\| \le r\}}}{\sum_{i=1}^n I_{\{\|x - X_i\| \le r\}}}$$

L. Györfi, M. Kohler, A. Krzyzak, H. Walk (2002) *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York.

Correspondence

$$X \quad \sim \quad \mathbf{X}_1^k$$

Correspondence

$$
\begin{aligned}
X &\sim \mathbf{X}_1^k \\
Y &\sim \ln \langle \mathbf{b}\,, \mathbf{X}_{k+1} \rangle
\end{aligned}
$$

Correspondence

$$
\begin{aligned}
X &\sim \mathbf{X}_1^k \\
Y &\sim \ln \langle \mathbf{b}\,, \mathbf{X}_{k+1} \rangle \\
m(x) = \mathbf{E}\{Y \mid X = x\} &\sim m_{\mathbf{b}}(\mathbf{x}_1^k) = \mathbf{E}\{\ln \langle \mathbf{b}\,, \mathbf{X}_{k+1} \rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\}
\end{aligned}
$$

choose the radius $r_{k,\ell} > 0$ such that for any fixed k,

$$\lim_{\ell \to \infty} r_{k,\ell} = 0.$$

choose the radius $r_{k,\ell} > 0$ such that for any fixed k,

$$\lim_{\ell \to \infty} r_{k,\ell} = 0.$$

for $n > k + 1$, define the expert $\mathbf{b}^{(k,\ell)}$ by

$$\mathbf{b}^{(k,\ell)}(\mathbf{x}_1^{n-1}) = \arg\max_{\mathbf{b}} \sum_{i \in J_n^{(k,\ell)}} \ln \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle\,,$$

Kernel-based portfolio selection

choose the radius $r_{k,\ell} > 0$ such that for any fixed k,

$$\lim_{\ell \to \infty} r_{k,\ell} = 0.$$

for $n > k+1$, define the expert $\mathbf{b}^{(k,\ell)}$ by

$$\mathbf{b}^{(k,\ell)}(\mathbf{x}_1^{n-1}) = \arg\max_{\mathbf{b}} \sum_{i \in J_n^{(k,\ell)}} \ln \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle\,,$$

where

$$J_n^{(k,\ell)} = \left\{ k < i < n : \|\mathbf{x}_{i-k}^{i-1} - \mathbf{x}_{n-k}^{n-1}\| \le r_{k,\ell} \right\}$$

if the sum is non-void, and $\mathbf{b}_0 = (1/d, \ldots, 1/d)$ otherwise.

Combining elementary portfolios

for fixed $k, \ell = 1, 2, \ldots$,
$\mathbf{B}^{(k,\ell)} = \{\mathbf{b}^{(k,\ell)}(\cdot)\}$, are called elementary portfolios

Combining elementary portfolios

for fixed $k, \ell = 1, 2, \ldots$,
$\mathbf{B}^{(k,\ell)} = \{\mathbf{b}^{(k,\ell)}(\cdot)\}$, are called elementary portfolios

How to choose k, ℓ

- small k or large $r_{k,\ell}$: large bias
- large k and small $r_{k,\ell}$: few matching, large variance

Combining elementary portfolios

for fixed $k, \ell = 1, 2, \ldots$,
$\mathbf{B}^{(k,\ell)} = \{\mathbf{b}^{(k,\ell)}(\cdot)\}$, are called elementary portfolios

How to choose k, ℓ

- small k or large $r_{k,\ell}$: large bias
- large k and small $r_{k,\ell}$: few matching, large variance

Machine learning: combination of experts

N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Aggregation

combine the elementary portfolio strategies $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}_n^{(k,\ell)}\}$

Aggregation

combine the elementary portfolio strategies $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}_n^{(k,\ell)}\}$
let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k, ℓ) such that for all k, ℓ, $q_{k,\ell} > 0$.

