Forecasting electricity consumption by aggregating specialized experts

Marie Devaine, Pierre Gaillard, Yannig Goude, Gilles Stoltz

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

Forecasting 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 · Pierre Gaillard · Yannig Goude · Gilles Stoltz

Received: 27 March 2011 / Revised: 10 February 2012 / Accepted: 4 June 2012

Abstract We consider the setting of sequential prediction of arbitrary sequences based on specialized experts. We first provide a review of the relevant literature and present two theoretical contributions: a general analysis of the specialist aggregation rule of Freund et al. [1997] and an adaptation of fixed-share rules of Herbster and Warmuth [1998] in this setting. We then apply these rules 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 follow a general methodology to perform the stated empirical studies and detail in particular tuning issues of the learning parameters. 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 expert advice · Specialized experts · Application to real data

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

P. Gaillard
Ecole Normale Supérieure, CNRS, INRIA, Paris, France
Tel.: +33-144-322-041
E-mail: pierre.gaillard@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

1 Introduction and motivation

We consider the sequential prediction of arbitrary sequences based on expert 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, for instance, the best constant convex combination of the experts. 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.

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 expert advice called prediction with specialized (or sleeping) experts. At each round only some of the experts output a prediction while the other ones are inactive. This more difficult scenario 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. 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 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].

The theory of prediction with expert advice has of course been already applied to real data in many fields; we provide a list and a classification of such empirical studies in Section 2.4. We only mention here that 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].

Contributions and outline of the paper

We review in Section 2 the framework of sequential prediction with specialized experts. Three families of aggregation rules are discussed, which were for two of them obtained by taking a new look at existing strategies; this new look corresponds to (slight or more important) adaptations of these existing strategies and to simpler or more general analyses of their theoretical performance bounds. Finally, a practical online tuning of these aggregation rules is developed and put in perspective with respect to theoretical methods to do so.

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 ("Electricité de France", a French electricity provider) 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 methodology described in Section 3: construction of the experts based on historical data; tabulation of the performance of some benchmark prediction methods; results obtained by the sequential aggregation rules, first with parameters optimally tuned in hindsight, and then when the tuning is performed sequentially according to the introduced online tuning. The section on French data is also followed by a note (Section 5.6) 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 the base experts.

2 Aggregation of specialized experts: A survey with some new results

The following framework was introduced in Blum [1997] and further studied in Freund et al. [1997].

A bounded sequence of observations (e.g., hourly or half-hourly electricity consumptions) $y_1, y_2, \ldots, y_T \in [0, B]$ is to be predicted element by element at time instances $t = 1, 2, \ldots, T$. A finite number N of base forecasting methods, henceforth referred to as experts, are available. 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 and 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.

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 $t + 1$ starts.

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 expert forecasts $f_{j,t}$ and in the outcome y_t. Our goal is to design sequential convex aggregation rules $\mathcal{A}$ with a small cumulative error $\sum_{t=1}^{T} \ell_t(\boldsymbol{p}_t)$. To do so,

we will ensure that quantities called regrets (with respect to fixed experts, to fixed convex combinations of experts, or to sequences of experts with few shifts) are small.

Possible loss functions are the square loss, defined by $\ell(x,y) = (x-y)^2$ for all $x, y \in [0, B]$, the absolute loss $\ell(x,y) = |x-y|$, and the absolute percentage of error $\ell(x,y) = |x-y|/y$, which are all three convex and bounded (so that their associated loss functions ℓ_t are convex and bounded as well).

2.1 Minimizing regret with respect to fixed experts

This notion of regret was introduced in Freund et al. [1997] and compares the error suffered by a rule $\mathcal{A}$ to the one of a given expert j only on time instances when j was active; formally, the regret of $\mathcal{A}$ with respect to expert j up to instance T equals

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

where $\delta_j \in \mathcal{X}$ is the Dirac mass on j (the convex weight vector with weight 1 on j).

The exponentially weighted average aggregation rule

It relies on a parameter $\eta > 0$ and will thus be denoted by $\mathcal{E}_\eta$. It chooses $\boldsymbol{p}_1$ to be the uniform distribution over E_1 and uses at time instance $t \geqslant 2$ the convex weight vector $\boldsymbol{p}_t$ given by

$$p_{j,t} = \frac{e^{\eta R_{t-1}(\mathcal{E}_\eta, j)}\, \mathbb{I}_{\{j \in E_t\}}}{\sum_{k \in E_t} e^{\eta R_{t-1}(\mathcal{E}_\eta, k)}}\,; \tag{2}$$

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 $R_{t-1}(\mathcal{E}_\eta, j)$.

The following performance bound is a straightforward consequence of the results presented in Cesa-Bianchi and Lugosi [2003] (its Corollary 2 and the methodology followed in its Sections 3 and 6.2).

Theorem 1 *We assume that the loss functions ℓ_t are convex and uniformly bounded; we denote by L a uniform bound on the quantities $|\ell_t(\delta_i) - \ell_t(\delta_j)|$ when i and j vary in E_t and t varies from 1 to T. The regret of $\mathcal{E}_\eta$ is bounded over all such sequences of expert forecasts and observations as*

$$\max_{j=1,\dots,N} R_T(\mathcal{E}_\eta, j) \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} L^2 T\,. \tag{3}$$

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 as far as theoretical bounds are concerned, see Auer et al. [2002], Cesa-Bianchi et al. [2007].

Remark 1 A slightly different family of aggregation rules based on exponentially weighted averages, referred to as $\mathcal{H}$ in the sequel (which stands for Hedge), was presented in [Blum and Mansour, 2007, Section 6]. It replaces the update (2) by

$$w_{j,t} = \exp\left(-\eta_j \sum_{s=1}^{t-1} \big(\ell_s(\delta_j) - e^{-\eta_j}\ell_s(\boldsymbol{p}_s)\big)\right)$$

$$\text{and} \qquad p_{j,t} = \frac{w_{j,t}\big(1-e^{-\eta_j}\big)\,\mathbb{I}_{\{j\in E_t\}}}{\sum_{k\in E_t} w_{k,t}\big(1-e^{-\eta_k}\big)\,\mathbb{I}_{\{k\in E_t\}}}\,,$$

where the learning rates η_j now depend on the experts $j = 1, \ldots, N$. 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. However, we checked in [Devaine et al., 2009, Section 2.1] that the empirical performance of the families of rules $\mathcal{H}$ and $\mathcal{E}$ were equal. This is why only the simplest of the two, $\mathcal{E}$, will be considered in the sequel.

The specialist aggregation rule

The content of this section revisits and (together with the gradient trick recalled in the next section) improves on the results of [Freund et al., 1997, Sections 3.2–3.4]. In the latter reference, aggregation rules designed to minimize the regret were introduced but their statement, analyses, and regret bounds heavily depended on the specific[1] loss functions at hand. Two special cases were worked out (absolute loss and square loss). In contrast, we provide a compact and general analysis, solely based on Hoeffding's lemma.

The specialist aggregation rule is described in Figure 1; it relies on a parameter $\eta > 0$ and will be denoted by $\mathcal{S}_\eta$. It is close in spirit to but different from the rule $\mathcal{E}_\eta$: 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 little in common (even though the achieved performance was often similar).

