Prévision séquentielle par agrégation d'ensemble : application à des prévisions météorologiques assorties d'incertitudes

Paul Baudin

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Université Paris-Sud

École Doctorale de mathématiques Jacques Hadamard
Laboratoire : équipe projet INRIA CLIME

Discipline : mathématiques et leurs interactions

Thèse de doctorat

Soutenue le 26 novembre 2015 par

Paul Baudin

Prévision séquentielle par agrégation d'ensemble :
application à des prévisions météorologiques assorties d'incertitudes

Co-directeur de thèse :	M. Gilles Stoltz	Chargé de recherche (CNRS)
Co-directeur de thèse :	M. Vivien Mallet	Chargé de recherche (INRIA)

Composition du jury :

Présidente du jury :	Mme. Élisabeth Gassiat	Professeur (Université Paris-Sud)
Rapporteurs :	M. Jean-Marc Bardet	Professeur (Université Panthéon-Sorbonne)
	M. Olivier Wintenberger	Professeur (Université Pierre et Marie Curie)
Examinateurs :	Mme. Liliane Bel	Professeur (Agroparitech)
	M. Olivier Mestre	Chercheur (Météo-France)
Invités :	M. Laurent Descamps	Chercheur (Météo-France)

Table des matières

1 Introduction

Dans cette thèse, nous nous intéressons à des problèmes de prévision tour après tour. L'objectif est d'imaginer et d'appliquer des stratégies automatiques, qui tirent de l'expérience du passé. Nous souhaitons que ces stratégies obtiennent des garanties mathématiques robustes et soient valables dans des cas de figure très généraux. Cela nous permet en pratique d'appliquer ces algorithmes à la prévision concrète de grandeurs météorologiques. Enfin, nous nous intéressons aux déclinaisons théoriques et pratiques dans un cadre de prévision de fonctions de répartition.

Sommaire

1.1 Présentation des suites arbitraires

1.1.1 Prévision de suite arbitraire

Cadre des suites arbitraires Considérons la situation de prévision suivante : à chaque échéance discrète $t \in \mathbb{N}^\star$, le statisticien cherche à prévoir la valeur y_t d'une *observation*, appartenant à un ensemble $\mathcal{Y}$ quelconque, avant que celle-ci ne soit révélée. La suite des observations $(y_t)_t$ est appelée *suite arbitraire* (on parle aussi de *suite individuelle*) car aucune hypothèse stochastique n'est imposée a priori sur cette suite. En particulier et contrairement au cadre statistique classique, les y_t ne sont pas nécessairement des réalisations d'une variable aléatoire. En toute généralité, la suite peut même être de nature adversariale et chercher activement à contrecarrer les prévisions du statisticien.

Soit $\mathcal{X}$, l'ensemble convexe d'où sont issues les prévisions, par exemple un segment de $\mathbb{R}$. À chaque échéance t, le statisticien propose une prévision de y_t, notée $\widehat{y}_t \in \mathcal{X}$. À cette fin, le statisticien dispose d'un *historique*, une réserve d'information h_t dont nous ne figeons pas la définition à ce stade de l'exposé. Si $\mathcal{H}_t$ est l'ensemble possible des historiques, $h_t \in \mathcal{H}_t$ en est une réalisation, constituée *a minima* des observations et des prévisions du passé et éventuellement d'autres sources à préciser (par exemple, les avis passés et présents des experts, comme nous les définirons en partie 1.1.2) : $\left\{(y_s, \widehat{y}_s) : s \in \{1, \dots, t-1\}\right\} \subset \mathcal{H}_t$. Pour simplifier l'écriture, nous omettons la dépendance de $\widehat{y}_t$ à l'historique h_t. L'objectif est de réaliser une suite $(\widehat{y}_t)_t$ dont chaque terme s'approche au mieux du terme correspondant de la suite $(y_t)_t$. Pour préciser ce « au mieux », il est nécessaire de définir une mesure de performance permettant de comparer une prévision et l'observation correspondante.

Mesure de performance Cette mesure quantifiée de performance est ce que l'on nomme une *fonction de perte*. Il s'agit d'une fonction réelle de deux variables $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, convexe par rapport à la première variable (c'est-à-dire par rapport aux prévisions). Ainsi, à une échéance t donnée, le statisticien prévoit $\widehat{y}_t$, puis l'observation y_t est révélée ce qui conduit à une perte instantanée $\ell(\widehat{y}_t, y_t)$. Voici quelques exemples de telles fonctions de perte :

- la perte quadratique $\ell(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$;
- la perte en valeur absolue $\ell(\widehat{y}_t, y_t) = |\widehat{y}_t - y_t|$;
- la perte *pinball*, pour $\alpha \in]0,1[$, $\ell_\alpha(\widehat{y}_t, y_t) = \left(\alpha - \mathbb{1}_{\{y_t \geqslant \widehat{y}_t\}}\right)(y_t - \widehat{y}_t)$;
- le *continuous ranked probability score* (CRPS) dont la description est donnée en partie 1.5 et plus en détail au chapitre 5.

Soit $T \in \mathbb{N}^\star$, un horizon temporel fixé, la *perte cumulée du statisticien* sur les T premières échéances est définie comme :

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

Pour une suite quelconque de prévisions $z = (z_t)_{t\in\mathbb{N}^\star} \in \mathcal{X}^{\mathbb{N}^\star}$, on écrit la perte cumulée correspondante :

$$L_T(z) = \sum_{t=1}^{T} \ell(z_t, y_t)\,,$$

et ainsi, la perte cumulée du statisticien n'est autre que la perte cumulée de la suite de prévisions : $\widehat{L}_T = L_T(\widehat{y})$ où $\widehat{y} = (\widehat{y}_t)_{t\in\mathbb{N}^\star}$. Même si sa formulation ne le met pas en valeur, la perte $L_T(z)$ dépend aussi des observations $(y_t)_{t\in\mathbb{N}^\star}$. Le déroulement d'une échéance de prévision est précisé à l'algorithme 1.

À chaque échéance $t = 1, 2, \ldots$

1. L'information h_t est fournie au statisticien.
2. Celui-ci propose une prévision $\widehat{y}_t$, construite à partir de h_t.
3. L'observation y_t est révélée et la perte instantanée $\ell(\widehat{y}_t, y_t)$ est encourue par le statisticien.

Algorithm 1: Déroulement d'une suite d'échéances de prévision.

Prévoir le pire des cas Cherchons à préciser l'objectif. Les observations sont toujours révélées après les prévisions et, dans le pire des cas, la suite est adversariale : elle peut alors systématiquement faire subir la perte la plus grande possible au statisticien au vu de sa prévision $\widehat{y}_t$. De ce fait, minimiser la perte cumulée $\widehat{L}_T$ uniformément par rapport à toutes les suites $(y_t)_t$ est illusoire. En revanche, nous pouvons désormais comparer deux suites de prévisions entre elles via leurs pertes cumulées ce qui permet de définir un critère relatif de performance : le *regret*.

Des garanties nécessairement relatives Soit une classe de prédicteurs $\mathcal{C}$ dont les éléments sont les suites de fonctions $\tilde{y}_t : \mathcal{H}'_t \to \mathcal{X}$ vérifiant certaines propriétés que nous préciserons au cas par cas et où $\mathcal{H}'_t$ est l'analogue de $\mathcal{H}_t$ pour $\tilde{y}_t$. Le *regret* $R_T^{\mathcal{C}}$, est la différence entre la perte cumulée du statisticien et la perte cumulée du meilleur prédicteur issu de la classe de référence $\mathcal{C}$ à l'horizon T :

$$R_T^{\mathcal{C}} = \widehat{L}_T - \inf_{\tilde{y}\in\mathcal{C}} L_T\left(\tilde{y}\right)\,,$$

où $\tilde{y} = (\tilde{y}_t)_{t\in\mathbb{N}^\star}$. Le regret peut être vu comme le manque à gagner que le statisticien subit de n'avoir pas choisi le meilleur prédicteur jusqu'à présent issu de la classe de référence $\mathcal{C}$. Le regret dépend donc de la stratégie de prévision du statisticien et crucialement de la classe de comparaison $\mathcal{C}$. Minimiser ce critère relatif est un objectif réalisable dans certains cas de figure. L'objectif du statisticien est donc de créer une suite de prévisions $(\widehat{y}_t)_{t\in\mathbb{N}^\star} \in \mathcal{Y}^{\mathbb{N}^\star}$ dont le regret moyen est négatif ou nul asymptoti-

quement :

$$\limsup_{T\to+\infty}\left(\sup\frac{R_T^{\mathcal{C}}}{T}\right)\leqslant 0\,; \tag{1.1}$$

où le supremum porte sur l'ensemble des suites d'observations $(y_t)_{t\in\mathbb{N}^\star}\in\mathcal{Y}^{\mathbb{N}^\star}$ et sur toute autre information dont dispose le statisticien pour réaliser ses prévisions, i.e. $\mathcal{H}_t\setminus\left\{\widehat{y}_s : s\in\{1,\dots,t-1\}\right\}$. Il s'agit d'un cadre de prévision robuste car il permet de construire des prévisions dont les garanties sont valables pour toutes les suites $(y_t)_t$ d'observations, pas seulement les plus typiques ou les plus probables.

Stratégies automatiques de prévisions Pour mener à bien ce projet, le statisticien conçoit des algorithmes qui produisent automatiquement les termes de la suite de prévisions $\widehat{y}_t$ à partir de l'information h_t. Lorsque nous notons $\mathcal{S}$ un de ces algorithmes, nous soulignons parfois la dépendance de la perte cumulée ou du regret à cet algorithme en les notant respectivement $\widehat{L}_T(\mathcal{S})$ et $R_T^{\mathcal{C}}(\mathcal{S})$. L'analyse de ces algorithmes permet d'assurer que le regret moyen est asymptotiquement nul, ou mieux encore, que le regret est sous-linéaire, c'est-à-dire négligeable devant T :

$$\sup R_T^{\mathcal{C}}=o(T)\,.$$

Cette dernière garantie de vitesse implique la garantie de consistance de l'équation 1.1.

Pour résumer, l'objectif du statisticien dans le domaine de la prévision de suites arbitraires est donc d'inventer, d'analyser et de mettre en pratique de tels algorithmes automatiques, aux garanties robustes. Déclinons ce cadre général aux cas classiques de prévision avec avis d'experts en précisant $\mathcal{H}_t$, $\mathcal{C}$, les stratégies automatiques et les garanties de vitesse.

1.1.2 Cas classiques de prévision avec avis d'experts

La *prévision avec avis d'expert* constitue le cadre de prévision le plus fréquent, que nous employons aux chapitres 3, 4 et 5.

Information disponible Soit $M\in\mathbb{N}^\star$. Dans ce cadre, afin de former sa prévision à chaque échéance, le statisticien dispose d'un ensemble de M *prédicteurs élémentaires* (appelés aussi *experts*). Ceux-ci constituent les briques fondamentales des prévisions $\widehat{y}_t$. À chaque échéance t, le $m^{\text{ème}}$ prédicteur fournit une prévision $x_{m,t}\in\mathcal{X}$ qui s'appuie potentiellement sur les observations passées $y_1,\dots,y_{t-1}$ et sur des informations à lui seul accessibles. L'information dont le statisticien dispose est alors :

$$h_t=\left\{y_s,\,\widehat{y}_s : s\in\{1,\dots,t-1\}\right\}\cup\left\{(x_{m,s})_{1\leqslant m\leqslant M} : s\in\{1,\dots,t\}\right\} \tag{1.2}$$

La théorie des suites arbitraires ne se préoccupe pas de l'origine des ces prédictions élémentaires. Dans cette introduction, nous considérons ces prédicteurs comme des boîtes noires. Nous verrons concrètement dans les chapitres 3, 4 et 5, comment les prévisions

météorologiques permettent de construire un ensemble de prédicteurs élémentaires.

La prévision, combinaison linéaire des experts Les stratégies automatiques consistent à choisir à chaque échéance t, un vecteur de poids $\boldsymbol{p}_t = (p_{m,t})_{1 \leqslant m \leqslant M} \in \mathbb{R}^M$. Le statisticien réalise alors une combinaison linéaire des prédicteurs, c'est-à-dire que sa prévision $\widehat{y}_t$ résulte de la pondération des prévisions élémentaires $x_{m,t}$:

$$\widehat{y}_t = \sum_{m=1}^{M} p_{m,t}\, x_{m,t}\,.$$

En toute généralité, les poids $\boldsymbol{p}_t$ appartiennent à $\mathbb{R}^M$, sans restriction. Certaines méthodes imposent cependant que ces poids soient convexes :

$$\forall\, m \in \{1, \dots, M\}, \qquad p_{m,t} \geqslant 0 \qquad \text{et} \qquad \sum_{m=1}^{M} p_{m,t} = 1\,.$$

Dans ce cadre général de prévision avec avis d'experts, lorsqu'il n'y a pas d'ambiguïté possible, nous employons les notations plus maniables suivantes : $\ell_{m,t} = \ell(x_{m,t},\, y_t)$ pour la perte instantanée du prédicteur élémentaire $x_{m,t}$ et

$$\widehat{\ell}_t = \ell(\widehat{y}_t, y_t) = \ell\left(\sum_{m=1}^{M} p_{m,t}\, x_{m,t},\; y_t\right)\,;$$

pour la perte instantanée de la prévision du statisticien.

1.1.3 Algorithme de pondération par poids exponentiels

L'algorithme de pondération par poids exponentiels a été introduit en apprentissage séquentiel par Littlestone et Warmuth [LW94] et Vovk [Vov90]. Intuitivement, cet algorithme de pondération convexe fournit un poids plus important aux prédicteurs élémentaires ayant eu les meilleures performances dans le passé et vice versa.

Classe de comparaison On définit la perte cumulée d'un expert m fixé sur les T premières échéances par :

$$L_{m,T} = \sum_{t=1}^{T} \ell_{m,t}\,.$$

La classe de comparaison $\mathcal{C}$ est ici constituée des M prédicteurs fondamentaux et le regret à l'échéance T est alors la différence entre la perte cumulée du statisticien et la perte cumulée du meilleur expert :

$$R_T^{\exp} = \widehat{L}_T - \min_{m=1,\dots,M} L_{m,T}\,; \tag{1.3}$$

où l'exposant exp de $R_T^{\exp}$ traduit justement que la classe de comparaison $\mathcal{C}$ est constituée des M experts. La stratégie $\mathcal{E}_\eta$ de pondération par poids exponentiels des pertes cumulées est décrite à l'algorithme 2 et les garanties théoriques associées sont données au théorème 1.

- **Paramétrisation :** choisir le *paramètre d'apprentissage* $\eta > 0$.
- **Initialisation :** $\boldsymbol{p}_1$ est le vecteur de mélange uniforme, $p_{m,1} = 1/M$, pour $m = 1, \dots, M$.
- **À chaque échéance** $t = 1, 2, \dots$, le vecteur des poids $\boldsymbol{p}_{t+1}$ est défini pour $m = 1, \dots, M$ par :

$$p_{m,t+1} = \frac{\mathrm{e}^{-\eta L_{m,t}}}{\sum_{k=1}^{M} \mathrm{e}^{-\eta L_{k,t}}} \,.$$

Algorithm 2: Algorithme de pondération par poids exponentiels, $\mathcal{E}_\eta$

Théorème 1. *On suppose que la fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ est bornée, à valeurs dans $[0, B]$ et convexe en son premier argument ; alors pour tout $\eta > 0$,*

$$\sup\Big\{ R_T^{\exp}(\mathcal{E}_\eta) \Big\} = \sup\left\{ \widehat{L}_T(\mathcal{E}_\eta) - \min_{m=1,\dots,M} L_{m,T} \right\} \leqslant \frac{\ln M}{\eta} + \eta \frac{B^2}{8} T \,,$$

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

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

Calibration du paramètre d'apprentissage Notons que le choix théorique de $\eta^\star$ n'est pas réalisable en pratique. En effet, d'une part, on ne connaît pas l'horizon temporel T puisque l'on souhaite en général exécuter l'algorithme de prévision indéfiniment, d'autre part, on ne connaît pas nécessairement la valeur de la borne B. En ce qui concerne l'adaptation par rapport à T, il existe plusieurs possibilités qui permettent de conserver la borne de vitesse à une constante multiplicative près : par exemple le « *doubling trick* » (voir Cesa-Bianchi et Lugosi [CL06]) consiste à relancer l'algorithme avec un nouveau paramètre $\eta^{(r)} = (1/B)\sqrt{(8 \ln M)/2^r}$ à chaque fois qu'une échéance 2^r est atteinte avec $r \in \mathbb{N}$. Une autre solution consiste à modifier la valeur de η_t à chaque échéance : $\eta_t = (1/B)\sqrt{(8 \ln M)/t}$. Enfin, il existe aussi des méthodes de grilles qui consistent à lancer l'algorithme avec plusieurs paramètres (sur une *grille*) et à calibrer directement le paramètre d'apprentissage sur les données à la volée. Cela est fait en détail pour un autre algorithme que les poids exponentiels aux chapitres 3 et 4.

Ce théorème dérive du Lemme 1 ci-dessous (cf. Stoltz [Sto10]), en remarquant qu'on a l'inégalité de convexité suivante :

$$\widehat{L}_T(\mathcal{E}_\eta) = \sum_{t=1}^{T} \ell\left(\sum_{m=1}^{M} p_{m,t}\, x_{m,t},\, y_t \right) \leqslant \sum_{t=1}^{T} \sum_{m=1}^{M} p_{m,t}\, \ell(x_{m,t}, y_t) = \sum_{t=1}^{T} \sum_{m=1}^{M} p_{m,t}\, \ell_{m,t} \,.$$

Lemme 1. *Soit $a, b \in \mathbb{R}$ avec $a < b$. Pour tout $\eta > 0$ et toute suite d'éléments $\ell_{j,t} \in [a, b]$, où $j \in \{1, \dots, M\}$ et $t \in \{1, \dots, T\}$,*

$$\sum_{t=1}^{T} \sum_{m=1}^{M} p_{m,t}\, \ell_{m,t} - \min_{m=1,\dots,M} \sum_{t=1}^{T} \ell_{m,t} \leqslant \frac{\ln M}{\eta} + \eta \frac{(b-a)^2}{8} T\,,$$

où on définit les poids de manière identique à l'algorithme 2, i.e. pour tout $m = 1, \dots, M$,

$$p_{m,t} = \frac{\exp\Bigl(-\eta \sum_{s=1}^{t-1} \ell_{m,s}\Bigr)}{\sum_{k=1}^{M} \exp\Bigl(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\Bigr)}\,.$$

Par convention, une somme sur un nombre nul d'élément est nulle ce qui implique pour tout $m = 1, \dots, M$ que

$$p_{m,1} = \frac{1}{M}\,.$$

Démonstration. La preuve (cf. Cesa-Bianchi et Lugosi [CL06, théorème 2.2]) repose sur le lemme de Hoeffding : soit $a, b \in \mathbb{R}$ avec $a < b$, si Z est une variable aléatoire bornée, à valeurs dans $[a, b]$, alors $\forall s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[\mathrm{e}^{sZ}\right] \leqslant s\, \mathbb{E}[Z] + \frac{s^2}{8}(b-a)^2\,.$$

En considérant la variable aléatoire qui prend comme valeur $\ell_{m,t}$ avec une probabilité discrète de $p_{m,t}$, pour tout $m = 1, \dots M$, on a

$$-\eta \sum_{m=1}^{M} p_{m,t}\, \ell_{m,t} \geqslant \ln \frac{\sum_{m=1}^{M} \exp(-\eta \sum_{s=1}^{t} \ell_{m,s})}{\sum_{k=1}^{M} \exp(-\eta \sum_{s=1}^{t-1} \ell_{k,s})} - \frac{\eta^2}{8}(b-a)^2\,;$$

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

$$\sum_{t=1}^{T} \sum_{m=1}^{M} p_{m,t}\, \ell_{m,t} \leqslant -\frac{1}{\eta} \ln \frac{\sum_{m=1}^{M} \exp\Bigl(-\eta \sum_{s=1}^{T} \ell_{m,s}\Bigr)}{M} + \eta \frac{(b-a)^2}{8} T\,. \tag{1.4}$$

Par ailleurs, nous remarquons que

$$\sum_{m=1}^{M} \exp\left(-\eta \sum_{s=1}^{T} \ell_{m,s}\right) \geqslant \exp\left(-\eta \min_{m=1,\dots,M} \sum_{s=1}^{T} \ell_{m,s}\right)$$

car la somme est plus grande que le plus grand de ses termes. La décroissance de la fonction $u \mapsto -\frac{1}{\eta}\ln(u)$ implique l'inégalité suivante

$$-\frac{1}{\eta} \ln \sum_{m=1}^{M} \exp\left(-\eta \sum_{s=1}^{T} \ell_{m,s}\right) \leqslant \min_{m=1,\dots,M} \sum_{s=1}^{T} \ell_{m,s}$$

que l'on réinjecte dans l'inégalité (1.4) ce qui conclut la preuve. □

La stratégie de pondération exponentielle $\mathcal{E}_\eta$ permet donc de fournir une garantie de vitesse par rapport à la classe des M prédicteurs élémentaires. Une modification de cet algorithme utilisant la convexité de la fonction de perte permet de dériver des garanties similaires face à une classe de comparaison plus vaste et donc plus intéressante : la classe de toutes les combinaisons convexes constantes des M prédicteurs élémentaires.

1.1.4 Astuce du gradient

Le simplexe correspond à l'ensemble des poids convexes :

$$\mathcal{P} = \left\{ \boldsymbol{q} : \sum_{m=1}^{M} q_m = 1 \text{ et } q_m \geqslant 0 \,, \forall\, m \in \{1, \dots, M\} \right\} .$$

Dans cette partie, la classe de référence $\mathcal{C}$ est l'ensemble des combinaisons convexes constantes des prédicteurs élémentaires :

$$\mathcal{C} = \left\{ \left(\sum_{m=1}^{M} q_m x_{m,t} \right)_{t \in \mathbb{N}^\star} : \boldsymbol{q} \in \mathcal{P} \right\} .$$

Pour atteindre un regret par rapport à $\mathcal{C}$ sous-linéaire par rapport à T, il est possible d'employer l'astuce du gradient (« *gradient trick* »), introduite par Kivinen et Warmuth [KW97] et dont l'analyse est due à Cesa-Bianchi [Ces99]. Cette astuce revient à remplacer la fonction de perte instantanée par l'expression de son gradient dans la formule des poids. Si l'on part de la stratégie de pondération exponentielle précédente, on obtient donc la stratégie de pondération exponentielle des sous-gradients des pertes cumulées $\mathcal{E}_\eta^{\mathrm{grad}}$ qui est alors compétitive face à toutes les combinaisons convexes constantes des M prédicteurs élémentaires. Nous supposons que $\mathcal{X} \subseteq \mathbb{R}^M$ et que les fonctions $\ell(\,\cdot\,, y)$ sont différentiables sur $\mathcal{X}$ pour tout $y \in \mathcal{Y}$. On note $\nabla \ell(\,\cdot\,, y)$ le gradient par rapport à la première variable au point y. Pour tout couple u, v de points de $\mathcal{X}$, l'inégalité des pentes pour la fonction convexe ℓ s'écrit

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

En particulier, on a pour des vecteurs de poids $\boldsymbol{p}$, $\boldsymbol{q} \in [0,1]^M$, pour toutes les prévisions $(x_1, \dots, x_M) \in \mathcal{X}^M$ et toute observation $y \in \mathcal{Y}$,

$$\ell\left(\sum_{m=1}^{M} p_m \, x_m, y \right) - \ell\left(\sum_{m=1}^{M} q_m \, x_m, y \right) \leqslant \nabla \ell\left(\sum_{k=1}^{M} p_k \, x_k, y \right) \cdot \left(\sum_{m=1}^{M} p_m \, x_m - \sum_{m=1}^{M} q_m \, x_m \right) . \tag{1.5}$$

On définit alors pour l'expert $m \in \{1, \dots, M\}$ à l'échéance $t \in \{1, \dots, T\}$, la pseudo-perte

$$\widetilde{\ell}_{m,t} = \nabla \ell \left(\sum_{k=1}^{M} p_{k,t} \, x_{k,t}, \, y_t \right) \cdot x_{m,t} \,, \tag{1.6}$$

qui intervient dans la stratégie $\mathcal{E}_\eta^{\mathrm{grad}}$ de pondération par poids exponentiels des pertes cumulées décrite à l'algorithme 2, nommée aussi *EG* dans la suite pour « *exponentiated gradient* ».

Le regret R_T^{cvx} de $\mathcal{E}_\eta^{\mathrm{grad}}$ par rapport à la meilleure combinaison de poids convexe est

- **Paramétrisation :** choisir la *paramètre d'apprentissage* $\eta > 0$.
- **Initialisation :** $\boldsymbol{p}_1$ est le vecteur de mélange uniforme, $p_{m,1} = 1/M$, pour $m = 1, \dots, M$.
- **À chaque échéance** le vecteur des poids $\mathbf{p}_{t+1}$ est défini composante par composante selon

$$p_{m,t+1} = \frac{\exp\Bigl(-\eta \sum_{s=1}^{t} \widetilde{\ell}_{m,s}\Bigr)}{\sum_{k=1}^{M} \exp\Bigl(-\eta \sum_{s=1}^{t} \widetilde{\ell}_{k,s}\Bigr)}\,,$$

où $\widetilde{\ell}_{m,s}$ est la pseudo-perte définie en (1.6).

Algorithm 3: Algorithme *EG* de pondération par poids exponentiels, $\mathcal{E}_\eta^{\mathrm{grad}}$

alors majoré selon

$$\begin{aligned} R_T^{\mathrm{cvx}}(\mathcal{E}_\eta^{\mathrm{grad}}) &= \sup_{\boldsymbol{q}\in\mathcal{P}} \sum_{t=1}^{T} \biggl(\ell\Bigl(\sum_{m=1}^{M} p_{m,t}\, x_{m,t}, y_t\Bigr) - \ell\Bigl(\sum_{m=1}^{M} q_m\, x_{m,t}, y_t\Bigr) \biggr) \\ &\leqslant \sup_{\boldsymbol{q}\in\mathcal{P}} \sum_{t=1}^{T} \biggl(\sum_{m=1}^{M} p_{m,t}\, \widetilde{\ell}_{m,t} - \sum_{m=1}^{M} q_m\, \widetilde{\ell}_{m,t} \biggr) \\ &= \sum_{t=1}^{T} \sum_{m=1}^{M} p_{m,t}\, \widetilde{\ell}_{m,t} - \min_{m=1,\dots,M} \sum_{t=1}^{T} \widetilde{\ell}_{m,t}\,; \end{aligned}$$

où l'inégalité procède de (1.5) et la seconde égalité du fait que le majorant ainsi obtenu est linéaire en $\boldsymbol{q}$ et est donc maximisé par une masse de Dirac. Le théorème 2 découle alors directement du Lemme 1.

Théorème 2. *On suppose que les pseudo-pertes sont bornées sur les domaines $\mathcal{X}$ et $\mathcal{Y}$ considérés, à valeurs dans $[-C, C]$. Alors pour tout $\eta > 0$,*

$$\sup\Bigl\{ R_T^{\mathrm{cvx}}(\mathcal{E}_\eta^{\mathrm{grad}}) \Bigr\} = \sup\Bigl\{ \widehat{L}_T(\mathcal{E}_\eta^{\mathrm{grad}}) - \inf_{\boldsymbol{q}\in\mathcal{P}} L_T(\boldsymbol{q}) \Bigr\} \leqslant \frac{\ln M}{\eta} + \eta \frac{C^2}{2} T\,,$$

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

$$\sup\Bigl\{ R_T^{\mathrm{cvx}}(\mathcal{E}_{\eta^\star}^{\mathrm{grad}}) \Bigr\} \leqslant C\sqrt{2\,T \ln M}\,.$$

En modifiant l'algorithme des poids exponentiels, l'astuce du gradient permet d'obtenir une stratégie compétitive $\mathcal{E}_\eta^{\text{grad}}$ face à la classe des combinaisons convexes des prédicteurs élémentaires. De plus, cette adaptation est réalisée sans détériorer la vitesse ce qui est très appréciable : les théorèmes 1 et 2 indiquent que le majorant est à chaque fois un $\mathcal{O}(\sqrt{T})$.

1.2 Arbre de régression déterministe

Le chapitre 2 étudie un problème de prévision auto-régressive, i.e. employant uniquement le passé proche, qui est traditionnellement abordé sous l'angle stochastique. Ici, les suites arbitraires permettent d'exhiber des garanties de vitesse face à une classe de référence de fonctions lipschitziennes puis de retrouver des garanties de consistance dans un cadre stochastique. On suppose dans toute la suite que $\mathcal{X} = \mathcal{Y} = [0,1]$.

Classe de comparaison Soit $d \in \mathbb{N}^\star$, soit $L \in \mathbb{R}_+^\star$. La classe de comparaison est l'ensemble des fonctions L-lipschitziennes : $\mathcal{C} = \mathcal{L}_L^d = \{f : [0,1]^d \to [0,1],\ f \text{ est } L-\text{lipschitzienne}\}$. $\mathcal{L}_L^d$ est une classe non-paramétrique et plus massive que celle des combinaisons convexes d'experts étudiée à la partie 1.1.4. Cette dernière est paramétrique car elle s'identifie essentiellement au simplexe $\mathcal{P}$. À la différence de la partie précédente, cette partie ne fait intervenir aucune prévision d'expert et, en ce sens, ce cadre atypique est proche de la calibration décrite par Foster et Vohra [FV98] et dans la monographie de Cesa-Bianchi et Lugosi [CL06], chapitre 4.5.

Nested EG Le chapitre décrit un algorithme de prévision par strate. La brique de base est l'algorithme *EG* détaillé dans l'algorithme 3, mis en œuvre sur un jeu de deux pseudo-experts prévoyant toujours 0 ou 1. Ainsi, les prévisions ont lieu dans le segment délimité par ces pseudo-experts : le segment $[0,1]$. Les garanties de vitesse en $\mathcal{O}(\sqrt{T})$ sont celles présentées dans la partie 1.1.4. Il s'agit de maintenir simultanément plusieurs de ces algorithmes de prévision dans des régions de l'espace $[0,1]^d$ dont la taille s'adapte en fonction des observations passées. Autrement dit, l'algorithme « *Nested EG* » emploie l'idée de l'arbre de régression, que l'on retrouve dans l'ouvrage de Breiman, Friedman, Stone et Olshen [Bre+84], et construit des partitions s'adaptant à la densité des données afin de fournir des prévisions plus fines dans les zones denses et vice versa.

