Forecast of the electricity consumption by aggregation of specialized experts; application to Slovakian and French country-wide hourly predictions

Marie Devaine, Yannig Goude, Gilles Stoltz

Forecast of the electricity consumption by aggregation of specialized experts; application to Slovakian and French country-wide hourly predictions

Marie Devaine*
Ecole Normale Supérieure
Paris, France
marie.devaine@ens.fr

Yannig Goude
EDF R&D
Clamart, France
yannig.goude@edf.fr

Gilles Stoltz†
Ecole Normale Supérieure‡, CNRS
Paris, France
&
HEC Paris, CNRS,
Jouy-en-Josas, France
gilles.stoltz@ens.fr

May 19, 2010

Gilles Stoltz is the corresponding author.

Abstract

We consider the sequential short-term forecast of electricity consumption based on ensemble methods. That is, we use several possibly independent base forecasters and design meta-forecasters which combine the base predictions that are output by them. These meta-forecasters are extracted from the literature of sequential learning, namely, from the field called prediction with expert advice, and come with strong theoretical guarantees. The forecasters considered here may be specialized and need not output a prediction at all time indexes while the meta-forecasters have to. We first motivate review and adapt existing meta-forecasters to our setting and then describe the improvements obtained by these techniques on two data sets, a Slovakian one and a French one, respectively concerned with hourly and half-hourly predictions. These improvements lie in a reduced mean squared error but also in a more robust behavior with respect to large occasional errors.

*Research carried out while being an M.Sc. student at Université Paris-Sud Orsay and completing an internship at EDF R&D, Clamart.

†Partially supported by the French "Agence Nationale pour la Recherche" under grant JCJC06-137444 "From applications to theory in learning and adaptive statistics" and by the PASCAL Network of Excellence under EC grant no. 506778.

‡Research carried out within the INRIA project CLASSIC hosted by Ecole normale supérieure and CNRS.

1 Introduction and motivation

We consider the sequential short-term forecast of electricity consumption based on ensemble methods. That is, we use several possibly independent base forecasters and design meta-forecasters which combine the base predictions that are output by them. These base forecasters can be given by various methods combining some side information, some stochastic estimation, as well as, possibly, some numerical simulation. The design of such forecasters is the focus of a large literature as electricity demand forecasting stands for a central point in power system scheduling; it is the core work of R&D departments of electricity providers like the French largest such company, EDF ("Electricité de France").

The side information used consists in all the features that were shown to have a strong effect on electricity load; see, e.g., Bunn and Farmer [1985]. Among others, one can cite seasonal effects and mostly seasonal variation of day lengths, calendar events like vacation periods or public holidays, weather conditions (temperature, cloud cover, wind), and weekly patterns of days. Classifying electricity load forecasting methods is a hard task and we only highlight some common statistical approaches. Seasonal ARIMA and state space models were introduced by Campo and Ruiz [1987] and are still in use nowadays (see, e.g., Harvey and Koopman [1993] or Dordonnat et al. [2008]). Then come regression or multivariate regression techniques; they are popular in industry as they lead to a convenient interpretation of the different effects driving the load consumption process. They are used in the extensive regression model by hour of the day built by Ramanathan et al. [1997] or in the nonlinear regression model developed by EDF R&D and presented in Bruhns et al. [2005]. The latter model will be used in this paper to provide some base forecasters. Semi-parametric Bayesian regressions with independant or correlated errors are applied to short-term electricity load forecasting in Smith [2000] and Cottet and Smith [2003], allowing not only to output point forecasts but also distributional ones. For highly short-term forecasting (less than one day ahead), univariate methods, without weather variables, are also popular. Among them, exponential smoothing seems to be a good choice as shown by Taylor et al. [2006] on two data sets (Rio de Janeiro; England and Wales) or in Taylor [2008] on minute-by-minute load data. Another univariate approach relying on nonparametric regression based on functional kernels considers the observed load demand as discrete recordings of an underlying stochastic curve; see Antoniadis et al. [2006] and Antoniadis et al. [2010]. Another base forecaster of the present paper is produced with this method. At the same time that these statistical methods were applied to the load consumption, new opportunities came from other multivariate methods based on artificial intelligence and machine learning techniques; for example, several papers have reported successful experiments about the use of neural networks to forecast the electricity load. We refer to Hippert et al. [2001] or Taylor et al. [2006] for an extended description.

In this paper we consider a bunch of base forecasters constructed by the methods reviewed above and study meta-forecasters that use them as sub-routines and combine their predictions; their goal is to perform as well as, or even outperform, the best base forecaster. For instance, a way to construct these base forecasters is to instantiate a prediction method based on several parameters with different sets of parameters (one then obtains a base forecaster per set of parameters); this avoids having to fully tune the parameters of the method, which is usually a delicate and critical issue. The difficulty and the underlying reason of considering several base forecasters is that it is often not clear in advance which of them will behave best. The high-level principle of our meta-forecasters is to output convex combinations of the predictions output by the base forecasters, where the values of the convex weights chosen vary with time according to the performance of

the base forecasters. This is why we speak of sequential aggregation rules.

This kind of rules were already successfully considered for the dayly prediction of the French electricity load, see Goude [2008a,b]). Experimental results reported in this paper demonstrate that not only the performance of the (unknown in advance) best base forecaster is achieved but also that the latter may even be outperformed.

Besides, the aggregation rules described here come with strong theoretical guarantees on their performance and these guarantees are not linked in any sense to a stochastic model. In fact, they hold for all sequences of consumptions, in a worst-case sense. Thus, the base forecasters might rely on some stochastic modeling but their aggregation and its performance do not. Of course, the type of guarantees that can be given here is with respect to the performance of the base forecasters; more precisely, of interest will be here the difference between the average errors encountered by the sequential aggregation rule and by the best base forecaster or the best constant convex combination of the base forecasters – a quantity called regret.

A good general reference for the framework considered here and the theoretical guarantees that can be proved is provided by the monography of Cesa-Bianchi and Lugosi [2006], which calls the base forecasters the experts so that the whole framework is also termed as prediction with experts advice. We will use indifferently the terminologies experts and base forecasters in the sequel.

The studied aggregation rules apply in theory to various other settings; however there has been only a small number of empirical studies on real data. The most famous line of experiments was in the direction of on-line investment in the stock market, that is, on-line aggregation of portfolios; this trend was initiated by the seminal paper of Cover [1991], with corresponding data set formed by the performance of 36 assets of the New-York stock exchange in the period 1963–1985; the experts here are identified with the base assets. A second line of work considers the predictions of outcomes of sports games, see Dani et al. [2006] or Vovk and Zhdanov [2008]. The experts therein are given by odds formed either by bookmakers or by individual participants outputting their bets on a web site. Finally, Mallet et al. [2009] is focused on the sequential prediction of ozone peaks with the help of 48 experts given by different physical, chemical, and numerical methods, as well as different sets of input parameters.

We are aware of no other context of application on real data of the techniques provided by the theory of prediction with expert advice and believe that the present paper contributes to making these techniques more popular – a notoriety that they deserve.

In this paper we consider actually a variant of the basic problem of prediction with expert advice called prediction with specialized (or sleeping) experts: at each round only some of the experts output a prediction while the other ones are inactive. This more difficult setting does not arise from experts being lazy but rather from them being specialized. Indeed, each expert is aware that it is accurate only in given external conditions, that can be known beforehand; when these conditions are not met it then refrains from forming a prediction. For instance, in the case of prediction of electricity consumption, experts can be specialized to winter or to summer, to working days or to public holidays, etc.

The literature on specialized experts is rather sparse. The seminal paper is Freund et al. [1997] and it was followed only by few other ones: two papers mention some results for the context of specialized experts only by passing ([Blum and Mansour, 2007, Sections 6–8] and [Cesa-Bianchi and Lugosi, 2003, Section 6.2]) while another one considers a somewhat different notion of regret, namely, Kleinberg et al. [2008].

Outline of the paper We present in Section 2 the framework of prediction with expert advice, by defining in general the notion of sequential aggregation rules (Section 2.1), by commenting on the chosen assessment criterion formed by the regret (Section 2.2), and by exhibiting three families of such rules (Section 2.3). We then study, respectively in Sections 3 and 4, the performance obtained by the rules on two data sets. The first one was provided by the Slovakian subbranch of the French largest electricity provider EDF and represents its local market; the second one deals with the French market for which EDF is still the overwhelming provider. These empirical studies are organized according to the same scheme: presentation of the data set and of its characteristics, including the performance of some benchmarking prediction methods; results obtained by the sequential aggregation rules with parameters optimally tuned (in hindsight); stability of the previous results when the tuning is performed sequentially (as it should be). The section on French data is also followed by a note (Section 4.3) on the individual performance of the aggregation rules, i.e., an indication that their behavior is not only good on average but also more frequently than any base expert.

2 Sequential aggregation of specialized experts

A sequence of electricity consumptions $y_1, y_2, \ldots, y_T$ is to be predicted at time indexes $t = 1, 2, \ldots, T$. For the sake of concreteness, we assume that the consumptions are all bounded by some constant B, so that the y_t lie in $[0, B]$. A finite number N of base forecasting methods, henceforth referred to as experts, are available; they are indexed by $j = 1, \ldots, N$. Before each time index t, some experts provide a forecast and the other ones do not. The first ones are said active and their forecasts are denoted by $f_{j,t} \in \mathbb{R}_+$, where j is the index of the considered active expert; the experts of the second group are said inactive. By convention, we extend the notation for the forecasts as follows: if expert j is inactive at round t, then $f_{j,t} = 0$.

We denote by $E_t \subset \{1, \ldots, N\}$ the set of active experts at a given time index t.

2.1 Definition of a sequential convex aggregation rule

At each time index t, a sequential convex aggregation rule produces a convex weight vector $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$ based on the past consumptions $y_1, \ldots, y_{t-1}$ and the past and present forecasts $f_{j,s}$, for all $s = 1, \ldots, t$ and $j \in E_s$. By convex weight vector, we mean a vector $\mathbf{p}_t \in \mathbb{R}^N$ such that $p_{j,t} \geqslant 0$ for all $j = 1, \ldots, N$ and $p_{1,t} + \ldots + p_{N,t} = 1$; we denote by $\mathcal{X}$ the set of all these convex weight vectors over N elements.

The final prediction at t is then obtained by linearly combining the predictions of the models according to the weights given by the components of the vector $\mathbf{p}_t$. More precisely, the aggregated prediction at time index t equals

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

(we recall that by convention, $f_{j,t} = 0$ if $j \notin E_t$).

2.2 Assessment of the quality of a sequential convex aggregation rule

To measure the accuracy of the prediction $\widehat{y}_t$ proposed at round t for the electricity consumption y_t we introduce a loss function $\ell : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$. In our experiments we used

the square loss, defined as $\ell(x,y) = (x-y)^2$ for all $x, y \in \mathbb{R}_+$. A popular criterion to assess the quality of a sequential aggregation rule $\mathcal{A}$ is then given by its root mean square error (RMSE), defined as

$$\text{RMSE}(\mathcal{A}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\widehat{y}_t - y_t\big)^2}\,.$$

We would like to compare the performance of the aggregation rules to the one of the experts or to the one of some simple rules based on the experts.

2.2.1 Comparison to any base expert

A difficulty is however that the RMSE of the j–th expert cannot be defined as the RMSE of the aggregation rule that would predict at each time index as the j–th expert; this is because the latter may be inactive, in which case its associated forecast, defined by convention as $f_{j,t} = 0$, is likely to be poor. We therefore only consider the indexes t where j is active, a fact denoted by $j \in E_t$, and define the RMSE of the j–th expert as

$$\text{RMSE}(j) = \sqrt{\frac{1}{\sum_{t=1}^{T}\mathbb{I}_{\{j\in E_t\}}}\sum_{t=1}^{T}(f_{j,t}-y_t)^2\,\mathbb{I}_{\{j\in E_t\}}}\,. \tag{1}$$

The methodology here is to ensure that the mean error suffered by a rule $\mathcal{A}$ is not much larger than the one of any expert j. However, to provide a fair comparison between the rule $\mathcal{A}$ and the expert j, we compare their performance only on time indexes when j was active, that is, we compare $\text{RMSE}(j)$ to

$$\text{RMSE}(\mathcal{A}, j) = \sqrt{\frac{1}{\sum_{t=1}^{T}\mathbb{I}_{\{j\in E_t\}}}\sum_{t=1}^{T}\big(\widehat{y}_t - y_t\big)^2\,\mathbb{I}_{\{j\in E_t\}}}\,.$$

To do so, we will actually compare their squares and consider a quantity called the regret; formally, the (cumulative) regret of $\mathcal{A}$ with respect to expert j up to T equals

$$R_T(\mathcal{A}, j) = \sum_{t=1}^{T}\Big(\big(\widehat{y}_t - y_t\big)^2 - (f_{j,t}-y_t)^2\Big)\,\mathbb{I}_{\{j\in E_t\}}\,.$$

Our methodology, which is to ensure that the performance of $\mathcal{A}$ is as much as good as any expert j, can then be rephrased as guaranteeing that the regrets $R_T(\mathcal{A}, j)$ are small (i.e., $o(T)$) for all experts j.