Theorem 2 *We assume that the loss functions ℓ_t are convex and uniformly bounded; we denote by L a constant such that the quantities $\ell_t(\delta_i)$ all belong to $[0, L]$ when i varies in E_t and t varies from 1 to T. The regret of $\mathcal{S}_\eta$ is bounded over all such sequences of expert forecasts and observations as*

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

The proof of this theorem is postponed to the appendix (Section A). The (theoretically) optimal choice $\eta^\star = \sqrt{(8\ln N)/(L^2T)}$ leads to the uniform bound $L\sqrt{(T/2)\ln N}$ on the regret of $\mathcal{S}_{\eta^\star}$. The same comments on the calibration of η as in the previous sections 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 [...]."

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 $\widehat{y}_t = \dfrac{1}{\sum_{k \in E_t} w_{k,t}} \sum_{j \in E_t} w_{j,t}\, f_{j,t}$;

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

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

Fig. 1 The specialist aggregation rule $\mathcal{S}_\eta$.

2.2 Minimizing regret with respect to fixed convex combinations of experts

This notion of regret was introduced in Freund et al. [1997] as well and compares the error suffered by a rule $\mathcal{A}$ to the one of a given convex combination $\boldsymbol{q} \in \mathcal{X}$ as follows. Formally, for a set $E \subset \{1, \ldots, N\}$ of active experts, we define

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

and denote by $\boldsymbol{q}^E = (q_1^E, \ldots, q_N^E)$ the convex weight vector obtained by "conditioning" $\boldsymbol{q}$ to E:

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

Now, the definition (1) can be generalized as

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

This is indeed a generalization as we have $R_T(\mathcal{A}, \delta_j) = R_T(\mathcal{A}, j)$.

We deal with this more ambitious goal by resorting to the so-called gradient trick, see [Cesa-Bianchi and Lugosi, 2006, Section 2.5] for more details. When the loss function $\ell : [0, B]^2 \to \mathbb{R}$ is convex and (sub)differentiable in its first argument, then the functions ℓ_t are convex and (sub)differentiable over $\mathcal{X}$; we denote by $\nabla \ell_t$ their (sub)gradient function. By denoting by $\cdot$ the inner product in $\mathbb{R}^N$ and viewing $\mathcal{X}$ as a subset of $\mathbb{R}^N$, we have 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 $\mathcal{X}$. Now, the gradient trick simply consists of replacing the loss functions ℓ_t by the pseudo-loss functions $\widetilde{\ell}_t$ in the definitions of the forecasters. In particular, this replacement in (2), where the loss functions are

hidden in the regret terms, respectively, in Figure 1, leads to an aggregation rule denoted by $\mathcal{E}_\eta^{\mathrm{grad}}$, respectively, $\mathcal{S}_\eta^{\mathrm{grad}}$.

Now, the above convexity inequality and the linearity of the $\widetilde{\ell}_t$ imply that for any rule $\mathcal{A}$,

$$\max_{\boldsymbol{q}\in\mathcal{X}} R_T\big(\mathcal{A},\boldsymbol{q}\big) \leqslant \max_{\boldsymbol{q}\in\mathcal{X}} \sum_{t=1}^{T}\Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t\big(\boldsymbol{q}^{E_t}\big)\Big)\,\boldsymbol{q}(E_t)$$
$$= \max_{\boldsymbol{q}\in\mathcal{X}} \sum_{j=1}^{N} q_j \sum_{t=1}^{T}\Big(\widetilde{\ell}_t(\boldsymbol{p}_t) - \widetilde{\ell}_t(\delta_j)\Big)\,\mathbb{I}_{\{j\in E_t\}}\,;$$

the following result is thus a corollary of Theorems 1 and 2.

Corollary 1 *We assume that the loss functions ℓ_t are convex and (sub)differentiable over $\mathcal{X}$, with (sub)gradient functions 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 expert forecasts and observations as*

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

while the one of $\mathcal{S}_\eta^{\mathrm{grad}}$ is also uniformly bounded as

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

2.3 Minimizing regret with respect 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 (abiding by the activeness constraints given by the E_t). 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 $j_1^T = (j_1,\ldots,j_T)$, 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 $\boldsymbol{q}_1^T = (\boldsymbol{q}_1,\ldots,\boldsymbol{q}_T)$, 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 $\{1,\ldots,T\}$ into integer subintervals corresponding to the use of the same expert or

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

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

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

(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 $i = 1, \ldots, N$,

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

(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. 2 The fixed-share aggregation rule $\mathcal{F}_{\eta,\alpha}$.

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 regrets of a rule $\mathcal{A}$ with respect to $j_1^T \in \mathcal{L}$ and $\boldsymbol{q}_1^T \in \mathcal{C}$ are respectively given by

$$R_T\big(\mathcal{A}, j_1^T\big) = \sum_{t=1}^T \Big(\ell_t(\boldsymbol{p}_t) - \ell_t\big(\delta_{j_t}\big)\Big) \qquad \text{and} \qquad R_T\big(\mathcal{A}, \boldsymbol{q}_1^T\big) = \sum_{t=1}^T \big(\ell_t(\boldsymbol{p}_t) - \ell_t(\boldsymbol{q}_t)\big)\,.$$

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 $\mathcal{L}_m$.

The aggregation rule presented in Figure 2 (when used directly on the losses ℓ_t) 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 B), for the sake of completeness and to show how the share update of Figure 2 was obtained.

Theorem 3 *We assume that the loss functions ℓ_t are convex and uniformly bounded; we denote by L a constant such that the quantities $\ell_t(\delta_i)$ all belong to $[0, L]$ when i varies in E_t 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 expert forecasts and observations as*

$$\max_{j_1^T \in \mathcal{L}_m} R_T\big(\mathcal{F}_{\eta,\alpha}, j_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m (1-\alpha)^{T-m-1}} + \frac{\eta}{8} L^2 T \,. \qquad (5)$$

The (theoretically almost) optimal bound in the theorem 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 carefully choosing parameters $\alpha^\star$ and $\eta^\star$ depending on m, L, and T:

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

which 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; see Section 2.4 for a discussion.

By resorting to the gradient trick defined in Section 2.2, i.e., by replacing the losses ℓ_t in the loss update of Figure 2 by the pseudo-losses $\widetilde{\ell}_t$, one obtains a variant of the previous forecaster, denoted by $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$. The following performance bound is a corollary of Theorem 3; a formal proof is provided in appendix (Section C).

Corollary 2 *We assume that the loss functions ℓ_t are convex and (sub)differentiable over $\mathcal{X}$, with (sub)gradient functions uniformly bounded in the supremum norm as t varies by G. For all $m \in \{0, \dots, T-1\}$, the regret of $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ is uniformly bounded over all such sequences of observations and of expert forecasts as*

$$\max_{\boldsymbol{q}_1^T \in \mathcal{C}_m} R_T\big(\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}, \boldsymbol{q}_1^T\big) \leqslant \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{\alpha^m (1-\alpha)^{T-m-1}} + \frac{\eta}{2} G^2 T \,. \qquad (6)$$

2.4 Sequential automatic tuning of the parameters on data

The aggregation rules discussed above are only semi-automatic strategies, as they rely on fixed-in-advance parameters η (and possibly α) that are not tuned on data. Fully sequential aggregation rules need to set these parameters online. Theoretically almost optimal ways of doing so exist; for instance, Auer et al. [2002], Cesa-Bianchi et al. [2007] indicate ways to online tune the learning rates η for exponentially weighted average rules $\mathcal{E}$ and $\mathcal{E}^{\mathrm{grad}}$ so as to achieve almost the same regret bounds as if the parameters L, G, and T were known in advance. However, the learning rates thus obtained usually perform poorly in practice; see Mallet et al. [2009] for an illustration of this fact on different data sets. The same is observed on our data sets (results not reported); this does not come as a surprise as the theoretically optimal parameters $\eta^\star$ themselves perform poorly, see Remarks 2 and 3 in the empirical studies. Therefore, in spite of the existence of theoretically satisfactory methods, other ones need to be designed based on more empirical considerations.

