

ROBUST ONLINE AGGREGATION OF FORECASTS

APPLICATION TO ELECTRICITY LOAD FORECASTING

Pierre Gaillard

October 21, 2015

University of Copenhagen

The framework of this talk

SEQUENTIAL prediction of **ARBITRARY** time-series based on expert forecasts:

- a time-series $y_1, \dots, y_n \in \mathbb{R}^d$ is to be predicted
- Expert forecasts are available: e.g., given by some stochastic or machine-learning models (for us: **BLACK BOXES**)



At each forecasting instance $t = 1, \dots, n$

- forecasting black-box $k \in \{1, \dots, K\}$ provides forecast $x_{k,t}$ of y_t
- we are ask to form a prediction \hat{y}_t of y_t with knowledge of
 - the **PAST** observations y_1, \dots, y_{t-1}
 - the **CURRENT** and **PAST** expert forecasts $(x_{k,s})_{s \leq t, 1 \leq k \leq K}$
- we observe y_t

The framework of this talk

SEQUENTIAL prediction of **ARBITRARY** time-series based on expert forecasts:

- a time-series $y_1, \dots, y_n \in \mathbb{R}^d$ is to be predicted
- Expert forecasts are available: e.g., given by some stochastic or machine-learning models (for us: **BLACK BOXES**)

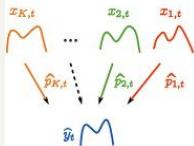


At each forecasting instance $t = 1, \dots, n$

- forecasting black-box $k \in \{1, \dots, K\}$ provides forecast $x_{k,t}$ of y_t
- **TYPICAL SOLUTION:** assign a weight $\hat{p}_{k,t}$ to each expert and predict

$$\hat{y}_t = \sum_{k=1}^K \hat{p}_{k,t} x_{k,t}$$

- we observe y_t



Evaluation criterion

We consider a convex loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, e.g., the square loss $\ell(x, y) = \|x - y\|^2$.

Goal: minimize our average loss

$$\widehat{L}_n = \frac{1}{n} \sum_{t=1}^n \ell(\widehat{y}_t, y_t).$$

Difficulty: **NO STOCHASTIC ASSUMPTION** on the time series

- neither on the observations (y_t)
- neither on the expert forecasts ($x_{k,t}$)

They are **ARBITRARY** and can be chosen by an **ADVERSARY**.

If all experts are bad, good performance is hopeless

➡ relative criterion

The regret: a relative criterion

We evaluate our performance relatively to the ones of the experts

$$\underbrace{\frac{1}{n} \sum_{t=1}^n \widehat{\ell}_t}_{\stackrel{\text{def}}{=} \widehat{L}_n} = \underbrace{\min_{k=1, \dots, K} \frac{1}{n} \sum_{t=1}^n \ell_{k,t}}_{\stackrel{\text{def}}{=} L_n^*} + \underbrace{\frac{1}{n} \sum_{t=1}^n \widehat{\ell}_t - \min_{k=1, \dots, K} \frac{1}{n} \sum_{t=1}^n \ell_{k,t}}_{\stackrel{\text{def}}{=} \text{Reg}_n}$$

our performance reference performance (approximation error) average regret (estimation error)

where $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$ and $\ell_{k,t} = \ell(x_{k,t}, y_t)$.

Goal

Perform almost as good as the best of the experts when $n \rightarrow \infty$

$$\limsup_{n \rightarrow \infty} \left(\sup_{(y_t), (x_{k,t})} \text{Reg}_n \right) \leq 0$$

Best convex combination

A more ambitious approximation error

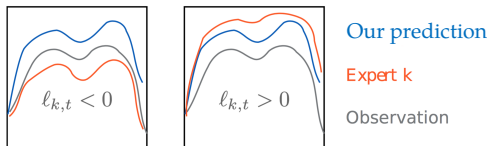
$$\min_{\mathbf{q} \in \Delta_K} \frac{1}{n} \sum_{t=1}^n \ell \left(\sum_{k=1}^K q_k x_{k,t}, y_t \right)$$

where $\Delta_K = \{\mathbf{q} \in \mathbb{R}_+^K : \sum_{k=1}^K q_k = 1\}$.

If an expert provides inaccurate forecasts which compensate other expert forecasts, we should increase its weight.