1.2.1 Garanties auto-régressives pour des suites individuelles

Les fonctions de la classe de référence $\mathcal{L}_L^d$ sont appliquées aux valeurs du d-passé à chaque échéance t : $y_{t-d}^{t-1} = (y_{t-d}, \dots, y_{t-1})$. Si l'on fixe la longueur d de ce passé, le lemme 2.3.1 fournit une borne supérieure de la perte cumulée de l'algorithme *Nested EG* esquissé dans la partie précédente 1.2.

Lemme 2. *On suppose que ℓ est convexe et κ-lipschitzienne. Soit $T \geqslant 1$ et $d \in \{1, \dots, T\}$. Alors, pour tout $y_1, \dots, y_T \in [0,1]$ et tout $L > 0$, la version auto-régressive*

de Nested EG *vérifie la majoration de regret suivante*

$$R_{L,T}^{d} \stackrel{\text{def}}{=} \sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) \; - \inf_{f \in \mathcal{L}_L^d} \sum_{t=d+1}^{T} \ell\big(f(y_{t-d}^{t-1}), y_t\big) \leqslant \kappa(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,. \tag{1.7}$$

Le lemme 2.3.1 fournit des garanties pour une longueur du passé d fixée. Afin de s'affranchir de ce paramètre, on définit une nouvelle couche d'agrégation qui emploit des algorithmes *Nested EG* auto-régressifs du lemme 2.3.1 lancés par régimes de tailles d croissantes. Ces différents algorithmes sont alors eux-mêmes considérés comme des experts dits spécialisés (voir Freund, Schapire, Singer et Warmuth [Fre+97]) et combinés via un algorithme *EG* approprié. Les garanties de cet algorithme final sont données au théorème 3. Soulignons de plus qu'il est valable pour tout $L > 0$.

Théorème 3. *On suppose que ℓ est convexe et κ-lipschitzienne. Il existe un algorithme construit sur les algorithmes présentés au lemme 2.3.1 qui vérifie pour tout $T \in \mathbb{N}^\star$, pour tout L, toute suite $y_1, \dots, y_T \in [0,1]$, et tout $d \in \{0, \dots, T-1\}$,*

$$R_{L,T}^{d} \stackrel{\text{def}}{=} \sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=d+1}^{T} \ell\left(f\left(y_{t-d}^{t-1}\right), y_t\right) \leqslant \mathcal{O}\left(T^{\frac{d+1}{d+2}}\right)\,.$$

Ainsi, pour tout $d \geqslant 1$, pour tout $L > 0$ et pour toute suite $y_1, \dots, y_T, \dots \in [0,1]$,

$$\limsup_{T \to \infty} \left(R_{L,T}^{d}/T\right) \leqslant 0\,.$$

1.2.2 Consistance de la stratégie dans le cadre stochastique

Nous montrons que la stratégie précédente de prévision de suites arbitraires permet d'obtenir des garanties optimales dans un cadre stochastique.

Performance optimale possible Considérons la situation de prévision stochastique suivante : à chaque échéance $t = 1, 2, \dots$, le statisticien doit former une prévision $\widehat{Y}_t$ de l'observation à venir $Y_t \in [0,1]$ issue d'un processus borné stationnaire et ergodique $(Y_t)_{t=-\infty,\dots,\infty}$ en employant sa connaissance du passé $Y_1, \dots, Y_{t-1}$. L'évaluation de la prévision est réalisée via une fonction de perte, $\ell : [0,1]^2 \to [0,1]$. Algoet [Alg94] a prouvé la borne inférieure fondamentale suivante pour toute fonction de perte continue et bornée. On note $\mathcal{B}^\infty$, l'ensemble des fonctions boréliennes de $[0,1]^{\mathbb{N}^\star}$ vers $[0,1]$. Pour toute stratégie de prévision, on a presque sûrement

$$\liminf_{T \to \infty} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell(\widehat{Y}_t, Y_t) \right\} \geqslant L^\star \qquad \text{p.s.}\,,$$

avec

$$L^\star = \mathbb{E}\left[\inf_{f \in \mathcal{B}^\infty} \mathbb{E}\left[\ell(f(Y_{-\infty}^{-1}), Y_0) | Y_{-\infty}^{-1} \right] \right]\,.$$

$L^\star$ est donc l'espérance de la perte minimale par rapport à toutes les estimations boréliennes de l'observation Y_0 à partir du passé infini.

Consistance Nous montrons que la stratégie esquissée au théorème 3 est *consistante* c'est-à-dire que la borne inférieure précédente $L^\star$ est atteinte :

$$\limsup_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}\ell(\widehat{Y}_t,Y_t)\right)\leqslant L^\star \qquad \text{p.s.}\,. \tag{1.8}$$

En utilisant des méthodes de plus proches voisins, les travaux de Györfi, Lugosi et Fargas [GLF01] montrent des résultats similaires dans le cas de la fonction de perte carrée et ceux de Biau et Patra [BP11] dans le cas de la fonction de perte *pinball*. Nos résultats sont plus généraux : il nous suffit que la fonction de perte ℓ soit bornée, convexe et lipschitzienne. Le théorème 4 présente cette garantie.

Théorème 4. *Soit $(Y_t)_{t=-\infty,\dots,\infty}$ un processus borné à valeur dans $[0,1]$ stationnaire et ergodique. On suppose que pour tout $d\geqslant 1$ la loi de probabilité de $Y_{-d}^{-1}=(Y_{-d},\dots,Y_{-1})$ est régulière, c'est-à-dire que pour tout borélien $S\subset[0,1]^d$ et pour tout $\varepsilon>0$, il existe un compact K et un ouvert V tels que*

$$K\subset S\subset V, \qquad \text{et} \qquad \mathbb{P}_{Y_{-d}^{-1}}(V\backslash K)\leqslant\varepsilon\,.$$

Soit $\ell:[0,1]^2\to[0,1]$ une fonction de perte lipschitzienne et convexe en son premier argument. Considérons une stratégie de prévision vérifiant presque sûrement,

$$\forall d\geqslant 1,\quad \forall L\geqslant 0, \qquad \limsup_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}\ell\Big(\widehat{Y}_t,Y_t\Big)\right)\leqslant\limsup_{T\to\infty}\left(\inf_{f\in\mathcal{L}_L^d}\frac{1}{T}\sum_{t=1}^{T}\ell\Big(f(Y_{t-d}^{t-1}),Y_t\Big)\right), \tag{1.9}$$

alors,

$$\limsup_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}\ell\Big(\widehat{Y}_t,Y_t\Big)\right)\leqslant L^\star \qquad \text{p.s.}$$

Comme la stratégie présentée en 1.2.1 vérifie en particulier l'hypothèse (1.9), la consistance est garantie.

1.3 Contexte de la prévision météorologique

Dans cette partie et la suivante, nous présentons le cadre pratique et les résultats de la théorie des suites arbitraires appliquée à des jeux de données météorologiques. Présentons tout d'abord les simulations d'ensemble dans le domaine de la prévision météorologique puis évoquons les aspects plus spécifiques des données traitées. Les éléments de cette présentation sont détaillés dans le chapitre 3.

1.3.1 Description des simulations d'ensemble

La théorie des suites arbitraires, précédemment décrite, a été appliquée lors de cette thèse à deux types de grandeurs physiques : la pression réduite au niveau de la mer et la vitesse du vent au niveau du sol. Voici les protagonistes entrant en jeu dans ces

prévisions.

Prédicteurs élémentaires La théorie des suites arbitraires impose uniquement aux prédicteurs élémentaires $(x_{m,t})_{t\in\mathbb{N}^\star,1\leqslant m\leqslant M}$ qu'ils appartiennent à un espace convexe $\mathcal{X}$. Ceux-ci sont vus comme des boîtes noires, c'est-à-dire des simulations dont la mécanique interne n'entre pas en ligne de compte dans les algorithmes. Concrètement, dans notre cas, il s'agit de M simulations numériques issues de différents centres météorologiques internationaux. Pour simplifier, chacune de ces simulations numériques peut être vue comme une carte entière de prévision météorologique à chaque échéance t. Chacun de ces centres régionaux (canadien, européen, coréen...) fournit un faisceau d'une ou plusieurs dizaines de prévisions qui diffèrent entre elles par leurs conditions initiales (c'est-à-dire l'état de l'atmosphère qui sert de point de départ aux prévisions) et éventuellement leur modèle sous-jacent. En pratique, toutes ces prévisions sont rassemblées et mises à disposition par une initiative internationale commune nommée « *TIGGE* » qui signifie « *the THORPEX Interactive Grand Global Ensemble* ». Au final, les prévisions agrégées sont réalisées à partir d'un ensemble d'au moins 150 prédicteurs élémentaires, aussi appelés membres de l'ensemble. Afin de conserver des poids ayant un sens, les membres de l'ensemble sont triés avant d'être utilisés, ce qui est légitime car ce sont des valeurs scalaires dans chaque cellule (les explications sont détaillées en partie 3.2.1). Cela signifie que le 1^{er} membre est systématiquement le minimum de l'ensemble et le $M^{\text{ème}}$ membre est systématiquement le maximum. Précisons que le tri des prédicteurs est réalisé cellule par cellule et échéance par échéance, de manière indépendante.

Analyse et prévision déterministe L'*analyse* est l'estimation optimale, en un certain sens, de la valeur d'une grandeur physique à une échéance t donnée. Bien que des observations, obtenues aux stations d'observation et par les satellites, soient disponibles, elles sont entâchées d'erreurs. L'analyse est alors le meilleur compromis possible entre, d'une part, ces observations, et d'autre part, une simulation de référence de cette même réalité terrain. Ce sont les analyses successives qui vont remplir le rôle des observations $y_t \in \mathcal{Y}$ de la partie 1.1. La *prévision déterministe*, $(x_t^{\text{det}})_{t\in\mathbb{N}^\star}$, est la prévision de référence, simulée avec une résolution fine et un modèle standard et rodé, dont dispose chaque centre météorologique international. Il s'agit souvent de la simulation dont les prévisions sont les plus précises. Les bulletins météorologiques s'appuient essentiellement sur cette dernière car elle obtient généralement les meilleurs scores en prévision. Nous ajoutons à l'ensemble les prévisions déterministes de Météo France et du Centre européen de prévision (*ECMWF*, *European Centre for Medium-Range Weather Forecasts*). Les prédicteurs élémentaires, l'analyse et la prévision déterministe de Météo France sont rassemblés dans la figure 1.1 qui représente les données en un point de l'espace pour la pression réduite au niveau de la mer.

Limites spatiales, discrétisation et cadre temporel Les caractéristiques suivantes sont communes aux deux grandeurs étudiées. Le domaine spatial s'étend entre les parallèles de latitudes 35° et 61° et entre les méridiens de longitudes −15° et 17°. La résolution horizontale est 0,10° selon les deux dimensions. La période étudiée est de 366 jours depuis le 1^{er} octobre 2011 et jusqu'au 1^{er} octobre 2012 exclu. L'*horizon* des

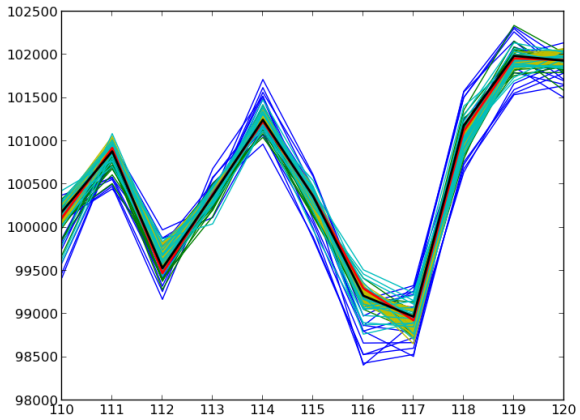

FIGURE 1.1 – Données de pression réduite au niveau de la mer, aux coordonnées $(-13°, 37°)$, entre les échéances 110 (19 janvier 2012) et 120 (29 janvier 2012) de la période temporelle considérée. Un échantillon de 100 membres de l'ensemble parmi les 150 est représenté. En rouge, la prévision déterministe, $x^{\mathrm{det}}_{t,(i,j)}$; en noir, l'analyse, $y_{t,(i,j)}$; le reste des couleurs est dédié aux membres de l'ensemble, $x_{m,\,t,\,(i,j)}$.

prévisions est de 6 ou 18 heures. Notons que les simulations sont disponibles sur une grille, constituée de cellules discrétisant le domaine de calcul. Les algorithmes d'agrégation sont appliqués cellule par cellule : l'indice (i, j) est ajouté le cas échéant. Nous notons $\mathcal{N}$ l'ensemble des cellules présentes sur le domaine, et N leur nombre total.

1.3.2 Application de la théorie des suites arbitraires

Score et perte quadratique La fonction de perte utilisée est la perte quadratique $\ell(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$. Par ailleurs, la fonction de score la plus répandue dans les sciences environnementales est la *RMSE*, qui signifie « *Root Mean Square Error* », et est définie comme

$$RMSE = \sqrt{\frac{1}{TN} \sum_{(i,j)\in\mathcal{N}} \sum_{t=1}^{T} \left(\widehat{y}_{t,\,(i,j)} - y_{t,\,(i,j)}\right)^2}. \tag{1.10}$$

Il existe un lien direct entre la perte quadratique et la *RMSE*, et assurer un regret sous-linéaire implique de maîtriser la *RMSE*. Pour rester cohérent avec le domaine d'application de la théorie, nous fournissons les résultats en termes de *RMSE*. De plus, nous définissons la *différence relative* qui mesure le degré d'amélioration de la prévision agrégée (répérée par l'indice A) par rapport à une prévision de référence sélectionnée au préalable (répérée par l'indice R) :

$$\Delta_{\%}(\mathrm{A}, \mathrm{R}) = \frac{RMSE_{\mathrm{R}} - RMSE_{\mathrm{A}}}{RMSE_{\mathrm{R}}}, \tag{1.11}$$

où $RMSE_{\mathrm{A}}$ est la *RMSE* de la prévision agrégée et $RMSE_{\mathrm{R}}$, celle de la référence en question. Cette différence relative est positive lorsque la *RMSE* de la prévision agrégée

est meilleure que celle de la référence. Par ailleurs, une *période d'entraînement*, déterminée empiriquement, est écartée de l'évaluation afin de fournir le temps de mise en marche de l'algorithme employé.

Performance de l'ensemble, oracle, point de référence Un *oracle* est le minimiseur sur la classe $\mathcal{C}$ de la perte cumulée, soit $\arg\min_{\tilde{y}\in\mathcal{C}} L_T(\tilde{y})$. Par exemple, si $\mathcal{C}$ est la classe contenant tous les prédicteurs élémentaires, alors l'oracle correspondant est le meilleur membre de l'ensemble jusqu'à l'échéance T. Si $\mathcal{C}$ est la classe contenant toutes les combinaisons convexes des prédicteurs élémentaires, l'oracle correspondant est la meilleure combinaison convexe des membres de l'ensemble jusqu'à l'échéance T. Les oracles pourraient apparaître comme des références naturelles pour calculer des différences relatives et les stratégies d'agrégation ont des performances qui s'en approchent sans les battre. Mais afin de comparer les résultats de l'agrégation à une unique prévision qui est employée en pratique, nous déterminons dans chaque étude empirique une prévision de référence, dont le score est bon. Parmi les candidats à cette place de prévision de référence figurent les prévisions déterministes Météo France et *ECMWF* et la moyenne d'ensemble qui est la prévision agrégée dont les poids sont constants et uniformes. Dans les deux cas, pression et vent, la prévision déterministe fournie par Météo France endosse ce rôle de référence et c'est donc celle utilisée dans la formule de différences relatives (1.11).

Algorithme de régression ridge L'algorithme de la *régression ridge* fournit des garanties face à la classe $\mathcal{C}$ de l'ensemble des prévisions linéaires formées à partir des prédicteurs élémentaires. Nous employons l'algorithme de régression *ridge escompté* dans toute la suite : la description précise de l'algorithme est donnée à la partie 3.1.3. Fixons deux paramètres de cet algorithme $\lambda \in \mathbb{R}_+$, le paramètre d'apprentissage, et $\gamma \in \mathbb{R}_+$, le paramètre d'escompte. On choisit un poids initial : $\boldsymbol{u}_1 = (u_{1,1}, \ldots, u_{M,1})^T$. Dans une cellule (i, j) donnée (dont on omet l'indice) et à une échéance t, le calcul des poids linéaires $\boldsymbol{u}_t$ est déterminé en résolvant l'équation de régression suivante :

$$\boldsymbol{u}_t = (u_{1,t}, \ldots, u_{M,t})^T = \arg\min_{\boldsymbol{u}\in\mathbb{R}^M} \left\{ \lambda\|\boldsymbol{u}\|_2^2 + \sum_{s=1}^{t-1}(1+\psi_{t-s}) \left(\sum_{m=1}^{M} u_m\, x_{m,\,s} - y_s \right)^2 \right\},$$

où nous avons noté $\|\cdot\|_2$ la norme euclidienne d'un vecteur de $\mathbb{R}^M$ et où la fonction $\psi_t = \gamma/t^2$ est décroissante. Plus le coefficient γ est grand, plus l'atténuation du passé lointain due à l'escompte est importante. Ce calcul des poids peut être vu comme la résolution d'une régression dont le premier terme est une régularisation proportionnelle à la norme L_2 des poids (qui donne son nom « *ridge* » à l'algorithme) et le second, une fonction de régression qui n'est autre que la perte cumulée du statisticien, modifiée par un terme d'escompte.

Esquisse du plan d'étude Le plan d'étude exhaustif, commun aux deux variables, est présenté à la partie 3.1.6. Il est similaire à celui réalisé par Devaine, Gaillard, Goude et Stoltz [Dev+13, p. 12]). Mentionnons les quatre étapes principales de cette méthode :

1. Présentation du jeu de données et construction des experts.
2. Calcul des oracles et sélection de la prévision de référence.
3. Agrégation avec paramètres fixés, en particulier avec les meilleurs paramètres rétrospectifs.
4. Agrégation réalisée dans des conditions similaires aux conditions opérationnelles, c'est-à-dire avec des paramètres d'agrégation fixés a priori ou adaptés séquentiellement.

La partie suivante, 1.4, explicite les résultats pour les deux variables considérées.

1.4 Résultats empiriques pour les prévisions ponctuelles

1.4.1 Pression réduite au niveau de la mer

La pression réduite au niveau de la mer est la pression atmosphérique observée à la surface du sol et normalisée en fonction de l'altitude du point. Ainsi, les valeurs calculées peuvent être comparées entre elles, quelles que soient les altitudes considérées.

Résultats Le tableau 1.1 rassemble les *RMSE* et les différences relatives associées à plusieurs types de prévision : la moyenne d'ensemble, la prévision déterministe de Météo France, l'oracle convexe, l'oracle linéaire et trois différentes formes d'agrégation ridge escomptée. La première forme est la régression ridge avec un unique jeu de paramètres optimaux rétrospectifs pour l'ensemble des cellules. La seconde régression ridge du tableau permet une adaptation du jeu de paramètres optimaux à chaque cellule. La troisième régression ridge permet une adaptation temporelle du jeu de paramètres dans chaque cellule, et ces paramètres sont cette fois déterminés séquentiellement, comme ils le seraient lors de véritables prévisions. Comme annoncé, la prévision déterministe est la référence pour le calcul de la différence relative. Chacune des trois stratégies ridge escomptée conduit à une différence relative de *RMSE* de plus de 17% ce qui est conséquent. La figure 1.2 rend visible ces bons résultats de l'agrégation ridge avec paramètres optimaux rétrospectifs.

1.4.2 Vitesse du vent à 10 mètres au-dessus du sol

La vitesse du vent est calculée comme le module du vent horizontal (vent méridional et vent zonal), à 10 mètres au-dessus du sol.

Résultats En ce qui concerne la norme de la vitesse du vent à 10 mètres au-dessus du sol, le tableau 1.2 rassemble les *RMSE* et les différences relatives associées à ces types de prévision : la moyenne d'ensemble, la prévision déterministe de Météo France, l'oracle convexe, l'oracle linéaire et deux différentes formes d'agrégation ridge escomptée. Chacune des deux stratégies ridge escomptée conduit à une différence relative de *RMSE* de plus de 6% ce qui est, là encore, intéressant. La figure 1.3 rend visible les résultats de l'agrégation ridge avec paramètres optimaux rétrospectifs.

Type de prévision	*RMSE* (Pa)	Différence relative (%)
Moyenne d'ensemble	52,39	−73,88
Déterministe	30,13	0,00
Oracle convexe	24,29	19,38
Oracle linéaire	18,02	40,18
Agrégation ridge (paramètres optimaux rétrospectifs)	24,90	17,35
Agrégation ridge (adaptation locale rétrospective des paramètres)	24,44	18,87
Agrégation ridge (adaptation locale en ligne sur une grille de paramètres)	24,80	17,69

TABLE 1.1 – Performances de différents oracles et stratégies d'agrégation sur les données de pression réduite au niveau de la mer. Scores : *RMSE* et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures. Dans la troisième et dernière stratégie de l'agrégation, les paramètres peuvent varier tous les $t_{\text{bascule}} = 30$ jours sur une certaine grille de paramètres. Les calculs sont réalisés sur 1000 cellules sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

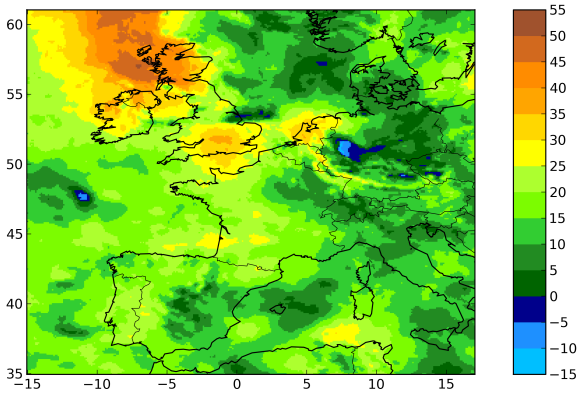

FIGURE 1.2 – Représentation graphique des différences relatives de *RMSE* entre l'agrégé avec paramètres optimaux rétrospectifs et le déterministe Météo France à un horizon de 6 heures pour la pression réduite au niveau de la mer. Dans chaque cellule (i, j) est représentée la différence relative : $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est la *RMSE* temporelle de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'*ECMWF* et de Météo France sont compris). Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

Type de prévision	*RMSE* ($\mathrm{m\,s^{-1}}$)	Différence relative (%)
Moyenne d'ensemble	2,32	−46,96
Déterministe	1,58	0,00
Oracle convexe	1,43	9,45
Oracle linéaire	1,38	12,45
Agrégation ridge (paramètres optimaux rétrospectifs)	1,43	8,95
Agrégation ridge (adaptation locale en ligne sur une grille de paramètres)	1,47	6,48

TABLE 1.2 – Performances de différents oracles et stratégies d'agrégation sur les données de vitesse du vent. *RMSE* et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures. Dans la troisième et dernière stratégie de l'agrégation, les paramètres peuvent varier tous les $t_{\text{bascule}} = 30$ jours sur une certaine grille de paramètres. Les calculs sont réalisés sur 1000 cellules sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

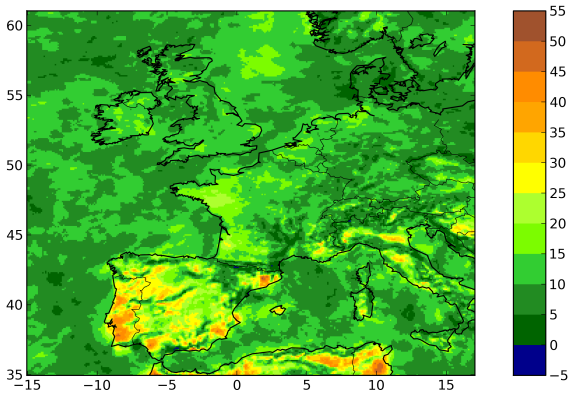

FIGURE 1.3 – Représentation graphique des différences relatives en *RMSE* entre l'agrégé et le déterministe Météo France à un horizon de 6 heures pour la vitesse du vent. Dans chaque cellule (i, j) est représentée la différence relative : $\left(RMSE^{\text{det}}_{(i,j)} - RMSE^{\text{agr}}_{(i,j)}\right)/RMSE^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est la *RMSE* moyenne de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'*ECMWF* et de Météo France sont compris). Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

1.5 Agrégation de fonctions de répartition

Dans les parties précédentes, les applications de la théorie des suites arbitraires mettaient en jeu la prévision de valeurs ponctuelles. Or un domaine en plein essor est la *prévision probabiliste* : nous présentons, au chapitre 5, une application de la théorie des suites arbitraires à l'agrégation de fonctions de répartition. À cette fin, le *CRPS*, une perte entre deux fonctions de répartition quelconques, est définie. Nous proposons alors une prévision probabiliste aux garanties intéressantes au sens de ce score à partir des mêmes informations que dans la partie précédente 1.3.

1.5.1 Aperçu du *continuous ranked probability score*

Ensemble et observations : Les données disponibles en prévision sont identiques à celles de la partie 1.3 : les même prévisions ponctuelles de l'ensemble $x_{m,s}$ et les mêmes analyses y_s. Soit $x \in \mathbb{R}$, la fonction échelon de Heaviside H_x est définie comme : $\mathbb{1}_{[x,+\infty[}$. Si, à chaque prévision élémentaire x_m, nous associons une fonction de répartition échelon, $H_{x_m} = \mathbb{1}_{[x_m,+\infty[}$, nous pouvons les agréger selon un vecteur de poids convexe quelconque $\boldsymbol{p}$:

$$\sum_{m=1}^{M} p_m \, H_{x_m} \, . \tag{1.12}$$

Autrement dit, les fonctions de Heaviside $H_{x_{m,s}}$ jouent désormais le rôle d'experts, à la place de $x_{m,s}$ et nous combinons ces experts par des algorithmes convexes. En effet, une fonction de répartition étant nécessairement croissante et de limite 1 en l'infini, on exclut les algorithmes non-convexes. Afin de comparer cette combinaison à l'échelon d'observation H_y, il est nécessaire de définir une nouvelle fonction de perte : le « *Continuous Ranked Probability Score* », voir partie 5.2.4. Dès lors, il devient possible de déterminer la prévision de référence par rapport à laquelle les prévisions agrégées se compareront, en partie 5.4.1, puis d'adapter et d'évaluer certains algorithmes d'agrégation classiques ou plus récents à la prévision probabiliste en parties 5.3.1 et 5.3.3.

Expression mathématique du *CRPS* : Supposons qu'il existe $\gamma < \Gamma$, tels que $y_s \in [\gamma, \Gamma]$, pour tout $s \in \{1, \dots, T\}$ et $x_{m,s} \in [\gamma, \Gamma]$, pour tout $(m, s) \in \{1, \dots, M\} \times \{1, \dots, T\}$. Pour une échéance s fixée et dans une cellule (i, j) donnée, les fonctions de répartition associées à l'observation H_{y_s} et à la prévision $\sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}$ interviennent dans le *CRPS* instantané selon :

$$CRPS_s^i(\boldsymbol{p}_s) = \int_{-\infty}^{\infty} \left(H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z) \right)^2 dz \, , \tag{1.13}$$

où nous soulignons la dépendance de ce score au vecteur de poids $\boldsymbol{p}_s$. Il s'agit d'une variante probabiliste de la fonction de perte quadratique ponctuelle. Une fois ce score défini, nous employons l'algorithme EG (voir la partie 3), ainsi qu'un algorithme récent « ML-poly-grad », dont la description précise est fournie à la partie 5.3.3. Ces deux algorithmes obtiennent des garanties théoriques de regret sous-linéaire par rapport à l'oracle convexe.

Présentation didactique du *CRPS* : La figure 1.4 propose une représentation de deux prévisions probabilistes distinctes et de leur évaluation dans un cas simple. Supposons que l'on se place dans une cellule donnée à un pas de temps fixé, que l'observation est $y_s = 0.65$, que les prévisions sont à valeurs dans $[0, 1]$ et que l'ensemble fournit les prévisions $\{0, 0.1, \ldots, 0.9, 1\}$. Dans le premier cas (en violet, à gauche), les poids affectés aux membres de l'ensemble sont uniformes, ce qui est visible par l'aspect régulier de la courbe dont toutes les marches sont de même hauteur. Dans le second cas (en rose, à droite), ils sont distincts les uns des autres, tout en restant convexes. Dans chaque graphique, la partie colorée (violette ou rose) correspond à la différence entre la fonction de répartition de l'observation $H_{y_s}(z)$ et de la fonction de répartition de la prévision d'ensemble probabiliste $\sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$. Dans chaque cas, le carré de cette différence est représenté par la fonction en jaune, et le CRPS est alors l'aire du graphique de cette fonction. Au vu de ces exemples, le CRPS dans le cas non-uniforme est plus faible car un accroissement relatif des poids autour de la valeur à prévoir y_s contribue à diminuer cette aire en jaune par rapport à une prévision uniforme.

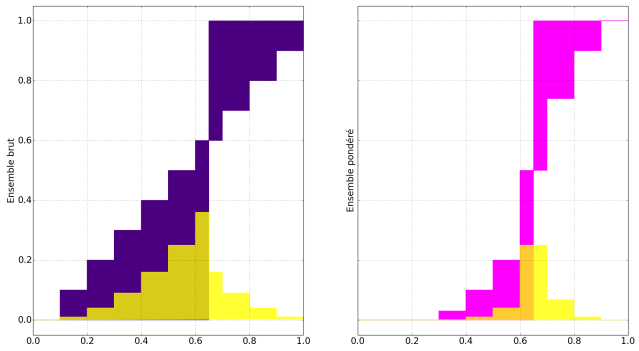

FIGURE 1.4 – Représentation de deux fonctions de répartition et visualisation de leur CRPS respectif par rapport à une observation $y_s = 0.65$. L'aire colorée (violette ou rose) correspond à la fonction $H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$, avec, à gauche, des poids uniformes convexes, et à droite, des poids non uniformes. La surface superposée en jaune dans chaque figure correspond au carré de la valeur précédente. Son aire est donc la valeur du $CRPS_s^i$.

