ROBUST ONLINE AGGREGATION OF FORECASTS

Pierre Gaillard

April 18, 2016

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 $\widehat{y}_t$ of y_t with knowledge of
  - the past observations $y_1, \dots, y_{t-1}$
  - the current and past expert forecasts $(x_{k,s})_{s \le t,\ 1 \le k \le K}$
- we observe y_t

THE FRAMEWORK OF THIS TALK

Sequential prediction of arbitrary time-series based on expert forecasts:

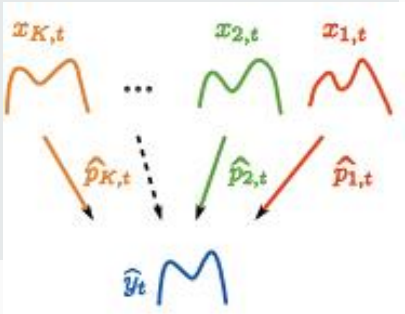
- a time-series $y_1, \ldots, y_n \in \mathbb{R}^d$ is to be predicted
- Expert forecasts are available: e.g., given by some stochastic or machine-learning models (for us: black boxes)

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

- forecasting black-box $k \in \{1, \ldots, K\}$ provides forecast $x_{k,t}$ of y_t
- typical solution: assign a weight $\widehat{p}_{k,t}$ to each expert and predict

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

- we observe y_t



We consider a convex loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, 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

We evaluate our performance relatively to the ones of the experts

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

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 $n \to \infty$

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

BEST CONVEX COMBINATION

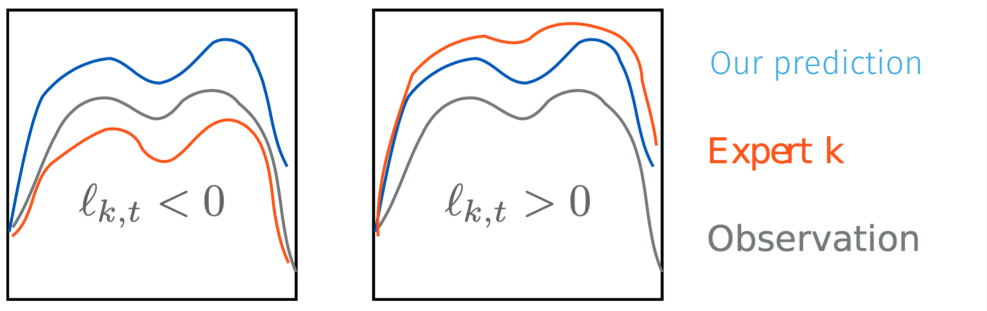
A more ambitious approximation error

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

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

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: $(x_{k,t} - y_t)^2 \quad \rightarrow \quad (\widehat{y}_t - y_t)(x_{k,t} - y_t)$



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

$$\widehat{L}_n = L_n^\star + \mathrm{Reg}_n$$

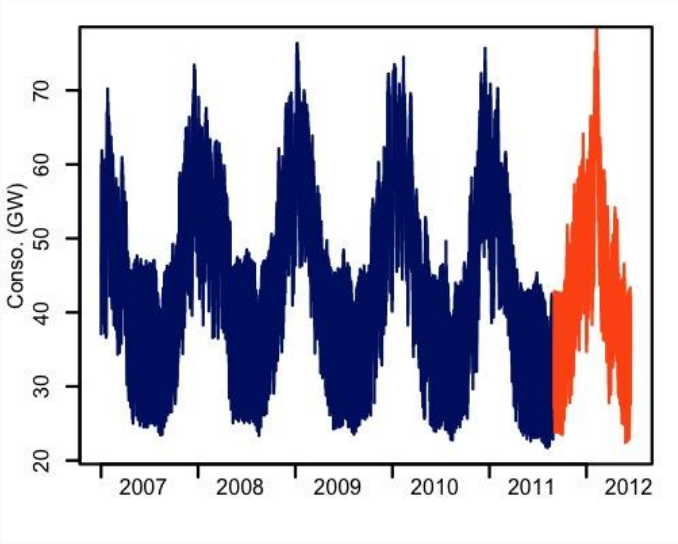
APPLICATION: ELECTRICITY LOAD FORECASTING