Of course, this implies that $\text{RMSE}(\mathcal{A})$ is small as well only if there are experts j which both show a good performance and are active often enough. Indeed, by crudely bounding the loss of the rule $\mathcal{A}$ on the rounds when the comparison expert is inactive, we get

$$\text{RMSE}(\mathcal{A})^2 \leqslant \min_{j=1,\dots,N}\ \frac{\sum_{t=1}^{T}\mathbb{I}_{\{j\in E_t\}}}{T}\big(\text{RMSE}(j)^2 + R_T(\mathcal{A}, j)\big) + B^2\left(1 - \frac{\sum_{t=1}^{T}\mathbb{I}_{\{j\in E_t\}}}{T}\right)\,. \tag{2}$$

2.2.2 Comparison to any fixed combination of experts

We extend the regret methodology by now allowing comparison to fixed convex combinations of the experts. The latter are each parameterized by a convex weight vector $\mathbf{q} \in \mathcal{X}$

and they sequentially aggregate their forecasts based on a renormalization of $\mathbf{q}$ to the set of active experts. Formally, for a set $E \subset \{1, \ldots, N\}$, we define

$$\mathbf{q}(E) = \sum_{j \in E} q_j$$

and denote by $\mathbf{q}^E = (q_1^E, \ldots, q_N^E)$ the following convex weight vector:

$$\mathbf{q}^E = \begin{cases} (0, \ldots, 0) & \text{if } \mathbf{q}(E) = 0; \\ \left(\dfrac{q_1 \mathbb{I}_{\{1 \in E\}}}{\mathbf{q}(E)}, \ldots, \dfrac{q_N \mathbb{I}_{\{N \in E\}}}{\mathbf{q}(E)} \right) & \text{if } \mathbf{q}(E) > 0. \end{cases}$$

Now, the definition (1) can be generalized as follows,

$$\textsc{rmse}(\mathbf{q}) = \sqrt{\frac{1}{\sum_{t=1}^T \mathbf{q}(E_t)} \sum_{t=1}^T \left(\sum_{j \in E_t} q_j^{E_t} f_{j,t} - y_t \right)^2 \mathbf{q}(E_t)}\,.$$

Indeed, when $\mathbf{q} = \delta_j$, the convex weight vector that puts a mass 1 on the j–th component, we recover $\textsc{rmse}(\delta_j) = \textsc{rmse}(j)$.

The notion of cumulative regret of a sequential aggregation rule $\mathcal{A}$ with respect to some fixed weight vector $\mathbf{q}$ up to T can also be generalized as

$$R_T(\mathcal{A}, \mathbf{q}) = \sum_{t=1}^T \left(\left(\widehat{y}_t - y_t \right)^2 - \left(\sum_{j \in E_t} q_j^{E_t} f_{j,t} - y_t \right)^2 \right) \mathbf{q}(E_t)\,.$$

Here also, we have $R_T(\mathcal{A}, \delta_j) = R_T(\mathcal{A}, j)$.

An argument similar to (2) shows that if the regrets $R_T(\mathcal{A}, \mathbf{q})$ can all be guaranteed to be small, then the performance of $\mathcal{A}$, as measured by $\textsc{rmse}(\mathcal{A})$, should be good as well. But of course, such a guarantee would in particular be stronger than (2) itself.

2.2.3 Comparison to any fixed sequence of experts with few shifts

In this third and last definition of regret, we compare the performance of a rule not only to the performance of a fixed expert or a fixed convex combination of the experts, but to a sequence of experts; that is, to prediction methods that would focus on a single expert on each element of a partition of the time indexes $\{1, \ldots, T\}$ into few integer subintervals.

Formally, we denote by $\mathcal{C}$ the set of legal sequences of expert indexes $j_1^T = (j_1, \ldots, j_T)$, where legality means that for all time indexes t, the considered expert j_t is in E_t. We call compound experts the elements of $\mathcal{C}$. For such a compound expert j_1^T, we denote by

$$\text{size}\big(j_1^T\big) = \sum_{t=2}^T \mathbb{I}_{\{j_{t-1} \neq j_t\}}$$

its number of switches (the number minus one of elements in the partition of $\{1, \ldots, T\}$ into integer subintervals corresponding to the use of the same expert). For $0 \leqslant m \leqslant T-1$, we then define $\mathcal{C}_m$ as the subset of $\mathcal{C}$ containing the compound experts with at most m shifts.

The RMSE of a compound expert $j_1^T \in \mathcal{C}$ is defined as

$$\mathrm{RMSE}\big(j_1^T\big) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(f_{j_t,t} - y_t)^2}$$

while the regret of a rule $\mathcal{A}$ with respect to $j_1^T \in \mathcal{C}$ equals

$$R_T\big(\mathcal{A}, j_1^T\big) = \sum_{t=1}^{T}\Big(\big(\widehat{y}_t - y_t\big)^2 - (f_{j_t,t} - y_t)^2\Big)\,.$$

Here, the relationship between the RMSE of a given rule and the ones of the elements of the comparison class formed by $\mathcal{C}_m$, for some m, is clearer; here, we fix an m because, as we will see below, the regret can only be guaranteed to be small if m itself is not too large. Indeed,

$$\mathrm{RMSE}(\mathcal{A})^2 \leqslant \min_{j_1^T \in \mathcal{C}_m} \mathrm{RMSE}\big(j_1^T\big)^2 + R_T\big(\mathcal{A}, j_1^T\big)\,.$$

2.3 Three families of aggregation rules minimizing the regret

We now recall how to ensure that the various notions of regret introduced above can be made uniformly small thanks to some explicit aggregation rules; by uniformity, we mean bounds that hold uniformly over all sequences of consumptions $y_1, \ldots, y_T$. The presented bounds are deterministic in the sense that they do not rely on any stochastic model that would generate the consumptions $y_1, \ldots, y_T$; they hold for all possible sequences of elements in $[0, B]$.

2.3.1 Exponentially weighted average aggregation rules

Basic version The basic version of the exponentially weighted average aggregation rules in this setting relies on a parameter $\eta > 0$ and will thus be denoted by $\mathcal{E}_\eta$. It uses at time index t the convex weight vector $\mathbf{p}_t$ given by

$$p_{j,t} = \frac{e^{\eta R_{t-1}(\mathcal{E}_\eta, j)}\, \mathbb{I}_{\{j \in E_t\}}}{\sum_{k \in E_t} e^{\eta R_{t-1}(\mathcal{E}_\eta, k)}}\,, \tag{3}$$

that is, it only puts mass on the experts j active at round t and does so by performing an exponentially weighted average of their past performance, measured by the regrets $R_{t-1}(\mathcal{E}_\eta, j)$. When $t = 1$, the latter quantity equals 0 by convention, so that $\mathbf{p}_1$ is simply the uniform distribution over E_1.

The following performance bound is a special case of the results presented, e.g., in [Cesa-Bianchi and Lugosi, 2003, Corollary 2 and Section 6.2]. For all sequences of consumptions $y_1, \ldots, y_T$ with values in $[0, B]$, the regret of $\mathcal{E}_\eta$ is bounded as

$$\max_{j=1,\ldots,N} R_T(\mathcal{E}_\eta, j) \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} B^4 T\,. \tag{4}$$

The (theoretically) optimal choice $\eta^\star = \sqrt{(2\ln N)/(B^4 T)}$ leads to the uniform bound $B^2\sqrt{2T\ln N}$ on the regret of $\mathcal{E}_{\eta^\star}$. This choice depends on the horizon T, which is not necessarily known in advance; in the simulation study, we present an on-line calibration technique that deals with this limitation.

Gradient version This version relies on the following convexity inequality: for all convex weight vectors $\mathbf{p}_t$, for all consumptions t, for all expert forecasts $f_{j,t}$,

$$\big(\widehat{y}_t - y_t\big)^2 - (f_{j,t} - y_t)^2 \leqslant 2\left(\widehat{y}_t - y_t\right)\left(\widehat{y}_t - f_{j,t}\right).$$

It leads to the following upper bound on the regret of an aggregation rule with respect to a given expert j up to round T,

$$\begin{aligned} R_T(\mathcal{A}, j) \leqslant \widetilde{R}_T(\mathcal{A}, j) &= 2\sum_{t=1}^{T}\Big(\left(\widehat{y}_t - y_t\right)\left(\widehat{y}_t - f_{j,t}\right)\Big)\,\mathbb{I}_{\{j\in E_t\}} \\ &= 2\sum_{t=1}^{T}\Big(\left(\widehat{y}_t - y_t\right)\widehat{y}_t - \left(\widehat{y}_t - y_t\right) f_{j,t}\Big)\,\mathbb{I}_{\{j\in E_t\}}. \end{aligned}$$

The gradient version of the previous aggregation rule relies also on a parameter $\eta > 0$, is denoted by $\mathcal{E}_\eta^{\mathrm{grad}}$, and aims at minimizing $\widetilde{R}_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, j\big)$. To do so, it mimics the previous rule and uses

$$p_{j,t} = \frac{e^{\eta\widetilde{R}_{t-1}(\mathcal{E}_\eta^{\mathrm{grad}},j)}\,\mathbb{I}_{\{j\in E_t\}}}{\sum_{k\in E_t} e^{\eta\widetilde{R}_{t-1}(\mathcal{E}_\eta^{\mathrm{grad}},k)}}. \tag{5}$$

A straightforward adaptation of the proof of the previous performance bound using the techniques shown in [Cesa-Bianchi and Lugosi, 2006, Section 2.5] (see [Devaine et al., 2009, Section 2.3] for the details) leads to the following guarantee. For all sequences of consumptions $y_1, \ldots, y_T$ with values in $[0, B]$, the regret of $\mathcal{E}_\eta^{\mathrm{grad}}$ is bounded as

$$\max_{j=1,\ldots,N} R_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, j\big) \leqslant \max_{j=1,\ldots,N} \widetilde{R}_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, j\big) \leqslant \frac{\ln N}{\eta} + 2\eta B^4 T.$$

The (theoretically) optimal choice $\eta^\star = \sqrt{(\ln N)/(2B^4T)}$ leads to the uniform bound $2B^2\sqrt{2T\ln N}$ on the regret of $\mathcal{E}_{\eta^\star}^{\mathrm{grad}}$. The same comments as above on the calibration of η apply.

2.3.2 The specialist aggregation rule

The content of this paragraph is extracted from [Freund et al., 1997, Section 3.4], see also [Devaine et al., 2009, Section 2.5]. The aggregation rule described in Figure 1, the specialist aggregation rule, relies on a parameter $\eta > 0$ and will be denoted by $\mathcal{S}_\eta$. It is close to but different from the rule $\mathcal{E}_\eta^{\mathrm{grad}}$.

Its performance guarantee is stronger than the ones of the previous family of aggregation rules, since it is with respect to all fixed convex combinations of the experts. More precisely, for all sequences of consumptions $y_1, \ldots, y_T$ with values in $[0, B]$, the regret of $\mathcal{S}_\eta$ is bounded as

$$\max_{\mathbf{q}\in\mathcal{X}} R_T\big(\mathcal{S}_\eta, \mathbf{q}\big) \leqslant \frac{\ln N}{\eta} + 2\eta B^4 T,$$

where the maximum is taken over all convex weights $\mathbf{q}$ over N elements. The (theoretically) optimal choice $\eta^\star = \sqrt{(\ln N)/(2B^4T)}$ leads to the uniform bound $2B^2\sqrt{2T\ln N}$ on the regret of $\mathcal{S}_{\eta^\star}$. The same comments as above on the calibration of η apply.

Parameters: learning rate $\eta > 0$
Initialization: $p_{i,1} = 1/N$ for $i = 1, \dots, N$
For each time index $t = 1, 2, \dots, T$,

(1) predict $\widehat{y}_t = \sum_{j \in E_t} p_{j,t}^{E_t} \, f_{j,t}$;

(2) observe y_t and compute $\mathbf{p}_{t+1}$ as

$$p_{i,t+1} = \begin{cases} p_{i,t} \, e^{-2\eta(\widehat{y}_t - y_t) f_{i,t}} \dfrac{\sum_{j \in E_t} p_{j,t}}{\sum_{k \in E_t} p_{k,t} \, e^{-2\eta(\widehat{y}_t - y_t) f_{k,t}}} & \text{if } i \in E_t, \\ p_{i,t} & \text{if } i \notin E_t. \end{cases}$$

Figure 1: The specialist aggregation rule of Freund et al. [1997].

2.3.3 Fixed-share aggregation rules

Basic version Here also we will present two versions of the rule, the first one being based on plain expert losses and the second one resorting to a gradient upper bound. The rule presented in Figure 2 is actually nothing but an efficient computation of the rule that would consider all compound experts and perform exponentially weighted averages on them in the spirit of the rule $\mathcal{E}_\eta$ but with a non-uniform starting distribution. We will call it the fixed-share rule; we denote it by $\mathcal{F}_{\eta,\alpha}$ as it depends on two parameters, $\eta > 0$ and $0 \leqslant \alpha \leqslant 1$. This rule is a straightforward adaptation to the setting of specialized experts of the original fixed-share forecaster of Herbster and Warmuth [1998], see also [Cesa-Bianchi and Lugosi, 2006, Section 5.2]. The details of the adaptation and of the proof of the performance bound presented below can be found in [Devaine et al., 2009, Section 2.10].