Formulation générale *CRPS* : De manière plus générale, le *CRPS* entre deux fonctions de répartition F et G quelconques est défini comme :

$$CRPS(F, G) = \int_{-\infty}^{\infty} (F(z) - G(z))^2 \, dz \,. \tag{1.14}$$

Pour des fonctions de répartition de la variable réelle z, il existe une formulation équivalente sous forme de *distance énergétique*, qui est la racine carrée de la quantité suivante :

$$D^2(F, G) = 2\mathbb{E}\Big[||X - Y||\Big] - \mathbb{E}\Big[||X - X'||\Big] - \mathbb{E}\Big[||Y - Y'||\Big] > 0, \tag{1.15}$$

où X et X' (respectivement Y et Y') sont des variables aléatoires indépendantes, identiquement distribuées de fonction de répartition F (respectivement G) et $||\cdot||$ désigne

la norme euclidienne. D définit bien une distance entre F et G, elle vérifie l'équivalence $D(F,G) = 0$ si et seulement si $F = G$. L'article de Székely et Rizzo [SR05b] donne une preuve, en dimension un, de l'égalité de la distance énergétique et du *CRPS* (appelé aussi distance de Harald Cramér). En dehors de la prévision probabiliste en météorologie (Gneiting et Raftery [GR07]), cette distance énergétique trouve des applications dans de nombreux domaines des statistiques : méthodes à noyau (Sejdinovic, Sriperumbudur, Gretton et Fukumizu [Sej+13]), classification hiérarchique (Szekely et Rizzo [SR05a]) et tests statistiques d'adéquation de deux distributions dans des espaces métriques généraux de dimensions quelconques (Székely et Rizzo [SR13]).

1.5.2 Résultats empiriques

Après recherche du meilleur paramètre d'apprentissage *a posteriori* dans chaque cas, les résultats empiriques sont rassemblés au tableau 1.3 et à la figure 1.5 pour la pression réduite au niveau de la mer et au tableau 1.4 et à la figure 1.6 pour la vitesse du vent. Dans les deux cas, à une échéance t fixée, la prévision de référence est l'échelon dont le seuil est la prévision déterministe proposée par Météo France x_t^d et la prévision d'ensemble correspond à la fonction de répartition uniforme associée aux points $\left\{x_{m,t} : 1 \leqslant m \leqslant M\right\}$. Les améliorations, en terme de différence relative par rapport à la prévision de référence, sont conséquentes : dans le cas de $\mathcal{E}_\eta^{\text{grad}}$, dont le paramètre η est calibré manuellement afin de sélectionner la meilleure prévision possible avec ce paramètre rétrospectif, 33,4 % pour la pression et 24,1 % pour la vitesse du vent, et dans le cas de l'algorithme automatique ML-poly-grad, 17,6 % et 13,1 %, respectivement.

Type de prévision	CRPS (Pa)	différence relative (%)
Ensemble *TIGGE*	27,98	−19,67
Ensemble *ECMWF*	32,77	−40,15
Échelon déterministe	23,38	0,00
Poids exponentiels (avec paramètre optimal a posteriori)	15,58	33,36
ML-poly-grad	19,27	17,60

Table 1.3 – Performances de différents oracles et stratégies d'agrégation sur les données de pression réduite au niveau de la mer. CRPS : Scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision probabiliste agrégée à un horizon de 6 heures. Le poids initial est uniforme dans les méthodes d'agrégation. La période d'entraînement est de 100 jours. Ces scores sont évalués sur l'intégralité de la carte.

1.6 Perspectives

Dans le cas de l'arbre de régression déterministe, certains cas de figure généraux n'ont pas été traités. Les observations et prévisions appartiennent par hypothèse à un ensemble $[0,1]$ et ce segment a pu être généralisé à tout segment borné quelconque. Mais,

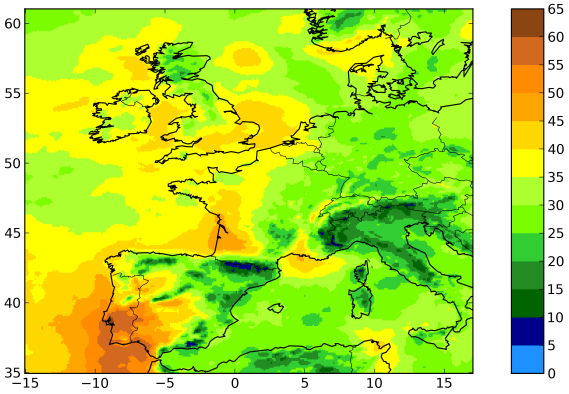

FIGURE 1.5 – Représentation graphique des différences relatives du CRPS entre la fonction de répartition agrégée par les poids exponentiels $\mathcal{E}_\eta^{\text{grad}}$ et l'échelon déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i,j) est représentée la différence relative : $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est le CRPS moyen de la prévision agrégée (respectivement déterministe) dans la cellule (i,j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les prévisions déterministes d'*ECMWF* et de Météo France sont compris). Les poids initiaux sont uniformes. La période d'entraînement est de 100 jours.

Type de prévision	CRPS ($\mathrm{m\,s^{-1}}$)	différence relative (%)
Ensemble *TIGGE*	1,15	−36,57
Ensemble *ECMWF*	1,41	−66,85
Échelon déterministe	0,84	0,00
Poids exponentiels (avec paramètre optimal a posteriori)	0,64	24,12
ML-poly-grad	0,73	13,07

TABLE 1.4 – CRPS : Scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision probabiliste agrégée à un horizon de 6 heures. Les poids initiaux sont uniformes. La période d'entraînement est de 100 jours. 1000 cellules sont sélectionnées aléatoirement pour évaluer ces scores.

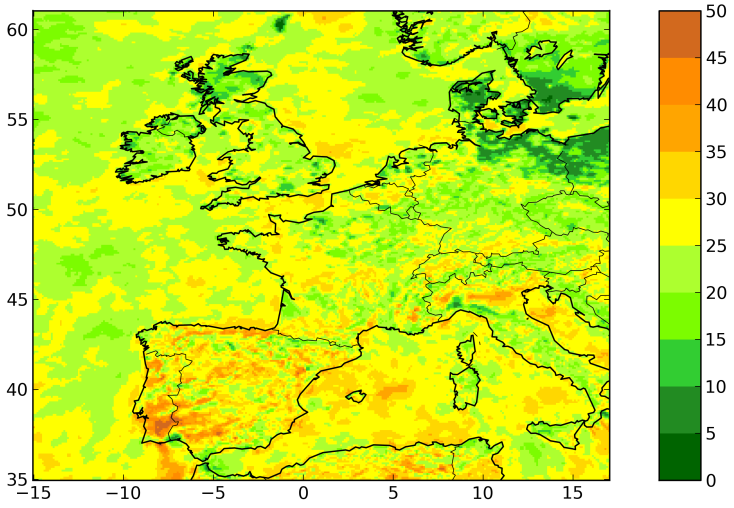

FIGURE 1.6 – Performances de différents oracles et stratégies d'agrégation sur les données de vitesse du vent. Représentation graphique des différences relatives du CRPS entre la fonction de répartition agrégée par les poids exponentiels $\mathcal{E}_{\eta}^{\text{grad}}$ et l'échelon déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i, j) est représentée la différence relative : $\left(r_{(i,j)}^{\text{det}} - r_{(i,j)}^{\text{agr}}\right)/r_{(i,j)}^{\text{det}}$ où $r_{(i,j)}^{\text{agr}}$ (respectivement $r_{(i,j)}^{\text{det}}$) est le CRPS moyen de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les prévisions déterministes d'*ECMWF* et de Météo France sont compris). Les poids initiaux sont uniformes. La période d'entraînement est de 100 jours.

nous avons laissé de côté le cas non-borné qui rendrait l'algorithme plus général encore.

L'application de la théorie des suites arbitraires à la météorologie fournit des prévisions dont la qualité surpasse quasiment systématiquement celle de la prévision de référence. Pour apprécier le gain réel apporté par les algorithmes d'agrégation séquentielle, il faut savoir qu'une différence relative de l'ordre de 15 ou 5 % (respectivement dans le cas de la pression réduite au niveau de la mer et de la vitesse du vent), telle que celle obtenue par l'agrégation ridge escomptée est à comparer à une différence relative de 1 ou 2 % entre un modèle météorologique et sa version améliorée de l'année suivante. Par conséquent, mettre en place l'algorithme de régression ridge en prévision opérationnelle semble logiquement l'étape suivante à réaliser.

En ce qui concerne la prévision avec incertitudes, les conclusions sont similaires à celles de la partie précédente. Ainsi, les travaux réalisés sur les données concrètes montrent que les algorithmes d'agrégation séquentielle sont performants pour combiner les membres de l'ensemble au sens du *CRPS*. Une perspective possible serait donc là encore d'appliquer les méthodes décrites dans cette thèse à la prévision opérationnelle de fonctions de répartition. Par ailleurs, en considérant que la fonction de répartition de chaque observation y_s est une fonction de Heaviside H_{y_s}, nous ignorons sciemment l'incertitude sur les observations, dont certaines estimations sont disponibles. Proposer des versions alternatives de fonctions de répartitions (fonction rampe ou fonction de répartition de loi normale sous-jacente par exemple) améliorerait la modélisation générale du problème de prévision et par conséquent la précision des prévisions probabilistes issues des algorithmes.

2 Deterministic Regression Tree

We study online prediction of bounded stationary ergodic processes. To do so, we consider the setting of prediction of individual sequences and propose a deterministic regression tree that performs asymptotically as well as the best L-Lipschitz predictor. Then, we show why the obtained regret bound entails the asymptotical optimality with respect to the class of bounded stationary ergodic processes.

Ce chapitre est composé de l'article « Deterministic Regression Tree », écrit avec Pierre Gaillard (GREGHEC, HEC Paris, CNRS ; EDF R&D).

Sommaire

2.1 Introduction

Consider that at each time step $t = 1, 2, \dots$, a learner is asked to form a prediction $\widehat{Y}_t$ of the next outcome $Y_t \in [0,1]$ of a bounded stationary ergodic process $(Y_t)_{t=-\infty,\dots,\infty}$ with knowledge of the past observations $Y_1, \dots, Y_{t-1}$. To evaluate the performance, a convex and M-Lipschitz (with respect to its first argument) loss function $\ell : [0,1]^2 \to [0,1]$ is considered. But, for any bounded and continuous loss functions ℓ, Algoet [Alg94] proved the following fundamental limit. For any prediction strategy, almost surely

$$\liminf_{T\to\infty} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell(\widehat{Y}_t, Y_t) \right\} \geqslant L^\star ,$$

where

$$L^\star = \mathbb{E}\Big[\inf_{f \in \mathcal{B}^\infty} \mathbb{E}\Big[\ell(f(Y_{-\infty}^{-1}), Y_0) | Y_{-\infty}^{-1} \Big] \Big]$$

is the expected minimal loss over all possible Borelian estimations of the outcome Y_0 based on the infinite past ($\mathcal{B}^\infty$ denotes the set of Borelian functions from $[0,1]^\infty$ to $[0,1]$). One may thus try to design *consistent* strategies that achieve the lower bound, that is,

$$\limsup_{T\to\infty} \left\{ \frac{1}{T} \sum_{t} \ell(\widehat{Y}_t, Y_t) \right\} \leqslant L^\star . \tag{2.1}$$

Our approach. To do so, we partition the analysis and the design into two separate layers: the setting of individual sequences and the one of stochastic time series. In Sections 2.2 and 2.3, we adopt first the point of view of individual sequences, where no stochastic assumption about the underlying process that generates the data is made, see the monograph of Cesa-Bianchi and Lugosi [CL06]. Only Section 2.4 assumes that the data comes from a stationary ergodic process and states that any strategy that satisfies some deterministic regret bound is consistent. Sections 2.2, 2.3, and 2.4 can be read independently.

Formally, our framework(s) is (are) the following. The setting of individual sequences (Section 2.2) assumes that a sequence $(\boldsymbol{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ is observed step by step, where $\mathcal{X} \subset [0,1]^d$ is the covariable space and $\mathcal{Y} \subset [0,1]$ a convex observation space*. The learner is asked at each time step t to predict the next observation y_t with knowledge of the past observations $y_1, \dots, y_{t-1}$ and of the past and present exogenous variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_t$. The accuracy of a prediction is measured by a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to [0,1]$ that we assume to be convex and M-Lipschitz in its first argument (typically the square loss, the pinball loss, ...). Given a d-dimensional input space $\mathcal{X}$, the goal of the forecaster is to minimize its cumulative regret against the classes $\mathcal{L}_L^d$ of L-Lipschitz functions with respect to the ℓ^2-norm (for all fixed $L > 0$†) from $[0,1]^d$ to $[0,1]$,

$$\widehat{R}_{L,T} = \sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^{T} \ell(f(\boldsymbol{x}_t), y_t) ,$$

*In Section 2.3, $\boldsymbol{x}_t$ will be replaced by $y_{t-d}^{t-1} = y_{t-d}, \dots, y_{t-1}$, then, the deterministic elements y_t will be replaced by random variables Y_t in Section 2.4.

†The strategy of the forecaster is independent of the Lipschitz constant L, which is not known.

that is, to ensure

$$\forall L > 0\,, \qquad \limsup_{T\to\infty} \left\{ \sup_{(\boldsymbol{x}_1,y_1),\dots,(\boldsymbol{x}_T,y_T)} \frac{\widehat{R}_{L,T}}{T} \right\} \leqslant 0\,.$$

Section 2.4 adresses actually the more challenging goal of competing against all Borelian functions.

Section 2.2 designs such a strategy. The nested EG strategy (Algorithm 5) follows the spirit of binary regression trees like CART (see Breiman, Friedman, Stone, and Olshen [Bre+84]), by performing a data-driven partition of the covariable space. Theorem 2.2.3 guarantees that the nested EG strategy satisfies for all $T \geqslant 1$ and all $L > 0$,

$$\widehat{R}_{L,T} \leqslant M(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,, \tag{2.2}$$

in the worst case, that is, for all possible values of $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_T, y_T)$.

In Section 2.3, we switch to an autoregressive setting: previous responses $y_{t-d}^{t-1} = (y_{t-d}, \dots, y_{t-1})$ are used as covariates to predict y_t. Basically, $\boldsymbol{x}_t$ is replaced by y_{t-d}^{t-1}. The number d of previous responses is the order of the autoregressive model. For each order $d \geqslant 1$, Inequality (2.2) remains valid. The challenge of Section 2.3 is to obtain the result simultaneously for all orders $d \geqslant 1$. This is done (Algorithm 6) by combining an increasing number of fixed order-d nested EG. Theorem 2.3.2 upper-bounds the regret of Algorithm 6 for all $d \leqslant T$, and for all $y_1, \dots, y_T \in [0,1]$,

$$\sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) \leqslant \inf_{f\in\mathcal{L}_L^d} \sum_{t=d+1}^{T} \ell\Big(f(y_{t-d}^{t-1}), y_t\Big) + \sqrt{(T+1)\log(T+1)} \\ + M(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,.$$

Consequently, for all $d \geqslant 0$ and all $L > 0$, and for all $y_1, \dots, y_T \in [0,1]$,

$$\limsup_{T\to\infty} \left\{ \frac{1}{T} \sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) \right\} \leqslant \limsup_{T\to\infty} \left\{ \inf_{f\in\mathcal{L}_L^d} \frac{1}{T} \sum_{t=d+1}^{T} \ell\Big(f(y_{t-d}^{t-1}), y_t\Big) \right\}\,. \tag{2.3}$$

Finally, we assume in Section 2.4 that the sequence of observations $y_1, \dots, y_T$ actually comes from a stationary ergodic process (Y_t): y_t is replaced by Y_t. Our main result is Theorem 2.4.1 which states that any strategy achieving (2.3) is actually consistent, i.e., satisfies (2.1).

Literature review. Many consistent forecasting strategies have been designed to achieve (2.1). The vast majority of these strategies are based on statistical techniques used for time-series prediction, ranging from parametric models like autoregressive models (see Brockwell and Davis [BD91]) to non-parametric methods (see the surveys of Györfi, Härdle, Sarda, and Vieu [Gyö+89], Bosq [Bos96], and Merhav and Feder [MF98]). In recent years, another collection of algorithms resolving related problems have been designed in Györfi, Lugosi, and Fargas [GLF01], Gyorfi and Ottucsak [GO07],

Biau, Bleakley, Györfi, and Ottucsák [Bia+10], and Biau and Patra [BP11]. At their cores, all these algorithms use some machine learning non-parametric prediction scheme (like histogram, kernel, or nearest neighbor estimation). These algorithms rely on two parameters: a window parameter h taking values in a discret countable set $\mathcal{H}$ (e.g., the number of the bins of the uniform histograms or the number of neighbors), and the order d of the autoregressive model to consider. Then, they output predictions by mixing the countably infinite set of experts indexed by $(h, d) \in \mathcal{H} \times \mathbb{N}^\star$ corresponding to strategies with fixed values of these two parameters. To do so, a prior distribution π on this infinite set of experts $\mathcal{H} \times \mathbb{N}$ has to be considered. All these studies only perform asymptotical analysis and thus only require that the prior π assign positive weight to each expert (h, d) without taking into account the complexity of the experts. The practical choice of π is however left to the user and not calibrated online. No finite-time analysis is performed. Besides, computational purposes require to consider nothing but finite grids in numerical experiments, which results in practice in approximation of the algorithm studied theoretically. In all these studies, assumption are made from the beginning about the underlying process generating the sequence of observations. In our method, consistency only comes as an additional feature of the algorithm via Theorem 2.4.1 in the special case of stochastic time series.

In contrast to statistics, little research has been devoted to online non-parametric prediction in the setting of individual sequences. Recently, Rakhlin and Sridharan [RS14] provided the first optimal rates for the square loss for arbitrary classes of regression functions. Surprisingly, these rates match the ones for statistical learning with square loss. However, the algorithm provided is generally not computationally feasible. Vovk [Vov05; Vov06] considered vast classes of regression function (such as Besov space) and proposed two approaches: defensive forecasting that uses the convexity of the functional space and perform directly a method of defensive forecasting (such as gradient descent), and another one based on aggregating a set of basis functions, that approximates well the space. However, Vovk does not explains how to build efficiently the set of basis functions to aggregate, and how to calibrate the precision of approximation. Our method nested EG can be seen as a data-driven procedure to do so in the particular case where the comparison class are Lipschitz functions. Theorem 2.4.1 provides the first link between this literature, and the statistical goal of getting consistent strategies that achieve (2.1).

Contributions. First, we clean up the standard analysis of prediction of ergodic processes by carrying out the aforementioned separation in two layers: first we perform a worst-case deterministic analysis, then in the particular case where data comes from stationary ergodic process, consistency is obtained as a direct consequence via Theorem 2.4.1. Our second main contribution is to build an efficient data-driven and fully automatic procedure (nested EG) to compete against Lipschitz functions. Our forecaster comes with robust (worst-case) finite-time guarantees. In Section 2.2.3, we consider two simulation studies which confirm that nested EG performs much better than simpler methods like uniform histograms. Our algorithm, which is purely sequential (unlike nearest neighbor estimation for instance), is computationally efficient in contrast to the one of Rakhlin and Sridharan [RS14]. Its time complexity can indeed be chosen arbitrarily close to linear in the number of time steps. A last benefit of our

approach is to be valid for a general class of loss functions when previous papers to our knowledge only treat particular cases like the square loss or the pinball loss.

2.2 A strategy that competes against Lipschitz functions

The nested EG strategy (Algorithm 5) incrementally builds an estimate of the best Lipschitz function (denoted by $f^\star$) with respect to the ℓ^2-norm. The core idea is to estimate $f^\star$ precisely in areas of the covariable space $\mathcal{X}$ with many occurrences of covariables $\boldsymbol{x}_t$, while estimating it loosely in other parts of the space. To implement this idea, Algorithm 5 maintains a deterministic binary tree whose nodes are associated with regions of the covariable space, such that the regions with nodes deeper in the tree represent increasingly smaller subsets of $\mathcal{X}$ (see Figure 2.1).

In the sequel, we assume for simplicity that $\mathcal{X} = [0,1]^d$ and $\mathcal{Y} = [0,1]$ and that the loss function ℓ is from $[0,1]^2$ to $[0,1]$. The case of unknown bounded sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ will be treated later in remarks.

This section is divided into three subsections. Section 2.2.2 is the core of the algorithm. It explains how to partition the space of covariables $\boldsymbol{x}_t$ in a clever data-driven fashion. Inside each region of the space, the algorithm estimates the best constant prediction by running a defensive forecasting. Section 2.2.1 and Section 2.2.1 control the error suffered by the defensive algorithm inside each region: Section 2.2.1 bounds the approximation error of a Lipschitz function by a constant, and Section 2.2.1 controls the estimation error of learning this best constant online.

2.2.1 Perfoming almost as well as the best constant

Approximation error of the best constant

If the number of observations such that $\boldsymbol{x}_t$ belong to a subset $\mathcal{X}^{\text{node}} \subset \mathcal{X}$ is small enough, one does not need to estimate $f^\star$ precisely over $\mathcal{X}^{\text{node}}$. Lemma 2.2.1 formalizes this idea by controlling the approximation error suffered by approximating $f^\star$ by the best constant in $[0,1]$. The control is expressed in terms of the number of observations T^{node} and of the size of the set $\mathcal{X}^{\text{node}}$, which is measured by its diameter defined as $\operatorname{diam}(\mathcal{X}^{\text{node}}) = \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}^{\text{node}}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$.

Lemma 2.2.1 (Approximation error)**.** *Let $T^{\text{node}} \geqslant 1$ and suppose that ℓ is M-Lipschitz in its first argument. Then, for all $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_{T^{\text{node}}}, y_{T^{\text{node}}}) \in [0,1]^2$ and all $L > 0$, the cumulative loss of the best constant is upper bounded as*

$$\inf_{y \in [0,1]} \sum_{t=1}^{T^{\text{node}}} \ell(y, y_t) \leqslant \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^{T^{\text{node}}} \ell\big(f(\boldsymbol{x}_t), y_t\big) + MLT^{\text{node}} \operatorname{diam}(\mathcal{X}^{\text{node}}),$$

where $\mathcal{X}^{\text{node}} \subset [0,1]^d$ is such that $\boldsymbol{x}_t \in \mathcal{X}^{\text{node}}$ for all $t = 1, \dots, T^{\text{node}}$.

Proof. Let $t \geqslant 1$. Using that ℓ is M-Lipschitz and f is L-Lipschitz with respect to the

ℓ^2 norm, we get

$$\begin{aligned}\ell(f(\boldsymbol{x}_1), y_t) - \ell(f(\boldsymbol{x}_t), y_t) &\leqslant M|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_t)| \\ &\leqslant ML\|\boldsymbol{x}_1 - \boldsymbol{x}_t\|_2 \\ &\leqslant ML\,\mathrm{diam}(\mathcal{X}^{\mathrm{node}})\,.\end{aligned}$$

Summing over t and noting that $\inf_y \sum_t \ell(y, y_t) \leqslant \sum_t \ell\big(f(\boldsymbol{x}_1), y_t\big)$ concludes. □

Estimation error of the best constant online

Parameter: $M > 0$

For time step $t = 1, 2, \ldots$

1. Define the learning parameter $\eta_t = M^{-1}\sqrt{(\log 2)/t}$
2. Predict
$$\widehat{y}_t = \frac{\exp\Big(-\eta_t \sum_{s=1}^{t-1} \ell'(\widehat{y}_s, y_s)\Big)}{1 + \exp\Big(-\eta_t \sum_{s=1}^{t-1} \ell'(\widehat{y}_s, y_s)\Big)} \in [0, 1]\,,$$
where ℓ' denotes the (sub)gradient of ℓ with respect to its first argument
3. Observe y_t

Algorithm 4: The gradient-based exponentially weighted average forecaster (EG) with two constant experts that predict respectively 0 and 1.

Lemma 2.2.1 implies that considering constant predictions is not bad when either the covariable region is small, or the number of observations is small. The next step consists thus of estimating online the best constant prediction in $[0, 1]$.

To do so, among many existing methods, we consider the well-known *gradient-based exponentially weighted average forecaster* (EG), introduced by Kivinen and Warmuth [KW97]. In the setting of prediction of individual sequences with expert advice—see the monograph by Cesa-Bianchi and Lugosi [CL06], EG competes with the best fixed convex combination of experts. In the case where two experts predict constant predictions respectively 0 and 1 at all time steps, EG ensures vanishing average regret with respect to any constant prediction in $[0, 1]$. Algorithm 4 implements this particular case of EG and Lemma 2.2.2 provides the associated regret bound. Lemma 2.2.2 is a particular case of the standard regret bound of EG, whose proof is available for instance in Cesa-Bianchi and Lugosi [CL06].

Lemma 2.2.2 (Estimation error)**.** *Let $T^{\mathrm{node}} \geqslant 1$. We assume that the loss function ℓ is convex and M-Lipschitz in its first argument. Then, for all $y_1, \ldots y_{T^{\mathrm{node}}} \in [0, 1]$, the*

cumulative loss of Algorithm 4 is upper bounded as follows:

$$\sum_{t=1}^{T^{\mathrm{node}}} \ell(\widehat{y}_t, y_t) \leqslant \inf_{y\in[0,1]} \sum_{t=1}^{T^{\mathrm{node}}} \ell(y, y_t) + 2M\sqrt{T^{\mathrm{node}} \log 2}\,.$$

Unknown value of M. Note that Algorithm 4 needs to know in advance a uniform bound M on ℓ'. This is the case if one considers (as we do) a bounded observation space $[0,1]$ with the absolute loss function, defined for all $y, y' \in [0,1]$ by $\ell(y', y) = |y - y'|$; the pinball loss, defined by $\ell_\alpha(y', y) = (\alpha - \mathbb{1}_{\{y \geqslant x\}})(y - y')$; or the square loss, defined by $\ell(y', y) = (y - y')^2$. However, in the case of an unknown observation space $\mathcal{Y}$ the bound on the gradient of the square loss is unknown and needs to be calibrated online at the small cost of the additional term $2M(2 + 4(\log 2)/3)$ in the regret bound, see Rooij, van Erven, Grünwald, and Koolen [Roo+14].

2.2.2 Performing almost as well as the best Lipschitz function: the nested EG strategy

The nested EG strategy (Algorithm 5) implements both the ideas of Lemma 2.2.1 and Lemma 2.2.2. It maintains a binary tree whose nodes are associated with regions of the covariable space $[0,1]^d$. The nodes in the tree are indexed by pairs of integers (h, i); where the first index $h \geqslant 0$ denotes the distance of the node to the root (also referred to as the depth of the node) and the second index i belongs to $\{1, \dots, 2^h\}$. The root is thus denoted by $(0, 1)$. By convention, $(h+1, 2i-1)$ and $(h+1, 2i)$ are used to refer to the two children of node (h, i). Let $\mathcal{X}^{(h,i)} \subset [0,1]^d$ be the region associated with node (h, i). By construction, these regions are hyper-rectangles and satisfy the constraints

$$\mathcal{X}^{(0,1)} = [0,1]^d \qquad \text{and} \qquad \mathcal{X}^{(h,i)} = \mathcal{X}^{(h+1,2i-1)} \sqcup \mathcal{X}^{(h+1,2i)}\,,$$

where $\sqcup$ denotes the disjoint union. The set of regions associated with terminal nodes (or leaves) thus forms a partition of $[0,1]^d$.

At time step t, when a new covariable $\boldsymbol{x}_t$ is observed, Algorithm 5 first selects the associated leaf (h_t, i_t) such that $\boldsymbol{x}_t \in \mathcal{X}^{(h_t,i_t)}$ (step 2). The leaf (h_t, i_t) then predicts the next observation y_t by updating a local version of Algorithm 4 (step 3). Namely, for node (h_t, i_t), Algorithm 5 runs Algorithm 4 on the sub-sequence of

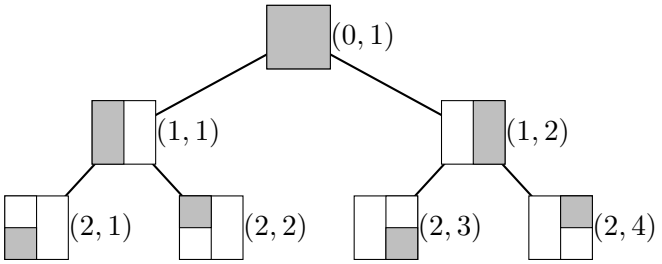

Figure 2.1: Representation of the binary tree in dimension $d = 2$. In this case, the regions associated to nodes $(0, 1)$ is $\mathcal{X}^{(0,1)} = [0,1]^2$ and to node $(2, 3)$ is $\mathcal{X}^{(2,3)} = [0.5, 1] \times [0, 0.5]$.

past observations $(\boldsymbol{x}_s, y_s)$ such that the associated leaf is (h_t, i_t), that is, $E^{(h_t,i_t)} \triangleq \{1 \leqslant s \leqslant t-1 : (h_s, i_s) = (h_t, i_t)\}$; then it forms the prediction (step 3) and finally updates the set of observations predicted by node (h_t, i_t): $E^{(h_t,i_t)} = E^{(h_t,i_t)} \cup \{t\}$ (step 4). When the number of observations $T^{(h_t,i_t)} \triangleq \#E^{(h_t,i_t)}$ received and predicted by leaf (h_t, i_t) becomes too large compared to the size of the region $\mathcal{X}^{(h_t,i_t)}$ (step 5), the tree is updated. To do so, the region $\mathcal{X}^{(h_t,i_t)}$ is divided in two sub-regions of equal volume by cutting along one given coordinate.

The coordinate $r_t + 1$ to be split is chosen in a deterministic order, where $r_t = (h_t \bmod d)$ and $\bmod$ denotes the modulo operation. Thus, at the root node $(0, 1)$ the first coordinate is split, then by going down in the tree we split the second one, then the third one and so on until we reach the depth d, in which case we split the first coordinate for the second time. Each sub-region is associated with a child of node (h_t, i_t). Consequently, (h_t, i_t) becomes an inner node and is thus no longer used to form predictions.

