

Robust sequential learning

with applications to the forecasting
of air quality and of exchange rates

Gilles Stoltz

CNRS — HEC Paris

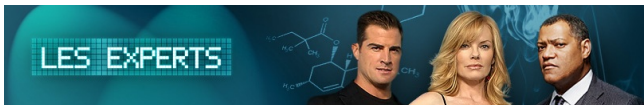


The framework of this talk

Sequential and worst-case deterministic prediction of time series
based on expert advice

A statistician has to predict a time series $y_1, y_2, \dots \in \mathcal{C}$, where $\mathcal{C}$ is a convex subset of $\mathbb{R}^d$.

Finitely many **expert** forecasts are available, e.g., given by some stochastic models.



At each instance t , expert $j \in \{1, \dots, N\}$ outputs a forecast

$$f_{j,t} = f_{j,t}(y_1^{t-1}) \in \mathcal{C}$$

Observations and predictions are made in a **sequential** fashion:

The prediction $\hat{y}_t$ of y_t is determined based

- on the **past** observations $y_1^{t-1} = (y_1, \dots, y_{t-1})$,
- and the **current** and **past** expert forecasts $f_{j,s}$, where $s \in \{1, \dots, t\}$ and $j \in \{1, \dots, N\}$,

before getting to know the actual value y_t .

A typical solution of the problem is to form **convex** (or **linear**) combinations of the expert forecasts, with weights $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ or $\mathbf{v}_t = (v_{1,t}, \dots, v_{N,t})$ adjusted over time.

The statistician then outputs the forecasts $\hat{y}_t = \sum_{j=1}^N p_{j,t} f_{j,t}$

The observations y_t will **not** be considered **stochastic** anymore at this stage; thus the performance criterion will be a relative one.

We consider a convex loss function $\ell : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}_+$, e.g., the square loss $\ell(x, y) = (x - y)^2$ when $\mathcal{C} \subseteq \mathbb{R}$.

The **cumulative losses** of the statistician and of the constant convex combinations $\mathbf{q} = (q_1, \dots, q_N)$ of the expert forecasts equal

$$\hat{L}_T = \sum_{t=1}^T \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) \quad \text{and} \quad L_T(\mathbf{q}) = \sum_{t=1}^T \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right)$$

First study: Forecasting of air quality

Starting date: September 2005

Academic partner: Vivien Mallet, INRIA, project-team CLIME

Industrial partner: Edouard Debry, INERIS (Institut National de l'EnviRonnement Industriel et des RisqueS)

M.Sc. students involved over time:

- Boris Mauricette (6 months in 2007; from M2 Pro Paris-Diderot and ENS de Lyon)
- Sébastien Gerchinovitz (5 months in 2008; from M2 Maths Paris-Sud)
- Karim Drifi (4 months in 2009; from M2 MVA ENS Cachan)
- Paul Baudin (4 months in 2012; from M2 MVA ENS Cachan)

Associated publication: in the Journal of Geophysical Research



*maîtriser le risque
pour un développement durable*

Some characteristics of one among the studied data sets:

- 126 days during summer '01; **one-day ahead** prediction
- 241 stations in France and Germany
- Typical **ozone** concentrations between $40 \mu\text{g m}^{-3}$ and $150 \mu\text{g m}^{-3}$; sometimes above the values $180 \mu\text{g m}^{-3}$ or $240 \mu\text{g m}^{-3}$
- **48 experts**, built in **Mallet et Sportisse '06** by choosing a physical and chemical formulation, a numerical approximation scheme to solve the involved PDEs, and a set of input data (among many)

→ Instead of trusting only one model/expert ("**selection**"), we proceed in a more greedy way and consider many models/experts, which we combine sequentially ("**aggregation**").

This leads to more accurate and more stable (meta-)predictions.

The stations of the network are indexed by $\mathcal{S}$.

Each model $j = 1, \dots, 48$ outputs a prediction $f_{j,t}^s$ for the ozone peak at station s and day t , which is then compared to the measured peak y_t^s . (We discard measurement errors.)

The statistician chooses at each round a single convex weight vector $\mathbf{p}_t$ or linear weight vector $\mathbf{v}_t$ to be used at **all stations**; this leads to **prediction fields**.