Aggregation

combine the elementary portfolio strategies $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}_n^{(k,\ell)}\}$
let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k,ℓ) such that for all k,ℓ, $q_{k,\ell} > 0$.
put

$$w_{n,k,\ell} = q_{k,\ell} S_{n-1}(\mathbf{B}^{(k,\ell)})$$

Aggregation

combine the elementary portfolio strategies $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}_n^{(k,\ell)}\}$
let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k,ℓ) such that for all k,ℓ, $q_{k,\ell} > 0$.
put

$$w_{n,k,\ell} = q_{k,\ell} S_{n-1}(\mathbf{B}^{(k,\ell)})$$

the aggregated portfolio $\mathbf{b}$:

$$\mathbf{b}_n(\mathbf{x}_1^{n-1}) = \frac{\sum_{k,\ell} w_{n,k,\ell} \mathbf{b}_n^{(k,\ell)}(\mathbf{x}_1^{n-1})}{\sum_{k,\ell} w_{n,k,\ell}}.$$

Theorem

The kernel-based portfolio scheme is universally consistent with respect to the class of all ergodic processes such that $\mathbf{E}\{|\ln X^{(j)}|\} < \infty$, for $j = 1, 2, \ldots, d$.

L. Györfi, G. Lugosi, F. Udina (2006) "Nonparametric kernel-based sequential investment strategies", *Mathematical Finance*, 16, pp. 337-357
www.szit.bme.hu/˜gyorfi/kernel.pdf

Conditions of the model:

Assume that

- the assets are arbitrarily divisible,
- the assets are available in unbounded quantities at the current price at any given trading period,
- there are no transaction costs,
- the behavior of the market is not affected by the actions of the investor using the strategy under investigation.

NYSE data sets

At `www.szit.bme.hu/~oti/portfolio` there are two benchmark data set from NYSE:

- The first data set consists of daily data of 36 stocks with length 24 years (1962–1985)
- The second data set contains 19 stocks and has length 45 years (1962–2006)

NYSE data sets

At `www.szit.bme.hu/~oti/portfolio` there are two benchmark data set from NYSE:

- The first data set consists of daily data of 36 stocks with length 24 years (1962–1985)
- The second data set contains 19 stocks and has length 45 years (1962–2006)

Our experiment is on the second data set

Experiments on average annual yields (AAY) for kernel based experts

Kernel based semi-log-optimal portfolio selection with

Experiments on average annual yields (AAY) for kernel based experts

Kernel based semi-log-optimal portfolio selection with finite array of size $K \times L$ such that $K = 5$ and $L = 10$. Choose the uniform distribution $\{q_{k,\ell}\} = 1/(KL)$ over the experts in use.

Experiments on average annual yields (AAY) for kernel based experts

Kernel based semi-log-optimal portfolio selection with finite array of size $K \times L$ such that $K = 5$ and $L = 10$. Choose the uniform distribution $\{q_{k,\ell}\} = 1/(KL)$ over the experts in use. $k = 1, \ldots, 5$ and $l = 1, \ldots, 10$

$$r_{k,l}^2 = 0.0002 \cdot d \cdot k + 0.00002 \cdot d \cdot k \cdot \ell,$$

Experiments on average annual yields (AAY) for kernel based experts

Kernel based semi-log-optimal portfolio selection with finite array of size $K \times L$ such that $K = 5$ and $L = 10$. Choose the uniform distribution $\{q_{k,\ell}\} = 1/(KL)$ over the experts in use. $k = 1, \ldots, 5$ and $l = 1, \ldots, 10$

$$r^2_{k,l} = 0.0002 \cdot d \cdot k + 0.00002 \cdot d \cdot k \cdot \ell,$$

AAY of kernel based semi-log-optimal portfolio is 31%

Experiments on average annual yields (AAY) for kernel based experts

Kernel based semi-log-optimal portfolio selection with finite array of size $K \times L$ such that $K = 5$ and $L = 10$. Choose the uniform distribution $\{q_{k,\ell}\} = 1/(KL)$ over the experts in use. $k = 1, \ldots, 5$ and $l = 1, \ldots, 10$

$$r_{k,l}^2 = 0.0002 \cdot d \cdot k + 0.00002 \cdot d \cdot k \cdot \ell,$$

AAY of kernel based semi-log-optimal portfolio is 31%
MORRIS had the best AAY, 20%

