Forecasting the electricity consumption by aggregating specialized experts; a review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions

Marie Devaine, Yannig Goude, Gilles Stoltz

Noname manuscript No.
(will be inserted by the editor)

Forecasting the electricity consumption by aggregating specialized experts

A review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions

Marie Devaine · Yannig Goude · Gilles Stoltz

Received: March 27, 2011

Abstract We consider a variant of the sequential prediction of arbitrary sequences based on experts advice, called prediction with specialized experts. We design aggregation rules, that sequentially combine the forecasts provided by the experts; the latter are specialized and need not output a prediction at all time instances while the aggregation rules have to. We provide first a review of the literature on specialized experts and take a new look at some aggregation rules (obtained as adaptations or extensions of earlier rules). We then consider an application to the sequential short-term (one-day-ahead) forecasting of electricity consumption; to do so, we consider two data sets, a Slovakian one and a French one, respectively concerned with hourly and half-hourly predictions. We introduce and develop a general methodology to perform the stated empirical studies. The introduced aggregation rules demonstrate an improved accuracy on the data sets at hand; the improvements lie in a reduced mean squared error but also in a more robust behavior with respect to large occasional errors.

Keywords Prediction with experts advice · Specialized experts · Individual sequences · Application to real data

M. Devaine
Ecole Normale Supérieure, Paris, France
E-mail: marie.devaine@ens.fr

Y. Goude
EDF R&D, Clamart, France
Tel.: +33-147-651-561
E-mail: yannig.goude@edf.fr

G. Stoltz
Ecole Normale Supérieure, CNRS, INRIA, Paris, France
&
HEC Paris, CNRS, Jouy-en-Josas, France
Tel.: +33-144-323-277
E-mail: gilles.stoltz@ens.fr

Contents

1 Introduction and motivation

We consider the sequential prediction of arbitrary sequences based on experts advice, the topic of a large literature summarized in the monography of Cesa-Bianchi and Lugosi [2006]. At each round of a repeated game of prediction, experts output forecasts, which are to be combined by an aggregation rule (usually based on their past performance); the true outcome is then revealed and losses, which correspond to prediction errors, are suffered by the aggregation rules and the experts. We are interested in aggregation rules that perform almost as well as the

best constant convex combination of the experts; more precisely, we will control the difference between the average errors encountered by the sequential aggregation rules and by the best constant convex combination of the base forecasters –a quantity called regret. The considered aggregation rules come thus with strong theoretical guarantees on their performance. In our setting, 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. This is why guarantees are said to be given for all individual sequences. (However, as is indicated below, the experts might rely on some stochastic modeling.) The high-level principle of the studied aggregation rules is to output convex combinations of the predictions formed by the experts, where the values of the convex weights chosen over time vary according to the past performance of the experts.

A variant of the general problem of prediction with experts advice: Specialized experts

The application we have in mind –the sequential short-term (one-day-ahead) forecasting of electricity consumption– will take place in a variant of the basic problem of prediction with experts 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 expected to provide accurate forecasts mostly in given external conditions, that can be known beforehand; it is designed to refrain from forming a prediction when these conditions are not met. For instance, in the case of the prediction of electricity consumption, experts can be specialized to winter or to summer, to working days or to public holidays, etc.

The specialization is usually a fortunate and desirable property of an expert, as the latter is likely to be more accurate on the prediction instances when it is active (and as, symmetrically, it gets inactive when poor performance is expected).

The literature on specialized experts is –to the best of our knowledge– rather sparse. The first references are Blum [1997] and Freund et al. [1997]; they respectively introduce and formalize the framework of specialized experts. They were followed only by few other ones: two papers mention some results for the context of specialized experts only in 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].

Previous applications of the sequential prediction of individual sequences to real data

In the applications the terminology of ensemble methods is often used; similarly, experts can also be called base forecasters (since the methodology at hand is to use several possibly independent base forecasters and design meta-forecasters which sequentially combine the base predictions that are output by them). As far as the forecasting of electricity consumption is concerned, a preliminary study of some aggregation rules for individual sequences was already performed for the daily prediction of the French electricity load in Goude [2008a,b]. But the aggregation rules at hand apply in theory to various other settings –actually, to virtually all settings where sequential prediction is to be performed with the help of experts.

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] 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 experts advice and believe that the present paper contributes to making these techniques more popular –a notoriety that they deserve.

Construction of the experts for the forecasting of short-term electricity consumption

The base forecasters to be used in our empirical studies 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 the short-term (one-day ahead) forecasting of electricity demand 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 of 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 (most importantly, the seasonal variations of day lengths), calendar events like vacation periods or public holidays, weather conditions (temperature, cloud cover, wind), and weekly patterns of days. It is difficult to provide an exhaustive list of all forecasting methods for the electricity load and we only highlight some popular 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 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; a presentation can be found in Bruhns et al. [2005]. The latter nonlinear regression model will indeed be used in this paper to provide some experts. Semi-parametric Bayesian regressions with independent 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 probability distribution functions over the set of possible forecasts. 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 models the observed load demand as discrete recordings of an underlying stochastic curve; see Antoniadis et al. [2006] and Antoniadis et al. [2010]. Another expert used in the present paper is indeed 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 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.

The methodology followed in this paper can be rephrased as follows. We consider a bunch of base forecasters constructed with the methods reviewed above and study aggregation rules (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 (or the best constant convex combination of the base forecasters). For instance, a way to construct these base forecasters is to instantiate a prediction method based on several parameters with different sets of such 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 thus constructed is that it is often not clear in advance which set of parameters will lead to the most accurate predictions.

Comparison to stochastic prediction methods

In this paper we only resort to techniques stemming from the theory of sequential prediction with expert advice, a subfield of machine learning. Other techniques are considered in the statistical literature to combine such experts forecasts. We briefly mention them here but point out that the present paper –in view of its current length– cannot perform a detailed comparison of the respective theoretical and empirical merits of these stochastic methods versus our approach; this is deferred to future work (to be published in the applied statistics community). In particular, the goal of the present paper is to construct efficient and fully adaptive strategies based on individual sequences techniques to be used in this future paper as outsiders to existing statistical techniques.

The statistical literature on combining forecasts is vast; one of the founding papers was Bates and Granger [1969], in which the instantaneous errors of each experts were essentially assumed to be independent and identically distributed, so that optimal combinations could be derived based on the variances of errors of each expert. An early application of techniques of this flavor to electricity demand forecasting was proposed by Smith [1989].

The closest, maybe, to our setting of prediction with expert advice would be the use of so-called Bayesian model averaging (BMA); it was introduced in Leamer [1978], Kass and Raftery [1995], Hoeting et al. [1999] and applied, e.g., to ensemble weather forecasting in Raftery et al. [2005]. This technique has a Bayesian flavor as its name indicates. Each experts is associated with a probability distribution function over the possible observations and BMA combines these distributions, by computing the weights as some posterior probabilities. Doing so, not only predictions but also uncertainties can be provided. This is one advantage over our techniques based on individual sequences. However, BMA is computationally more

involved (the computation of the weights is based on EM algorithms) and is not sequential by nature; in turn, its performance bounds are rather for stochastic processes and are not comparable to the theoretical bounds developed in the context of individual sequences.

Lastly, virtually all other stochastic methods –like CART, random forests, etc.– are based on contextual data and should rather be used as experts. Our approach is therefore absolutely not incompatible with the use of such stochastic methods, it simply appears as a second layer of prediction, via aggregation.

Contributions and outline of the paper

We review in Section 2 the framework of sequential prediction with specialized experts, by defining 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 such families of aggregation rules (Section 2.3). These rules are obtained by taking a new look at existing strategies; this new look corresponds to (slight or more important) adaptations of these existing strategies and/or to simpler and/or more general analyses of their theoretical performance bounds.

We then study, respectively in Sections 4 and 5, the performance obtained by the developed aggregation rules on two data sets. The first one was provided by the Slovakian subbranch of 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 standardized methodological scheme (described in Section 3), which consists of four steps:

- presentation of the data sets and of the experts, in Sections 4.1 and 5.1;
- performance of some benchmark prediction methods, in Sections 4.2 and 5.2;
- results obtained by the sequential aggregation rules with parameters optimally tuned in hindsight, in Sections 4.3 and 5.3;
- stability of the previous results when the tuning is performed sequentially (as it should be, leading to fully operational rules), in Sections 4.4 and 5.4.

The section on French data is also followed by a note (Section 5.5) on the individual performance of the aggregation rules, i.e., an indication that their behavior is not only good on average but also that the large prediction errors occur less frequently for the aggregation rules than for any base expert.

Section 6 concludes the paper, by summarizing the empirical evidence and by providing some research perspectives, while an appendix (Section 7) contains some proofs omitted from the main text because they correspond to immediate (but sometimes lengthy) adaptations of well-known techniques.

2 Sequential aggregation of specialized experts: A survey with some new results

A sequence of observations (e.g., hourly or half-hourly electricity consumptions) $y_1, y_2, \ldots, y_T$ is to be predicted element by element at time instances $t = 1, 2, \ldots, T$. For the sake of concreteness, we assume that the observations 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 instance 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. We assume that the experts know the bound B so that they only produce forecasts $f_{j,t} \in [0, B]$.

Finally, we denote by $E_t \subset \{1, \ldots, N\}$ the set of active experts at a given time instance t and assume that it is always non empty.

2.1 Definition of a sequential convex aggregation rule

At each time instance $t \geqslant 1$, a sequential convex aggregation rule produces a convex weight vector $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ based on the past observations $y_1, \ldots, y_{t-1}$ 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 $\boldsymbol{p}_t \in \mathbb{R}^N$ 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 experts in E_t according to the weights given by the components of the vector $\boldsymbol{p}_t$. More precisely, the aggregated prediction at time instance t equals

$$\widehat{y}_t = \sum_{j \in E_t} p_{j,t} f_{j,t} \,.$$

The observation y_t is then revealed and instance $t + 1$ then starts.

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 observation y_t we consider a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. At each time instance t, the convex combination $\boldsymbol{p}_t$ output by the rule is thus evaluated by the loss function $\ell_t : \mathcal{X} \to \mathbb{R}$ defined by

$$\ell_t(\boldsymbol{p}) = \ell\left(\sum_{j \in E_t} p_j f_{j,t},\, y_t\right)$$

for all $\boldsymbol{p} \in \mathcal{X}$. The subscript t in the notation ℓ_t encompasses the dependencies in the experts forecasts $f_{j,t}$ and in the outcome y_t. Our goal is to design sequential convex aggregation rules $\mathcal{A}$ with a small mean error

$$\overline{\mathrm{ERR}}(\mathcal{A}) = \frac{1}{T} \sum_{t=1}^{T} \ell_t(\boldsymbol{p}_t) \,.$$

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

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

We have that $\textsc{rmse}(\mathcal{A})^2 = \overline{\textsc{err}}(\mathcal{A})$ when ℓ is the square loss.

Actually, the only features of the square loss that we will need below are that it is such that all loss functions ℓ_t are convex and bounded on $[0, B]^2$, with a subgradient that is also bounded on this region. The absolute error $\ell(x, y) = |x-y|$ and the absolute percentage of error $\ell(x, y) = |x-y|/y$ would thus be suitable loss functions in the sequel as well. This is why we will formulate all the algorithms below and their regret bounds first in terms of general loss functions $\ell_t : \mathcal{X} \to \mathbb{R}$ and then indicate how they can be instantiated to the case of the square loss.

We start in this respect with the definition of the regret. The regret compares the performance of the aggregation rules to the one of the experts or to the one of some simple rules based on the experts. The first two notions of regret recalled below were introduced in Freund et al. [1997] while the third one is a straightforward extension of a definition provided by Herbster and Warmuth [1998].

2.2.1 Comparison to a fixed expert

A difficulty in the setting of specialized experts is that the mean error of an expert is not necessarily well defined; e.g., the RMSE of the j–th expert cannot be defined in general as the RMSE of the aggregation rule that would predict at each time instance as the j–th expert, simply because the latter is not defined at a given time instance when the expert is inactive.

We therefore only consider the instances t where j is active, a fact denoted by $j \in E_t$, and define the RMSE of the j–th expert as

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

A general definition (i.e., for general loss functions) of the mean error of an expert would be

$$\overline{\textsc{err}}(j) = \frac{1}{\sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}}} \sum_{t=1}^{T} \ell_t(\delta_j) \, \mathbb{I}_{\{j \in E_t\}} \, , \tag{1}$$

where $\delta_j \in \mathcal{X}$ is the convex weight vector that puts a mass 1 on the j–th component.

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 instances when j was active, that is, we compare $\overline{\textsc{err}}(j)$ to

$$\overline{\textsc{err}}(\mathcal{A}, j) = \frac{1}{\sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}}} \sum_{t=1}^{T} \ell_t(\boldsymbol{p}_t) \, \mathbb{I}_{\{j \in E_t\}} \, .$$

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

$$R_T(\mathcal{A}, j) = \sum_{t=1}^{T} \big(\ell_t(\boldsymbol{p}_t) - \ell_t(\delta_j) \big) \, \mathbb{I}_{\{j \in E_t\}} \, . \tag{2}$$

Our methodology, which is to ensure that the performance of $\mathcal{A}$ is almost 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 $\overline{\mathrm{ERR}}(\mathcal{A})$ is small as well as soon as there are experts j that *both* exhibit 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

$$\overline{\mathrm{ERR}}(\mathcal{A}) \leqslant \min_{j=1,\dots,N} \left\{ \frac{\sum_{t=1}^T \mathbb{I}_{\{j\in E_t\}}}{T}\, \overline{\mathrm{ERR}}(j) + \frac{R_T(\mathcal{A}, j)}{T} + L\left(1 - \frac{\sum_{t=1}^T \mathbb{I}_{\{j\in E_t\}}}{T}\right)\right\}, \tag{3}$$

where L is a bound on $|\ell_t|$; for instance, $L = B^2$ in the case of the square loss $\ell(x, y) = (x - y)^2$ under the assumption that the experts forecasts and the observations all lie in some bounded interval $[0, B]$.

