

# 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  $\mathcal{X}$ .

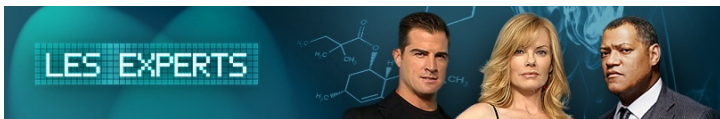
Observations and predictions (1) are made in a **sequential** fashion and (2) rely on **no stochastic modeling**.

(1) means 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**.

## A simple strategy

Let's do some maths. But simple maths, and for 5 minutes only!

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  $\ell$  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  $\mu_t$  and  $\ell_t$  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  $\eta$ ).

## 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

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

We now discuss the obtained bound.

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

The stated bound can be **optimized** in  $\eta$ :

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

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... but this can be taken care of.

## Non stationarity

Competing against sequences of experts with few shifts

In **changing environments** the performance of a given fixed convex combination  $\underline{p}$  can be poor.

A more ambitious goal is to mimic the performance of sequences of the form

$$\underline{p} = (\underline{p}^1, \dots, \underline{p}^1, \underline{p}^2, \dots, \underline{p}^2, \dots, \underline{p}^{m+1}, \dots, \underline{p}^{m+1}),$$

where among the  $T$  rounds up to  $m$  **shifts** can occur.

The cumulative loss  $L_{T,m}^*$  of the best such sequence  $\underline{p}$  is usually much smaller than the cumulative loss of the best fixed convex combination in hindsight,  $\min_{\underline{q}} L_T(\underline{q})$ .

The **cumulative loss** can be decomposed as

$$\hat{L}_T = L_{T,m}^* + R_{T,m},$$

where  $R_{T,m}$  is the corresponding **regret**. And the question is:

How much larger gets the regret bound?

The **fixed-share** algorithm resembles the exponentially weighted average algorithm, except that at the end of each round the weights are **redistributed**, via a **mixing** with the uniform distribution:

$$p_{i,t} \quad \text{becomes} \quad \alpha + (1 - N\alpha)p_{i,t}$$

Fixed-share thus relies on two parameters  $\alpha \geq 0$  and  $\eta > 0$ .

When these are optimally tuned, the regret bound is

$$R_{T,m} \leq \square \sqrt{T m \ln N} + \dots$$

where  $\square$  is some constant depending on the scale of the problem.

We will see that in practice –when indeed breaks occur– this worsening of the **regret** (by a factor of  $\sqrt{m}$ ) is more than **compensated** by the better **approximation error**.

## 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; **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)

## RMSE / Performance of the experts

---

Uniform mean	Best expert	Best $p$
24.41	22.43	21.45

---

RMSE / 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

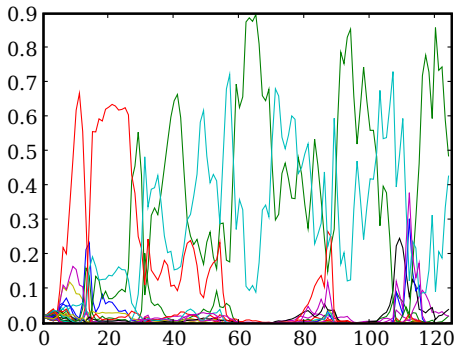
---

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 exponentially weighted average strategy.

## Second study:

### Forecasting of the electricity consumption

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

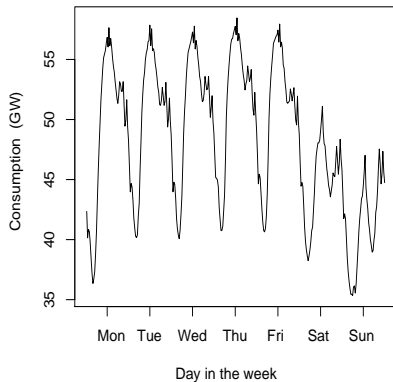
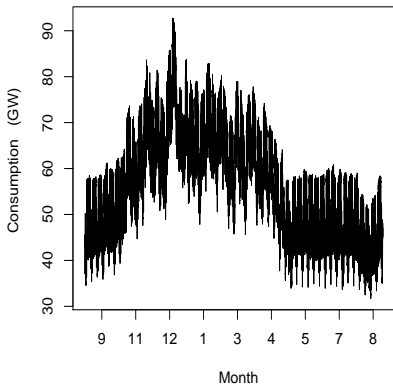
**Specialized experts** are available: each of them only outputs a forecast when specific conditions are met (working day vs. week end, temperature, etc.).

