Journal de la Société Française de Statistique
Vol. 151 No. 2 (2010)

Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique

Title: Sequential aggregation of predictors: General methodology and application to air-quality forecasting and to the prediction of electricity consumption

Gilles Stoltz *

Résumé : Cet article fait suite à la conférence que j'ai eu l'honneur de donner lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des XL^{e} Journées de Statistique à Ottawa, en 2008. Il passe en revue les résultats fondamentaux, ainsi que quelques résultats récents, en prévision séquentielle de suites arbitraires par agrégation d'experts. Il décline ensuite la méthodologie ainsi décrite sur deux jeux de données, l'un pour un problème de prévision de qualité de l'air, l'autre pour une question de prévision de consommation électrique. La plupart des résultats mentionnés dans cet article reposent sur des travaux en collaboration avec Yannig Goude (EDF R&D) et Vivien Mallet (INRIA), ainsi qu'avec les stagiaires de master que nous avons co-encadrés : Marie Devaine, Sébastien Gerchinovitz et Boris Mauricette.

Abstract: This paper is an extended written version of the talk I delivered at the "XL^{e} Journées de Statistique" in Ottawa, 2004, when being awarded the Marie-Jeanne Laurent-Duhamel prize. It is devoted to surveying some fundamental as well as some more recent results in the field of sequential prediction of individual sequences with expert advice. It then performs two empirical studies following the stated general methodology: the first one to air-quality forecasting and the second one to the prediction of electricity consumption. Most results mentioned in the paper are based on joint works with Yannig Goude (EDF R&D) and Vivien Mallet (INRIA), together with some students whom we co-supervised for their M.Sc. theses: Marie Devaine, Sébastien Gerchinovitz and Boris Mauricette.

Classification AMS 2000 : primaire 62-02, 62L99, 62P12, 62P30

Mots-clés : Agrégation séquentielle, prévision avec experts, suites individuelles, prévision de la qualité de l'air, prévision de la consommation électrique

Keywords: Sequential aggregation of predictors, prediction with expert advice, individual sequences, air-quality forecasting, prediction of electricity consumption

Ecole normale supérieure, CNRS, 45 rue d'Ulm, 75005 Paris
& HEC Paris, CNRS, 1 rue de la Libération, 78350 Jouy-en-Josas
E-mail : gilles.stoltz@ens.fr
URL : http://www.math.ens.fr/∼stoltz

* L'auteur remercie l'Agence nationale de la recherche pour son soutien à travers le projet JCJC06-137444 ATLAS ("From applications to theory in learning and adaptive statistics").

† Ces recherches ont été menées dans le cadre du projet CLASSIC de l'INRIA, hébergé par l'Ecole normale supérieure et le CNRS.

 ISSN: 2102-6238

1. Introduction

On se concentre sur un problème de prévision séquentielle, que l'on abordera sous un angle méta-statistique. Plus précisément, il s'agit de prévoir des observations $y_1, \ldots, y_t$, qui viennent l'une après l'autre. On ne suppose pas que ces observations sont la réalisation d'un certain processus stochastique sous-jacent dont il faudrait estimer les caractéristiques afin d'en déduire une bonne manière de prévoir. Autrement dit, il ne s'agit pas d'un problème statistique et l'on considérera l'ensemble des suites d'observations possibles.

Cependant, on suppose disposer d'un nombre fini de prédicteurs fondamentaux, indexés par $j = 1, \ldots, N$ et qui à chaque échéance t, lorsqu'il s'agit de prévoir y_t, proposent chacun une prévision $f_{j,t}$. Ces prédicteurs peuvent, eux, reposer sur une modélisation stochastique et dériver de méthodes statistiques. L'objectif est d'agréger séquentiellement leurs prévisions $f_{j,t}$ afin d'en déduire une prévision finale $\widehat{y}_t$ la plus performante possible. C'est en ce sens que l'on parle de cadre méta-statistique.

Les prédicteurs fondamentaux sont appelés experts, parce qu'en plus de pouvoir reposer sur des techniques statistiques, ils peuvent éventuellement reposer sur des informations contextuelles, utiliser des ressources numériques importantes, et même faire appel à une expertise humaine. En fait, on les traitera essentiellement comme des boîtes noires prédictives dans la première partie de cet article, au moment de décrire comment, de manière théorique, utiliser leurs prévisions. Ensuite, dans la seconde partie applicative, on expliquera bien sûr, pour chaque jeu de données, comment les experts ont été construits.

1.1. Historique de ce cadre de prévision séquentielle de suites arbitraires

La première mention des problèmes de prévision séquentielle de suites arbitraires remonte aux années 50, et plus précisément aux travaux de Hannan [29] et Blackwell [6], deux statisticiens qui ont énoncé des résultats fondateurs en théorie des jeux dans ces articles. Cover [14] figure également parmi les pionniers du domaine. Il faut également citer l'étude de la compression de suites arbitraires de données en théorie de l'information, où les recherches d'avant-garde ont été menées par Ziv [48, 49] et Lempel et Ziv [33, 50] ; ils ont résolu la question de compresser une suite arbitraire de données presqu'aussi bien que le meilleur automate fini. Enfin, en théorie de l'apprentissage, l'introduction du problème de prévision séquentielle de suites arbitraires a été effectuée par Littlestone et Warmuth [34] et Vovk [43] ; Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire et Warmuth [10], Foster [21], Freund et Schapire [23] et Vovk [44] ont présenté quelques-un des résultats fondamentaux.

Pour un exposé de l'état de l'art du domaine des années 50 jusqu'en 2006, on pourra se reporter à l'ouvrage de synthèse [12]. Pour le cas particulier de la prévision séquentielle randomisée (lorsque les prévisions $\widehat{y}_t$ peuvent être aléatoires), un article de survol [35] est paru dans le présent journal. Cependant, on s'intéresse uniquement ici à la prévision séquentielle par agrégation convexe, où les $\widehat{y}_t$ doivent être déterministes.

1.2. Applications précédentes de ces techniques à des jeux de données réelles

Les articles de prévision de suites arbitraires relevant de l'apprentissage éprouvent souvent les nouvelles méthodes sur des données artificielles. A vrai dire, au mieux de notre connaissance,

 ISSN: 2102-6238

seules peu d'études empiriques ont été menées jusqu'à ce jour, alors qu'en principe le cadre méta-statistique d'agrégation des prévisions d'experts a vocation à s'appliquer à tout problème pour lequel on peut construire plusieurs tels experts – autant dire, presque à tous les problèmes de prévision séquentielle.

La première lignée de tels travaux empiriques a porté sur l'investissement séquentiel dans le marché boursier et a été initiée par [15], les observations étant formées par les évolutions journalières de 36 valeurs boursières de la bourse de New-York sur la période couvrant 1963 à 1985 et les experts étant simplement identifiés à ces valeurs. Une seconde lignée plus récente considère la prévision de résultats sportifs, voir [16] ou [46]. Les prévisions des experts y sont données par les cotes indiquées par différents bookmakers ou par celles résultant des paris de nombreux participants sur un site web de paris en ligne. Enfin, les méthodes d'agrégation d'experts en prévision de consommation électrique ont été étudiées en premier lieu par Goude [27, 28]. Nous avons ajouté à tout cela un dernier cadre applicatif, la prévision de la qualité de l'air ; [38] fournit des experts qu'on utilisera dans la suite du présent article et mentionne l'intérêt du recours à des stratégies d'agrégation d'experts.

1.3. Présentation, objectifs et plan de l'exposé

Cet article présente à la fois un survol de la méthodologie de la prévision séquentielle de suites arbitraires et quelques résultats théoriques nouveaux ; les résultats fondamentaux sont démontrés en détails afin que le lecteur ait une idée des techniques mathématiques en jeu, tandis les résultats plus avancés sont simplement énoncés avec mention d'une référence où leur preuve peut être trouvée. Le tout est précédé d'une introduction philosophique au cadre d'agrégation des experts, où l'on essaie d'établir des liens avec des problèmes classiques en statistique, notamment l'équilibre entre erreur d'approximation et difficulté d'estimation. On effectue ensuite le résumé de deux études empiriques sur deux jeux de données distincts, l'un pour un problème de prévision de qualité de l'air, l'autre pour une question de prévision de consommation électrique. Ce n'est qu'après cette étude que l'on pourra discuter sereinement des liens entre les techniques d'agrégation séquentielle de prédicteurs et les autres outils de théorie de l'apprentissage déjà utilisés avec succès dans le passé en statistique (notamment, les arbres de régression).

2. Cadre mathématique de l'agrégation de prédicteurs

Cette partie décrit en détails le cadre de la prévision séquentielle par agrégation de prédicteurs fondamentaux. Elle introduit notamment un critère de performance appelé regret ; elle justifie et interprète ce regret comme une difficulté d'estimation, qu'il faut évidemment mesurer à l'aune d'une erreur d'approximation. Elle présente ensuite des stratégies de prévision assurant que ce regret est faible.

2.1. Formulation générale comme un méta-problème de prévision séquentielle

L'objet du problème est la prévision séquentielle d'observations $y_1, y_2, \ldots$ issues d'un ensemble $\mathscr{Y}$, sur lequel on n'impose aucune condition particulière. Contrairement au cadre habituel de la statistique, nous ne supposerons pas que cette suite d'observations est la réalisation d'un certain

 ISSN: 2102-6238

processus stochastique sous-jacent ; il ne s'agit donc pas ici d'estimer au mieux les caractéristiques d'un tel processus afin d'en prévoir le comportement et d'en tirer des prévisions les plus précises possibles. Au contraire, nous considérerons que toutes les suites possibles de $\mathscr{Y}$ peuvent se produire et exhiberons des garanties de performance uniformes en ces suites. C'est pourquoi l'on parle de suites arbitraires ou encore, de *suites individuelles*.

Le problème est séquentiel : à chaque échéance t, le statisticien doit produire une prévision (déterministe) $\widehat{y}_t$ fondée sur les observations passées (déterministes) $y_1, \ldots, y_{t-1}$. Cette prévision appartient à un ensemble convexe $\mathscr{X}$, qui peut être différent de $\mathscr{Y}$; un cas typique est celui où $\mathscr{X}$ est l'enveloppe convexe de $\mathscr{Y}$. La prévision est ensuite comparée à l'observation y_t grâce à une fonction de perte $\ell : \mathscr{X} \times \mathscr{Y} \longrightarrow \mathbb{R}$; cette dernière est le plus souvent positive. On définit ainsi la perte cumulée du statisticien sur les T premières échéances comme

$$\sum_{t=1}^{T} \ell\big(\widehat{y}_t, y_t\big)$$

et on veut assurer que cette dernière est la plus faible possible.

Il manque un dernier ingrédient pour que ce problème de prévision non stochastique ait un sens : il s'agit de l'aide procurée par des experts. Ces derniers sont des prédicteurs fondamentaux qui proposent à chaque échéance une prévision fondée sur l'observation du passé. Plus précisément, ils sont en nombre fini N et on les indexe par $j = 1, \ldots, N$; l'expert j procure à l'échéance t une prévision notée $f_{j,t} \in \mathscr{X}$ et qui dépend de $y_1, \ldots, y_{t-1}$ et éventuellement d'autres informations auxquelles il aurait accès lui seul. Le statisticien peut alors former sa prévision $\widehat{y}_t$ en se fondant non seulement sur les observations passées $y_1, \ldots, y_{t-1}$ mais aussi sur les prévisions présentes et passées des experts, $f_{j,s}$ pour $1 \leqslant 1 \leqslant t-1$ et $j = 1, \ldots, N$. La considération des prévisions passées des experts est utile pour suivre l'intuition selon laquelle il est sage de faire d'autant plus confiance à la prévision présente d'un expert qu'il s'est montré efficace dans le passé.

Dans cet article, on se restreint aux stratégies de prévision par agrégation convexe. Une telle stratégie $\mathscr{S}$ associe à l'information disponible au début de chaque échéance t (aux observations passées et prévisions passées et présentes des experts) un vecteur de mélange $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$, qui est ensuite utilisé pour former la combinaison convexe dans $\mathscr{X}$

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

ainsi, on impose ici que les vecteurs de mélange soient choisis dans le simplexe $\mathscr{P}$ de $\mathbb{R}^N$, c'est-à-dire qu'ils vérifient les conditions

$$\forall j \in \{1, \ldots, N\}, \;\; p_{j,t} \geqslant 0 \qquad \text{et} \qquad \sum_{k=1}^{N} p_{k,t} = 1 \,.$$

Remarque. *Lorsque l'ensemble des prévisions $\mathscr{X}$ n'est pas convexe, au lieu d'agréger les prévisions selon un vecteur de mélange convexe $\mathbf{p}_t$, on en tire une au hasard selon la probabilité donnée par $\mathbf{p}_t$; on parle de prévision randomisée. On ne s'arrête pas sur cette variante, notamment parce qu'un article précédent de ce journal l'a déjà traitée en détails [35].*

 ISSN: 2102-6238

2.1.1. Critère de qualité d'une stratégie : le regret

L'évaluation d'une stratégie $\mathscr{S}$ ne peut être effectuée de manière absolue : si tous les experts sont mauvais, il est vraisemblable qu'aucune stratégie de prévision par agrégation convexe ne pourra avoir de bons résultats. On retient donc un critère relatif qui quantifie la proximité de la précision de prévision d'une stratégie $\mathscr{S}$ à celle de la meilleure combinaison convexe des experts ; à cet effet, on définit les pertes cumulées de $\mathscr{S}$ et de chaque vecteur de mélange $\mathbf{q} \in \mathscr{P}$ comme, respectivement

$$\widehat{L}_T(\mathscr{S}) = \sum_{t=1}^{T} \ell\big(\widehat{y}_t, y_t\big) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t\right)$$

$$\text{et} \qquad L_T(\mathbf{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j f_{j,t}, y_t\right).$$

Le regret de $\mathscr{S}$ sur les T premières échéances est alors la différence entre ces pertes cumulées,

$$R_T(\mathscr{S}) = \widehat{L}_T(\mathscr{S}) - \inf_{\mathbf{q}\in\mathscr{P}} L_T(\mathbf{q}).$$

Bien sûr, les quantités $\widehat{L}_T(\mathscr{S})$, $L_T(\mathbf{q})$ et $R_T(\mathscr{S})$ dépendent également des observations $y_1, \ldots, y_T$ et des prévisions des experts même si, dans un souci d'allégement des notations, cette dépendance n'est pas explicitement rappelée.

Le regret $R_T(\mathscr{S})$ est de l'ordre au plus de T lorsque la fonction de perte est bornée. On recherche ici des stratégies telles que leur regret, rapporté au nombre d'échéance, tende uniformément vers 0, quelles que soient les observations et les prévisions des experts.

Objectif 2.1. *Construire des stratégies de prévision $\mathscr{S}$ telles que*

$$\limsup_{T\to\infty} \sup \left\{ \frac{R_T(\mathscr{S})}{T} \right\} \leqslant 0,$$

où le supremum porte sur l'ensemble des suites d'observations et de prévisions possibles.

A cause de cette uniformité sur le contrôle du regret, on parle d'agrégation robuste.

2.1.2. Interprétation comme un méta-problème statistique

L'objectif précisé ci-dessus se rapporte à la minimisation du regret, alors que l'on rappelle que l'objectif initial est d'assurer que la perte cumulée du statisticien est petite. Or, la décomposition

$$\widehat{L}_T(\mathscr{S}) = \inf_{\mathbf{q}\in\mathscr{P}} \big\{ L_T(\mathbf{q}) \big\} + R_T(\mathscr{S})$$

indique que cette perte cumulée est la somme d'une erreur d'approximation, donnée par la perte cumulée de la meilleure combinaison convexe constante des experts, et d'une erreur d'estimation, donnée par le regret et qui mesure la difficulté à se rapprocher, à cause de la contrainte séquentielle, de la performance de cette meilleure combinaison convexe constante. On notera d'ailleurs à cet

 ISSN: 2102-6238

égard que la valeur du vecteur de mélange optimal (ou quasi-optimal) pour les échéances 1 à T peut fortement varier avec T, s'agissant de suites individuelles.

En pratique, il faudrait donc arbitrer entre la considération d'un nombre N suffisamment grand d'experts aux comportements suffisamment différents afin de rendre l'erreur d'approximation la plus faible possible, et le fait qu'évidemment le regret $R_T(\mathscr{S})$ croît avec N. Cependant, cette croissance est, comme on le verra, très modérée en général : elle est de l'ordre de $\sqrt{\ln N}$. On a donc souvent intérêt à considérer un nombre important d'experts.

La question essentielle est alors de construire des experts ; pour l'instant, nous avions simplement formulé le problème en identifiant chaque expert à une boîte noire prédictive. Nous expliquerons sur les deux jeux de données considérés plus loin comment nous avons obtenu les experts mais illustrons par un exemple générique pourquoi le problème décrit dans cette partie est un méta-problème statistique. Dans un problème statistique classique où les observations (y_t) sont les réalisations d'un certain processus stochastique (Y_t), des méthodes stochastiques permettent d'obtenir des prévisions aléatoires ; on note $f_{j,t}$ la réalisation de la prévision de la j–ème méthode à l'échéance t, c'est-à-dire que l'on identifie cette méthode à un expert. Au lieu de sélectionner une méthode précise, on peut ici en considérer plusieurs et agréger leurs prévisions. Cette agrégation est effectuée, elle, de manière robuste, sans prendre en compte l'éventuel caractère stochastique des observations. En fait, comme les méthodes stochastiques habituellement utilisées dépendent d'un ou plusieurs paramètres, on peut considérer pour chacune plusieurs instances obtenues avec des jeux de paramètres différents, ce qui rend le réglage précis des paramètres moins crucial.