For all sequences of consumptions $y_1, \dots, y_T$ with values in $[0, B]$, for all $m \in \{0, \dots, n-1\}$, the regret of $\mathcal{F}_{\eta,\alpha}$ with respect to all elements of $\mathcal{C}_m$ is uniformly bounded as

$$\max_{j_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}_{\eta,\alpha}, j_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m \, (1-\alpha)^{T-m-1}} + \frac{\eta}{8} B^4 T \, . \tag{6}$$

The (theoretically almost) optimal bound can be obtained by defining the binary entropy H as $H(x) = x \ln x + (1-x) \ln(1-x)$ for $x \in [0,1]$, by fixing a value of m, and by choosing $\alpha^\star = m/(T-1)$ and

$$\eta^\star = \frac{1}{B^2} \sqrt{\frac{8}{T} \Big((m+1) \ln N + (T-1) H\big(m/(T-1)\big) \Big)} \, ;$$

it is given by

$$\max_{j_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}_{\eta^\star,\alpha^\star}, j_1^T\big) \leqslant B^2 \sqrt{\frac{T}{2} \Big((m+1) \ln N + (T-1) H\big(m/(T-1)\big) \Big)} \, .$$

This optimal upper bound is $o(T)$ as desired as soon as $m = o(T)$; of course, the theoretical optimal choices depend on T and m, so that here also sequential adaptive choices are necessary.

Parameters: $\eta > 0$ and $0 \leqslant \alpha \leqslant 1$
Initialization: $(w_{1,0}, \ldots, w_{N,0}) = \big(\mathbb{I}_{\{1\in E_1\}}, \ldots, \mathbb{I}_{\{1\in E_N\}}\big)$
For each round $t = 1, 2, \ldots, T$,

(1) $\widehat{y}_t = \dfrac{1}{\sum_{k=1}^N w_{k,t-1}} \sum_{j=1}^N w_{j,t-1}\, f_{j,t}$;

(2) [loss update] observe y_t and define for each $i = 1, \ldots, N$,

$$v_{i,t} = \begin{cases} w_{i,t-1}\, e^{\eta(\widehat{\ell}_t - \ell_{i,t})} & \text{if } i \in E_t, \\ \text{undefined} & \text{if } i \notin E_t; \end{cases}$$

(3) [share update] let $w_{i,t} = 0$ if $i \notin E_{t+1}$ and

$$w_{i,t} = \frac{1}{|E_{t+1}|} \sum_{j \in E_t \setminus E_{t+1}} v_{j,t} + \frac{\alpha}{|E_{t+1}|} \sum_{j \in E_t \cap E_{t+1}} v_{j,t} + (1-\alpha)\, \mathbb{I}_{\{i\in E_t \cap E_{t+1}\}}\, v_{i,t}$$

if $i \in E_{t+1}$ (with the convention that an empty sum is null).

Figure 2: The fixed-share aggregation rule $\mathcal{F}_{\eta,\alpha}$ (basic version).

Gradient version We proceed here as in Section 2.3.1 and consider a variant of the previous forecaster that is based on the gradient of the losses rather than on the losses themselves. This variant has the same form as $\mathcal{F}_{\eta,\alpha}$ and will be denoted by $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$.

The only modification to be performed in Figure 2 to define the gradient version is to replace the update in step (2) by

$$v_{i,t} = w_{i,t-1}\, e^{-2\eta(\widehat{y}_t - y_t)(\widehat{y}_t - f_{i,t})}\,.$$

The theoretical bounds of $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$ are the same of the ones for $\mathcal{F}_{\eta,\alpha}$ up to the replacement of the bounds B^2 on the losses by the bound $2B^2$ on the gradient of the losses. That is, for all sequences of consumptions $y_1, \ldots, y_T$ with values in $[0, B]$, for all $m \in \{0, \ldots, n-1\}$, the regret of $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$ with respect to all elements of $\mathcal{C}_m$ is uniformly bounded as

$$\max_{j_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}_{\eta,\alpha}, j_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m\, (1-\alpha)^{T-m-1}} + \frac{\eta}{2} B^4 T\,,$$

which can be (almost) optimized, via suitable choices $\eta^\star$ and $\alpha^\star$ for η and α, as

$$\max_{j_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}_{\eta^\star,\alpha^\star}, j_1^T\big) \leqslant 2B^2 \sqrt{\frac{T}{2}\Big((m+1)\ln N + (T-1) H\big(m/(T-1)\big)\Big)}\,.$$

Comments The fixed-share aggregation rules update the convex combinations they use in two steps, a loss update and a share update, as indicated in Figure 2. The loss update follows the logic behind the exponentially weighted average aggregation rules, where experts are weighted according to their past performance through an exponential reweighting. The share update redistributes the weights over the active experts and ensures that

Time intervals	Only 11:00–12:00
Number of days	1 095
Time indexes T	1 095
Number of experts N	35
Median of the y_t	702.6
Bound B on the y_t	1020.0

Table 1: Some characteristics of the observations y_t of the Slovakian data set for the time intervals 11:00–12:00.

each of them is played with a sufficient probability; this is the key for the rule to be competitive with respect to compound experts.

Note also that compound experts in our non-stochastic setting can be related to breaks in a sequence of stochastic observations in a more classical statistical framework where the observations are the realizations of some underlying stochastic process whose parameters can changer over time.

3 A first data set: Slovakian data

3.1 Presentation and characteristics of the data set

The data set concerns the consumption encountered by the Slovakian subbranch of the French provider EDF. It is formed by the hourly predictions of 35 experts and the corresponding observations on the period from January 1, 2005 to December 31, 2007. That is, there are 24 series (one for each hour) of 1 095 observations and of at most $35 \times 1\,095 = 38\,325$ expert predictions. Actually, there are fewer such predictions since some of the experts are specialized and were not able to deliver a prediction at all time indexes. In this part and unlike for the French data set of the next part, we have absolutely no information on how the experts were built and we merely consider them as black boxes. The observations are given by the hourly mean consumptions.

The reason why we parsed the data set into 24 subsets is that the behavior of electricity consumption depends heavily on the hour (much more than on the given day in the week); for instance, the consumption is low at nights and some peaks can be observed during the day, e.g., around 19:00.

We therefore ran 24 parallel aggregation rules, one for each fixed hour interval. We report in this section the characteristics of the data and the results of several aggregation rules for the interval 11:00–12:00. The observed mean consumptions are plotted in Figure 3.

3.1.1 Characteristics of the data set

The characteristics of the observations y_t are described in Table 1. In this section, we will omit the unit MW (megawatt) of the observations and predictions of the electricity consumption, as well as the one of their corresponding RMSE.

The characteristics of the experts are depicted in Figure 4. The bar plot represents the values of the RMSE of the 35 available experts; we computed the 35 values of RMSE(j), one for each expert j, and ordered them. The scatter plot relates the RMSE of each of the

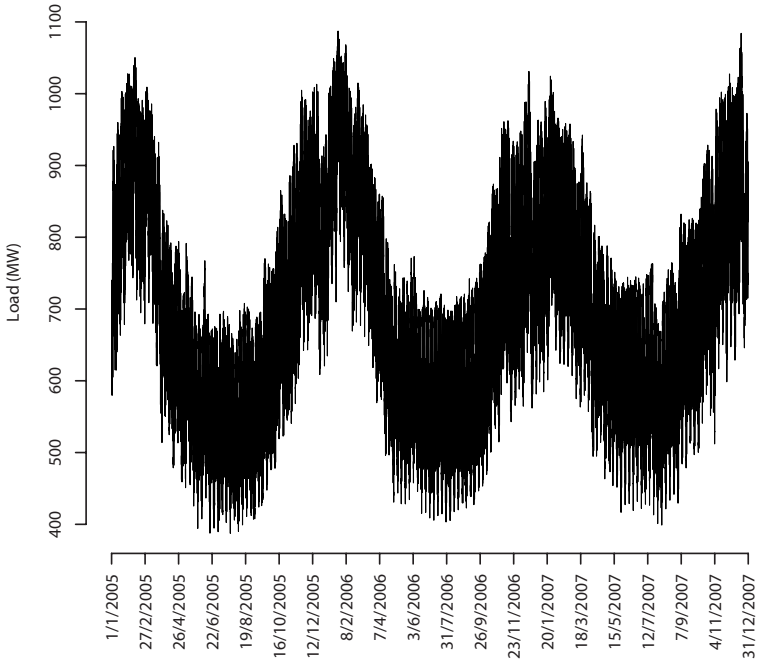

Figure 3: The observed hourly electricity consumptions encountered by the Slovakian subbranch between January 1, 2005 and December 31, 2007.

expert to its frequency of activity, that is, it plots the pairs

$$\left(\text{RMSE}(j),\ \frac{\sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}}}{T} \right) \tag{7}$$

for all experts j.

3.1.2 Benchmarking values: performance of some oracles

We present in Table 2 the values[1] of the RMSE of several procedures that can work off-line using the whole data set (i.e., observations and predictions) and are only constrained by the fact that at each round they need to output as a prediction a convex combination of the predictions of the experts. None of them, except the use of the uniform convex weight vector in $\mathcal{X}$ and the uniform sequential aggregation rule $\mathcal{U}$, can be implemented sequentially, and this is why they are called oracles.

The rule $\mathcal{U}$ simply chooses, at each time index t, the uniform convex weight on E_t. Its RMSE differs from the one of the uniform convex weight vector $(1/35, \ldots, 1/35)$, as the

[1] All of them have been computed exactly, except the ones that involve minimizations over simplexes of convex weights, for which a Monte-Carlo stochastic approximation method was used.

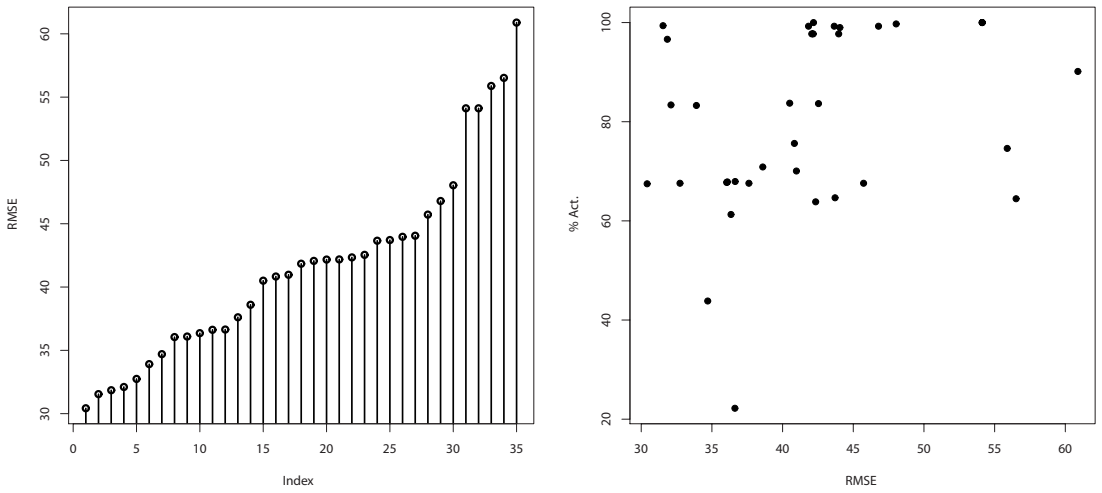

Figure 4: Graphical representations of the performance of the experts of the Slovakian data set: sorted RMSE (left) and RMSE–frequency of activity pairs (right).

general definitions instantiate here as

$$\text{RMSE}(\mathcal{U}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in E_t} f_{j,t}}{|E_t|} - y_t\right)^2}$$

$$\text{and} \qquad \text{RMSE}\big((1/35,\dots,1/35)\big) = \sqrt{\frac{1}{\sum_{t=1}^{T}|E_t|}\sum_{t=1}^{n}|E_t|\left(\frac{\sum_{j\in E_t} f_{j,t}}{|E_t|} - y_t\right)^2}.$$

The oracle with the least error is the one that can pick at each time index the expert that will perform best. This oracle corresponds to the choice of the best compound expert with size at most $T-1$; it suffers a non-zero RMSE of 9.4 since none of the $f_{j,t}$ is exactly equal to the observation y_t to come. This oracle indicates a lower bound on the best performance that can be achieved by a sequential aggregation rule using the experts predictions. Of course, this lower bound is very optimistic.

More reasonable comparison points are given, on the one hand, by best compound experts of smaller[2] size m, and on the other hand, by the individual performance of some fixed convex weight vectors (not necessarily evaluated on all time indexes): the uniform weight vector, which is the best a priori constant choice of a weight vector, and two oracle weight vectors, the best single expert in hindsight and the best fixed convex weight vector in hindsight.

Remark 1. The fact that the RMSE of the best compound expert with size at most 10 is larger than the RMSE of the best single expert is explained by the fact that some overall good experts may refrain from predicting at some time indexes when they know that the prediction will be difficult to them; this is a side-effect of the specialist framework. On the contrary, compound experts are required to output a prediction at each time index even when they guess that an accurate prediction will be difficult to perform.

We also wanted to assess whether gains in performance could be hoped for by partitioning time into subsets of indexes with constant sets of active experts; that is, by

[2] $m = 1$ would have been a suitable value since two experts are active at all time indexes; it however leads to a bad performance, which explains why we considered the minimal value of $m = 10$.