We do so below but for the sake of completeness we discuss first the symmetric case of the tuning of the parameter α of the fixed-share type rules. These rules need

actually to tune two parameters, η and α; the two tunings are equally important, as is illustrated by the performance reported in Tables 4 and 10. The tuning of η could be done according to the same theoretical methods as mentioned above (e.g., [Auer et al., 2002, Cesa-Bianchi et al., 2007]) but the same issues of practical performance arise. As for α, it is possible in theory not to tune it but to aggregate instances of the rule corresponding to different values of α, where these values lie in a thin enough grid; again, the rule performing this aggregation, e.g., an exponentially weighted average rule, needs to be properly tuned as far as its learning rate η' is concerned. Such a double-layer aggregation was proposed by Monteleoni and Jaakkola [2003], see also de Rooij and van Erven [2009]. We implemented it on our second data set and it turned out to have a performance similar to the empirical method we detail now, as long as the learning rates η and η' were properly set both in the base rules and in the second-layer aggregation, e.g., as follows.

An empirical online tuning of the parameters

We describe the method in a general framework; it is due to Vivien Mallet and was proposed in the technical report by Gerchinovitz et al. [2008] (but never published elsewhere to the best of our knowledge). Let $\mathcal{A}_\lambda$ be a family of sequential aggregation rules relying each 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, the rule index by λ prescribes at time instance t a convex weight vector which we denote by $\boldsymbol{p}_t\big(\mathcal{A}_\lambda\big)$.

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 $t \geqslant 2$,

$$\widehat{\boldsymbol{p}}_t = \boldsymbol{p}_t\Big(\mathcal{A}_{\widehat{\lambda}_{t-1}}\Big) \qquad \text{where} \qquad \widehat{\lambda}_{t-1} \in \operatorname*{argmin}_{\lambda\in\Lambda} \sum_{s=1}^{t-1} \ell_s\Big(\boldsymbol{p}_s\big(\mathcal{A}_\lambda\big)\Big)\,; \tag{7}$$

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. We can however offer no theoretical guarantee for the performance of the meta-rule in terms of the performance of the underlying family.

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 $\Lambda = (0, +\infty)$ and $\Lambda = (0, +\infty) \times [0, 1]$. This is why, in practice, we only consider a finite grid $\widetilde{\Lambda}$ over Λ and perform the minimization of (7) only on the elements of $\widetilde{\Lambda}$ instead of performing it on the whole set Λ. A final choice still seems to be left to the user, namely, how to design this finite grid $\widetilde{\Lambda}$. For the first data set (in Section 4.3) we fix it somewhat arbitrarily. Based on the observed behaviors, we then propose for the second data set (in Section 5.5) a way to construct online the grid $\widetilde{\Lambda}$, finally leading to a fully sequential meta-rule.

Literature review of empirical studies in our framework

Several articles report applications of prediction based on expert advice to real data. They do not investigate the online tuning issues discussed above and can be clustered into three categories as far as the tuning of the parameters is concerned (there is often only a learning rate η to be set).

The first group chooses in the experiments the theoretically optimal parameters (sometimes, for instance, in the case of square losses, these are given by the rates η such that a property of exp-concavity holds). This would be possible as well in our context with improved regret bounds but only for the basic versions of our forecasters, not for their gradient versions (which will be seen to obtain a much improved performance in practice). Furthermore, even such choices of η are slightly suboptimal on our data sets with respect to the fully sequential tuning described above. Actually, tuning η in such a way, one only targets the performance of the best expert, not the one of the best convex combination of the experts (which is significantly better). Examples of such articles and fields of application include the management of the tradeoff between energy consumption and performance in wireless networks ([Monteleoni and Jaakkola, 2003]), the tracking of climate models ([Monteleoni et al., 2011, Jacobs, 2011]), the network traffic demand ([Dashevskiy and Luo, 2011]), the prediction of GDP data ([Jacobs, 2011]), and also the online aggregation of portfolios (e.g., [Cover, 1991, Stoltz and Lugosi, 2005], but the literature is vast). In particular, as far as the latter application is concerned, we note that Borodin et al. [2000] indicates that the studied forecasters do not differ significantly from uniform averages of the experts; this is because the parameter η is not set large enough. This is why we designed a method to tune it automatically based on past data to get the right scale of the problem.

The second group of articles only reports results of optimal-in-hindsight parameters (and sometimes argues that the performance is not very sensitive to the tuning, a fact that we do not observe on the data sets studied in this paper). The studied topics are, for instance, the forecasting of air quality ([Mallet et al., 2009, Mallet, 2010]) and the prediction of outcomes of sports games ([Dani et al., 2006]).

The third group reports the performance of various values of the parameters without choosing between them in advance, for instance, Vovk and Zhdanov [2008] for the latter application or Stoltz and Lugosi [2005] already mentioned above.

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.

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

1. Design experts (based on some historical data).
2. Choose a loss function and evaluate the performance of the experts (on new data).
3. For each family of strategies compute the performance corresponding to the best constant choices of the parameters in hindsight.
4. Assess the quality of the operational performance, i.e., the performance obtained after some automatic and sequential tuning (see Section 2.4).
5. Provide additional results and comments (e.g., a robustness study).

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 (a strategy easily implementable online) or by some oracles, like the best single expert or 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 forecast output by the set of experts; it indicates a bound on the performance that no aggregation strategy can improve on given the data set (given the expert forecasts and the observations). It corresponds to the best element in $\mathcal{L}_{T-1}$.

4 A first data set: Slovakian consumption data

The data was provided by the Slovakian subbranch of the French electricity 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. 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.

As the behavior of electricity consumption depends heavily on the hour of the day and the data set is large enough, we parsed it set into 24 subsets (one per hour interval of the day) and only report 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 are described in Table 1 while all observations (for all hour frames) are plotted in Figure 3.

The considered loss function is the square loss and we will not report cumulative losses but root mean square errors (RMSE), i.e., roots of the per-round cumulative losses. For instance, for a given convex combination $\boldsymbol{q} \in \mathcal{X}$,

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

while for an aggregation rule $\mathcal{A}$,

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

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.

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

The characteristics of the experts are depicted in Figure 4. The bar plot represents the values of the RMSE of the 35 available experts. The scatter plot relates the RMSE of each of the expert to its frequency of activity, that is, it plots the pairs, for all experts j,

$$\left(\text{RMSE}(j),\ \frac{\sum_{t=1}^{T}\mathbb{I}_{\{j\in E_t\}}}{T}\right). \tag{8}$$

We present in Table 2 the values[2] of the RMSE of several procedures, all of them but the first two being oracles. The procedure $\mathcal{U}$ is an aggregation rule that 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 $(1/35,\dots,1/35)$ as the RMSE of the latter gives a weight to each instance t that depends on the cardinality of E_t.