The strategies are assessed based on their RMSEs, which amounts to considering the **convex losses**

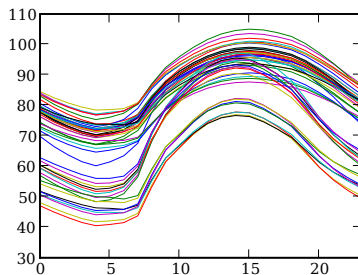
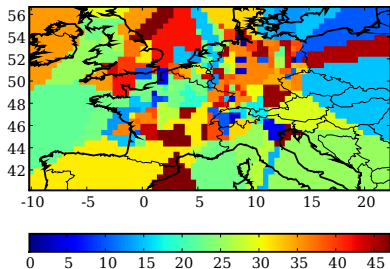
$$\ell_t(\mathbf{p}_t) \stackrel{\text{def}}{=} \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} p_{j,t} f_{j,t}^s - y_t^s \right)^2$$

where $\mathcal{S}_t$ is the subset of active stations at day t .

The RMSE equals $\sqrt{\frac{\sum_{t=t_0}^T \ell_t(\mathbf{p}_t)}{\sum_{t=t_0}^T |\mathcal{S}_t|}}$ for $t_0 = 31$ (short training)

Left: There are several good and useful experts.

Right: Their forecasting profiles are quite different (the experts are not clones the ones of the others!).



Left: Coloring of Europe according to the index of the locally best expert

Right: Average forecasting profiles during a day (averages over time and space)

The framework of this talk

(continued)

The **regret** R_T is defined as the difference

$$\hat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q}) = \sum_{t=1}^T \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) - \min_{\mathbf{q}} \sum_{t=1}^T \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right)$$

We are interested in aggregation rules with (uniformly) **vanishing per-round regret**,

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sup \left\{ \hat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q}) \right\} \leq 0$$

where the supremum is over **all possible sequences** of observations and of expert forecasts. (Not just over most of these sequences!)

Remarks:

- Hence the name “prediction of **individual sequences**” (or **robust** aggregation of expert forecasts).
- The best convex combination $\mathbf{q}^*$ is known **in hindsight** whereas the statistician has to predict in a **sequential** fashion.

This framework leads to a **meta-statistical** interpretation:

- each series of **expert** forecasts is given by a **statistical** forecasting method, possibly tuned with some given set of parameters;
- these base forecasts relying on some stochastic model are then **combined** in a **robust** and **deterministic** manner.

The **cumulative loss** of the statistician can be decomposed as

$$\hat{L}_T = \min_{\mathbf{q}} L_T(\mathbf{q}) + R_T$$

This leads to the following interpretations:

- the term indicating the performance of the best convex combination of the expert forecasts is an **approximation error**;
- the regret term measures a **sequential estimation error**.

First study, continued

Forecasting of the air quality

How good are our experts? See the “oracles” below.

Do we **expect** the aggregation methods to provide significant **improvements**? Yes, whenever the best convex and/or linear combinations significantly outperform the best expert.

Uniform mean	Best expert	Best p	Best u
24.41	22.43	21.45	19.24

Performance, in terms of RMSE, of (some combinations of) the **experts**

Disclaimer

We could also consider **batch learning** methods to aggregate models/experts, like

- BMA (Bayesian model averaging),
- CART (classification and regression trees),
- random forests, etc.,

or even **selection** methods, and apply them online, by running a batch analysis at each step.

We instead resort to **“real” online** techniques that, in addition, come up with theoretical guarantees even in non-stochastic scenarios.

We will also see that calibrating their parameters can be done in a more satisfactory way, using the sequential character of the prediction.

A strategy to pick convex weights

Let's do some maths!