To facilitate the formal study of the algorithm, we will need some additional notation. In particular, we will introduce time-indexed versions of several quantities. $\mathcal{T}_t$ denotes the tree stored by Algorithm 5 at the beginning of time step t. The initial tree is thus the root $\mathcal{T}_0 = \{(0, 1)\}$ and it is expanded when the splitting condition (step 5) holds, as

$$\mathcal{T}_{t+1} = \mathcal{T}_t \cup \big\{(h_t + 1, 2i_t - 1), (h_t + 1, 2i_t)\big\}$$

(step 5.2.3) and remains unchanged otherwise. We denote by N_t the number of nodes of $\mathcal{T}_t$ and by H_t the height of $\mathcal{T}_t$, that is, the maximal depth of the leaves of $\mathcal{T}_t$. A performance bound for Algorithm 5 is provided below.

Theorem 2.2.3. *Let $T \geqslant 1$ and $d \geqslant 1$. Then, the cumulative regret $\widehat{R}_{L,T}$ of Algorithm 5 is upper bounded as*

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^{T} \ell\big(f(\boldsymbol{x}_t), y_t\big) \leqslant M\,(L+3)\,\sqrt{N_T T}$$
$$\leqslant M(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,.$$

The proof of Theorem 2.2.3 is deferred to Section 2.5.1.

Time and storage complexity. The following lemma, whose proof is postponed to Section 2.5.1, provides time and storage complexity guarantees for Algorithm 5. It upper bounds the maximal size of $\mathcal{T}_T$, that is, its number of nodes N_T and its depth H_T, which yields in particular the regret bound of order $O\big(T^{(d+1)/(d+2)}\big)$ stated in Theorem 2.2.3.

Lemma 2.2.4. *Let $T \geqslant 1$ and $d \geqslant 1$. Then the depth H_T and the number of nodes N_T of the binary tree $\mathcal{T}_T$ stored by Algorithm 5 after T time steps are upper bounded as follows:*

$$H_T \leqslant 1 + \frac{d}{2}\log_2(4dT) \qquad \text{and} \qquad N_T \leqslant 1 + 8\,(dT)^{\frac{d}{d+2}}\,.$$

Initialization:

- $\mathcal{T} = \{(0,1)\}$ a tree (for now reduced at a root node)
- Define the bin $\mathcal{X}^{(0,1)} = [0,1]^d$ and the corresponding set of points $E^{(0,1)} = \emptyset$.

For $t = 1, \dots, T$

1. Observe $\boldsymbol{x}_t \in [0,1]^d$
2. Select the leaf (h_t, i_t) such that $\boldsymbol{x}_t \in \mathcal{X}^{(h_t,i_t)}$
3. Predict
$$\widehat{y}_t = \frac{\exp\Big(-\eta^{(h_t,i_t)} \sum_{s \in E^{(h_t,i_t)}} \ell'(\widehat{y}_s, y_s)\Big)}{1 + \exp\big(-\eta^{(h_t,i_t)} \sum_{s \in E^{(h_t,i_t)}} \ell'(\widehat{y}_s, y_s)\big)} \in [0,1]\,,$$
where $\eta^{(h_t,i_t)} = \sqrt{(\log 2)/(\#E^{(h_t,i_t)} + 1)}$.
4. Update the set of observations predicted by node (h_t, i_t)
$$E^{(h_t,i_t)} \leftarrow \{1 \leqslant s \leqslant t, \quad (h_s, i_s) = (h_t, i_t)\}$$
5. **If** the splitting condition $\#E^{(h_t,i_t)} + 1 \geqslant \Big(\mathrm{diam}(\mathcal{X}^{(h_t,i_t)})\Big)^{-2}$ holds **then** extend the binary tree $\mathcal{T}$ as follows:
   5.1. Compute the decomposition $h_t = k_t d + r_t$ with $r_t \in \{0, \dots, d-1\}$
   5.2. Split coordinate $r_t + 1$ for node (h_t, i_t)
      5.2.1. Define the splitting threshold $\tau = (x^- + x^+)/2$, where $x^- = \inf_{\boldsymbol{x} \in \mathcal{X}^{(h_t,i_t)}} \{x_{r_t+1}\}$ and $x^+ = \sup_{\boldsymbol{x} \in \mathcal{X}^{(h_t,i_t)}} \{x_{r_t+1}\}$.
      5.2.2. Define two children leaves for node (h_t, i_t):
      - the left leaf $(h_t + 1, 2i_t - 1)$ with corresponding bin
      $$\mathcal{X}^{(h_t+1,2i_t-1)} = \{\boldsymbol{x} \in \mathcal{X}^{(h_t,i_t)} \ : \ x_{r_t+1} \in [x^-, \tau[\}$$
      - the right leaf $(h_t + 1, 2i_t)$ with corresponding bin
      $$\mathcal{X}^{(h_t+1,2i_t-1)} = \left\{ \boldsymbol{x} \in \mathcal{X}^{(h_t,i_t)} \ : \ \begin{array}{ll} x_{r_t+1} \in [\tau, x^+[& \text{if } x_+ < 1 \\ x_{r_t+1} \in [\tau, 1] & \text{if } x_+ = 1 \end{array} \right\}$$
      5.2.3. Initialize their sets of observations
      $$E^{(h_t+1,2i_t-1)} = E^{(h_t+1,2i_t-1)} = \emptyset.$$
      5.2.3. Update $\mathcal{T} \leftarrow \mathcal{T} \cup \{(h_t + 1, 2i_t - 1), (h_t + 1, 2i_t)\}$

Algorithm 5: Sequential prediction of function via Nested EG

Indeed, Algorithm 5 needs to store a constant number of parameters at each node of the tree. Thus the space complexity is of order $O(N_T) = O(T^{d/(d+2)})$. Besides at each time step t, Algorithm 5 needs to perform $O(H_t) = O(\log t)$ binary test operations in order to select the leaf (h_t, i_t). It then only needs constant time to update both the local version of Algorithm 4 associated with node (h_t, i_t) and the tree $\mathcal{T}$. Thus the per-round time complexity of Algorithm 5 is of order $O(\log t)$ and the global time complexity is of order $O(T \log T)$. Therefore, we can summarize:

$$\text{Storage complexity:} \quad O(T^{d/(d+2)})\,, \qquad \text{Time complexity:} \quad O(T \log T)\,.$$

Unknown bounded sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. As we mentioned in the end of Section 2.2.1, the generalization of Algorithm 4 and thus of Algorithm 5 to an unknown set $\mathcal{Y} \subset \mathbb{R}$ can be obtained by using standard tools of individual sequences—see for instance Rooij, van Erven, Grünwald, and Koolen [Roo+14]. To adapt Algorithm 5 to any unknown compact set $\mathcal{X} \subset \mathbb{R}^d$, one can first divide the covariable space $\mathbb{R}^d$ in hyper-rectangle subregions of the form $[n_1, n_1 + 1] \times \cdots \times [n_d, n_d + 1]$ and then run independent versions of Algorithm 5 on all of these subregions. If $\operatorname{diam}(\mathcal{X}) \leqslant \sqrt{d}B$ with an unknown value of $B > 0$, then the number of initial subregions is upper-bounded by $\lceil B \rceil^d$ and by Jensen's inequality, this adaptation would lead to a multiplicative cost of $\lceil B \rceil^{d/(d+2)}$ in the upper-bound of Theorem 2.2.3.

Comparison with other methods. One may want to obtain similar guarantees by considering other strategies like uniform histograms, kernel regression, or nearest neighbors, which were studied in the context of stationary ergodic processes by Györfi, Lugosi, and Fargas [GLF01], Gyorfi and Ottucsak [GO07], Biau, Bleakley, Györfi, and Ottucsák [Bia+10], and Biau and Patra [BP11]. We were unfortunately unable to provide any finite-time and worst-case analysis neither for kernel regression nor for nearest neighbors estimation.

The regret bound of Theorem 2.2.3 can however be obtained in an easier manner with uniform histograms. To do so, one can consider the class of uniform histograms $\mathcal{H}_N$. We divide the covariable space $[0,1]^d$ in a partition $(I_j)_{j=1,\dots,N}$ of N hyper-rectangle subregions of equal size. $\mathcal{H}_N$ is the set of functions that takes constant values in $[0,1]$ in each subregion I_j. We consider the class of 2^N prediction strategies that predict the constant values 0 or 1 in each bin of the partition. Competing with this class of 2^N functions by resorting for instance to EG gives the regret bound

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) \leqslant \min_{h \in \mathcal{H}_N} \sum_{t=1}^{T} \ell\Big(h(\boldsymbol{x}_t), y_t\Big) + 2M\sqrt{TN \log 2}\,.$$

Now, considering the approximation of $\mathcal{H}_N$ to capture $\mathcal{L}_L^d$ and optimizing the number N of bins in hindsight setting $N \approx (dT)^{\frac{d}{d+2}}$ (or by resorting to the doubling trick, see Cesa-Bianchi and Lugosi [CL06]) provides a regret bound similar to the one of Theorem 2.2.3 of order $d^{d/(4d+4)}T^{(d+1)/(d+2)}$ against any Lipschitz function. The details are provided in the appendix.

In the worst case the nested EG strategy provides no better guarantee. Such worst case occurs for large number N_T of nodes, which happens in particular when the trees

are height-balanced, that is, when the covariables $\boldsymbol{x}_t$ are uniformly distributed in $[0,1]^d$. But the nested EG strategy adapts better to data as it is depicted in the simulation studies below. If the covariables $\boldsymbol{x}_t$ are non-uniformly allocated (with regions of the space $[0,1]^d$ associated with much more observations than in other regions of similar size), the resulting tree $\mathcal{T}_T$ will be un-balanced, leading to a smaller number of nodes. In the best case, $N_T = O(H_T)$, which yields a regret of order $O(\sqrt{T \log T})$.

By improving the definition of Algorithm 5, one can even obtain the optimal and expected $O(\sqrt{T})$ regret if $(\boldsymbol{x}_t)$ is constant. To do so, it only needs to compute online the effective range of the data that belongs to each node (h,i),

$$\delta_t^{(h,i)} = \operatorname{diam}\{\boldsymbol{x}_s, \quad 0 \leqslant s \leqslant t \text{ and } (h_s, i_s) = (h,i)\}$$

and substitute the diameter $\operatorname{diam} \mathcal{X}^{(h,i)}$ by $\delta_{t+1}^{(h,i)}$ in the splitting condition of the algorithm (step 5).

2.2.3 Simulation studies

We consider in this section two simulated data sets so as to compare the performance obtained by three procedures:

- histEG corresponds to run Algorithm 4 on uniform histograms, whose size is theoretically calibrated by doubling trick with approximatively $T^{1/3}$ bins (this follows from the theoretical tuning of the previous section with $d = 1$);
- nestedEG is Algorithm 5;
- nestedEG(+) corresponds to Algorithm 5 where the splitting condition (step 5) is replaced by the condition:

$$2 \min_{c \in [0,1]} \left\{ \sum_{s \in E^{(h_t, i_t)}} (y_s - c)^2 \right\} > \sum_{s \in E^{(h_t, i_t)}} (y_s - \widehat{y}_s)^2$$

  which can be rephrased as: "the approximation error of the best constant in the local region is larger than the estimation error". This condition is better than the one suggested in Algorithm 5 since it adapts not only to the structure of covariates $(\boldsymbol{x}_t)$ but also to the structure of the objective variable (y_t). In particular, it adapts to easy areas of the space where the link function g between $\boldsymbol{x}_t$ and y_t does not vary much. We conjecture that the theoretical guarantees can be retrieved, the proof is however left for future research.

We performed two simulations studies. In the first one, the data are independent and identically distributed, while in the second one the data are distributed from an Hidden Makov Model (HMM), see the monograph of Cappé, Moulines, and Rydén [CMR05]. In both studies, we sampled sequences $\{(X_t, Y_t)\}_{t=1,\dots,1000}$ of $T = 1000$ observations.

Experiment 1: I.I.D. data. (X_t) is independent and identically distributed from a mixture of two Gaussian distribution restricted to $[0,1]$. Its density is proportional to

$$\frac{1}{2}\Big(\mathcal{N}_{(0,0.1)}(t) + \mathcal{N}_{(0.7,0.1)}(t)\Big)\, \mathbb{1}_{t \in [0,1]}\,,$$

where $\mathcal{N}_{(\mu,\sigma)}$ is the density function of the normal distribution of mean μ and standard deviation σ. (Y_t) is independent and identically distributed and follows the normal distribution $\mathcal{N}(g(X_t), 0.1)$ restricted to $[0,1]$, where $g(x) = (\cos(10x) + \sin(15x) + \cos(20x) + \sin(25x) + \cos(30x) + 2)/6$. The data are represented in Figure 2.3. The choices of the distribution of X and Y are quite arbitrary. The goal was to obtain X not uniformly distributed over $[0,1]$ and to have a function g with large variations in some areas and small variations in others.

Experiment 2: HMM. (X_t, Y_t) follows a 2-states HMM with transition probabilities $a_{11} = a_{22} = 0.9$ and $a_{12} = a_{21} = 0.1$ and uniform initial distribution. For each $t \geqslant 1$, $X_t \sim \text{Beta}(2,5)$ for state 1 and $X_t \sim \text{Beta}(0.7, 0.3)$ for state 2. Besides, for state $i \in \{1,2\}$, $Y_t \sim \mathcal{N}(g_i(X_t), 0.1)$ restricted to $[0,1]$, where $g_1(x) = x$ and $g_2(x) = |\cos(2\pi x)|$.

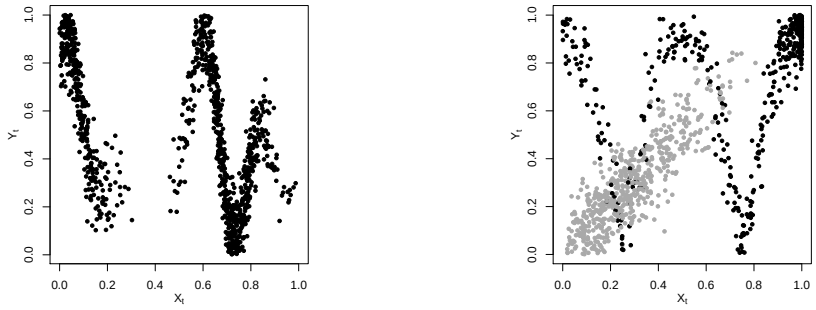

Figure 2.2: Plots of the observations (X_t, Y_t) of Experiment 1 [left] and Experiment 2 [right]. In Experiment 2, gray points corresponds to state 1 and black points to state 2.

The data are depicted in Figure 2.2. Figure 2.3 displays the boxplots (over 100 independent runs of the experiments) of the root mean square errors (RMSEs) obtained by the three forecasting procedures. We observe much better performance of the data-driven calibration procedures in comparison to uniform histograms with theoretical calibration of the number of bins. We could observe that in area with large variations of the link function and many data, the bins of nested EG are significantly smaller than in areas with few points and low variation of the link function g. Our method however suffers from the non-smoothness of histogram procedures. The extension of our deterministic analysis to smoother strategies (such as nearest neighbors, or kernel regression) is left for future research.

2.3 Autoregressive framework

We present in this section a technical result that will be useful for later purposes. Here, the forecaster still sequentially observes from time $t = 1$ an arbitrary bounded sequence $(y_t)_{t=-\infty,\dots,+\infty}$. However, at time step t, it is asked to forecast the next outcome $y_t \in [0,1]$ with knowledge of the past observations $y_1^{t-1} = y_1, \dots, y_{t-1}$ only.

We are interested in a strategy that performs asymptotically as well as the best model that considers the last d observations to form the predictions, and does this

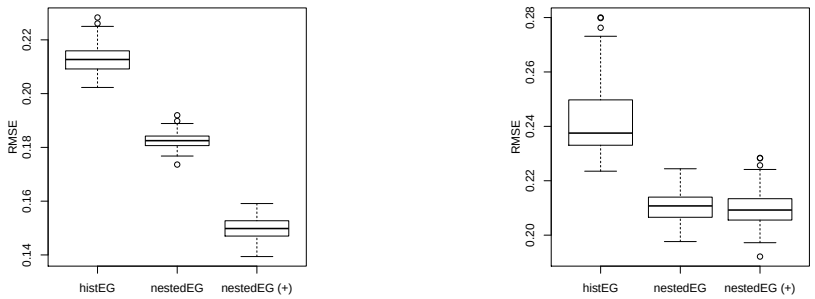

Figure 2.3: Boxplots of the RMSE*s* suffered by the three forecasting procedures over 100 independent replications of Experiment 1 [left] and Experiment 2 [right].

simultaneously for all values of $d \geqslant 1$. More formally, we denote

$$\widehat{R}_{L,T}^{d} \triangleq \sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=d+1}^{T} \ell\big(f(y_{t-d}^{t-1}), y_t\big)\,,$$

and we want that for all d, the average regrets $\widehat{R}_{L,T}^{d}/T$ vanish as $T \to \infty$. We show how it can be obtained via a meta-algorithm (Algorithm 6) that combines an increasing sequence of orders d of autoregressive nested EG forecasters.

Fixed order *d*. For each order $d \geqslant 0$, let $\mathcal{A}_d$ denote the autoregressive forecaster of order d that forms prediction for $t \geqslant d+1$ by performing nested EG (Algorithm 5) on the sequence $\big\{(y_t, y_{t-d}^{t-1})\big\}_{t \geqslant d+1}$. We denote by $f_{d,t}$ the prediction provided by $\mathcal{A}_d$ at time step $t \geqslant d+1$. By substituting $\boldsymbol{x}_t = y_{t-d}^{t-1}$ in Theorem 2.2.3, $\mathcal{A}_d$ satisfies the following regret bound

$$\begin{aligned}\sum_{t=d+1}^{T} \ell(f_{d,t}, y_t) - \inf_{f \in \mathcal{L}_L^d} \sum_{t=d+1}^{T} \ell\big(f(y_{t-d}^{t-1}), y_t\big) \\ \leqslant M(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,,\end{aligned} \tag{2.4}$$

which is valid for all $T \geqslant 1$, all $L > 0$ and for all $y_1, \ldots, y_T \in [0,1]$.

All orders *d*. Now, we show how to obtain the above regret bound simultaneously for all orders $d \geqslant 1$. To do so, Algorithm 6 combines via EG the predictions formed by all forecasters $\mathcal{A}_d$ for $d \geqslant 0$. Note that at time step t, only the t first forecasters $\mathcal{A}_0, \ldots, \mathcal{A}_{t-1}$ suggest predictions.

Parameter:

- $(\mathcal{A}_d)_{d\geqslant 1}$ a sequence of forecasters such that $\mathcal{A}_d$ forms predictions for time steps $t \geqslant d+1$

For $t = 1, \ldots, T$

1. **For** each $d = 0, \ldots, t-1$, denote by $f_{d,t}$ the prediction formed by $\mathcal{A}_d$
2. predict $\widehat{y}_t = \sum_{d=0}^{t-1} \widehat{p}_{d,t} f_{d,t}$
3. initialize the weight of the new forecaster: $p_{t,t+1} = 1/(t+1)$
4. observe Y_t and perform exponential weight update component-wise for $d = 0, \ldots, t-1$ as

$$\widehat{p}_{d,t+1} = \frac{t}{t+1} \, \frac{\widehat{p}_{d,t}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell(f_{d,t},y_t)}}{\sum_{k=1}^{t} \widehat{p}_{k,t}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell(f_{k,t},y_t)}} \,,$$

where $\eta_t = 2\sqrt{(\log t)/t}$ for all $t \geqslant 1$.

Algorithm 6: Extension of the Algorithm 5 to unknown order d of autoregressive model.

Lemma 2.3.1 controls the cumulative loss of Algorithm 6 by the cumulative loss of the best strategy $\mathcal{A}_d$.

Lemma 2.3.1. *Let $T \geqslant 1$. Then, Algorithm 6 satisfies for all $d \in 1, \dots, T$, for all $L > 0$, and for all $y_1, \dots, y_T \in [0,1]$,*

$$\sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) - \ell(f_{d,t}, y_t) \leqslant \sqrt{(T+1)\log(T+1)}\,.$$

The proof of Lemma 2.3.1 follows the standard one of the exponentially weighted average forecaster. It is postponed to Section 2.5.2. It could also be recovered by noting that our setting with starting experts is a particular case of the setting of sleeping experts introduced in Freund, Schapire, Singer, and Warmuth [Fre+97].

Theorem 2.3.2. *Let $T \geqslant 1$, $L > 0$. Then, for all $d \in \{0, \dots, T-1\}$, Algorithm 6 run with the sequence of forecasters $(\mathcal{A}_d)$ satisfies for all $L > 0$ and for all $y_1, \dots, y_T \in [0,1]$,*

$$\begin{aligned}\widehat{R}^d_{L,T} &= \sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{L}^d_L} \sum_{t=d+1}^{T} \ell(f(y_{t-d}^{t-1}), y_t) \\ &\leqslant \sqrt{(T+1)\log(T+1)} + M(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,.\end{aligned}$$

Consequently, for all $d \geqslant 1$, $\limsup_{T\to\infty} \left(\widehat{R}^d_{L,T}/T\right) \leqslant 0$.

Proof. The regret bound is by combining (2.4) and Lemma 2.3.1. The second part is obtained by dividing by T and letting T grow to infinity. □

2.4 From individual sequences to ergodic processes: convergence to $L^\star$

In this section, we present our main result by deriving from Theorem 2.3.2 similar results obtained in a stochastic setting by Györfi, Lugosi, and Fargas [GLF01], Gyorfi and Ottucsak [GO07], Biau, Bleakley, Györfi, and Ottucsák [Bia+10], and Biau and Patra [BP11].

We leave here the setting of individual sequences of the previous sections and we assume that the sequence of observations $y_1, \dots, y_T$ is now generated by some stationary ergodic process. More formally, we assume that a stationary bounded ergodic process $(Y_t)_{t=-\infty,\dots,\infty}$ is sequentially observed. At time step t, the learner is asked to form a prediction $\widehat{Y}_t$ of the next outcome $Y_t \in [0,1]$ of the sequence with knowledge of the past observations $Y_1^{t-1} = Y_1, \dots, Y_{t-1}$. The nested EG strategy, as a consequence of the deterministic regret bound of Theorem 2.2.3, will be shown to be consistent, i.e., satisfies Equation (2.1).

Theorem 2.4.1 shows that any strategy that achieves a deterministic regret bound for individual sequences as in Theorem 2.3.2 predicts asymptotically as well as the best strategy defined by a Borelian function.

Theorem 2.4.1. *Let $(Y_t)_{t=-\infty,\dots,\infty}$ be a stationary bounded ergodic process. We assume that for all t, $Y_t \in [0,1]$ almost surely and that for all $d \geqslant 1$ the law of $Y_{-d}^{-1} = (Y_{-d},\dots,Y_{-1})$ is regular. Let $\ell : [0,1]^2 \to [0,1]$ be a loss function M-Lipschitz in its first argument. Assume that a prediction strategy satisfies for all $d \geqslant 1$, almost surely,*

$$\forall L \geqslant 0 \qquad \limsup_{T\to\infty} \left(\frac{1}{T} \sum_{t=1}^{T} \ell\Big(\widehat{Y}_t, Y_t\Big) \right) \leqslant \limsup_{T\to\infty} \left(\inf_{f\in\mathcal{L}_L^d} \frac{1}{T} \sum_{t=1}^{T} \ell\Big(f(Y_{t-d}^{t-1}), Y_t\Big) \right) ,$$

then, almost surely,

$$\limsup_{T\to\infty} \left(\frac{1}{T} \sum_{t=1}^{T} \ell\Big(\widehat{Y}_t, Y_t\Big) \right) \leqslant L^\star .$$

By Theorem 2.3.2, Algorithm 6 satisfies the assumption of Theorem 2.4.1 (by replacing the deterministic terms y_t by the random variables Y_t). Our deterministic strategy is thus asymptotically optimal for any stationary bounded ergodic process satisfying the assumptions of Theorem 2.4.1. Here we only give the main ideas in the proof of Theorem 3. The complete argument is given in Section 2.5.3.

Sketch of proof. Basically, the proof of Theorem 2.4.1 consists first in applying Breiman's generalized ergodic theorem[‡] (see Breiman [Bre57]), so that

$$\limsup_{T\to\infty} \left(\frac{1}{T} \sum_{t=1}^{T} \ell\Big(f(Y_{t-d}^{t-1}), Y_t\Big) \right) = \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big] .$$

Then, by exchanging lim sup and inf in the right-term of the assumption and by letting $L \to \infty$, we can compete against any Lipschitz function:

$$\limsup_{T\to\infty} \left(\frac{1}{T} \sum_{t=1}^{T} \ell\Big(\widehat{Y}_t, Y_t\Big) \right) \leqslant \inf_{f\in\mathcal{L}_L^d} \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big] .$$

The proof is then completed by approximating the best Borelian function by the best Lipschitz function (see Lemma 2.4.2 below). □

Lemma 2.4.2. *Let $\mathcal{X}$ be a convex and compact subset of a normed space. Let $\ell : [0,1]^2 \to [0,1]$ be a loss function M-Lipschitz in its first argument. Let X be a random variable on $\mathcal{X}$ with a regular law $\mathbb{P}_X$ and let Y be a random variable on $[0,1]$. Then,*

$$\inf_{f\in\mathcal{L}^{\mathcal{X}}} \mathbb{E}\big[\ell(f(X),Y)\big] = \inf_{f\in\mathcal{B}^{\mathcal{X}}} \mathbb{E}\big[\ell(f(X),Y)\big] ,$$

where $\mathcal{L}^{\mathcal{X}}$ denotes the set of Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{B}^{\mathcal{X}}$ the one of Borelian functions from $\mathcal{X}$ to $\mathbb{R}$.

The proof of Lemma 2.4.2 is postponed to Section 2.5.3. It follows from the Stone-Weierstrass theorem, used to approximate continuous functions, and from Lusin's theorem, to approximate Borelian functions.

[‡]Here, we use the assumption that (Y_t) is a stationary ergodic process.

The assumptions. Theorem 2.4.1 makes two main assumptions on the ergodic sequence to be predicted. First, the sequence is supposed to lie in $[0,1]$. As earlier, this assumption can be easily relaxed to any bounded subset of $\mathbb{R}$—see remarks of Sections 2.2.1 and 2.2.2. The generalization to unbounded sequence is left to future work and should follow from the same techniques as in Gyorfi and Ottucsak [GO07]. Second, Theorem 2.4.1 assumes that for all $d \geqslant 1$ the law of Y_{-d}^{-1} is regular, that is, for any Borelian set $S \subset [0,1]^d$ and for any $\varepsilon > 0$, one can find a compact set K and an open set V such that

$$K \subset S \subset V, \qquad \text{and} \qquad \mathbb{P}_{Y_{-d}^{-1}}(V \backslash K) \leqslant \varepsilon \,.$$

This second assumption is considerably weaker than the assumptions required by Biau and Patra [BP11] for quantile prediction. The authors indeed imposed that the random variables $\|Y_{-d}^{-1} - s\|$ have continuous distribution functions for all $s \in \mathbb{R}^d$ and the conditional distribution function $F_{Y_0|Y_{-\infty}^{-1}}$ to be increasing. One can however argue that their assumptions are thus hardly comparable with ours because they consider unbounded ergodic processes. We aim at obtaining in the future minimal assumptions for any generic convex loss function ℓ in the case of unbounded ergodic process, see Morvai and Weiss [MW11].

Computational efficiency. The space complexity of Algorithm 6 is $O(T^2)$. Previous algorithms of Györfi, Lugosi, and Fargas [GLF01], Gyorfi and Ottucsak [GO07], Biau, Bleakley, Györfi, and Ottucsák [Bia+10], and Biau and Patra [BP11] exhibit consistent strategies as well. However, in practice, these algorithms involve choices of parameters somewhere in their design (by choosing the a priori weight of the infinite set of experts). Then, the consideration of an infinite set of experts makes the exact algorithm computationally inefficient. For practical purpose, it needs to be approximated. This can be obtained by MCMC or for instance by restricting the set of experts to some finite subset at the cost, however, of loosing theoretical guarantees, see Biau and Patra [BP11].

Generic loss function. Theorem 2.4.1 assumes ℓ to be bounded, convex, and M-Lipschitz in its first argument. In contrast, the results of Györfi, Lugosi, and Fargas [GLF01], Gyorfi and Ottucsak [GO07], Biau, Bleakley, Györfi, and Ottucsák [Bia+10], and Rakhlin and Sridharan [RS14] only hold for the square loss (while Biau and Patra [BP11] extend them to the pinball-loss).

2.5 Technical proofs

We gather in this section the proofs, which were omitted from the previous sections.

2.5.1 Proofs of Section 2

The proofs of Theorem 2.2.3 and Lemma 2.2.4 are based on the following lemma, which controls the size of the regions associated with nodes located at depth h in the tree $\mathcal{T}_T$.

Lemma 2.5.1. *Let $h \geqslant 0$. Then, for all indices $i = 1, \dots, 2^h$, the diameter of the region $\mathcal{X}^{(h,i)}$ associated with node (h, i) in Algorithm 5 is upper bounded as*

$$\operatorname{diam}\left(\mathcal{X}^{(h,i)}\right) \leqslant \sqrt{2d}2^{-h/d}\,.$$

Proof. Basically, the proof of Lemma 2.5.1 consists of an induction on the depth h. It suffices to prove that for all $h \geqslant 0$, for all indexes $i \in \{1, \dots, 2^h\}$ and all coordinates $j \in \{1, \dots, d\}$, the ranges $\delta_j^{(h,i)} \triangleq \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}^{(h,i)}} |x_j - x'_j|$ satisfies

$$\delta_j^{(h,i)} = \begin{cases} 2^{-(k+1)} & \text{if } j \leqslant r \\ 2^{-k} & \text{otherwise} \end{cases}, \tag{2.5}$$

where $h = kd + r$ is the decomposition with $r \in \{0, \dots, d-1\}$. Indeed, we then have

$$\operatorname{diam}\left(\mathcal{X}^{(h,i)}\right) = \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}^{(h,i)}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leqslant \sqrt{\sum_{j=1}^{d} \left(\delta_j^{(h,i)}\right)^2}\,.$$

But by (2.5), for r coordinates $j \in \{1, \dots, r\}$ among the d coordinates $\delta_j^{(h,i)}$ equals $2^{-(k+1)}$ while the $d - r$ remaining coordinates $j \in \{r+1, \dots, d\}$ satisfy $\delta_j^{(h,i)} = 2^{-k}$. Thus, by routine calculations

$$\begin{aligned} \operatorname{diam}\left(\mathcal{X}^{(h,i)}\right) &\leqslant \sqrt{r\left(2^{-(k+1)}\right)^2 + (d-r)\left(2^{-k}\right)^2} \\ &= 2^{-k}\sqrt{\frac{r}{4} + d - r} \\ &= \sqrt{d}2^{-k}\sqrt{1 - \frac{3r}{4d}} \\ &= \sqrt{d}\left(2^{1/d}\right)^{-(dk+r)} 2^{r/d}\sqrt{1 - \frac{3r}{4d}} \end{aligned}$$

But,

$$2^{r/d}\sqrt{1 - \frac{3r}{4d}} \leqslant \max_{0 \leqslant u \leqslant 1} \left\{2^u \sqrt{1 - \frac{3u}{4}}\right\} \approx 1.12 \leqslant \sqrt{2}\,.$$

The proof is concluded by substituting in the previous bound.