2.2.2 Comparison to a fixed convex combination of experts

The regret methodology can be extended by now allowing comparison to rules based on fixed convex combinations of the experts. The latter are each parameterized by a convex weight vector $\boldsymbol{q} \in \mathcal{X}$ and they sequentially aggregate the forecasts of the experts based on a normalization of $\boldsymbol{q}$ to the set of active experts.

Formally, for a set $E \subset \{1, \dots, N\}$, we define

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

and denote by $\boldsymbol{q}^E = (q_1^E, \dots, q_N^E)$ the following convex weight vector, which can be interpreted as the convex weight vector obtained by "conditioning" $\boldsymbol{q}$ to E:

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

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

$$\overline{\mathrm{ERR}}(\boldsymbol{q}) = \frac{1}{\sum_{t=1}^T \boldsymbol{q}(E_t)} \sum_{t=1}^T \ell\big(\boldsymbol{q}^{E_t}\big)\, \boldsymbol{q}(E_t)\,.$$

Indeed, when $\boldsymbol{q} = \delta_j$ we recover $\overline{\mathrm{ERR}}(\delta_j) = \overline{\mathrm{ERR}}(j)$.

The notion of cumulative regret of a sequential aggregation rule $\mathcal{A}$ with respect to some fixed weight vector $\boldsymbol{q}$ up to T can be generalized from (2) as

$$R_T(\mathcal{A}, \boldsymbol{q}) = \sum_{t=1}^T \Big(\ell_t(\boldsymbol{p}_t) - \ell_t\big(\boldsymbol{q}^{E_t}\big)\Big)\, \boldsymbol{q}(E_t)\,. \tag{4}$$

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

An argument similar to (3) shows that if the regrets $R_T(\mathcal{A}, \boldsymbol{q})$ can all be guaranteed to be small, then the performance of $\mathcal{A}$, as measured by $\overline{\mathrm{ERR}}(\mathcal{A})$, is good as well. Of course, ensuring that all quantities $R_T(\mathcal{A}, \boldsymbol{q})$ are small is more difficult than simply controlling the regrets $R_T(\mathcal{A}, j)$ with respect to the experts.

2.2.3 Comparison to sequences of (convex combinations of) experts with few shifts

This third and last definition of regret was introduced by Herbster and Warmuth [1998] and compares the performance of a rule not to the performance of a fixed expert or a fixed convex combination of the experts, but to sequences of experts or of convex combinations of experts. The comparison can be made for all time instances provided that the sequences of (convex combinations of) experts are well chosen. To the best of our knowledge, this approach of considering sequences of experts had not been used before to deal with specialized experts.

Formally, we denote by $\mathcal{L}$ the set of all legal sequences of expert instances $j_1^T = (j_1, \dots, j_T)$, where legality means that for all time instances t, the considered expert j_t is active (i.e., is in E_t). We call compound experts the elements of $\mathcal{L}$. Similarly, we denote by $\mathcal{C}$ the set of all legal sequences of convex weight vectors $\boldsymbol{q}_1^T = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_T)$, where legality means that for all time instances t, the considered convex weight vector $\boldsymbol{q}_t$ puts positive masses only on elements in E_t. We call compound convex weight vectors the elements of $\mathcal{C}$.

For such compound experts j_1^T or compound convex weight vectors $\boldsymbol{q}_1^T$, we denote by

$$\mathrm{size}\big(j_1^T\big) = \sum_{t=2}^{T} \mathbb{I}_{\{j_{t-1} \neq j_t\}} \qquad \text{and} \qquad \mathrm{size}\big(\boldsymbol{q}_1^T\big) = \sum_{t=2}^{T} \mathbb{I}_{\{\boldsymbol{q}_{t-1} \neq \boldsymbol{q}_t\}}$$

their numbers of switches (the number minus one of elements in the partition of $\{1, \dots, T\}$ into integer subintervals corresponding to the use of the same expert or convex weight vector). For $0 \leqslant m \leqslant T-1$, we then respectively define $\mathcal{L}_m$ and $\mathcal{C}_m$ as the subsets of $\mathcal{L}$ and of $\mathcal{C}$ containing the compound experts and compound convex weight vectors with at most m shifts. When m is too small, the subsets $\mathcal{L}_m$ and $\mathcal{C}_m$ might be empty.

The definition of the mean error of a compound expert $j_1^T \in \mathcal{L}$,

$$\overline{\mathrm{ERR}}\big(j_1^T\big) = \frac{1}{T} \sum_{t=1}^{T} \ell_t\big(\delta_{j_t}\big)\,,$$

and of the regret of a rule $\mathcal{A}$ with respect to $j_1^T \in \mathcal{L}$,

$$R_T\big(\mathcal{A}, j_1^T\big) = \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{p}_t) - \ell_t\big(\delta_{j_t}\big) \Big)\,, \tag{5}$$

are immediate in this setting.

The relationship between the mean error of a given rule and the ones of the elements of the comparison class formed by $\mathcal{L}_m$, for some m, is simpler than in (3). Indeed,

$$\overline{\mathrm{ERR}}(\mathcal{A}) \leqslant \min_{j_1^T \in \mathcal{L}_m} \left\{ \overline{\mathrm{ERR}}\big(j_1^T\big) + \frac{R_T\big(\mathcal{A}, j_1^T\big)}{T} \right\};$$

here, we fixed a number m of switches because, as we will see below, the regret can only be guaranteed to be small if m itself is not too large.

The definitions of the mean error and of the regret, as well as the above inequality, can be extended in a straightforward manner to compound convex weight vectors, as follows; for all $\boldsymbol{q}_1^T \in \mathcal{C}$,

$$\overline{\mathrm{ERR}}\big(\boldsymbol{q}_1^T\big) = \frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{q}_t)$$

and

$$R_T\big(\mathcal{A}, \boldsymbol{q}_1^T\big) = \sum_{t=1}^{T} \big(\ell_t(\boldsymbol{p}_t) - \ell_t(\boldsymbol{q}_t)\big)\,. \tag{6}$$

Since $\mathcal{L}_m \subseteq \mathcal{C}_m$ (up to the identification of expert indexes j to convex weight vectors δ_j), it is more difficult to control the regret with respect to all elements of $\mathcal{C}_m$ than the one with respect to simply $\mathcal{L}_m$.

2.2.4 Bounds on the regret obtained by considering (sub)gradients

Most of the algorithms discussed below will have two forms: a basic version (using the losses ℓ_t) and a gradient version, based on the following remark, which exploits a fundamental result in convex analysis. When the loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$ is convex in its first argument –as is the case for all specific loss functions mentioned above–, then the functions ℓ_t are convex and admit a least one subgradient at all points $\boldsymbol{p} \in \mathcal{X}$; we denote by $\nabla\ell_t(\boldsymbol{p})$ such a subgradient. By denoting by $\cdot$ the inner product in $\mathbb{R}^N$ (and viewing $\mathcal{X}$ as a subset of $\mathbb{R}^N$) we thus get the following inequality: for all t, for all $\boldsymbol{q} \in \mathcal{X}$,

$$\ell_t(\boldsymbol{p}_t) - \ell_t(\boldsymbol{q}) \leqslant \nabla\ell_t(\boldsymbol{p}_t) \cdot \big(\boldsymbol{p}_t - \boldsymbol{q}\big) = \widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\boldsymbol{q})\,,$$

where we denoted by $\widetilde{\ell}_t(\boldsymbol{q}) = \nabla\ell_t(\boldsymbol{p}_t) \cdot \boldsymbol{q}$ the pseudo-loss function associated with time instance t. It is linear over $\mathcal{X}$.

The above inequality shows that the cumulative regrets defined in (2), (4), and (5)–(6) can be upper bounded by quantities of the same form where the functions ℓ_t are simply replaced by the pseudo-losses $\widetilde{\ell}_t$. The so-called gradient versions of the algorithms defined below minimize the upper bounds on the regrets defined in terms of the $\widetilde{\ell}_t$, and which we denote respectively by

$$\widetilde{R}_T(\mathcal{A}, j)\,, \qquad \widetilde{R}_T(\mathcal{A}, \boldsymbol{q})\,, \qquad \widetilde{R}_T\big(\mathcal{A}, j_1^T\big)\,, \qquad \text{and} \qquad \widetilde{R}_T\big(\mathcal{A}, \boldsymbol{q}_1^T\big)\,.$$

Example 1 In the case of the square loss, we have

$$\ell_t(\boldsymbol{q}) = \left(\sum_{j \in E_t} q_j f_{j,t} - y_t\right)^2,$$

so that

$$\widetilde{\ell}_t(\boldsymbol{q}) = 2\left(\sum_{j \in E_t} p_{j,t} f_{j,t} - y_t\right) \sum_{j \in E_t} q_j f_{j,t}\,.$$

Parameter: learning rate $\eta > 0$
Initialization: $(w_{1,0}, \ldots, w_{N,0}) = (1, \ldots, 1)$
For each round $t = 1, 2, \ldots, T$,

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

$$\boldsymbol{p}_t = \frac{1}{\sum_{k \in E_t} w_{k,t-1}} \Big(w_{1,t-1} \mathbb{I}_{\{1 \in E_t\}},\ \ldots,\ w_{N,t-1} \mathbb{I}_{\{N \in E_t\}} \Big) ;$$

(2) observe y_t and perform the update, for each $j = 1, \ldots, N$,

$$w_{j,t} = \begin{cases} w_{j,t-1}\, e^{\eta \left(\ell_t(\boldsymbol{p}_t) - \ell_t(\delta_j) \right)} & \text{if } j \in E_t, \\ w_{j,t-1} & \text{if } j \notin E_t. \end{cases}$$

Fig. 1 The exponentially weighted average aggregation rules $\mathcal{E}_\eta$ and $\mathcal{E}_\eta^{\mathrm{grad}}$; the first rule corresponds to the choice in (2) of the loss function ℓ_t and the second rule, to the replacement in (2) of the two occurrences of ℓ_t by the pseudo-loss $\widetilde{\ell}_t$ defined in Section 2.2.4.

2.3 Three families of aggregation rules minimizing the regret

We now recall or show 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 observations $y_1, \ldots, y_T$ and of experts forecasts. In some sense, the presented bounds are deterministic as they do not rely on any stochastic model that would generate the observations $y_1, \ldots, y_T$.

2.3.1 Exponentially weighted average aggregation rules

Basic version The exponentially weighted average aggregation rule relies on a parameter $\eta > 0$ and will thus be denoted by $\mathcal{E}_\eta$. It uses at time instance t the convex weight vector $\boldsymbol{p}_t$ given by

$$p_{j,t} = \frac{e^{\eta R_{t-1}(\mathcal{E}_\eta, j)}\, \mathbb{I}_{\{j \in E_t\}}}{\sum_{k \in E_t} e^{\eta R_{t-1}(\mathcal{E}_\eta, k)}} , \tag{7}$$

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 $\boldsymbol{p}_1$ is simply the uniform distribution over E_1. Its implementation is recalled in Figure 1.

The following performance bound was almost stated in Cesa-Bianchi and Lugosi [2003]; in any case it is a straightforward consequence of the results presented therein (Corollary 2 and the methodology followed in Sections 3 and 6.2).

Proposition 1 *We assume that the loss functions ℓ_t are convex and uniformly bounded; we denote by L a uniform bound on the quantities $|\ell_t(\boldsymbol{p}) - \ell_t(\boldsymbol{q})|$ when $\boldsymbol{p}$ and $\boldsymbol{q}$ vary in $\mathcal{X}$ and t varies from 1 to T. The regret of $\mathcal{E}_\eta$ is bounded over all such sequences of*

experts forecasts and of observations as

$$\max_{j=1,\dots,N} R_T(\mathcal{E}_\eta, j) \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} L^2 T \,. \tag{8}$$

The (theoretically) optimal choice $\eta^\star = \sqrt{(2\ln N)/(L^2 T)}$ leads to the uniform bound $L\sqrt{2T\ln N}$ on the regret of $\mathcal{E}_{\eta^\star}$. This choice depends on the horizon T and of the bound L, which are not always known in advance; standard techniques, like the doubling trick or time-varying learning rates η_t can be used to cope with these limitations, see Auer et al. [2002], Cesa-Bianchi et al. [2007].

Proof The cited performance bound is exactly the result stated in [Cesa-Bianchi and Lugosi, 2003, Corollary 2]; we therefore only need to check that its assumption is satisfied, that is, that for all $t \geqslant 2$, the combined forecast

$$\widehat{y}_t = \sum_{j \in E_t} p_{j,t} f_{j,t}$$

is such that for all $f_{j,t}$ and y_t,

$$\sum_{j=1}^{N} e^{\eta R_{t-1}(\mathcal{E}_\eta, j)} \Big(\big(\ell_t(\boldsymbol{p}_t) - \ell_t(\delta_j) \big) \mathbb{I}_{\{j \in E_t\}} \Big) \leqslant 0 \,.$$

This is immediate by the definition (7) of $\boldsymbol{p}_t$ and the convexity of ℓ_t.