The fact that the RMSE of the best compound expert with size at most 10 is larger than the RMSE of the best single expert is explained by the fact that some overall good experts 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. 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 uniform average rule is evaluated with unequal weights put on the different time instances (more weight put on instances when more experts are active).

A final series of oracles is given 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

$$\min\left\{\sqrt{\frac{1}{T}\sum_{k=1}^{K}\sum_{t:E_t=E^{(k)}}\big(f_{j^k,t}-y_t\big)^2},\ \text{ with } j^k\in E^{(k)}\ \text{ for all } k=1,\dots,K\right\}, \tag{9}$$

[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.

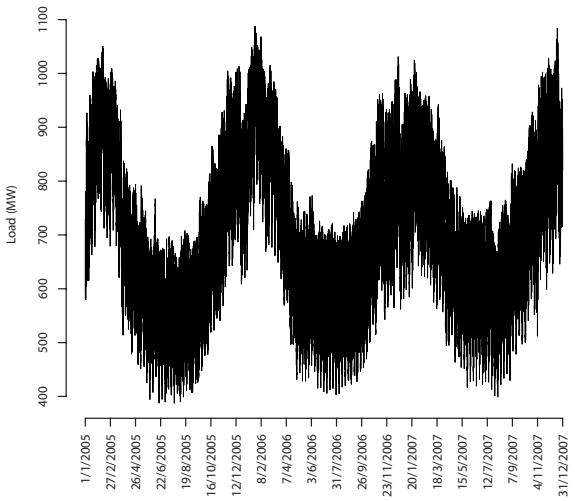

Fig. 3 The observed hourly electricity consumptions encountered by the Slovakian subbranch between January 1, 2005 and December 31, 2007.

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

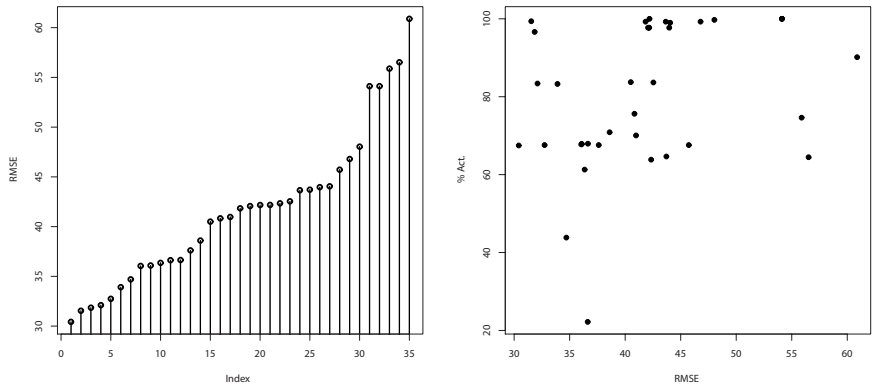

Fig. 4 Graphical representations of the performance of the experts of the Slovakian data set: sorted RMSE (left) and RMSE–frequency of activity pairs (right).

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 $m = 200$	$\min_{j_1^T\in\mathcal{L}_{200}} \textsc{rmse}\big(j_1^T\big)$	$= 15.2$
Prescient strategy (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 (9)	$= 29.1$
Best convex weight vector on each element	See (10)	$= 24.5$

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

$$\min\left\{\sqrt{\frac{1}{T}\sum_{k=1}^{K}\sum_{t:E_t=E^{(k)}}\left(\sum_{j\in E^{(k)}} q_j^{(k)} f_{j,t} - y_t\right)^2}\,,\right.$$
$$\left.\text{with } \boldsymbol{q}^{(k)} \text{ a convex weight vector on } E^{(k)} \text{ for all } k=1,\ldots,K\right\}, \quad (10)$$

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 $m = 50$).

4.2 Results obtained with constant values of the parameters

We now detail the practical performance of the sequential aggregation rules introduced in Section 2, for fixed values of the parameters η and α of the rules. We report for each rule the best performance obtained; the corresponding parameters are said the best constant choices in hindsight. The performance of the families

$\mathcal{E}_\eta$, $\mathcal{E}_\eta^{\text{grad}}$, and $\mathcal{S}_\eta^{\text{grad}}$ is summarized in Table 3. We note that $\mathcal{E}_\eta^{\text{grad}}$ and $\mathcal{S}_\eta^{\text{grad}}$, 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 its respective comparison oracle, the best single expert (RMSE of 30.4 versus 30.5). The performance of the fixed-share type rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}^{\text{grad}}$ is reported in Table 4.

Remark 2 As in Mallet et al. [2009], the best constant choices in hindsight are far away from the theoretically optimal ones, given by $\eta^\star \approx 8 \times 10^{-8}$ for $\mathcal{E}_\eta$, $\eta^\star \approx 4 \times 10^{-8}$ for $\mathcal{S}_\eta^{\text{grad}}$, and $\eta^\star \approx 2 \times 10^{-8}$ for $\mathcal{E}_\eta^{\text{grad}}$.

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

4.3 Results obtained with an online tuning of the parameters

We show in this section how the meta-rules constructed in Section 2.4 can get performance close to the one of the rules based on the best constant parameters in hindsight; we do so, for this data set only, by fixing somewhat arbitrarily the used grids. Based on the observed behaviors we then indicate for the second data set (in Section 5.5) how these grids can be constructed online. For the exponentially weighted average rules $\mathcal{E}_\eta$ and $\mathcal{E}_\eta^{\text{grad}}$, 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 is summarized in Table 5. We note that the good performance obtained for the best choices of the parameters in hindsight is preserved by the adaptive meta-rules resorting to the grids. The sequences of choices of η on the largest grid $\widetilde{\Lambda}_\ell$ are depicted in Figure 6.

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

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

The performance on this grid is summarized in Table 6 while the sequences of choices of η and α on the grid $\widetilde{\Lambda}_{\text{FS}}$ are depicted in Figure 7. The same comments as above on the preservation of the good performance apply.

Table 3 Performance obtained by the sequential aggregation rules $\mathcal{E}_\eta$, $\mathcal{E}_\eta^{\mathrm{grad}}$, and $\mathcal{S}_\eta^{\mathrm{grad}}$ 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	
	$\mathcal{E}_\eta^{\mathrm{grad}}$		31.3	30.9		29.8	<u>28.2</u>	33.5
	$\mathcal{S}_\eta^{\mathrm{grad}}$		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 $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ 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	$\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$	28.0	28.9	29.3	29.2	28.7	28.5	<u>27.2</u>	

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

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

Table 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

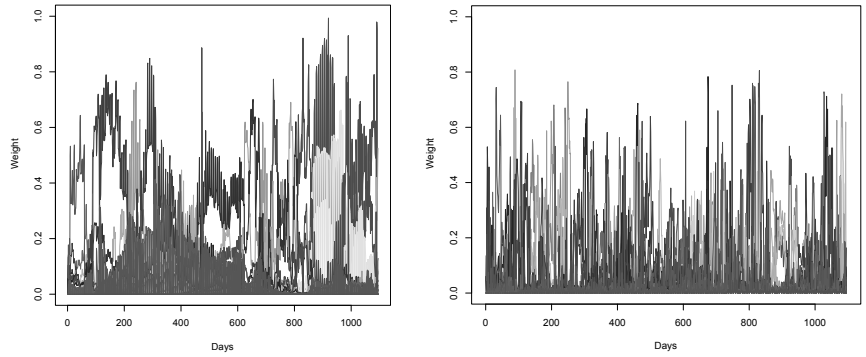

Fig. 5 Graphical representations of the convex weights associated at each time instance with the 35 experts by the rules $\mathcal{E}^{\mathrm{grad}}_{10^{-4}}$ (left) and $\mathcal{F}_{2\times10^{-3},0.2}$ (right).