Name of the benchmark procedure	Formula	Value
Uniform sequential aggregation rule	$\textsc{rmse}(\mathcal{U})$	$= 31.1$
Uniform convex weight vector	$\textsc{rmse}\big((1/35,\dots,1/35)\big)$	$= 30.7$
Best single expert	$\min_{j=1,\dots,35} \textsc{rmse}(j)$	$= 30.4$
Best convex weight vector	$\min_{\mathbf{q}\in\mathcal{X}} \textsc{rmse}(\mathbf{q})$	$= 29.2$
Best compound expert		
Size at most $m = 10$	$\min_{j_1^T\in\mathcal{C}_{10}} \textsc{rmse}\big(j_1^T\big)$	$= 32.1$
Size at most $m = 50$	$\min_{j_1^T\in\mathcal{C}_{50}} \textsc{rmse}\big(j_1^T\big)$	$= 23.1$
Size at most $m = 200$	$\min_{j_1^T\in\mathcal{C}_{200}} \textsc{rmse}\big(j_1^T\big)$	$= 15.2$
Size at most $m = T - 1 = 1\,094$	$\min_{j_1^T\in E_1\times E_2\times\dots\times E_T} \textsc{rmse}\big(j_1^T\big)$	$= 9.4$
On the $K = 74$ elements of a partition of time according to the values of the active sets E_t		
Best expert on each element	See (8)	$= 29.1$
Best convex weight vector on each element	See (9)	$= 24.5$

Table 2: Definition and performance of several (possibly off-line) benchmarking procedures on the Slovakian data set; they serve as comparison points for on-line procedures.

defining

$$\big\{E^{(1)},\dots,E^{(K)}\big\} = \big\{E_t,\ t\in\{1,\dots,T\}\big\}$$

and by partitioning time according to the values $E^{(k)}$ taken by the sets of active experts E_t. The corresponding natural oracles are

$$\min\left\{\sqrt{\frac{1}{T}\sum_{k=1}^{K}\sum_{t:E_t=E^{(k)}}\big(f_{j^k,t}-y_t\big)^2},\ \text{ with } j^k\in E^{(k)}\ \text{ for all } k=1,\dots,K\right\}, \tag{8}$$

which corresponds to the choice of the best expert on each element of the partition, and

$$\min\left\{\sqrt{\frac{1}{T}\sum_{k=1}^{K}\sum_{t:E_t=E^{(k)}}\left(\sum_{j\in E^{(k)}}q_j^{(k)}f_{j,t}-y_t\right)^2},\right.$$
$$\left.\text{with } \mathbf{q}^{(k)} \text{ a convex weight vector on } E^{(k)}\ \text{ for all } k=1,\dots,K\right\}, \tag{9}$$

which corresponds to the choice of the best convex weight vector on each element of the partition. Despite the fact that there are relatively many elements in this partition,

Value of	η	10^{-8}	10^{-7}	10^{-6}	4×10^{-6}	10^{-5}	10^{-4}	10^{-3}
RMSE of	$\mathcal{E}_\eta$	31.3	31.2	30.8	<u>30.5</u>	30.9	32.7	
	$\mathcal{E}_\eta^{\mathrm{grad}}$		31.3	30.9		29.8	<u>28.2</u>	33.5
	$\mathcal{S}_\eta$		31.3	30.9		29.8	<u>28.2</u>	34.7

Table 3: Performance obtained by the sequential aggregation rules $\mathcal{E}_\eta$, $\mathcal{E}_\eta^{\mathrm{grad}}$, and $\mathcal{S}_\eta$ for various choices of η; the smallest value for each rule is underlined.

namely, $K = 74$, the gain with respect to constant choices throughout time exists (RMSE of 29.1 versus 30.4 and 24.5 versus 29.2) but is less significant than the one achieved with compound experts (which achieve a smaller RMSE of 23.1 already with a size $m = 50$).

3.2 Results obtained by the considered sequential aggregation rules

We now detail the practical performance of the introduced sequential aggregation rules and compare it to the one of the oracles. We will proceed in two steps.

First, we report the results obtained for fixed values of the parameters η and α of the rules; to do so, we considered a grid of the possible values and computed the RMSE for each value of the parameters. We report for each rule the best performance obtained; the corresponding parameters are said the best constant choices in hindsight. This assesses the potential performance of the rules but does not lead to fully sequential procedures yet.

We then explain how fully sequential procedures can get performance close to the one obtained by these best constant choices in hindsight; as we will see, these procedures calibrate on-line the parameters by running several instances of the base sequential aggregation rules.

3.2.1 Performance with constant values of the parameters

The performance of $\mathcal{E}_\eta$, $\mathcal{E}_\eta^{\mathrm{grad}}$, and $\mathcal{S}_\eta$ is summarized in Table 3. As indicated in Section 2.3, they should be compared, respectively, to the performance of the best single expert (for $\mathcal{E}_\eta$) and to the one of the best convex weight vector (for $\mathcal{E}_\eta^{\mathrm{grad}}$ and $\mathcal{S}_\eta$). We recall that these are indicated in Table 2. We note that $\mathcal{E}_\eta^{\mathrm{grad}}$ and $\mathcal{S}_\eta$, when tuned with the best parameter η in hindsight, outperform their comparison oracle, the best convex weight vector (with a relative improvement of 3 % in terms of the RMSE), while the performance of the best $\mathcal{E}_\eta$ comes very close to the one of the best single expert (RMSE of 30.4 versus 30.5).

Here, as in Mallet et al. [2009], the best constant choices in hindsight are far away from the theoretically optimal ones, given by $\eta^\star \approx 8 \times 10^{-8}$ for the $\mathcal{E}_\eta$ and $\eta^\star \approx 4 \times 10^{-8}$ for $\mathcal{E}_\eta^{\mathrm{grad}}$ and $\mathcal{S}_\eta$. The computation of these values of $\eta^\star$ however served us to set the grid used in Table 3; we started basically at $\eta^\star$ and then performed logarithmic increments.

The performance of the fixed-share type rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ is reported in Table 4. Here, it is trickier to speak of a specific comparison class and to compute the values of theoretically almost optimal parameters (the choice of m is crucial for these issues). The most important remark is thus probably that these rules, when tuned properly (in hindsight), improve on the already good results of the previously cited rules. Of course, this might be because these methods are a bit more flexible, since they rely on two parameters instead of one.

We close this preliminary review of performance by showing in Figure 5 that the considered rules fully exploit the whole set of experts and do not concentrate on a limited

Value of	η	10^{-4}	10^{-4}	10^{-3}	10^{-3}	10^{-2}	10^{-2}	2×10^{-4}	2×10^{-3}
	α	0.05	0.2	0.1	0.2	0.05	0.2	0.07	0.2
RMSE of	$\mathcal{F}_{\eta,\alpha}$	29.3	29.5	27.5	27.2	28.0	27.8		<u>27.0</u>
	$\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$	28.0	28.9	29.3	29.2	28.7	28.5	<u>27.2</u>	

Table 4: Performance obtained by the sequential aggregation rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$ for various choices of η and α; the smallest value for each rule is underlined.

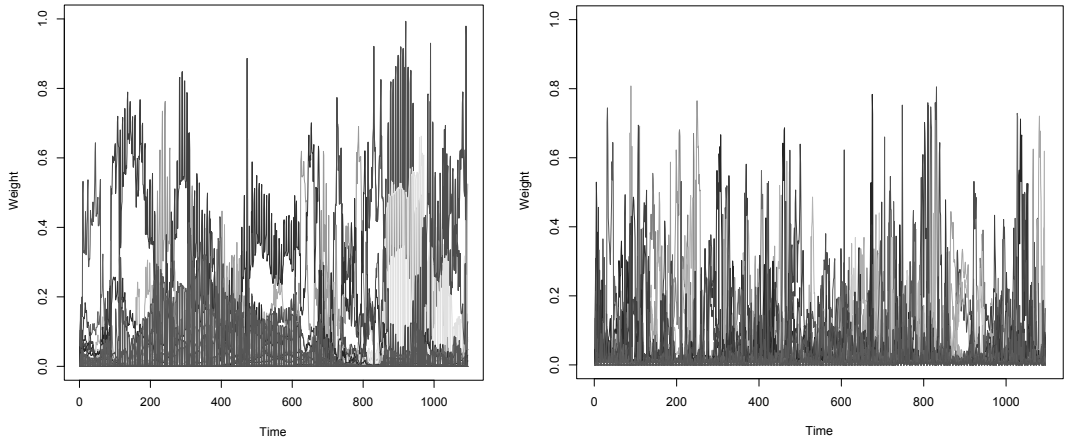

Figure 5: Graphical representations of the convex weights associated at each time index with the 35 experts by $\mathcal{E}^{\mathrm{grad}}_{10^{-4}}$ (left) and $\mathcal{F}_{2\times 10^{-3},0.2}$ (right).

subset of the experts. They carefully adapt their convex weights as time evolves and remain reactive to changes of performance; in particular, the sequences of these weights do not converge to a limit vector.

3.2.2 On-line calibration of the parameters

We explain in this section how fully sequential aggregation rules can get performance close to the one of the best constant choices of the parameters in hindsight studied above.

We describe first the method in a general framework. Let $\mathcal{A}_\lambda$ be a sequential aggregation rule relying on some real parameter λ (possibly vector-valued) taking its values in some set Λ. Given the past observations and the past and present forecasts of the experts, it prescribes at time index t a convex weight vector which we denote by $\mathbf{p}_t\big(\mathcal{A}_\lambda\big)$. As illustrated above, a crucial issue is to find a suitable value of λ. Since no obvious a priori choice is available (the optimal values to minimize the bounds on the regret showing poor practical performance), we will resort to the following method, due to Vivien Mallet and proposed in the technical report Gerchinovitz et al. [2008] but never published elsewhere.

On-line calibration of the parameters We describe here the weights $\widehat{\mathbf{p}}_t$ used by a meta-sequential aggregation rule based on the family of rules $\mathcal{A}_\lambda$, where $\lambda \in \Lambda$. We assume that the considered family is such that $\mathbf{p}_1\big(\mathcal{A}_\lambda\big)$ is independent of λ, so that $\widehat{\mathbf{p}}_1$ equals this

common value. Then, at time indexes $t \geqslant 2$,

$$\widehat{\mathbf{p}}_t = \mathbf{p}_t\big(\mathcal{A}_{\widehat{\lambda}_{t-1}}\big) \qquad \text{where} \qquad \widehat{\lambda}_{t-1} \in \operatorname*{argmin}_{\lambda\in\Lambda} \sum_{s=1}^{t-1} \left(\sum_{j\in E_s} p_{j,s}\big(\mathcal{A}_\lambda\big)\, f_{j,s} - y_s \right)^2 ; \qquad (10)$$

that is, we consider, for the prediction of the next time index, the aggregated forecast proposed by the best so far member of the family of aggregation rules.

Computationally speaking, we need to run in parallel all the instances of $\mathcal{A}_\lambda$, together with the meta-rule. This of course is impossible as soon as Λ is not finite; for the families considered above we had $\Lambda = (0, +\infty)$ (exponentially weighted average rules) and $\Lambda = (0, +\infty) \times [0, 1]$ (fixed-share type rules). This is why, in practice, we only consider a finite grid $\widetilde{\Lambda}$ over Λ and perform the minimization of (10) only on the elements of $\widetilde{\Lambda}$ instead of performing it on the whole set Λ.

Some choices are still left to the user, namely, how to design this grid $\widetilde{\Lambda}$ and thus, the proposed procedure is not fully automatic yet. We checked (see [Devaine et al., 2009, Section 3.1] for the details) however that the performance was not too sensitive to the design of $\widetilde{\Lambda}$. A reasonable rule is, e.g., to start the grid around $\lambda^\star$, the optimal value prescribed by theory, and then take logarithmically spaced points till a given upper bound. Below, in the case of the tuning of the parameter η of the exponentially weighted average rules, we take the upper bound 1. A way to set this upper bound is explained in Section 4.2.4; for the reader to fully appreciate it we however need first to illustrate several times the on-line calibration of the parameters with finite grids and this is why we consider some seemingly arbitrary grids for the time being.

In what follows, we do not report the results obtained with the family of the specialist aggregation rules $\mathcal{S}_\eta$ because these were very similar to the ones obtained by the exponentially weighted average rules, while the latter are more easily implemented.

Application to the exponentially weighted average rules $\mathcal{E}_\eta$ and $\mathcal{E}_\eta^{\text{grad}}$ Following the methodology described above and the order of magnitude of the optimal values $\eta^\star$ being around 10^{-8}, we considered two finite grids for the tuning of η, both with endpoints 10^{-8} and 1: a smaller grid, with 9 equally logarithmically spaced points,

$$\widetilde{\Lambda}_s = \big\{10^{-k},\ \text{for } k \in \{0, 1, \ldots, 8\}\big\}\,,$$

and a larger grid, with 25 equally logarithmically spaced points,

$$\widetilde{\Lambda}_\ell = \big\{m \times 10^{-k},\ \text{for } k \in \{1, \ldots, 8\} \text{ and } m \in \{1,\, 2.5,\, 5\}\big\} \cup \{1\}\,.$$

The performance on these grids with respect to the best constant choice of η in hindsight (as discussed in Table 3) is summarized in Table 5. We note that the good performance obtained for the best choices of the parameters in hindsight is preserved by the adaptive meta-rules resorting to the grids. The sequences of choices of η on the largest grid $\widetilde{\Lambda}_\ell$ are depicted in Figure 6.