Remark 1 We also studied another aggregation rule based on exponentially weighted averages, called $\mathcal{H}$ (which stands for Hedge). In [Blum and Mansour, 2007, Section 6] it was originally stated in the setting of randomized prediction, which corresponds to loss functions linear in the convex weight vectors $\boldsymbol{p}$. It can however be extended in a straightforward manner to convex losses (while preserving the regret bound); see, e.g., [Devaine et al., 2009, Section 2.1] for the details. We checked in the mentioned reference that the practical performance of both rules were equal. This is because the rules are almost identical: the rule $\mathcal{H}$ simply replaces the update in step (2) of Figure 1 by

$$w_{j,t-1}\, e^{\eta_j \left(e^{-\eta_j \ell_t(\boldsymbol{p}_t) - \ell_t(\delta_j)} \right)} \,,$$

where the learning rates η_j now depend on the experts $j = 1, \dots, N$. By carefully setting these rates, uniform regret bounds of the form

$$R_T(\mathcal{H}, j) = \mathcal{O}\left(L \sqrt{\sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}} \ln N} + L \ln N \right)$$

can be obtained. They are more precise than the bound of Proposition 1, which yields a bound of order $L\sqrt{T\ln N}$ that is uniform over experts and does not take into account how many times each expert was active. However, first, the rule stated in Figure 1 seems more natural in practice since it gives the same weight to the losses encountered by the experts and by itself; and second, in view of the goal (3), the uniform bound of order $\sqrt{T}$ is enough for our purpose.

Gradient version 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 uses

$$p_{j,t} = \frac{e^{\eta \widetilde{R}_{t-1}(\mathcal{E}_\eta^{\mathrm{grad}}, j)}\, \mathbb{I}_{\{j \in E_t\}}}{\sum_{k \in E_t} e^{\eta \widetilde{R}_{t-1}(\mathcal{E}_\eta^{\mathrm{grad}}, k)}}\,. \tag{9}$$

By Section 2.2.4, the following result is almost a corollary of Proposition 1; note that the obtained regret bound is however with respect to all convex weight vectors.

Corollary 1 *We assume that the loss functions ℓ_t have subgradients at all points of $\mathcal{X}$, uniformly bounded in the supremum norm (as t varies) by G. The regret of $\mathcal{E}_\eta^{\mathrm{grad}}$ is bounded over all such sequences of experts forecasts and of observations as*

$$\max_{\boldsymbol{q} \in \mathcal{X}} R_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, \boldsymbol{q}\big) \leqslant \frac{\ln N}{\eta} + 2\eta G^2 T\,.$$

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

Proof First, we recall that by Section 2.2.4,

$$\max_{\boldsymbol{q} \in \mathcal{X}} R_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, \boldsymbol{q}\big) \leqslant \max_{\boldsymbol{q} \in \mathcal{X}} \widetilde{R}_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, \boldsymbol{q}\big) = \max_{\boldsymbol{q} \in \mathcal{X}} \sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t\big(\boldsymbol{q}^{E_t}\big)\Big)\, \boldsymbol{q}(E_t)\,.$$

Since the $\widetilde{\ell}_t$ are linear over $\mathcal{X}$ and by definition of the $\boldsymbol{q}^{E_t}$, the last expression simplifies to

$$\max_{\boldsymbol{q} \in \mathcal{X}} \widetilde{R}_T\big(\mathcal{E}_\eta^{\mathrm{grad}}, \boldsymbol{q}\big) = \max_{\boldsymbol{q} \in \mathcal{X}} \sum_{j=1}^{N} q_j \sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\delta_j)\Big)\, \mathbb{I}_{\{j \in E_t\}}\,.$$

The proof is concluded by noting that Proposition 1 ensures that the rule $\mathcal{E}_\eta^{\mathrm{grad}}$ is such that

$$\max_{j=1,\dots,N} \sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\delta_j)\Big)\, \mathbb{I}_{\{j \in E_t\}} \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2}(2G)^2 T\,.$$

Instantiation to the square loss The loss functions ℓ_t and $\widetilde{\ell}_t$ were indicated in Example 1; the constants appearing in Proposition 1 and Corollary 1 equal in this case $L = B^2$ and $G = 2B^2$.

Parameters: learning rate $\eta > 0$

Initialization: $\boldsymbol{w}_1$ is the uniform convex weight vector, $w_{i,1} = 1/N$ for $i = 1, \ldots, N$

For each time instance $t = 1, 2, \ldots, T$,

(1) predict $\widehat{y}_t = \sum_{j \in E_t} w_{j,t}^{E_t} f_{j,t}$, that is, resort to the convex weight vector $\boldsymbol{p}_t = \boldsymbol{w}_t^{E_t}$;

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

$$w_{i,t+1} = \begin{cases} w_{i,t}\, e^{-\eta \widetilde{\ell}_t(\delta_i)} \dfrac{\sum_{j \in E_t} w_{j,t}}{\sum_{k \in E_t} w_{k,t}\, e^{-\eta \widetilde{\ell}_t(\delta_k)}} & \text{if } i \in E_t, \\ p_{i,t} & \text{if } i \notin E_t. \end{cases}$$

Fig. 2 The specialist aggregation rule $\mathcal{S}_\eta$; it uses the pseudo-loss $\widetilde{\ell}_t$ defined in Section 2.2.4 and based on the loss functions ℓ_t.

2.3.2 The specialist aggregation rule

Gradient version The content of this paragraph provides a new look at the results of [Freund et al., 1997, Sections 3.2–3.4]. In the latter reference, a general rule was introduced but it had to be instantiated to each specific loss function; in particular, its analysis and its regret bound heavily depended on the specific loss function at hand and possibly on the learning rate η.

We show how the algorithm SEG described therein for the case of the square loss can be generalized to any convex loss function, via the subgradient trick explained in Section 2.2.4. To achieve this, we replace the ad hoc inequalities needed therein[1] by a new general bound, provided by Lemma 1 below and based on Hoeffding's lemma. The rest of the structure of the proof is borrowed from Freund et al. [1997].

The specialist aggregation rule is described in Figure 2; it relies on a parameter $\eta > 0$ and will be denoted by $\mathcal{S}_\eta$. It is close to but different from the rule $\mathcal{E}_\eta^{\mathrm{grad}}$: as we will see below, the two rules have comparable theoretical guarantees, their statements might be found to exhibit some similarity as well, but we noted that in practice the output convex weight vectors $\boldsymbol{p}_t$ had nothing in common (even though their achieved performance were almost equal).

The performance guarantees of the rule $\mathcal{S}_\eta$ are, like the ones of $\mathcal{E}_\eta^{\mathrm{grad}}$, with respect to all fixed convex combinations of the experts; the regret bound that we could prove for $\mathcal{S}_\eta$ is however slightly smaller than the one exhibited for $\mathcal{E}_\eta^{\mathrm{grad}}$.

Theorem 1 *We assume that the loss functions ℓ_t have subgradients at all points of $\mathcal{X}$, uniformly bounded in the supremum norm (as t varies) by G. The regret of $\mathcal{S}_\eta$ is bounded over all such sequences of experts forecasts and of observations as*

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

The (theoretically) optimal choice $\eta^\star = \sqrt{(2 \ln N)/(G^2 T)}$ leads to the uniform bound $G\sqrt{2T \ln N}$ on the regret of $\mathcal{S}_{\eta^\star}$. The same comments on the calibration of η as in the previous section apply.

[1] See equation (6) in Freund et al. [1997] and the comments after its statement: "Here, a and b are positive constants which depend on the specific on-line learning problem [...]."

In the course of the proof of the theorem above, we will need the following lemma, which is the key to a general analysis independent of the specific loss functions at hand.

Lemma 1 *Let $d \geqslant 2$ be an integer. Fix two convex weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ over d elements and a vector $\gamma \in \mathbb{R}^d$; we assume that $v_i > 0$ for all $i = 1, \dots, d$ but do not constrain $\boldsymbol{u}$. Denote by G a positive real number such that $-G \leqslant \gamma_i \leqslant G$ for all $i = 1, \dots, d$. Consider the convex weight vector $\boldsymbol{v}'$ defined as follows: for all $i = 1, \dots, d$,*

$$v'_i = \frac{v_i \, e^{-\eta\gamma_i}}{\sum_{k=1}^d v_k \, e^{-\eta\gamma_k}} \, .$$

Then

$$\eta \sum_{i=1}^d (v_i - u_i)\gamma_i \leqslant \frac{\eta^2 G^2}{2} + \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) - \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}') \, .$$

Proof By direct calculations and since v_i, $v'_i > 0$ for all $i = 1, \dots, d$,

$$\begin{aligned} \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) - \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}') = \sum_{i=1}^d u_i \ln \frac{v'_i}{v_i} &= \sum_{i=1}^d u_i \ln \frac{e^{-\eta\gamma_i}}{\sum_{k=1}^d v_k \, e^{-\eta\gamma_k}} \\ &= -\ln \left(\sum_{k=1}^d v_k \, e^{-\eta\gamma_k} \right) - \eta \sum_{i=1}^d u_i \gamma_i \, . \end{aligned}$$

The first term in the last expression can be bounded by Hoeffding's lemma (see, e.g., [Cesa-Bianchi and Lugosi, 2006, Lemma A.1]),

$$\ln \left(\sum_{k=1}^d v_k \, e^{-\eta\gamma_k} \right) \leqslant -\eta \sum_{k=1}^d v_k \gamma_k + \eta^2 \frac{(2G)^2}{8} \, ,$$

and this concludes the proof.

Proof (of Theorem 1) We first note that by induction, $w_{j,t} > 0$ for all j and t, and we recall that by Section 2.2.4,

$$\max_{\boldsymbol{q} \in \mathcal{X}} R_T(\mathcal{S}_\eta, \boldsymbol{q}) \leqslant \max_{\boldsymbol{q} \in \mathcal{X}} \widetilde{R}_T(\mathcal{S}_\eta, \boldsymbol{q}) = \max_{\boldsymbol{q} \in \mathcal{X}} \sum_{t=1}^T \Big(\widetilde{\ell}_t\big(\boldsymbol{w}_t^{E_t}\big) - \widetilde{\ell}_t\big(\boldsymbol{q}^{E_t}\big) \Big) \, \boldsymbol{q}(E_t) \, ,$$

where we used the notation of Figure 2. Now, for all instances $t \geqslant 1$, by definition of the rule $\mathcal{S}_\eta$, we have $\boldsymbol{w}_{t+1}(E_t) = \boldsymbol{w}_t(E_t)$ and thus, for all $j \in E_t$,

$$w_{j,t+1}^{E_t} = \frac{w_{j,t}^{E_t} \, e^{-\eta \widetilde{\ell}_t(\delta_j)}}{\sum_{k \in E_t} w_{k,t}^{E_t} \, e^{-\eta \widetilde{\ell}_t(\delta_k)}} \, ,$$

so that Lemma 1 shows that

$$\widetilde{\ell}_t\big(\boldsymbol{w}_t^{E_t}\big) - \widetilde{\ell}_t\big(\boldsymbol{q}^{E_t}\big) \leqslant \frac{\eta G^2}{2} + \frac{1}{\eta} \Big(\mathcal{K}\big(\boldsymbol{q}^{E_t}, \boldsymbol{w}_t^{E_t}\big) - \mathcal{K}\big(\boldsymbol{q}^{E_t}, \boldsymbol{w}_{t+1}^{E_t}\big) \Big) \, .$$

Substituting this inequality in the first upper bound on the regret, we get

$$\max_{\boldsymbol{q}\in\mathcal{X}} R_T(\mathcal{S}_\eta,\boldsymbol{q}) \leqslant \max_{\boldsymbol{q}\in\mathcal{X}} \frac{\eta G^2 T}{2} + \frac{1}{\eta}\sum_{t=1}^{T} \boldsymbol{q}(E_t)\left(\mathcal{K}\big(\boldsymbol{q}^{E_t},\boldsymbol{w}_t^{E_t}\big) - \mathcal{K}\big(\boldsymbol{q}^{E_t},\boldsymbol{w}_{t+1}^{E_t}\big)\right).$$

Using again, for the second equality below, that $\boldsymbol{w}_{t+1}(E_t) = \boldsymbol{w}_t(E_t)$, we rewrite the summands as

$$\begin{aligned}
&\boldsymbol{q}(E_t)\left(\mathcal{K}\big(\boldsymbol{q}^{E_t},\boldsymbol{w}_t^{E_t}\big) - \mathcal{K}\big(\boldsymbol{q}^{E_t},\boldsymbol{w}_{t+1}^{E_t}\big)\right)\\
&\quad= \quad \boldsymbol{q}(E_t)\sum_{j\in E_t} q_j^{E_t}\ln\frac{w_{j,t+1}^{E_t}}{w_{j,t}^{E_t}}\\
&\quad= \quad \sum_{j\in E_t} q_j \ln\frac{w_{j,t+1}}{w_{j,t}} = \sum_{j=1}^{N} q_j\ln\frac{w_{j,t+1}}{w_{j,t}} = \mathcal{K}(\boldsymbol{q},\boldsymbol{w}_t) - \mathcal{K}(\boldsymbol{q},\boldsymbol{w}_{t+1}),
\end{aligned}$$

where the third equality follows from the fact that, by definition of the rule, $w_{j,t+1} = w_{j,t}$ whenever $j \notin E_t$. After substitution, a telescoping sum appears and we are thus left with

$$\max_{\boldsymbol{q}\in\mathcal{X}} R_T(\mathcal{S}_\eta,\boldsymbol{q}) \leqslant \frac{\eta G^2 T}{2} + \frac{1}{\eta}\max_{\boldsymbol{q}\in\mathcal{X}}\big\{\mathcal{K}(\boldsymbol{q},\boldsymbol{w}_1) - \mathcal{K}(\boldsymbol{q},\boldsymbol{w}_{T+1})\big\} \leqslant \frac{\eta G^2 T}{2} + \frac{\ln N}{\eta},$$

where the last inequality is by nonnegativity of the Kullback-Leibler divergence and by the fact that $\boldsymbol{w}_1$ is the uniform convex weight vector, hence $\mathcal{K}(\boldsymbol{q},\boldsymbol{w}_1) \leqslant \ln N$ for all $\boldsymbol{q}\in\mathcal{X}$.