En conclusion, on considère ici l'agrégation robuste et non stochastique de prédicteurs fondamentaux qui, eux, peuvent reposer sur des techniques stochastiques. C'est en ce sens que l'on a affaire à un problème méta-statistique : on ne cherche pas à améliorer les performances individuelles des prédicteurs mais on vise à bien combiner leurs prévisions.

2.2. Une famille simple de stratégies d'agrégation : par poids exponentiels

Une stratégie naturelle mais qui échoue en général (au sens où son regret est de l'ordre de T) de est prédire à l'échéance t comme le meilleur des N experts sur les échéances 1 à $t-1$. Avec un peu de recul, on voit que le problème ici est que deux experts aux performances parfois très proches, en l'occurrence, les deux meilleurs experts sur le passé, ont des poids très différents, 0 pour le moins bon des deux et 1 pour le meilleur des deux. Une idée plus raisonnable est d'attribuer un poids $p_{j,t}$ simplement d'autant plus grand à l'expert j pour l'échéance t que ses performances ont été meilleures sur les échéances précédentes $1,\ldots,t-1$, sans qu'aucun de ces poids ne soit nul.

Notre première stratégie, notée $\mathscr{E}_\eta$ et dite de pondération par poids exponentiels des pertes cumulées, est décrite à la figure 1. On y note δ_j la masse de Dirac en j, de sorte que pour toute échéance $t \geqslant 2$,

$$L_{t-1}(\delta_j) = \sum_{s=1}^{t-1} \ell(f_{j,s}, y_s)$$

désigne la perte cumulée de l'expert j aux échéances précédentes.

Les garanties théoriques qu'offre $\mathscr{E}_\eta$ sont plus faibles que celles visées par l'objectif 2.1, puisque le théorème ci-dessous ne compare $\widehat{L}_T(\mathscr{E}_\eta)$ qu'à la perte cumulée du meilleur expert, et non pas de la meilleure combinaison convexe constante d'experts. On verra cependant plus loin quelle transformation simple effectuer dans la stratégie de la figure 1 pour atteindre cet objectif.

 ISSN: 2102-6238

Paramètre : Vitesse d'apprentissage $\eta > 0$

Initialisation : $\mathbf{p}_1$ est le mélange uniforme, soit $p_{j,1} = 1/N$ pour $j = 1,\dots,N$

Pour les échéances $t = 2,3,\dots,T$, le vecteur de mélange $\mathbf{p}_t$ est défini par la valeur de ses composantes $j = 1,\dots,N$ selon

$$p_{j,t} = \frac{e^{-\eta L_{t-1}(\delta_j)}}{\sum_{k=1}^N e^{-\eta L_{t-1}(\delta_k)}}\,.$$

FIGURE 1. *La stratégie $\mathscr{E}_\eta$ de pondération par poids exponentiels des pertes cumulées.*

Le théorème ci-dessous est l'un des résultats les plus fondamentaux et les plus connus en prévision de suites individuelles. Plusieurs versions en ont été données, par [34, 43, 44, 10, 23]. Nous reprenons ci-dessous un schéma de démonstration élémentaire (Lemme 2.1) suggéré par [9] et que l'on pourra retrouver également dans [12, paragraphe 2.2].

Théorème 1. *On suppose que la fonction de perte $\ell : \mathscr{X} \times \mathscr{Y}$ est bornée, à valeurs dans $[0,M]$, et convexe en son premier argument : pour tout $y \in \mathscr{Y}$, l'application $x \in \mathscr{X} \longmapsto \ell(x,y)$ est convexe. Alors, pour tout $\eta > 0$,*

$$\sup\left\{\widehat{L}_T(\mathscr{E}_\eta) - \min_{j=1,\dots,N} L_T(\delta_j)\right\} \leqslant \frac{\ln N}{\eta} + \eta\frac{M^2}{8}T\,,$$

où le supremum porte sur toutes les suites possibles d'observations et de prévisions des experts. En particulier, le choix de $\eta^\star = (1/M)\sqrt{(8\ln N)/T}$ conduit à la majoration

$$\sup\left\{\widehat{L}_T(\mathscr{E}_{\eta^\star}) - \min_{j=1,\dots,N} L_T(\delta_j)\right\} \leqslant M\sqrt{\frac{T}{2}\ln N}\,.$$

Ce théorème découle en fait du résultat plus générique proposé par le Lemme 2.1 ci-dessous, *via* la majoration linéaire (qui procède de la convexité de ℓ en son premier argument)

$$\widehat{L}_T(\mathscr{E}_\eta) \leqslant \sum_{t=1}^T \sum_{j=1}^N p_{j,t}\ell(f_{j,t},y_t)\,. \tag{2.1}$$

Lemme 2.1. *On fixe deux réels $m \leqslant M$. Pour tout $\eta > 0$ et pour toute suite d'éléments $\ell_{j,t} \in [m,M]$, où $j \in \{1,\dots,N\}$ et $t \in \{1,\dots,T\}$,*

$$\sum_{t=1}^T \sum_{j=1}^N \mu_{j,t}\ell_{j,t} - \min_{k=1,\dots,N} \sum_{t=1}^T \ell_{k,t} \leqslant \frac{\ln N}{\eta} + \eta\frac{(M-m)^2}{8}T\,, \tag{2.2}$$

où pour tout $j = 1,\dots,N$, on définit $\mu_{j,1} = 1/N$ et pour $t \geqslant 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta\sum_{s=1}^{t-1}\ell_{j,s}\right)}{\sum_{k=1}^N \exp\left(-\eta\sum_{s=1}^{t-1}\ell_{k,s}\right)}\,.$$

 ISSN: 2102-6238

Démonstration. La preuve repose sur le lemme de Hoeffding, dont on rappelle l'énoncé : si X est une variable aléatoire bornée, à valeurs dans $[m,M]$, alors pour tout $s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{sX}\right] \leqslant s\,\mathbb{E}[X] + \frac{s^2}{8}(M-m)^2 .$$

En particulier, pour tout $t = 1,2,\ldots$, en utilisant (pour le cas $t = 1$) la convention qu'une somme sur aucun élément est nulle,

$$-\eta \sum_{j=1}^{N} \mu_{j,t}\ell_{j,t} \geqslant \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{t} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)} - \frac{\eta^2}{8}(M-m)^2 ;$$

en sommant ces inégalités sur t et en divisant les deux membres par $-\eta < 0$, il vient

$$\sum_{t=1}^{T}\sum_{j=1}^{N} \mu_{j,t}\ell_{j,t} \leqslant -\frac{1}{\eta} \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{T} \ell_{j,s}\right)}{N} + \eta \frac{(M-m)^2}{8} T .$$

La preuve est alors conclue en minorant la somme de termes positifs restant dans le logarithme du membre de droite par le plus grand de ces termes. □

Remarque. *Lorsque η est mal calibré, la borne du Théorème 1 peut être grande : le choix optimal de η pour minimiser la borne théorique dépend de M et T, deux quantités que l'on n'a aucune raison de connaître. De plus, ce choix suppose que T soit fixé et on n'a donc pas encore assuré qu'il existait une stratégie dont le regret rapporté au nombre de tours T est asymptotiquement négatif lorsque $T \to \infty$. On expliquera ci-dessous, au paragraphe 2.4, comment calibrer, en théorie et en pratique, la vitesse d'apprentissage η et obtenir la garantie asymptotique énoncée à l'objectif 2.1.*

2.3. Pondération par poids exponentiels des gradients des pertes

Les stratégies $\mathscr{E}_\eta$ ne sont compétitives, comme l'indique le Théorème 1, que face au meilleur expert et non pas, comme le requiert l'objectif imposé, face à la meilleure combinaison convexe constante des experts. On explique ici comment remédier à cela. Ce qui suit a été inspiré par [32, 9] et peut également être retrouvé sous une forme proche dans [12, paragraphe 2.5].

Il s'agit toujours de rendre le regret linéaire en les vecteurs de mélange afin de pouvoir appliquer le Lemme 2.1 ; c'est donc l'étape (2.1) qui va être modifiée. A cet effet, on va employer des techniques d'analyse convexe en dimension réelle finie, dont la mise en œuvre est justifiée par l'hypothèse suivante.

Hypothèse 2.1. *On suppose que $\mathscr{X}$ est un sous-ensemble convexe de $\mathbb{R}^d$ et que soit les fonctions $\ell(\cdot,y)$ soient différentiables sur $\mathscr{X}$ pour tout $y \in \mathscr{Y}$, soit qu'il existe un ouvert U tel que ℓ puisse être étendue en une fonction $U \times \mathscr{Y}$ convexe en son premier argument.*

C'est un résultat élémentaire d'analyse convexe que sous la seconde version de l'hypothèse, il existe, pour tout $y \in \mathscr{Y}$, un (sous-)gradient de $\ell(\cdot,y)$ en tout point de U ; on note $\partial\ell(\cdot,y)$ un tel élément. Par définition, il vérifie que pour tout couple u, v de points de U,

$$\ell(u,y) - \ell(v,y) \leqslant \partial\ell(u,y) \cdot (u-v) ,$$

 ISSN: 2102-6238

Paramètre : Vitesse d'apprentissage $\eta > 0$

Initialisation : $\mathbf{p}_1$ est le mélange uniforme, soit $p_{j,1} = 1/N$ pour $j = 1, \ldots, N$

Pour les échéances $t = 2, 3, \ldots, T$, le vecteur de mélange $\mathbf{p}_t$ est défini par la valeur de ses composantes $j = 1, \ldots, N$ selon

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)},$$

où l'on a posé

$$\widetilde{\ell}_{j,s} = \partial \ell\left(\sum_{k=1}^{N} p_{k,s} f_{k,s}, y_s\right) \cdot f_{j,s}\,.$$

FIGURE 2. *La stratégie $\mathscr{E}_\eta^{\mathrm{grad}}$ de pondération par poids exponentiels des (sous-)gradients pertes cumulées.*

où $\cdot$ désigne le produit scalaire de $\mathbb{R}^d$.

En particulier, pour les deux versions de l'hypothèse, on a l'inégalité, pour tous les vecteurs de mélange $\mathbf{p}$ et $\mathbf{q}$, pour toutes les prévisions $f_1, \ldots, f_N$ et toute observation $y \in \mathscr{Y}$,

$$\ell\left(\sum_{j=1}^{N} p_j f_j, y\right) - \ell\left(\sum_{j=1}^{N} q_j f_j, y\right) \leqslant \partial \ell\left(\sum_{j=1}^{N} p_j f_j, y\right) \cdot \left(\sum_{j=1}^{N} p_j f_j - \sum_{j=1}^{N} q_j f_j\right). \tag{2.3}$$

Définissons les pseudo-pertes suivantes, pour l'expert $j \in \{1, \ldots, N\}$ à l'échéance $t \in \{1, \ldots, T\}$:

$$\widetilde{\ell}_{j,t} = \partial \ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t}, y_t\right) \cdot f_{j,t} \tag{2.4}$$

et considérons la famille de stratégies d'agrégation de la figure 2, appelées $\mathscr{E}_\eta^{\mathrm{grad}}$. Le regret de $\mathscr{E}_\eta^{\mathrm{grad}}$ est alors majoré selon

$$\begin{aligned} R_T\left(\mathscr{E}_\eta^{\mathrm{grad}}\right) &= \sup_{\mathbf{q} \in \mathscr{P}} \sum_{t=1}^{T} \left(\ell\left(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t\right) - \ell\left(\sum_{j=1}^{N} q_j f_{j,t}, y_t\right)\right) \\ &\leqslant \sup_{\mathbf{q} \in \mathscr{P}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \sum_{j=1}^{N} q_j \widetilde{\ell}_{j,t}\right) = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{k=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{k,t}\,, \end{aligned}$$

où l'inégalité procède de (2.3) et la seconde égalité du fait que le majorant ainsi obtenu est linéaire en $\mathbf{q}$ et est donc maximisé par un vecteur de mélange égal à une masse de Dirac. Le résultat suivant découle ensuite immédiatement du Lemme 2.1.

Théorème 2. *On suppose que l'hypothèse 2.1 est vérifiée et que les pseudo-pertes définies en (2.4) sont bornées, à valeurs dans un intervalle $[-C, C]$. Alors, pour tout $\eta > 0$,*

$$\sup\left\{R_T\left(\mathscr{E}_\eta^{\mathrm{grad}}\right)\right\} = \sup\left\{\widehat{L}_T\left(\mathscr{E}_\eta^{\mathrm{grad}}\right) - \inf_{\mathbf{q} \in \mathscr{P}} L_T(\mathbf{q})\right\} \leqslant \frac{\ln N}{\eta} + \eta \frac{C^2}{2} T\,,$$

où le supremum porte sur toutes les suites possibles d'observations et de prévisions des experts. En particulier, le choix de $\eta^\star = (1/C)\sqrt{(2 \ln N)/T}$ conduit à la majoration

$$\sup\left\{R_T\left(\mathscr{E}_{\eta^\star}\right)\right\} \leqslant C\sqrt{2T \ln N}\,.$$

 ISSN: 2102-6238

Ici encore, un problème de calibration se pose pour le choix de η, qui dépend de la borne C sur les pseudo-pertes et du nombre d'échéances T, deux quantités éventuellement inconnues par avance. C'est pourquoi nous traitons maintenant de la calibration de la vitesse d'apprentissage.

2.4. Adaptation aux paramètres : calibration de la vitesse d'apprentissage η

Nous allons étudier deux techniques, l'une à vocation théorique, l'autre à vocation pratique.

2.4.1. Méthode théorique

Dans les figures 1 et 2, on remplace les occurrences de η dans la définition de $p_{j,t}$ par η_t, où (η_t) est une suite décroissante de vitesses d'apprentissage qui peuvent dépendre du passé ; c'est-à-dire qu'à l'échéance $t \geqslant 2$, la vitesse η_t est choisie en fonction de la même information que celle disponible pour choisir le vecteur de mélange $\mathbf{p}_t$, à savoir les observations passées et les conseils passés et présents des experts.

Le premier tel choix est proposé dans [3] ; il suppose la connaissance d'une borne M sur les pertes (ou C sur les pseudo-pertes) et s'adapte à T. Plus précisément, en choisissant

$$\eta_t = \frac{1}{M}\sqrt{\frac{8\ln N}{t-1}}\,,$$

[12, Théorème 2.3] a montré que le regret sur les T premières échéances est uniformément majoré par $M\sqrt{2T\ln N} + M\sqrt{(\ln N)/8}$, et ce, pour tout $T \geqslant 2$.

Un problème un peu plus délicat est de traiter également l'adaptation en la borne M sur les pertes. Nous avons proposé dans [13, Théorème 6] de choisir η_t en fonction, notamment, de t et

$$M_{t-1} = \max\left\{\ell(f_{j,s}, y_s),\ j \in \{1,\dots,N\} \text{ et } s \in \{1,\dots,t-1\}\right\},$$

qui représente la meilleure information disponible sur la borne M au début du tour t. (On ne reporte pas ici la forme précise de η_t, par souci de simplicité du propos.) Nous avons alors obtenu la borne uniforme sur le regret $2M\sqrt{T\ln N} + 6M\ln N + 6M$. Ce résultat a été récemment amélioré, [24] proposant la borne uniforme

$$\sqrt{(e-2)\left(\sqrt{2}+1\right)}\,M\sqrt{T\ln N} + 2M\ln N + 6M \leqslant 1.87\,M\sqrt{\frac{T}{2}\ln N} + 2M\ln N + 6M\,.$$

La borne de regret du Théorème 1 est optimale, comme le prouve [10], tant dans les ordres de grandeur en T et N qu'en ce qui concerne la constante $1/\sqrt{2}$. Le coût actuel pour l'adaptation en M et T est donc, au vu de la borne de [24], légèrement inférieur à 2. Nous ne connaissons pas encore la valeur optimale de ce coût (en particulier, nous ne savons pas s'il diffère de 1, *id est*, si coût il doit y avoir).

2.4.2. Méthode pratique

Comme on le verra sur les données réelles, les valeurs optimales de η pour minimiser les bornes théoriques des Théorèmes 1 et 2, de même que celles obtenues par les méthodes adaptatives

 ISSN: 2102-6238

présentées ci-dessus, conduisent en général à de mauvaises performances. On peut expliquer cela par le fait que ces bornes sont conçues par rapport à l'ensemble de toutes les suites individuelles et correspondent ainsi à des algorithmes trop précautionneux et qui ont un temps de réaction un peu trop long. Une idée naturelle consiste donc à augmenter la valeur des vitesses d'apprentissage η_t, toute la question étant d'exhiber une manière adaptative et performante de le faire.