➔ The **GRADIENT TRICK** formalizes this idea

Example for the square loss: $(x_{k,t} - y_t)^2 \rightarrow (\hat{y}_t - y_t)(x_{k,t} - y_t)$



A meta-statistical interpretation:

- **expert forecasts** are given by some **statistical** forecasting methods, each possibly tuned with a different given set of parameters. They may rely on some stochastic model.
- these ensemble forecasts are then **combined** in a **robust and deterministic manner**

A trade-off: **our final performance** expresses these two parts

$$\widehat{L}_n = L_n^* + \text{Reg}_n$$

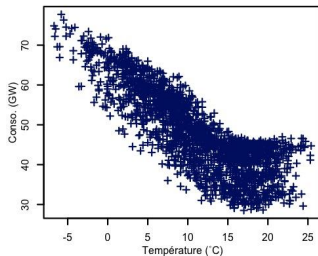
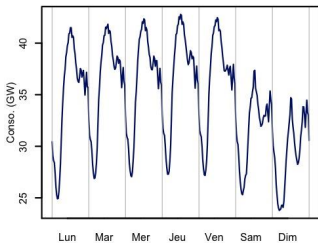
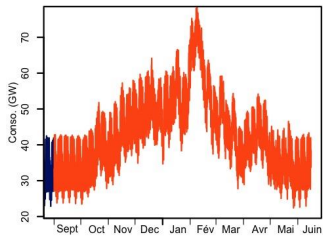
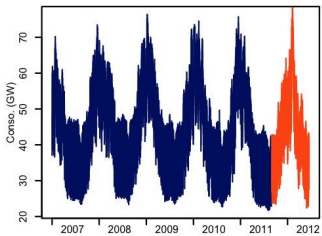
Application: electricity load forecasting

Goal: a day-ahead forecasting of the French electricity load

Data characteristics:

- January 1, 2008 – August 31, 2011 as a **TRAINING** data set
- September 1, 2011 – June 15, 2012 (excluding some special days) as **TESTING** set
- Electricity demand for EDF clients, at a half-hour step
- Typical values: median = 43 496 MW,
maximum = 78 922 MW
- Three expert forecasters: GAM, CLR, KWF

Data looks like...



Application: electricity load forecasting

Convex loss functions considered:

- squareloss: $\ell(x, y) = (x - y)^2 \rightarrow$ RMSE
- absolute percentage of error: $\ell(x, y) = |x - y|/y \rightarrow$ MAPE

Operational constraint:

One-day ahead prediction at a half-hour step, i.e., 48 aggregated forecasts

Expert forecasters:

- GAM / **generalized additive models**
(see Wood 2006; Wood, Goude, Shaw 2014)
- CLR / **curve linear regression**
(see Cho, Goude, Brossat, Yao 2013, 2014)
- KWF / **functional wavelet-kernel approach**
(see Antoniadis, Paparoditis, Sapatinas 2006; Antoniadis, Brossat, Cugliari, Poggi 2012, 2013)

How good are our expert?

Loss: RMSE and MAPE on the testing sets (with no warm-up period)

$$\sqrt{\frac{1}{n} \sum_{t=1}^n (\hat{y}_t - y_t)^2}$$

$$\frac{1}{n} \sum_{t=1}^n \frac{|\hat{y}_t - y_t|}{y_t}$$

We look at the performance of the **ORACLES**:

	Uniform mean	Best forecaster	Best convex p	Best linear u
RMSE (MW)	725	744	629	629
MAPE (MW)	1.18	1.29	1.06	1.06

A strategy to pick the convex weights

The exponentially weighted average forecaster (EWA)

Parameter: $\eta > 0$

Initialization: $\widehat{\mathbf{p}}_1 = (1/K, \dots, 1/K)$

At each time step t , we assign to expert k the weight

$$\widehat{p}_{k,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}$$

Performance: if the loss is convex and bounded by B ,

$$\begin{aligned} \text{Reg}_n &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^n \widehat{\ell}_t - \min_k \frac{1}{n} \sum_{t=1}^n \ell_{k,t} \leq \frac{\log K}{\eta n} + \frac{\eta B^2}{8} \\ &\leq B \sqrt{\frac{\log K}{2n}} \quad \text{for } \eta = B^{-1} \sqrt{\frac{8 \log K}{n}} \end{aligned}$$

A strategy to pick the convex weights

The exponentially weighted average forecaster (EWA)

Parameter: $\eta > 0$

Initialization: $\widehat{\mathbf{p}}_1 = (1/K, \dots, 1/K)$

At each time step t , we assign to expert k the weight

$$\widehat{p}_{k,t} = \frac{\widehat{p}_{k,t-1} e^{-\eta \ell_{k,t-1}}}{\sum_{j=1}^K \widehat{p}_{j,t-1} e^{-\eta \ell_{j,t-1}}}$$

Performance: if the loss is convex and bounded by B ,

$$\begin{aligned} \text{Reg}_n &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^n \widehat{\ell}_t - \min_k \frac{1}{n} \sum_{t=1}^n \ell_{k,t} \leq \frac{\log K}{\eta n} + \frac{\eta B^2}{8} \\ &\leq B \sqrt{\frac{\log K}{2n}} \quad \text{for } \eta = B^{-1} \sqrt{\frac{8 \log K}{n}} \end{aligned}$$

Proof: let's do some maths...