Instantiation to the square loss The loss functions ℓ_t and $\widetilde{\ell}_t$ were indicated in Example 1; the constant appearing in Theorem 1 equals in this case $G = 2B^2$.

2.3.3 Fixed-share aggregation rules

Here, as in Section 2.3.1, 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.

Basic version The rule presented in Figure 3 (when used directly on the losses) 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 prior distribution. We will call it the fixed-share rule for specialized experts; 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].

Its performance bound is stated below; it follows from a straightforward but lengthy adaptation of the techniques used in Herbster and Warmuth [1998] and [Cesa-Bianchi and Lugosi, 2006, Section 5.2]. We thus provide it in the appendix of this paper (Section 7), for the sake of completeness and to show how the share update of Figure 3 was obtained.

Parameters: learning rate $\eta > 0$ and mixing rate $0 \leqslant \alpha \leqslant 1$

Initialization: $(w_{1,0}, \dots, w_{N,0}) = \dfrac{1}{|E_1|}\big(\mathbb{I}_{\{1\in E_1\}}, \dots, \mathbb{I}_{\{N\in E_1\}}\big)$

For each round $t = 1, 2, \dots, T$,

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

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

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

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

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

if $j \in E_{t+1}$, with the convention that an empty sum is null and denoting by $|E_{t+1}|$ the cardinality of E_{t+1}.

Fig. 3 The fixed-share aggregation rules $\mathcal{F}_{\eta,\alpha}$ (basic version, when implemented as defined above) and $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$ (gradient version, when the loss ℓ_t in the loss update is replaced by the pseudo-losses $\widetilde{\ell}_t$ defined in Section 2.2.4).

Proposition 2 *We assume that the loss functions ℓ_t are convex and uniformly bounded; we denote by L a uniform bound on the quantities $|\ell_t(\boldsymbol{p}) - \ell_t(\boldsymbol{q})|$ when $\boldsymbol{p}$ and $\boldsymbol{q}$ vary in $\mathcal{X}$ and t varies from 1 to T. For all $m \in \{0, \dots, T-1\}$, the regret of $\mathcal{F}_{\eta,\alpha}$ is uniformly bounded over all such sequences of observations and of experts forecasts as*

$$\max_{j_1^T \in \mathcal{L}_m} R_T\big(\mathcal{F}_{\eta,\alpha}, j_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m\, (1-\alpha)^{T-m-1}} + \frac{\eta}{8} L^2 T\,. \tag{10}$$

The (theoretically almost) optimal bound in the proposition above 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

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

it is given by

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

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.

Gradient version We proceed here as in Section 2.3.1 and consider a variant of the previous forecaster that is based on the (sub)gradients 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 3 to define this gradient version is to replace the update in step (2) by

$$v_{i,t} = w_{i,t-1}\, e^{-\widetilde{\ell}_t(\delta_i)}$$

when $i \in E_t$. The following performance bound is almost a corollary of Proposition 2; here again, for the sake of completeness, a proof is provided in appendix (in Section 7).

Corollary 2 *We assume that the loss functions ℓ_t have subgradients at all points of $\mathcal{X}$, uniformly bounded in the supremum norm (as t varies) by G. For all $m \in \{0, \ldots, T-1\}$, the regret of $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$ is uniformly bounded over all such sequences of observations and of experts forecasts as*

$$\max_{\boldsymbol{q}_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}, \boldsymbol{q}_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m\,(1-\alpha)^{T-m-1}} + \frac{\eta}{2} G^2 T\,. \qquad (11)$$

The above bound can here also be (almost) optimized as before, via suitable choices $\eta^\star$ and $\alpha^\star$ for η and α,

$$\max_{\boldsymbol{q}_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}^{\mathrm{grad}}_{\eta^\star,\alpha^\star}, \boldsymbol{q}_1^T\big) \leqslant G\sqrt{2T\Big((m+1)\ln N + (T-1)H\big(m/(T-1)\big)\Big)}\,.$$

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

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 change over time.

Instantiation to the square loss The loss functions ℓ_t and $\widetilde{\ell}_t$ were indicated in Example 1; the constants appearing in Proposition 1 and Corollary 1 equal in this case $L = B^2$ and $G = 2B^2$.

3 Methodology followed in the empirical studies

We provide a standardized outline of the treatment of the two data sets discussed in the next sections.

3.1 A methodology in four steps

The aggregation rules discussed above are only semi-automatic strategies, as they rely on fixed-in-advance parameters that are not tuned on data. Our ultimate aim in this article is to design operational (i.e., fully sequential) aggregation rules, which set these parameters online, as is explained below.

To evaluate the experts used and the performance of the semi-automatic and fully automatic strategies, we of course use data sets with all observations available for the period of interest but proceed as if we had to do so sequentially. In particular, the experts are constructed independently of the data sets used for the assessment of the strategies.

We now state and describe in details our methodology, which takes place in four steps.

Outline of the empirical studies of performance of the sequential aggregation rules

1. Describe the data set and design some experts.
2. Choose a loss function and evaluate the performance of the experts.
3. For each family of strategies compute the performance corresponding to the best constant choices of the parameters in hindsight.
4. Measure the cost of some automatic and sequential tuning and assess the quality of the operational performance.

1. Design some experts The guideline is to design them so that –as much as possible– they exhibit varied enough behaviors (for the aggregation strategies to have a sufficient flexibility in the output aggregated forecasts). Constructing the experts is usually the responsibility of the partner of the learning theorist because of his knowledge of the field of application and of the methods –classical and more modern ones– that are likely to exhibit a good performance. These methods can rely on some tuning parameters that were set on data sets anterior to the data set at hand (see, for instance, the construction of the experts in Section 5.1.1).

2. Choose a loss function and evaluate the performance of the experts By evaluation of the performance of the experts we mean the assessment of the accuracy obtained by some simple strategies like the uniform average of the forecasts of the active experts (which is a strategy that is easy to implement online) or by some oracles; this assessment is given by their cumulative losses. By oracles we mean strategies that cannot be defined online and that require the beforehand knowledge of the whole data set: e.g., the best single expert and the best constant convex combination of the experts. Finally, the so-called prescient strategy is the strategy that picks at each time instance the best experts forecast; it indicates a bound on the performance that no aggregation strategy can improve given the data set (given the experts forecasts and the observations). It corresponds to the best element in $\mathcal{L}_{T-1}$.

3. For each family of strategies compute the performance corresponding to the best constant choices of the parameters in hindsight The aggregation strategies described in Section 2 require the tuning of a small number (one or two) of user parameters. What we do here is to tabulate the performance on a thin grid of possible parameters and compare the best accuracy obtained in this way to the performance of the reference strategies and oracles discussed above –with the hope that the aggregation strategies will perform almost as well as or even better than these oracles in addition of being implementable online.

4. Measure the cost of the automatic tuning and assess the quality of the operational performance We then implement the meta-rule discussed in the next section and which is based on the families considered in the previous step: instead of considering fixed-in-advance parameters, it tunes them sequentially. We measure how different is its performance with respect to the best member of the underlying family. This is the most crucial step of the empirical study since it indicates the performance that would have been achieved for real by outputting sequentially aggregated forecasts based on the experts constructed in the first step –hence the notion of operational performance.

3.2 Sequential automatic tuning of the parameters on data

We explain in this section how fully sequential aggregation rules can be designed; we describe first the method in a general framework. Let $\mathcal{A}_\lambda$ be a sequential aggregation rule relying on some 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 instance t a convex weight vector which we denote by $\boldsymbol{p}_t\big(\mathcal{A}_\lambda\big)$. 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 theoretical bounds on the regret tend to show poor practical performance), we will resort in practice to the following method, due to Vivien Mallet and proposed in the technical report Gerchinovitz et al. [2008] (but never published elsewhere to the best of our knowledge).

The weights used by the fully sequential aggregation rule based on the family of rules $\mathcal{A}_\lambda$, where $\lambda \in \Lambda$, will be denoted by $\widehat{\boldsymbol{p}}_t$. We assume that the considered family is such that $\boldsymbol{p}_1\big(\mathcal{A}_\lambda\big)$ is independent of λ, so that $\widehat{\boldsymbol{p}}_1$ equals this common value. Then, at time instances $t \geqslant 2$,

$$\widehat{\boldsymbol{p}}_t = \boldsymbol{p}_t\Big(\mathcal{A}_{\widehat{\lambda}_{t-1}}\Big) \qquad \text{where} \qquad \widehat{\lambda}_{t-1} \in \operatorname*{argmin}_{\lambda\in\Lambda} \sum_{s=1}^{t-1} \ell_s\Big(\boldsymbol{p}_s\big(\mathcal{A}_\lambda\big)\Big)\,; \tag{12}$$

that is, we consider, for the prediction of the next time instance, the aggregated forecast proposed by the best so far member of the family of aggregation rules. Because of this formulation, we will speak of a meta-rule in the sequel. For the time being, we can offer no theoretical guarantee for the performance of the meta-rule in terms of the performance of the underlying family (this is a work in progress).

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 aggregation rules and the specialist aggregation rule) 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 (12) only on the elements of $\widetilde{\Lambda}$ instead of performing it on the whole set Λ.

Of course, some choices are still left to the user, namely, how to design this finite grid $\widetilde{\Lambda}$ and thus, the proposed procedure is not fully automatic yet. We however checked (see [Devaine et al., 2009, Section 3.1] for the details) that in practice the performance was not too sensitive to the design of $\widetilde{\Lambda}$, as soon as the latter covers a range large enough. Since the optimal value $\lambda^\star$ prescribed by theory is usually too conservative in practice, a reasonable procedure is, e.g., to start the grid around some small value likely to be close to $\lambda^\star$ and then take logarithmically evenly 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 adaptively these lower and upper bounds is explained in Section 5.4; for the reader to appreciate it we however need first to illustrate several times the on-line calibration of the parameters with fixed finite grids and this is why we will consider some seemingly arbitrary grids in the first empirical studies.

4 A first data set: Slovakian consumption data

The study is divided in four subsections, following the methodology in four steps described above.

4.1 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 (formed by hourly mean consumptions) on the period from January 1, 2005 to December 31, 2007. That is, there are 24 series (one for each hour) of 1 095 hourly mean consumptions 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 instances. 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 reason why we parsed the data set into 24 subsets (one per hour interval of the day) is that first, for this data set we have enough observations to do so (we have 1 095 observations per given hour frame, that is, we spit the data set into 24 data subsets of size 1 095); and second, 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 mostly report in this section and in the next ones the results obtained for one-day-ahead prediction on a given (somewhat arbitrarily chosen) hour interval: the interval 11:00–12:00. The characteristics of the observations y_t of this hour frame

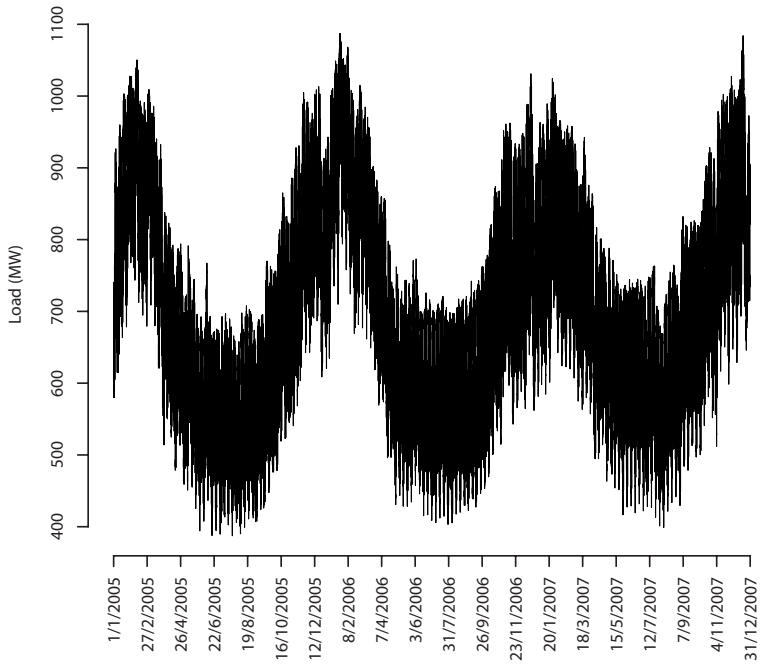

Fig. 4 The observed hourly electricity consumptions encountered by the Slovakian subbranch between January 1, 2005 and December 31, 2007.

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

Number of days D	1 095
Time intervals	Only 11:00–12:00
Number of instances T	1 095 ($= 1\,095 \times 1$)
Number of experts N	35
Unit	MW
Median of the y_t	702.6
Bound B on the y_t	1020.0

are described in Table 1 while all observations (for all hour frames) are plotted in Figure 4.