Notons, avant de continuer, que ces remarques sur un apprentissage plus rapide ne remettent pas en cause notre méthodologie d'agrégation de prédicteurs. Là encore, les résultats pratiques montreront que des stratégies d'agrégation calibrées avec des vitesses suffisamment rapides obtiennent des performances bien meilleures que celles du meilleur expert ou du meilleur vecteur de mélange constant (qui, en outre, ne sont connus qu'à la fin de la période de prévision) et que cela se traduit, au niveau des vecteurs de mélange calculés par les stratégies, par des vecteurs qui ne ressemblent absolument pas à des masses de Dirac.

Ce qui suit est à porter au crédit de Vivien Mallet, a été présenté la première fois dans [25] puis a été ré-utilisé et approfondi dans [17]. On présente la méthode de calibration pratique par exemple pour la famille $\mathscr{E}_\eta$ des stratégies de pondération par poids exponentiels des pertes cumulées ; celle-ci repose sur la considération de toutes les stratégies de cette famille. C'est pourquoi il nous faudra écrire explicitement la dépendance du vecteur de mélange $\mathbf{p}_t$ en la stratégie $\mathscr{E}_\eta$ qui le prescrit, ce que l'on fait en le notant $\mathbf{p}_t(\mathscr{E}_\eta)$.

La méta-stratégie calibrée choisit alors, à l'échéance t, le paramètre η_t ayant obtenu les meilleurs performances dans le passé puis recourt au vecteur de mélange $\mathbf{p}_t\left(\mathscr{E}_{\eta_t}\right)$. Formellement,

$$\eta_t \in \operatorname*{arg\,min}_{\eta>0} \widehat{L}_{t-1}(\mathscr{E}_\eta) \tag{2.5}$$

(ou un antécédent du minimum à un facteur de tolérance près si le minimum n'est pas atteint).

Plusieurs remarques s'imposent. Premièrement, même si cette technique obtient d'excellentes performances pratiques au sens où sa perte cumulée jusqu'à l'échéance T se rapproche souvent de celle de $\mathscr{E}_{\eta_T}$, *id est*, du meilleur algorithme de la famille $\mathscr{E}_\eta$ sur les données, nous n'avons pour l'instant pas encore de garanties théoriques à proposer sur son regret. Deuxièmement, le calcul pratique d'un antécédent du minimum, même à un facteur de tolérance près, est une opération délicate. Plusieurs idées peuvent être avancées pour la mettre en œuvre : une dichotomie ou l'utilisation d'une grille. La méthode de dichotomie repose sur un fait souvent observé en pratique : pour toute échéance t, la fonction $\eta \mapsto L_t(\mathscr{E}_\eta)$ est décroissante sur un premier intervalle puis croissante sur son complémentaire.

Pour la méthode reposant sur une grille, on construit cette dernière à partir d'une échelle logarithmique en devant fixer un certain pas de discrétisation, ainsi qu'une borne inférieure et une borne supérieure, ce qui permet de ne conserver qu'un nombre fini de points η. A chaque échéance, on ne réalise alors la minimisation (2.5) que sur la grille et non pas sur tout $\mathbb{R}_+^\star$. Le choix de ces bornes inférieure et supérieure peut être effectué adaptativement, en exploitant à nouveau le fait observé de monotonie sur deux intervalles. A chaque échéance, on complète éventuellement la grille en ajoutant un point en-deçà de sa borne inférieure (respectivement, au-delà de sa borne supérieure) si le point précédent (respectivement, suivant) obtient une erreur cumulée plus petite que la borne inférieure (respectivement, supérieure) actuelle. Nous avons vérifié en pratique que le choix du paramètre de discrétisation logarithmique n'influe en revanche pas trop fortement sur les performances de la méta-stratégie calibrée.

 ISSN: 2102-6238

2.5. *Deux variantes : fenêtrage et escompte*

Une critique souvent opposée aux stratégies des figures 1 et 2, et qui n'est pas sans lien avec la mention ci-dessus de leur tempérament parfois précautionneux, est qu'elles tiennent trop compte du passé. Le passé proche semble nécessaire et utile mais l'utilisation du passé lointain paraît souvent moins profitable à tous ceux qui sont familiers du cadre stochastique, notamment lorsque ce dernier est non stationnaire. Une variante de nos stratégies ne reposant que sur le passé proche est donnée par la considération d'une fenêtre maximale d'historique H et du remplacement des définitions des composantes de $\mathbf{p}_t$ aux figures 1 et 2 par, respectivement,

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \ell(f_{j,s}, y_s)\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \ell(f_{k,s}, y_s)\right)}$$

$$\text{et} \qquad p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

pour tous $t \geqslant 2$ et $j = 1, \ldots, N$. On parle de fenêtrage du passé ; il semble peu vraisemblable que des bornes sur le regret uniformes en les suites individuelles soient conservées par fenêtrage.

Une manière de réconcilier tous les points de vue est de dire que le passé est d'autant plus significatif qu'il est proche sans toutefois considérer que le passé lointain soit inutile ; cela se traduit mathématiquement en escomptant les pertes passées par un facteur multiplicatif strictement positif mais d'autant plus petit que le passé est lointain. Formellement, on fixe deux suites décroissantes de réels strictement positifs, les escomptes (β_t) et les vitesses d'apprentissage (η_t) ; on remplace alors les définitions des composantes de $\mathbf{p}_t$ aux figures 1 et 2 par, respectivement,

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \ell(f_{j,s}, y_s)\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \ell(f_{k,s}, y_s)\right)} \tag{2.6}$$

$$\text{et} \qquad p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \widetilde{\ell}_{k,s}\right)} \tag{2.7}$$

pour tous $t \geqslant 2$ et $j = 1, \ldots, N$. Des bornes uniformes sur le regret de ces stratégies peuvent alors être prouvées ; elles dépendent évidemment des suites (η_t) et (β_t), cette dernière ne devant pas décroître trop rapidement vers 0. On a plus précisément le résultat suivant.

Théorème 3. *L'objectif 2.1 est atteint pour les deux stratégies par escomptes présentées en (2.6) et (2.7) lorsque les vitesses d'apprentissage et les escomptes vérifient que*

$$t\,\eta_t \to \infty \qquad \text{et} \qquad \eta_t \sum_{s=1}^{t-1} \beta_s \longrightarrow 0$$

lorsque $t \to \infty$ et que la fonction de perte ℓ est bornée et convexe en son premier argument.

Démonstration. Une preuve détaillée est fournie par [37, chapitre 6]. L'idée consiste en un schéma par approximation, où l'on quantifie les écarts (faibles) entre les vecteurs de mélange définis avec escomptes et les vecteurs de mélange sans escomptes puis l'on ajoute ces écarts aux bornes des Théorèmes 1 et 2. □

 ISSN: 2102-6238

Il est à noter que ces techniques d'escompte ont été introduites en apprentissage séquentiel dans [12, paragraphe 2.11], à ceci près qu'il est absolument crucial pour l'analyse qui y est effectuée que le nombre d'échéances T ait une valeur fixée et connue à l'avance.

3. Deux autres types de stratégies répondant à des attentes et questions habituelles

On présente dans cette partie deux autres types de stratégies de prévision (deux familles par type, soit quatre familles au total). Les premières reposent sur une agrégation donnée par une régression linéaire séquentielle effectuée avec facteur de régularisation. Les secondes arrivent à maîtriser le regret face non plus seulement au meilleur expert ou au meilleur vecteur de mélange constant, mais face à la meilleure suite d'experts ou de vecteurs de mélanges avec un nombre de ruptures sous-linéaire.

3.1. Régressions linéaires séquentielles avec facteurs de régularisation

On se limite ici au cas où les ensembles d'observations et de prévisions sont donnés par la droite réelle, $\mathscr{Y} = \mathscr{X} = \mathbb{R}$, et où la fonction de perte ℓ est la perte quadratique, $\ell(x,y) = (x-y)^2$.

On s'autorise dans ce paragraphe à choisir à chaque échéance t des vecteurs de mélange arbitraires $\mathbf{u}_t = (u_{1,t}, \ldots, u_{N,t}) \in \mathbb{R}^N$; on n'impose aucune contrainte de positivité ou de somme égale à 1. La prévision du statisticien est alors la combinaison linéaire

$$\widehat{y}_t = \sum_{j=1}^{N} u_{j,t} f_{j,t} \,. \tag{3.1}$$

L'avantage de ce cadre est qu'en ôtant la condition de somme égale à 1, on peut espérer que les stratégies de prévision compensent de manière automatique un biais éventuel commun à tous les experts ; c'est d'ailleurs ce que l'on observe en pratique sur certains jeux de données. L'inconvénient est qu'en général les vecteurs de mélange $\mathbf{u}_t$ ont de nombreuses composantes négatives, ce qui les rend bien moins facilement interprétables que leurs homologues convexes $\mathbf{p}_t$.

3.1.1. Régularisation ℓ^2 : la régression ridge

La première stratégie de prévision est la régression ridge, introduite par [31] dans un contexte stochastique et étudiée par [4] et [45] dans le cadre des suites individuelles. Elle repose sur une régularisation ℓ^2 ; on note à cet effet

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{j=1} u_j^2}$$

la norme euclidienne d'un vecteur $\mathbf{u} \in \mathbb{R}^N$. Elle dépend d'un paramètre $\lambda > 0$ et on la note $\mathscr{R}_\lambda$; elle choisit, à toute échéance $t \geqslant 1$, un vecteur de mélange $\mathbf{u}_t$ vérifiant

$$\mathbf{u}_t \in \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^N} \left\{ \lambda \|\mathbf{v}\|_2^2 + \sum_{s=1}^{t-1} \left(y_s - \sum_{j=1}^{N} v_j f_{j,s} \right)^2 \right\} \tag{3.2}$$

 ISSN: 2102-6238

avec la convention habituelle qu'une somme sur aucun élément est nulle (de sorte que $\mathbf{u}_1$ est le vecteur nul).

On précise tout d'abord la borne de performance générale, qui est un peu implicite, mais on indique ensuite comment des hypothèses de domaines bornés pour $\mathscr{X}$ et $\mathscr{Y}$ permettent (presque) de pallier ce manquement. Pour énoncer cette borne de manière compacte, on définit le vecteur (ligne) $\mathbf{f}_t = (f_{1,t}, \ldots, f_{N,t})$ des prévisions des experts à l'échéance t.

Théorème 4. *On définit la matrice $M_T = \sum_{t=1}^{T} \mathbf{f}_t^{\mathrm{T}} \mathbf{f}_t$ et on note $\mu_{1,T}, \ldots, \mu_{N,T}$ ses valeurs propres. Pour toutes les suites de prévisions $\mathbf{f}_1, \ldots, \mathbf{f}_T$ et d'observations $y_1, \ldots, y_T$, le regret de $\mathscr{R}_\lambda$ par rapport à tout vecteur de mélange $\mathbf{v} \in \mathbb{R}^N$ est contrôlé de la manière suivante :*

$$\sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} u_{j,t} f_{j,t} \right)^2 - \sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} v_j f_{j,t} \right)^2 \leqslant \frac{\lambda}{2} \|\mathbf{v}\|_2^2 + \left(\sum_{j=1}^{N} \ln\left(1 + \frac{\mu_{j,T}}{\lambda} \right) \right) \max_{t \leqslant T} \left(y_t - \sum_{j=1}^{N} u_{j,t} f_{j,t} \right)^2 .$$

Une preuve compacte du théorème peut être trouvée dans [12, paragraphe 11.7], de même que la discussion qui suit. La restriction de $\mathscr{X}$ et $\mathscr{Y}$ à un domaine borné $[-B, B]$ permet (presque) d'ôter toute dépendance de la borne en la suite des prévisions $f_{j,t}$ des experts et des observations y_t. En effet, par un argument simple d'algèbre linéaire, il vient alors

$$\sum_{j=1}^{N} \ln\left(1 + \frac{\mu_{j,T}}{\lambda} \right) \leqslant N \ln\left(1 + \frac{B^2 T}{\lambda N} \right) .$$

On en déduit le contrôle suivant sur la perte cumulée de $\mathscr{R}_\lambda$:

$$\sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} u_{j,t} f_{j,t} \right)^2 \leqslant \inf_{\mathbf{v} \in \mathbb{R}^N} \left\{ \frac{\lambda}{2} \|\mathbf{v}\|_2^2 + \sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} v_j f_{j,t} \right)^2 \right\} + N \ln\left(1 + \frac{B^2 T}{\lambda N} \right) \max_{t \leqslant T} \left(y_t - \sum_{j=1}^{N} u_{j,t} f_{j,t} \right)^2 .$$

Dans cette borne de performance, il n'est pas évident de majorer uniformément le maximum en $t \leqslant T$; il suffirait d'être sûr que toutes les prévisions $\widehat{y}_t$ définies en (3.1) soient, comme les prévisions des experts $f_{j,t}$, dans l'intervalle $[-B, B]$, mais cela n'est pas certain.

A cet égard, il faut noter qu'en pratique, on pourrait lancer $\mathscr{R}_\lambda$ en stratégie maître et seuiller ses prévisions $\widehat{y}_t$ à $-B$ ou B si, respectivement, on avait $\widehat{y}_t \leqslant -B$ ou $\widehat{y}_t \geqslant B$. La stratégie seuillée ainsi obtenue obtient des performances meilleures que celles de $\mathscr{R}_\lambda$; mais il n'est pas possible pour autant d'obtenir pour elle une meilleure borne que celle du Théorème 4 (la borne reste en fonction des vecteurs de mélange prescrits par la stratégie maître $\mathscr{R}_\lambda$).

Remarque. *Il est aisé d'exhiber une solution explicite au problème de minimisation (3.2) : en désignant par $^{-1}$ un pseudo-inverse, on a*

$$\mathbf{u}_t = \left(\lambda I_N + M_{t-1} \right)^{-1} \sum_{s=1}^{t-1} y_s \mathbf{f}_s .$$

 ISSN: 2102-6238

On conclut ce paragraphe en notant qu'ici encore, un problème de calibration de λ se pose. La détermination de la valeur théorique optimale qui minimiserait la borne du Théorème 4 n'est pas évidente ; en pratique, on recourra plutôt aux techniques de calibration séquentielles du paragraphe 2.4.2.

3.1.2. Régularisation ℓ^1 : Lasso séquentiel

Une version plus moderne de la régression linéaire régularisée remplace la régularisation ℓ^2 utilisée dans (3.2) par une régularisation ℓ^1 : on note à cet effet

$$\|\mathbf{u}\|_1 = \sum_{j=1} |u_j|$$

la norme ℓ^1 d'un vecteur $\mathbf{u} \in \mathbb{R}^N$. Pour un facteur de régularisation $\lambda > 0$ fixé, la stratégie choisit alors, à chaque échéance $t \geqslant 1$,

$$\mathbf{u}_t \in \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^N} \left\{ \lambda \|\mathbf{v}\|_1 + \sum_{s=1}^{t-1} \left(y_s - \sum_{j=1}^{N} v_j f_{j,s} \right)^2 \right\} . \tag{3.3}$$

On l'appelle la stratégie de Lasso séquentiel et on la note $\mathscr{L}_\lambda$.

Le Lasso a été introduit et étudié dans un cadre stochastique [42] ; il a depuis connu un succès fulgurant et s'est révélé remarquablement adapté aux problèmes de régression en grande dimension. En effet, l'avantage de la régularisation ℓ^1 de (3.3) est qu'elle permet de retenir des vecteurs $\mathbf{u}_t$ ayant peu de composantes non nulles.

Un inconvénient, cependant, est qu'il n'existe pas d'expression explicite pour les solutions de (3.3), même s'il existe un algorithme, l'algorithme LARS de [20], qui permet de calculer efficacement leur valeur sur des données. Ceci est à comparer au problème (3.2), pour lequel on a vu qu'il était aisé d'exhiber une expression explicite de $\mathbf{u}_t$ en fonction de λ et du passé.

A ce jour et au meilleur de notre connaissance, il n'existe pas de borne de suites individuelles pour le Lasso séquentiel (la forme souhaitée de la borne serait celle du Théorème 4).

3.2. Regret face à des suites d'experts avec un nombre pas trop élevé de ruptures

Une remarque fréquente est que se rapprocher des performances du meilleur expert (ou de la meilleure combinaison convexe constante des experts) peut constituer un objectif manquant d'ambition. Cette remarque n'est réellement fondée que lorsqu'un expert domine les autres de manière quasi-permanente dans le temps, auquel cas il est évidemment facile de déterminer rapidement qui est cet expert plus efficace que tous les autres et de ne plus suivre que lui. L'histoire est moins claire lorsque plusieurs experts sont au coude-à-coude en termes de performances cumulées ; à tout moment l'un d'entre eux peut connaître une subite amélioration ou détérioration de ses performances et il faut combiner leurs prévisions de manière suffisamment prudente. C'est essentiellement cette situation qui motivait les bornes exhibées jusqu'ici dans cet article.