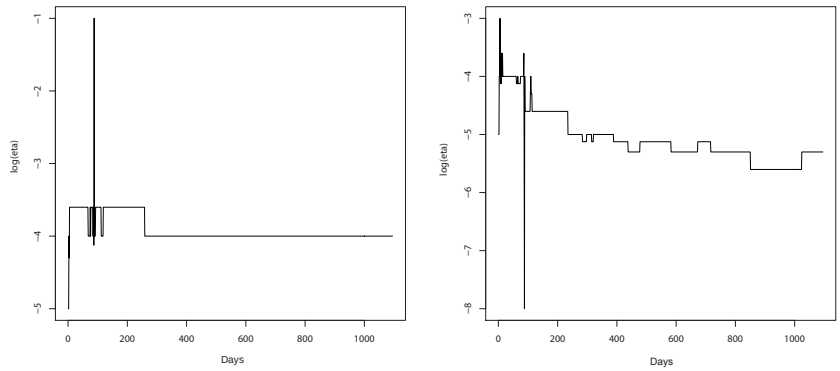

Fig. 6 Graphical representations of the sequences of tuning parameters η chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_\ell$; the base rules are $\mathcal{E}^{\mathrm{grad}}_{\eta}$ (left) and $\mathcal{E}_\eta$ (right).

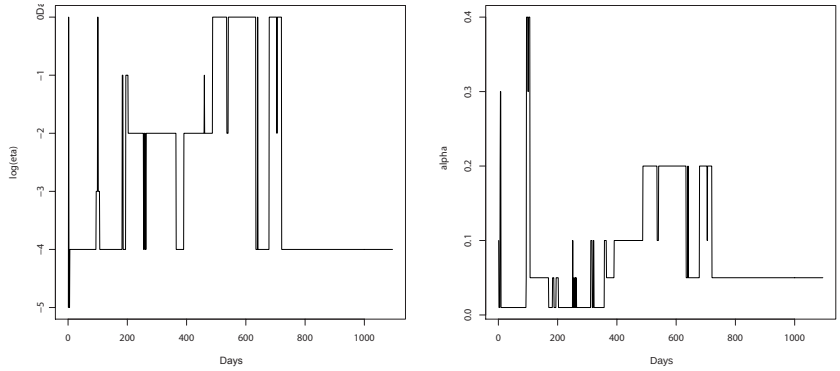

Fig. 7 Graphical representations of the sequences of tuning parameters η (left) and α (right) chosen by the meta-rule selecting sequentially the values on the grid $\widetilde{\Lambda}_{\mathrm{FS}}$; the base rule is $\mathcal{F}^{\mathrm{grad}}_{\eta,\alpha}$.

5 A second data set: Operational forecasting on French data

The data set used in this part is a standard data set used by EDF R&D department. It contains the observed electricity consumptions as well as some side information, which 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. We summarize below some of its characteristics and we refer the interested reader to Dordonnat et al. [2008] for a more detailed description.

It is divided into two sets. The first set ranges from September 1, 2002 to August 31, 2007. We call it the estimation set and use it to design the experts, which then provide forecasts throughout the period corresponding to the second set. This second set covers the period from September 1, 2007 to August 31, 2008. We call it the validation set and use it to evaluate the performance of the considered aggregation rules. 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. The characteristics of the observations y_t of the validation set (formed by half-hourly mean consumptions) 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 we do not split anymore the data set into subsets by the half-hours; this is explained in detail 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 as far as the forecasting in France is concerned.

5.1 Brief description of the construction of the considered experts

The experts we consider here come from three main categories of statistical models: parametric, semi-parametric, and non-parametric models. We do so to get experts that are heterogenous and exhibit varied enough behaviors.

The parametric model used to generate the first group of 15 experts is described in Bruhns et al. [2005] and is implemented in an EDF software called "Eventail." (For conciseness we refer to them as the Eventail experts.) 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 led to the indicated 15 experts.

The second group of 8 experts comes from a generalized additive model presented in Pierrot et al. [2009], Pierrot and Goude [2011] and implemented in the

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

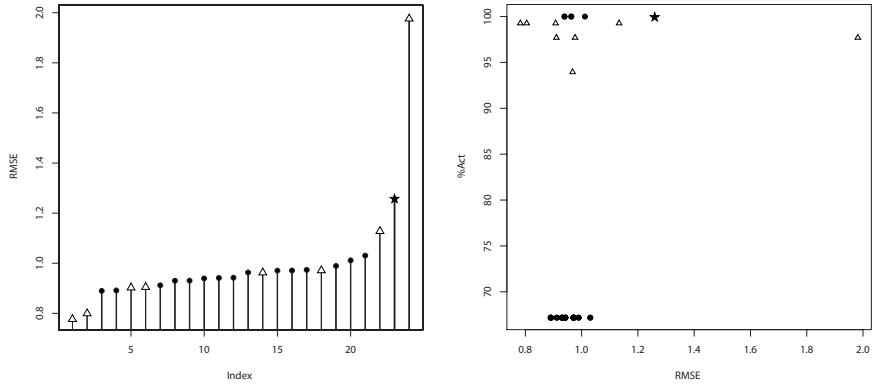

Fig. 8 Graphical representations of the performance of the experts of the French data set: sorted RMSE (left) and RMSE–frequency of activity pairs (right); Eventail experts are depicted by the symbols •, GAM experts are represented by △, while ⋆ stands for the similarity expert.

Table 8 Definition and performance of several (possibly off-line) benchmark procedures on the French data set; 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	$\text{RMSE}\big((1/24, \ldots, 1/24)\big)$	$= 0.748$
Best single expert	$\min_{j=1,\ldots,24} \text{RMSE}(j)$	$= 0.782$
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 $m = T - 1 = 15\,359$	$\min_{j_1^T \in E_1 \times E_2 \times \ldots \times E_T} \text{RMSE}\big(j_1^T\big)$	$= 0.223$

software R by the mgcv package developed by Wood [2006]. (We refer to them as the GAM experts.) The considered generalized additive model imports the idea of the parametric modeling presented above into a semi-parametric modeling. One of its key advantages 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 derived the 8 GAM experts 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.

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 factor like weather conditions); this method is presented in Antoniadis et al. [2006, 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.

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

The characteristics of the experts presented above are depicted in Figure 8, here again with a bar plot representing the (sorted) values of the RMSE of the 24 available experts and a scatter plot relating the RMSE of each of the expert to its frequency of activity. 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 experts 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.

We report in Table 8 the performance obtained by most of the oracles already discussed in Section 4.1. 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 4.1 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, as the value of K is small here ($K = 7$); these two facts explain that the performance of the oracles based on partitions is to expected 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 4.1, 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, which will turn out to be the case).