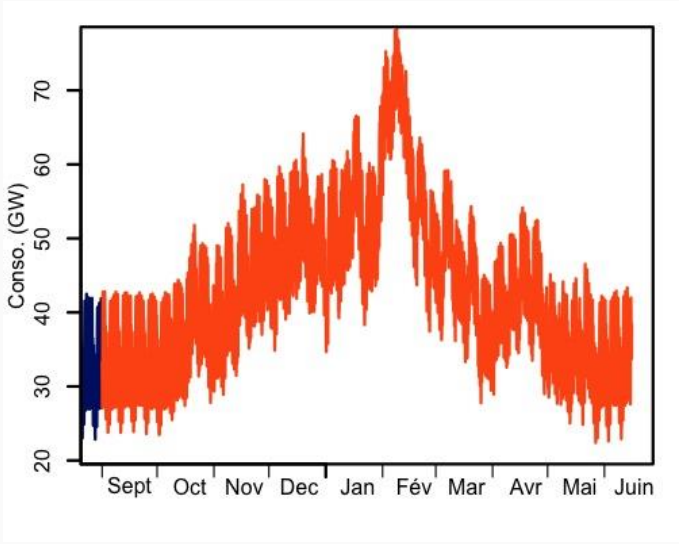
Goal: a day-ahead forecasting of the French electricity load

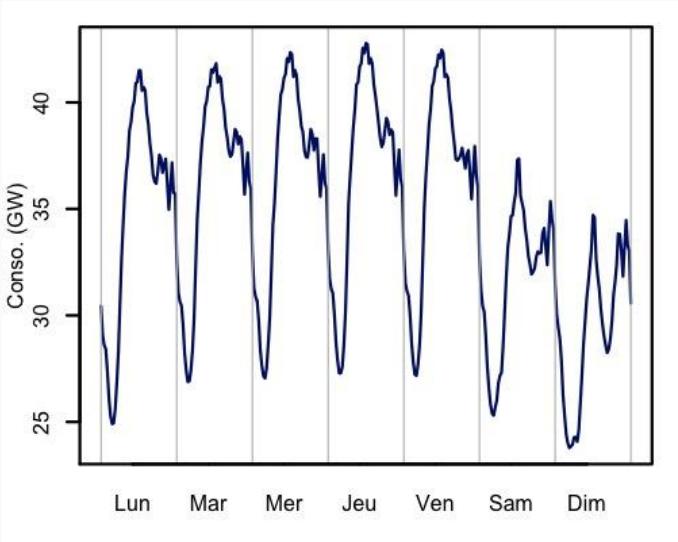
Data characteristics:

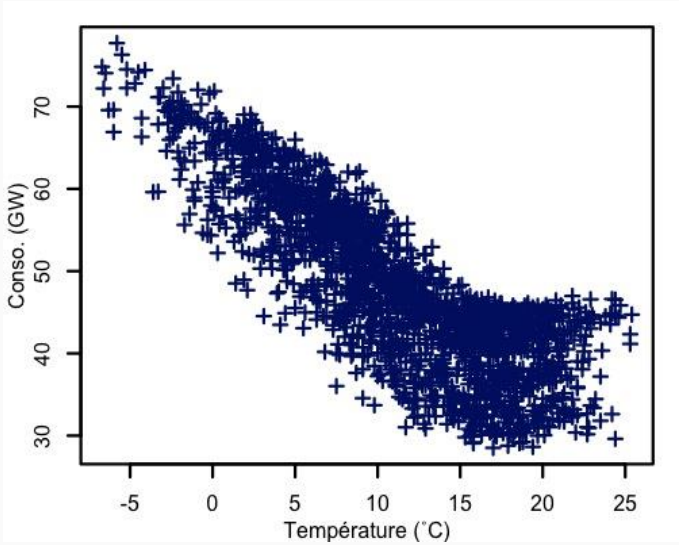
- January 1, 2008 – August 31, 2011 as a training data set
- September 1, 2011 – June 15, 2012 (excluding some special days) as testing set
- Electricity demand for EDF clients, at a half-hour step
- Typical values: median = 43 496 MW,
  maximum = 78 922 MW
- Three expert forecasters: GAM, CLR, KWF

DATA LOOKS LIKE...