Reminder of the aim and setting:

Given a loss function $\ell : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$, where $\mathcal{C} \subseteq \mathbb{R}^d$ is convex

Choose sequentially the convex weights $p_{j,t}$

To **uniformly bound** the **regret** with respect to all sequences of observations y_t and expert predictions $f_{j,t}$:

$$\sum_{t=1}^T \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) - \min_{\mathbf{q}} \sum_{t=1}^T \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right)$$

When ℓ is convex in its first argument, **sub-gradients** exist, i.e.:

For all $x, y \in \mathcal{C}$, there exists $\nabla \ell(x, y)$ such that

$$\forall x' \in \mathcal{C}, \quad \ell(x, y) - \ell(x', y) \leq \nabla \ell(x, y) \cdot (x - x')$$

To **uniformly bound** the **regret** with respect to all convex weight vectors $\mathbf{q}$, we write

$$\begin{aligned}
 & \max_{\mathbf{q}} \sum_{t=1}^T \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) - \sum_{t=1}^T \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right) \\
 & \leq \max_{\mathbf{q}} \sum_{t=1}^T \nabla \ell \left(\sum_{k=1}^N p_{k,t} f_{k,t}, y_t \right) \cdot \left(\sum_{j=1}^N p_{j,t} f_{j,t} - \sum_{j=1}^N q_j f_{j,t} \right) \\
 & = \max_{\mathbf{q}} \sum_{t=1}^T \left(\sum_{j=1}^N p_{j,t} \tilde{\ell}_{j,t} - \sum_{j=1}^N q_j \tilde{\ell}_{j,t} \right) \\
 & = \sum_{t=1}^T \sum_{j=1}^N p_{j,t} \tilde{\ell}_{j,t} - \min_{i=1, \dots, N} \sum_{t=1}^T \tilde{\ell}_{i,t}
 \end{aligned}$$

where we denoted

$$\tilde{\ell}_{j,t} = \nabla \ell \left(\sum_{k=1}^N p_{k,t} f_{k,t}, y_t \right) \cdot f_{j,t}$$

Via the (**signed**) pseudo-losses

$$\tilde{\ell}_{j,t} = \nabla \ell \left(\sum_{k=1}^N p_{k,t} f_{k,t}, y_t \right) \cdot f_{j,t}$$

it suffices to consider the following **simplified** framework.

At each round $t = 1, 2, \dots$,

- the experts provide forecasts $f_{1,t}, \dots, f_{N,t}$;
- the statistician picks convex weights $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$;
- the environment **then** determines, possibly with the knowledge of $\mathbf{p}_t$, a loss vector $(\tilde{\ell}_{1,t}, \dots, \tilde{\ell}_{N,t})$

The aim is to **bound uniformly** the **regret**

$$R_T = \sum_{t=1}^T \sum_{j=1}^N p_{j,t} \tilde{\ell}_{j,t} - \min_{i=1,\dots,N} \sum_{t=1}^T \tilde{\ell}_{i,t}$$

For all $j \in \{1, \dots, N\}$, we pick $p_{j,1} = 1/N$ and for all $t \geq 2$,

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{k,s}\right)}$$

This strategy is known as performing **exponentially weighted averages** of the past cumulative losses of the experts (with fixed learning rate $\eta > 0$).

Lemma. Consider two real numbers $m \leq M$.

For all $\eta > 0$ and for all **individual sequences** $\tilde{\ell}_{j,t} \in [m, M]$,

$$R_T = \sum_{t=1}^T \sum_{j=1}^N p_{j,t} \tilde{\ell}_{j,t} - \min_{i=1, \dots, N} \sum_{t=1}^T \tilde{\ell}_{i,t} \leq \frac{\ln N}{\eta} + \eta \frac{(M - m)^2}{8} T$$

References: Vovk '90; Littlestone and Warmuth '94

Proof of the regret bound

It relies on **Hoeffding's lemma**: for all random variables X with range $[m, M]$, for all $s \in \mathbb{R}$,

$$\ln \mathbb{E}[e^{sX}] \leq s \mathbb{E}[X] + \frac{s^2}{8} (M - m)^2$$

For all $t = 1, 2, \dots$,

$$\begin{aligned} -\eta \sum_{j=1}^N p_{j,t} \tilde{\ell}_{j,t} &= -\eta \sum_{j=1}^N \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{k,s}\right)} \tilde{\ell}_{j,t} \\ &\geq \ln \frac{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^t \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{k,s}\right)} - \frac{\eta^2}{8} (M - m)^2 \end{aligned}$$