Application to the fixed-share type rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}^{\text{grad}}$ Two parameters have to be tuned here; we take a finite grid in $\Lambda = (0, +\infty) \times [0, 1]$, e.g., following the methodology above,

$$\widetilde{\Lambda}_{\text{FS}} = \big\{(10^{-k},\, \alpha),\ \text{for } k \in \{0, 1, \ldots, 8\} \text{ and } \alpha \in \{0.01,\, 0.05,\, 0.1,\, 0.2,\, 0.3,\, 0.4\}\big\}\,.$$

		Best constant η	Grid $\widetilde{\Lambda}_s$	Grid $\widetilde{\Lambda}_\ell$
RMSE of	$\mathcal{E}_\eta$	30.5	31.1	30.7
	$\mathcal{E}_\eta^{\mathrm{grad}}$	28.2	28.2	28.4

Table 5: Performance obtained by the rules $\mathcal{E}_\eta$ and $\mathcal{E}_\eta^{\mathrm{grad}}$ for the best constant choice of η in hindsight (left) and when used as keystones of a meta-rule selecting sequentially the values of η on the chosen grids (middle and right).

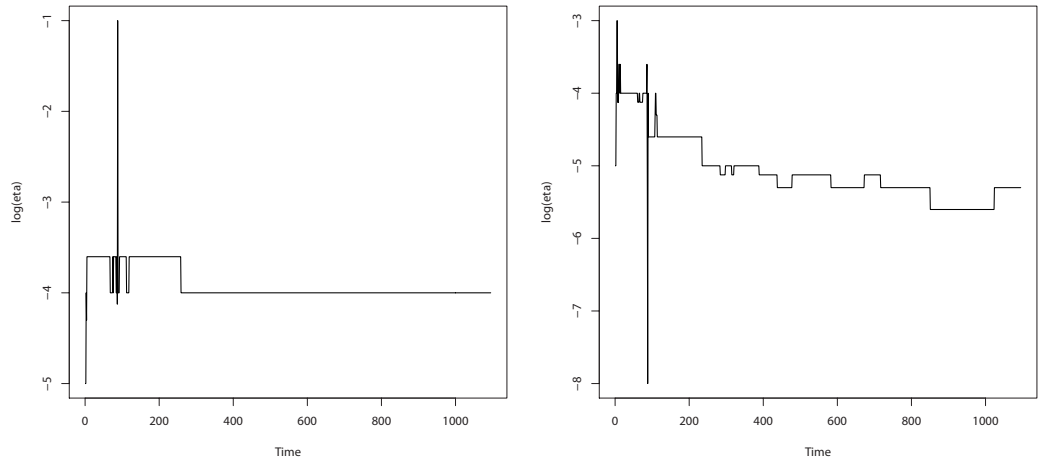

Figure 6: Graphical representations of the sequences of tuning parameters η chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_\ell$; the base rules are $\mathcal{E}_\eta^{\mathrm{grad}}$ (left) and $\mathcal{E}_\eta$ (right).

The performance on this grid with respect to the best constant choices of η and α in hindsight (as discussed in Table 4) is summarized in Table 6. Here also, we note that the good performance obtained for the best choices of the parameters in hindsight is preserved by the adaptive meta-rules resorting to the grids. The sequences of choices of η and α on the grid $\widetilde{\Lambda}_{\mathrm{FS}}$ for the meta-rule based on $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ are depicted in Figure 7.

4 A second data set: Operational forecasting on French data

4.1 Presentation and characteristics of the data set

4.1.1 Characteristics of the data set

The data set used in this part is the standard data set used for the calibration of the EDF short-term models for the French electricity load. It includes half-hourly electricity data and meteorological observations (temperature and cloud cover) throughout the French territory. Load data are built by EDF from the French load data measured and provided by the French national grid company, RTE ("Réseau de transport d'électricité"). Meteorological data is issued by the French weather-forecasting institution Météo-France.

This data set is divided into two parts: the first part ranges from September 1, 2002 to August 31, 2007 – we call it the estimation set; the second part covers the period from September 1, 2007 to August 31, 2008 – we call it the validation set. The experts we consider in this part are trained over the estimation set and then provide base forecasts

		Best constant pair (η, α)	Grid $\widetilde{\Lambda}_{\mathrm{FS}}$
RMSE of	$\mathcal{F}_{\eta,\alpha}$	27.0	27.8
	$\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$	27.2	28.5

Table 6: Performance obtained by the rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$ for the best constant choices of η and α in hindsight (left) and when used as keystones of a meta-rule selecting sequentially the values of η and α on the grid $\widetilde{\Lambda}_{\mathrm{FS}}$ (right).

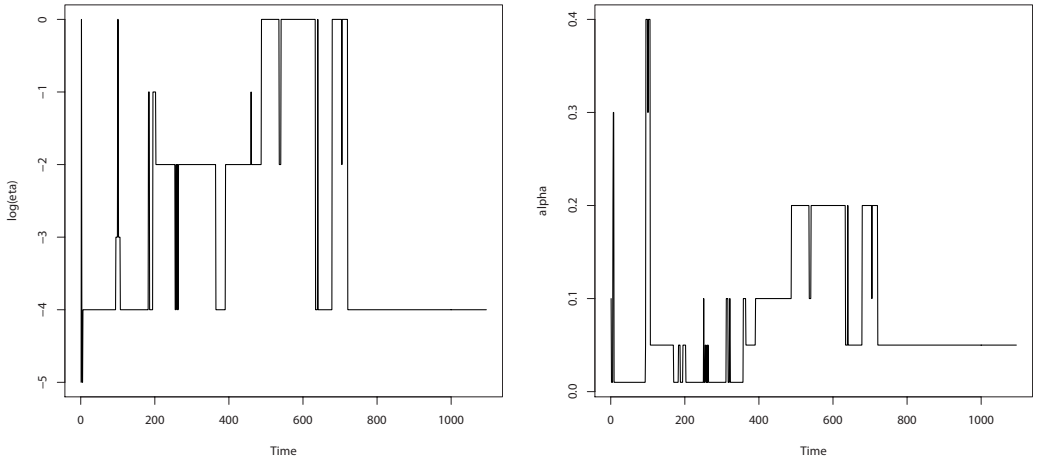

Figure 7: Graphical representations of the sequences of tuning parameters η (left) and α (right) chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_{\mathrm{FS}}$; the base rule is $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$.

throughout the period corresponding to the validation set, which we aggregate. Actually, we exclude some special days from the validation set. Out of the 366 days between September 1, 2007 and August 31, 2008, we keep 320 days. The excluded days correspond to public holidays (the day itself, as well as the days before and after it), daylight saving days and winter holidays (that is, the period between December 21, 2007 and January 4, 2008); however, we include the summer break (August 2008) in our analysis as we have access to experts that are able to produce forecasts for this period. Other particular days exist and correspond to temporary changes of the fare prices in order to reduce expected high consumption (mainly due to low temperature); they are included in the validation set whenever a preprocessing based on EDF commercial data was available. For a more detailed description of this data set we refer the interested reader to Dordonnat et al. [2008].

The characteristics of the observations on the validation set y_t are described in Table 7. In this part as well, we omit the unit GW (gigawatt) of the observations and predictions of the electricity consumption, as well as the one of their corresponding RMSE.

The experts we consider here come from three main categories of statistical models: parametric, semi-parametric, and non-parametric models. The reason for this choice is two-fold: first, we believe that combining base forecasters is particularly useful when they are heterogenous and exhibit sensibly different behaviors; and second, EDF could provide these three types of models. We provide below a short description of them but refer the reader to Devaine et al. [2009] for more details.

Time intervals	Every 30 minutes
Number of days D	320
Time indexes T	15 360
Number of experts N	24 $(= 15 + 8 + 1)$
Median of the y_t	56.33
Bound B on the y_t	92.76

Table 7: Some characteristics of the observations y_t of the French data set of operational forecasting.

The parametric model used to generate the first group of experts is described in Bruhns et al. [2005] and is implemented in an EDF software called "Eventail." We mention briefly that this model is based on a nonlinear regression approach that consists in decomposing the electricity load into a main component including all the seasonality of the process together with a weather-dependant component. To this nonlinear regression model is added an autoregressive correction of the error of the short-term forecasts of the last seven days. Changing the parameters (the gradient of the temperature, the short-term correction) of this model, we derive 15 experts. For conciseness we refer to them as the Eventail experts.

The second group of experts comes from a generalized additive model (henceforth referred to as the GAM model) implemented in the software `R` by the `mgcv` package developed by Wood [2006]. This model is presented in Pierrot et al. [2009] and imports the idea of the parametric modeling presented above into a semi-parametric modeling. One of the key advantages of this model is its ability to adapt to changes in consumption habits where parametric models like Eventail need some a priori knowledge on customers behaviors. Here again, we derive different experts from the GAM model by changing the trend extrapolation effect (which accounts for the yearly economic growth) or the short-term effects like the one-day-lag effect; these changes affect the reactivity to changes along the run. Doing so, we obtained 8 experts, which we call the GAM experts.

The last expert is drastically different from the two previous groups of experts as it relies on a univariate method (i.e., not requiring any exogenous factor like weather conditions). The method is presented in Antoniadis et al. [2006] and Antoniadis et al. [2010]. Its key idea is to assume that the load is driven by an underlying stochastic curve and to see each day as a discrete recording of this functional process. Forecasts are then performed according to a similarity measure between days. We call this expert the similarity expert.

The characteristics of the experts presented above are depicted in Figure 8. Similarly to the study of Slovakian data, the bar plot represents the (sorted) values of the RMSE of the 24 available experts. The scatter plot relates the RMSE of each of the expert to its frequency of activity, that is, it plots the pairs indicated in (7).

Out of the 15 Eventail experts, 3 are active all the time; they correspond to the operational model and two variants on it based on different short-term corrections. The other 12 Eventail experts are sleeping during the summer as their predictions are redundant (they were obtained by changing the gradient of the temperature for the heating part of the load consumption, which generates differences to the operational model in winter only). GAM expert are active on an overwhelming fraction of the time and are sleeping only

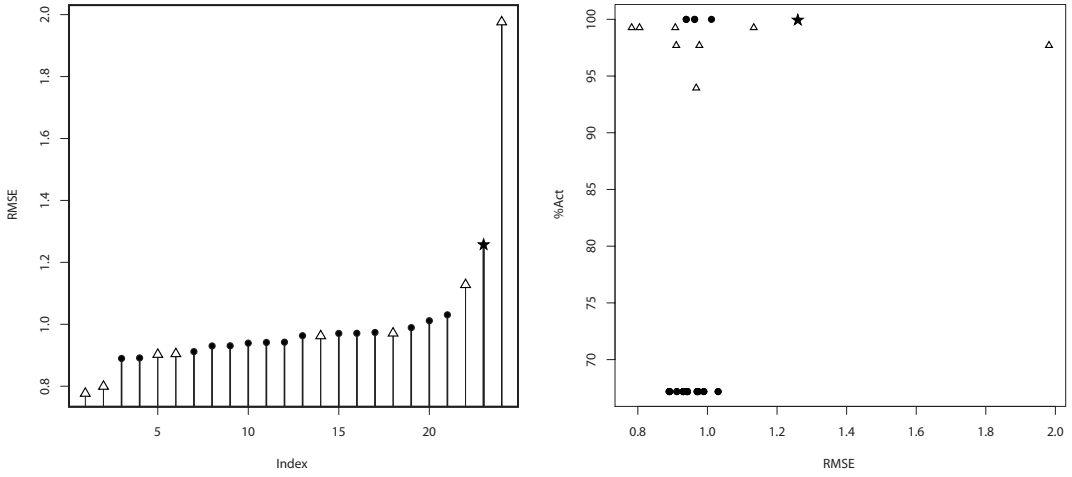

Figure 8: Graphical representations of the performance of the experts of the French data set of operational forecasting: sorted RMSE (left) and RMSE–frequency of activity pairs (right); Eventail experts are depicted by the symbols •, GAM experts are represented by △ while ⋆ stands for the similarity expert.

during periods when practitioners know they will perform poorly (e.g., close to public holidays); the lengths of these periods depend on the parameters of the expert. Finally, the similarity expert is always active.

4.1.2 An operational constraint

The operational constraint consists in this part in producing half-hourly forecasts every day at 12:00 for the next 24 hours. In this respect, all models presented above are also based on an half-hourly decomposition of the data (excepted the functional similarity estimation procedure).

Of course, we could mimic the methodology indicated in Section 3.1 and decompose the French data set into 48 smaller data sets of equal size. However, Table 7 indicates that doing so, this common size would be of 320 observations, which is rather small. Hence, we are not willing this time to decompose the forecasting problem into parallel sub-forecasting procedures and require that the proposed aggregation rules only output predictions (i.e., weight vectors) every 48 steps but when doing so, give 48 such predictions. Section 4.2.1 explains the modifications that need to be performed for the rules described in Section 2.3 to abide by this constraint.

4.1.3 Benchmarking values: performance of some oracles

Before doing so, we report in Table 8 the performance obtained by most of the oracles described in Section 3.1.2. We do not report here the performance obtained by considering partitions of the time in terms of the values of the active sets E_t, as, on the one hand, the study of Section 3.1.2 showed that even when the number of elements K in the partition was large, the compound experts had better performance, and on the other hand, the value of K is small here ($K = 7$); these two facts explain that the performance of the oracles based on partitions is to be poor on this data set.