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)

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

We look at the performance of the oracles:

	Uniform mean	Best forecaster	Best convex $\boldsymbol{p}$	Best linear $\boldsymbol{u}$
RMSE (MW)	725	744	629	629
MAPE (MW)	1.18	1.29	1.06	1.06

The exponentially weighted average forecaster (EWA)

Parameter: $\eta > 0$ Initialization: $\widehat{p}_1 = (1/K, \dots, 1/K)$

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

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

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

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

The exponentially weighted average forecaster (EWA)

Parameter: $\eta > 0$ Initialization: $\widehat{p}_1 = (1/K, \dots, 1/K)$

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,

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

PROOF: LET'S DO SOME MATHS...

Lemma (Hoeffding)

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

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

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

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

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

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

Issue: n and B are not known in advance!

Solutions:

- "doubling trick"
- adaptive learning rate η_t picked according to some theoretical value
- calibrate on a grid by choosing

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

Benchmark and oracles (RMSE)

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

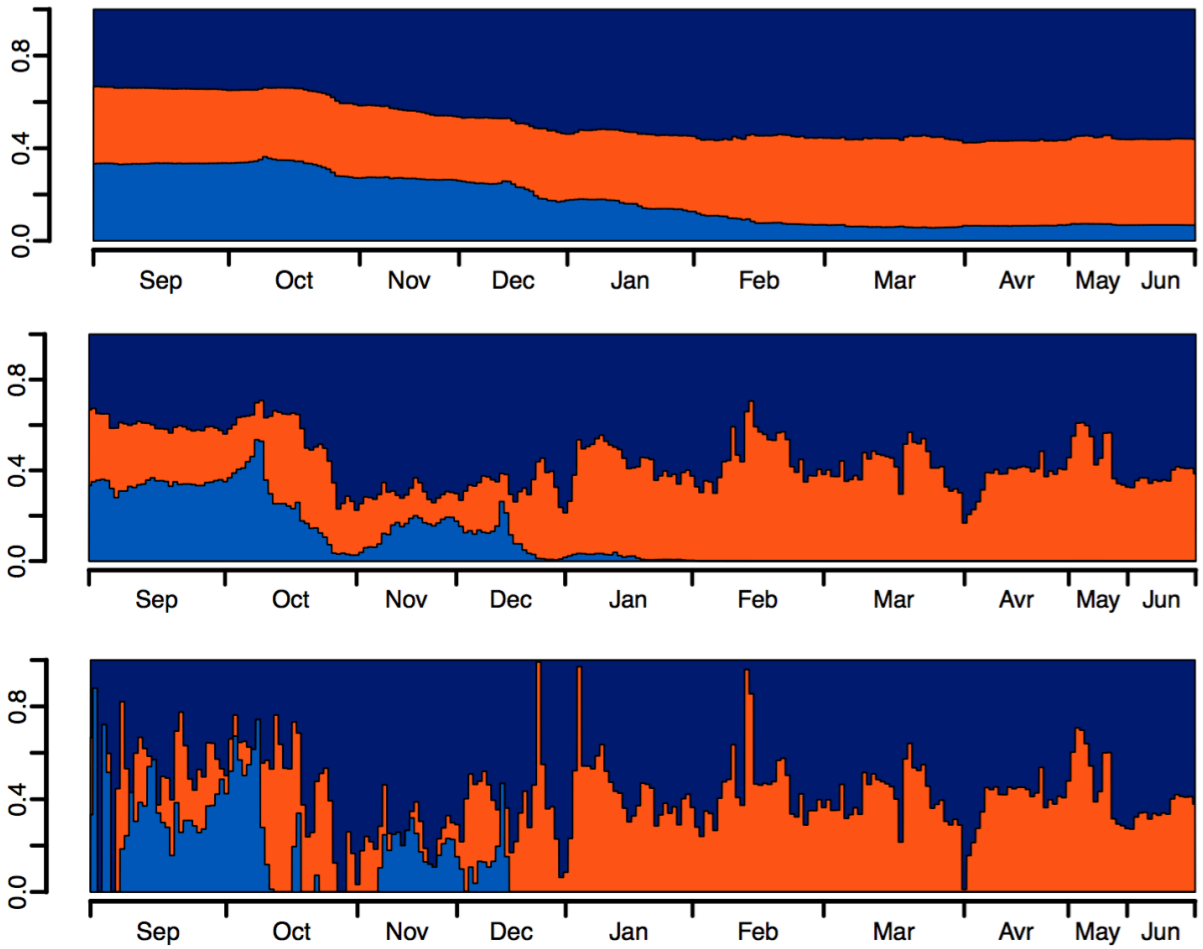
vs.

