

Robust sequential learning with applications to the forecasting of air quality and of electricity consumption

Gilles Stoltz

CNRS — École normale supérieure — INRIA, équipe CLASSIC
& HEC Paris



A statistician has to predict a sequence y_1, y_2, \dots of observations lying in some set \mathcal{Y} .

His predictions $\hat{y}_1, \hat{y}_2, \dots$ are picked in a set \mathcal{X} .

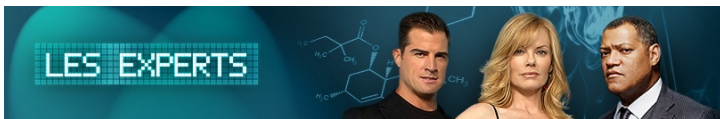
Observations and predictions (1) are made in a **sequential** fashion and (2) rely on **no stochastic modeling**.

(1) mean that for each instance, the prediction \hat{y}_t of y_t is determined

- solely based on the past observations $y_1^{t-1} = (y_1, \dots, y_{t-1})$,
- before getting to know the actual value y_t .

(2) indicates that the methods at hand will not resort to the estimation of some parameters of some stochastic process to build a good model and get some accurate forecasts from it.

To make the problem meaningful, finitely many **expert** forecasts are called for.



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

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

The statistician now determines \hat{y}_t 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 $j \in \{1, \dots, N\}$.

We assume that the set \mathcal{X} of predictions is convex and we restrict the statistician to form **convex combinations** of the expert forecasts.

At each instance t , the statistician thus picks a convex weight vector $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ and forms

$$\hat{y}_t = \sum_{j=1}^N p_{j,t} f_{j,t}$$

The **aim** of the statistician is to predict –on average– as well as the **best constant convex combination** of the expert forecasts.

... But we need first to indicate how to assess the accuracy of a given prediction!

To that end, we consider a **convex loss function** $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$.

When $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$, possible choices are

- the square loss $\ell(x, y) = (x - y)^2$;
- the absolute loss $\ell(x, y) = |x - y|$;
- the absolute percentage of error $\ell(x, y) = |x - y|/|y|$.

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

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

The **regret** is defined as the difference

$$R_T = \widehat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q})$$

Recall that the **regret** R_T is defined as the difference

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

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

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

where the supremum is over **all possible sequences** of observations and of expert forecasts.

This is why this framework is referred to as prediction of **individual sequences** or as **robust** aggregation of expert forecasts.

Note that the best convex combination \mathbf{q}^* can only be determined **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 may be 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**.

Reminder of the aim:

Uniformly bound the regret with respect to all convex weight vectors \mathbf{q} ,

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

When $\mathcal{X} \subseteq \mathbb{R}^d$ and when ℓ is convex in its first argument, sub-gradients exist, i.e.:

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

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

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 $\tilde{\ell}_{j,t}$, it suffices to consider the following simplified framework.

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

- the statistician picks a convex weight vector $\mu_t = (\mu_{1,t}, \dots, \mu_{N,t})$;
- the environment simultaneously determines a loss vector $\ell_t = (\ell_{1,t}, \dots, \ell_{N,t})$;
- the values of μ_t and ℓ_t are both revealed.

The aim is to bound uniformly the regret

$$R_T = \sum_{t=1}^T \sum_{j=1}^N \mu_{j,t} \ell_{j,t} - \min_{i=1, \dots, N} \sum_{t=1}^T \ell_{i,t}$$

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

For all $\eta > 0$ and for all **individual sequences** of elements $\ell_{j,t} \in [m, M]$, where $j \in \{1, \dots, N\}$ and $t \in \{1, \dots, T\}$,

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

where for all $j \in \{1, \dots, N\}$, we picked $\mu_{j,1} = 1/N$ and for all $t \geq 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

This strategy is known as performing **exponentially weighted averages** of the past cumulative losses of the experts (with fixed learning rate η).

We will prove this bound in a minute but for now, we discuss it.

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 not necessarily 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.

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 $t \geq 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

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.

References: Auer, Cesa-Bianchi and Gentile '02; Cesa-Bianchi, Mansour and Stoltz '07

However, these theoretically satisfactory solutions would not work well **in practice**. This is what we do instead.

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

$$\mu_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

We denote its cumulative loss $\hat{L}_t(\eta) = \sum_{s=1}^t \sum_{j=1}^N \mu_{j,s}(\eta) \ell_{j,s}$

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

$$\mu_t(\eta_t) \quad \text{where} \quad \eta_t \in \underset{\eta > 0}{\arg \min} \hat{L}_{t-1}(\eta)$$

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 $t = 1, 2, \dots$,

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

A **telescoping sum** appears and leads to

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

Two empirical studies

- Prediction of air quality
- Forecasting of the electricity consumption

Two empirical studies

The methodology of our studies is in four steps:

- 1 **Build the experts** (possibly on a training data set) and pick another data set for the evaluation of our methods;
- 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**.