We note the disappointing performance of the best single expert with respect to the naive rule $\mathcal{U}$. This comes here from our experts being more active in challenging situations, as can be seen as follows. The rule $\mathcal{U}$ also performs better than the uniform convex weight

Name of the benchmark procedure	Formula	Value
Uniform sequential aggregation rule	$\textsc{rmse}(\mathcal{U})$	$= 0.724$
Uniform convex weight vector	$\textsc{rmse}\big((1/24,\dots,1/24)\big)$	$= 0.748$
Best single expert	$\min_{j=1,\dots,24} \textsc{rmse}(j)$	$= 0.782$
Best convex weight vector	$\min_{\mathbf{q}\in\mathcal{X}} \textsc{rmse}(\mathbf{q})$	$= 0.683$
Best compound expert		
Size at most $m = 50$	$\min_{j_1^T\in\mathcal{C}_{50}} \textsc{rmse}\big(j_1^T\big)$	$= 0.534$
Size at most $m = 100$	$\min_{j_1^T\in\mathcal{C}_{100}} \textsc{rmse}\big(j_1^T\big)$	$= 0.474$
Size at most $m = T - 1 = 15\,359$	$\min_{j_1^T\in E_1\times E_2\times\dots\times E_T} \textsc{rmse}\big(j_1^T\big)$	$= 0.223$

Table 8: Definition and performance of several (possibly off-line) benchmarking procedures on the French data set of operational forecasting; they serve as comparison points for on-line procedures.

vector, which induces at each time index the same forecast as the rule $\mathcal{U}$ but for which the losses incurred at time indexes is more weighted as more experts are active. Thus, this is because the considered specialized experts are more active and more helpful when needed.

From Table 8 we mostly conclude the following. The true benchmarking values from the first part of the table are the RMSE of the rule $\mathcal{U}$ – that all fancy rules have to outperform to be considered worth the trouble – and the RMSE of the best convex weight vector.

4.2 Results obtained by the sequential aggregation rules

4.2.1 Extension of the previous rules to operational forecasting

Before describing the detailed performance of the sequential aggregation rules (among which we consider only here the families of exponentially weighted average and fixed-share type rules), we need to describe how we extended them so as to deal with the constraint that predictions need to be output for all time intervals of the next 24 hours, i.e., for the next 48 time indexes. (In the setting of Slovakian data, forecasts also needed to be made 24 hours ahead of time but this constraint could be somewhat discarded by running in parallel a rule per time interval.)

The high-level idea is to run the original rules on the data (called below the base rules), access to the proposed convex weight vectors only at time indexes of the form $t_k = 48k+1$, and use these vectors for the next 48 time indexes, by adapting them via a renormalization or a mixing to the values of the active sets $E_{t_k+1},\dots,E_{t_k+48}$.

We do so to be able to guarantee theoretical bounds on the regret. Indeed, as will be clear from the algorithmic statements of the extensions the weights output by the base rules are, for all t, not too far from the adaptations that have to be made of them (and of course, coincide at the time indexes t_k). This is because in the studied rules a fixed number of losses, namely the ones between the last t_k and the current index t, count much

less than the past losses (the ones encountered between the indexes 1 and $t_k - 1$).

We also propose another extension, which is related to the structure of the set of experts. The latter are of three different types and experts of the same type are obtained as variants of a given prediction method (GAM, Eventail, or functional similarity estimation). It would be fair to allocate an initial weight of 1/3 to the group of GAM experts, which turns into an initial weight of 1/24 to each of the 8 GAM expert; a weight of 1/3 to the group formed by the 15 Eventail experts, that is, an initial weight of 1/45 to each of them; and an initial weight of 1/3 to the similarity expert.

We denote by $p_{j,0}$ the initial weight of an expert j. We will call fair initial weights the convex weight vector described above (with components equal to 1/3, 1/24, or 1/45) and uniform initial weights the vector defined by $p_{j,0} = 1/24$ for all experts j. The effect of this on the regret bounds, e.g., (4) or (6), is the replacement of $\ln N$ by $\max_j \ln 1/p_{j,0}$. This does not change the order of magnitude in T of the regret bounds but only increases them by a multiplicative factor.

Adaptation of the exponentially weighted average rules We will denote the adaptations of the rules of Section 2.3.1 by $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\text{grad}}$ to distinguish them from the base versions $\mathcal{E}_\eta$ and $\mathcal{E}_\eta^{\text{grad}}$.

For instance, $\mathcal{W}_\eta$ uses, at time $t = 1, 2, \ldots, T$, the weight vector $\mathbf{p}_t$ defined by

$$p_{j,t} = \frac{p_{j,0}\ e^{\eta R_{48\lfloor (t-1)/48 \rfloor}(\mathcal{E}_\eta, j)}\ \mathbb{I}_{\{j \in E_t\}}}{\sum_{k \in E_t} p_{k,0}\ e^{\eta R_{48\lfloor t/48 \rfloor}(\mathcal{E}_\eta, k)}}\,, \tag{11}$$

for all experts j, with the usual convention that empty sums equal 0.

In particular, the only difference between the definitions (3) and (11) in the case of uniform initial weights $p_{j,0} = 1/24$ is that not all past losses are used at time index t, but only the ones that were available at the time index when the forecast of the observation at round t was to be output, that is, after round $48\lfloor (t-1)/48 \rfloor$ and before round $48\lfloor (t-1)/48 \rfloor + 1$. (The notation $\lfloor x \rfloor$ denotes the lower integer part of a real number x.) This ensures that the weights $\mathbf{p}_t(\mathcal{W}_\eta)$ output by the adaptation are not too far from the weights $\mathbf{p}_t(\mathcal{E}_\eta)$ of the base version, thus preserving a $o(T)$ bound on the regret of $\mathcal{W}_\eta$, as desired.

A similar adaptation is considered for the gradient version of the exponentially weighted average rule; it suffices to consider in (5) the values of the regrets at rounds $48\lfloor (t-1)/48 \rfloor$ instead of the rounds $t - 1$.

Adaptation of the fixed-share type rules We only describe in detail the extension $\mathcal{G}_{\eta,\alpha}$ of the (basic) fixed-share aggregation rule $\mathcal{F}_{\eta,\alpha}$ to operational forecasting; the methodology is the same for the gradient versions $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ and $\mathcal{F}_{\eta,\alpha}^{\text{grad}}$.

As is illustrated in its statement in Figure 9, $\mathcal{G}_{\eta,\alpha}$ basically needs to run an instance of $\mathcal{F}_{\eta,\alpha}$ and to access its proposed weight vector every 48 rounds. We stated the extension in this way to highlight that it gets synchronized with the base rule $\mathcal{F}_{\eta,\alpha}$ every 48 time indexes, but of course more efficient implementations of $\mathcal{G}_{\eta,\alpha}$ could exist.

The interesting and crucial point is the behavior of the rule $\mathcal{G}_{\eta,\alpha}$ between two such synchronizations. In Figure 2, the base rule was performing, at each time index, a loss update in step (2) and a share update in step (3); the latter was used to deal with the specialized setting, i.e., with the fact that some experts become inactive at the next time index and some others become active again, while the former was to set the weights in accordance to the performance of each of the experts. Of course, in operational forecasting, the adjustments with respect to the individual performance of the experts can only be

Parameters: $\eta > 0$ and $0 \leqslant \alpha \leqslant 1$, as well as an initial convex weight vector $\big(p_{1,0}, \ldots, p_{N,0}\big)$
Initialization: $(w_{1,0}, \ldots, w_{N,0}) = \big(p_{1,0}\,\mathbb{I}_{\{1\in E_1\}}, \ldots, p_{N,0}\,\mathbb{I}_{\{1\in E_N\}}\big)$
For each round $t = 1, 2, \ldots, T$,

(1) $\widehat{y}_t = \dfrac{1}{\sum_{k=1}^N w_{k,t-1}} \sum_{j=1}^N w_{j,t-1}\, f_{j,t}\,;$

(2) [loss and share updates]
if $t = 48k$ for some k, observe $y_{t-47}, \ldots, y_t$ and take[a] $(w_{1,t}, \ldots, w_{N,t}) = \mathbf{p}_{t+1}(\mathcal{F}_{\eta,\alpha})$;

(3) [share update]
otherwise (when t is not a multiple of 48), let $w_{i,t} = 0$ if $i \notin E_{t+1}$ and

$$w_{i,t} = \frac{1}{|E_{t+1}|} \sum_{j\in E_t\setminus E_{t+1}} w_{j,t-1} + \frac{\alpha}{|E_{t+1}|} \sum_{j\in E_t\cap E_{t+1}} w_{j,t-1} + (1-\alpha)\,\mathbb{I}_{\{i\in E_t\cap E_{t+1}\}}\, w_{i,t-1}$$

if $i \in E_{t+1}$ (with the convention that an empty sum is null).

[a] $\mathbf{p}_{t+1}(\mathcal{F}_{\eta,\alpha})$ is the convex weight vector chosen by the rule $\mathcal{F}_{\eta,\alpha}$ after seeing the sequence of observations $y_1, \ldots, y_t$ and the corresponding experts predictions; we use here the same notation as in Section 3.2.2, where we indicated in parentheses the name of the rule whenever it was needed. Here, the rule $\mathcal{G}_{\eta,\alpha}$ thus synchronizes again with $\mathcal{F}_{\eta,\alpha}$ at steps t of the form $t = 48k$ for some k.

Figure 9: The extension $\mathcal{G}_{\eta,\alpha}$ of the (basic) fixed-share aggregation rule $\mathcal{F}_{\eta,\alpha}$ to operational forecasting.

performed every 48 time indexes but the share updates still need to be performed at each time index, since the values of the sets of active experts E_t can vary within a day (i.e., within a round of 48 time indexes). When losses for a round become available, the loss and share updates are done as they should have been done without the operational constraint: this is the meaning of the call to $\mathcal{F}_{\eta,\alpha}$ in step (2) of Figure 9.

4.2.2 Performance of exponentially weighted average rules

The performance of the extensions $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\text{grad}}$ described above is summarized in Table 9. As indicated in Section 2.3, it should be compared, respectively, to the performance of the best single expert (for $\mathcal{W}_\eta$) and to the one of the best convex weight vector (for $\mathcal{W}_\eta^{\text{grad}}$). We recall that these two values are reported in Table 8 but we indicated in the comments to it that the only interesting benchmarking value among the oracles of the first part of the table was the RMSE of the best convex weight vector, equal to 0.696.

We note that $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\text{grad}}$, when run with a fair initial allocation rather than a uniform one and when tuned with the best parameter η in hindsight, outperform this comparison point. It is also worth noting that the performance of the gradient version $\mathcal{W}_\eta^{\text{grad}}$ is not sensitive to the prior weights and that in all cases a relative improvement of about 6 % is obtained with respect to the performance of the best convex weight vector.

Here again, as already mentioned for the Slovakian data in Section 3.2.1, the best constant choices in hindsight are far away from the theoretically optimal ones, given by values $\eta^\star$ of the order of 10^{-6} on this data set. For such small values of η, the rules are basically equivalent to the uniform aggregation rule $\mathcal{U}$, as is indicated by the performance

Value of	η		10^{-6}	10^{-5}	10^{-4}	2×10^{-4}	10^{-3}	5×10^{-3}	10^{-2}
RMSE of	$\mathcal{W}_\eta$	(unif.)	0.724	0.722	<u>0.718</u>		0.731		0.788
	$\mathcal{W}_\eta$	(fair)	0.736	0.731	0.695	<u>0.683</u>	0.722		0.789
	$\mathcal{W}_\eta^{\text{grad}}$	(unif.)	0.724	0.722	0.712		0.683	<u>0.650</u>	0.668
	$\mathcal{W}_\eta^{\text{grad}}$	(fair)	0.737	0.733	0.711		0.674	<u>0.651</u>	0.670

Table 9: Performance obtained by the sequential aggregation rules $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\text{grad}}$ based on exponentially weighted averages for various choices of η; the smallest value for each rule is underlined.

			Best constant η	Grid $\widetilde{\Lambda}_{\mathcal{W}}$
RMSE of	$\mathcal{W}_\eta$	(unif.)	0.718	0.723
	$\mathcal{W}_\eta$	(fair)	0.683	0.696
	$\mathcal{W}_\eta^{\text{grad}}$	(unif.)	0.650	0.654
	$\mathcal{W}_\eta^{\text{grad}}$	(fair)	0.651	0.662

Table 10: Performance obtained by the rules $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\text{grad}}$ for the best constant choice of η in hindsight (left) and when used as keystones of a meta-rule selecting sequentially the values of η on the grid $\widetilde{\Lambda}_{\mathcal{W}}$ (right).

reported in Table 9.

However, the computation of the order of magnitude of the $\eta^\star$ served again to set the grid used in Table 9; we started around the theoretical optimal value and then performed logarithmic increments. Following the methodology of Section 3.2.2 we use it also to set the grid $\widetilde{\Lambda}_{\mathcal{W}}$ of on-line tuning of the η as

$$\widetilde{\Lambda}_{\mathcal{W}} = \left\{m\times10^{-k},\ \text{for } k\in\{1,\dots,6\} \text{ and } m\in\{1,\,2.5,\,5\}\right\}\cup\{1\}\,,$$

which contains 19 equally logarithmically spaced points. The choice of the upper bound 1 considered here will be explained and made automatic in Section 4.2.4. The performance on this grid with respect to the best constant choice of η in hindsight (as discussed in Table 9) is summarized in Table 10. Again, the fully on-line character of the meta-rule comes almost at no cost in the performance.