Now, we prove (2.5) by induction on the depth h. This is true for $h = 0$ as the bin of the root node $\mathcal{X}^{(0,1)}$ equals $[0,1]^d$ by definition. Besides, let $h \geqslant 0$ and $i \in \{1, \dots, 2^h\}$. We compute the decomposition $h = kd + r$ with $r \in \{0, \dots, d-1\}$. We have by step 5.4 of Algorithm 5 that the range of each coordinate $j \neq r+1$ of the bin of the child node $(h+1, 2i)$ remains the same

$$\delta_j^{(h+1,2i)} = \delta_j^{(h,i)} = \begin{cases} 2^{-(k+1)} & \text{if } j \leqslant r \\ 2^{-k} & \text{if } j \geqslant r+2 \end{cases}, \tag{2.6}$$

and the range of coordinate $r+1$ is divided by 2,

$$\delta_{r+1}^{(h+1,2i)} = \delta_{r+1}^{(h,i)}/2 = 2^{-(k+1)}\,. \tag{2.7}$$

Equations (2.6) and (2.7) are also true for the second child $(h+1, 2i-1)$, and this concludes the induction. □

Proof of Lemma 2.2.4

Upper bound for N_T. For each node (h,i), we recall that $T^{(h,i)} = \sum_{t=1}^{T} \mathbb{1}_{\{(h_t,i_t)=(h,i)\}}$ denotes the number of observations predicted by using the local version of Algorithm 4 associated with the sequence of observations $E^{(h,i)}$. The total number of observations T is the sum of $T^{(h,i)}$ over all nodes (h,i). That is,

$$T = \sum_{h=0}^{H_T} \sum_{i=1}^{2^h} T^{(h,i)} \mathbb{1}_{\{(h,i)\in\mathcal{T}_T\}} \geqslant \sum_{h=0}^{H_T} \sum_{i=1}^{2^h} T^{(h,i)} \mathbb{1}_{\left\{(h,i) \text{ is an inner node in } \mathcal{T}_T\right\}}.$$

Now we use the fact that each inner node (h,i) has reached its splitting condition (step 5 of Algorithm 5), that is, $T^{(h,i)} + 1 \geqslant \left(\operatorname{diam}(\mathcal{X}^{(h,i)})\right)^{-2}$. Using that $\operatorname{diam}\left(\mathcal{X}^{(h,i)}\right) \leqslant \sqrt{2d}2^{-h/d}$ by Lemma 2.5.1, we get

$$\begin{aligned} T &\geqslant \sum_{h=0}^{H_T} \sum_{i=1}^{2^h} \left[-1 + \left(\operatorname{diam}\left(\mathcal{X}^{(h,i)}\right)\right)^{-2}\right] \mathbb{1}_{\{(h,i) \text{ is an inner node }\}} \\ &\geqslant \sum_{h=0}^{H_T} \underbrace{\left(-1 + \frac{2^{2h/d}}{2d}\right)}_{g(h)} \underbrace{\sum_{i=1}^{2^h} \mathbb{1}_{\{(h,i) \text{ is an inner node }\}}}_{n_h}. \end{aligned} \tag{2.8}$$

Because $g : \mathbb{R}_+ \to \mathbb{R}$ is convex in h, by Jensen's inequality

$$T \geqslant N_T^{\text{in}}\, g\left(\frac{1}{N_T^{\text{in}}} \sum_{h=0}^{H_T} h n_h\right),$$

where $N_T^{\text{in}} = \sum_h n_h$ is the total number of inner nodes. Now, by Lemma 2.5.2 available in Section 2.5.1, because $\mathcal{T}_T$ is a binary tree with N_T nodes in total, it has exactly $N_T^{\text{in}} = (N_T - 1)/2$ inner nodes and the average depth of its inner nodes is lower-bounded as

$$\frac{1}{N_T^{\text{in}}} \sum_{h=0}^{H_T} h n_h \geqslant \log_2\left(\frac{N_T - 1}{8}\right).$$

Substituting in the previous bound, it implies

$$\begin{aligned} T &\geqslant \frac{N_T - 1}{2} g\left(\log_2\left(\frac{N_T - 1}{8}\right)\right) \\ &= \frac{N_T - 1}{2}\left(-1 + \frac{1}{2d} 2^{\frac{2}{d}\log_2\left((N_T-1)/8\right)}\right) \\ &= -\frac{N_T - 1}{2} + \frac{N_T - 1}{4d}\left(\frac{N_T - 1}{8}\right)^{2/d} \\ &\geqslant \underbrace{-\frac{N_T - 1}{2}}_{\geqslant -T/2} + \frac{2}{d}\left(\frac{N_T - 1}{8}\right)^{1+2/d}. \end{aligned}$$

By reorganizing the terms, it entails $dT \geqslant (3/4)dT \geqslant \big((N_T - 1)/8\big)^{1+2/d}$. Thus, $(N_T - 1)/8 \leqslant (dT)^{d/(d+2)}$, which yields the desired bound for N_T.

Upper bound for H_T. We start from (2.8) and we use the fact that for all $h = 0, \dots, H_T - 1$, there exists at least one inner node of depth h in $\mathcal{T}$. Thus,

$$T \geqslant \sum_{h=0}^{H_T-1} \left(-1 + \frac{2^{2h/d}}{2d}\right) = -H_T + \frac{1}{2d}\frac{2^{2H_T/d} - 1}{2^{2/d} - 1} \geqslant -H_T + \frac{2^{2(H_T-1)/d}}{2d}$$

where the last inequality is because $(a-1)/(b-1) \geqslant a/b$ for all numbers $a \geqslant b > 1$. Therefore, by upper-bounding $T \geqslant H_T$, we get $4T \geqslant 2^{2(H_T-1)/d}/d$ and thus $2(H_T - 1)/d \leqslant \log_2(4dT)$ which concludes the proof.

Proof of Theorem 2.2.3

The cumulative regret suffered by Algorithm 5 is controlled by the sum of all cumulative regrets incurred by all local versions of Algorithm 4, each associated with a subsequence of observations $E^{(h,i)}$. That is,

$$\widehat{R}_{L,T} \leqslant \sum_{(h,i)\in\mathcal{T}_T} \left[\sum_{t\in E^{(h,i)}} \ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{L}_L^d} \sum_{t\in E^{(h,i)}} \ell(f(\boldsymbol{x}_t), y_t)\right],$$

where $E^{(h,i)} = \{1 \leqslant t \leqslant T \,:\, (h_t, i_t) = (h,i)\}$ is the set of time steps assigned to node (h,i). Now, by Lemma 2.2.2, the cumulative loss incurred by nested EG associated with node (h,i) satisfies

$$\begin{aligned}\sum_{t\in S^{(h,i)}} \ell(\widehat{y}_t, y_t) &\leqslant \inf_{y\in[0,1]} \sum_{t\in S^{(h,i)}} \ell(y, y_t) + M\sqrt{T^{(h,i)}\log 2} \\ &\leqslant \inf_{f\in\mathcal{L}_L^d} \sum_{t\in S^{(h,i)}} \ell(f(\boldsymbol{x}_t), y_t) + ML \underbrace{\operatorname{diam}\Big(\mathcal{X}^{(h,i)}\Big)}_{\leqslant 1/\sqrt{T^{(h,i)}} \text{ by step 5 of Algorithm 5}} T^{(h,i)} + 2M\sqrt{T^{(h,i)}\log 2}\end{aligned}$$

where the second inequality is by Lemma 2.2.1. Thus,

$$\widehat{R}_{L,T} \leqslant M\Big(L + \underbrace{2\sqrt{\log 2}}_{\leqslant 3}\Big) \sum_{(h,i)\in\mathcal{T}_T} \sqrt{T^{(h,i)}}\,.$$

Then, by Jensen's inequality,

$$\frac{1}{N_T} \sum_{(h,i)\in\mathcal{T}_T} \sqrt{T^{(h,i)}} \leqslant \sqrt{\frac{1}{N_T}\sum_{(h,i)} T^{(h,i)}} = \sqrt{\frac{T}{N_T}}\,,$$

which concludes the first statement of the theorem. The second statement follows from Lemma 2.2.4 and because for all $a, b \geqslant 0$, $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$,

$$\begin{aligned}M(L+3)\sqrt{N_T T} &\leqslant M(L+3)\sqrt{\Big(1 + 4(3dT)^{d/(d+2)}\Big)T} \\ &\leqslant M(L+3)\Big(\sqrt{T} + \sqrt{4(3dT)^{d/(d+2)}T}\Big) \\ &= M(L+3)\Big(\sqrt{T} + 2(3d)^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\Big)\,.\end{aligned}$$

Lemma 2.5.2 and its proof

Lemma 2.5.2. *Let $N \geqslant 1$ be an odd integer. Let $\mathcal{T}$ be a binary tree with N nodes. Then,*

- *its number of inner-nodes equals $N^{\mathrm{in}} = (N-1)/2$.*
- *the average depth (i.e., distance to the root) of its inner nodes is lower-bounded as*
$$\frac{1}{N^{\mathrm{in}}} \sum_{h=0}^{\infty} h \, \#\{\textit{inner nodes in } \mathcal{T} \textit{ of depth } h\} \geqslant \log_2\left(\frac{N-1}{8}\right).$$

Proof. First statement. We proceed by induction. If $N = 1$, there is only one binary tree with one node, the lone leaf, so that $N^{\mathrm{in}} = 0$. Now, if $\mathcal{T}$ is a binary tree with $N \geqslant 3$ nodes, select an inner node n which is parent of two leaf nodes. Then, replace the subtree rooted at n by a leaf node. The resulting subtree $\mathcal{T}'$ of $\mathcal{T}$ has $N-2$ nodes, so that by induction hypothesis $\mathcal{T}'$ has $(N-3)/2$ inner nodes. But, $\mathcal{T}'$ has also $N^{\mathrm{in}}-1$ inner nodes. Therefore $N^{\mathrm{in}} = (N-1)/2$.

Second statement. We note that the average depth is minimized for the equilibrated binary trees, that are such that

- all depths $h \in \{0, \dots, \lfloor \log_2 N^{\mathrm{in}} \rfloor\}$ have exactly 2^h inner nodes;
- no inner nodes has depth $h > \lceil \log_2 N^{\mathrm{in}} \rceil$.

Therefore,

$$\frac{1}{N^{\mathrm{in}}} \sum_{h=0}^{\infty} h \, \#\{\text{inner nodes in } \mathcal{T} \text{ of depth } h\} \geqslant \frac{1}{N^{\mathrm{in}}} \sum_{h=0}^{\lfloor log_2 N^{\mathrm{in}} \rfloor} h 2^h$$

Now, we use that $\sum_{i=0}^{n-1} i2^i = 2^n(n-2)+2$ for all $n \geqslant 1$, which implies because $\lfloor \log_2 N^{\mathrm{in}} \rfloor \geqslant \log_2 N^{\mathrm{in}} - 1$ and by substituting in the previous bound,

$$\frac{1}{N^{\mathrm{in}}} \sum_{h=0}^{\infty} h \, \#\{\text{inner nodes in } \mathcal{T} \text{ of depth } h\} \geqslant \underbrace{\frac{2^{\log_2 N^{\mathrm{in}}}}{N^{\mathrm{in}}}}_{=1} \Big(\log_2 N^{\mathrm{in}} - 2\Big) + \underbrace{\frac{2}{N^{\mathrm{in}}}}_{\geqslant 0}.$$

This concludes the proof by substituting $N^{\mathrm{in}} = (N-1)/2$. □ □

2.5.2 Proofs of Section 2.3

The proof of Lemma 2.3.1 follows from a simple adaptation of the proof of the regret bound of the exponentially weighted average forecaster—see for instance Cesa-Bianchi and Lugosi [CL06]. By convexity of ℓ and by Hoeffding's inequality, we have at each time step t

$$\ell(\widehat{y}_t, y_t) \leqslant \sum_{d=0}^{t-1} \widehat{p}_{d,t} \ell(f_{d,t}, y_t) \leqslant -\frac{1}{\eta_t} \log \sum_{d=0}^{t-1} \widehat{p}_{d,t} e^{-\eta_t \ell(f_{d,t}, y_t)} + \frac{\eta_t}{8}$$

By Jensen's inequality, since $\eta_{t+1} \leqslant \eta_t$ and thus $x \mapsto x^{\eta_t/\eta_{t+1}}$ is convex

$$\begin{aligned}\frac{1}{t}\sum_{d=0}^{t-1} \widehat{p}_{d,t} e^{-\eta_t \ell(f_{d,t},y_t)} &= \frac{1}{t}\sum_{d=0}^{t-1} \left(\widehat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1}\ell(f_{d,t},y_t)}\right)^{\frac{\eta_t}{\eta_{t+1}}} \\ &\geqslant \left(\frac{1}{t}\sum_{d=1}^{t-1} \widehat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1}\ell(f_{d,t},y_t)}\right)^{\frac{\eta_t}{\eta_{t+1}}}\end{aligned}$$

Substituting in Hoeffding's bound we get

$$\ell(\widehat{y}_t, y_t) \leqslant \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\log t - \frac{1}{\eta_{t+1}}\log\left(\sum_{d=0}^{t-1} \widehat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1}\ell(f_{d,t},y_t)}\right) + \frac{\eta_t}{8}$$

Now, by definition of the loss update in step 3 of Algorithm 6, for all $d = 0, \dots, t-1$

$$\sum_{k=0}^{t-1} \widehat{p}_{k,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1}\ell(f_{k,t},y_t)} = \frac{t}{t+1}\,\frac{\widehat{p}_{d,t}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1}\ell(f_{d,t},y_t)}}{\widehat{p}_{d,t+1}}$$

which after substitution in the previous bound leads to the inequality

$$\ell(\widehat{y}_t, y_t) \leqslant \ell(f_{d,t}, y_t) + \frac{1}{\eta_{t+1}}\log\left((t+1)\,\widehat{p}_{d,t+1}\right) - \frac{1}{\eta_t}\log(t\,\widehat{p}_{d,t}) + \frac{\eta_t}{8}\,.$$

By summing over $t = d+1, \dots, T$, the sum telescopes; using that $\widehat{p}_{d,d+1} = 1/(d+1)$ by step 3.1.

$$\begin{aligned}&\sum_{t=d+1}^{T} \ell(\widehat{y}_t, y_t) - \sum_{t=d+1}^{T} \ell(f_{d,t}, y_t) \\ &\qquad\leqslant \frac{1}{\eta_{T+1}}\log\left((T+1)\,\underbrace{\widehat{p}_{d,T+1}}_{\leqslant 1}\right) - \frac{1}{\eta_t}\log\left(\underbrace{(d+1)\,\widehat{p}_{d,d+1}}_{=1}\right) + \frac{1}{8}\sum_{t=d+1}^{T}\eta_t \\ &\qquad\leqslant \frac{1}{\eta_{T+1}}\log(T+1) + \frac{1}{8}\sum_{t=d+1}^{T}\eta_t\,.\end{aligned}$$

Finally, by routine calculation

$$\begin{aligned}\sum_{t=d+1}^{T}\eta_t \leqslant 2\sum_{t=1}^{T}\sqrt{\frac{\log t}{t}} &\leqslant 2\sqrt{\log T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\ &= 2\sqrt{\log T}\left(1 + \sum_{t=2}^{T}\frac{1}{\sqrt{t}}\right) \\ &\leqslant 2\sqrt{\log T}\left(1 + \int_1^T \frac{1}{\sqrt{t}}dt\right) \\ &\leqslant 4\sqrt{T(\log T)}\,,\end{aligned}$$

which concludes the proof.

2.5.3 Proofs of Section 2.4

Proof of Lemma 2.4.2

The proof is performed in two steps.

Step 1: Lipschitz → Continuous. The set $\mathcal{L}$ of Lipschitz functions, from a compact metric space $\mathcal{X}$ to $\mathbb{R}$, is a subalgebra of the set $\mathcal{C}$ of continuous functions. Besides, $\mathcal{L}$ contains the constant functions and separates the points of $\mathcal{X}$. Therefore, the Stone-Weierstrass theorem, recalled in Theorem 2.5.3, entails that any continuous function $f : \mathcal{X} \to \mathbb{R}$ is the uniform limit of Lipschitz functions. Thus, the dominated convergence theorem yields

$$\inf_{f\in\mathcal{L}} \mathbb{E}\Big[\ell(f(X),Y)\Big] = \inf_{f\in\mathcal{C}} \mathbb{E}\Big[\ell(f(X),Y)\Big]\,.$$

.

Step 2: Continuous → Borelian. Second, by the version of Lusin's theorem stated in Theorem 2.5.4, we can approximate any mesurable function by continuous functions (this is where regularity is used).

Let $\delta, \varepsilon > 0$ and $f : \mathcal{X} \to [0,1]$ be a Borelian function. By Theorem 2.5.4, there exists a continuous function $g : \mathcal{X} \to [0,1]$ such that

$$\mathbb{P}_X\Big\{|f-g| \geqslant \delta\Big\} \leqslant \varepsilon\,.$$

Then by Jensen's inequality, and since

$$\begin{aligned}\Delta \triangleq \Big|\mathbb{E}\Big[\ell(f(X),Y)\Big] - \mathbb{E}\Big[\ell(g(X),Y)\Big]\Big| &\leqslant \mathbb{E}\Big[\Big|\ell(f(X),Y) - \ell(g(X),Y)\Big|\Big] \\ &\leqslant \underbrace{\mathbb{P}_X\Big\{|f-g| \geqslant \delta\Big\}}_{\leqslant\varepsilon} + \underbrace{\mathbb{E}\Big[M|f(X)-g(X)|\,\mathbb{1}_{\{|f(X)-g(X)|\leqslant\delta\}}\Big]}_{\leqslant M\delta}\,,\end{aligned}$$

where the second inequality is because ℓ takes values in $[0,1]$ and is M-Lipschitz in its first argument. Thus $\Delta \leqslant \varepsilon + M\delta$, which concludes the proof since this is true for arbitrary small values of ε and δ.

Theorem 2.5.3 (Stone-Weierstrass)**.** *Let $\mathcal{C}(X,\mathbb{R})$ be the ring of continuous function on a compact X with the topology of uniform convergence i.e. the topology generated by the norm*

$$||f|| = \max_{x\in X} |f(x)| \qquad f \in \mathcal{C}(X,\mathbb{R})\,.$$

Let $A \subseteq \mathcal{C}(X,\mathbb{R})$ be a subring containing all constant functions and separating the points of X, that is for any two different points $x_1, x_2 \in X$, there exists a function $f \in A$ for which $f(x_1) \neq f(x_2)$. Then A is dense in $\mathcal{C}(X,\mathbb{R})$: every continuous function on X is the limit of a uniformly converging sequence of functions in A.

Proof. The proof is carried out in several references, for instance Rudin [Rud91] □

Theorem 2.5.4 (Lusin)**.** *If $\mathcal{X}$ is a convex and compact subset of a normed space, equipped with a regular probability mesure μ, then for every measurable function $f : \mathcal{X} \to [0,1]$ and for every $\delta, \varepsilon > 0$, there exists a continuous function $g : \mathcal{X} \to [0,1]$ such that*

$$\mu\left\{|f - g| \geqslant \delta\right\} \leqslant \varepsilon\,.$$

Proof. The proof of Theorem 2.5.4 can be easily derived from the proof of Stoltz and Lugosi [SL07, Proposition 25]. □

Proof of Theorem 2.4.1

In this proof, apart from the use of Breiman's generalized ergodic theorem in the beginning and the martingale convergence theorem in the end (as exhibited in Györfi, Lugosi, and Fargas [GLF01], Gyorfi and Ottucsak [GO07], Biau, Bleakley, Györfi, and Ottucsák [Bia+10], and Biau and Patra [BP11]), we resort to new arguments.

Let $d \geqslant 1$ and $L \geqslant 0$. Then, by assumption and by exchanging lim sup and inf,

$$\limsup_{T\to\infty} \frac{1}{T}\left(\sum_{t=1}^{T} \ell(\widehat{Y}_t, Y_t)\right) \leqslant \inf_{f\in\mathcal{L}_L^d} \limsup_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \ell(f(Y_{t-d}^{t-1}), Y_t)\right).$$

Because ℓ is bounded over $[0,1]^2$ and thus integrable, Breiman's generalized ergodic theorem (see Breiman [Bre57]) entails that the right-term converges: almost surely,

$$\lim_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \ell(f(Y_{t-d}^{t-1}), Y_t)\right) = \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big]$$

and thus,

$$\limsup_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \ell(\widehat{Y}_t, Y_t)\right) \leqslant \inf_{f\in\mathcal{L}_L^d} \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big]\,.$$

By letting $L \to \infty$ in the inequality above, we get

$$\limsup_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \ell(\widehat{Y}_t, Y_t)\right) \leqslant \inf_{f\in\mathcal{L}^d} \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big]\,.$$

By Lemma 2.4.2 the infimum over all continuous functions equals the infimum over the set $\mathcal{B}^d$ of Borelian functions. Therefore,

$$\begin{aligned}\limsup_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \ell(\widehat{Y}_t, Y_t)\right) &\leqslant \inf_{f\in\mathcal{B}^d} \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big] \\ &\leqslant \mathbb{E}\Big[\underbrace{\inf_{f\in\mathcal{B}^d} \mathbb{E}\Big[\ell(f(Y_{-d}^{-1}), Y_0)\Big|Y_{-d}^{-1}\Big]}_{\triangleq Z_d}\Big]\,,\end{aligned}$$

where the second inequality is by the measurable selection theorem—see Theorem 8 in Appendix I of Algoet [Alg94]. Now, we remark that (Z_d) is a bounded super-martingale with respect to the family of sigma algebras $\big(\sigma(Y_{-d}^{-1})\big)_{d\geqslant 1}$. Indeed, the

function $\inf_{f\in\mathcal{B}^{d+1}}(.)$ is concave, thus conditional Jensen's inequality

$$\begin{aligned}\mathbb{E}[Z_{d+1}|Y_{-d}^{-1}] &\leqslant \inf_{f\in\mathcal{B}^{d+1}} \mathbb{E}\Big[\mathbb{E}\Big[\ell\Big(f(Y_{-(d+1)}^{-1}),Y_0\Big)\Big|Y_{-(d+1)}^{-1}\Big]\Big|Y_{-d}^{-1}\Big]\\ &= \inf_{f\in\mathcal{B}^{d+1}} \mathbb{E}\Big[\ell\Big(f(Y_{-(d+1)}^{-1}),Y_0\Big)\Big|Y_{-d}^{-1}\Big]\end{aligned}$$

Now, we note that

$$\inf_{f\in\mathcal{B}^{d+1}} \mathbb{E}\Big[\ell\Big(f(Y_{-(d+1)}^{-1}),Y_0\Big)\Big|Y_{-d}^{-1}\Big] \leqslant \inf_{f'\in\mathcal{B}^{d}} \mathbb{E}\Big[\ell\Big(f'(Y_{-d}^{-1}),Y_0\Big)\Big|Y_{-d}^{-1}\Big] = Z_d\,,$$

which yields $\mathbb{E}[Z_{d+1}|Y_{-d}^{-1}] \leqslant Z_d$. Thus, the martingale convergence theorem (see e.g. Chow [Cho65]) implies that Z_d converges almost surely and in $\mathbb{L}_1$. Thus,

$$\lim_{d\to\infty}\mathbb{E}[Z_d] = \mathbb{E}\Big[\inf_{f\in\mathcal{B}^\infty}\mathbb{E}\Big[\ell(f(Y_{-\infty}^{-1}),Y_0)\Big|Y_{-\infty}^{-1}\Big]\Big] = L^\star\,,$$

which yields the stated result $\limsup_T \sum_{t=1}^T \ell(\widehat{Y}_t,Y_t)/T = L^\star$.

2.6 Uniform histograms

We detail in this appendix the performance bound obtained by competing against uniform histograms over the input space $\mathcal{X}=[0,1]^d$. We denote by $\mathcal{H}_N$ the class of uniform histograms with N hyper-rectangle subregions of equal size. Note that this class exists for $N\in\{i^d,\text{ such that } i\geqslant 1\}$.

Approximation error. First, we bound the approximation error made by the best uniform histogram in $\mathcal{H}_N$ to approximate the unknown best L-Lipschitz objective function $f^\star:[0,1]^d\to[0,1]$. The diameter $\mathrm{diam}(\mathcal{H}_N)$ with respect to the ℓ^2-norm of the bins of a uniform histogram in $\mathcal{H}_N$ equals

$$\begin{aligned}\mathrm{diam}(\mathcal{H}_N) &\triangleq \max_{\substack{\boldsymbol{x}_i,\boldsymbol{x}_j\in[0,1]^d\\ \forall h\in\mathcal{H}_N,\ h(\boldsymbol{x}_i)=h(\boldsymbol{x}_j)}} \|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2\\ &= \sqrt{d}\;N^{-1/d}\,.\end{aligned}$$

Therefore, by applying Lemma 2.2.1 on each bin and summing over all N bins, the cumulative approximation error of $\mathcal{H}_N$ satisfies for all $L>0$

$$\begin{aligned}\inf_{h\in\mathcal{H}_N}\sum_{t=1}^T \ell(h(\boldsymbol{x}_t),y_t) &\leqslant \inf_{f\in\mathcal{L}_L^d}\sum_{t=1}^T \ell(f(\boldsymbol{x}_t),y_t) + MLT\,\mathrm{diam}(\mathcal{H}_N)\\ &= \inf_{f\in\mathcal{L}_L^d}\sum_{t=1}^T \ell(f(\boldsymbol{x}_t),y_t) + MLT\sqrt{d}\;N^{-1/d}\,. \qquad (2.9)\end{aligned}$$

Estimation error. Now, we bound the additional error obtained by estimating the best histogram in $\mathcal{H}_N$ online. To do so, we resort to EG on the set of 2^N functions that predict the constant values 0 or 1 in each bin of the partition of $\mathcal{H}_N$, we obtain the upper-bound

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) \leqslant \inf_{h \in \mathcal{H}_N} \sum_{t=1}^{T} \ell\Big(h(\boldsymbol{x}_t), y_t\Big) + 2M\sqrt{T \log\left(2^N\right)}\,. \tag{2.10}$$

Total error. By summing the approximation error (2.9) and the estimation error (2.10), we finally get the regret bound

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) \leqslant \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^{T} \ell(f(\boldsymbol{x}_t), y_t) + \underbrace{MLT\sqrt{d}\; N^{-1/d}}_{\text{Approximation}} + \underbrace{2M\sqrt{TN\log 2}}_{\text{Estimation}}\,. \tag{2.11}$$

The optimal number of bins N that balances the approximation and the estimation errors need to be optimized. Solving the equality

$$MLT\sqrt{d}\; N^{-1/d} = 2M\sqrt{TN\log 2}\,,$$

in N yields the optimal value

$$N = \left(\frac{L}{2}\right)^{\frac{2d}{d+2}} \left(\frac{dT}{\log 2}\right)^{\frac{d}{d+2}}\,.$$

Substituting in (2.11), we get

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) \leqslant \inf_{f \in \mathcal{L}_L^d} \sum_{t=1}^{T} \ell(f(\boldsymbol{x}_t), y_t) + \underbrace{2(4L\log 2)^{\frac{d}{d+2}}}_{\leqslant 6L} M d^{\frac{d}{2(d+2)}} T^{\frac{d+1}{d+2}}\,.$$

Online calibration of N. To achieve the above regret bound, the forecaster need to know the Lipschitz constant L and the time horizon T in advance. We could not find a solution for the calibration of the Lipschitz constant. Therefore, we assumed the constant to be $L = 1$ which resulted in the suboptimal linear dependency in L instead of $L^{\frac{d}{d+2}}$. However, we can avoid the assumption that T is known in advance at the cost of a constant factor by resorting to the well-known doubling trick, see Cesa-Bianchi and Lugosi [CL06]. The idea is to restart the algorithm whenever we reach a time step t such that t is a power of 2. At each restart, we forget all the information gained in the past and we set $N \approx (dt)^{\frac{d}{d+2}}$[§].

[§]We recall that for uniform histograms the number of bins should lie in $\{i^d,\ \text{such that } i \geqslant 1\}$

3 Pression réduite au niveau de la mer

Nous présentons une méthode d'agrégation séquentielle des simulations météorologiques de pression réduite au niveau de la mer. L'objectif est d'obtenir, grâce à l'algorithme ridge, de meilleures performances en prévision qu'une certaine prévision de référence, à déterminer. Dans la partie 3.1, nous rappelons le cadre mathématique et les fondamentaux des sciences environnementales. Puis, les jeux de données utilisés et les performances pratiques de l'algorithme sont détaillés en partie 3.2. Enfin, la partie 3.3 décrit plus en détail certains aspects du jeu de données et de l'algorithme.