5.3 Extension of the considered rules to the operational forecasting constraint

We consider prediction with an operational constraint required by EDF consisting of producing half-hourly forecasts every day at 12:00 for the next 24 hours; that is, of forecasting simultaneously the next 48 time instances. (The experts presented above also abide by this constraint.) 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 share update to the values of the active sets $E_{t_k+1}, \ldots, E_{t_k+48}$.

We also propose another extension 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., (3) or (5), 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.

All in all, we denote by $\mathcal{W}_\eta$ and $\mathcal{W}_\eta^{\text{grad}}$ the adaptations to the operational constraint of the rules $\mathcal{E}_\eta$ and $\mathcal{E}_\eta^{\text{grad}}$ of Sections 2.1 and 2.2; by $\mathcal{T}_\eta$ and $\mathcal{T}_\eta^{\text{grad}}$ the ones of the rules $\mathcal{S}_\eta$ and $\mathcal{S}_\eta^{\text{grad}}$ of Sections 2.1 and 2.2; and by $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ the ones of the rules $\mathcal{F}_{\eta,\alpha}$ and $\mathcal{F}_{\eta,\alpha}^{\text{grad}}$ described in Section 2.3. For instance, $\mathcal{W}_\eta$ uses, at time $t = 1, 2, \ldots, T$, the weight vector $\boldsymbol{p}_t$ defined by

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

for all experts j, with the usual convention that empty sums equal 0. (The notation $\lfloor x \rfloor$ denotes the lower integer part of a real number x.)

Similarly, as is illustrated in its statement in Figure 9, $\mathcal{G}_{\eta,\alpha}$ basically needs to run an instance of $\mathcal{F}_{\eta,\alpha}$ and to access to its proposed weight vector every 48 rounds. Between two such synchronizations, only share updates (and no loss update) are performed, to deal with the fact that experts are specialized. Indeed, the values of

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 expert predictions; we use here the same notation as in Section 2.4, 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 $t_k = 48k$ for some k.

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

the sets of active experts E_t may (and do) vary within a one-day-ahead period of time.

Theoretical bounds on the regret can be proved since, as is clear from the algorithmic statements of the extensions, the weights output by the base rules are, for all t, close to the ones of their adaptations (and of course, coincide with them at the time instances t_k). This is because these weights are computed on almost the same sets of losses; these sets differ by at most 47 losses, the ones between the last t_k and the current instance t. A quantification of this fact and a sketch of a regret bound, e.g., for $\mathcal{W}_\eta$, are provided in the appendix (Section D).

5.4 Results obtained with constant values of the parameters

The performance of the extensions $\mathcal{W}_\eta$, $\mathcal{W}_\eta^{\text{grad}}$, $\mathcal{T}_\eta$, and $\mathcal{T}_\eta^{\text{grad}}$ described above is summarized in Table 9. We note that the gradient versions of the forecasters (for both priors) outperform the comparison point formed by the RMSE of the best convex weight vector, equal to 0.696, and which was the only interesting benchmark value among the oracles of the first part of Table 8. They do so by a relative factor of about 5 %; on the other hand, their basic versions (in case of a fair prior) get only a slightly improved performance with respect to this comparison point. It is also worth noting that the performance of the gradient versions is not sensitive to the initial allocation of weights.

Remark 3 Here again, as already mentioned for the Slovakian data set in Section 4.2, 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.

The performance of the extensions $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}^{\text{grad}}_{\eta,\alpha}$ described above is summarized in Table 10. (It turned out that the performance of the algorithms did not depend much on whether the initial weight allocation was fair or uniform and we report only the results obtained by the latter in the sequel.) The comparison points are given by the best compound experts studied in Table 8, which exhibited an excellent performance. This is why we expected and actually see a significant gain of performance for the aggregation rules when resorting to forecasters tracking the performance of the compound experts. Table 10 shows a relative improvement in the performance of about 5 % with respect to the results of Table 9.

5.5 Results obtained with a fully online tuning of the parameters

In Sections 2.4 and 4.3 we indicated that our simulations showed that the step of the grid was not too crucial parameter and that the results were not too sensitive to it; we however did not clarify how to choose the maximal (and also the minimal) possible value(s) of η in the considered grids, i.e., how to determine the right scaling for η. The procedure is based on the observation that in Figures 6 and 7 of Section 4.3 the selected parameters $\widehat{\eta}_t$ are eventually constant or vary in a small range. It thus simply suffices to ensure that the constructed grid covers a large enough span. This can be implemented by extending online the considered grid as follows. We let the user fix an arbitrary finite starting grid, say, reduced to $\{1\}$. At any time t when the selected parameter $\widehat{\eta}_{t-1}$ is an endpoint of the grid, we enlarge it by adding the values $2^r\,\widehat{\eta}_{t-1}$, for $r \in \{1,2,3\}$, respectively, for $r \in \{-1,-2,-3\}$, if the endpoint was the upper limit, respectively, the lower limit of the grid. (We tested different factors than the factor of 2 considered here and also tried to increase the grid with more than three points; no such change had an important impact on the performance.) The possible choices for α are in the (known) bounded range $[0,1]$ and therefore no scaling issue takes place. We considered a fixed grid of possible α given by

$$\alpha \in \big\{0,\ 0.005,\ 0.01,\ 0.05,\ 0.1,\ 0.2,\ 0.5,\ 1\big\}\,.$$

The performance of this adaptive construction of the grids used by the meta-rules with respect to the best constant choices in hindsight is summarized in Tables 11 and 12. We observe that the now fully sequential character of the meta-rule comes at a limited cost in the performance. (That cost would be almost insignificant if a training period was allowed, so as to start the evaluation period with a grid already large enough.)

5.6 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

Table 9 Performance obtained by the sequential aggregation rules $\mathcal{W}_\eta$, $\mathcal{W}_\eta^{\text{grad}}$, $\mathcal{T}_\eta$, and $\mathcal{T}_\eta^{\text{grad}}$ for various choices of η; the smallest RMSE obtained for each rule is underlined.

Values	of η	Prior	10^{-6}	10^{-5}	2×10^{-4}	10^{-3}	2×10^{-2}	10^{-1}	2
RMSE	$\mathcal{W}_\eta$	(unf.)	0.724	0.722	<u>0.718</u>	0.731	0.784	0.783	0.784
	$\mathcal{W}_\eta$	(fair)	0.736	0.731	<u>0.684</u>	0.722	0.785	0.784	0.785
	$\mathcal{W}_\eta^{\text{grad}}$	(unf.)	0.724	0.722	0.705	0.683	0.631	0.640	<u>0.629</u>
	$\mathcal{W}_\eta^{\text{grad}}$	(fair)	0.737	0.733	0.697	0.674	<u>0.633</u>	0.641	0.640
	$\mathcal{T}_\eta$	(unf.)	0.724	0.722	<u>0.718</u>	0.731	0.785	0.783	0.752
	$\mathcal{T}_\eta$	(fair)	0.736	0.731	<u>0.684</u>	0.721	0.786	0.784	0.753
	$\mathcal{T}_\eta^{\text{grad}}$	(unf.)	0.724	0.712	0.705	0.683	<u>0.631</u>	0.640	0.741
	$\mathcal{T}_\eta^{\text{grad}}$	(fair)	0.737	0.733	0.697	0.674	<u>0.633</u>	0.641	0.855

Table 10 Performance obtained by the sequential aggregation rules $\mathcal{G}_{\eta,\alpha}$ and $\mathcal{G}_{\eta,\alpha}^{\text{grad}}$ run with an initial uniform allocation of the weights for various choices of η and α; the smallest RMSE obtained 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>
	$\mathcal{G}_{\eta,\alpha}^{\text{grad}}$	0.646	0.669	0.700	0.622	<u>0.598</u>	0.637	0.683	0.675	0.671

Table 11 Performance obtained by the rules $\mathcal{W}_\eta$, $\mathcal{W}_\eta^{\text{grad}}$, $\mathcal{T}_\eta$, and $\mathcal{T}_\eta^{\text{grad}}$ for the best constant choice of η in hindsight and when used as keystones of a meta-rule selecting sequentially the values of η based on an adaptive grid; results are reported for both the uniform and fair priors.