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 loss function used is the square loss and we will mostly report root mean square errors (RMSE), whose definition was given in Section 2.2.

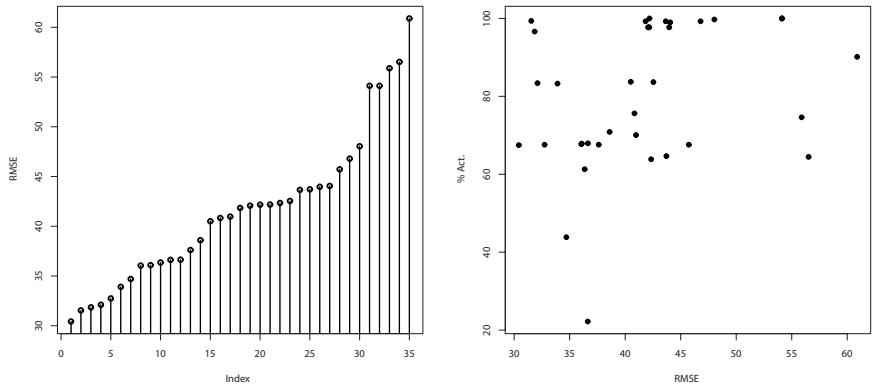

Fig. 5 Graphical representations of the performance of the experts of the Slovakian data set: sorted RMSE (left) and RMSE–frequency of activity pairs (right).

4.2 Benchmark values: performance of the experts and of some oracles

We consider here only the observations and predictions that correspond to the hour frame 11:00–12:00. The characteristics of the experts are depicted in Figure 5. The bar plot represents the values of the RMSE of the 35 available experts; we computed the 35 values of $\text{RMSE}(j)$, one for each expert j, and ordered them. The scatter plot relates the RMSE of each of the expert to its frequency of activity, that is, it plots the pairs

$$\left(\text{RMSE}(j),\ \frac{\sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}}}{T} \right) \tag{13}$$

for all experts j.

We present in Table 2 the values[2] of the RMSE of several procedures, some of them being off-line procedures using the whole data set (i.e., observations and predictions) in hindsight. Actually, except the use of the uniform convex weight vector in $\mathcal{X}$ and the uniform sequential aggregation rule $\mathcal{U}$, none of these procedures can be implemented sequentially, and this is why they are called oracles.

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

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

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

[2] 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.

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

Name of the benchmark procedure	Formula	Value
Uniform sequential aggregation rule	$\textsc{rmse}(\mathcal{U})$	$= 31.1$
Uniform convex weight vector	$\textsc{rmse}\big((1/35, \ldots, 1/35)\big)$	$= 30.7$
Best single expert	$\min_{j=1,\ldots,35} \textsc{rmse}(j)$	$= 30.4$
Best convex weight vector	$\min_{\boldsymbol{q}\in\mathcal{X}} \textsc{rmse}(\boldsymbol{q})$	$= 29.2$
Best compound expert		
Size at most $m = 10$	$\min_{j_1^T\in\mathcal{L}_{10}} \textsc{rmse}\big(j_1^T\big)$	$= 32.1$
Size at most $m = 50$	$\min_{j_1^T\in\mathcal{L}_{50}} \textsc{rmse}\big(j_1^T\big)$	$= 23.1$
Size at most $m = 200$	$\min_{j_1^T\in\mathcal{L}_{200}} \textsc{rmse}\big(j_1^T\big)$	$= 15.2$
Size at most $m = T - 1 = 1\,094$	$\min_{j_1^T\in E_1\times E_2\times\ldots\times E_T} \textsc{rmse}\big(j_1^T\big)$	$= 9.4$
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 (14)	$= 29.1$
Best convex weight vector on each element	See (15)	$= 24.5$

The oracle with the least error is the one that can pick at each time instance 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 typically, 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[3] size m, and on the other hand, by the individual performance of some fixed convex weight vectors (not necessarily evaluated on all time instances): 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 2 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

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

some overall good experts refrain from predicting at some time instances when all active experts perform poorly, while compound experts are required to output a prediction at each time instance even when an accurate prediction is likely to be difficult to perform. The fact that such good experts tend not to form predictions at instances that are more difficult to cope with can also be seen from the fact that $\text{RMSE}(\mathcal{U})$ is larger than $\text{RMSE}\big((1/35,\dots,1/35)\big)$, since the second average puts unequal weights to the time instances, with more weight to time instances when more experts are active.

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

$$\min\left\{\sqrt{\frac{1}{T}\sum_{k=1}^{K}\sum_{t:E_t=E^{(k)}}\big(f_{j^k,t}-y_t\big)^2},\ \text{ with } j^k\in E^{(k)}\ \text{ for all } k=1,\dots,K\right\}, \tag{14}$$

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

$$\min\left\{\sqrt{\frac{1}{T}\sum_{k=1}^{K}\sum_{t:E_t=E^{(k)}}\left(\sum_{j\in E^{(k)}} q_j^{(k)} f_{j,t}-y_t\right)^2},\right.$$
$$\left.\text{with } \boldsymbol{q}^{(k)} \text{ a convex weight vector on } E^{(k)}\ \text{ for all } k=1,\dots,K\right\}, \tag{15}$$

which corresponds to the choice of the best convex weight vector on each element of the partition. Even if there are relatively many elements in this partition, 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 $m=50$).

4.3 Results obtained by the considered sequential aggregation rules: With constant values of the parameters

We now detail the practical performance of the sequential aggregation rules introduced in Section 2 and compare it to the one of the oracles. As indicated in Section 3, we will proceed in two steps. First, we report in this subsection the results obtained for fixed values of the parameters η and α of the rules; to do so, we considered a grid of their 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

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

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

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

procedures yet. We will study the fully sequential procedures in the next subsection.

The performance of the families $\mathcal{E}_\eta$, $\mathcal{E}_\eta^{\text{grad}}$, and $\mathcal{S}_\eta$ are 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 $\mathcal{E}_\eta^{\text{grad}}$ and $\mathcal{S}_\eta$). We recall that these are indicated in Table 2. We note that $\mathcal{E}_\eta^{\text{grad}}$ 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 $\eta^\star \approx 8\times10^{-8}$ for $\mathcal{E}_\eta$ and $\eta^\star \approx 4\times10^{-8}$ for $\mathcal{E}_\eta^{\text{grad}}$ 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 $\mathcal{F}_{\eta,\alpha}^{\text{grad}}$ 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 6 that the considered rules fully exploit the whole set of experts and do not concentrate on a limited 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 weights do not converge to a limit vector.

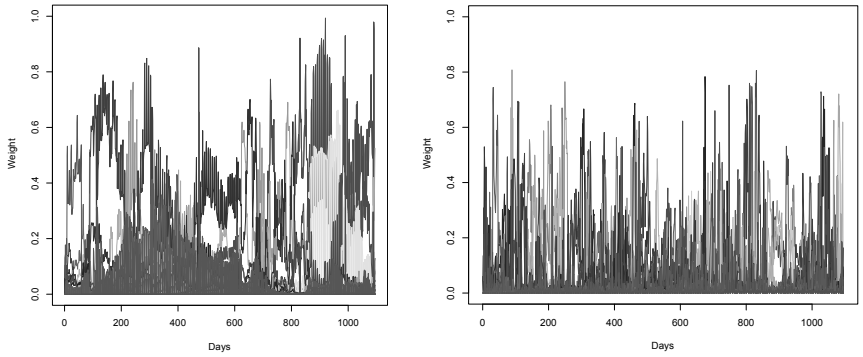

Fig. 6 Graphical representations of the convex weights associated at each time instance with the 35 experts by the rules $\mathcal{E}^{\mathrm{grad}}_{10^{-4}}$ (left) and $\mathcal{F}_{2\times 10^{-3},0.2}$ (right).

Table 5 Performance obtained by the rules $\mathcal{E}_\eta$ and $\mathcal{E}^{\mathrm{grad}}_\eta$ 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).

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

4.4 Results obtained with an on-line calibration of the parameters

We show in this section that the fully sequential aggregation rules (the meta-rules) constructed in Section 3.2 can get performance close to the one of the rules studied in the previous section and which were based on the choices of the best constant parameters in hindsight.

Application to the exponentially weighted average rules $\mathcal{E}_\eta$ and $\mathcal{E}^{\mathrm{grad}}_\eta$ 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 logarithmically evenly spaced points,

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

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

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

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 7.

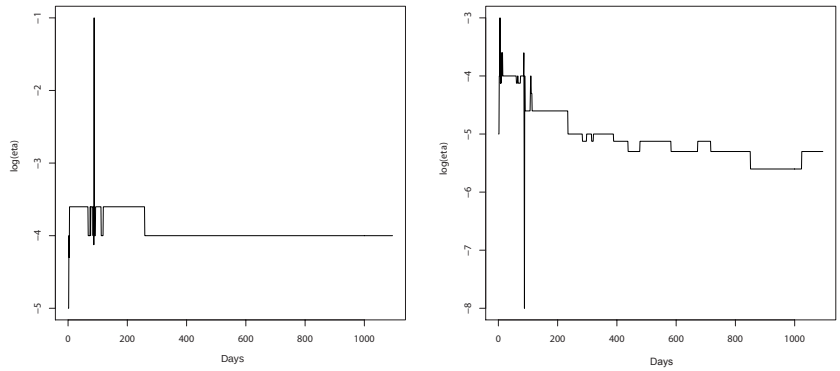

Fig. 7 Graphical representations of the sequences of tuning parameters η chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_\ell$; the base rules are $\mathcal{E}_\eta^{\mathrm{grad}}$ (left) and $\mathcal{E}_\eta$ (right).

Table 6 Performance obtained by the rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ 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).

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

Application to the fixed-share type rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ 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,

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

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 8.

5 A second data set: Operational forecasting on French data

We use again the methodology in four steps described in Section 3, whose results are reported in Sections 5.1–5.4. We also provide a robustness study in Section 5.5.

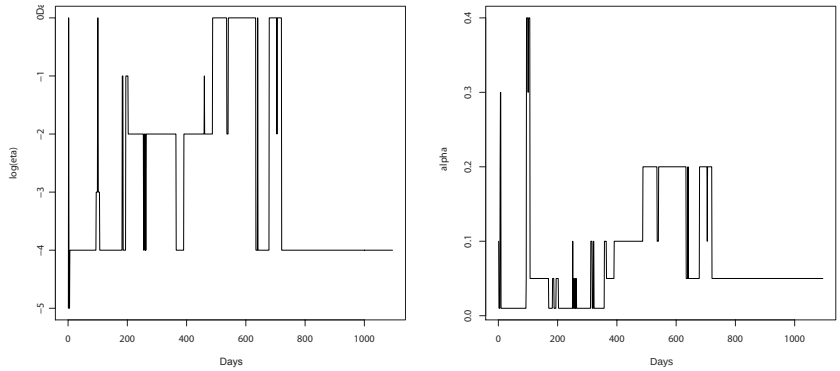

Fig. 8 Graphical representations of the sequences of tuning parameters η (left) and α (right) chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_{\mathrm{FS}}$; the base rule is $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$.

5.1 Presentation and characteristics of the data set; design of the experts

5.1.1 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 based on 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 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 (formed by half-hourly mean consumptions) 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. Note that this time

Table 7 Some characteristics of the observations y_t (half-hourly mean consumptions) of the French data set of operational forecasting.

Number of days D	320
Time intervals	Every 30 minutes
Time instances T	15 360 ($= 320 \times 48$)
Number of experts N	24 ($= 15 + 8 + 1$)
Unit	GW
Median of the y_t	56.33
Bound B on the y_t	92.76

we do not split anymore the data set into subsets; this is explained in details below and comes from two facts: the data set is smaller (and thus the data subsets would be too small) and we need to abide by an operational constraint.

5.1.2 Design of the experts

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 varied enough behaviors; and second, EDF could provide these three types of models. We report below a short description of the experts of each category but refer the reader to Devaine et al. [2009] for more details.

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 of decomposing the electricity load into a main component including all the seasonality effects 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 while 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., a method not requiring any exogenous

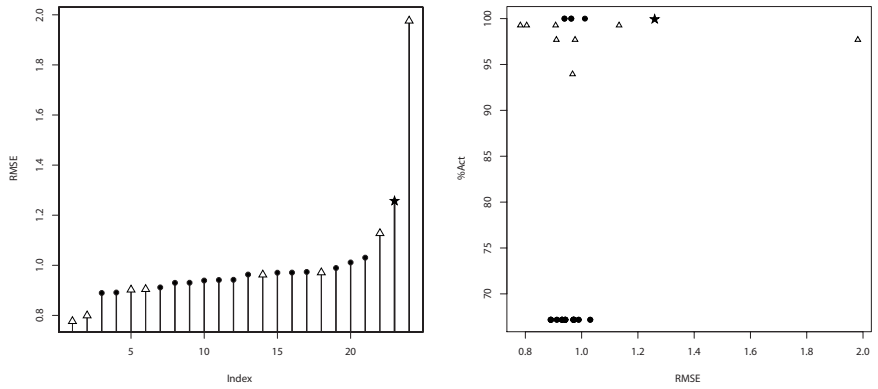

Fig. 9 Graphical representations of the performance of the experts of the French data set of operational forecasting: sorted RMSE (left) and RMSE–frequency of activity pairs (right); Eventail experts are depicted by the symbols •, GAM experts are represented by △ while ⋆ stands for the similarity expert.

factor like weather conditions); this 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 view 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 9. 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 (13).