A **telescoping sum** appears and leads to

$$\begin{aligned} \sum_{t=1}^T \sum_{j=1}^N p_{j,t} \tilde{\ell}_{j,t} &\leq \underbrace{-\frac{1}{\eta} \ln \frac{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^T \tilde{\ell}_{j,s}\right)}{N}}_{\leq \min_{i=1, \dots, N} \sum_{t=1}^T \tilde{\ell}_{i,t}} + \eta \frac{(M - m)^2}{8} T \\ &\leq \min_{i=1, \dots, N} \sum_{t=1}^T \tilde{\ell}_{i,t} + \frac{\ln N}{\eta} \end{aligned}$$

We now discuss the obtained bound.

Recall that $[m, M]$ is the loss range.

The stated bound can be **optimized** in η :

$$R_T \leq \min_{\eta > 0} \left\{ \frac{\ln N}{\eta} + \eta \frac{(M - m)^2}{8} T \right\} = (M - m) \sqrt{\frac{T}{2} \ln N}$$

for the (theoretical) optimal choice

$$\eta^* = \frac{1}{M - m} \sqrt{\frac{8 \ln N}{T}}$$

This choice depends on M and m , which are sometimes not known beforehand, as well as on T , which may not be bounded (if the prediction game goes forever).

Since no fixed value of $\eta > 0$ ensures that $R_T = o(T)$, we still have no **fully sequential** strategy... but this can be taken care of.

The possible patches are, first, to resort to the “**doubling trick**.”

Alternatively, the learning rates of the exponentially weighted average strategy may **vary over time**, depending on the past: for $t \geq 2$,

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta_t \sum_{s=1}^{t-1} \tilde{\ell}_{k,s}\right)}$$

By a careful such adaptive choice of the η_t , the following regret bound can be obtained:

$$R_T \leq \square (M - m) \sqrt{T \ln N} + \square (M - m) \ln N$$

where the $\square$ denote some universal constants.

We thus recover the **same orders of magnitude** for the regret bound.

However, these theoretically satisfactory solutions would not work well **in practice**.

This is what we do instead. (It is very **different** from techniques like **cross-validation**: we exploit the sequential fashion.)

The exponentially weighted average strategy $\mathcal{E}_\eta$ with fixed learning rate η picks the convex combination $\mu_t(\eta)$, where

$$p_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{k,s}\right)}$$

We denote its cumulative loss $\hat{L}_t(\eta) = \sum_{s=1}^t \ell\left(\sum_{j=1}^N p_{j,s}(\eta) f_{j,s}, y_s\right)$

Based on the family of the $\mathcal{E}_\eta$, we build a **data-driven meta-strategy** which at each instance $t \geq 2$ resorts to

$$\mathbf{p}_{t+1}(\eta_t) \quad \text{where} \quad \eta_t \in \arg \min_{\eta > 0} \hat{L}_t(\eta)$$

Other natural variants: Focus on the most recent losses

Moving sums (with window of size H):

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1, t-H\}}^{t-1} \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=\max\{1, t-H\}}^{t-1} \tilde{\ell}_{k,s}\right)}$$

One can prove that the regret is $\geq \Omega(T)$ in the worst case.

Discounted losses (with discounts given by a sequence $\beta_t \searrow 0$):

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1 + \beta_{t-s}) \tilde{\ell}_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta_t \sum_{s=1}^{t-1} (1 + \beta_{t-s}) \tilde{\ell}_{k,s}\right)}$$

Sublinear regret bounds hold for suitable sequences (β_t) and (η_t) :

$$t\eta_t \longrightarrow 0 \quad \text{and} \quad \eta_t \sum_{s \leq t} \beta_s \longrightarrow 0$$

(We often take $\beta_s = \Omega(1/s^2)$ in the experimental studies.)