		Uniform prior		Fair prior	
		Best constant η	Adaptive grid	Best constant η	Adaptive grid
RMSE of	$\mathcal{W}_\eta$	0.718	0.724	0.684	0.696
	$\mathcal{W}_\eta^{\text{grad}}$	0.629	0.640	0.633	0.644
	$\mathcal{T}_\eta$	0.718	0.723	0.684	0.698
	$\mathcal{T}_\eta^{\text{grad}}$	0.631	0.640	0.633	0.645

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 based on an adaptive grid and the values of α according to a fixed grid (right).

		Best constant pair (η, α)	Adaptive grid
RMSE of	$\mathcal{G}_{\eta,\alpha}$	0.632	0.658
	$\mathcal{G}_{\eta,\alpha}^{\text{grad}}$	0.598	0.623

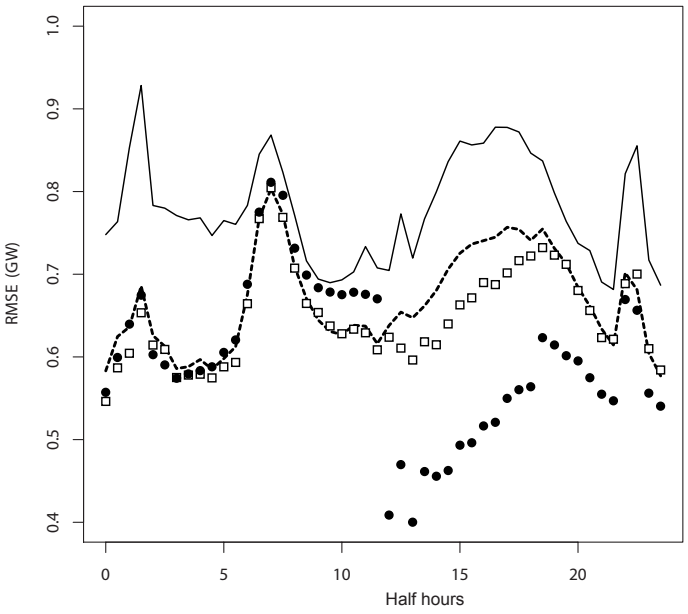

Fig. 10 Half-hourly RMSE of the meta-rules based on the rules $\mathcal{W}_{\eta}^{\mathrm{grad}}$ (symbol: □) and $\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$ (symbol: •); as well as the ones of the best overall single expert (solide line) and of the best overall convex weight vector (dashed line).

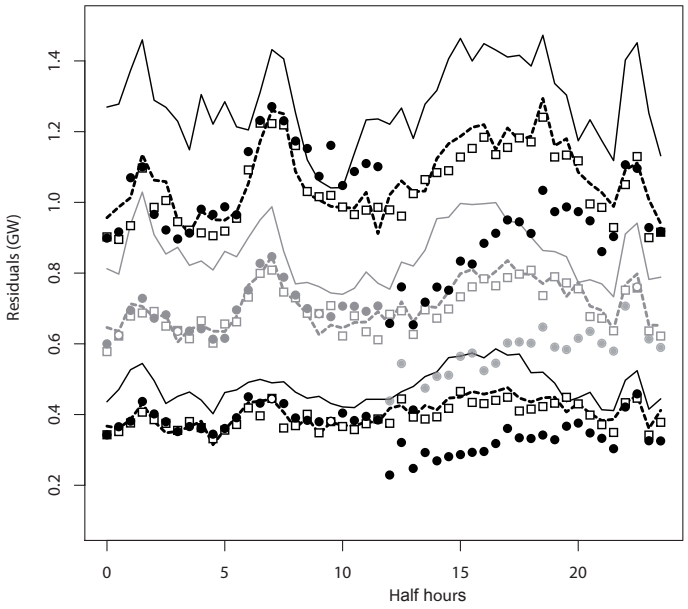

Fig. 11 Using the same rules and benchmarks as in Figure 10, with the same legend: 50 % (black), 75 % (grey), and 90 % (black) quantiles of the absolute values of the residuals, grouped per half hours.

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. To do so we consider two fully sequential aggregation rules, namely, the meta-rules based on families of $\mathcal{W}_{\eta}^{\mathrm{grad}}$ and $\mathcal{G}_{\eta,\alpha}^{\mathrm{grad}}$ run with initial uniform weight allocations. We use as benchmarks the (overall) best single expert and the (overall) best convex weight vector, whose performance was reported in Table 8.

Figure 10 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 expert 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 between 6:00 and 12:00. It thus seems that this rule has excellent performance on very short-term horizon and would probably strongly benefit from an intermediate update around midnight (this is however not the purpose of the present study: intra-day forecasting is left for future research). A similar behavior is observed in Figure 11, which depicts the medians, the third quartiles, and the 90 % quantiles of the absolute values of the residuals grouped by half hours. 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 are seen to reduce the risk of prediction, which is one important pro for operational forecasting.

6 Conclusions

On the theoretical side, we reviewed and extended known aggregation rules for the case of specialized (sleeping) experts. First, we provided a general analysis of the specialist aggregation 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.

We then followed a general methodology to study the performance of these rules on real data of electricity consumption. In particular, we provided fully adaptive

methods that can tune online their parameters based on adaptive grids; doing so, they outperform clearly the rules tuned with the theoretically optimal parameters. All in all, for the two data sets at hand the best rules, given by fixed-share type rules, improve on the accuracy of 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 always improved the performance of the underlying aggregation rule. Finally, the raw improvement in terms of the global performance, as measured by the RMSE, of the sequential aggregation rules over the (convex combinations of) 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.

Acknowledgements We thank the anonymous reviewers and associated editor for their valuable comments and feedback, which improved drastically the exposition of our results and conclusions. Marie Devaine and Pierre Gaillard carried out this research while completing internships at EDF R&D, Clamart; this article is based on the technical reports ([Devaine et al., 2009, Gaillard et al., 2011]) written therefor. 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.

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. Borodin, R. El-Yaniv, and V. Gogan. On the competitive theory and practice of portfolio selection. In *Proceedings of the Fourth Latin American Symposium on Theoretical Informatics (LATIN'00)*, pages 173–196, 2000.

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.

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.

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

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. Dashevskiy and Z. Luo. Time series prediction with performance guarantee. *IET Communications*, 5:1044–1051, 2011.

S. de Rooij and T. van Erven. Learning the switching rate by discretising Bernoulli sources online. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

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.