Out of the 15 Eventail experts, 3 are active all the time; they correspond to the operational model actually used at the R&D center of EDF and to two variants of it based on different short-term corrections. The other 12 Eventail experts are inactive during the summer as their predictions are redundant with the 3 main Eventail experts (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 during periods when R&D practitioners know beforehand that they will perform poorly (e.g., in time periods close to public holidays); the lengths of these periods depend on the parameters of the expert. Finally, the similarity expert is always active.

5.1.3 Addition of an operational constraint

We still consider one-day-ahead prediction, with an operational constraint in this part: it consists of producing half-hourly forecasts every day at 12:00 for the next 24 hours (as required by the R&D services of EDF); that is, of forecasting simultaneously the next 48 time instances. The experts presented above abide by this constraint.

Of course, we could still mimic the methodology indicated in Section 4.1 and decompose the French data set into 48 smaller data sets of equal size. However,

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

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

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; we require instead that the proposed aggregation rules only output predictions (i.e., weight vectors) every 48 steps and that when doing so, they provide 48 such predictions. Section 5.3.1 explains the (slight) modifications that need to be performed for the rules described in Section 2.3 to abide by this constraint.

Note that the 48 weight vectors simultaneously output can be different; they even must be different when some experts get inactive or active during the set of 48 time instances for which predictions are to be performed.

5.2 Benchmark values: performance of the experts and of some oracles

Before doing so, we report in Table 8 the performance obtained by most of the oracles already discussed in Section 4.2. 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 4.2 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 $\mathcal{U}$. Unlike in Section 4.2, this comes from our experts being more active in challenging situations. Indeed, the rule $\mathcal{U}$ also performs better than the uniform convex weight vector, which induces at each time instance the same

forecast as the rule $\mathcal{U}$ but for which the loss incurred at a given time instance is more weighted as more experts are active. All in all, the poor performance of the best single expert or of the uniform convex weight vector are caused by the considered specialized experts being more active and more helpful when needed.

From Table 8 we mostly conclude the following. The true benchmark 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. The second part of the table indicates that important gains in accuracy are obtained with compound experts (and therefore, fixed-share type rules are expected to perform well).

5.3 Results obtained by the sequential aggregation rules: With constant parameters

5.3.1 Extension of the previous rules to operational forecasting

Before describing the detailed performance of the sequential aggregation rules (among which we consider only in this section 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 the next 24 hours, i.e., for the next 48 time instances. (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 an aggregation 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 instances of the form $t_k = 48k + 1$, and use these vectors for the next 48 time instances, by adapting them via a renormalization or a mixing to the values of the active sets $E_{t_k+1}, \ldots, E_{t_k+48}$.

We do so to be able to provide 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 (and of course, coincide with them at the time instances 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 instance t, count much less than the past losses (the ones encountered between the instances 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 experts; 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., (8) or (10), 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.

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

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

$$p_{j,t} = \frac{p_{j,0}\ e^{\eta R_{48\lfloor (t-1)/48 \rfloor}(\mathcal{E}_\eta, j)}\, \mathbb{I}_{\{j \in E_t\}}}{\sum_{k \in E_t} p_{k,0}\ e^{\eta R_{48\lfloor t/48 \rfloor}(\mathcal{E}_\eta, k)}}\,, \tag{16}$$

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

In particular, the only difference between the definitions (7) and (16) in the case of uniform initial weights $p_{j,0} = 1/24$ is that not all past losses are used at time instance t, but only the ones that were available at the time instance 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 $\boldsymbol{p}_t(\mathcal{W}_\eta)$ output by the adaptation are not too far from the weights $\boldsymbol{p}_t(\mathcal{E}_\eta)$ of the base version, thus preserving a sublinear bound on the regret of $\mathcal{W}_\eta$, as desired. We quantity this fact in the proof below.

A similar adaptation is considered for the gradient version of the exponentially weighted average rule; it suffices to consider in (9) the values of the regrets at rounds $48\lfloor (t-1)/48 \rfloor$ instead of the values at rounds $t-1$.

Proof (of a regret bound on $\mathcal{W}_\eta$; sketched) We provide a proof by approximation and show that the regret of $\mathcal{W}_\eta$ is bounded by the regret of $\mathcal{E}_\eta$ plus some small term. To do so, we compare the definitions (7) and (16), e.g., in the case when $p_{j,0} = 1/24$ for all experts j. Since $R_{48\lfloor (t-1)/48 \rfloor}(\mathcal{E}_\eta, j)$ and $R_{t-1}(\mathcal{E}_\eta, j)$ differ by at most 47 instantaneous regrets, each of which is bounded between $-B^2$ and B^2, the ratio between the numerators of (7) and (16), as well as the one between their denominators, lie in the interval $\left[e^{-47\eta B^2},\ e^{47\eta B^2}\right]$. Therefore, the ratios of the weights defined in (7) and (16) are in the interval $\left[e^{-94\eta B^2},\ e^{94\eta B^2}\right]$. Thus, using a subgradient bound, the difference between the regrets of interest can be bounded as

$$R_T(\mathcal{W}_\eta, j) - R_T(\mathcal{E}_\eta, j) \leqslant 2B^2 \max\left\{e^{\eta 94 B^2} - 1,\ 1 - e^{-\eta 94 B^2}\right\} T\,,$$

which, for η small enough, is of the order of $B^4 \eta T$. Taking η of the the order of $1/\sqrt{T}$, which is also the optimal order of magnitude for the bound on $R_T(\mathcal{E}_\eta, j)$ stated in Proposition 1, entails that $R_T(\mathcal{W}_\eta, j) = O\big(\sqrt{T}\big) = o(T)$, as asserted above.

Adaptations 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 $\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$ and $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$.

As is illustrated in its statement in Figure 10, $\mathcal{G}_{\eta,\alpha}$ basically needs to run an instance of $\mathcal{F}_{\eta,\alpha}$ and to access to its proposed weight vector every 48 rounds. We stated the extension in this way to highlight that it gets synchronized with the base

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: $(w_{1,0}, \ldots, w_{N,0}) = \big(p_{1,0}\, \mathbb{I}_{\{1 \in E_1\}}, \ldots, p_{N,0}\, \mathbb{I}_{\{N \in E_1\}}\big)$
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] $(w_{1,t}, \ldots, w_{N,t}) = \boldsymbol{p}_{t+1}(\mathcal{F}_{\eta,\alpha})$;

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

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

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

[a] $\boldsymbol{p}_{t+1}(\mathcal{F}_{\eta,\alpha})$ 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, 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.

Fig. 10 The extension $\mathcal{G}_{\eta,\alpha}$ of the (basic) fixed-share aggregation rule $\mathcal{F}_{\eta,\alpha}$ to operational forecasting.

rule $\mathcal{F}_{\eta,\alpha}$ every 48 time instances, 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 3, the base rule was performing, at each time instance, a loss update in step (2) and a share update in step (3); the latter update was used to deal with the setting of specialized experts, i.e., with the fact that some experts become inactive at the next time instance and some others become active again, while the former update 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 performed every 48 time instances but the share updates still need to be performed at each time instance, since the values of the sets of active experts E_t may (and do) vary within a day (i.e., within a round of 48 time instances). When losses 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 10.

Finally, we note that a proof by approximation totally similar to the one provided above shows that the regrets of $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}$ do not differ by much, at least when η is small.

5.3.2 Performance of exponentially weighted average rules: With constant parameters

The performance of the extensions $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\mathrm{grad}}$ 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

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.

Values	of η		10^{-6}	10^{-5}	10^{-4}	2×10^{-4}	10^{-3}	5×10^{-3}	10^{-2}
RMSE	$\mathcal{W}_\eta$	(unf.)	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
	$\mathcal{W}_\eta^{\text{grad}}$	(unf.)	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

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 benchmark 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 of weights 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 initial allocation of 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 set in Section 4.3, 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 the present 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 reported in Table 9.

5.3.3 Performance of fixed-share type rules: With constant parameters

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

$$\widetilde{\Lambda}_{\text{FS-France}} = \Big\{ \big(m\times 10^k,\, \alpha\big), \quad \text{for } m\in\{1,\,5\},\ k\in\{-6,\,\ldots,\,0,\,\ldots,\,4\}, \\ \text{and } \alpha\in\{0,\ 0.001,\ 0.01,\, 0.05,\, 0.1,\, 0.2\}\Big\}.$$

(Section 5.4 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 $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ to operational forecasting described above in Section 5.3.1 is summarized in Table 10. 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.

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

Table 10 Performance obtained by the sequential aggregation rules $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}^{\text{grad}}_{\eta,\alpha}$ run with an initial uniform allocation of the weights for various choices of η and α on the grid $\widetilde{\Lambda}_{\text{FS-France}}$; the smallest value for each rule is underlined.

Values	of η	0.01	0.01	0.01	1	1	1	500	500	500
	of α	0.001	0.01	0.05	0.001	0.01	0.05	0.001	0.01	0.05
RMSE	$\mathcal{G}_{\eta,\alpha}$	0.678	0.683	0.704	0.711	0.659	0.652	0.674	0.633	<u>0.632</u>
	$\mathcal{G}^{\text{grad}}_{\eta,\alpha}$	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 rules $\mathcal{W}_\eta$ and $\mathcal{W}^{\text{grad}}_\eta$ 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).

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

we read in Table 10, where the RMSE can get slightly better than 0.6; the relative improvement in the performance with respect to the results of Table 11 is almost 10 %.

5.4 Construction and performance of fully adaptive aggregation rules

5.4.1 Performance of adaptive aggregation rules using given grids

Exponentially weighted average rules Like for the previous data set, 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 we use it also to set the grid $\widetilde{\Lambda}_{\mathcal{W}}$ of on-line tuning of the η as

$$\widetilde{\Lambda}_{\mathcal{W}} = \left\{ m \times 10^{-k},\ \text{for } k \in \{1, \dots, 6\} \text{ and } m \in \{1,\, 2.5,\, 5\} \right\} \cup \{1\}\,,$$

which contains 19 logarithmically evenly spaced points. The choice of the upper bound 1 considered here will be explained and made automatic in Section 5.4. 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 11. Again, the fully sequential 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 $\mathcal{W}^{\text{grad}}_\eta$ run with a fair initial allocation of the weights, as well as the sequence of η chosen at each step, are depicted in Figure 11. Again, the sequence of the weights exhibits absolutely no convergence; large changes in the allocated weights occur over time.

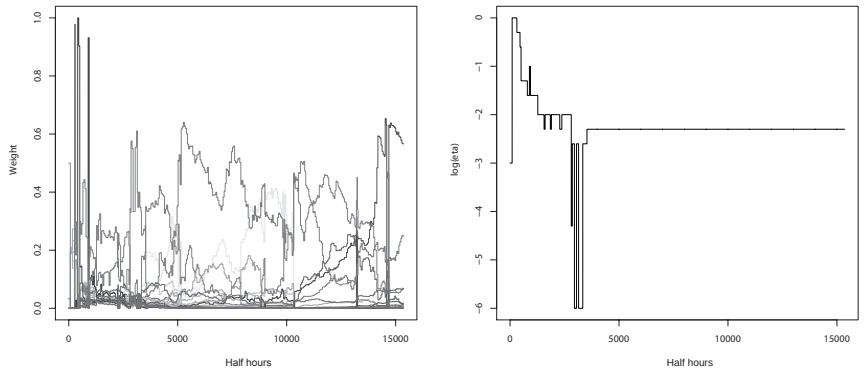

Fig. 11 Graphical representations of the sequences of weights (left) and tuning parameters η (right) chosen by the meta-rule based on the $\mathcal{W}_{\eta}^{\text{grad}}$ run with an initial fair weight allocation and selecting sequentially the values on the grid $\widetilde{\Lambda}_{\mathcal{W}}$.

Table 12 Performance obtained by the rules $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ run with an initial uniform weight 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).

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

Fixed-share-type rules The excellent performance for the best-in-hindsight parameters 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 12.

5.4.2 Adaptive constructions of the grids

In Sections 3.2, 5.3.2, and 5.3.3 we did not clarify how to choose the maximal (and also the minimal) possible value(s) of η in the considered grids; the minimal values were defined in some intrinsic way (at a value close to $\eta^\star$) but it is true that in practice, the order of magnitude of $\eta^\star$ may be unknown. We however 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 most critical issue that remains to be dealt with is the choice of the two endpoint values of η. In Sections 3.2 and 5.3.2 we stopped the grid somewhat arbitrary at the value 1; Section 5.3.3 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 (and also, when needed, of the lower bound).