Cependant, pour rejoindre la remarque souvent formulée, on peut imaginer que dans certaines applications, le temps peut être divisé en un nombre pas trop élevé d'intervalles sur chacun

 ISSN: 2102-6238

desquels un expert domine nettement les autres en termes de performances. Si l'on note m ce nombre d'intervalles, que ces derniers correspondent alors à $1,\dots,T_1$, puis $T_1+1,\dots,T_2$, etc., jusque $T_{m-1}+1,\dots,T_m$, et que $j_1^\star,\dots,j_m^\star$ désignent les indices du meilleur expert dans chacun de ces intervalles, alors on a envie que les performances de la stratégie du statisticien ne soient pas trop éloignées de la perte cumulée suivante, où par convention $T_0=1$:

$$\sum_{r=1}^{m}\sum_{t=T_{r-1}+1}^{T_r}\ell\big(f_{j_r^\star},y_t\big)\,. \tag{3.4}$$

C'est pourquoi [30] a introduit la classe des experts composés. Pour un nombre d'échéances T, cette classe peut être identifiée à $\mathscr{C}_T=\{1,\dots,N\}^T$. Pour tout élément $j_1^T=(j_1,\dots,j_T)$ de $\mathscr{C}_T$, on note

$$L_T\big(j_1^T\big)=\sum_{t=1}^{T}\ell(f_{j_t},y_t)$$

la perte cumulée de l'expert composé lui correspondant.

Il est évidemment impossible, pour toute stratégie, d'être compétitive face au meilleur expert composé : cela reviendrait essentiellement à connaître l'indice du meilleur expert pour la prochaine échéance. Ainsi, il faut contraindre un peu les experts composés, en requérant par exemple que leur nombre de ruptures ne soit pas trop grand, où ce nombre est défini, pour un expert composé j_1^T, par

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

On parle également de sauts (d'où la définition par une fonction notée s). Par exemple, l'expert composé de (3.4) admet $m-1$ ruptures.

Nous allons présenter deux stratégies de prévision, l'une fondée uniquement sur les pertes (version dite initiale) et l'autre, sur leurs gradients. Toutes deux forment, à chaque échéance t, des combinaisons convexes précisées par des vecteurs de mélange notés par conséquent $\mathbf{p}_t$.

3.3. Stratégie de redistribution des poids, version initiale

On appelle la stratégie présentée à la figure 3 la stratégie de redistribution des poids, à cause de son étape (3). Elle est fondée sur l'utilisation de poids exponentiels, cf. étape (2), que l'on mélange entre eux à l'étape (3). Elle a été introduite par [30] mais on pourra également se référer à [12, paragraphe 5.2]. On la note $\mathscr{F}_{\eta,\alpha}$ parce qu'elle dépend de deux paramètres η et α, la lettre $\mathscr{F}$ venant pour sa part du nom anglais de la stratégie ("fixed share"). Dans le cas où $\alpha=0$, on retrouve la stratégie de pondération par poids exponentiels : $\mathscr{F}_{\eta,0}=\mathscr{E}_\eta$; on ne considérera donc dans la suite de ce paragraphe que le cas $\alpha>0$.

En fait, le résultat essentiel ici, qui conduit à l'obtention d'une borne sur le regret, est que $\mathscr{F}_{\eta,\alpha}$ effectue simplement une mise en œuvre efficace et séquentielle de la stratégie de pondération par poids exponentiels sur l'ensemble des experts composés, chacun étant affecté d'un poids initial non pas uniforme mais fonction du nombre de ses sauts. Plus précisément, le poids initial de j_1^T est égal à

$$\frac{1}{N}\left(\frac{\alpha}{N-1}\right)^{s(j_1^T)}(1-\alpha)^{T-s(j_1^T)-1}\,,$$

 ISSN: 2102-6238

Paramètres : Vitesse d'apprentissage $\eta > 0$ et fraction de redistribution $\alpha \in]0,1[$

Initialisation : Vecteur de poids uniforme $(w_{1,0},\dots,w_{N,0}) = (1,\dots,1)$

Pour les échéances $t = 1,2,\dots,T$,

(1) $\mathbf{p}_t$ est défini par la valeur de ses composantes $j = 1,\dots,N$ selon

$$p_{j,t} = \frac{w_{j,t-1}}{\sum_{k=1}^N w_{k,t-1}}\,,$$

ce qui correspond à la prévision

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

(2) Prise en compte des pertes : le statisticien accède à l'observation y_t et définit, pour tout $j = 1,\dots,N$,

$$v_{j,t} = w_{j,t-1} e^{-\eta \ell(f_{j,t},y_t)}\,;$$

(3) Redistribution des poids : on note $V_t = v_{1,t} + \dots + v_{N,t}$ et on définit, pour tout $j = 1,\dots,N$,

$$w_{j,t} = (1-\alpha) v_{j,t} + \frac{\alpha}{N} V_t\,.$$

FIGURE 3. *La stratégie $\mathscr{F}_{\eta,\alpha}$ de redistribution des poids.*

c'est-à-dire qu'on lui attribue la probabilité de sa réalisation sous le processus markovien qui choisirait le premier élément au hasard puis avec probabilité $1-\alpha$ à chaque pas conserverait la valeur précédente et avec probabilité α tirerait une nouvelle valeur au hasard. L'idée est alors d'en revenir à la borne (2.2) ; on peut montrer, en reprenant les calculs, que le facteur N y est en fait l'inverse du poids initial de l'expert j auquel on se compare. Ici, par imitation, on obtient alors la borne suivante, lorsque les pertes sont à valeurs dans $[0,M]$ et convexes en leur premier argument :

$$\widehat{L}_T(\mathscr{F}_{\eta,\alpha}) - L_T\left(j_1^T\right) \leqslant \frac{1}{\eta} \ln \frac{1}{(1/N)\left(\alpha/(N-1)\right)^{s(j_1^T)} (1-\alpha)^{T-s(j_1^T)-1}} + \eta \frac{M^2}{8} T\,.$$

Afin d'avoir la forme habituelle des bornes de regret, où l'on se compare au meilleur expert d'une certaine classe d'experts, on se restreint par exemple à la classe $\mathscr{C}_{T,m}$ des experts composés j_1^T tels que $s\left(j_1^T\right) \leqslant m$ pour un entier $m \geqslant 1$.

En utilisant le fait que pour $\alpha \leqslant 1/2$, la suite $k \mapsto \left(\alpha/(1-\alpha)\right)^k$ est décroissante, on obtient alors le résultat suivant.

Théorème 5. *On suppose que la fonction de perte $\ell : \mathscr{X} \times \mathscr{Y}$ est bornée, à valeurs dans $[0,M]$, et convexe en son premier argument. Alors, pour tout $0 < \alpha \leqslant 1/2$ et $\eta > 0$, le regret de $\mathscr{F}_{\eta,\alpha}$ est uniformément borné par rapport à la classe $\mathscr{C}_{T,m}$, où $m \geqslant 1$, selon*

$$\sup \left\{ \widehat{L}_T(\mathscr{F}_{\eta,\alpha}) - \min_{j_1^T \in \mathscr{C}_{T,m}} L_T\left(j_1^T\right) \right\} \leqslant \frac{1}{\eta} \ln \frac{N^{m+1}}{\alpha^m (1-\alpha)^{T-m-1}} + \eta \frac{M^2}{8} T\,,$$

où le supremum porte sur toutes les suites possibles d'observations et de prévisions des experts.

 ISSN: 2102-6238

On précise maintenant comment minimiser la borne uniforme en calibrant α et η en fonction de T, M et m, avec bien sûr le problème désormais habituel que le statisticien n'a aucune raison de connaître à l'avance T et M d'une part, et que d'autre part, le choix *a priori* de m est délicat, puisque ce dernier s'effectue en considérant un compromis entre la performance de la classe $\mathscr{C}_{T,m}$ et la borne de regret exhibée ci-dessous. Les choix (quasi-)optimaux sont donnés en termes de la fonction d'entropie binaire $H : x \in [0,1] \longmapsto x\ln x + (1-x)\ln(1-x)$; plus précisément,

$$\eta^\star = \frac{1}{M}\sqrt{\frac{8}{T}\left((m+1)\ln N + (T-1)\,H\left(\frac{m}{T-1}\right)\right)}$$

et $\alpha^\star = m/(T-1)$ conduisent à la borne

$$\sup\left\{\widehat{L}_T(\mathscr{F}_{\eta,\alpha}) - \min_{j_1^T \in \mathscr{C}_{T,m}} L_T\left(j_1^T\right)\right\} \leqslant M\sqrt{\frac{T}{2}\left((m+1)\ln N + (T-1)\,H\left(\frac{m}{T-1}\right)\right)}.$$

En pratique, on recourt aux techniques de grilles du paragraphe 2.4.2 pour calibrer α et η (il faut donc évidemment une grille à deux dimensions ici).

3.4. Stratégie de redistribution des poids, sur les (sous-)gradients des pertes

On applique ici les mêmes arguments qu'au paragraphe 2.3 pour passer de garanties par rapport à des experts individuels à des garanties par rapport à des combinaisons convexes fixes d'experts. La classe de comparaison sera cette fois-ci formée des éléments $\mathbf{q}_1^T = (\mathbf{q}_1, \ldots, \mathbf{q}_T)$, où chaque $\mathbf{q}_t = (q_{1,t}, \ldots, q_{N,t})$ est un vecteur de mélange issu du simplexe $\mathscr{P}$; on note $\mathscr{C}_T^{\text{cvx}}$ l'ensemble de ces éléments et on parlera de vecteurs de mélange composés. La perte cumulée d'un tel élément est définie par

$$L_T\left(\mathbf{q}_1^T\right) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_{j,t} f_{j,t}, y_t\right).$$

De même que précédemment on définit le nombre de ruptures d'un vecteur de mélange composé comme

$$s\left(\mathbf{q}_1^T\right) = \sum_{t=2}^{T} \mathbb{I}_{\{\mathbf{q}_{t-1} \neq \mathbf{q}_t\}}.$$

Pour tout entier $m \geqslant 0$, on note alors $\mathscr{C}_{T,m}^{\text{cvx}}$ la sous-classe de $\mathscr{C}_T^{\text{cvx}}$ formée par les vecteurs de mélange composés admettant au plus m ruptures.

On remplace les occurrences des pertes $\ell(f_{j,t}, y_t)$ dans l'étape (2) de la figure 3 par les pseudo-pertes $\widetilde{\ell}_{j,t}$ définies en (2.4). On appelle la stratégie ainsi obtenue la stratégie de redistribution des poids par (sous-)gradients et on la note $\mathscr{F}_{\eta,\alpha}^{\text{grad}}$.

Les majorations du paragraphe 2.3 montrent que le Théorème 5 entraîne le résultat suivant.

Corollaire 3.1. *On suppose que l'hypothèse 2.1 est vérifiée et que les pseudo-pertes définies en (2.4) sont bornées, à valeurs dans un intervalle* $[-C, C]$. *Alors, pour tout* $0 < \alpha \leqslant 1/2$ *et* $\eta > 0$,

 ISSN: 2102-6238

le regret de $\mathscr{F}_{\eta,\alpha}^{\mathrm{grad}}$ *est uniformément borné par rapport à la classe* $\mathscr{C}_{t,m}^{\mathrm{cvx}}$*, où* $m \geqslant 1$*, selon*

$$\sup\left\{\widehat{L}_T\left(\mathscr{F}_{\eta,\alpha}^{\mathrm{grad}}\right) - \min_{\mathbf{q}_1^T \in \mathscr{C}_{T,m}^{\mathrm{cvx}}} L_T\left(\mathbf{q}_1^T\right)\right\} \leqslant \frac{1}{\eta}\ln\frac{N^{m+1}}{\alpha^m(1-\alpha)^{T-m-1}} + \eta\frac{C^2}{2}T\,,$$

où le supremum porte sur toutes les suites possibles d'observations et de prévisions des experts.

Les mêmes résultats et commentaires qu'au paragraphe précédent s'appliquent en ce qui concerne l'optimisation en les paramètres α et η en fonction de T, C et m.

4. Application à la prévision de la qualité de l'air

Dans cette partie, on présente le contexte et les données relatives au problème de prévision de qualité de l'air. On explique alors brièvement comment adapter les stratégies générales présentées ci-dessus à ce cadre applicatif, avant d'en détailler les performances pratiques. On conclut par quelques perspectives de recherche et par une synthèse des liens et différences avec l'utilisation d'autres approches d'apprentissage (notamment, les arbres de régression).

Par souci de concision, on ne présentera que les grandes lignes des résultats obtenus ; davantage de détails peuvent être obtenus en consultant l'article [39] ainsi que les rapports techniques [37, 25].

4.1. Description du jeu de données et des experts utilisés

4.1.1. Jeu de données

Les données étudiées correspondent en temps à la période du 28 avril au 31 août 2001, qui comporte donc $T = 126$ jours, et en espace à un ensemble de 241 sites (appelés également stations dans la suite) en France et en Allemagne : 116 sites en France métropolitaine, 81 sites en Allemagne, uniformément répartis dans chacun des deux pays. Dans cet article, on ne discutera que des résultats obtenus pour la prévision des pics journaliers d'ozone : à chaque jour t et à chaque site s, on associe la quantité y_t^s, qui est la valeur maximale de la concentration en ozone au cours de la journée en ce site. Les indices t et s prennent respectivement leurs valeurs dans les ensembles $\{1,\ldots,126\}$ et $\mathscr{N} = \{1,\ldots,241\}$.

Les mesures de concentration sont données en microgrammes par mètre cube ($\mu\mathrm{g\,m^{-3}}$), une unité généralement omise par la suite. On rappelle à cet égard que les concentrations typiques sont de l'ordre de 40 $\mu\mathrm{g\,m^{-3}}$ à 150 $\mu\mathrm{g\,m^{-3}}$ et que les seuils légaux d'information et d'alerte sont respectivement de 180 $\mu\mathrm{g\,m^{-3}}$ et 240 $\mu\mathrm{g\,m^{-3}}$.

On a donc affaire ici à la prévision d'environ 30 000 pics, mais seuls 27 500 d'entre eux ont été effectivement mesurés au cours de la période étudiée (les valeurs manquantes étant à attribuer, notamment, à des pannes de mesure ponctuelles en certaines stations). On notera dans la suite $\mathscr{N}_t$ l'ensemble des stations actives au jour t, de sorte que pour t fixé, seules les observations y_t^s avec $s \in \mathscr{N}_t$ sont disponibles.

D'autres jeux de données sont étudiés dans [37, 39] : l'un européen et l'autre français (fondé sur la base de données BDQA gérée par l'ADEME et regroupant des observations effectuées par une quarantaine d'associations agréées pour la surveillance de la qualité de l'air). Par ailleurs, on y étudie également la prévision horaire des concentrations, pour les trois jeux de données décrits.

 ISSN: 2102-6238

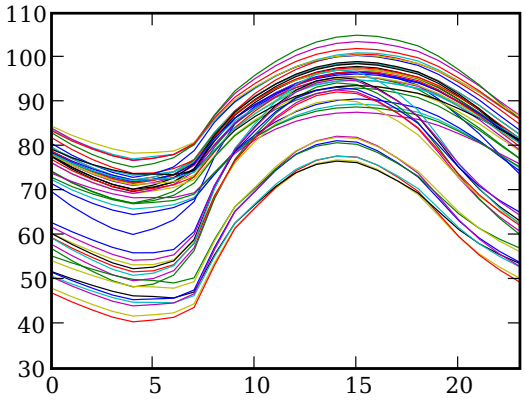

FIGURE 4. *Profils de prévisions de concentration ($\mu g\, m^{-3}$) en ozone proposées par les 48 experts : moyennes des prévisions horaires en espace et en les jours de prévisions. Abscisses : heures de la journée ; ordonnées : concentrations.*

4.1.2. Construction des experts utilisés

On dispose ici de $N = 48$ experts, construits précédemment dans [38] et intégrés dans la plateforme de prévision Polyphemus [1]. En fait, chaque expert est le résultat de trois choix : un modèle de diffusion physico-chimique des polluants atmosphériques ; un jeu de données d'entrée (notamment, des données météorologiques et d'émission de polluants) ; un schéma de résolution numérique des équations aux dérivées partielles en jeu (choix d'une discrétisation spatiale et temporelle par exemple). Les choix considérés et leur combinaison pour former les 48 experts sont expliqués en détails dans [38, paragraphe 2.2]. La figure 4 montre que les prévisions des experts sont fortement dispersées : bien que l'on prenne la moyenne de leurs prévisions horaires en temps (en les jours de la période de prévision) et en espace, il y a un écart d'un facteur multiplicatif 2 entre les experts proposant les prévisions de concentration les plus fortes et les plus faibles. On notera que la forme des courbes correspond au profil typique de la concentration d'ozone au cours de la journée (creux de concentration à la fin de la nuit, pic en fin d'après-midi) mais qu'en aucun cas les experts ne sont donnés par des translations d'un expert de référence ; c'est uniquement la moyenne en temps et en espace qui concourt à produire ces profils similaires. Comme on le verra même par la suite, les experts ont des comportements et performances variables en temps et en espace.

Dans cet article, on se contente de décrire les performances des experts et on les considère essentiellement comme des boîtes noires prédictives, dont il s'agit d'améliorer la qualité de prévision de manière automatique. Les experts sont indexés par $j \in \{1,\dots,48\}$ et proposent chacun, pour chaque jour t et chaque station s, une prévision de pic notée $f^s_{j,t}$.

En réalité, ils proposent même un champ de prévisions sur tout l'Europe, *id est*, une prévision pour chaque point d'une grille fine de l'espace européen.