Sommaire

3.1 Généralités

3.1.1 Notations

$\boldsymbol{x}_t$	Vecteur d'état de la prévision, à valeur dans $\mathbb{R}^M$
y_t	Observation scalaire
$\boldsymbol{u}$	Vecteur de poids
$\boldsymbol{u}_t$	Vecteur de poids dépendant du temps t
$\boldsymbol{p}_t$	Vecteur de poids dépendant du temps t
M	Taille de l'ensemble : nombre de membres
λ_t	Paramètre de régularisation de la régression ridge
γ_t	Paramètre d'escompte temporelle de la régression ridge

3.1.2 Rappels théoriques

Nous cherchons à réaliser une agrégation séquentielle des simulations issues des membres de l'ensemble de prévisions, afin d'obtenir les meilleures performances possibles en prévision. L'objectif est de prévoir une suite arbitraire d'observations $y_1, y_2, \ldots$, chacune étant issue d'un espace donné $\mathcal{Y}$. Arbitraire signifie qu'on ne pose aucun a priori sur cette suite, i.e. y_t n'est pas la réalisation d'un processus stochastique sous-jacent. L'avantage de ce cadre est qu'il permet d'exhiber des bornes théoriques qui valent pour toutes les suites $(y_t)_{t\in\mathbb{N}}$, pas seulement les plus typiques ou les plus probables. Ce que l'on appelle en agrégation séquentielle un *expert* correspond dans notre cas précisément à une simulation d'une grandeur météorologique issue d'un membre de l'ensemble et constitue la brique fondamentale de notre prévision. Plus formellement, il s'agit d'un nombre fini M de prédicteurs élémentaires issus d'un espace $\mathcal{X}$, convexe ou muni d'une structure d'espace vectoriel. Cet ensemble $\mathcal{X}$ peut différer de $\mathcal{Y}$. À chaque échéance t, le m-ème prédicteur fournit une prévision $x_{m,t} \in \mathcal{X}$ qui s'appuie potentiellement sur les observations passées $y_1, y_2, \ldots, y_{t-1}$ et sur des informations à lui seul accessibles. À chaque échéance t, notre prévision de y_t est notée $\widehat{y}_t$ et appartient à l'ensemble $\mathcal{X}$. Cette prévision est ensuite comparée à l'observation y_t et une perte est encourue via la fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. On définit la perte cumulée d'un expert m fixé sur les T premières échéances par

$$L_{m,T} = \sum_{t=1}^{T} \ell(x_{m,t}, y_t)\,,$$

et la perte cumulée du statisticien sur les T premières échéances comme

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

et on souhaite minimiser cette dernière.

Pour cela, la stratégie consiste à combiner les prévisions fournies par les experts, à l'échéance courante,

$$\widehat{y}_t = \sum_{m=1}^{M} u_{m,t} x_{m,t}\,.$$

En toute généralité, les poids $\boldsymbol{u}_t$ appartiennent à $\mathbb{R}^M$, sans restriction. Certaines méthodes imposent cependant que ces poids évoluent dans des ensembles donnés, par exemple, le simplexe $\mathcal{P}$ des combinaisons convexes :

$$\mathcal{P} = \left\{ \boldsymbol{q} : \sum_{m=1}^{M} q_m = 1 \text{ et } q_m \geqslant 0 \,, \forall\, m \in \{1, \dots, M\} \right\} .$$

La méthode de calcul des poids $\boldsymbol{u}_t$ constitue la stratégie $\mathcal{S}$, à laquelle est associée la perte cumulée du statisticien

$$\widehat{L}_T(\mathcal{S}) = \sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) = \sum_{t=1}^{T} \ell\left(\sum_{m=1}^{M} u_{m,t} x_{m,t}, y_t \right) .$$

Similairement, la perte cumulée pour un vecteur $\boldsymbol{u}$ de poids constants est définie comme :

$$L_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell\left(\sum_{m=1}^{M} u_m x_{m,t}, y_t \right) .$$

Cependant, la qualité de chacun des experts n'est pas assurée : on peut imaginer une situation où tous les experts prévoient des valeurs systématiquement erronées $x_{m,t}$. La tâche de prévision du statisticien est alors nécessairement plus ardue que si tous les experts sont performants à chaque échéance. La perte cumulée $\widehat{L}_T(\mathcal{S})$ peut alors être conséquente même si la stratégie $\mathcal{S}$ choisie est la plus performante dans ce cas de figure précis (c'est-à-dire au vu des hypothèses données). Définir un critère absolu de performance de la prévision réalisée empêche l'appréhension des performances générales de $\mathcal{S}$. Par conséquent, on définit un critère relatif, le *regret*, qui marque le gain ou la perte de performance de notre combinaison linéaire séquentielle par rapport à l'instance de la classe de comparaison choisie. Les classes de comparaison les plus classiques sont celles constituées respectivement de tous les experts, de toutes les combinaisons convexes ou encore de toutes les combinaisons linéaires constantes, dont les pertes cumulées associées sont :

$$\min_{m=1,\dots,M} L_{m,T} \,, \quad \inf_{\boldsymbol{u}\in\mathcal{P}} L_T(\boldsymbol{u}), \quad \text{ou} \quad \inf_{\boldsymbol{u}\in\mathbb{R}^M} L_T(\boldsymbol{u}).$$

Le regret de la stratégie $\mathcal{S}$ sur les T premières échéances est alors la différence entre la perte cumulée du statisticien et la perte cumulée du meilleur expert,

$$R_T^{\text{exp}}(\mathcal{S}) = \widehat{L}_T(\mathcal{S}) - \min_{m=1,\dots,M} L_{m,T} \,; \tag{3.1}$$

ou bien de la meilleure combinaison convexe d'experts,

$$R_T^{\text{cvx}}(\mathcal{S}) = \widehat{L}_T(\mathcal{S}) - \inf_{\boldsymbol{u}\in\mathcal{P}} L_T(\boldsymbol{u}) \,; \tag{3.2}$$

ou encore de la meilleure combinaison linéaire d'experts,

$$R_T^{\text{lin}}(\mathcal{S}) = \widehat{L}_T(\mathcal{S}) - \inf_{\boldsymbol{u}\in\mathbb{R}^M} L_T(\boldsymbol{u}) \,. \tag{3.3}$$

Attention, la forme du regret et donc les performances atteignables dépendent crucialement de la classe de comparaison choisie.
Lorsque la fonction de perte est bornée, le regret de la stratégie la plus naïve est de l'ordre au plus de T, l'horizon temporel. Nous nous intéressons donc aux stratégies dont le regret croît plus lentement qu'un $\mathcal{O}(T)$ dans le pire des cas. Vu autrement, le regret rapporté au nombre d'échéances, tend uniformément vers 0, pour toutes les observations et toutes les prévisions des experts. Plus formellement, on cherche à concevoir des stratégies $\mathcal{S}$ telles que

$$\limsup_{T\to+\infty} \sup \frac{R_T(\mathcal{S})}{T} \leqslant 0\,;$$

où le supremum porte sur l'ensemble des suites d'observations $(y_t)_{t\in\mathbb{N}} \in \mathcal{Y}^{\mathbb{N}}$ et des prévisions d'expert possibles $(x_{m,t})_{t\in\mathbb{N}} \in \mathcal{X}^{\mathbb{N}}$, pour tout $m = 1, \dots, M$. Cette exigence est trop forte et ne peut jamais être réalisée sans ajouter certaines conditions supplémentaires. Ainsi, la stratégie de régression ridge qui fournit les résultats pratiques en 3.2.3 et minimise le regret linéaire requiert que les poids évoluent dans une boule de rayon fixé en norme L_2.

3.1.3 Régression ridge

La stratégie de régression ridge escomptée, $\mathcal{R}_{\lambda,\gamma}$, avec les paramètres λ de régularisation et γ d'escompte, est la suivante :

1. Initialisation : $\boldsymbol{u}_1$ est le vecteur de mélange uniforme, $u_{m,1} = 1/M$, pour $m = 1, \dots, M$.

2. À chaque échéance $t \geqslant 2$, le vecteur de poids est défini selon

$$\boldsymbol{u}_t = (u_{1,t}, \dots, u_{M,t})^T = \arg\min_{\boldsymbol{u}\in\mathbb{R}^M} \left\{ \lambda\|\boldsymbol{u}\|_2^2 + \sum_{s=1}^{t-1} (1+\psi_{t-s}) \left(y_s - \boldsymbol{u}\cdot\boldsymbol{x}_s\right)^2 \right\},$$

où nous avons noté $\|\cdot\|_2$ la norme euclidienne d'un vecteur de $\mathbb{R}^M$ et où la fonction $\psi_t = \gamma/t^2$, avec $\gamma > 0$, est décroissante. Celle-ci traduit que l'escompte apportée aux écarts quadratiques passés est d'autant plus grande que ceux-ci sont plus lointains. Plus le coefficient γ est grand, plus l'atténuation due à l'escompte est importante.

On peut montrer que, sous l'hypothèse de restreindre les poids linéaires $\boldsymbol{u}$ à un compact, l'algorithme de régression ridge vérifie une égalité de type oracle (3.3), dans un cadre sans escompte (voir Cesa-Bianchi et Lugosi [CL06, théorème 11.7]).

Dans cet algorithme, les poids sont séquentiellement déterminés comme les minimiseurs d'une perte cumulée, régularisée par un terme proportionnel à la norme euclidienne des poids. Régulariser la perte de cette façon permet d'assurer que la solution est unique et que les poids ne s'éloignent pas trop du vecteur nul. Dans ce cadre, les vecteurs successifs de poids $\boldsymbol{u}_t$ sont à nouveau pris en toute généralité, i.e., dans $\mathbb{R}^M$: leurs composantes peuvent ainsi prendre des valeurs négatives et ne pas se sommer à 1.

3.1.4 Glossaire des sciences environnementales

Discrétisation spatiale

La prévision météorologique cherche à prévoir les grandeurs physico-chimiques d'intérêt pour la compréhension de l'atmosphère et de son évolution. Il est évidemment impossible de connaître rigoureusement ces grandeurs en tout point de l'espace. Le météorologue se contente en pratique de moyennes spatiales. Une *cellule* est un volume élémentaire parallélépipédique dans lequel sont moyennées les grandeurs considérées. La *résolution* est la donnée des dimensions de ce parallélépipède. Tout comme une haute résolution d'image donne une représentation plus fidèle de la réalité, une résolution spatiale plus grande, c'est-à-dire des mailles plus resserrées, donne une vision potentiellement plus fidèle de l'évolution des grandeurs atmosphériques. La *grille* est l'ensemble des cellules sur le domaine de l'atmosphère considéré. Dans la suite, la résolution est de 0,10 ° (latitude, longitude) dans les deux directions horizontales. Cette résolution correspond à des cellules parallélépipédique d'environ 10 km de largeur et de longueur. La résolution verticale est plus fine mais nous ne nous intéressons ici qu'à la couche la plus basse de la grille.

Ensemble de simulations

Les météorologues réalisent des modèles physiques et thermodynamiques rigoureux depuis la fin du XIX^{e} siècle. Et les connaissances en ce domaine ne cessent de s'améliorer et de se raffiner, à tel point qu'il existe souvent divers modèles paramétriques, avec différentes plages de paramètres admissibles pour modéliser un phénomène. Néanmoins, certaines limites apparaissent systématiquement lorsque l'on essaye de prédire l'état de l'atmosphère. Ces limites résultent, entre autres, des sources d'incertitude suivantes : la stratégie numérique de résolution des équations, i.e. le schéma d'intégration choisi et des données brutes en entrée. Via l'imprécision sur les conditions initiales, la partie chaotique du système atmosphérique est aussi source de fortes incertitudes.

La prévision d'ensemble permet d'apporter un élément de réponse pertinent à ces problématiques. On note $\mathbf{X}_t$, le vecteur d'état, c'est-à-dire le vecteur rassemblant toutes les informations sur l'atmosphère disponibles à une date donnée sur la grille spatiale considérée. Nous verrons comment l'obtenir un peu plus loin dans cette partie. Dans ce vecteur sont conservées toutes les grandeurs physiques et thermodynamiques que le météorologue juge nécessaire au vu des équations d'évolution. On définit un modèle paramétrique, que l'on note $\mathcal{M}_t$, qui résout les équations d'évolution. Dans notre optique de prévision, on met à jour à chaque échéance un vecteur d'état $\mathbf{X}_t$ de l'atmosphère à l'aide de $\mathcal{M}_t$, le nouveau vecteur d'état $\mathbf{X}_{t+1}$ constituant notre prévision.

Pour la clarté, explicitons la dépendance du modèle aux paramètres d'entrée $\boldsymbol{p}_t$. L'intégration du schéma numérique d'une échéance à l'échéance suivante se formalise donc selon

$$\mathbf{X}_{t+1} = \mathcal{M}_t(\mathbf{X}_t, \boldsymbol{p}_t)\,.$$

En modifiant les paramètres employés $\boldsymbol{p}_t$ ou bien la condition initiale à l'échéance t, $\mathbf{X}_t$, nous sommes capables de construire plusieurs simulations reposant sur différentes

estimations de l'état initial et des données d'entrée. Chacune de ces combinaisons génère une diversité nouvelle et leur ensemble permet ainsi de refléter l'incertitude de la simulation. Comment faire varier ces paramètres et ces conditions initiales ? La démarche naturelle revient à se placer dans un cadre stochastique, où l'on considère les entrées comme des variables aléatoires que l'on échantillonne. On est donc en mesure de générer plusieurs tirages, comme dans une approche Monte Carlo, et d'obtenir autant de simulations que l'on a de tirages de perturbations sur les champs. Soulignons qu'au tirage aléatoire d'un jeu de paramètres d'entrée donné correspond une simulation. Ainsi, on construit un ensemble de simulations. Cet ensemble peut être enrichi par l'emploi de modèles $\mathcal{M}_t^{(i)}$ qui diffèrent entre eux par leur formulation physique et numérique. Actuellement, les centres de prévision mondiaux génèrent des ensembles d'une ou plusieurs dizaines de membres.

On peut écrire de manière compacte le modèle numérique, qui met à jour d'une échéance à l'autre l'ensemble des grandeurs physiques (discrétisées spatialement) :

$$\mathbf{X}_{t+1}^{(i,j)} = \mathcal{M}_t^{(i)}\left(\mathbf{X}_t^{(i,j)},\, \boldsymbol{p}_t + \delta_t^{(j)}\right).$$

où l'exposant i rend compte d'un changement de formulation physique et numérique du modèle tandis que l'exposant j rend compte des tirages aléatoires sur les données d'entrée du modèle. Cette égalité signifie donc que le modèle dépend des données d'entrées brutes et qu'on modifie les données d'entrée par une perturbation $\delta_t^{(j)}$ qu'on écrit ici comme additive. Afin de simplifier les notations, on s'affranchit des exposants i et j pour les regrouper dans l'exposant m. À chaque m correspond alors un seul modèle :

$$\mathbf{X}_{t+1}^{m} = \mathcal{M}_t^{m}\left(\mathbf{X}_t^{m},\, \boldsymbol{p}_t + \delta_t^{m}\right).$$

Analyse et simulation déterministe

Afin d'offrir une vision globale de l'évolution de l'atmosphère, la météorologie combine les différentes sources d'information dont elle dispose. Ces dernières sont les suivantes : les observations prises aux stations de mesures, que l'on peut considérer comme ponctuelles, les observations satellitaires et bien sûr les instantanés issus des simulations des modèles d'évolution. Intervient alors l'assimilation de données, un domaine qui cherche à calculer la meilleure représentation possible de l'état d'un système à partir de différentes sources d'information. Ce que l'on appelle une analyse est un état à un instant donné, compromis entre les observations disponibles et les contraintes du modèle d'évolution. Une analyse combine souvent un instantané d'un modèle d'évolution atmosphérique donné, nommé l'ébauche, et corrigé par les observations disponibles. Les centres de prévision météorologique confient le rôle de cette ébauche à la prévision déterministe réalisée pour cette échéance. Cette prévision déterministe est formée d'un modèle d'évolution qui, à partir des équations d'évolution atmosphérique et thermodynamique, fournit un unique scénario d'évolution des variables atmosphériques. Pour résumer, l'assimilation de données corrige la prévision déterministe par les observations ponctuelles pour fournir la meilleure estimation a posteriori possible de l'état atmosphérique.

Prévision d'ensemble

Comme exposé précédemment, nous disposons d'un ensemble de modèles susceptibles de décrire l'évolution de l'atmosphère. Cet ensemble traduit potentiellement les incertitudes en entrée et dans le calcul de ces simulations. Mais on peut aussi se donner comme objectif de construire à partir de cet ensemble une simulation unique pour approcher au mieux l'analyse. Historiquement en météorologie, l'idée première et naturelle a été de prendre la *moyenne d'ensemble* : s'il y a M modèles, on définit cette moyenne d'ensemble comme

$$\overline{X}_t = \frac{1}{M} \sum_{m=1}^{M} X_t^m \, .$$

Les résultats de cette méthode sont mitigés mais indiquent le chemin vers d'autres techniques plus prometteuses, par exemple réaliser une moyenne pondérée de ces modèles. Reprenons ici de manière plus pragmatique les résultats de la partie 3.1.2. On mesure la performance de chaque modèle par rapport à l'analyse (le meilleur état de l'atmosphère connu a posteriori) via une fonction de perte quadratique. Cette mesure d'écart est employée pour mettre à jour les poids liés à chaque simulation de l'ensemble. La prévision à l'échéance suivante prendra alors en entrée les simulations de chaque modèle et les poids associés mis à jour. Nous réalisons donc une combinaison linéaire ou convexe, dont la pondération change au cours du temps. Cet aspect dynamique est essentiel. En effet, il se peut qu'une simulation ou un sous-ensemble de simulations augmentent soudainement en fiabilité. Dans le cas de simulations convexes où l'interprétation des poids est directe, la stratégie d'agrégation séquentielle convexe permet d'augmenter les poids de ces modèles. Pensons à des simulations efficaces durant une saison particulière, inefficaces le reste du temps : celles-ci seront alors mises à profit au moment opportun. Cette méthode de combinaison convexe et linéaire dynamique et automatique, l'*agrégation d'ensemble*, est au cœur de l'étude empirique suivante.

3.1.5 Variables considérées

Nous allons dans l'étude empirique considérer les deux variables suivantes :

- La pression réduite au niveau de la mer (en anglais, MSLP : Mean Sea Level Pressure) ;
- La vitesse du vent au niveau du sol.

La pression réduite au niveau de la mer est la pression atmosphérique observée à la surface du sol et réduite dans le but de rendre comparable les pressions entre elles. Une pression mesurée à une station est ajustée en considérant que la température décroît à un taux de 6.5K par tranche de mille mètres d'altitude. Outre qu'il s'agit de deux variables typiques et intéressantes pour la météorologie, leurs natures diffèrent. La pression réduite au niveau de la mer est une variable synoptique : elle reflète des phénomèmes se produisant à l'échelle planétaire tandis que la vitesse du vent au niveau du sol est une variable locale dont l'échelle de variation peut être faible. Il est donc intéressant d'étudier les deux afin d'apréhender la sensibilité et la robustesse éventuelle de la méthode d'agrégation séquentielle au type de variable météorologique étudiée.

3.1.6 Méthodologie

Voici le plan d'étude que nous allons suivre, similaire à celui réalisé dans Devaine, Gaillard, Goude et Stoltz [Dev+13, p. 12]. Ce plan d'étude se décline pour chacune des deux variables envisagées, la pression réduite au niveau de la mer et le module de la vitesse du vent au niveau du sol.

1. Nous décrivons les limites spatiales et temporelles du jeu de données ainsi que les divers acteurs : l'ensemble de prévisions, l'analyse et la prévision déterministe.

2. Nous choisissons la fonction de perte et évaluons les pertes des membres de l'ensemble. Puis nous calculons les oracles convexes et linéaires ainsi que leurs performances. Parmi les candidats disponibles, nous repérons alors le point de référence pour l'évaluation des performances.

3. Commencent alors les phases d'agrégation des membres de l'ensemble. Nous choisissons une grille fixe de paramètres de régression. Puis, pour chaque couple de paramètres, nous effectuons une régression ridge et sélectionnons le meilleur résultat auquel correspond le jeu de paramètres optimaux a posteriori.

4. Une fois cette étape réalisée, nous pouvons alors opérer l'automatisation du choix de paramètres de la méthode de régression. Pour ce faire, il s'agit de modifier l'algorithme précédent :

   a) tout d'abord selon une stratégie de sélection des paramètres semi-automatique. Nous définissons une période $T \in \mathbb{N}$ et la stratégie peut sauter d'un jeu de paramètres à un autre tous les $k \times T$, $\forall k \in \mathbb{N}$. Le nouveau jeu de paramètres pour la période suivante $k \times T, \ldots, (k+1) \times T$ est celui qui a obtenu la meilleure performance globale durant la période $0, \ldots, k \times T$.

   b) puis selon une stratégie complètement automatique. Durant une première période exclue de l'évaluation, dynamiquement créer la grille en explorant itérativement l'espace des paramètres à la manière de Devaine, Gaillard, Goude et Stoltz [Dev+13], section 5.5. Employer sur cette grille la stratégie précédente durant le reste de la période.

5. Enfin, il est alors intéressant de vérifier la robustesse de la méthode vis-à-vis des paramètres ainsi que du nombre de membres.

3.2 Étude empirique de la pression réduite au niveau de la mer

Nous allons appliquer la méthodologie définie en 3.1.6 à l'étude de la variable de pression réduite au niveau de la mer.

3.2.1 Description du jeu de données

Limite spatiale et cadre temporel

La résolution horizontale est 0,10° selon la latitude, de même selon la longitude. Le domaine s'étend à l'Europe occidentale entre les parallèles de latitudes 35° et 61° et entre les méridiens de longitudes −15° et 17°. Cela recouvre une zone allant du nord du Maroc jusqu'au sud de la Norvège, et de l'Italie jusqu'au large de l'Atlantique. L'agrégation est réalisée sur une période de 366 jours depuis le 2011-10-01 jusqu'au 2012-10-01 exclu.

Prévisions déterministes et analyse

Nous disposons des prévisions déterministes émanant de deux sources différentes : Météo France et le Centre européen de prévision météorologique à moyen terme (ECMWF pour *European Centre for Medium-Range Weather Forecasts*). Les prévisions déterministes de ces deux sources sont réputées de bonne qualité et apparaissent, avant même le moindre calcul, comme des points de références potentiels auxquels comparer les résultats des stratégies d'agrégation. Le calcul de leurs performances en prévision sera fait dans la partie 3.2.2. Notons qu'en pratique, les prévisions déterministes de Météo France sont précisément celles présentées dans les prévisions à plus court terme (6 à 48 heures) des bulletins météorologiques. Au contraire, les simulations d'ensemble interviennent pour les prévisions à échéances plus lointaines, par exemple pour estimer l'incertitude de la simulation déterministe.

L'analyse est, rappelons-le, la meilleure estimation d'un vecteur d'état à une échéance donnée, qui combine deux sources d'information : les mesures de la variable considérée et une ébauche, ici, la prévision déterministe. Deux centres de prévision produisent des analyses légèrement différentes puisqu'ils ne possèdent pas les mêmes prévisions déterministes. C'est donc cette réalité terrain que l'on cherche à approcher le plus précisément possible à l'aide de nos prévisions agrégées. Dans le cadre formel des suites individuelles de la partie 3.1.2, l'analyse est appelée observation, y_t. Étant donné la visée opérationnelle de cette étude, l'analyse employée est très naturellement celle que Météo France génère. La résolution horizontale de 0,10° est la résolution à laquelle sont générées la prévision déterministe et l'analyse. Une représentation graphique de la moyenne temporelle de l'analyse est donnée en figure 4.1.

Ensemble considéré

TIGGE, qui est l'acronyme pour « the THORPEX Interactive Grand Global Ensemble » présenté dans Richardson [Ric05], est un projet d'envergure internationale visant à mettre en commun et à rendre disponible les ensembles de prévisions des centres météorologiques partenaires. Il s'agit d'une implémentation opérationnelle réalisée dans le cadre du projet THORPEX. L'un des objectifs est de permettre au chercheur en météorologie d'améliorer les prévisions en s'appuyant sur la diversité d'origine des

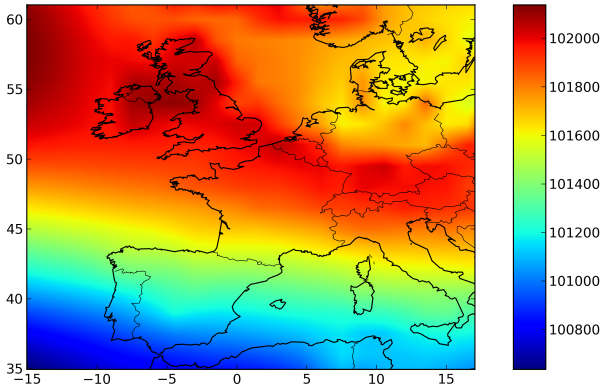

FIGURE 3.1 – Représentation graphique de la moyenne temporelle de l'analyse, dans le cas de la pression réduite au niveau de la mer (Pmer, MSLP). Dans chaque cellule de coordonnées (i, j) est représentée la moyenne $\frac{1}{T}\sum_{t=1}^{T} y_{t,(i,j)}$. L'unité est le Pascal et ces moyennes sont réparties autour de l'atmosphère standard (101 325 Pa). L'étendue de ces valeurs moyennes est de 2300 Pa.

membres des ensembles. En effet, les deux sources d'incertitudes majeures en prévision météorologique sont l'incertitude sur les conditions initiales et l'incertitude sur les modèles de prévision. Chaque centre de prévision régional (européen (ECMWF), américain (NCEP), chinois (CMA), etc. voir la table 3.1) génère un faisceau de simulations météorologiques qui lui est propre et diffère des autres par les perturbations apportées aux conditions initiales et la constitution des modèles physiques et thermodynamiques. Dans la communauté météorologique, ces prévisions sont nommées *prévisions perturbées* pour rappeler la manière dont elles sont conçues.

La résolution originelle est de 0,25°. Les simulations des membres de l'ensemble sont donc plus grossières que celles fournies par l'analyse ou le déterministe. S'il en allait autrement, vu que chaque centre de prévision produit une vingtaine de simulations perturbées, les temps de calculs seraient prohibitifs. Les simulations, mises à disposition et partagées par TIGGE, sont interpolées linéairement pour atteindre une résolution de 0,10° au moment où nous récupérons les données. Avec une différence telle de résolution native entre la prévision déterministe et les membres de l'ensemble, il semble intuitif que les performances de tous les membres de l'ensemble soient moins bonnes que celles des prévisions déterministes.

À cet ensemble déjà conséquent, nous nous laissons la possibilité d'ajouter ou non les prévisions déterministes fournies par ECMWF et Météo France. Dans toute la suite, nous préciserons s'il s'agit de l'ensemble constitué uniquement des membres fournis par TIGGE ou bien de l'ensemble de simulations augmenté, enrichi du déterministe ECMWF ou Météo France ou encore des deux à la fois.

Les informations concernant les centres de prévision dont sont originaires les membres de l'ensemble sont rassemblées dans la table 3.1. On note qu'un tiers des membres de l'ensemble est issu du centre européen (ECMWF) tandis que les cent autres membres sont répartis de manière relativement équilibrée entre les cinq autres centres de prévisions.

Origine	Nombre de membres	Acronyme
Chine	14	CMA
Canada	20	CMC
Europe	50	ECMWF
Corée du Sud	23	KMA
États-Unis	20	NCEP
Grande-Bretagne	23	UKMO
Total	150	

TABLE 3.1 – Table donnant les divers centres internationaux de prévision météorologique, les nombres de membres associés qui sont employés dans l'ensemble de prévisions ainsi que leurs acronymes.

Certaines simulations issues de centres de prévisions régionaux n'ont pas été inclus dans l'ensemble. Ainsi, les centres australiens (BoM), brésiliens (CPTEC) et japonais (JMA), qui sont membres de l'ensemble TIGGE n'apparaissent pas dans la table 3.1. La raison en est que les données fournies par ces centres sont incomplètes durant la période étudiée et ne sont donc pas exploitables. Par ailleurs, pour des raisons d'incompatibilités entre leurs échéances et celles du déterministe, les prévisions d'ensemble de Météo France n'ont pas pu non plus être exploitées.

Bien que nous puissions considérer les membres issus d'un centre météorologique donné comme des boîtes noires lorsque nous souhaitons les combiner, nous possédons néanmoins les informations suivantes sur la manière dont sont calculées les simulations. Chaque centre de prévision météorologique conçoit plusieurs modèles d'évolution météorologique, en sélectionnant les modèles physiques et thermodynamiques et leurs paramétrisations, les conditions aux limites, les schémas de discrétisation. Chaque modèle d'évolution est appliqué à un vecteur d'état donné à une échéance de départ (spécifié par le jour et l'heure). À certaines échéances données dans le cours de cette évolution (par exemple 6h, 12h, 18h, 24h, 32h...), on relève le vecteur d'état ayant évolué en suivant le modèle sélectionné : ces résultats constituent les prévisions aux échéances en question pour ce modèle.

Soulignons une pratique commune à tous les centres de météorologie que nous nommerons le rebattage des cartes. Parmi les modèles d'évolution météorologique, il y en a parfois un qui est meilleur que les autres. Si les équipes en opérationnel le repèrent, elles peuvent alors choisir de mettre à l'écart les autres membres de l'ensemble. Cela est néfaste car ces derniers recellent de l'information sur l'incertitude de la prévision à échéance de plus de quelques jours via la dispersion de l'ensemble. Pour pallier ce risque, les modèles météorologiques sont régulièrement (par exemple tous les trois jours)

redistribués parmi les différents membres pour que le meilleur modèle soit intraçable. Ce rebattage des cartes impliquent que les membres de l'ensemble n'ont pas de modèle d'évolution associé en continu. Étant donné que l'on cherche à pondérer un ensemble le plus cohérent possible, il est plus logique de trier les membres de l'ensemble à chaque pas de temps et de recoller sur l'ensemble de la période les membres selon leur rang. Ainsi, le 1$^{\text{er}}$ membre est systématiquement le minimum de l'ensemble et le $M^{\text{ème}}$ membre est systématiquement le maximum. Dans la suite, « l'ensemble » désignera par défaut l'ensemble trié de la sorte et nous préciserons « non trié » pour préciser l'alternative. Notons que dans tous les cas, l'ajout éventuel de la prévision du déterministe ECMWF ou Météo France ou encore des deux à la fois s'opère après ce tri.