First study, continued

Forecasting of the air quality

Oracles

(RMSE of the experts and of fixed combinations thereof)

Uniform mean	Best expert	Best $\mathbf{p}$
24.41	22.43	21.45

Semi-sequential strategies

(RMSE of the strategies tuned with best parameters in hindsight)

Original version	Moving sums ($H = 83$)	Discounts ($\beta_s = 1/s^2$)
21.47	21.37	21.31

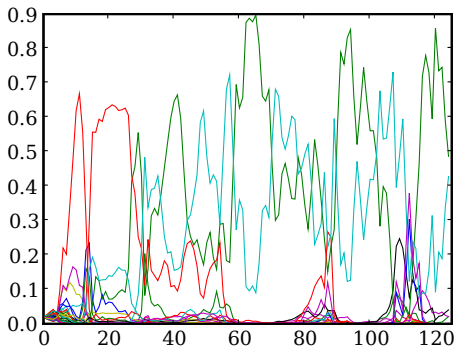
Fully sequential strategies

(RMSE of the original version of the strategy)

Best parameter	Data-driven η_t
21.47	21.77

Our strategies do **not** focus on a single expert. We knew it from the numerical performance.

But actually, the weights associated with the experts **change** quickly and **significantly over time** and do not converge (which illustrates in passing that the performance of the considered experts varies over time).



Convex weights output by the (original) strategy with best parameter η in hindsight

A strategy to pick linear weights

It will ring a bell to the statisticians among you!

Linear combinations: Ridge regression (and the LASSO?)

The ridge regression was introduced in the 70s by Hoerl and Kennard; it was intensively studied since then in a stochastic setting.

We consider the case where $\mathcal{C} \subseteq \mathbb{R}$ and $\ell(x, y) = (x - y)^2$.

The ridge regression resorts to linear combinations of the experts:

$$\mathbf{v}_t \in \arg \min_{\mathbf{u} \in \mathbb{R}^N} \left\{ \lambda \|\mathbf{u}\|_2^2 + \sum_{s=1}^{t-1} \left(y_s - \sum_{j=1}^N u_j f_{j,s} \right)^2 \right\}$$

for some regularization parameter $\lambda > 0$.

It also exhibits a **sublinear regret** against **individual sequences**.

We do not know any such regret bounds for the LASSO **yet**.

Theorem. Consider a bound $B > 0$.

For all $\lambda > 0$, for all **individual sequences** of observations $y_t \in [-B, B]$ and of expert predictions $f_{j,t} \in [-B, B]$,
for **all** $\mathbf{u} \in \mathbb{R}^N$,

$$\begin{aligned} \sum_{t=1}^T \ell \left(\sum_{j=1}^N v_{j,t} f_{j,t}, y_t \right) - \sum_{t=1}^T \ell \left(\sum_{j=1}^N u_j f_{j,t}, y_t \right) \\ \leq \lambda \|\mathbf{u}\|_2^2 + 2NB^2 \left(1 + \frac{NTB^2}{\lambda} \right) \ln \left(1 + \frac{TB^2}{N\lambda} \right) \end{aligned}$$

λ of the order of $1/\sqrt{T}$ is thus a good theoretical choice and leads to $O(\sqrt{T} \ln T)$ regret bounds.

Time-varying or data-driven parameters λ_t can be considered (both for theoretical bounds or for the sake of practical performance).

References: Vovk '01; Azoury and Warmuth '01; Gerchinovitz '11

The interest of this method is that it can compensate for **biases** (in either direction) as the weights do not need to sum up to 1.

Even better, we can/should use it as a **pre-treatment** on each single expert and

- turn it into a modified expert with predictions $\gamma_t f_{j,t}$,
- performing on average almost as well as the best expert of the form $\gamma f_{j,t}$ for some constant $\gamma \in \mathbb{R}$.

This would improve greatly the predictions if there existed, for instance, an almost constant multiplicative bias of $1/\gamma$.

First study, continued

Forecasting of the air quality

In our application to the prediction of air quality, we have extra sums over the stations of the network.

E.g., the ridge regression is defined as

$$\mathbf{v}_t \in \arg \min_{\mathbf{u} \in \mathbb{R}^N} \left\{ \lambda \|\mathbf{u}\|_2^2 + \sum_{\tau=1}^{t-1} \sum_{s \in \mathcal{S}_\tau} \left(y_\tau^s - \sum_{j=1}^N u_j f_{j,\tau}^s \right)^2 \right\}$$

One can show that $O(\sqrt{T} \ln T)$ regret bounds are still preserved.