[1] Voir `http://cerea.enpc.fr/polyphemus/`

 ISSN: 2102-6238

4.1.3. Agrégation uniforme en espace mais variable en temps

On se restreint ici aux stratégies d'agrégation proposant le même vecteur de mélange convexe $\mathbf{p}_t$ ou linéaire $\mathbf{u}_t$ des prévisions des experts en tous les sites ; c'est-à-dire que ce vecteur dépend de t uniquement, mais pas de s. C'est une contrainte que l'on peut lever afin de gagner en performance (voir [39, paragraphe A.1]) mais qui a l'avantage de faire gagner en interprétabilité et en force de prévision : en effet, les experts proposant chacun un champ de prévisions, on obtient alors un champ agrégé de prévisions, ce qui permet de proposer des prévisions même en dehors des stations (même si l'évaluation de la qualité de la prévision ne peut, elle, avoir lieu qu'en ces stations).

Avant de continuer, il nous faut définir la fonction de perte utilisée pour effectuer cette évaluation de la qualité de la prévision, et à cet effet, préciser au préalable les ensembles d'observations $\mathscr{Y}$ et de prévisions $\mathscr{X}$. Les observations en chaque station sont situées dans l'intervalle $[0,300]\cup\{\perp\}$, où le symbole $\perp$ désigne une absence d'observation correspondant à une station en panne et où la valeur 300 est prise pour fixer les idées : c'est une borne sur la concentration maximale d'ozone. De même, les prévisions individuelles en un jour donné et un site fixé sont supposées être dans l'intervalle $[0,300]$. Les stations étant indexées par l'ensemble $\mathscr{N}$, on a donc les ensembles d'observations et de prévisions :

$$\mathscr{Y} = \big([0,300]\cup\{\perp\}\big)^{\mathscr{N}} \qquad \text{et} \qquad \mathscr{X} = [0,300]^{\mathscr{N}}\,.$$

On retient une perte quadratique : la fonction de perte $\ell : \mathscr{X}\times\mathscr{Y}\to\mathbb{R}$ est égale, pour tout couple $\mathbf{y} = (y_s)_{s\in\mathscr{N}}$ et $\mathbf{x} = (x_s)_{s\in\mathscr{N}}$ d'éléments de $\mathscr{Y}$ et $\mathscr{X}$, à

$$\ell(\mathbf{x},\mathbf{y}) = \sum_{s:y_s\neq\perp} (x_s - y_s)^2\,.$$

Cette fonction est bien bornée et convexe en son premier argument ; elle vérifie également l'hypothèse 2.1 (elle est différentiable en tout point, de sorte qu'un gradient existe en tout point). On peut donc dans la suite instancier toutes les stratégies du paragraphe 2.

La perte au jour t d'une stratégie d'agrégation proposant le vecteur de mélange (convexe ou linéaire) $\mathbf{v}_t = (v_{1,t},\dots,v_{N,t})$ des prévisions des experts est alors égale, avec les notations précédentes, à une quantité que l'on note, pour simplifier, par $\ell_t(\mathbf{v}_t)$:

$$\ell_t(\mathbf{v}_t) = \sum_{s\in\mathscr{N}_t} \left(y_t^s - \sum_{j=1}^N v_{j,t} f_{j,t}^s \right)^2 . \tag{4.1}$$

On cache donc dans la notation ℓ_t à la fois les prévisions des experts $f_{j,t}^s$ et les observations y_t^s et on n'explicite que la dépendance en le vecteur de mélange $\mathbf{v}_t$ uniforme en l'espace proposé par la stratégie de prévision.

On s'intéresse à l'erreur quadratique moyenne (EQM) des experts et des stratégies de référence. Dans cette partie, cette dernière est calculée non pas sur toute la période de prévision mais uniquement sur ses 96 derniers jours ; cela laisse 30 jours aux différentes stratégies comme période d'apprentissage sans évaluation. C'est notamment utile pour les stratégies de régression séquentielle du paragraphe 3.1, qui proposent le vecteur de mélange linéaire nul $\mathbf{u}_1 = (0,\dots,0)$

 ISSN: 2102-6238

pour le premier jour de prévision et commettent donc une erreur importante ; elles mettent quelques jours à passer dans un régime plus satisfaisant. On note $\{t_0,\dots,T\} = \{31,\dots,126\}$ les indices des jours où l'évaluation aura donc lieu.

Formellement, étant donné les vecteurs de mélange (linéaires ou convexes) $\mathbf{v}_{t_0},\dots,\mathbf{v}_T$ choisis par une stratégie d'agrégation $\mathscr{S}$ sur les $T+1-t_0 = 96$ derniers jours de prévision, l'erreur quadratique moyenne de $\mathscr{S}$ est définie comme

$$\mathrm{EQM}(\mathscr{S}) = \sqrt{\frac{1}{\sum_{t=t_0}^{T} |\mathscr{N}_t|} \sum_{t=t_0}^{T} \ell_t(\mathbf{v}_t)}$$

(où $|\mathscr{N}_t|$ désigne le cardinal de $\mathscr{N}_t$). On passe de la perte cumulée à l'erreur quadratique par renormalisation et racine carrée, et vice versa par les inverses de ces transformations. Ainsi, garantir qu'une stratégie a un regret faible, c'est garantir que son erreur quadratique moyenne est proche, par exemple, de celle du meilleur expert ou de la meilleure combinaison convexe constante des prévisions des experts. Mais dans la suite, nous reporterons les résultats uniquement en termes d'EQM plutôt que de regret, puisque c'est cette première qui est le critère de performance le plus couramment utilisé en pratique.

4.1.4. Performances des experts considérés et de certaines stratégies d'agrégation de référence

Le diagramme en bâtons de la figure 5 montre les erreurs quadratiques moyennes des experts sur le jeu de données considéré ; elles sont comprises entre 22.43 et 35.79. On pourrait penser qu'un meilleur expert ou groupe d'experts se dégage nettement, mais cela n'est pas le cas. La carte de l'Europe fournie à la figure 5, coloriée en fonction de l'indice de l'expert ayant la plus faible erreur quadratique moyenne sur chaque zone de l'espace, indique qu'il n'y a pas d'expert qui serait uniformément en espace le meilleur et qu'on ne peut parler au mieux que de meilleur expert local ; on note également que de nombreux experts sont meilleur expert local pour une partie de l'espace au moins. Cela illustre que tous les experts sont utiles et apportent de l'information, et qu'en outre, leur comportement et leurs performances sont variables en espace (on expliquera plus bas comment on peut affirmer qu'ils sont également variables en temps).

On définit l'erreur quadratique moyenne d'une suite de vecteurs de mélange linéaires ou convexes $\mathbf{v}_1,\dots,\mathbf{v}_T \in \mathbb{R}^N$, notée $\mathbf{v}_1^T$, comme

$$\mathrm{EQM}\left(\mathbf{v}_1^T\right) = \sqrt{\frac{1}{\sum_{t=t_0}^{T} |\mathscr{N}_t|} \sum_{t=t_0}^{T} \ell_t(\mathbf{v}_t)}$$

(le choix des $t_0 - 1$ premiers vecteurs n'a pas d'importance) ; la différence par rapport à une vraie stratégie de prévision est qu'ici les vecteurs $\mathbf{v}_t$ sont fixés à l'avance et non pas déterminés jour après jour en fonction des performances des experts sur le passé. L'erreur quadratique moyenne de la suite constante de vecteurs $\mathbf{v}$, qui agrège les prévisions des experts selon $\mathbf{v}$ uniformément en espace mais également en temps, est simplement notée $\mathrm{EQM}(\mathbf{v})$. On rappelle par ailleurs que l'on avait désigné par δ_j la masse de Dirac en j, qui correspond à la stratégie de prévision donnée par l'expert j.

 ISSN: 2102-6238

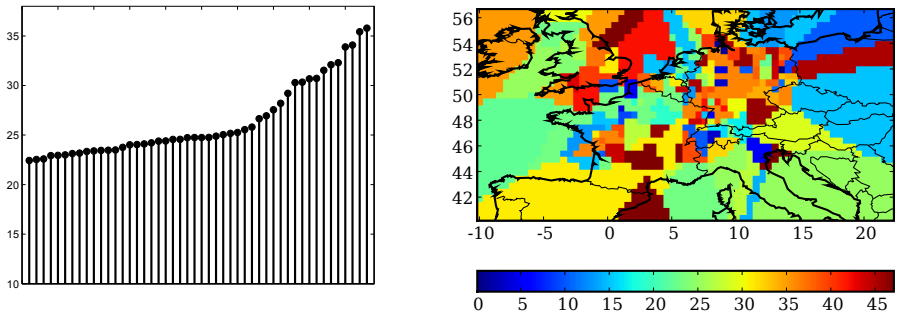

FIGURE 5. *Représentation graphique des performances des experts : erreurs quadratiques moyennes des experts sur les données considérées, classées par ordre croissant (gauche) et coloration de la carte de toute l'Europe en fonction de l'indice du meilleur expert local (droite).*

Nom de la stratégie de référence	Formule	Valeur
Moyenne uniforme	$\mathrm{EQM}\big((1/48,\ldots,1/48)\big)$	$= 24.41$
Meilleur expert	$\min_{j=1,\ldots,48} \mathrm{EQM}(\delta_j)$	$= 22.43$
Meilleure combinaison convexe	$\min_{\mathbf{q}\in\mathscr{P}} \mathrm{EQM}(\mathbf{q})$	$= 21.45$
Meilleure combinaison linéaire	$\min_{\mathbf{u}\in\mathbb{R}^N} \mathrm{EQM}(\mathbf{u})$	$= 19.24$
Stratégie presciente	$\min_{\mathbf{u}_1,\ldots,\mathbf{u}_T\in\mathbb{R}^N} \mathrm{EQM}\big(\mathbf{u}_1^T\big)$	$= 11.99$

TABLE 1. *Performances de quelques stratégies de référence pour le jeu de données de prévision de pics d'ozone.*

On reporte alors au tableau 1 les performances de quelques stratégies de référence : la moyenne uniforme des prévisions des experts (qui est une stratégie facile à mettre en œuvre de manière séquentielle), de même que les oracles suivants. Par oracles, on entend des stratégies impossibles à déterminer à l'avance sans avoir vu toutes les données : le meilleur expert global, la meilleure combinaison convexe des experts, la meilleure combinaison linéaire des experts. Enfin, la stratégie dite presciente, qui n'est contrainte que par l'obligation de choisir chaque jour une combinaison linéaire des experts, indique la borne de performance qu'aucune stratégie de prévision par agrégation ne peut améliorer.

On rappelle que les stratégies décrites au paragraphe 2 garantissaient que leur perte quadratique moyenne n'étaient pas trop éloignée de celles du meilleur expert ou de la meilleure combinaison convexe constante des experts, selon que l'on utilisait les pertes ou les pseudo-pertes définies par considération des sous-gradients. Les stratégies du paragraphe 3.1 se comparaient, elles, à la meilleure combinaison linéaire constante des experts, un objectif dont on voit au tableau 1 qu'il est bien plus ambitieux.

 ISSN: 2102-6238

Valeur de η	5×10^{-7}	5×10^{-6}	2×10^{-5}	10^{-4}	Grille
EQM de $\mathscr{E}_\eta^{\text{grad}}$	22.89	21.70	<u>21.47</u>	22.10	21.77

Valeur de λ	0	100	10^4	10^6	Grille
EQM de $\mathscr{R}_\lambda$	20.79	<u>20.77</u>	21.13	21.80	20.81

TABLE 2. *Performances des stratégies $\mathscr{E}_\eta^{\text{grad}}$ et $\mathscr{R}_\lambda$ pour différentes valeurs de leurs paramètres η et λ, ainsi que celles de la méthode de calibration par grille utilisant ces familles de stratégies comme briques élémentaires. La plus faible* EQM *pour un paramètre constant est soulignée pour chacune des deux stratégies.*

4.2. Performances de différentes stratégies de prévision

On montre dans cette partie que les performances des différents oracles du tableau 1 sont atteintes et même dépassées par les stratégies étudiées dans la partie théorique de cet article – un fait dont on peut se féliciter sachant que cela correspond à l'obtention d'un regret négatif alors que la théorie ne prévoit que la majoration du regret par une quantité pas trop grande. Dans [37, 39], nous avions étudié une vingtaine de stratégies : ici, nous ne reproduisons qu'un bref résumé des bonnes performances obtenues par trois stratégies (et leurs variantes).

4.2.1. Deux premières familles de stratégies : pondération par poids exponentiels des gradients des pertes et régression ridge

On étudie dans ce paragraphe les stratégies $\mathscr{E}_\eta^{\text{grad}}$ du paragraphe 2.3 et $\mathscr{R}_\lambda$ du paragraphe 3.1.1. La première repose sur la considération de pseudo-pertes associées aux gradients des pertes introduites en (4.1) et ne nécessite pas davantage de précisions. Pour la seconde en revanche, on notera que dans sa description au paragraphe 3.1.1, il était supposé qu'à chaque échéance, n'était subie qu'une et une seule perte quadratique avant que le nouveau vecteur de mélange ne soit choisi ; or ici, l'idée naturelle est d'étendre la définition (3.2) à un cas où non pas une mais $|\mathscr{N}_t|$ pertes quadratiques (celles en chaque site actif) sont encourues. Cette extension est effectuée dans [37, chapitre 12] et la forme de la borne du Théorème 4 y est préservée.

Les quatre premières colonnes du tableau 2 présentent les performances de ces deux stratégies pour diverses valeurs constantes des paramètres. La valeur 5×10^{-7} correspond approximativement à la valeur théorique optimale $\eta^\star$ précisée par le Théorème 2, mais elle n'est de loin pas la meilleure valeur en pratique. C'est pourquoi, ainsi qu'expliqué au paragraphe 2.4.2, on recourt à une méthode de calibration par grille. Ici, on utilise une version simplifiée de cette calibration, où la grille est fixée et immuable ; celle pour la famille des stratégies $\mathscr{E}_\eta^{\text{grad}}$, respectivement, $\mathscr{R}_\lambda$, consiste en 11 points logarithmiquement uniformément répartis entre 10^{-8} et 10^{-4}, respectivement, entre 1 et 10^6. Les résultats obtenus par calibration sur cette grille sont précisés dans la dernière colonne du tableau 2.

On rappelle que les valeurs de référence sont 21.45, EQM de la meilleure combinaison convexe constante, et 19.24, EQM de la meilleure combinaison linéaire constante. La famille des stratégies $\mathscr{E}_\eta^{\text{grad}}$ obtient des performances proches de cette première valeur de référence, alors même que

 ISSN: 2102-6238

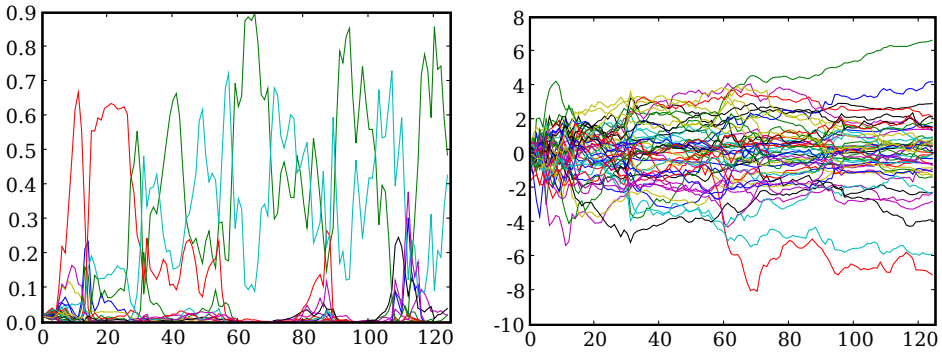

FIGURE 6. *Représentation graphique des vecteurs de mélange convexes choisis par* $\mathscr{E}^{\mathrm{grad}}_{2\times 10^{-5}}$ *(à gauche) et linéaires choisis par* $\mathscr{R}_{100}$ *(à droite), en fonction du temps. L'ensemble des 126 vecteurs est répresenté mais on rappelle que l'évaluation des performances n'est effectuée que sur les 96 derniers jours.*

le nombre d'échéances de prévision est assez faible ici. Quant à la famille des $\mathscr{R}_\lambda$, elle n'arrive certes pas à obtenir des performances suffisamment proches de celle de la meilleure combinaison linéaire constante, mais elle bat assez largement la meilleure combinaison convexe constante. Dans les deux cas, on voit que l'adaptation a un coût assez réduit par rapport aux meilleurs choix rétrospectifs de paramètres, précisés par les valeurs soulignées dans le tableau 2.

On conclut ce paragraphe en notant que les stratégies d'agrégation considérées ne se concentrent pas sur un seul expert et qu'au contraire, les poids attribués aux experts dans les vecteurs de mélange retenus peuvent changer rapidement et de manière significative au cours du temps. Ceci est illustré à la figure 6, où l'on a considéré les paramètres η et λ rétrospectivement optimaux, et est à attribuer au fait que les performances des experts changent au cours du temps (le paragraphe 4.1.4 avait déjà insisté sur le fait qu'elles variaient également en espace). On peut également ajouter au passage que les vecteurs de mélange linéaires retenus par la régression ridge $\mathscr{R}_{100}$ sont très loin d'être des vecteurs de mélange convexes, vu les composantes fortement négatives obtenues. Cependant, la somme des poids de ces vecteurs est souvent proche de 1. Malgré tout, ces vecteurs linéaires sont moins facilement interprétables que les vecteurs de mélange convexes proposés par $\mathscr{E}^{\mathrm{grad}}_{2\times 10^{-5}}$. En outre, si l'on en revient à notre souhait initial d'obtenir un champ de prévisions agrégées, on aura moins de scrupules à effectuer une telle agrégation par mélange convexe plutôt que par mélange linéaire : même si ce dernier donne de meilleures performances en les sites retenus, on ne peut exclure l'existence de prévisions désastreuses (ou même dénuées de sens, par exemple négatives) entre les sites.