Les figures 3.2 et 3.3 montrent des exemples d'évolution de la pression réduite au niveau de la mer dans une cellule pour une sous-partie des membres de l'ensemble, l'analyse et le déterministe Meteo France. Il est intéressant de remarquer que, bien qu'il existe une variabilité intra-centre, la variabilité inter-centre est plus importante. Cela est visible à la figure 3.3 et davantage encore à la figure 4.4 que nous verrons lors de l'étude des performances de l'ensemble, en partie 3.2.2. Par ailleurs, une représentation graphique de la moyenne temporelle de l'analyse est donnée en figure 3.4.

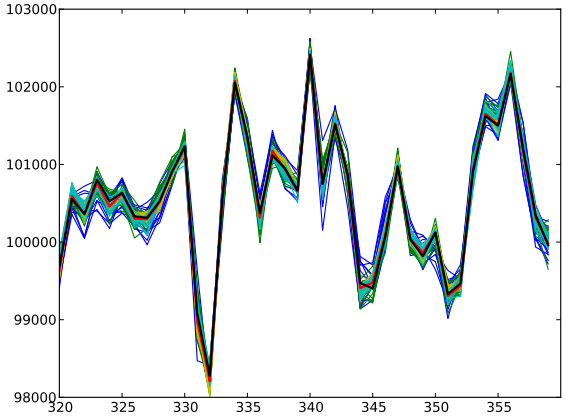

FIGURE 3.2 – Données de pression réduite au niveau de la mer dans une cellule typique (de coordonnées $(-13°, 37°)$), entre les échéances 320 (16 août 2012) et 360 (25 septembre 2012) de la période temporelle considérée. Un échantillon de 100 membres de l'ensemble parmi les 150 est représenté. En rouge, la prévision déterministe, $x^{\text{det}}_{t,(i,j)}$; en noir, l'analyse, $y_{t,(i,j)}$; le reste des couleurs est dédié aux membres de l'ensemble, $x^{m}_{t,(i,j)}$.

3.2.2 Performance de l'ensemble, oracle et point de référence

Performance de l'ensemble

La figure 4.4 permet de visualiser la performance moyenne de chacun des membres de l'ensemble en *RMSE*. Les membres de l'ensemble issus d'un même centre météorologique

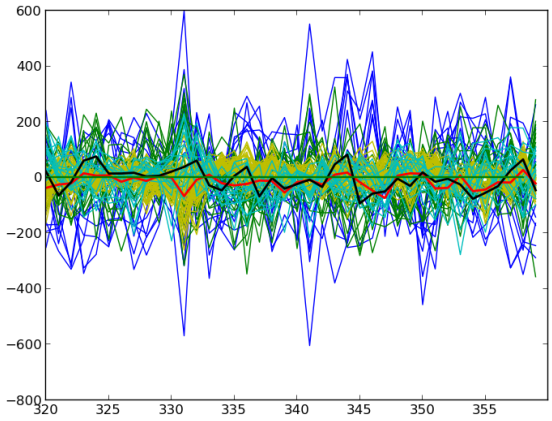

FIGURE 3.3 – Représentation graphique supplémentaire des données de 3.2, ici centrées autour de la moyenne d'ensemble, $\overline{\boldsymbol{x}}_{t,(i,j)}$, représentée par la ligne horizontale en vert foncé et d'ordonnée 0, afin d'assurer une plus grande visibilité de la variabilité. Un échantillon de 100 membres de l'ensemble parmi les 150 est représenté. En rouge, la prévision déterministe centrée, $x^{\text{det}}_{t,(i,j)} - \overline{\boldsymbol{x}}_{t,(i,j)}$; en noir, l'analyse centrée, $y_{t,(i,j)} - \overline{\boldsymbol{x}}_{t,(i,j)}$; le reste des couleurs est dédié aux membres de l'ensemble, $x^{m}_{t,(i,j)} - \overline{\boldsymbol{x}}_{t,(i,j)}$.

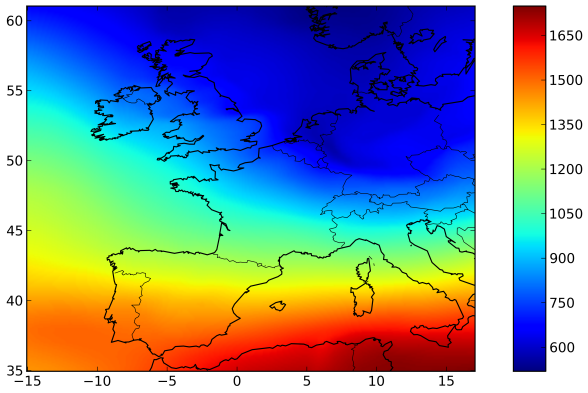

FIGURE 3.4 – Écart type de la moyenne temporelle de l'ensemble. Dans chaque cellule de coordonnées (i, j) est représentée la grandeur $\frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{M}\sum_{m=1}^{M}(x^{m}_{t,(i,j)} - \overline{x}_{t,(i,j)})^2}$. L'étendue sur toute la zone considérée de cet écart type est de 1200 Pa. On constate que les régions dans lesquelles se concentre la dispersion de l'ensemble (sud de la zone étudiée, couleurs chaudes) se trouvent être les régions dans lesquelles la pression moyenne est relativement basse (couleurs froides).

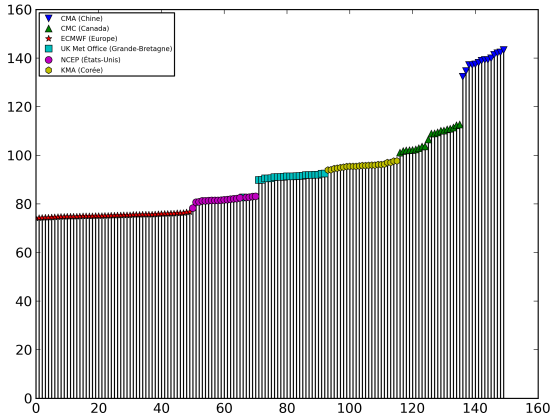

FIGURE 3.5 – Représentation graphique de la performance moyenne de chacun des membres de l'ensemble issus de TIGGE, triés par ordre croissant de *RMSE*, selon la formule $\frac{1}{Nx \times Ny} \sum_{(i,j) \in \text{Carte}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x^m_{(i,j),t} - y_{t,(i,j)})^2}$. Les prévisions sont réalisées à une échéance de 6 heures. Le symbole en haut de chacune des barres verticales indique l'origine (centre météorologique) du membre correspondant. Il s'agit de l'ensemble à 150 membres, ne comprenant pas les prévisions déterministes de Météo France ou d'ECMWF. Remarquons que ce sont les prévisions du centre européen qui réalisent les erreurs les plus basses.

font des erreurs comparables, ce qui donne cette répartition des scores sous forme d'un escalier où chaque marche correspond à un centre de prévision. Les valeurs extrêmes de la *RMSE* sont 74,17 Pa et 143,33 Pa et la moyenne arithmétique est 91,76 Pa. Il est à noter que ces valeurs indiquent que les prévisions semblent déjà précises :

$$\frac{91{,}76}{101325} = 0{,}09\% \,.$$

Mais il semble plus pertinent de comparer cette *RMSE* moyenne à l'étendue des données :

$$\frac{91{,}76}{2300} = 3{,}99\% \,.$$

Bien que déjà précises, ces prévisions peuvent encore être améliorées et d'éventuels gains pratiques ont un intérêt immédiat dans les prévisions opérationnelles.
En complément, les figures 3.6 et 3.7 reflètent la diversité spatiale et temporelle en ce qui concerne les meilleurs membres par cellule. Dans ces graphiques sont représentés respectivement le meilleur membre sur toute la période et l'origine de la meilleure moyenne d'ensemble à la première échéance. Bien que ECMWF fournit les membres ayant les *RMSE* les plus basses lorsqu'on les évalue globalement (cf. figure 4.4), les meilleurs membres cellules par cellules peuvent varier au vu de la figure 3.6. Nous remarquons aussi que cette variation n'est pas constante en temps vu les différences entre 3.6 et 3.7. En résumé, cet ensemble de prévision TIGGE est marqué par une diversité spatiale et temporelle.

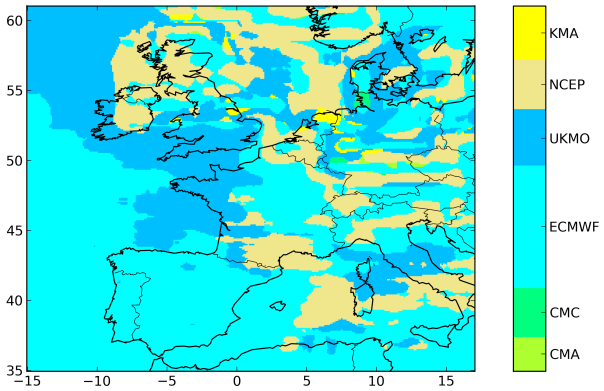

FIGURE 3.6 – Représentation graphique dans chaque cellule (i, j) de la carte de $m^\star(i, j)$, du meilleur membre annuel de l'ensemble issu de TIGGE (constitué de $M = 150$ membres) pour les prévisions à horizon de six heures. Ce meilleur membre est celui qui possède la *RMSE* la plus basse dans la cellule (i, j). Les centres de prévisions ont les origines suivantes : CMA - Chine, CMC - Canada, ECMWF - Europe, UKMO - Grande-Bretagne, NCEP - États-Unis, KMA - Corée. Les meilleurs modèles sont issus d'ECMWF (50% de la carte), UKMO (25%) et NCEP (20%).

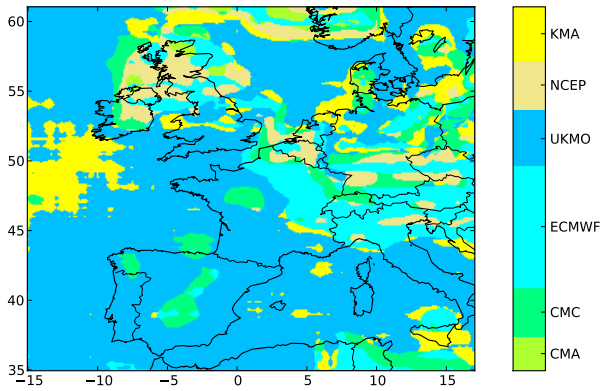

FIGURE 3.7 – Représentation graphique dans chaque cellule (i, j) de la carte de $e^\star(i, j)$, de la meilleure moyenne d'ensemble parmi les moyennes d'ensemble des centres régionaux de TIGGE (constitué de $M = 150$ membres, répartis dans 6 centres régionaux), lors du premier jour de la période considérée, pour les prévisions à horizon de six heures. Cette meilleure moyenne d'ensemble est celle qui possède la *RMSE* la plus basse dans la cellule (i, j). Les centres de prévision ont les origines suivantes : CMA - Chine, CMC - Canada, ECMWF - Europe, UKMO - Grande-Bretagne, NCEP - États-Unis, KMA - Corée.

Oracle

Nous mentionnons, en partie 3.1.2, que le calcul du regret dépend de la classe de référence choisie. Les classes de référence les plus classiques sont celles constituées des membres de l'ensemble (dont les performances sont déterminées a posteriori), des combinaisons convexes et des combinaisons linéaires. Une fois fixés l'ensemble de prévision et les observations sur une période de temps donnée, l'oracle issu de la classe de référence est entièrement déterminé. Dans le cas où ce sont des combinaisons, il existe alors un unique jeu de poids qui engendre la prévision convexe ou linéaire. Les trois classes de comparaisons d'intérêt sont les suivantes :

1. le meilleur membre de l'ensemble de prévision, $\arg\min_{m=1,\dots,M} L_{m,T}$. Dans notre cas de figure, avec l'ensemble à 152 membres (comprenant donc les membres de TIGGE et les deux déterministes Météo France et les prévisions d'ECMWF), il s'agit systématiquement de la prévision déterministe de Météo France ;

2. la meilleure combinaison convexe d'experts, $\arg\min_{\boldsymbol{u}\in\mathcal{P}} L_T(\boldsymbol{u})$ où $\mathcal{P}$ est le simplexe de $\mathbb{R}^M$;

3. la meilleure combinaison linéaire d'experts, $\arg\min_{\boldsymbol{u}\in\mathbb{R}^M} L_T(\boldsymbol{u})$.

Les figures 3.8 et 3.9 représentent les poids constants des oracles, respectivement convexe et linéaire.

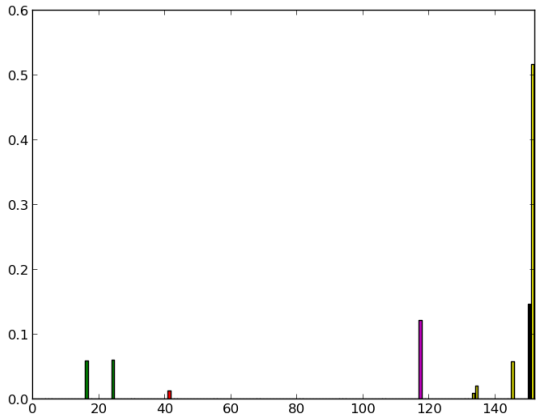

FIGURE 3.8 – Représentation graphique des poids de l'oracle convexe pour l'ensemble complet (152 membres). La meilleure combinaison convexe constante tire parti uniquement de quelques membres de l'ensemble : seuls les poids d'une dizaine de membres diffèrent de zéro à 10^{-6} près. Notons que ces poids sont bien tous compris entre 0 et 1 et leur somme est égale à 1. Notons que les deux derniers poids sont affectés aux prévisions déterministes de l'ECMWF et de Météo France, dans cet ordre.

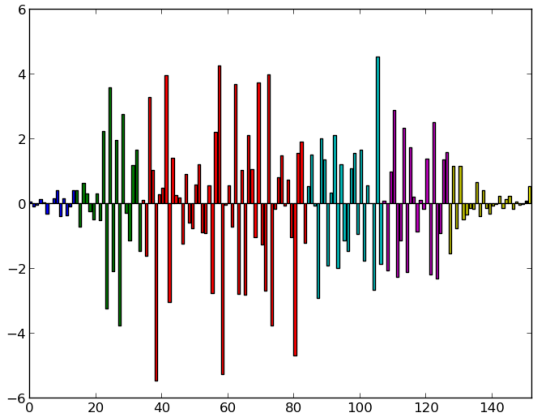

FIGURE 3.9 – Représentation graphique des poids de l'oracle linéaire pour l'ensemble complet (152 membres). La meilleure combinaison linéaire constante tire parti de tous les membres de l'ensemble : ils sont tous visiblement différents de zéro. Notons que ces poids varient dans $\mathbb{R}$. La somme de ces poids est proche de 1 (à 10^{-4} près), et cela n'est pas nécessairement attendu par la théorie dans le cas des poids linéaires.

Prévision de référence

Les garanties théoriques de maîtrise du regret assurent des performances robustes par rapport à des classes d'approximation intéressantes, quelles que soient les prévisions en entrée et les analyses à prévoir. Néanmoins, la stratégie de régression ridge assure un score de regret négatif ou nul par rapport à toute combinaison linéaire, comme mentionné dans la partie 3.1.2, uniquement dans le cas asymptotique. Ce qui signifie que pour un horizon infini, les prévisions issues de la stratégie ridge sont assurées d'être au moins aussi performantes que celles de l'oracle linéaire si l'on ajoute l'hypothèse supplémentaire que les poids sont contenus dans une boule de rayon arbitraire, cf. 3.1.3. Or, ces conditions ne sont jamais vérifiées en pratique, tout simplement, puisque l'on cherche à prévoir des analyses durant une période de temps bien délimitée. Il est donc possible que la transposition de ces garanties théoriques ne soit pas vérifiée pour une période finie. Dans les autres domaines d'application où ont été employées ces méthodes, par exemple DEVAINE, GAILLARD, GOUDE et STOLTZ [Dev+13], lorsque les garanties asymptotiques assurent de battre l'oracle linéaire, les prévisions agrégées non-asymptotiques ont des performances typiques comprises entre celles de l'oracle convexe et celles de l'oracle linéaire.

Donc, en pratique, pour obtenir une appréciation non-asymptotique des performances de la stratégie ridge, il est pertinent de sélectionner une prévision qu'un spécialiste de la communauté météorologique reconnaît comme étant assez précise pour la considérer comme raisonnable pour de la prévision à court terme dans un cadre déterministe. Notons qu'étant donné qu'il s'agit de stratégie naturelle pour un spécialiste, nous ne faisons pas rentrer en ligne de compte les oracles en supposant que le météorologue n'y a pas accès. Trois choix s'offrent à nous dans la détermination de ce point de référence :

- la moyenne d'ensemble comprenant les 150 membres issus de TIGGE ;
- la prévision déterministe fournie par le centre Européen (ECMWF) ;
- la prévision déterministe fournie par Météo France.

On aurait pu ajouter les moyennes d'ensembles de chacun des centres régionaux. Cependant, comme le montre le graphique 4.4, chacun des centres régionaux est strictement moins bon que les prévisions des membres d'ECMWF et la prévision de la moyenne d'ensemble ECMWF est moins bonne que la prévision déterministe ECMWF.

Étant donné qu'on s'attache à prévoir l'analyse Météo France, on pourrait souhaiter mettre en concurrence avec les prévisions précédentes la moyenne d'ensemble des prévisions de Météo France, mais ces données ne sont pas aisément disponibles et ont donc été écartées.

Pour la prévision à horizon de 6 heures, les performances sont reportées dans le tableau 3.2.

origine	*RMSE* (Pa)
moyenne d'ensemble	52,39
déterministe ECMWF	34,19
déterministe Météo France	30,13

TABLE 3.2 – Performances des trois prévisions candidates au rôle de prévision de référence dans le cadre non-asymptotique.

Parmi ces trois choix, la prévision déterministe de Météo France obtient le meilleur score en *RMSE*. C'est donc celle que nous choisissons comme point de référence. Il ne s'agit pas pour autant de la « réalité terrain » que constitue l'analyse et qui entre dans le calcul direct de la *RMSE*. La prévision de référence joue un rôle similaire dans un cadre non-asymptotique à celui que joue l'oracle dans un cadre asymptotique. Dans la suite de ce chapitre, les évaluations des différences relatives de *RMSE* ont donc lieu par rapport à la prévision déterministe de Météo France (voir par exemple 3.3). Si MF correspond à la prévision déterministe de Météo France et A à la prévision issue d'un algorithme donné, La différence relative des gains est définie selon la formule

$$\Delta_{\%}(\mathrm{A}, \mathrm{MF}) = \frac{RMSE_{\mathrm{MF}} - RMSE_{\mathrm{A}}}{RMSE_{\mathrm{MF}}} .$$

Cette référence étant désormais déterminée, il ne nous reste plus qu'à chercher à battre cette prévision. C'est la condition *sine qua non* qui permettra d'intéresser potentiellement les acteurs opérationnels. En ce qui concerne les oracles (à horizon fini), leurs scores sont reportés au tableau 3.3. Dépasser les performances des oracles convexes ou linéaires va s'avérer un objectif des plus ambitieux : nous ne pouvons qu'espérer nous rapprocher de la prévision de l'oracle convexe.

3.2.3 Résultats

À ce stade de la description, maintenant qu'ont été déterminées les limites de la période, sélectionnés les acteurs en jeu, et calculés les oracles, il nous reste à réaliser les stratégies d'agrégation proprement dites. Les variantes possibles dans les stratégies qui vont suivre portent sur les différentes manières d'employer les deux paramètres de la régression ridge escomptée : λ, le coefficient de régularisation et γ, le coefficient multiplicatif d'escompte. Les trois variantes possibles sont décrites dans les paragraphes suivants :

- stratégie ridge escomptée menée sur une grille fixe de paramètres ;
- stratégie ridge escomptée avec adaptation cellule par cellule des paramètres ;
- stratégie ridge escomptée en suivant une méthode semi-automatique, adaptative en temps et en espace de choix des paramètres.

Stratégie ridge sur une grille fixe

Dans l'étude décrite dans ce paragraphe, les paramètres sont vus comme constants $\lambda_t = \lambda$ et $\gamma_t = \gamma$ et communs à toutes les cellules. Lorsque, pour la première fois, nous menons les calculs de régression sur le jeu de données, nous ne savons rien au sujet des ordres de grandeur des paramètres de régularisation et d'escompte qui conduiront aux meilleures performances. La procédure préliminaire cherche à déterminer des ordres de grandeurs raisonnables pour ces paramètres.

Un raisonnement qualitatif à partir de la formule donnant l'expression des poids de la régression ridge 3.1.3 permet de déterminer un premier ordre de grandeur du paramètre de régularisation. En effet, un raisonnement sur la dimension, sachant que les poids $\boldsymbol{u}_t$ sont sans dimension, λ est du même ordre de grandeur que la somme en temps des erreurs. Dans l'exemple présent, la pression réduite au niveau de la mer prend des valeurs autour de 10^5Pa. Les premières estimations de performances sont donc faites en faisant varier λ autour de cet ordre de grandeur.

De tels indices qualitatifs pour le facteur multiplicatif d'escompte γ n'existent pas. En pratique, on commence donc par évaluer manuellement les performances avec plusieurs couples (λ, γ) afin de réduire le champ des valeurs possibles. Calculer de nombreuses fois les scores sur toutes les $261 \times 321 = 83781$ cellules de la carte s'avère trop long : on se restreint donc à un nombre limité de cellules choisies pour estimer les performances. Afin d'obtenir une vue spatiale globale de la performance malgré tout, on sélectionne ces cellules aléatoirement parmi toutes les cellules de la carte. Et afin de pouvoir comparer raisonnablement ces premières estimations entre elles, on fixe la graine du générateur pseudo-aléatoire déterminant ces cellules : toutes les estimations sont donc fondées sur le même ensemble de cellules.

Après cette étude préliminaire et une première calibration du paramètre γ d'escompte sur une grille logarithmique, reporté à la figure 3.11, nous déduisons la grille de paramètres suivants : γ varie dans l'ensemble $\{k \times 5, \forall k \in \{0, \ldots, 40\}\}$ et, indépendamment,

λ varie entre $0{,}40 \times 10^5$ et $1{,}50 \times 10^5$ par pas de $0{,}10 \times 10^5$. Les résultats sont reportés graphiquement à la figure 3.12. Une fois ces paramètres optimaux a posteriori trouvés, nous opérons les calculs sur toute la carte avec ces derniers : les résultats associés sont reportés dans la table 3.3 et la figure 3.10. Concernant les choix des poids initiaux, les performances sont systématiquement optimales pour un choix de mesure de Dirac avec un poids initial unité sur la prévision déterministe de Météo France. Par ailleurs, il se trouve qu'il s'agit aussi du choix le plus cohérent : c'est en effet cette prévision déterministe qui est employée en opérationnel. Parmi les choix mis à l'épreuve et écartés, le vecteur de poids initial est systématiquement convexe afin de confèrer une cohérence physique aux prévisions agrégées dès les premières échéances. Concernant les choix de la période d'entraînement pendant laquelle les calculs des performances sont ignorés, voir 3.3.3. Des représentations complémentaires des résultats sous forme d'histogrammes sont fournies aux figures 3.13 pour la comparaison entre la moyenne d'ensemble et la régression ridge, 3.14 pour la comparaison des oracles avec la régression ridge, 3.15 pour la visualisation des différences relatives de *RMSE* par rapport au déterministe Météo France et 3.16 pour une visualisation de ces mêmes données pour tous les acteurs en jeu sous forme de diagrammes en boîtes.

Type de prévision	*RMSE* (Pa)	différence relative (%)
Moyenne d'ensemble	53,12	−74,79
Déterministe	30,39	0,00
Oracle convexe	24,44	19,59
Oracle linéaire	18,10	40,45
Agrégation ridge (avec paramètres optimaux rétrospectifs)	24,90	17,35

TABLE 3.3 – *RMSE* : Scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres optimisés au préalable : $(\lambda; \gamma) = (60\,000; 15)$. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

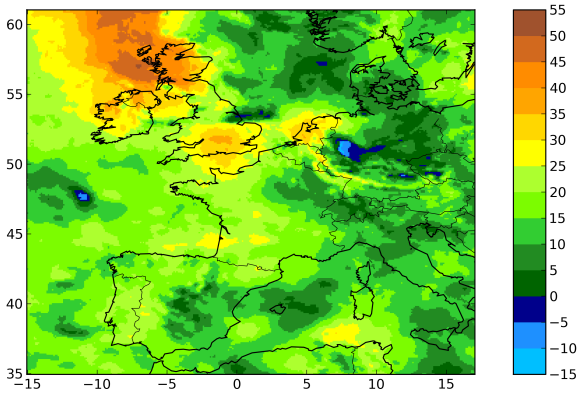

FIGURE 3.10 – Représentation graphique des différences relatives de *RMSE* entre l'agrégé avec paramètres optimaux rétrospectifs et le déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i,j) est représentée la différence relative : $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est la *RMSE* moyenne de la prévision agrégée (respectivement déterministe) dans la cellule (i,j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

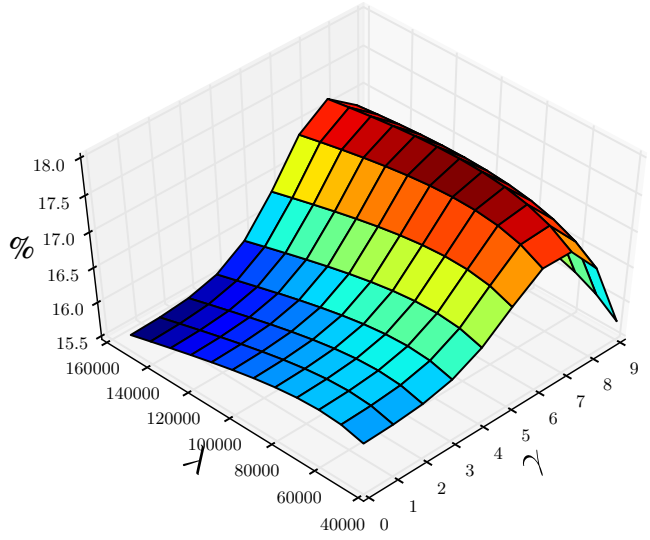

FIGURE 3.11 – Représentation graphique de l'influence des paramètres sur les performances de l'algorithme de régression ridge escompté pour la prévision à horizon 6 heures de la pression réduite au niveau de la mer. La cote représente la différence relative (en pourcentage), entre d'une part, la prévision agrégée avec les paramètres en abscisse et ordonnée et, d'autre part, la prévision du déterministe Météo France. Le facteur d'escompte γ varie en puissance de 10 entre 10^{-7} et 10^{4} et le facteur de régularisation λ varie par pas linéaires de 10000 dans le segment $[40000, 160000]$. Le maximum sur cette plage de paramètre a pour valeur 17,31% et a pour antécédent le couple $(\lambda, \gamma) = (100\,000, 100)$.

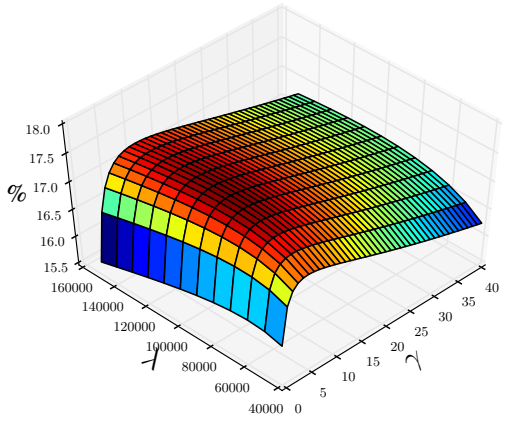

FIGURE 3.12 – Représentation graphique de l'influence des paramètres sur les performances de l'algorithme de régression ridge escompté pour la prévision à horizon 6 heures de la pression réduite au niveau de la mer. La cote représente la différence relative (en pourcentage), entre d'une part, la prévision agrégée avec les paramètres en abscisse et ordonnée et, d'autre part, la prévision du déterministe Météo France. Le facteur d'escompte γ varie dans l'ensemble $\{k \times 5, \forall k \in \{0, \ldots, 40\}\}$ (en abscisse sont représentés les k correspondants) et le facteur de régularisation λ varie par pas de 10000 dans le segment $[40000, 160000]$. Le maximum sur cette plage de paramètre (dont tout laisse à penser qu'il s'agit d'un maximum global) a pour valeur 17,61% et a pour antécédent le couple $(\lambda, \gamma) = (100\,000, 35)$.

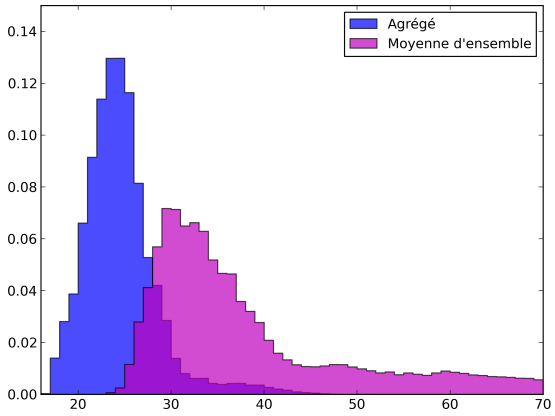

FIGURE 3.13 – Distribution des *RMSE* de l'agrégé (en bleu) et de la moyenne d'ensemble (en magenta). Les individus représentés sont les *RMSE* pour l'agrégé $r^{\text{agr}}_{(i,j)}$ et les *RMSE* pour la moyenne d'ensemble $r^{\text{ens}}_{(i,j)}$ des cellules (i, j) de la carte dans son intégralité. Les prévisions sont réalisées à une échéance de 6 heures, pour des paramètres optimaux de la régression ridge. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). La *RMSE* est prise par rapport à l'analyse Météo France. Le poids initial est une mesure de Dirac sur le déterministe Météo France.