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:] **one-day ahead** prediction at a **half-hour step**, i.e., the next 48 half-hour instances are to be predicted every day at noon



## Electricity consumption in France

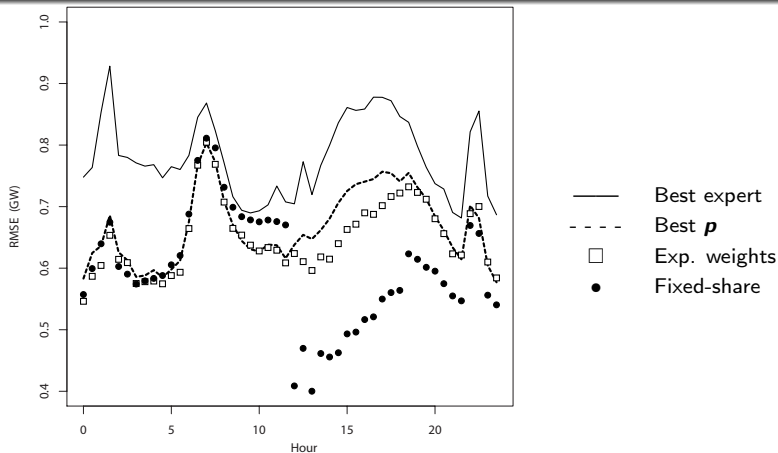
- Year 2007-08 (left)
- Typical summer week (right)

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

Time intervals	Every 30 minutes
Number of days $D$	320
Time instances $T$	15 360 (= 320 × 48)
Number of experts $N$	24 (= 15 + 8 + 1)
Median of the $y_t$	56 330 MW
Bound $B$ on the $y_t$	92 760 MW

We indicate RMSE (average errors and 95 % standard errors).

	Best expert	Uniform mean	Best $p$
	$782 \pm 10$	$724 \pm 11$	$658 \pm 9$
	Exp. weights	Best parameter	Adaptive
		$629 \pm 8$	$637 \pm 9$
Shifts	$m = T - 1 = 15\,359$	$m = 200$	$m = 50$
	$223 \pm ?$	$414 \pm ?$	$534 \pm ?$
	Fixed-Share	Best parameter	Adaptive
		$599 \pm 9$	$629 \pm 8$



Average RMSEs (in GW / not in MW) according to the half hours

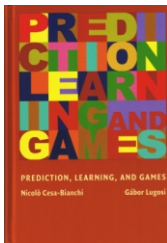
A picture is worth thousand tables, right?

The average RMSE were similar but the behaviors seem **different** by the **half-hours**.

## 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**



Journal de la Société Française de Statistique  
Vol. 151 No. 2 (2010)

**Agrégation séquentielle de prédicteurs :  
méthodologie générale et applications à la  
prévision de la qualité de l'air et à celle de la  
consommation électrique**

**Title:** Sequential aggregation of predictors: General methodology and application to air-quality forecasting and to the prediction of electricity consumption

Gilles Soltz \*

**Résumé :** Cet article fait suite à la conférence que j'ai eu l'honneur de donner lors de la réception de prix Marie-Anne Laurens-Duhouail, dans le cadre des XL<sup>e</sup> Journées de Statistique à Ottawa, en 2008. Il passe en revue les résultats fondamentaux, ainsi que quelques résultats récents, en prévision séquentielle de séries observées par agrégation d'experts. Il discute ensuite la méthodologie ainsi décrite sur deux jeux de données. Y en a pour un problème de prévision de qualité de l'air. L'autre pose une question de prévision de consommation électrique. Le plupart des résultats mentionnés dans cet article reposent sur des travaux en collaboration avec Yanning Guo (IESE BOLD) et Victor Malhotra (DRIIA), ainsi qu'avec les ingénieurs de master que j'ai eu comme co-encadrés : Marie Devaine, Sébastien Gauthier et Brice Mauricette.

**Abstract:** This paper is an extended written version of the talk I delivered at the "XL<sup>e</sup> Journées de Statistique" in Ottawa, 2008, when being awarded the Marie-Anne Laurens-Duhouail prize. It is devoted to surveying some fundamental as well as some more recent results in the field of sequential prediction of individual sequences with expert advice. It then performs two empirical studies following the stated general methodology: the first one to air-quality forecasting and the second one to the prediction of electricity consumption. Most results mentioned in the paper are based on joint works with Yanning Guo (IESE BOLD) and Victor Malhotra (DRIIA), together with some students whom we co-supervised for their M.Sc. thesis: Marie Devaine, Sébastien Gauthier and Brice Mauricette.

**Classification AMS 2000 :** primaire 62-05, 62J99, 62P12, 62P30

**Mots-clés :** Agrégation séquentielle, prévision avec experts, séries individuelles, prévision de la qualité de l'air, prévision de la consommation électrique

**Keywords:** Sequential aggregation of predictors, prediction with expert advice, individual sequences, air-quality forecasting, prediction of electricity consumption

Bureau national supérieures, CNRS, 45 rue d'Ulm, 75005 Paris  
de - HEC Paris, CNRS, 1 rue de la Libération, 75002 Paris  
E-mail : gilles.soltz@hec.fr

URL : <http://www.hec.fr/~soltz>

\* L'auteur remercie l'Agence nationale de la recherche pour son soutien à travers le projet JC036-137444 ATLAS ("From applications to theory in learning and adaptive statistics").

© Ce travail a été réalisé dans le cadre du projet CLASSIC de l'INRIA, hébergé par l'École normale supérieure et le CNRS.

Journal de la Société Française de Statistique, Vol. 151 No. 2 46-116

<http://www.elsevier.com/locate/jfrst>

© Société Française de Statistique et Société Mathématique de France (2010) ISSN: 1202-6238

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