The sequence of weights chosen by the meta-rule based on the $\mathcal{W}_\eta^{\text{grad}}$ run with a fair initial allocation of the weights, as well as the sequence of η chosen at each step, are depicted in Figure 10.

4.2.3 Performance of fixed-share type rules

For both off-line and on-line performance, we considered uniform and fair initial allocations of the weights and resorted to the grid

$$\widetilde{\Lambda}_{\text{FS-France}} = \Big\{\big(m\times10^k,\,\alpha\big),\quad \text{for } m\in\{1,\,5\},\ k\in\{-6,\,\dots,\,0,\,\dots,\,4\},\\ \text{and } \alpha\in\{0,\ 0.001,\ 0.01,\,0.05,\,0.1,\,0.2\}\Big\}\,.$$

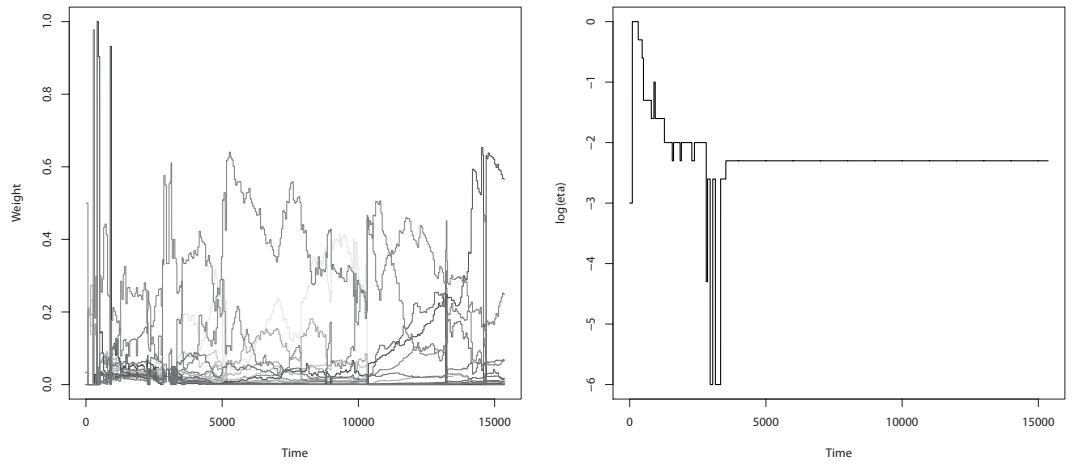

Figure 10: Graphical representations of the sequences of weights (left) and tuning parameters η (right) chosen by the meta-rule based on the $\mathcal{W}_{\eta}^{\mathrm{grad}}$ run with initial fair weight allocation and selecting sequentially the values on the grid $\widetilde{\Lambda}_{\mathcal{W}}$.

Value of	η	0.01	0.01	0.01	1	1	1	500	500	500
	α	0.001	0.01	0.05	0.001	0.01	0.05	0.001	0.01	0.05
RMSE of	$\mathcal{G}_{\eta,\alpha}$	0.678	0.683	0.704	0.711	0.659	0.652	0.674	0.633	<u>0.632</u>
	$\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$	0.646	0.669	0.700	0.622	<u>0.598</u>	0.637	0.683	0.675	0.671

Table 11: Performance obtained by the sequential aggregation rules $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$ run with an initial uniform allocation for various choices of η and α on the grid $\widetilde{\Lambda}_{\text{FS-France}}$; the smallest value for each rule is underlined.

(Section 4.2.4 will explain how to construct this grid in some adaptive way; we take it as given for the time being.) Actually, it turned out that the performance of the algorithms did not depend on whether the initial weight allocation was fair or uniform so that we report only the results obtained by the latter in the sequel.

The performance of the extensions $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$ to operational forecasting described above in Section 4.2.1 is summarized in Table 11. The comparison points are given by the best compound experts studied in Table 8; the best compound expert with 50 (unconstrained) shifts is already an excellent competitor with respect to our forecasters since the latter are evaluated on 320 days and can thus update essentially only 319 times their chosen convex weights (and in a constrained manner, since these updates can only occur at noon every day).

From Table 8 we expect a significant gain of performance when resorting to forecasters tracking the performance of the compound experts and this is what we read in Table 11, where the RMSE can get slightly better than 0.6 GW; the relative improvement in the performance with respect to the results of Table 10 is almost 10 %.

This excellent performance is almost unaffected when the adaptive version based on the grid $\widetilde{\Lambda}_{\text{FS-France}}$ is considered; see Table 12. A graphical representation of the weights and of the tuning parameters chosen by the meta-rule based on the $\mathcal{G}_{\eta,\alpha}$ is provided in Figure 11.

		Best constant pair (η, α)	Grid $\widetilde{\Lambda}_{\text{FS-France}}$
RMSE of	$\mathcal{G}_{\eta,\alpha}$	0.632	0.644
	$\mathcal{G}^{\text{grad}}_{\eta,\alpha}$	0.598	0.599

Table 12: Performance obtained by the rules $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}^{\text{grad}}_{\eta,\alpha}$ run with an initial uniform allocation for the best constant choices of η and α in hindsight (left) and when used as keystones of a meta-rule selecting sequentially the values of η and α on the grid $\widetilde{\Lambda}_{\text{FS-France}}$ (right).

4.2.4 Final elements in the construction of fully adaptive aggregation rules

In Sections 3.2.2, 4.2.2, and 4.2.3 we did not clarify how to choose the maximal possible value of η in the considered grids; only their starting points were defined in an intrinsic way (at a value close to $\eta^\star$) and we indicated therein that our simulations showed that the step of the grid was not a crucial parameter and that the results were not too sensitive to it. Therefore, the last issue that remains to be dealt with is the choice of the maximal possible value of η. In Sections 3.2.2 and 4.2.2 we stopped the grid somewhat arbitrary at the value 1; Section 4.2.3 showed however that values of η larger than 1 could yield some improvements. We propose the following procedure to perform automatic and efficient choices of the upper bound of the grid.

The procedure is based on the observation that in Figures 6, 7, 10, and 11 the sequences of values chosen on-line for the parameters using the method described in Section 3.2.2 are asymptotically (essentially) constant; that is, provided that the grid is large enough and the best constant choice in hindsight of the parameter lies in the grid, this best constant choice will be achieved after a certain point and the average performance will be close to the one of this parameter.

It thus suffices to ensure that the grid is large enough, i.e., that it covers a large enough spectrum. This can be implemented by extending the considered grid on-line as follows. We let the user fix a starting grid (e.g., a grid between $\eta^\star$ and 1) and check at each time index whether the next increment of the maximal value of the parameter of the current grid would have obtained a better performance when being used as a constant choice than the one of any other point in the current grid; if this is the case, this next increment is added to the current grid for the remaining time indexes. The upper bounds of the grids can therefore only increase over time but in practice, once the stationarity level of the sequences of on-line calibrated parameters is reached no further addition is made to the grid.

For example, in the setting of Section 4.2.3, if the initial grid had been set to

$$\widetilde{\Lambda}_{\text{FS-small}} = \Big\{ \big(m \times 10^k,\, \alpha\big), \quad \text{for } m \in \{1,\, 5\},\ k \in \{-6,\, \ldots,\, 0,\, 1\}, \\ \text{and } \alpha \in \{0,\ 0.001,\ 0.01,\, 0.05,\, 0.1,\, 0.2\} \Big\},$$

the above method would quickly have added 10, 50, 100, 500, 10 000 to the grid of the η; it would then have achieved almost the same performance as the one discussed in Section 4.2.3, where the initial grid already contained these values. Note that Figure 11 (bottom, left) shows that no larger values than 10 000 were considered in Section 4.2.3 despite the fact that the grid contained one larger such value (50 000).

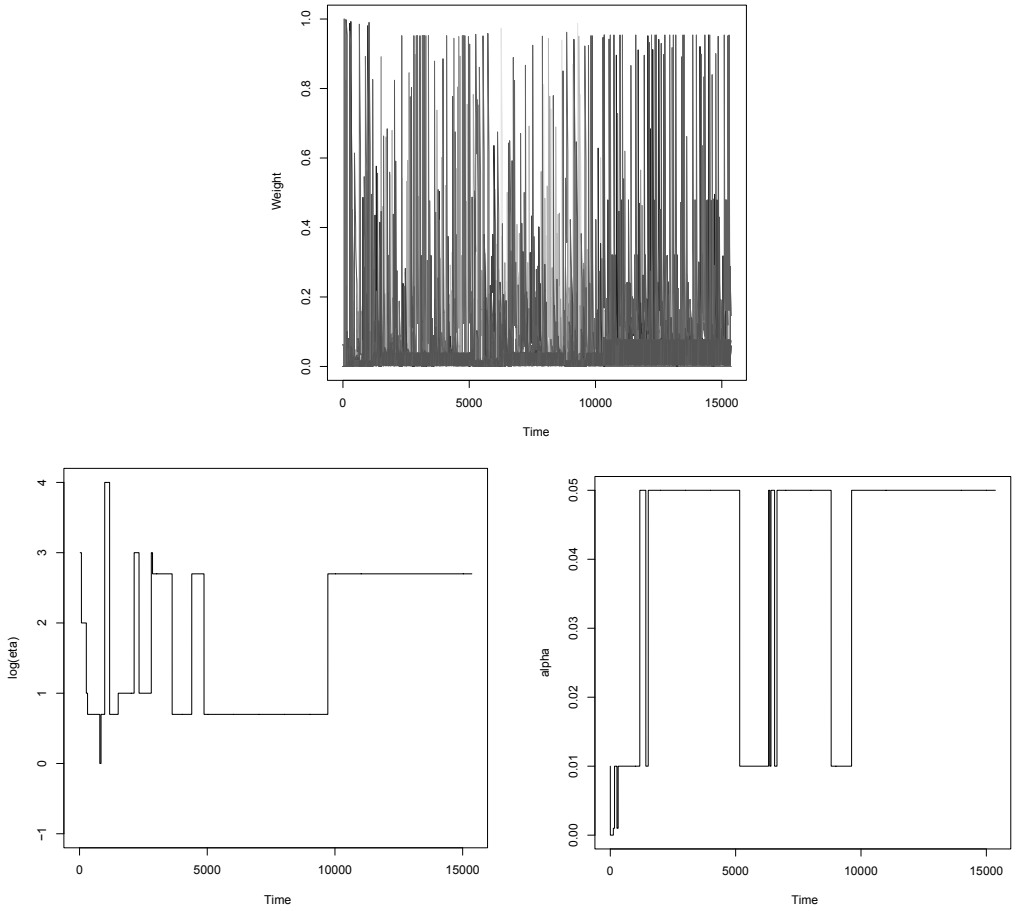

Figure 11: Graphical representations of the sequences of weights (top) and tuning parameters η (bottom, left) and α (bottom, right) chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_{\text{FS-France}}$; the base rule is $\mathcal{G}_{\eta,\alpha}$ run with an initial uniform allocation.

4.3 Robustness of the considered aggregation rules

In this section we move from the study of global average behaviors of the aggregation rules (as measured by their RMSE) to a more individual analysis. In particular we study the performance as a function of the hours of the days and compare the individual extreme behaviors of the base experts and of the aggregation rules designed above. This is to see whether the good average behavior comes or not at the cost of some disastrous forecasts from time to time.

To do so we consider the respective best fully sequential aggregation rules of Sections 4.2.2 and 4.2.3, that is, the meta-forecasters using the families of $\mathcal{W}_{\eta}^{\text{grad}}$ and $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ run with initial uniform weight allocations and calibrating their parameters on a grid. We use as benchmarks the (overall) best single expert and the (overall) best convex weight vector, whose performance was reported in Table 8.

Figure 12 plots the half-hourly RMSE of these two aggregation rules and two benchmarks. It shows that the performance of the rule based on exponential weighted averages is, uniformly over the 48 elements of the partition of days in half hours, at least as good as the one of the best constant convex combination of the experts forecasts. The rule based on fixed-share aggregation rules exhibits a performance a bit worse than the latter benchmark on the period 7:00–12:00 but the strongly improved performance in the period

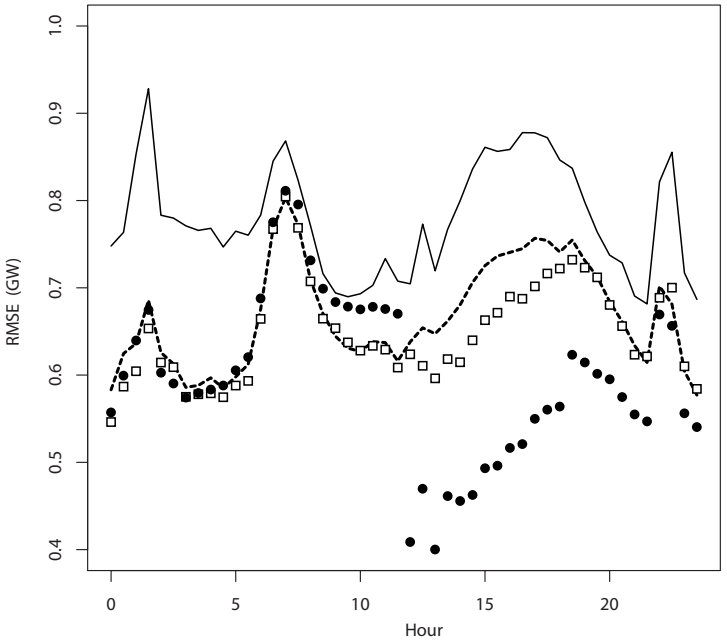

Figure 12: Half-hourly RMSE of the meta-rules based on the rules $\mathcal{W}_{\eta}^{\text{grad}}$ (symbol: •) and $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ (symbol: □) and calibrating them respectively on the grids $\widetilde{\Lambda}_{\mathcal{G}}$ and $\widetilde{\Lambda}_{\text{FS-France}}$; as well as the ones of the best overall single expert (solide line) and of the best overall convex weight vector (dashed line).