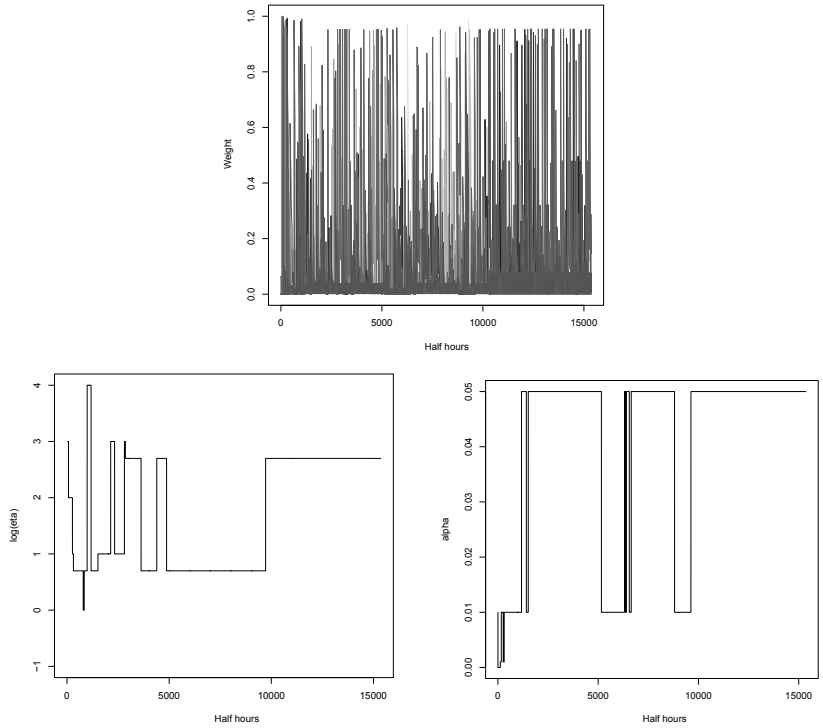

Fig. 12 Graphical representations of the sequences of weights (top) and tuning parameters η (bottom, left) and α (bottom, right) chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_{\text{FS-France}}$; the base rule is $\mathcal{G}_{\eta,\alpha}$ run with an initial uniform weight allocation.

The procedure is based on the observation that in Figures 7, 8, 11, and 12 the sequences of values chosen on-line for the parameters using the method described in Section 3.2 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, it was always the case in our experiments that a highly-performing constant choice was found after a period of time and the measured average performance was close to the one of this final parameter.

It thus seems that it suffices to ensure that the grid is large enough, i.e., that it covers a large enough spectrum. This can be implemented by extending on-line the grid considered originally as follows. We let the user fix a finite starting grid (with logarithmically evenly spaced points) and check at each time instance whether the next increment of the maximal value –or the previous increment of the minimal value– of the parameter of the current grid would have obtained a better performance when being used as a constant choice than any other point in the current grid; if this is the case, this increment is added to the current grid for the remaining time instances. The upper, respectively, lower bounds, of the grids can therefore only increase, respectively, decrease, 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 5.3.3, 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 (some of the values) 100, 500, 1 000, 5 000, 10 000 to the grid of the η; it would then have achieved almost the same performance as the one discussed in Section 5.3.3, where the initial grid already contained all these values. Note that Figure 12 (bottom, left) shows that no larger values than 10 000 were considered in Section 5.3.3 despite the fact that the grid contained one larger such value (50 000).

5.5 Robustness study 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, based on the scattering of the prediction residuals $\widehat{y}_t - y_t$. The RMSE is indeed a global criterion and we want to check that the overall good performance does not come at the cost of local disasters in the accuracy of the aggregated forecasts. To that end we split the data set by the half hours into 48 sub-data sets; for each of these subsets we compute the RMSEs of some of the benchmarks and aggregation rules discussed above and study also the scattering of the (absolute values of the) prediction residuals. This is to see whether the good average behavior exhibited 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 5.3.2 and 5.3.3, 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 13 plots the half-hourly RMSE of these two aggregation rules and of these 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 performance of the rule based on fixed-share aggregation rules is intriguing: its accuracy is significantly improved with respect to the one of the latter benchmark between 12:00 and 21:00 but is also slightly worse than the latter between 6:00 and 12:00. 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). We can provide no reason for this behavior yet; a further study of this behavior and its benefits is left to future research.

As a measure of robustness, we plot in Figures 14 and 15 quantities related to the distribution of the residuals of the forecasts, that is, the difference between the actual consumptions y_t and the forecasts output by the rules and benchmarks described above. Here again, we grouped the residuals by half hours. Figure 14 represents the medians, the third quartiles, and the 90 % quantiles of the absolute values of the residuals. The graphs of Figure 15 are concerned with the behavior

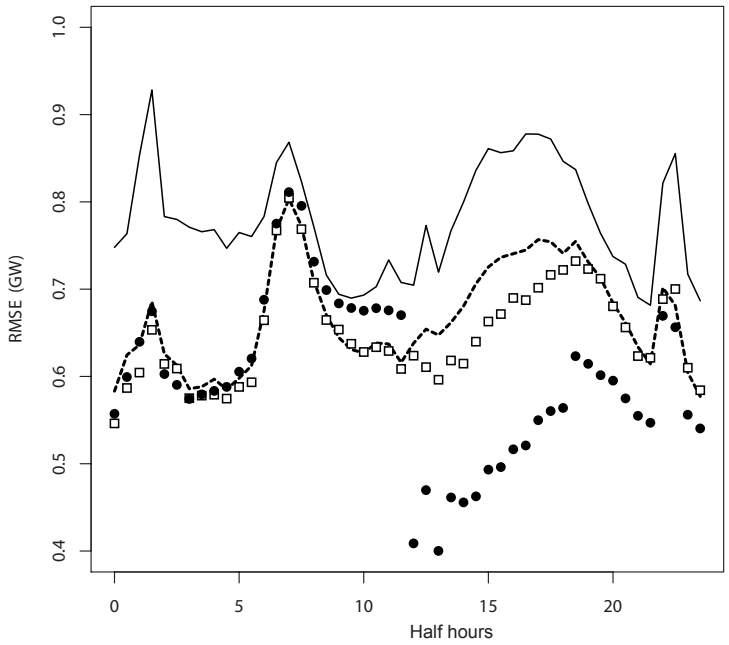

Fig. 13 Half-hourly RMSE of the meta-rules based on the rules $\mathcal{W}_{\eta}^{\mathrm{grad}}$ (symbol: □) and $\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$ (symbol: •) and calibrating them respectively on the grids $\widetilde{\Lambda}_{\mathcal{G}}$ and $\widetilde{\Lambda}_{\text{FS-France}}$; as well as the ones of the best overall single expert (solide line) and of the best overall convex weight vector (dashed line).

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 confirm the story stated above based on the study of the average half-hourly RMSE. In addition, we see that the distributions of the errors of the aggregation rules are more concentrated than the ones of the 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 (potential issues range 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.

6 Conclusions and open questions

6.1 Theoretical achievements and open questions

Achievements We reviewed and extended known aggregation rules for the case of sleeping experts. First, we provided a general analysis of the specialist aggregation

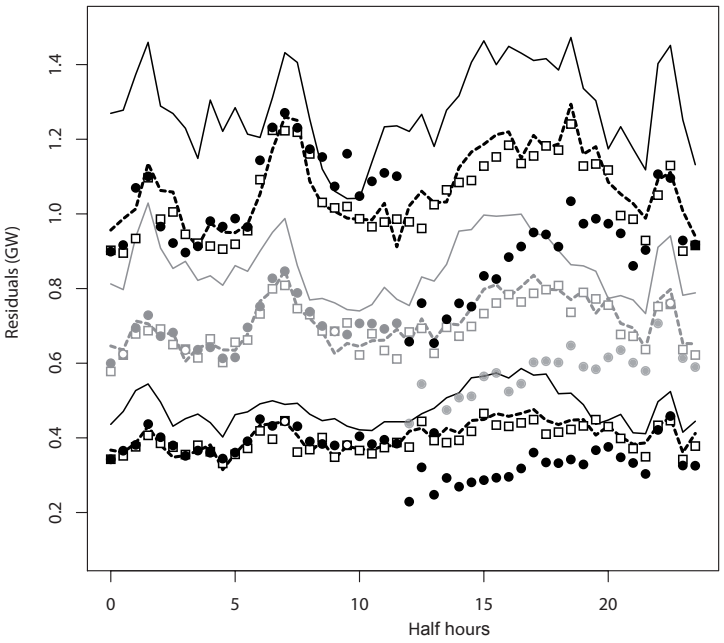

Fig. 14 Using the same rules and benchmarks as in Figure 13, with the same legend: 50 % (black), 75 % (grey), and 90 % (black) quantiles of the absolute values of the residuals, grouped per half hours.

rules of Freund et al. [1997] for all convex loss functions, while the original reference needed an ad hoc analysis for each loss function of interest. Second, we showed how the fixed-share rules of Herbster and Warmuth [1998] can accommodate specialized experts: they form a natural and efficient alternative to the specialist aggregation rules. Finally, for all these rules, as well as the exponentially weighted average ones, we indicated how to extend them so as to take into account some operational constraint of outputting simultaneous forecasts for a fixed number of future time instances.

Open questions All studied rules perform convex aggregation while in practice (see, e.g., Mallet et al. [2009]) linear combinations of the experts forecasts can lead to significant improvements in the accuracy of the aggregated forecasts. For instance, a useful set of aggregation rules in the context of non-specialized experts is given by regularized least-square aggregation rules (like the sequential ridge regression); but –to the best of our knowledge– no extension of them to the case of specialized experts is known. In the technical report Devaine et al. [2009] we reported some preliminary attempts towards such an extension but largely failed achieving reasonable rules with satisfactory performance.

In addition –as is possible with the Bayesian model averaging techniques discussed in the introduction– 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 or by a mere inspection of the dispersion of the base forecasts themselves. Indeed,

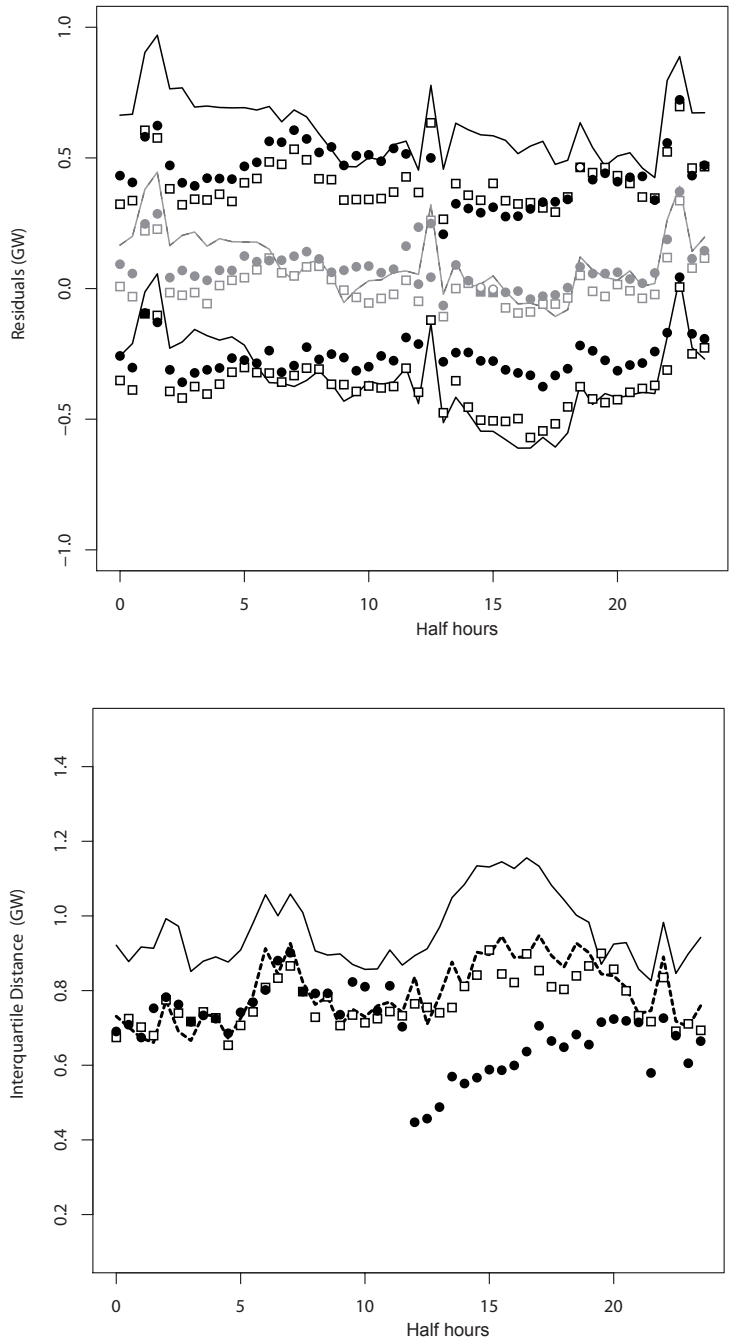

Fig. 15 Using the same rules and benchmarks as in Figure 13, with the same legend: 25 % (black), 50 % (grey), and 75 % (black) quantiles (top graph) and interquartile distances (bottom graph) of the signed values of the residuals, grouped per half hours. For the sake of readability, the values for the best constant convex combination are not displayed on the top graph.

the more concentrated are the latter around a given value, the more confident a strategy should be about its own aggregated forecast. Getting a quantification of these uncertainties in the framework of prediction of individual sequences will be our next move in this line of research.

6.2 Methodological contributions to the applications of prediction with experts advice to real data

We introduced a general methodology (in four steps, see Section 3) to apply the aggregation rules to real data sets. Previous empirical studies (like the ones mentioned in the introduction) often followed its first three steps and did not devote enough energy to the fourth and most important step, in which the true operational performance is studied.

This fourth step consists, indeed, of a general way of efficiently tuning on-line the few (one or two) parameters needed for each aggregation rule, based on the past performance of all members of the family of rules at hand. Good practical performance was demonstrated but a theoretical guarantee of performance (with respect to the best member of the family in hindsight) would be appreciated, as indicated in Section 3.2; this does not seem to be a trivial task and is the subject of an on-going work.