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

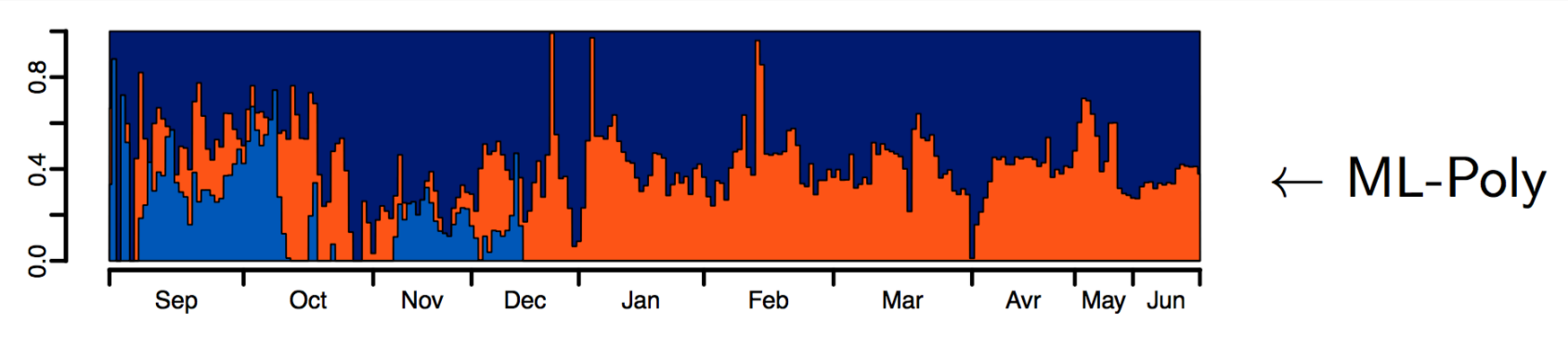
(illustrate that performance of forecasters varies over time)

ARE ALL FORECASTERS USEFUL?

Definitely yes!

3 forecasters → only best 2	
Exp. weights	625 → 644
ML-Poly	626 → 646

Forecasters not considered anymore can come back if needed



BETTER CONVERGENCE RATES ?

If the loss is convex and bounded

$$\mathrm{Reg}_n \overset{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \widehat{\ell}_t - \min_k \frac{1}{n}\sum_{t=1}^{n} \ell_{k,t} \quad \lesssim \quad \sqrt{\log K}\, n^{-1/2}$$

The rate $n^{-1/2}$ is optimal.

Can we do better?

- With additional assumption such as exp-concavity ($e^{-\eta \ell(x,y)}$ is concave)

$$\mathrm{Reg}_n \lesssim (\log K)\, n^{-1}$$

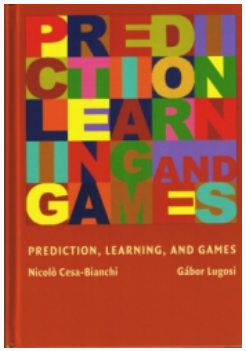
- If one of the experts performs well

$$\mathrm{Reg}_n \lesssim \sqrt{\min_k \frac{1}{n}\sum_{t=1}^{n} \ell_{k,t} (\log K)}\, n^{-1/2}$$

Here, with O. Wintenberger we aim at proving

- such improvements for other scenarios (stochastic setting with condition on the loss function).
- data-dependent lower-bounds

This was only a small glimpse into the setting of prediction with expert advice.
For more details:



N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

G. "Contributions to online robust aggregation: work on the approximation error and on probabilistic forecasting. Applications to forecasting for energy market". PhD thesis. Université Paris-Sud 11, 2015.

The method was applied to many other data sets with good results.

I developed the R-package OPERA implementing the methods (will be presented at useR in June)

THANKS