12:00–21:00 does more than compensate for that. It thus seems that this rules has excellent performance on very short-term horizon and would strongly benefit from an intermediate update around midnight (that goes however, as it stands now, against the operational constraint). A further study of this behavior and its benefits is left to future research.

As a measure of robustness, we plot in Figures 13 and 14 quantities related to the distribution of the residuals of the forecasts, that is, the difference between the actual consumptions y_t and the forecasts operated by the rules and benchmarks described above. Here again, we grouped the residuals by half hours. Figure 13 represents the median, the third quartile, and the 90 % quantiles of the absolute values of the residuals. The graphs of Figures 14 are concerned with the behavior of the signed residuals, with the medians and the first and third quartiles on the top graph and the interquartile distances on the bottom graph. They all show the same story as described above for the half-hourly RMSE. In particular, the distributions of the errors of the aggregation rules are more concentrated than the ones for best benchmarks, which indicates that their good overall performance does not come at the cost of some local disasters in the quality of the predictions.

All in all, we conclude that the best aggregation rules never encounter large prediction errors in comparison to the best expert or to the best convex combination of experts and often encounter much smaller such errors. This is strongly in favor of their use in an industrial context where large errors can be highly prejudicious (from financial penalties to black outs). In a nutshell, aggregation rules can reduce the risk of prediction, which is one important pro for operational forecasting.

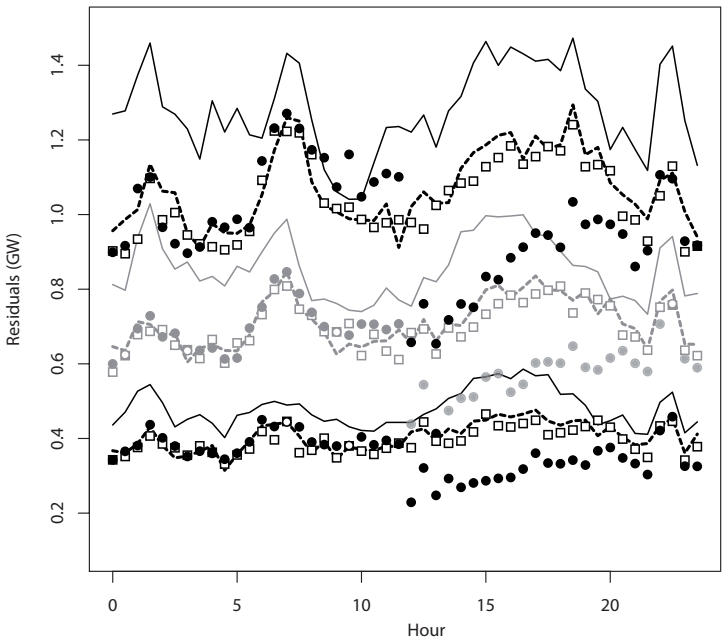

Figure 13: Using the same rules and benchmarks as in Figure 12, with the same legend: 50 % (black), 75 % (grey), and 90 % (black) quantiles of the absolute values of the residuals, grouped per half hours.

5 Conclusion

In this paper we showed the interest of ensemble methods in the prediction of electricity load; the sequential aggregation rules we discussed are black-box ways to improve on a bunch of base forecasters. In particular, on the two data sets the best rules, given by fixed-share type rules, improve on the best constant convex combination of the experts by about 5 % (Slovakian data set) to about 15 % (French data set). Thus, this line of research comes as a complement to the design of good base forecasters; if benefits from the consideration of several as different as possible base forecasters. It is not in opposition but on top of the more classical problem of constructing the latter. Note that it suffices to design specialized (thus sleeping) forecasters; they only output forecasts in the contexts when the methods they rely on are known to be efficient.

The raw improvement in terms of the global performance, as measured by the RMSE, from the experts on the sequential aggregation rules, also comes together with a reduction of the risk of large errors: the studied aggregation rules are more robust than the base forecasters they are using.

By passing, we also contributed to the methodology of the field of prediction with expert advice. First, we illustrated how to efficiently tune on-line the few (one or two) parameters needed for each aggregation rule. Second, we extended to the setting of specialized experts the fixed-share type rules; and for these rules and the exponentially weighted average ones, we indicated how to extend them so as to take into account the operational constraint of outputting simultaneous forecasts for a fixed number of future time indexes.

We also detail two research perspectives, in addition to the one already mentioned in Section 4.3. On the methodological side, a useful set of strategies is usually given by regularized least-square aggregation rules, but no simple extension of them to the case

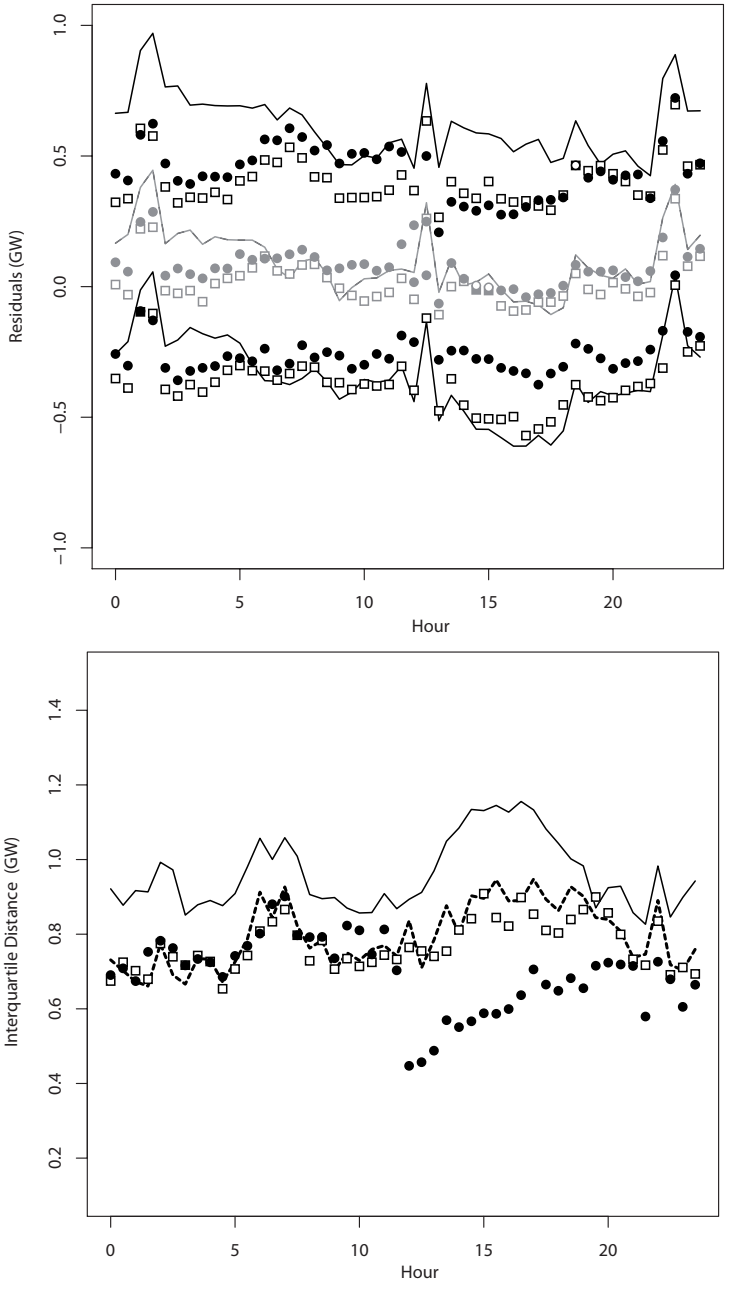

Figure 14: Using the same rules and benchmarks as in Figure 12, with the same legend: 25 % (black), 50 % (grey), and 75 % (black) quantiles (top graph) and interquartile distances (bottom graph) of the signed values of the residuals, grouped per half hours. For the sake of readability, the values for the best constant convex combination are not displayed on the the top graph.

of specialized experts is known; the technical report Devaine et al. [2009] shows some preliminary attempts to do so but largely fails achieving reasonable rules with satisfactory performance. On the operational side, the aggregated forecasts should come with a measure of their uncertainties; the latter would be provided either by the aggregation of some measures of uncertainties related to the base forecasts provided by the experts (ongoing work at EDF R&D) or by a mere inspection of the dispersion of the base forecasts themselves. Indeed, the more concentrated are the latter around a given value, the more confident a strategy should be about its own aggregated forecast.

References

A. Antoniadis, E. Paparoditis, and T. Sapatinas. A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B*, 68(5):837–857, 2006.

A. Antoniadis, X. Brossat, J. Cugliari, and J.M. Poggi. Clustering functional data using wavelets. In *Proceedings of the Nineteenth International Conference on Computational Statistics (COMPSTAT)*, 2010.

A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.

A. Bruhns, G. Deurveilher, and J.-S. Roy. A non-linear regression model for mid-term load forecasting and improvements in seasonnality. In *Proceedings of the Fifteenth Power Systems Computation Conference (PSCC)*, 2005.

D. W. Bunn and E. D. Farmer. *Comparative Models for Electrical Load Forecasting*. John Wiley and Sons Inc., New York, 1985.

R. Campo and P. Ruiz. Adaptive weather-sensitive short-term load forecasting. *IEEE Transactions on Power Systems*, 3:592–600, 1987.

N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51:239–261, 2003.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

R. Cottet and M. Smith. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849, 2003.

T.M. Cover. Universal portfolios. *Mathematical Finance*, 1:1–29, 1991.

V. Dani, O. Madani, D. Pennock, S. Sanghai, and B. Galebach. An empirical comparison of algorithms for aggregating expert predictions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

M. Devaine, Y. Goude, and G. Stoltz. Aggregation of sleeping predictors to forecast electricity consumption. Technical report, École normale supérieure, Paris and EDF R&D, Clamart, July 2009. Available at `http://www.math.ens.fr/%7Estoltz/DeGoSt-report.pdf`.

V. Dordonnat, S.J. Koopman, M. Ooms, A. Dessertaine, and J. Collet. An hourly periodic state space model for modelling French national electricity load. *International Journal of Forecasting*, 24:566–587, 2008.

Y. Freund, R. Schapire, Y. Singer, and M. Warmuth. Using and combining predictors that specialize. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 334–343, 1997.

S. Gerchinovitz, V. Mallet, and G. Stoltz. A further look at sequential aggregation rules for ozone ensemble forecasting. Technical report, INRIA Paris-Rocquencourt and École normale supérieure, Paris, September 2008. Available at `http://www.math.ens.fr/%7Estoltz/GeMaSt-report.pdf`.

Y. Goude. *Mélange de prédicteurs et application à la prévision de consommation électrique*. PhD thesis, Université Paris-Sud XI, January 2008a.

Y. Goude. Tracking the best predictor with a detection based algorithm. In *Proceedings of the Joint Statistical Meetings (JSP)*, 2008b. See the section on Statistical Computing.

A. Harvey and S. Koopman. Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association*, 88(424):1228–1253, 1993.

M. Herbster and M. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

H.S. Hippert, C.E. Pedreira, and R.C. Souza. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, 16(1):44–55, 2001.

R.D. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. In *Proceedings of the Twenty-First Annual Conference on Learning Theory (COLT)*, pages 425–436, 2008.

Vivien Mallet, Gilles Stoltz, and Boris Mauricette. Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research*, 114(D05307), 2009.

A. Pierrot, N. Laluque, and Y. Goude. Short-term electricity load forecasting with generalized additive models. In *Proceedings of the Third International Conference on Computational and Financial Econometrics (CFE)*, 2009.

R. Ramanathan, R. Engle, C.W.J. Granger, F. Vahid-Araghi, and C. Brace. Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, 13:161–174, 1997.

M. Smith. Modeling and short-term forecasting of New South Wales electricity system load. *Journal of Business and Economic Statistics*, 18:465–478, 2000.

J. Taylor, L.M. de Menezesb, and P.E. McSharry. A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, 22:1–16, 2006.

J.W. Taylor. An evaluation of methods for very short term electricity demand forecasting using minute-by-minute British data. *International Journal of Forecasting*, 24:645–658, 2008.

V. Vovk and F. Zhdanov. Prediction with expert advice for the Brier game. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML)*, 2008.

S.N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2006.