First study:

Prediction of air quality

Joint work with Vivien Mallet (INRIA) and M.Sc. students;
published in the Journal of Geophysical Research

Some characteristics of one among the studied data sets:

- 126 days during summer '01
- 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)

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 actual peak y_t^s .

The statistician chooses at each round a single convex weight vector $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ of the experts 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 **loss functions**

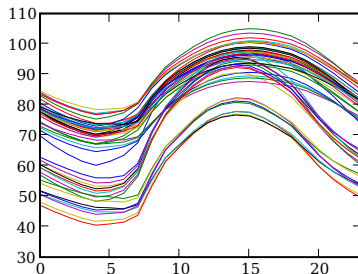
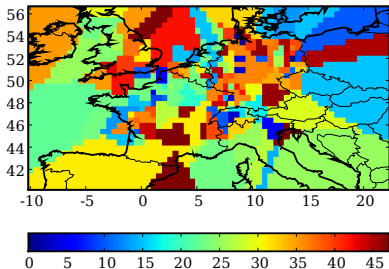
$$l_t(\mathbf{p}_t) = l(\mathbf{p}_t, (y_t^s)_{s \in \mathcal{S}_t}) = \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} p_{j,t} f_{j,t}^s - y_t^s \right)^2$$

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

Then, the RMSE equals $\sqrt{\frac{\sum_{t=1}^T l_t(\mathbf{p}_t)}{\sum_{t=1}^T |\mathcal{S}_t|}}$

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)

Performance of the **experts**

Uniform mean	Best expert	Best p
24.41	22.43	21.45

Performance of the **exponentially weighted average** strategies
(tuned with optimal parameters in hindsight)

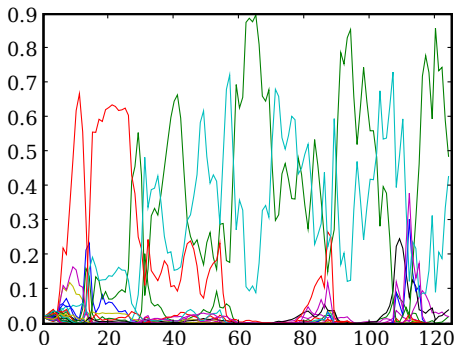
Original version	Fixed history length	Discounted version
21.47	21.37	21.31

The version with **fixed history length** H only uses the losses encountered in the past H rounds.

The version with **discounted losses** puts more weight on more recent losses (while still considering all past losses).

Our strategies do **not** focus on a single expert.

The weights associated with the experts can **change** quickly and **significantly over time** (which illustrates in passing that the performance of the considered experts varies over time).



Convex weight vectors output by the original version of the exponentially weighted average strategy.

Comments and extensions

The good performance is preserved for the **data-driven meta-strategies**.

Other strategies are useful and exhibit good performance (e.g., **ridge regression**).

We should also compare our robust methods to more **classical stochastic ones**.

But we are running out of time to detail this... and instead will switch to a last couple of slides with a different application.

Second study:

Forecasting of the electricity consumption

Joint work with Yannig Goude (EDF R&D) and Marie Devaine; submitted

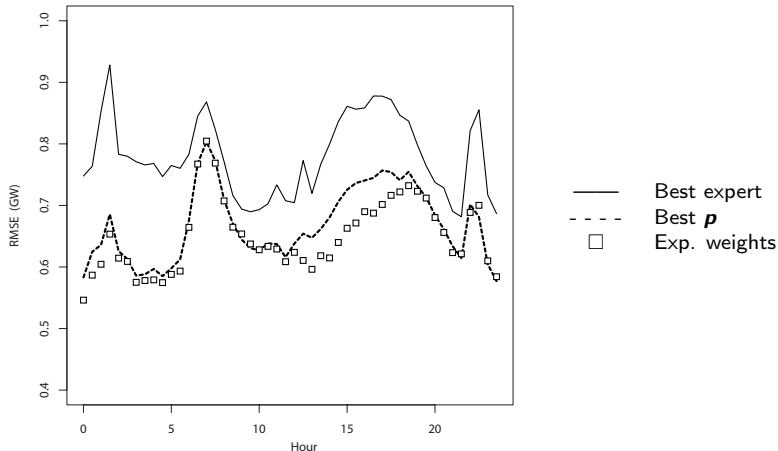
Specialized experts are available: each of them only outputs a forecast when specific conditions are met (working day vs. week end, temperature, etc.).

The definitions and strategies need to be generalized to this setting.

Exhaustive list of references: Blum '97; Freund et al. '97; Cesa-Bianchi and Lugosi '03; Blum and Mansour '07... This is it!

On our data set,

- 3 families of experts, 24 experts in total;
- [operational constraint] the next 48 half-hour instances are to be predicted every day at noon



Average RMSEs (in GW) according to the half hours

Again, the considered strategies achieve in a **sequential** fashion the performance of the best constant convex combination **in hindsight**.