Experiments on average annual yields (AAY) for kernel based experts

Kernel based semi-log-optimal portfolio selection with finite array of size $K \times L$ such that $K = 5$ and $L = 10$. Choose the uniform distribution $\{q_{k,\ell}\} = 1/(KL)$ over the experts in use. $k = 1, \ldots, 5$ and $l = 1, \ldots, 10$

$$r^2_{k,l} = 0.0002 \cdot d \cdot k + 0.00002 \cdot d \cdot k \cdot \ell,$$

AAY of kernel based semi-log-optimal portfolio is 31%
MORRIS had the best AAY, 20%
the BCRP had average AAY 21%

The average annual yields of the of the kernel based experts.

k / ℓ	1	2	3	4	5
1	31%	30%	24%	21%	26%
2	34%	31%	27%	25%	22%
3	35%	29%	26%	24%	23%
4	35%	30%	30%	32%	27%
5	34%	29%	33%	24%	24%
6	35%	29%	28%	24%	27%
7	33%	29%	32%	23%	23%
8	34%	33%	30%	21%	24%
9	37%	33%	28%	19%	21%
10	34%	29%	26%	20%	24%

Prediction for squared loss

Y_i real valued

X_i vector valued

Prediction for squared loss

Y_i real valued

X_i vector valued

At time instant i the predictor is asked to guess Y_i

Prediction for squared loss

Y_i real valued

X_i vector valued

At time instant i the predictor is asked to guess Y_i with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$

Prediction for squared loss

Y_i real valued

X_i vector valued

At time instant i the predictor is asked to guess Y_i with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$ and $(Y_1, \ldots Y_{i-1}) = Y_1^{i-1}$

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$
and $(Y_1, \ldots Y_{i-1}) = Y_1^{i-1}$
The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$
and $(Y_1, \ldots Y_{i-1}) = Y_1^{i-1}$
The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$
$g_i(X_1^i, Y_1^{i-1})$ is the estimate of Y_i

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \dots, X_i) = X_1^i$
and $(Y_1, \dots Y_{i-1}) = Y_1^{i-1}$
The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$
$g_i(X_1^i, Y_1^{i-1})$ is the estimate of Y_i
After n time instant the empirical squared error

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2.$$

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process
The best predictor is the conditional expectation

$$g_i^*(X_1^i, Y_1^{i-1}) = \mathbf{E}\{Y_i \mid X_1^i, Y_1^{i-1}\}.$$

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process
The best predictor is the conditional expectation

$$g_i^*(X_1^i, Y_1^{i-1}) = \mathbf{E}\{Y_i \mid X_1^i, Y_1^{i-1}\}.$$

The fundamental limit: for any predictor g

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{almost surely,}$$

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process
The best predictor is the conditional expectation

$$g_i^*(X_1^i, Y_1^{i-1}) = \mathbf{E}\{Y_i \mid X_1^i, Y_1^{i-1}\}.$$

The fundamental limit: for any predictor g

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{almost surely,}$$

where

$$L^* = \mathbf{E}\{(Y_0 - \mathbf{E}\{Y_0 | X_{-\infty}^0, Y_{-\infty}^{-1}\})^2\}$$

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $X_{-\infty}^0, Y_{-\infty}^{-1}$.

Universal consistency

there are universally consistent prediction sequence g_n:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 = L^*$$

a.s. for a class of stationary and ergodic sequences.

Universal consistency

there are universally consistent prediction sequence g_n:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 = L^*$$

a.s. for a class of stationary and ergodic sequences.
Such prediction sequence is called **universally consistent**.

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$.