Lemma (Hoeffding)

Let X be a random variable taking value in $[0, B]$. Then for any $s \in \mathbb{R}$

$$\log \mathbb{E}[e^{sX}] \leq s\mathbb{E}[X] + \frac{s^2 B^2}{8}$$

1. Upper bound the instantaneous loss $\widehat{\ell}_t$

$$\begin{aligned} \widehat{\ell}_t = \ell(\widehat{\mathbf{p}}_t \cdot \mathbf{x}_t, y_t) & \stackrel{\text{by convexity}}{\leq} \widehat{\mathbf{p}}_t \cdot \ell(\mathbf{x}_t, y_t) \\ & \stackrel{\text{by Hoeffding}}{\leq} -\frac{1}{\eta} \log \left(\sum_{k=1}^K \widehat{p}_{k,t} e^{-\eta \ell_{k,t}} \right) + \frac{\eta B^2}{8} \\ & \stackrel{\text{by definition of } \widehat{p}_{k,t+1}}{=} -\frac{1}{\eta} \log \left(\frac{\widehat{p}_{k,t}}{\widehat{p}_{k,t+1}} e^{-\eta \ell_{k,t}} \right) + \frac{\eta B^2}{8} \\ & = \ell_{k,t} + \frac{1}{\eta} \log \frac{\widehat{p}_{k,t+1}}{\widehat{p}_{k,t}} + \frac{\eta B^2}{8} \end{aligned}$$

2. Sum over all t , the sum telescopes

$$\sum_{t=1}^n \widehat{\ell}_t - \ell_{k,t} \leq \frac{1}{\eta} \log \frac{\widehat{p}_{k,n+1}}{\widehat{p}_{k,1}} + \frac{\eta n B^2}{8} \leq \frac{\log K}{\eta n} + \frac{\eta B^2}{8}$$

Best theoretical value

$$\eta^* = B^{-1} \sqrt{\frac{8 \log K}{n}}$$

Issue: n and B are not known in advance!

Solutions:

- “doubling trick”
- adaptive learning rate η_t picked according to some theoretical value
- use simultaneously several learning rates. . .
- calibrate on a grid by choosing

$$\eta_t \in \arg \min_{\eta} \left\{ \text{Loss of Exp. weights with } \eta \text{ until time } t - 1 \right\}$$

Application to electricity load forecasting (continued)

Benchmark and oracles (RMSE)

	Uniform mean	Best forecaster	Best convex p	Best linear u
RMSE (MW)	725	744	629	629

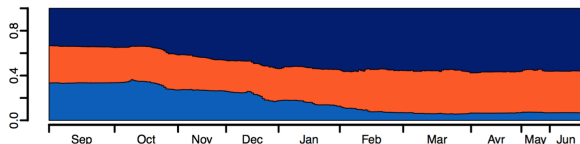
vs.

Aggregated forecasts with convex weights

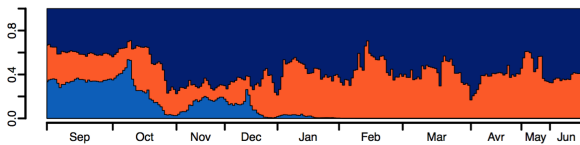
Exp. weights (best η for theory)	644
Exp. weights (best η on data)	644
Exp. weights (best η tuned on data)	625
ML-Poly (tuned according to theory)	626

Evolution of the weights

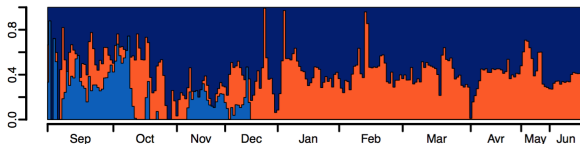
No focus on a single member!



← Exp. weights (theory)



← Exp. weights (best η)



← ML-Poly (theory)

Weights change significantly over time and do not converge
(illustrate that performance of forecasters varies over time)

Are all forecasters useful?

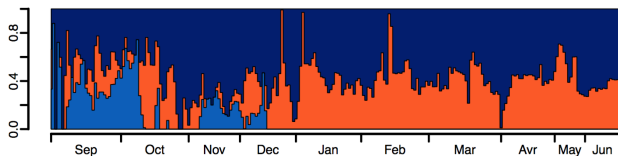
Definitely yes!

3 forecasters \rightarrow only best 2

Exp. weights 625 \rightarrow 644

ML-Poly 626 \rightarrow 646

Forecasters not considered anymore can come back if needed



\leftarrow ML-Poly

Conclusion

This was only a **SMALL GLIMPSE** into the work performed during my PhD at EDF R&D.

I applied the method to many other data sets with good results
→ **UNIVERSALITY** of the method

Here, with Olivier we aim at working on

- huge number of experts → **SPARSE** and **EFFICIENT** methods
- better calibration of the learning parameter to get **FASTER RATES**
- lower bounds
- **PROBABILISTIC FORECASTS** by using the pinball loss

THANKS