4.2.2. Variantes des deux familles de stratégies précédentes : fenêtrage et escompte

On applique ici les variantes présentées au paragraphe 2.5 aux deux familles de stratégies étudiées ci-dessus, en précisant au lecteur que bien sûr, ces variantes uniquement décrites dans le cas des stratégies de pondération par poids exponentiels s'étendent naturellement au cas des régressions séquentielles : avec les notations des paragraphes 2.5 et 3.1.1, la version escomptée de la régression ridge repose sur un facteur de régularisation λ et une suite décroissante (β_t) et choisit, à chaque

 ISSN: 2102-6238

Famille	Originelle	Fenêtrée	Escomptée
$\mathscr{E}_\eta^{\text{grad}}$	21.47	21.37	21.31
$\mathscr{R}_\lambda$	20.77	20.03	19.45

TABLE 3. EQM *de différentes familles de stratégies, calibrées chacune avec le(s) meilleur(s) paramètre(s) rétrospectifs : versions originelles, fenêtrées et escomptées.*

échéance t,

$$\mathbf{u}_t \in \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^N} \left\{ \lambda \|\mathbf{v}\|_2^2 + \sum_{t'=1}^{t-1} (1+\beta_{t-t'}) \sum_{s\in\mathscr{N}_t} \left(y_{t'}^s - \sum_{j=1}^N v_j f_{j,t'}^s \right)^2 \right\}. \tag{4.2}$$

Le tableau 3 reporte les résultats obtenus. On y entend par versions originelles celles qui avaient été décrites au tableau 2 ; on reporte la plus faible EQM qui y avait été obtenue pour un choix contant des paramètres. Les versions fenêtrées sont paramétrées d'une part par une vitesse d'apprentissage η ou un facteur de régularisation λ constants et d'autre part, par une largeur de fenêtre H. Les versions escomptées sont paramétrées par $\gamma > 0$ d'une part et $\eta > 0$ ou $\lambda > 0$ d'autre part, de telle sorte les conditions du Théorème 3 soient remplies ; par exemple, pour la stratégie de pondération par poids exponentiels des gradients, on prend les suites (η_t) et (β_t) de la forme

$$\eta_t = \frac{\eta}{\sqrt{t}} \qquad \text{et} \qquad \beta_t = \frac{\gamma}{t^2}.$$

Pour la régression ridge, on utilise la même forme de suite (β_t) et un paramètre λ constant. On reporte au tableau 3 les EQM obtenues pour les meilleurs choix respectifs de (η, H), (λ, H), (η, γ) et (λ, γ).

On constate que tenir moins compte du passé par fenêtrage ou escompte améliore les performances, comme nous le suggéraient les praticiens. Cela étant, les résultats obtenus par la considération d'escomptes étant meilleurs que ceux par fenêtrage, il ne faut pas pour autant totalement oublier le passé lointain.

4.2.3. Régression ridge escomptée : propriétés de robustesse et de correction automatique du biais

Robustesse On effectue dans [39, paragraphes 4.3.2 et 4.3.3] une étude de robustesse de la meilleure stratégie obtenue jusque-là, la régression ridge escomptée. Cette dernière obtient certes une excellente performance globale, sur l'ensemble des stations et des jours de prévision, puisque son EQM de 19.45 est très proche de celle de la meilleure combinaison linéaire constante, égale à 19.24. Mais on peut cependant se demander si cela ne cacherait pas de mauvaises performances locales çà et là (en temps ou en espace) et/ou si du point de vue de la prévision de dépassements de seuils, les résultats sont également à la hauteur. Les détails de ces réponses, négative à la première question et positive à la seconde, sont omis de cet article de survol pour les données de pics d'ozone ; on les détaillera uniquement pour le jeu de données de la partie suivante, à propos de consommation électrique.

 ISSN: 2102-6238

Expert	EQM originel	EQM après débiaisement
Meilleur	22.43	21.66
De référence	24.01	22.43
Pire	35.79	24.78

TABLE 4. *Réductions d'*EQM *par application du pré-traitement de débiaisement consistant à lancer la régression ridge escomptée optimale du tableau 3 sur l'expert seul.*

Correction automatique du biais En revanche, une propriété tout à fait intéressante et que l'on va développer est le fait que la régression ridge et plus encore, la régression ridge escomptée, peut être utilisée comme un pré-traitement de correction automatique du biais des experts. Cette faculté à corriger le biais était d'ailleurs la motivation à recourir à des vecteurs de mélange linéaires plutôt que convexes et elle devait compenser leur moindre interprétabilité.

Formellement, la propriété est mise en œuvre de la manière suivante. Rien n'empêche de fixer un expert k et de lancer ces stratégies sur lui seul : l'objectif est alors, comme l'indique le Théorème 4, de faire presqu'aussi bien que le meilleur des méta-experts proposant $bf^s_{j,t}$ en chaque site et à chaque échéance, pour un paramètre scalaire positif b représentant un facteur multiplicatif de débiaisement. Par exemple, à chaque échéance, la régression ridge escomptée propose les prévisions $b_t f^s_{k,t}$ en lieu des $f^s_{k,t}$, où b_t est le scalaire de débiaisement

$$b_t \in \operatorname*{arg\,min}_{b\in\mathbb{R}} \left\{ \lambda |b|^2 + \sum_{t'=1}^{t-1} (1+\beta_{t-t'}) \sum_{s\in\mathscr{N}_t} \left(y^s_{t'} - b f^s_{j,t'} \right)^2 \right\} ;$$

évidemment, dans ce cas, b_t sera toujours positif et tendra à être d'autant plus proche de 1 que l'expert k a un biais originel faible.

Le tableau 4 illustre l'intérêt de ce pré-traitement sur trois experts parmi les 48 : le meilleur et le pire expert au sens de leur EQM, ainsi qu'un expert de référence formé par les valeurs les plus courantes des choix du paragraphe 4.1.2 (voir [38, paragraphe 2.2] pour plus de détails). Dans tous les cas, une réduction d'EQM est obtenue.

Une idée que nous avons eue mais n'avons pas encore exploitée serait d'appliquer ce pré-traitement à tous les experts avant d'appliquer des stratégies de prévision par agrégation.

4.2.4. Stratégie de Lasso séquentiel escomptée

Afin de préparer des avancées futures qui consisteraient par exemple en la considération d'un grand nombre d'experts, nous avons voulu voir dans [25] s'il était possible de réaliser agrégation et sélection d'experts simultanément, *id est*, d'agréger les prévisions uniquement d'un sous-ensemble d'experts de cardinal petit. Ce sous-ensemble a bien sûr vocation à changer au cours du temps. Pour cela, nous avons recouru à la stratégie de Lasso séquentiel du paragraphe 3.1.2, et plus précisément, à sa version escomptée, définie par le remplacement de la norme ℓ^2 dans (4.2) par une norme ℓ^1 et par le choix de trois paramètres λ, β, γ vu que cette fois-ci nous nous autorisons à prendre la suite des escomptes (β_t) sous la forme $\beta_t = \gamma t^{-\beta}$. Les paramètres optimaux sont $\lambda = 2\times 10^4$, $\beta = 1.5$ et $\gamma = 150$ et ils conduisent aux performances indiquées au tableau 5, de même qu'à la sélection–agrégation décrite à la figure 7.

 ISSN: 2102-6238

Ridge, escomptée	Lasso, escomptée	Oracle linéaire
19.45	19.31	19.24

TABLE 5. *Comparaison des* EQM *des régressions séquentielles escomptées (calibrées avec les paramètres rétrospectivement optimaux) à l'oracle linéaire du tableau 1.*

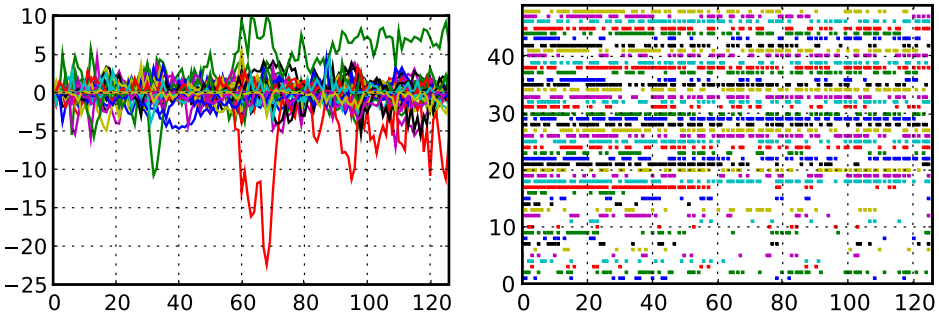

FIGURE 7. *Représentation graphique des vecteurs de mélange linéaires choisis par la version escomptée optimale de la stratégie de Lasso séquentiel (à gauche) et des experts sélectionnés (à droite : un carré plein signifie que l'expert est absent du mélange). L'ensemble des 126 échéances est représenté mais on rappelle que l'évaluation des performances n'est effectuée que sur les 96 derniers jours.*

Plus précisément, on note que typiquement, à chaque échéance une vingtaine de modèles est éliminée et que l'agrégation ne s'effectue plus que sur un sous-ensemble d'une trentaine d'experts. Par ailleurs, cette sélection assez marquée n'obère pas les performances et au contraire, elle les améliore même légèrement ; en particulier, on se rapproche encore de la performance de l'oracle linéaire du tableau 1.

4.3. Perspectives et comparaison avec d'autres approches issues de l'apprentissage

Dans ce dernier paragraphe de la partie consacrée au jeu de données de pics d'ozone, on indique d'une part quelques perspectives de recherche et développement et d'autre part, on discute les différences entre l'approche décrite ici et d'autres méthodes d'apprentissage ayant déjà été utilisées en prévision de la qualité de l'air.

4.3.1. Perspectives

La première perspective serait d'étudier les performances lorsque la période de prévision est plus longue, de l'ordre d'un an, avec ou sans augmentation du nombre d'experts. Des résultats préliminaires nous ont montré que, comme attendu, les gains de performances par rapport au meilleur expert ou à la meilleure combinaison convexe constante des experts sont encore plus importants.

La seconde est de s'intéresser non pas à la valeur même du pic d'ozone attendu mais de déterminer s'il dépassera ou non un des seuils réglementaires de 180 $\mu g\,m^{-3}$ et 240 $\mu g\,m^{-3}$. Nous avons lancé là aussi quelques études préliminaires mais n'avons pu faire mieux pour l'instant que

 ISSN: 2102-6238

la procédure qui consiste à comparer les valeurs prévues par des stratégies d'agrégation aux seuils – ce qui est surprenant malgré tout, s'agissant d'un problème de classification, généralement censé être plus facile parce que plus direct qu'un problème de prévision pure. Une manière intermédiaire de procéder serait par exemple de remplacer la perte quadratique uniforme $\ell(x,y) = (x-y)^2$ par une perte de la forme $\ell(x,y) = \gamma(y)\,(x-y)^2$ où γ est une fonction croissante à déterminer : cela indique que les erreurs de prévisions (à la hausse ou à la baisse) sont d'autant plus graves que les observations y_t sont proches des seuils réglementaires ou supérieures à eux.

4.3.2. Comparaison avec d'autres approches issues de l'apprentissage

On effectue cette discussion à partir de l'article [5], qui illustre l'application de techniques d'apprentissage stochastique non séquentiel (réseaux de neurones, arbres de régression, *bagging*, forêts aléatoires, séparateurs à vaste marge) à la prévision de pics d'ozone, de même que de l'article antérieur [26], qui emploie essentiellement des arbres de régression. Dans les deux cas, l'objectif est de construire un modèle prédictif qui, à partir de données d'entrées, forme une prévision de concentration pour l'échéance suivante. Ces entrées sont diverses ; par exemple, dans [5], il s'agit de différentes prévisions (concentration en ozone mais aussi force du vent, température, concentrations d'autres polluants) formées par le modèle MOCAGE de Météo-France. Dans [26], il s'agit de 41 variables explicatives et notamment, des concentrations d'ozone de la veille et de la nuit, ainsi que de divers facteurs climatiques mesurés au cours du début de la journée où il s'agit d'effectuer la prévision. Ainsi, ces résultats sont peut-être à voir, dans notre cadre, comme des manières possibles de construire des experts à partir de données d'entrée. Il est à noter que la plupart de ces méthodes nécessitent le réglage de paramètres (par exemple, la profondeur des arbres de régression) ; une manière de moins se préoccuper de ces réglages serait de considérer divers jeux de paramètres pour construire plusieurs experts reposant sur la même méthode sous-jacente et d'ensuite agréger séquentiellement leurs prévisions, ainsi que nous l'expliquons dans cet article.

Pour résumer, au moins d'un point de vue *théorique*, les techniques d'apprentissage stochastique non séquentiel sont une manière efficace de construire des experts, que les techniques d'apprentissage séquentiel présentées dans cet article agrègent alors. On notera que si la construction des experts peut utiliser avec profit des méthodes stochastiques, l'agrégation, au moins telle que nous la mettons en œuvre ici, est robuste parce que ne reposant sur aucune hypothèse stochastique.

Cependant, d'un point de vue *pratique*, l'absence de garanties théoriques formelles ne devrait pas nous empêcher de considérer une mise en œuvre séquentielle de techniques d'apprentissage stochastique non séquentiel (et notamment, les réseaux de neurones, les forêts aléatoires d'arbres de régression, ou encore, le mélange de modèles bayésien [41]) : à chaque échéance $t \geqslant 2$, on lance la technique stochastique sur les $t-1$ séries de données disponibles et on récupère une prévision agrégée. Nous n'avons pas encore étudié les performances de prévision pouvant être obtenues ainsi, essentiellement par manque de temps – mais ce sera sera l'objet de travaux futurs. Pour l'heure, le présent article est essentiellement une illustration des bonnes performances des techniques d'agrégation séquentielle robuste sur deux jeux de données, mais ne propose pas de comparaison pratique raisonnée aux approches stochastiques populaires mentionnées ci-dessus.

 ISSN: 2102-6238

5. Application à la prévision de consommation électrique

Dans cette partie, on s'intéresse à la prévision demi-horaire de la consommation électrique des clients d'EDF, le fournisseur historique français. Les résultats sont tirés de l'article soumis [18] ainsi que du rapport technique correspondant [17], qui étudient également la prévision de consommation des clients de la filiale slovaque d'EDF.

La principale difficulté ici (mais qui est également une chance), c'est que l'on a affaire à des experts spécialisés à certains contextes et qui ne fournissent par conséquent que des prévisions par intermittence. Par exemple, certains experts peuvent être construits pour fournir de bonnes prévisions en hiver et sont inactifs en été ; il peut également y avoir des experts de jours travaillés ou des experts de fin de semaine. C'est une chance parce que ces experts spécialisés sont susceptibles d'être bien plus précis ; c'est une difficulté à première vue parce qu'il faut adapter les définitions et résultats de la partie 2, ce qui requiert un peu de travail.

On commence par expliquer brièvement comment traiter mathématiquement ce nouveau cadre, avant de présenter le jeu de données et les experts retenus, puis de décrire les performances obtenues par les stratégies de prévision par agrégation considérées.

5.1. Comment tirer parti d'experts spécialisés

A notre connaissance, ce cadre a été peu traité dans le domaine de la prévision séquentielle par agrégation de prédicteurs fondamentaux et nous ne pouvons citer comme références que l'article fondateur [22], de même que deux autres articles principalement consacrés à d'autres sujets mais mentionnant des résultats pour les experts spécialisés en passant, à savoir [7, paragraphes 6–8] et [11, paragraphe 6.2]). Tous ont pour objet l'agrégation par vecteurs de mélange convexes ; il ne semble pas exister à ce jour de formulation et de résultats théoriques pour celle par vecteurs linéaires. Des tentatives en ce sens ont été considérées dans [17] mais elles ne sont pas satisfaisantes en l'état et nécessitent au moins un sérieux travail d'approfondissement.

Formellement (et en revenant aux notations de la partie 2), l'ensemble des prévisions $\mathscr{X}$ est étendu pour contenir le point $\perp$, qui a la signification suivante. Lorsqu'à l'échéance t, l'expert $j \in \{1,\dots,N\}$ propose $f_{j,t} = \perp$, c'est que les conditions externes liées à sa spécialisation ne sont pas remplies et qu'il s'abstient de former une prévision. On dira qu'il est inactif. Par opposition, les experts proposant comme prévision un élément de $\mathscr{X}$ sont dits actifs.