6.3 Empirical conclusions

In this paper we showed the interest of ensemble methods in the prediction of electricity consumption; the sequential aggregation rules we discussed are black-box ways to improve on a set of base forecasting methods. In particular, for the two data sets studied 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). In addition, we noted that resorting to the gradient trick described in Section 2.2.4 always improved the performance of the underlying aggregation rule.

All in all, the line of research developed in this paper 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 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, of the sequential aggregation rules over the experts, 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. (As far as robustness is concerned, we recall that we raised an open question on the empirical performance of the fixed-share type rules in Section 5.5.)

Acknowledgements Marie Devaine carried out this research while being an M.Sc. student at Université Paris-Sud Orsay and completing an internship at EDF R&D, Clamart. Gilles Stoltz was 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.

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.

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

J.M. Bates and C.W.J. Granger. The combination of forecasts. *Operational Research Quarterly*, 20:451–468, 1969.

A. Blum. Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain. *Machine Learning*, 26:5–23, 1997.

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.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order inequalities for prediction under expert advice. *Machine Learning*, 66:321–352, 2007.

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.
J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.
R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
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.
E.E. Leamer. *Specification Searches*. 1978.
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.
A.E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174, 2005.
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.
D.G.C. Smith. Combination of forecasts in electricity demand prediction. *Journal of Forecasting*, 8:349–356, 1989.
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.

7 Omitted proofs

7.1 Proof of Proposition 2

The following proof is a straightforward adaptation of the techniques presented in [Cesa-Bianchi and Lugosi, 2006, Section 5.2]. Its only merit is to show how the share update was obtained in Figure 3.

Proof (of Proposition 2) We first note that by convexity of the ℓ_t,

$$\max_{j_1^T \in \mathcal{L}_m} R_T\bigl(\mathcal{F}_{\eta,\alpha}, j_1^T\bigr) \leqslant \sum_{t=1}^{T} \left(\sum_{i \in E_t} p_{i,t} \ell_t(\delta_i) - \ell_t\bigl(\delta_{j_t}\bigr) \right). \tag{17}$$

We now use the same proof scheme as in [Cesa-Bianchi and Lugosi, 2006, Section 5.2] and show that the rule $\mathcal{F}_{\eta,\alpha}$ is simply an efficient implementation of the rule that would, at each round t, choose a convex weight vector $\boldsymbol{p}'_t$ with components proportional to

$$p'_{j,t} \;\propto\; w'_{j,t-1} = \begin{cases} 0 & \text{if } j \notin E_t, \\ \displaystyle\sum_{j_1^T \in \mathcal{L}} \nu\bigl(j_1^T\bigr)\, e^{-\eta \sum_{s=1}^{t-1} \ell_s(j_s)}\, \mathbb{I}_{\{j_t = j\}} & \text{if } j \in E_t, \end{cases}$$

where ν is some prior probability distribution over $\mathcal{L}$, to be defined below. It then follows from [Cesa-Bianchi and Lugosi, 2006, Lemma 5.1] that for all $j_1^T \in \mathcal{L}$,

$$\sum_{t=1}^{T} \left(\sum_{i \in E_t} p'_{i,t} \ell_t(\delta_i) - \ell_t\big(\delta_{j_t}\big) \right) \leqslant \frac{1}{\eta} \ln \frac{1}{\nu\big(j_1^T\big)} + \frac{\eta L^2 T}{8} \,. \tag{18}$$

To get the stated bound, we thus need, one the one hand, to define the distribution ν, and on the other hand, to show that $\mathcal{F}_{\eta,\alpha}$ indeed performs the efficient implementation indicated above.

[First part: *Definition of* ν] In the sequel we denote by $|E|$ the cardinality of a subset E of $\{1,\dots,N\}$. We fix a real number $\alpha \in [0,1]$ and consider the following probability distribution ν over the sequences of (legal and illegal) experts, i.e., over $\{1,\dots,N\}^T$. For each element $j_1^T \in \mathcal{L}$, we denote by m its size, by $t_1,\dots,t_m$ the instances $1 \leqslant t \leqslant T-1$ such that $j_t \neq j_{t+1}$, and by $\mathcal{T}$ the set of instances $1 \leqslant t \leqslant T-1$ such that $j_t = j_{t+1}$; we then set

$$\nu\big(j_1^T\big) = \frac{1}{|E_1|} \prod_{t \in \mathcal{T}} \left(1 - \alpha + \frac{\alpha}{|E_{t+1}|}\right) \prod_{s=1}^{m} \left(\frac{\alpha}{|E_{t_s+1}|} \mathbb{I}_{\{j_{t_s} \in E_{t_s+1}\}} + \frac{1}{|E_{t_s+1}|} \mathbb{I}_{\{j_{t_s} \notin E_{t_s+1}\}} \right);$$

for $j_1^T \notin \mathcal{L}$, we set $\nu\big(j_1^T\big) = 0$. This application ν indeed defines a probability distribution as can be seen by introducing the uniform distribution μ_1 over E_1 and the following transition functions $\mathrm{Tr}_t : \{1,\dots,N\}^2 \to [0,1]$; for all i,j,

$$\mathrm{Tr}_t(i \to j) = \begin{cases} 0 & \text{if } j \notin E_{t+1}; \quad (19) \\ (1-\alpha) + \alpha/|E_{t+1}| & j \in E_{t+1} \text{ and } i = j; \quad (20) \\ \alpha/|E_{t+1}| & j \in E_{t+1},\ i \in E_{t+1}, \text{ and } i \neq j; \quad (21) \\ 1/|E_{t+1}| & j \in E_{t+1} \text{ and } i \notin E_{t+1}. \quad (22) \end{cases}$$

Its interpretation is as follows. We never switch to an inactive expert, as is ensured by (19). If we can stay on the same expert (if the current expert remains active), then we do so with a probability slightly larger than $1-\alpha$, see (20). If we could have stayed on the same expert, then (19) indicates that we switch with probability $\alpha/|E_{t+1}|$ to a different expert in E_{t+1}. Finally, (22) controls the case when the current expert becomes inactive and we need to switch to a new expert for the compound expert to be legal.

Now, we note that for all i and t, by distinguishing whether $i \in E_{t+1}$ or $i \notin E_{t+1}$,

$$\sum_{j=1}^{N} \mathrm{Tr}_t(i \to j) = 1$$

and that, for all $j_1^T \in \{1,\dots,N\}^T$ (all of them–the legal and the illegal ones),

$$\nu\big(j_1^T\big) = \mu_1(j_1) \prod_{t=1}^{T-1} \mathrm{Tr}_t\big(j_t \to j_{t+1}\big) \,. \tag{23}$$

To prove the stated bound, assuming we have proven as well that $\boldsymbol{p}_t = \boldsymbol{p}'_t$ for all t (which we do below, in the second part of the proof), it suffices to combine

(17) and (18) with the following immediate lower bound on the $\nu(j_1^T)$,

$$\nu(j_1^T) \geqslant \frac{1}{N}\left(\prod_{t\in\mathcal{T}}(1-\alpha)\right)\left(\prod_{s=1}^{m}\frac{\alpha}{N}\right) = \frac{1}{N}\,(1-\alpha)^{T-m-1}\left(\frac{\alpha}{N}\right)^m,$$

which we obtained by upper bounding all cardinalities $|E_t|$ by N in the definition of ν and by using $0 \leqslant \alpha \leqslant 1$. (The obtained bound is actually exactly the one of [Cesa-Bianchi and Lugosi, 2006, Theorem 5.2], due to the loose way we lower bounded ν.)

[Second part: *Proof of the efficient implementation*] The proof goes by induction and mimics exactly the one of [Cesa-Bianchi and Lugosi, 2006, Theorem 5.1]. It suffices to show that for all $j \in \{1,\dots,N\}$ and $t \in \{0,\dots,T-1\}$, one has $w_{j,t} = w'_{j,t}$. To do so, we first note that thanks to (23), the distribution ν can be interpreted as the distribution of an inhomogeneous Markov process, hence (23) indicates the distribution that ν induces over $\{1,\dots,N\}^s$, for all $1 \leqslant s \leqslant T$; the latter is given by simply replacing T by s in (23). We can therefore rewrite $w'_{j,t}$ as

$$w'_{j,t} = \sum_{j_1,\dots,j_{t+1}} \nu(j_1^{t+1})\, e^{-\eta\sum_{s=1}^{t}\ell_s(j_s)}\,\mathbb{I}_{\{j_{t+1}=j\}}\,, \tag{24}$$

where the first sum is (indifferently) taken over $\{1,\dots,N\}^{t+1}$ or $E_1\times\ldots\times E_{t+1}$. For $t=0$, we get

$$w'_{j,0} = \sum_{j_1=1}^{N}\nu(j_1)\,\mathbb{I}_{\{j_1=j\}} = \mu_1(j) = w_{j,0}\,,$$

by definition of ν and of the $w_{j,0}$ (we recall that μ_1 denotes the uniform distribution over E_1). Now, we assume that for some $t \geqslant 1$, we have proved that $w_{i,t-1} = w'_{i,t-1}$ for all $i \in \{1,\dots,N\}$. For $j \in E_{t+1}$, by the share update in Figure 3 and by the induction hypothesis,

$$\begin{aligned} w_{j,t} = {} & \frac{1}{|E_{t+1}|}\sum_{i\in E_t\setminus E_{t+1}} w'_{i,t-1}\,e^{-\eta\ell_t(\delta_i)} + \frac{\alpha}{|E_{t+1}|}\sum_{i\in E_t\cap E_{t+1}} w'_{i,t-1}\,e^{-\eta\ell_t(\delta_i)} \\ & + (1-\alpha)\,\mathbb{I}_{\{j\in E_t\cap E_{t+1}\}}\,w'_{j,t-1}\,e^{-\eta\ell_t(\delta_j)}\,. \end{aligned}$$

By definition of the transition functions (19)–(22), this equality can be rewritten as

$$w_{j,t} = \sum_{i\in E_t} w'_{i,t-1}\,e^{-\eta\ell_t(\delta_i)}\,\mathrm{Tr}_t(i\to j)\,.$$

Substituting (24) in this equality, we get

$$\begin{aligned} w_{j,t} &= \sum_{j_1,\dots,j_t}\sum_{i\in E_t}\nu(j_1^t)\mathbb{I}_{\{j_t=i\}}\mathrm{Tr}_t(i\to j)\,e^{-\eta\sum_{s=1}^{t-1}\ell_s(j_s)}e^{-\eta\ell_t(\delta_i)} \\ &= \sum_{j_1,\dots,j_t}\nu(j_1^t)\mathrm{Tr}_t(j_t\to j)\,e^{-\eta\sum_{s=1}^{t}\ell_s(j_s)} \\ &= \sum_{j_1,\dots,j_t,j_{t+1}}\nu(j_1^{t+1})\,\mathbb{I}_{\{j_{t+1}=j\}}\,e^{-\eta\sum_{s=1}^{t}\ell_s(j_s)} \;=\; w'_{j,t}\,, \end{aligned}$$

where the last but one equality follows from (23). For $j \notin E_{t+1}$, by definitions, $w_{j,t} = 0$ and $w'_{j,t} = 0$. This concludes this proof.

7.2 Proof of Corollary 2

This proof uses the same methodology as the one of Corollary 1.

Proof (of Corollary 2) We fix a compound weight vector $\boldsymbol{q}_1^T \in \mathcal{C}_m$ and denote by $\mathcal{L}\big(\boldsymbol{q}_1^T\big) \subseteq \mathcal{L}_m$ the set of compound experts j_1^T that are compatible with $\boldsymbol{q}_1^T$ in the following sense: denoting by $t_1, \dots, t_m$ the time instances $1 \leqslant s \leqslant T-1$ such that $\boldsymbol{q}_s \neq \boldsymbol{q}_{s+1}$, the elements j_1^T in $\mathcal{L}\big(\boldsymbol{q}_1^T\big)$ are characterized by the fact that $j_s \neq j_{s+1}$ only if $s = t_k$ for some $k \in \{1, \dots, m\}$. We insist on the fact that this is a "only if" statement and not an "if and only if" statement; this means that the switches in the sequences $j_1^T \in \mathcal{L}\big(\boldsymbol{q}_1^T\big)$ can only occur (but are not bound to occur) at the indexes of the switches in $\boldsymbol{q}_1^T$.

Now, we recall that by Section 2.2.4,

$$R_T\big(\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}, \boldsymbol{q}_1^T\big) \leqslant \widetilde{R}_T\big(\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}, \boldsymbol{q}_1^T\big) = \sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\boldsymbol{q}_t) \Big) .$$

Since the $\widetilde{\ell}_t$ are linear over $\mathcal{X}$, the last expression can be upper bounded by

$$\sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\boldsymbol{q}_t) \Big) \leqslant \max_{j_1^T \in \mathcal{L}\left(\boldsymbol{q}_1^T\right)} \sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t\big(\delta_{j_t}\big) \Big) ,$$

which shows that in particular,

$$\sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\boldsymbol{q}_t) \Big) \leqslant \max_{j_1^T \in \mathcal{L}_m} \sum_{t=1}^{T} \Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t\big(\delta_{j_t}\big) \Big) = \max_{j_1^T \in \mathcal{L}_m} \widetilde{R}_T\big(\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}, j_1^T\big) .$$

The proof is concluded by noting that Proposition 2 exactly ensures that the rule $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ is such that

$$\max_{j_1^T \in \mathcal{L}_m} \widetilde{R}_T\big(\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}, j_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m \, (1-\alpha)^{T-m-1}} + \frac{\eta}{8} (2G)^2 T .$$