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$,
$k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$,
$k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$,
$k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

and

$$\lim_{\ell \to \infty} r'_{k,\ell} = 0.$$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$,
$k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

and

$$\lim_{\ell \to \infty} r'_{k,\ell} = 0.$$

location of the matches

$$J_n^{(k,\ell)} = \Big\{ k < i < n : \|x_{i-k}^i - x_{n-k}^n\| \leq r_{k,\ell},$$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$,
$k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

and

$$\lim_{\ell \to \infty} r'_{k,\ell} = 0.$$

location of the matches

$$J_n^{(k,\ell)} = \left\{ k < i < n : \|x_{i-k}^i - x_{n-k}^n\| \le r_{k,\ell}, \ \|y_{i-k}^{i-1} - y_{n-k}^{n-1}\| \le r'_{k,\ell} \right\}$$

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent
for small k, the bias is large and, for large k, the variance is large because of the few matchings.

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent
for small k, the bias is large and, for large k, the variance is large because of the few matchings.
for large radius, the bias is large and, for small radius, the variance is large because of the few matchings.

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent
for small k, the bias is large and, for large k, the variance is large because of the few matchings.
for large radius, the bias is large and, for small radius, the variance is large because of the few matchings.

The problem is how to choose k, $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$ in a data dependent way.

Combination of experts: bounded Y

Machine learning: exponential weighting

Combination of experts: bounded Y

Machine learning: exponential weighting

$|Y| \leq B$, and B is known.

Combination of experts: bounded Y

Machine learning: exponential weighting

$|Y| \leq B$, and B is known.

Let $\{q_{k,\ell}\}$ be a probability distribution over (k, ℓ),

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k, ℓ),
and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)}$$

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k, ℓ),
and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)} = q_{k,\ell} e^{-\sum_{i=1}^{t-1}(h_i^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2/(8B^2)}$$

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k, ℓ),
and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)} = q_{k,\ell} e^{-\sum_{i=1}^{t-1}(h_i^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2/(8B^2)}$$

and

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^{\infty} w_{t,i,j}} \ .$$

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k, ℓ),
and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)} = q_{k,\ell} e^{-\sum_{i=1}^{t-1}(h_i^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2/(8B^2)}$$

and

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^{\infty} w_{t,i,j}} \ .$$

Then the combined prediction

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1}) \ .$$

Consistency theorem for bounded Y

Theorem. (Györfi, Lugosi (2001)) If $|Y_0| \leq B$, then the combined predictor g is universally consistent.

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers.

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_k(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$.

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_k(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$,

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_k(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_k(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

If the prediction strategy $\tilde{g}$ is defined by

$$\tilde{g}_t(x_1^t, y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(x_1^t, y_1^{t-1})$$

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_k(x_1^n, y_1^{n-1})| \le B$ and $|y_n| \le B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \ge 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

If the prediction strategy $\tilde{g}$ is defined by

$$\tilde{g}_t(x_1^t, y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(x_1^t, y_1^{t-1})$$

then for every $n \ge 1$,

$$L_n(\tilde{g}) \le \inf_k \left(L_n(\tilde{h}_k) - \frac{c \ln q_k}{n} \right).$$

Sketch of the proof of the Theorem

Because of the fundamental limit

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{a.s.}$$

Sketch of the proof of the Theorem

Because of the fundamental limit

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{a.s.}$$

it is enough to show that

$$\limsup_{n\to\infty} L_n(g) \leq L^* \qquad \text{a.s.}$$

By the Expert Lemma

$$L_n(g) \leq \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

By the Expert Lemma

$$L_n(g) \leq \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\limsup_{n\to\infty} L_n(g) \quad \leq \quad \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right)$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned} \limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\ &\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)})
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&=: \inf_{k,\ell} \epsilon_{k,\ell}
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&=: \inf_{k,\ell} \epsilon_{k,\ell} \\
&= \lim_{k,\ell\to\infty} \epsilon_{k,\ell}
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&=: \inf_{k,\ell} \epsilon_{k,\ell} \\
&= \lim_{k,\ell\to\infty} \epsilon_{k,\ell} \\
&= L^*
\end{aligned}$$