On suppose qu'à chaque échéance t, au moins un expert est actif et on notera E_t l'ensemble non vide de ces experts actifs. Comme indiqué plus haut, on se restreint au choix de vecteurs de mélange convexes. En particulier, une stratégie de prévision $\mathscr{S}$ choisit à l'échéance t un vecteur de mélange $\mathbf{p}_t$ de support inclus dans E_t et forme la prévision

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

On considère ici encore la perte quadratique : on définit alors la perte cumulée et l'EQM d'une stratégie $\mathscr{S}$ sur les T premières échéances de la même manière que précédemment, c'est-à-dire selon

$$\widehat{L}_T(\mathscr{S}) = \sum_{t=1}^{T} \big(\widehat{y}_t - y_t\big)^2 \qquad \text{et} \qquad \mathrm{EQM}(\mathscr{S}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \big(\widehat{y}_t - y_t\big)^2}\,.$$

 ISSN: 2102-6238

Il est à noter que dans cette partie, contrairement à la précédente, l'évaluation se fera bien sur toute la période de prévision (sans période d'entraînement, contrairement au cas de la prévision agrégée de pics d'ozone) ; on verra que c'est le cas essentiellement parce que cette période de prévision a une longueur T grande cette fois-ci.

5.1.1. Comparaison au meilleur expert ou à la meilleure combinaison convexe constante des experts

Les choses se compliquent au moment de définir les quantités correspondantes pour les experts et leurs combinaisons convexes constantes. La perte cumulée d'un expert j donné a peu de sens, mais son EQM est facile à définir :

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

Il est également facile de définir une notion de regret, qui sera cette fois-ci fortement dépendante de l'expert j auquel on compare la stratégie $\mathscr{S}$ et c'est pourquoi on indexera ce regret par T, $\mathscr{S}$ mais aussi j ; en fait, pour que la comparaison entre j et $\mathscr{S}$ soit honnête, on ne l'effectue que sur les échéances où j était actif :

$$R_T(\mathscr{S}, j) = \sum_{t\leqslant T : j\in E_t} \left(\left(\widehat{y}_t - y_t\right)^2 - (f_{j,t}-y_t)^2 \right).$$

La dernière difficulté est maintenant d'étendre ces définitions au cas non pas d'un expert mais d'une combinaison convexe constante de ces experts, donnée par le vecteur de mélange convexe $\mathbf{q}$, et de les étendre de telle manière à ce que lorsque $\mathbf{q} = \delta_j$, la masse de Dirac en j, on retrouve exactement les définitions précédentes. A cet effet, on introduit la renormalisation $\mathbf{q}^E$ de $\mathbf{q}$ à un sous-ensemble E de $\{1,\dots,N\}$ en définissant d'abord le poids de E sous $\mathbf{q}$,

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

puis

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

On propose alors les extensions

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

et

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

 ISSN: 2102-6238

[18, paragraphe 2.3] effectue un survol de la littérature montrant qu'il est possible d'assurer que le regret face aux vecteurs de mélange convexes $\mathbf{q}$ soit uniformément borné par une quantité de l'ordre de $\sqrt{T}$, où l'uniformité porte sur $\mathbf{q}$ mais aussi sur les suites d'observations y_t et de prévisions des experts $f_{j,t}$. Les stratégies assurant cela sont obtenues par des extensions des stratégies des paragraphes 2.2 et 2.3. Par souci de concision, nous ne détaillons pas davantage cela et nous contentons de noter $\mathscr{W}_\eta$ et $\mathscr{W}_\eta^{\text{grad}}$ ces stratégies.

5.1.2. Comparaison à des experts composés

L'extension des définitions du paragraphe 3.2 est plus aisée : il suffit de ne retenir désormais comme classe d'experts composés non pas la classe identifiée à tout $\{1,\dots,N\}^T$ mais celle identifiée à $\mathscr{C}'_T = E_1 \times \dots \times E_T$ (on notera le symbole prime supplémentaire). La définition du nombre de ruptures ne changeant pas, on considère ensuite les sous-ensembles $\mathscr{C}'_{T,m}$ d'experts composés contenant au plus m ruptures. Bien évidemment, pour m petit, ces sous-ensembles peuvent être vides ici. Les notions de regret par rapport à un expert composé j_1^T ou d'EQM d'un tel expert composé sont claires.

Là encore, il est facile de modifier les stratégies du paragraphe 3.2 afin de les adapter à ce nouveau cadre et en leur permettant d'être compétitives face à tous les experts composés admettant une borne fixée sur leur nombre de ruptures ; les détails de l'adaptation sont omis et on renvoie le lecteur à [18, paragraphe 2.3], tout en précisant cependant que l'idée de cette extension est nouvelle, mais très naturelle. On obtient ainsi deux familles de stratégies, notées $\mathscr{G}_{\eta,\alpha}$ et $\mathscr{G}_{\eta,\alpha}^{\text{grad}}$.

5.2. Description du jeu de données et des experts utilisés

5.2.1. Jeu de données

On utilise un jeu de données couramment employé à EDF pour la calibration des modèles de prévision à court terme de la consommation électrique des Français (particuliers et industries). Il est décrit en détails dans [19] et nous ne résumons qu'à grands traits ses caractéristiques.

Il est constitué d'une part, de données de consommation à pas demi-horaire, et d'autre part, de prévisions météorologiques (température et couverture nuageuse) effectuées sur l'ensemble du territoire français. Les données de consommation sont construites par EDF à partir des mesures procurées par une entité filiale responsable de la distribution d'électricité, RTE, tandis que les prévisions météorologiques sont fournies par Météo-France.

Ce jeu de données est divisé en deux parties, la première couvrant la période entre le 1er septembre 2002 et le 31 août 2007 (l'ensemble d'apprentissage) et la seconde, celle entre le 1er septembre 2007 et le 31 août 2008 (l'ensemble de validation). Les experts construits ci-dessous le seront sur l'ensemble d'apprentissage, après quoi ils procureront des prévisions tout au long de la période correspond à l'ensemble de validation. En réalité, pour être tout à fait précis, nous excluons certains jours particuliers de l'ensemble de validation : des 366 jours que couvre sa période nous n'en conservons finalement que 320. Ces jours particuliers sont les jours fériés, ainsi que les jours situés immédiatement avant ou après eux ; les deux jours de changement d'heure ; et les vacances de Noël (du 21 décembre 2007 au 4 janvier 2008). Nous conservons en revanche les grandes vacances (et notamment, août 2008) parce que cette période étant suffisamment longue, il

 ISSN: 2102-6238

Echéances	Toutes les 30 minutes
Nombre de jours D	320
Nombre d'échéances T	15 360
Nombre d'experts N	24 ($= 15 + 8 + 1$)
Médiane des y_t	56.33
Maximum B des y_t	92.76

TABLE 6. *Quelques caractéristiques des consommations électriques y_t sur l'ensemble de validation.*

est possible de construire des experts dédiés à elle. D'autres jours particuliers auraient pu devoir être exclus : ceux qui correspondent à des changements tarifaires très ponctuels destinés à inciter à la réduction temporaire de consommation électrique lors de pics hivernaux de consommation consécutifs à des températures très basses. Ils ne l'ont pas été lorsqu'un pré-traitement fondé sur des données commerciales a pu être mis en œuvre.

Les caractéristiques des consommations y_t sur l'ensemble de validation sont précisées au tableau 6. Dans cette partie également, nous omettons l'unité (GW, gigawatt) des observations et des prévisions de la consommation, de même que celle de leur EQM correspondante.

5.2.2. Construction de trois familles d'experts

Les trois familles d'experts que nous avons construites sont issues des trois grandes familles de modèles statistiques : paramétriques, semi-paramétriques et non-paramétriques ; le but recherché en faisant varier les méthodes de construction est d'obtenir des experts aussi dissemblables que possible. Nous décrivons ci-dessous les traits principaux de leur création et adressons le lecteur à [17, paragraphe 4.1] pour plus de détails.

Le modèle paramétrique utilisé pour engendrer le premier groupe d'experts est décrit dans [8] et est mis en œuvre dans le logiciel de prévision d'EDF appelé «Eventail». On se contente de mentionner que ce modèle est fondé sur une approche du type régression non linéaire consistant à décomposer la consommation électrique en deux termes, un terme principal incluant les variations saisonnières de cette consommation auquel s'ajoute un terme dépendant des conditions météorologiques. A cette régression non linéaire est ajouté un terme de correction auto-régressif des erreurs de prévision à court terme sur les sept derniers jours. C'est en faisant varier les différents paramètres nécessaires (le gradient de température, les coefficients d'auto-régression) que nous avons défini 15 experts, que nous appellerons les experts Eventail.

Le second groupe d'experts a été engendré par un modèle additif généralisé (abrégé en MAG dans la suite) mis en œuvre sous `R` par la boîte à outils `mgcv` créée par [47]. Sa déclinaison dans notre contexte a été réalisée par [40] et importe les idées de la modélisation paramétrique présentée ci-dessous dans un contexte semi-paramétrique. Un des avantages-clés de ce modèle est sa faculté à s'adapter automatiquement aux changements d'habitudes de consommation des clients alors que les modèles paramétriques comme Eventail ont besoin d'informations externes pour y parvenir. Ici encore, c'est en faisant varier les différents paramètres du modèle additif généralisé que nous avons créé 8 experts, les experts dits MAG.

 ISSN: 2102-6238

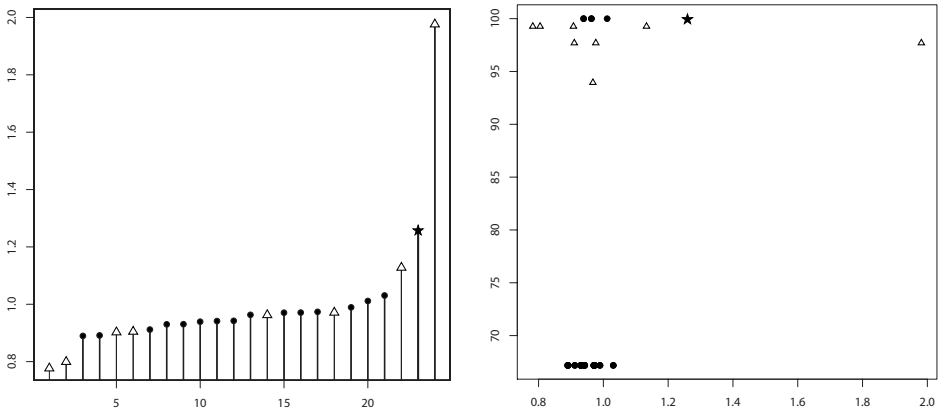

FIGURE 8. *Représentations graphiques des performances des experts :* EQM *classées par ordre croissant (à gauche) et couples* EQM*–fréquence d'activité (à droite) ; la correspondance entre les experts et les symboles est la suivante : experts Eventail (•), experts MAG (△) et expert fonctionnel de similitude (⋆).*

Enfin, nous avons considéré un dernier expert, issu d'une modélisation non-paramétrique proposée par [2] et [1]. Cette dernière consiste à voir la consommation électrique au cours de la journée comme la réalisation d'une certaine courbe stochastique sous-jacente, dont on dispose chaque jour d'une discrétisation de pas demi-horaire. Il s'agit alors d'estimer la distribution de la courbe sous-jacente, ce que l'on fait en exploitant l'historique des réalisations en affectant à chacune de ses journées un poids dépendant de leur similitude à la journée pour laquelle on veut effectuer une prévision. L'expert ainsi obtenu sera appelé l'expert fonctionnel de similitude.

Les performances des experts présentés ci-dessus sont résumées à la figure 8. Le diagramme en bâtons montre leurs EQM, classées dans l'ordre croissant, tandis que le diagramme bi-dimensionnel relie ces EQM aux fréquences d'activité, *id est*, représente les couples

$$\left(\mathrm{EQM}(j),\ \frac{\sum_{t=1}^{T} \mathbb{I}_{\{j \in E_t\}}}{T} \right)$$

pour tous les experts j.

On voit que trois experts Eventail sur les quinze sont actifs sur toute la période ; ils correspondent au modèle de prévision opérationnelle et à deux variations sur les paramètres de correction auto-régressive à court terme. Les douze autres experts Eventail sont inactifs durant l'été : en effet, leurs prévisions sont redondantes avec le modèle opérationnel car ils sont obtenus par variations d'un paramètre (le gradient de température) qui n'est lié qu'aux prévisions de la période hivernale. Les experts MAG sont actifs la plupart du temps, sauf à certaines périodes (par exemple, dans les semaines avec des jours fériés) où par expérience, les services d'EDF R&D savent qu'ils seront peu précis ; l'importance de ces périodes dépend par ailleurs des paramètres précis, c'est pour cela que tous ces experts MAG n'ont pas le même taux d'activité. Enfin, l'expert de similarité fonctionnelle est actif en permanence.

 ISSN: 2102-6238

Nom de la stratégie de référence	Formule	Valeur
Stratégie d'agrégation uniforme	$\mathrm{EQM}(\mathscr{U})$	$=0.724$
Combinaison convexe uniforme	$\mathrm{EQM}\big((1/24,\ldots,1/24)\big)$	$=0.748$
Meilleur expert	$\min_{j=1,\ldots,24} \mathrm{EQM}(j)$	$=0.782$
Meilleur combinaison convexe	$\min_{\mathbf{q}\in\mathscr{P}} \mathrm{EQM}(\mathbf{q})$	$=0.683$
Meilleur expert composé		
Au plus $m=50$ ruptures	$\min_{j_1^T\in\mathscr{C}'_{T,50}} \mathrm{EQM}\big(j_1^T\big)$	$=0.534$
Au plus $m=100$ ruptures	$\min_{j_1^T\in\mathscr{C}'_{T,100}} \mathrm{EQM}\big(j_1^T\big)$	$=0.474$
Stratégie presciente	$\min_{j_1^T\in E_1\times E_2\times\ldots\times E_T} \mathrm{EQM}\big(j_1^T\big)$	$=0.223$

TABLE 7. *Définition et performances de différentes stratégies de référence pour le jeu de données de consommation électrique.*

5.2.3. Ajout d'une contrainte opérationnelle

On requiert que les stratégies de prévision proposent chaque jour à midi des prévisions pour les 24 prochaines heures, c'est-à-dire, pour les 48 prochaines échéances demi-horaires. On suppose à cet effet que les prévisions des experts sont elles aussi disponibles pour toutes ces échéances futures. En revanche, on n'impose pas de contrainte d'égalité à une valeur commune pour les 48 vecteurs de mélanges convexes qu'une stratégie doit ainsi proposer.

5.3. Performances des experts et des stratégies de prévision considérés

On suivra ici la méthodologie développée dans le cadre de la prévision de pics d'ozone ; à savoir : premièrement, le calcul des performances de certains oracles et stratégies de référence, afin d'avoir une idée de ce que l'on attend d'une stratégie d'agrégation de prévisions ; ensuite, la tabulation des performances de ces dernières pour des choix constants des paramètres ; enfin, la mise en œuvre de la règle de calibration séquentielle des paramètres sur une grille.

5.3.1. Performances de certaines stratégies de référence

Le tableau 7 révèle que les experts construits sont très bons, au vu des ordres de grandeur typiques des y_t précisés au tableau 6. Quelques stratégies qui y sont introduites appellent des commentaires.

Premièrement, on entend ici par stratégie presciente la meilleure des stratégies qui peuvent avoir connaissance à l'avance de la suite des consommations $y_1,\ldots,y_T$ et ne sont contraintes

 ISSN: 2102-6238

que par l'obligation de choisir à chaque échéance, une combinaison convexe des prévisions des experts actifs ; la telle meilleure stratégie est naturellement donnée par des combinaisons convexes égales à des masses de Dirac, de sorte qu'on peut l'interpréter comme un expert composé avec au plus $m = T - 1 = 15\,359$ ruptures.

Deuxièmement, dans le cadre des experts spécialisés, il y a une différence subtile entre l'utilisation de la combinaison convexe uniforme $\mathbf{q} = (1/35, \ldots, 1/35)$ et la stratégie d'agrégation uniforme $\mathscr{U}$. Cette dernière est en effet définie comme employant, à chaque échéance, le vecteur de mélange convexe donné par la loi uniforme sur l'ensemble E_t des experts actifs, de sorte que

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

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

Ainsi, pour l'évaluation des performances de la combinaison convexe uniforme, les pertes associées à des échéances pour lesquelles de nombreux experts sont actifs comptent davantage que celles pour lesquelles peu d'experts sont actifs.

Ici, la combinaison convexe uniforme a une plus faible EQM que la stratégie $\mathscr{U}$, ce qui indique que les experts ont tendance à être davantage actifs dans les situations de prévision difficile. C'est un avantage dont sauront tirer parti les stratégies d'agrégation même si, pour l'instant, cela se traduit par le fait que le meilleur expert a une EQM plus grande que celui de la stratégie naïve $\mathscr{U}$. On retient donc du tableau 7 qu'il serait bon que les performances de nos stratégies d'agrégation sophistiquées dépassent de loin celles de la stratégie $\mathscr{U}$; les performances de la meilleure combinaison convexe constante sont déjà un peu meilleures mais les EQM obtenues pour les experts composés montrent qu'effectivement, un fort gain de performance est souhaité par rapport à $\mathscr{U}$.

La fin de cet article va montrer que c'est bien le cas pour la stratégie de pondération par poids exponentiels $\mathscr{W}_\eta^{\mathrm{grad}}$, de même que pour celles par redistribution des poids $\mathscr{G}_{\eta,\alpha}$ et $\mathscr{G}_{\eta,\alpha}^{\mathrm{grad}}$, auxquelles on a fait allusion au paragraphe 5.1.