The experts are indeed improved via the **ridge pre-treatment**. We illustrate this on the worst and best experts.

Original	Pre-treated	Original	Pre-treated
35.79	24.78	22.43	21.66

Oracles

(RMSE of the experts and of fixed combinations thereof)

Uniform mean	Best expert	Best $\mathbf{p}$	Best $\mathbf{u}$
24.41	22.43	21.45	19.24

Semi-sequential ridge regression

(RMSE of the strategies tuned with best parameters in hindsight)

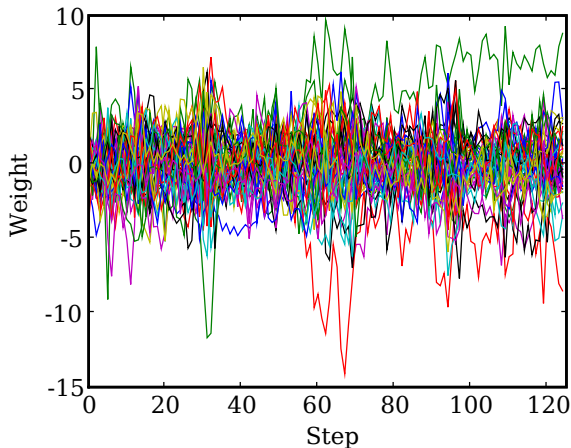
Original version	Moving sums ($H = 45$)	Discounts ($\beta_s = 100/s^2$)
20.77	20.03	19.45

Fully sequential ridge regression

(RMSE of the original version of the strategy)

Best parameter	Data-driven η_t
20.77	20.81

Our strategies do **not** focus on a single expert and the weights associated with the experts do not converge. [...]



Linear weights output by the (discounted) version of the ridge regression

Methodological summary

Methodological summary

- ① Build the N experts (possibly on a training data set) and pick another data set for the evaluation of our methods, with T instances;
- ② Compute some benchmarks and some reference oracles;
- ③ Evaluate our strategies when run with fixed parameters (i.e., with the best parameters in hindsight);
- ④ The performance of interest is actually the one of the data-driven meta-strategies.

We typically expect $T \geq 5N$ (or even $T \geq 10N$).

Hope arises when the oracles are 10% or 20% better than the methods used so far (e.g., the best expert when the latter is known in advance).

This usually requires the experts to be as different as possible.



Second study: Forecasting of exchange rates

Starting date: March 2012

Academic partner: Tomasz Michalski, HEC Paris

M.Sc. student involved over time:

- Christophe Amat (5 months in 2013; from Ecole Polytechnique)

Associated publication: In preparation

The goal is to predict monthly averages r_{t+1} of exchange rates based on **few macro-economic indicators** $x_{j,t}$ describing the state of the world:

- consumer price indexes (**CPI**);
- industrial production (**Prod**);
- monetary mass (**Mon**);
- required rates of return (“interest rates”, **3R**).

They will give rise to four experts.

The prediction horizon is **1-month ahead**.

A classical **stochastic modeling** is

$$\ln r_{t+1} = \ln r_t + \sigma (W_{t+1} - W_t)$$

for some **Brownian** motion W .

It is considered difficult to improve on it (Meese and Rogoff '83).

It will give rise to the final expert (“random walk, **RW**”).

We denote by r_t the averaged exchange rate of currency A with respect to currency B .

We focus on the **log-variation** $y_{t+1} = \Delta_{t+1} = \ln r_{t+1} - \ln r_t$.

The **stochastic modeling** suggested the prediction $f_{0,t+1} = 0$.