P. Gaillard, Y. Goude, and G. Stoltz. A further look at the forecasting of the electricity consumption by aggregation of specialized experts. Technical report, École normale supérieure, Paris and EDF R&D, Clamart, July 2011. Updated February 2012; available at `http://ulminfo.fr/%7Epgaillar/doc/GaGoSt-report.pdf`.

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.

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

A.Z. Jacobs. Adapting to non-stationarity with growing predictor ensembles. Master's thesis, Northwestern University, 2011.

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.

V. Mallet. Ensemble forecast of analyses: coupling data assimilation and sequential aggregation. *Journal of Geophysical Research*, 115(D24303), 2010.

V. Mallet, G. Stoltz, and B. Mauricette. Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research*, 114(D05307), 2009.

C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, pages 1093–1100, 2003.

C. Monteleoni, G. Schmidt, S. Saroha, and E. Asplund. Tracking climate models. *Journal of Statistical Analysis and Data Mining*, 4:372–392, 2011. Special issue "Best of CIDU 2010".

A. Pierrot and Y. Goude. Short-term electricity load forecasting with generalized additive models. In *Proceedings of the Sixteenth International Conference on Intelligent System Application to Power Systems (ISAP)*, 2011.

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.

G. Stoltz and G. Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59: 125–159, 2005.

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.

A Proof of Theorem 2

Proof One can show by induction that the vectors $\boldsymbol{w}_t$ are convex weight vectors. We use the notation defined in Section 2.2 for the normalization $\boldsymbol{q}^E$ of convex weight vectors $\boldsymbol{q}$ to a given set of active experts E; then, the convex combination used by $\mathcal{S}_\eta$ at round t can be written as $\boldsymbol{p}_t = \boldsymbol{w}_t^{E_t}$.

By convexity of the loss functions ℓ_t, the regret with respect to some expert j can be bounded as

$$R_T(\mathcal{S}_\eta, j) \leqslant \sum_{t=1}^{T} \left(\sum_{i \in E_t} w_{i,t}^{E_t}\, \ell_t(\delta_i) - \ell_t(\delta_j) \right) \mathbb{I}_{\{j \in E_t\}}\,.$$

Hoeffding's lemma (see, e.g., [Cesa-Bianchi and Lugosi, 2006, Lemma A.1]) entails that for all t such that $j \in E_t$,

$$\begin{aligned} \sum_{i \in E_t} w_{i,t}^{E_t}\, \ell_t(\delta_i) \quad &\leqslant \quad -\frac{1}{\eta} \ln \left(\sum_{i \in E_t} w_{i,t}^{E_t}\, e^{-\eta \ell_t(\delta_i)} \right) + \frac{\eta}{8} L^2 \\ &= \quad -\frac{1}{\eta} \ln \frac{w_{j,t}\, e^{-\eta \ell_t(\delta_j)}}{w_{j,t+1}} + \frac{\eta}{8} L^2 = \ell_t(\delta_j) - \frac{1}{\eta} \ln \frac{w_{j,t}}{w_{j,t+1}} + \frac{\eta}{8} L^2\,, \end{aligned}$$

where we used that the update of the weight of an expert $j \in E_t$ can be rewritten by definition as

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

For $j \notin E_t$, we have that $w_{j,t+1} = w_{j,t}$, again by definition of the rule. Thus a telescoping sum appears and we get

$$\sum_{t=1}^{T} \left(\sum_{i \in E_t} w_{i,t}^{E_t}\, \ell_t(\delta_i) - \ell_t(\delta_j) \right) \mathbb{I}_{\{j \in E_t\}} \leqslant -\frac{1}{\eta} \ln \frac{w_{j,1}}{w_{j,T+1}} + \frac{\eta}{8} L^2 \sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}}\,.$$

The proof is concluded by noting that $w_{j,1}/w_{j,T+1} \geqslant 1/N$ as $w_{j,1} = 1/N$ and $w_{j,T+1} \leqslant 1$.

B Proof of Theorem 3

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 2.

Proof We first note that by convexity of the ℓ_t,

$$\max_{j_1^T \in \mathcal{L}_m} R_T\big(\mathcal{F}_{\eta,\alpha}, j_1^T\big) \leqslant \sum_{t=1}^{T} \left(\sum_{i \in E_t} p_{i,t} \ell_t(\delta_i) - \ell_t\big(\delta_{j_t}\big) \right). \tag{12}$$

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

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

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 $j_1^T \in \mathcal{L}$,

$$\sum_{t=1}^{T} \left(\sum_{i \in E_t} p'_{i,t} \ell_t(\delta_i) - \ell_t\big(\delta_{j_t}\big) \right) \leqslant \frac{1}{\eta} \ln \frac{1}{\nu\big(j_1^T\big)} + \frac{\eta L^2 T}{8}\,. \tag{13}$$

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,

$$\mathrm{Tr}_t(i\to j) = \begin{cases} 0 & \text{if } j \notin E_{t+1}; \quad (14)\\ (1-\alpha)+\alpha/|E_{t+1}| & j\in E_{t+1} \text{ and } i=j; \quad (15)\\ \alpha/|E_{t+1}| & j\in E_{t+1},\ i\in E_{t+1}, \text{ and } i\neq j; \quad (16)\\ 1/|E_{t+1}| & j\in E_{t+1} \text{ and } i\notin E_{t+1}. \quad (17)\end{cases}$$

Its interpretation is as follows. We never switch to an inactive expert, as is ensured by (14). 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 (15). If we could have stayed on the same expert, then (14) indicates that we switch with probability $\alpha/|E_{t+1}|$ to a different expert in E_{t+1}. Finally, (17) 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),

$$\nu\big(j_1^T\big) = \mu_1(j_1) \prod_{t=1}^{T-1} \mathrm{Tr}_t\big(j_t \to j_{t+1}\big)\,. \tag{18}$$

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 (12) and (13) with the following immediate lower bound on the $\nu\big(j_1^T\big)$,

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

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 (18), the distribution ν can be interpreted as the distribution of an inhomogeneous Markov process, hence (18) 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 (18). We can therefore rewrite $w'_{j,t}$ as

$$w'_{j,t} = \sum_{j_1,\dots,j_{t+1}} \nu\big(j_1^{t+1}\big)\, e^{-\eta\sum_{s=1}^{t}\ell_s(j_s)}\,\mathbb{I}_{\{j_{t+1}=j\}}\,, \tag{19}$$

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 2 and by the induction hypothesis,

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

By definition of the transition functions (14)–(17), 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 (19) in this equality, we get

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

where the last but one equality follows from (18). For $j \notin E_{t+1}$, by definitions, $w_{j,t} = 0$ and $w'_{j,t} = 0$. This concludes this proof.

C Proof of Corollary 2

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

Proof 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 the gradient trick recalled in Section 2.2,

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

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 Theorem 3 exactly ensures that the rule $\mathcal{F}_{\eta,\alpha}^{\mathrm{grad}}$ is such that

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

D Sketch of a regret bound on the operational adaptation $\mathcal{W}_\eta$ of $\mathcal{E}_\eta$

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 (2) and (11), 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 (2) and (11), 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 (2) and (11) are in the interval $\left[e^{-94\eta B^2},\ e^{94\eta B^2}\right]$. Thus, using a gradient 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 Theorem 1, entails that $R_T(\mathcal{W}_\eta, j) = O\big(\sqrt{T}\big) = o(T)$, as asserted above.