5.3.2. Performances et robustesse des stratégies d'agrégation étudiées

Pour la tabulation des performances des familles de stratégies $\mathscr{W}_\eta$ et $\mathscr{W}_\eta^{\mathrm{grad}}$, nous avons recouru à la grille de 19 paramètres η donnée par les points logarithmiquement uniformément répartis

$$\widetilde{\Lambda}_{\mathscr{W}} = \big\{m \times 10^{-k},\ \text{pour } k \in \{1, \ldots, 6\} \text{ et } m \in \{1, 2.5, 5\}\big\} \cup \{1\},$$

grille que nous avons utilisé également pour le calcul de la méta-stratégie calibrée décrite au paragraphe 2.4.2 et construite sur ces deux familles. Pour les familles $\mathscr{G}_{\eta,\alpha}$ et $\mathscr{G}_{\eta,\alpha}^{\mathrm{grad}}$, il s'est agi respectivement de la grille

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

 ISSN: 2102-6238

		Meilleur paramètre fixé	Calibration sur la grille
EQM des	$\mathscr{W}_\eta$	0.718	0.723
	$\mathscr{W}_\eta^{\mathrm{grad}}$	0.650	0.654
	$\mathscr{G}_{\eta,\alpha}$	0.632	0.644
	$\mathscr{G}_{\eta,\alpha}^{\mathrm{grad}}$	0.598	0.599

TABLE 8. EQM *de quatre familles de stratégies de prévision par agrégation pour les données de consommation électrique française, calculées sur les grilles indiquées au paragraphe 5.3.2 : pour le meilleur paramètre fixé (colonne de gauche) et pour l'adaptation sur la grille (droite).*

(En fait ces grilles, ici encore fixées à l'avance, sont également les grilles construites par la méthode de construction totalement adaptative du paragraphe 2.4.2 au bout de quelques échéances.)

Les performances obtenues sont résumées au tableau 8, où l'on ne reporte pour chaque famille que les EQM correspondant au meilleur choix constant d'un point des grilles ou à l'adaptation sur les grilles. On rappelle que le tableau 7 avait indiqué que les valeurs de référence étaient 0.724, l'EQM de la stratégie d'agrégation uniforme, qu'il s'agissait de battre à plate couture, et 0.683, l'EQM de la meilleure combinaison convexe constante. Le contrat est rempli pour toutes les familles de stratégies sauf $\mathscr{W}_\eta$. Ici encore, on observe que les versions fondées sur les gradients sont plus efficaces en pratique que les versions initiales, ce qui était attendu au vu des résultats généraux du paragraphe 2.3. En fait, nous avons été agréablement surpris par les performances de la famille $\mathscr{G}_{\eta,\alpha}^{\mathrm{grad}}$, et même un peu intrigués, au vu des remarques de robustesse énoncées ci-dessous.

Cette étude de robustesse, que nous décrivons ici pour les données de consommation électrique alors que nous l'avions omise au paragraphe 4.2.3 pour celles de pics d'ozone, consiste à comparer les performances des stratégies par agrégation à celles du meilleur expert ou de la meilleure combinaison convexe constante des experts, non pas de manière globale comme c'est le cas pour le critère d'EQM mais de manière plus locale. A cet effet, nous avons d'une part découpé le jeu de données en les 48 sous-jeux correspondant à chaque demi-heure ; d'autre part, pour chacune de ces demi-heures, nous reportons non seulement l'EQM encourue mais aussi une idée de la dispersion des valeurs absolues des résidus de prévision.

Ces dernières sont définies comme $\left|\widehat{y}_t - y_t\right|$, où y_t mesure la consommation réelle à l'échéance t et $\widehat{y}_t$ la prévision qui en avait été faite. L'étude des quantiles des résidus permet de déterminer si les bonnes performances globales des stratégies de prévision par agrégation viennent ou non au prix de quelques ratés spectaculaires ; en particulier, ce sont les queues de distributions empiriques (quantiles à 75 % ou à 90 %) qui nous intéressent le plus. La figure 9 étudie les méta-stratégies calibrées formées respectivement à partir des familles $\mathscr{W}_\eta^{\mathrm{grad}}$ et $\mathscr{G}_{\eta,\alpha}^{\mathrm{grad}}$. Les performances de la première collent exactement à celles de la meilleure combinaison convexe constante des experts ou les améliorent un peu, tant au niveau des EQM que des quantiles demi-horaires. Le comportement de la seconde méta-stratégie peut intriguer : les performances sont nettement améliorées dans la période entre 12 heures et 21 heures mais peut-être un peu dégradées entre 6 heures et 12 heures par rapport à celles de la meilleure combinaison convexe constante – la période d'amélioration compensant de loin celle de légère dégradation. C'est pour nous une question ouverte que de mieux comprendre ce comportement et de tirer mieux parti des excellentes performances sur la fenêtre située juste après la mise à jour des poids effectuée à midi.

 ISSN: 2102-6238

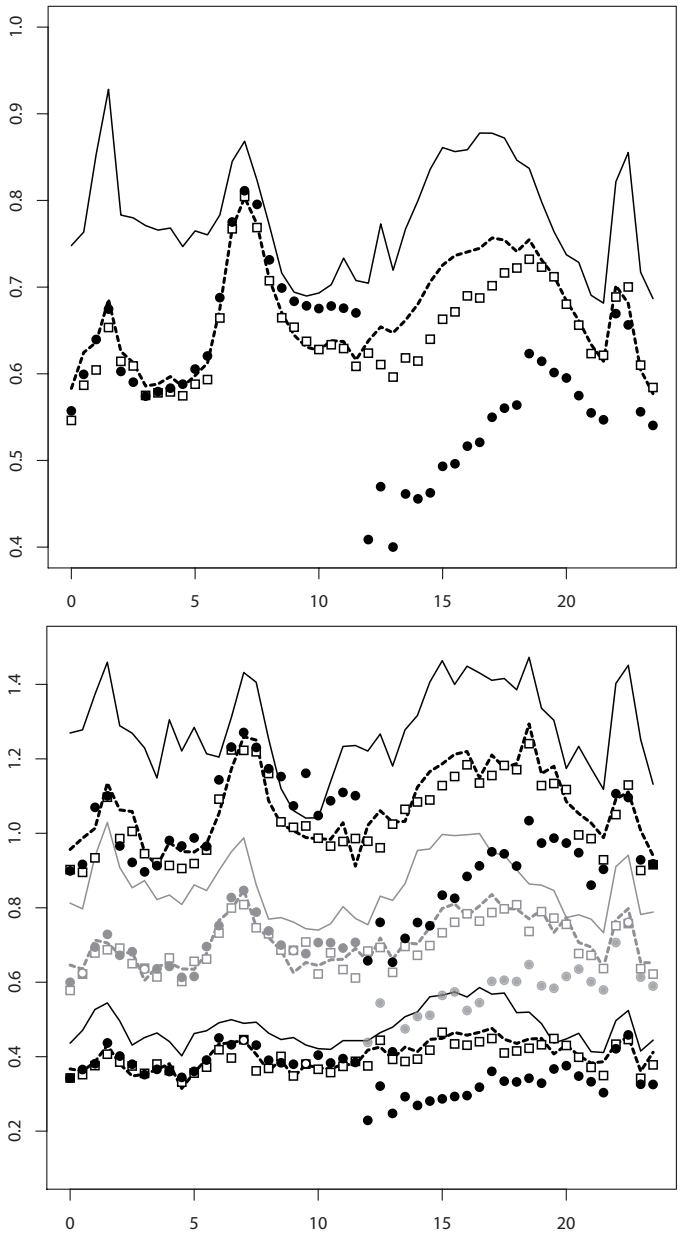

FIGURE 9. *Résultats demi-horaires du meilleur expert global (trait plein) et de la meilleure combinaison convexe globale (trait en pointillés), ainsi que des méta-stratégies calibrées fondées sur les familles* $\mathscr{W}_{\eta}^{\mathrm{grad}}$ *(symbole :* •*) et* $\mathscr{G}_{\eta,\alpha}^{\mathrm{grad}}$ *(symbole :* □*) ;* EQM *(en haut) et quantiles à* 50 % *(noir),* 75 % *(gris) et* 90 % *(noir) des valeurs absolues des résidus (en bas). Abscisses : demi-heures de la journées : ordonnées :* EQM.

 ISSN: 2102-6238

6. Conclusions et perspectives

Dans cet article portant sur la prévision séquentielle par agrégation de prévisions d'experts, on a tout d'abord décrit un cadre méthodologique et des stratégies de prévision générales. On a ensuite montré que ces stratégies pouvaient être appliquées avec succès, au prix d'adaptations mineures, à deux cadres applicatifs : la prévision de pics journaliers d'ozone et la prévision de consommation électrique à pas demi-horaire. Ce faisant, on encourage le lecteur à employer les stratégies décrites ici dans tout cadre de prévision séquentielle où il disposerait d'un ensemble d'experts dont il ne peut dire à l'avance qui sera le meilleur. En particulier, ces experts peuvent être donnés par des méthodes issues de modélisations stochastiques nécessitant le réglage de différents paramètres : une alternative au réglage peut être la considération de plusieurs instances de la méthode correspondant à différents jeux de paramètres suffisamment différents. Ou encore, on peut toujours ajouter des experts supplémentaires correspondant à des prédicteurs sans garantie théorique formelle mais dont l'intuition soutient qu'ils devraient bien se comporter. Les deux études empiriques ont démontré une variante de l'adage bien connu "Garbage in, garbage out" : lorsqu'il existe quelques bons experts, les stratégies d'agrégation auront elles aussi de bonnes performances. En particulier, il n'est pas besoin de ne construire que de bons experts, mais il faut pouvoir assurer qu'un petit nombre eux le sera (sans avoir besoin pour autant de savoir à l'avance lesquels).

Sur le plan applicatif, des perspectives et problèmes ouverts apparaissent dans chacune des deux études empiriques. En ce qui concerne la prévision des pics d'ozone, les projets à court terme sont d'évaluer l'impact de pré-traitement de débiaisement automatique sur les experts et de faire tourner les stratégies séquentielles en mode opérationnel et/ou sur de plus longues périodes. A moyen terme, il faudra s'intéresser à la prévision des dépassements de seuils réglementaires. On mentionne également un travail récent de Vivien Mallet sur les liens entre assimilation de données et agrégation d'experts [36]. Pour la prévision de consommation électrique, il s'agit à court terme de mieux comprendre pourquoi les stratégies de prévision par redistribution des poids sont si précises à horizon très court mais obtiennent des performances un peu plus décevantes juste avant la remise à jour des poids. Il faudra également tirer encore plus parti de la spécialisation en augmentant l'ensemble des experts. Enfin, il serait intéressant, ici comme en prévision de pics d'ozone, de pouvoir intégrer la quantification des incertitudes sur leurs prévisions que procurent les experts, afin de mieux choisir les poids des vecteurs de mélange mais aussi de pouvoir garantir une incertitude sur la prévision agrégée.

Par ailleurs, un projet d'application à un nouveau jeu de données, la prévision du taux de change euro-dollar, est en cours d'étude.

Références

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

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

 ISSN: 2102-6238

[3] P. AUER, N. CESA-BIANCHI et C. GENTILE : Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

[4] K.S. AZOURY et M. WARMUTH : Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43:211–246, 2001.

[5] P. BESSE, H. MILHEM, O. MESTRE, A. DUFOUR et V.-H. PEUCH : Comparaison de techniques de data mining pour l'adaptation statistique des prévisions d'ozone du modèle de chimie-transport MOCAGE. *Pollution Atmosphérique*, 195:285–292, 2007.

[6] D. BLACKWELL : An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

[7] A. BLUM et Y. MANSOUR : From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.

[8] A. BRUHNS, G. DEURVEILHER et 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.

[9] N. CESA-BIANCHI : Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999.

[10] N. CESA-BIANCHI, Y. FREUND, D. HAUSSLER, D.P. HELMBOLD, R. SCHAPIRE et M. WARMUTH : How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[11] N. CESA-BIANCHI et G. LUGOSI : Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51:239–261, 2003.

[12] N. CESA-BIANCHI et G. LUGOSI : *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[13] N. CESA-BIANCHI, Y. MANSOUR et G. STOLTZ : Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

[14] T. COVER : Behavior of sequential predictors of binary sequences. *In Proceedings of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 263–272. Maison d'édition de l'Académie des sciences de Tchécoslovaquie, Prague, 1965.

[15] T.M. COVER : Universal portfolios. *Mathematical Finance*, 1:1–29, 1991.

[16] V. DANI, O. MADANI, D. PENNOCK, S. SANGHAI et 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.

[17] M. DEVAINE, Y. GOUDE et G. STOLTZ : Aggregation of sleeping predictors to forecast electricity consumption. Rapport technique, EDF R&D et École normale supérieure, Paris, août 2009. Voir `http://www.math.ens.fr/%7stoltz/DeGoSt-report.pdf`.

[18] M. DEVAINE, Y. GOUDE et G. STOLTZ : Forecasting of the eletrical consumption by aggregation of sleeping experts ; application to Slovakian and French country-wide hourly predictions. 2010. Voir `http://www.math.ens.fr/%7Estoltz/publications`.

[19] V. DORDONNAT, S.J. KOOPMAN, M. OOMS, A. DESSERTAINE et J. COLLET : An hourly periodic state space model for modelling French national electricity load. *International Journal of Forecasting*, 24:566–587, 2008.

[20] B. EFRON, T. HASTIE, I. JOHNSTONE et R. TIBSHIRANI : Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[21] D. FOSTER : Prediction in the worst-case. *Annals of Statistics*, 19:1084–1090, 1991.

[22] Y. FREUND, R. SCHAPIRE, Y. SINGER et 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.

[23] Y. FREUND et R.E. SCHAPIRE : A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[24] S. GERCHINOVITZ : Communication personnelle, 2010.

[25] S. GERCHINOVITZ, V. MALLET et G. STOLTZ : A further look at sequential aggregation rules for ozone ensemble forecasting. Rapport technique, INRIA Paris-Rocquencourt et École normale supérieure, Paris, septembre 2008. Voir `http://www.math.ens.fr/%7Estoltz/GeMaSt-report.pdf`.

[26] B. GHATTAS : Prévisions des pics d'ozone par arbres de régression, simples et agrégés par *bootstrap*. *Revue de statistique appliquée*, 47(2):61–80, 1999.

 ISSN: 2102-6238

[27] Y. GOUDE : *Mélange de prédicteurs et application à la prévision de consommation électrique*. Thèse de doctorat, Université Paris-Sud, janvier 2008. Effectuée en convention avec EDF R&D.

[28] Y. GOUDE : Tracking the best predictor with a detection based algorithm. *In Proceedings of the Joint Statistical Meetings*. American Statistical Association, 2008. Voir la section de "Statistical Computing".

[29] J. HANNAN : Approximation to Bayes risk in repeated play. *In* M. DRESHER, A. TUCKER et P. WOLFE, éditeurs : *Contributions to the Theory of Games*, volume III, pages 97–139. Princeton University Press, 1957.

[30] M. HERBSTER et M. WARMUTH : Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

[31] A.E. HOERL et R.W. KENNARD : Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[32] J. KIVINEN et M. WARMUTH : Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[33] A. LEMPEL et J. ZIV : On the complexity of an individual sequence. *IEEE Transactions on Information Theory*, 22:75–81, 1976.

[34] N. LITTLESTONE et M. WARMUTH : The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[35] G. LUGOSI : Prédiction randomisée de suites individuelles. *Journal de la Société Française de Statistique*, 147:5–37, 2006.

[36] V. MALLET : Ensemble forecast of analyses : Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research*, 2010. Sous presse.

[37] V. MALLET, B. MAURICETTE et G. STOLTZ : Description of sequential aggregation methods and their performances for ozone ensemble forecasting. Rapport technique DMA-07-08, École normale supérieure, Paris, 2007.

[38] Vivien MALLET et Bruno SPORTISSE : Ensemble-based air quality forecasts : A multimodel approach applied to ozone. *Journal of Geophysical Research*, 111(D18), 2006.

[39] Vivien MALLET, Gilles STOLTZ et Boris MAURICETTE : Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research*, 114(D05307), 2009.

[40] A. PIERROT, N. LALUQUE et 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.

[41] A.E. RAFTERY, T. GNEITING, F. BALABDAOUI et M. POLAKOWSKI : Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174, 2005.

[42] R. TIBSHIRANI : Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[43] V. VOVK : Aggregating strategies. *In Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pages 372–383, 1990.

[44] V. VOVK : A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[45] V. VOVK : Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[46] V. VOVK et F. ZHDANOV : Prediction with expert advice for the Brier game. *In Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML)*, 2008.

[47] S.N. WOOD : *Generalized Additive Models : An Introduction with R*. Chapman and Hall/CRC, 2006.

[48] J. ZIV : Coding theorems for individual sequences. *IEEE Transactions on Information Theory*, 24:405–412, 1978.

[49] J. ZIV : Distortion-rate theory for individual sequences. *IEEE Transactions on Information Theory*, 26:137–143, 1980.

[50] J. ZIV et A. LEMPEL : A universal algorithm for sequential data-compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.

 ISSN: 2102-6238