The **economic theory** indicates that a given macro-economic indicator $j \in \{1, 2, 3, 4\}$ can be used to forecast the exchange rate according to

$$\hat{\Delta}_{j,t+1} = \ln x_{j,t}^B - \ln x_{j,t}^A \stackrel{\text{def}}{=} f_{j,t+1}$$

Using our methods we propose convex or linear combinations of the log-variations:

$$\hat{\Delta}_{t+1} = \sum_{j=0}^4 u_{j,t+1} f_{j,t+1} = \sum_{j=1}^4 u_{j,t+1} f_{j,t+1}$$

The predicted log-variations $\hat{\Delta}_t$ and exchange rates $\hat{\Delta}_t + \ln r_{t-1}$ are evaluated via their **common RMSE**:

$$\begin{aligned}\widehat{\text{RMSE}}_T &= \sqrt{\frac{1}{T - t_0 + 1} \sum_{t=t_0}^T (\hat{\Delta}_t - \Delta_t)^2} \\ &= \sqrt{\frac{1}{T - t_0 + 1} \sum_{t=t_0}^T ((\hat{\Delta}_t + \ln r_{t-1}) - \ln r_t)^2}\end{aligned}$$

where $t_0 = 30$ allows a short **training period**.

We apply two (families of) strategies:

- **EG** (exponential weights on the gradients of the losses), as it leads to interpretable weights;
- the **ridge regression**, as it pushes in favor of the **RW** expert.

Some **orders of magnitude** for the prediction problems at hand are indicated below.

Time intervals	Every month
Period	April 1973 – May 2013
Time instances T	about 480
Number of experts N	5 ($= 1 + 4$)
USD / GBP	
Median of the Δ_t	1.48×10^{-2}
Maximum of the $ \Delta_t $	11.08×10^{-2}
JPY / USD	
Median of the Δ_t	1.57×10^{-2}
Maximum of the $ \Delta_t $	10.52×10^{-2}

Results for USD / GBP

Experts	RMSE	Oracle	RMSE
RW	2.47×10^{-2}	Best expert	2.47×10^{-2}
CPI	2.71×10^{-2}	Best p	2.47×10^{-2}
3R	2.84×10^{-2}	Best u	2.46×10^{-2}
Prod	2.59×10^{-2}		
Mon	2.68×10^{-2}		

vs.

EG disc.	Semi-seq.	2.39×10^{-2}	
	Fully seq.	2.42×10^{-2}	(-2.1%)
Ridge disc.	Semi-seq.	2.35×10^{-2}	
	Fully seq.	2.36×10^{-2}	(-4.3%)

Results for JPY / USD

Experts	RMSE	Oracle	RMSE
RW	2.82×10^{-2}	Best expert	2.82×10^{-2}
CPI	2.94×10^{-2}	Best p	2.82×10^{-2}
3R	3.05×10^{-2}	Best u	2.79×10^{-2}
Prod	2.88×10^{-2}		
Mon	3.18×10^{-2}		

vs.

EG disc.	Semi-seq.	2.73×10^{-2}	
	Fully seq.	2.79×10^{-2}	(-0.9%)
Ridge disc.	Semi-seq.	2.71×10^{-2}	
	Fully seq.	2.73×10^{-2}	(-3.0%)

Other empirical studies

- Forecasting of the electricity consumption (EDF R&D)
- Forecasting of the production data of oil reservoirs (IFP-EN)



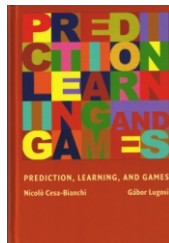
But time is over...

(Please consider inviting Pierre Gaillard for more details on our work with EDF R&D!)

References

In case you're not bored to death (yet) by this topic!

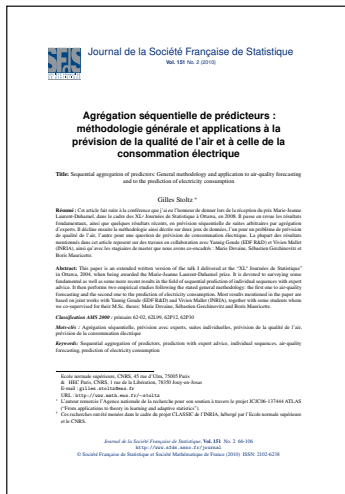
The so-called “red bible!”



Prediction, Learning, and Games

Nicolò Cesa-Bianchi et Gábor Lugosi

I published a survey paper (containing this talk!) one year ago in the **Journal de la Société Française de Statistique**



Even better (or worse)—it is in **French!**