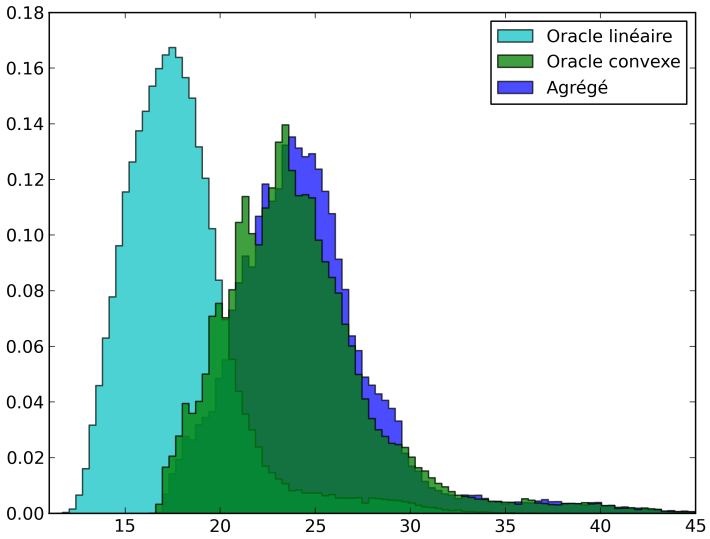

FIGURE 3.14 – Histogramme des *RMSE* de l'oracle linéaire (en cyan), de l'oracle convexe (en vert) et de l'agrégé (en bleu). Les individus représentés sont les *RMSE* pour l'oracle linéaire $r^{\text{lin}}_{(i,j)}$, les *RMSE* pour l'oracle convexe $r^{\text{cvx}}_{(i,j)}$ et les *RMSE* pour l'agrégé $r^{\text{agr}}_{(i,j)}$ des cellules (i,j) de la carte dans son intégralité. Les prévisions sont réalisées à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). La *RMSE* est prise par rapport à l'analyse Météo France. Le poids initial est une mesure de Dirac sur le déterministe Météo France.

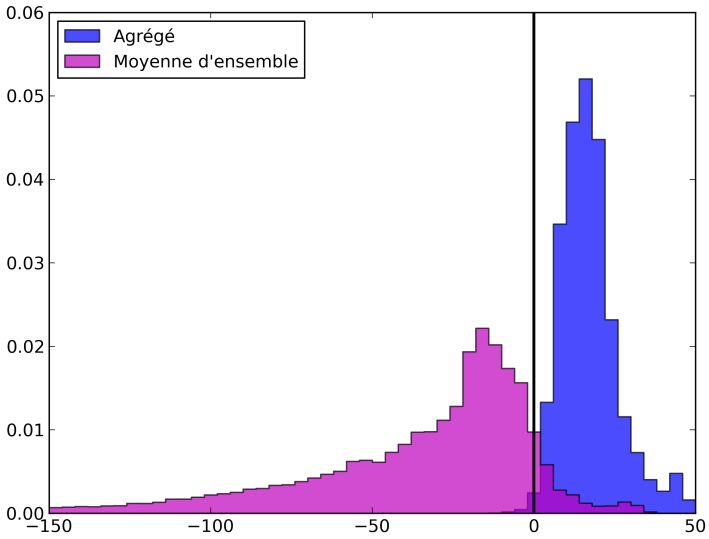

FIGURE 3.15 – Histogramme de la différence relative par rapport au déterministe de la *RMSE* de la moyenne d'ensemble (en magenta) et de l'agrégé (en bleu). Les individus représentés sont les différences relatives en *RMSE* pour l'agrégé $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ et les différences relatives en *RMSE* pour la moyenne d'ensemble $\left(r^{\text{det}}_{(i,j)} - r^{\text{ens}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$ et $r^{\text{det}}_{(i,j)}$) est la *RMSE* moyenne de la prévision agrégée (respectivement déterministe et prévision de la moyenne d'ensemble) dans la cellule (i, j). Les prévisions sont à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). La *RMSE* est prise par rapport à l'analyse Météo France. Le poids initial est une mesure de Dirac sur le déterministe Météo France.

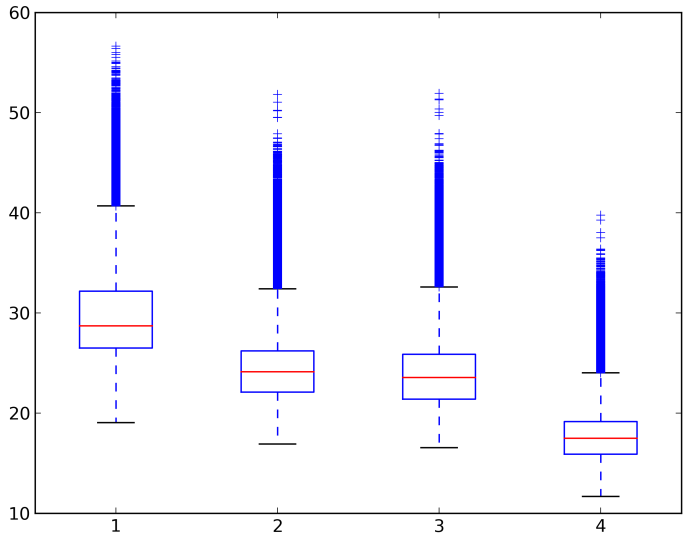

FIGURE 3.16 – Diagramme en boîte des *RMSE* à horizon de 6 heures entre l'analyse Météo France et respectivement, le déterministe (1), l'agrégé issu de la régression ridge (2), l'oracle convexe (3) et l'oracle linéaire (4). Les individus représentés sont les *RMSE* pour l'agrégé $r^{\text{agr}}_{(i,j)}$, les *RMSE* pour la moyenne d'ensemble $r^{\text{ens}}_{(i,j)}$, les *RMSE* pour l'oracle convexe $r^{\text{cvx}}_{(i,j)}$ et les *RMSE* pour l'oracle linéaire $r^{\text{lin}}_{(i,j)}$ des cellules (i,j) de la carte dans son intégralité. À un individu plus bas correspond une *RMSE* plus basse donc de meilleures performances. Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est une mesure de Dirac sur le déterministe Météo France.

Stratégie avec optimisation locale des paramètres

Dans cette partie, de la même manière que précédemment, les paramètres sont constants en temps $\lambda_t = \lambda$ et $\gamma_t = \gamma$, mais cette fois-ci, ils sont locaux, c'est-à-dire spécifiques à chacune des cellules. La grille de paramètres employée est la même que celle construite précédemment. Les poids initiaux et la période d'entraînement sont choisis identiques à ceux de la partie précédente et on emploie, là encore 1000 cellules pour mener cette étude. Cette stratégie est directement comparable à la stratégie précédente 3.2.3, pour laquelle les paramètres sont globaux selon les deux dimensions : nous ne modifions qu'un degré de liberté à la fois puisque si le couple de paramètres optimaux peut varier entre une cellule et sa voisine, il ne peut pas varier échéance après échéance. Étant donné cette souplesse nouvelle de la méthode par rapport à la précédente, nous anticipons une amélioration des performances en *RMSE*. Les résultats sont reportés pour les scores de *RMSE* et de biais dans les tables 3.4 et 3.5 où, dans le cas de la *RMSE*, sont aussi rappelés les résultats de la stratégie précédente avec meilleur paramètre a posteriori pour faciliter la comparaison.

Type de prévision	*RMSE* (Pa)	Différence relative (%)
Moyenne d'ensemble	52,39	−73,88
Déterministe	30,13	0,00
Oracle convexe	24,29	19,38
Oracle linéaire	18,02	40,18
Agrégation ridge (avec adaptation locale rétrospective des paramètres)	24,44	18,87
Agrégation ridge (avec paramètres optimaux rétrospectifs)	24,90	17,35

TABLE 3.4 – Stratégie de régression ridge avec optimisation locale des paramètres (excepté pour la dernière ligne). *RMSE* : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour un jeu de paramètres appartenant à la grille (4.2). 1000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

Pour cette stratégie, locale en espace, l'amélioration des différences relatives de performance en *RMSE* par rapport à la stratégie de la partie précédente 3.2.3 est de $18{,}87 - 18{,}07 = 0{,}80\%$. Cette amélioration est conforme à l'intuition et est d'ampleur modérée.

Stratégie avec optimisation locale et temporelle des paramètres

Dans la partie 3.2.3, qui dévoile les premiers résultats pratiques de la stratégie de régression ridge escomptée, nous réalisons la régression pour plusieurs couples de paramètres, ce qui nous permet de découvrir quels sont les paramètres optimaux a posteriori, c'est-à-dire une fois que tous les calculs ont été menés et que l'on connaît les scores associés à chaque couple de paramètres. Nous employons le même espace de paramètre que celui de la partie 3.2.3. Nous cherchons désormais à nous rapprocher d'une situation

Type de prévision	Biais (Pa)	Différence relative (%)
Moyenne d'ensemble	35,64	−53,70
Déterministe	23,19	0,00
Oracle convexe	18,54	20,05
Oracle linéaire	14,10	39,21
Agrégation ridge (avec adaptation locale rétrospective des paramètres)	18,72	19,27
Agrégation ridge (avec paramètres optimaux rétrospectifs)	18,84	18,76

TABLE 3.5 – Stratégie de régression ridge avec optimisation locale des paramètres (excepté pour la dernière ligne). Biais : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour un jeu de paramètres appartenant à la grille (4.2). 1000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

plus réaliste dans laquelle les paramètres sont mis à jour régulièrement à la volée en fonction des performances qu'ils induisent, c'est-à-dire introduire le degré de liberté temporel. En pratique, en effet, le prévisionniste ne dispose évidemment pas du jeu de paramètres optimal avant d'avoir réalisé la prévision. Il s'agit d'une version partiellement automatisée de la stratégie précédente. Pour mettre en oeuvre cette variante, il est nécessaire de déterminer au préalable une grille dans laquelle les couples de paramètres vont pouvoir évoluer séquentiellement. Les résultats reportés dans la figure 3.12, donnant la sensibilité aux paramètres, indiquent que la stratégie de régression ridge escomptée y est relativement peu sensible dès lors qu'une grille raisonnable a été fixée. En effet, la grille retenue dans 3.12 conduit à des différences relatives variant entre 15 et 18%. L'algorithme 7 réalise l'agrégation ridge escomptée semi-automatique dans chaque cellule.

Notons que l'algorithme 7 offre deux degrés de libertés nouveaux à l'espace des paramètres. En effet, ceux-ci peuvent varier non seulement en temps, mais aussi différer entre cellules, c'est-à-dire, varier en espace. Cependant, la grille de paramètres générée au début de l'algorithme reste commune à toutes les cellules. Enfin, en ce qui concerne l'initialisation des paramètres, en supposant n'avoir aucune information supplémentaire, on sélectionne le milieu de la grille de paramètres, de coordonnées $(\lfloor p/2 \rfloor; \lfloor q/2 \rfloor)$.

Le paramètre t_{bascule} donne le nombre de pas de temps durant lesquels l'agrégation séquentielle a lieu et est évaluée avec un certain couple de paramètres avant que ne soit donnée à l'algorithme la possibilité de changer de couple de paramètres. Remarquons que plus la période t_{bascule} est courte, plus s'accroît la capacité d'adaptation de l'algorithme. Néanmoins, il faut noter que le temps de calcul de cet algorithme partiellement automatisé est lui aussi largement augmenté.

Dans la partie 3.2.3, les paramètres sont globaux en temps et locaux en espace alors que dans celle-ci, ils peuvent varier selon les deux dimensions à la fois. La régression ridge escomptée partiellement automatique de la présente partie 3.2.3 peut donc direc-

---

Paramètres :

1. $t_{\text{bascule}} \in \{1, \dots, T\}$, la période de changement du couple de paramètres.
2. $p, q \in \mathbb{N}$, le nombre de paramètres différents de régularisation et d'escompte.
3. $\{\lambda_1, \dots, \lambda_p\} \times \{\gamma_1, \dots, \gamma_q\}$: la grille de paramètres retenue.

Initialisation : Le couple

$$(\lambda, \gamma)_0^* = (\lambda_{\lfloor p/2 \rfloor}, \gamma_{\lfloor q/2 \rfloor})$$

est sélectionné pour la première période.
Déroulement : Pour $k \in \{1, \dots, \lfloor T/t_{\text{bascule}} \rfloor\}$,

1. Pour tout $(\lambda, \gamma) \in \{\lambda_1, \dots, \lambda_p\} \times \{\gamma_1, \dots, \gamma_q\}$, la prévision agrégée issue de la régression ridge escomptée avec les paramètres (λ, γ) est calculée.
2. Celle des prévisions qui obtient la meilleure *RMSE* globale sur la période

   $$\{1, \dots, k \cdot t_{\text{bascule}}\}$$

   fournit le couple de paramètres retenu $(\lambda, \gamma)_k^*$, pour la période suivante

   $$\{k \cdot t_{\text{bascule}} + 1, \dots, \max((k+1) \cdot t_{\text{bascule}}, T)\}.$$

3. Pour tout $s \in \{(k-1) \cdot t_{\text{bascule}} + 1, \dots, \max(k \cdot t_{\text{bascule}}, T)\}$, $\widehat{y}_s\big((\lambda, \gamma)_{k-1}^*\big)$ constitue la prévision agrégée sur cette période.

Fin : Si $T \geqslant \lfloor T/t_{\text{bascule}} \rfloor \cdot t_{\text{bascule}} + 1$, alors,

$$\forall s \in \{\lfloor T/t_{\text{bascule}} \rfloor \cdot t_{\text{bascule}} + 1, \dots, T\},$$

$\widehat{y}_s\big((\lambda, \gamma)_{\lfloor T/t_{\text{bascule}} \rfloor}^*\big)$ constitue la prévision agrégée sur cette période.

Algorithm 7: Stratégie ridge escomptée partiellement automatique sur une grille de paramètres.

tement être comparée à la version précédente de la partie 3.2.3. De plus, étant donné que cet algorithme réalise une adaptation des paramètres en ligne contrairement au précédent qui sélectionnait le meilleur paramètre rétrospectif, nous pouvons anticiper une diminution des performances. Reste à déterminer l'amplitude de cette diminution. Les performances de l'algorithme sur un échantillon de 1000 cellules pour la grille de paramètres ci-dessous (4.2) (régularisation, escompte) sont résumées dans les tableaux 3.6 et 3.7 pour la *RMSE* et le biais.

$$\{40000; 70000; 100000; 130000; 160000\} \times \{0; 25; 50; 75; 100\} \tag{3.4}$$

Type de prévision	*RMSE* (Pa)	Différence relative (%)
Moyenne d'ensemble	52,39	−73,88
Déterministe	30,13	0,00
Oracle convexe	24,29	19,38
Oracle linéaire	18,02	40,18
Agrégation ridge (paramètres optimaux rétrospectifs)	24,90	17,35
Agrégation ridge (adaptation locale rétrospective des paramètres)	24,44	18,87
Agrégation ridge (adaptation locale en ligne sur une grille de paramètres)	24,80	17,69

Table 3.6 – Stratégie partiellement automatique. *RMSE* : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres pouvant varier tous les $t_{\text{bascule}} = 30$ jours sur la grille (4.2). 1000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

Type de prévision	Biais (Pa)	Différence relative (%)
Moyenne d'ensemble	35,64	−53,70
Déterministe	23,19	0,00
Oracle convexe	18,54	20,05
Oracle linéaire	14,10	39,21
Agrégation ridge (paramètres optimaux rétrospectifs)	18,84	18,76
Agrégation ridge (adaptation locale rétrospective des paramètres)	18,72	19,27
Agrégation ridge (adaptation locale en ligne sur une grille de paramètres)	18,68	19,42

Table 3.7 – Stratégie partiellement automatique. Biais : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres pouvant varier tous les $t_{\text{bascule}} = 30$ jours sur la grille (4.2). 1000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

Cette automatisation partielle conduit à une dégradation légères des performances. En effet, le gain algébrique des différences relatives de performance en *RMSE* par rapport à la stratégie de la partie précédente 3.2.3 est de $17{,}69 - 18{,}87 = -1{,}18\%$. En adaptant l'algorithme de régression ridge escompté avec paramètres locaux en espace à un cadre réaliste où les paramètres optimaux sont inconnus à l'avance, on s'expose donc à une perte de $-1{,}18\%$ ce qui est là encore modéré. Notons que nous prenons bien soin de comparer des quantités comparables en mettant en regard les performances de l'algorithme 7 et de la stratégie de régression ridge avec paramètres locaux de la partie 3.2.3 : seul le degré de liberté temporel a été ajouté.

3.3 Réactivité de l'algorithme à différentes formulations

3.3.1 Performance des centres régionaux de prévision

Comme cela a été précisé en partie 3.2.1, l'ensemble employé jusqu'à présent est constitué de tous les membres issus des centres régionaux de prévision procurés par TIGGE, ce qui donne un total de 150 membres auxquels s'ajoutent les éventuelles prévisions déterministes issues d'ECMWF et de Météo France. Cet ensemble est le plus complet qui soit à notre disposition pour prévoir la pression réduite au niveau de la mer dans la région étudiée. Aucune sélection parmi les centres régionaux n'a lieu car on suppose que tous les centres régionaux sont susceptibles d'apporter de l'information à l'ensemble. Cette supposition raisonnable mérite néanmoins d'être évaluée quantitativement et c'est l'objet de la présente partie. Nous cherchons à répondre à la question : quels sont les potentiels de prévision de chaque centre régional de prévision indépendamment les uns des autres ?
Tour à tour, nous sélectionnons donc les centres de prévision régionaux et ajoutons les deux prévisions déterministes à l'ensemble formé. Le poids initial est une mesure de Dirac centrée sur la prévision déterministe de Météo France. Cette prévision constitue, en effet, la meilleure prévision à notre disposition. Si l'on considère la situation avec beaucoup de recul, les membres de l'ensemble peuvent être vus comme des adjuvants de la prévision déterministe de Météo France. Les résultats de ces procédures sont résumés dans le tableau 3.8.

Au vu de la table 3.8, il n'apparaît pas de lien clair et direct entre les résultats de la régression ridge escomptée appliquée à un ensemble régional et la performance de la moyenne d'ensemble de ce centre régional.
La zone étudiée contient l'Europe, ce qui conduit à se poser la question : quels sont les résultats dans la procédure précédente du centre Européen (ECMWF) ? Nous observons que la moyenne d'ensemble d'ECMWF obtient la *RMSE* la plus basse parmi toutes les moyennes d'ensemble issues des centres régionaux : 55,83 Pa. En revanche, la *RMSE* de la prévision agrégée est de 24,88 Pa : cela fait d'elle une prévision de légèrement moindre qualité que la plupart de ses homologues. Les prévisions des membres ECMWF sont proches de la prévision déterministe Météo France, ce qui explique le bon résultat de la moyenne d'ensemble et le fait que la régression ridge associée ne parvienne pas à

Origine (acronyme)	*RMSE* (Pa)	Différence relative (%)
Ridge escompté		
Chine (CMA)	24,92	17,29
Canada (CMC)	24,70	18,01
Europe (ECMWF)	24,88	17,42
Corée (KMA)	24,72	18,32
États-Unis (NCEP)	24,18	19,73
Grande-Bretagne (UKMO)	24,61	17,94
Ensemble complet	24,69	18,04
Moyenne d'ensemble		
Chine (CMA)	67,17	−122,96
Canada (CMC)	65,00	−115,73
Europe (ECMWF)	55,83	−85,29
Corée (KMA)	69,96	−130,87
États-Unis (NCEP)	70,74	−134,78
Grande-Bretagne (UKMO)	69,56	−132,21
Ensemble complet	52,39	−73,88

TABLE 3.8 – *RMSE* : scores et différences relatives par rapport à la prévision déterministe de Météo France pour différentes prévisions agrégées à un horizon de 6 heures pour des paramètres optimisés au préalable : $(\lambda; \gamma) = (100\,000; 50)$. 1000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

atteindre l'un des meilleurs scores. Le centre régional ECMWF formule des prévisions proches de celle de Météo France mais qui apportent peu d'information complémentaire à cette prévision déterministe.
Le résultat le plus notable concerne la performance du centre NCEP : la *RMSE* de sa moyenne d'ensemble est 70,74 Pa, la plus élevée parmi les six possibilités. Dans le même temps, la *RMSE* de la stratégie ridge escomptée correspondante est 24,18 Pa, ce qui en fait la moins élevée. Autrement dit, les prévisions du centre régional NCEP engendrent l'agrégation ridge la plus précise malgré une moyenne d'ensemble qui obtient, au contraire, le score le moins bon de tous. Au contraire du cas précédent, les prévisions des membres NCEP sont clairement distinctes de la prévision déterministe Météo France. Cela explique le résultat de la moyenne d'ensemble médiocre et le fait que la régression ridge associée atteigne le meilleur score : le centre régional NCEP formule des prévisions fortement distinctes de celle de Météo France et qui, de ce fait, apportent beaucoup d'information complémentaire à cette prévision déterministe.
Il est à noter que les stratégies ridges mettant en jeu uniquement l'ensemble britannique (UKMO) ou nord-américain (NCEP) ont des scores en *RMSE* meilleurs que ceux de l'ensemble complet.

3.3.2 Dynamique des poids

Les figures 3.17, 3.18 et 3.19 représentent l'évolution des poids de la régression dans une cellule donnée, ici de coordonnées (5° O, 55° N), sur l'ensemble de la période consi-

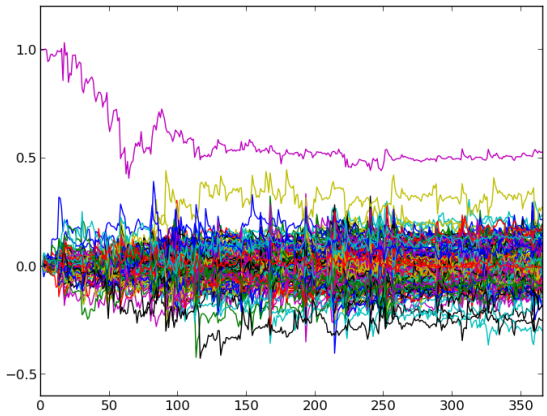

FIGURE 3.17 – Représentation graphique de l'évolution des poids en fonction du temps dans la cellule de coordonnées (5° E, 55° N). La régularisation λ de l'algorithme d'agrégation ridge est ici de 10 000 Pa. Comparativement à l'optimum empirique pour cette valeur, il s'agit d'une valeur plus faible d'un ordre de grandeur. Dans cette configuration, la variabilité des poids est importante lorsqu'on la compare aux figures 3.18 et 3.19. .

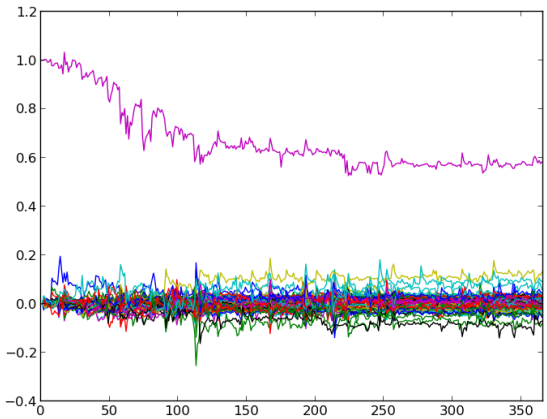

FIGURE 3.18 – Représentation graphique de l'évolution des poids en fonction du temps dans la cellule de coordonnées (5° E, 55° N). La régularisation λ de l'algorithme d'agrégation ridge est ici de 100 000 Pa. C'est la valeur du paramètre λ de la régression ridge qui fournit le score optimal. Dans cette configuration, la variabilité des poids est plus importante que dans le cas de la figure 3.19 et légèrement que dans la figure 3.17.

dérée pour différentes valeurs du paramètre de régularisation. La cellule, sélectionnée aléatoirement, est retenue car elle est typique : les poids n'y font pas montre d'une évolution pathologique par rapport à la majorité des autres cellules. Le paramètre d'escompte γ est fixé à 35 et le paramètre de régularisation λ de la régression ridge varie sur une échelle de deux ordres de grandeur autour de la valeur empiriquement optimale : $\lambda \in \{10\,000\,;\,100\,000\,;\,1\,000\,000\}$.

Le poids initial est fixé à 1 pour le déterministe Météo France et à 0 pour tous les autres membres de l'ensemble. C'est la configuration qui, toutes choses égales par ailleurs, entraîne les meilleures performances.
Avec un tel choix de poids initiaux, les gains sont modérément sensibles au paramètre de régularisation et très peu sensibles au paramètre d'escompte. Cela est visible à la figure 3.12. Les scores et différences relatives par rapport à la prévision déterministe en fonction de la régularisation sont donnés au tableau 3.9.

Type de prévision	*RMSE* (Pa)	Différence relative (%)
Moyenne d'ensemble	29,82	−0,07
Déterministe	27,88	0,00
Oracle convexe	21,31	23,56
Oracle linéaire	14,44	48,20
Paramètre de régularisation		
6×10^4	22,51	19,26
6×10^5	22,15	20,56
6×10^6	23,99	13,97

TABLE 3.9 – Scores et différences relatives par rapport à la prévision déterministe de Météo France en fonction de la régularisation.

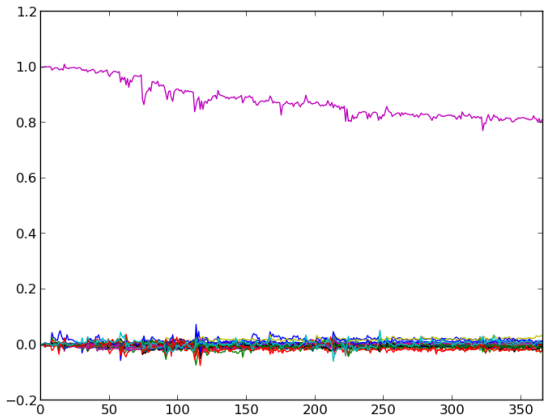

FIGURE 3.19 – Représentation graphique de l'évolution des poids en fonction du temps dans la cellule de coordonnées (5° O, 45° N). La régularisation λ de l'algorithme d'agrégation ridge est ici de 1 000 000 Pa. Comparativement à l'optimum empirique pour cette valeur, il s'agit d'une valeur plus forte d'un ordre de grandeur. Dans cette configuration, la variabilité des poids est faible comparativement aux figures 3.18 et 3.17.

3.3.3 Influence et choix de la période d'entraînement

Comment choisir la période d'entraînement pendant laquelle les calculs des performances sont ignorés ? Notre réponse à cette question se fonde sur deux constations. La première est la stagnation relative des poids remarquées à la figure 3.17. En effet, sur cette figure, les poids semblent entrer dans une forme de régime permanent, fluctuant autour d'une valeur moyenne (distincte pour chaque poids) après environ 100 jours. L'étude empirique dont les résultats sont reportés à la figure 3.20 conduit à la seconde constatation. Cette figure montre l'évolution de la différence relative de *RMSE* entre la prévision déterministe d'une part et la prévision agrégée (en bleu), l'oracle convexe (en rouge) et l'oracle linéaire (en noir) d'autre part, en fonction de la période d'entraînement sur un échantillon de 1000 cellules. La pente de la dérivée des différences relatives de *RMSE* diminue après environ 40 jours, ce qui signifie que toute période d'entraînement de plus de 40 jours affectera peu les différences relatives par rapport à des périodes de plus courtes durées. Ces deux remarques faites, nous choisissons une période d'entraînement de 100 jours.

3.4 Conclusion

Dans ce premier chapitre applicatif, nous présentons le jeu de données météorologiques de pression réduite au niveau de la mer en détaillant les origines et les spécificités des simulations et des analyses. Cela nous permet alors d'exposer puis de mettre en pratique un plan d'étude de ces simulations. À cet effet, différentes versions de l'algorithme d'agrégation ridge sont décrites (partie 3.1.6) puis implémentées (partie 3.2.3) par ordre croissant d'automatisation. En effet, en opérationnel, l'algorithme, si l'on veut qu'il soit pérenne, doit être le plus autonome possible et se fonder le moins possible sur la calibration manuelle des paramètres par un statisticien.

Les résultats sont satisfaisants et ont un potentiel pratique convaincant : la régression ridge avec paramètres optimaux rétrospectifs améliore les résultats de prévision opérationnel de 17,35% (partie 3.2.3) tandis que la méthode complètement automatique, avec adaptation locale en ligne sur une grille de paramètres, nécessairement plus coûteuse en temps de calcul, conduit à une différence relative de 17,69% (partie 3.2.3). Pour mettre en perspective ces nombres, soulignons que les améliorations des meilleurs modèles de simulation par les recherches fondamentales en météorologie apportent un gain de 1 à 2% par an. L'un des intérêts des algorithmes d'agrégation est qu'ils accroisseront la précision des prévisions quelles que soient les simulations sous-jacentes. Nous détaillons pour finir (partie 3.3) les performances des centres régionaux ainsi que la dynamique des poids en fonction du paramètre de régularisation et justifions les choix de la période d'entraînement.

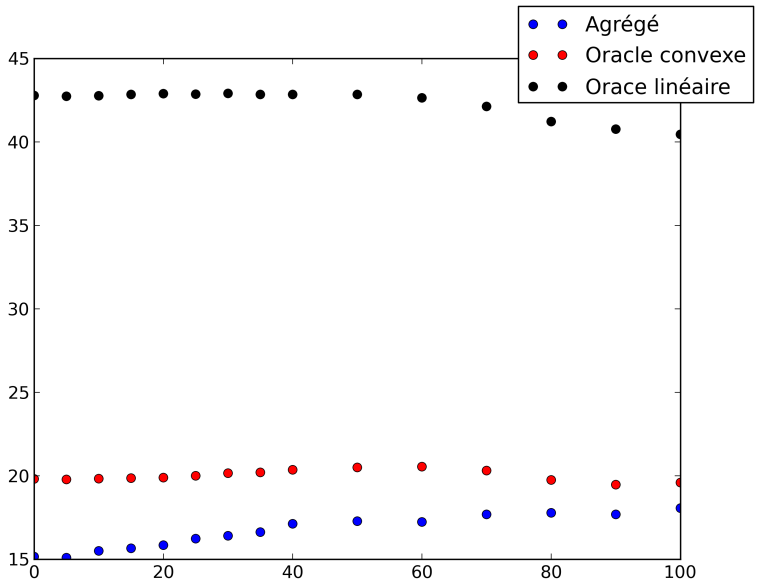

FIGURE 3.20 – Évolution de de la différence relative de *RMSE* entre la prévision déterministe d'une part et, d'autre part, la prévision agrégée (en bleu), l'oracle convexe (en rouge) et l'oracle linéaire (en noir) en fonction de la période d'entraînement sur un échantillon de 1000 cellules.

4 Vent

Nous déclinons la méthode précédente à l'étude d'une seconde grandeur physique : la norme de la vitesse du vent à 10 *mètres au-dessus du sol. Plusieurs remarques d'ordre physique sont faites au passage concernant ce jeu de données.*

Sommaire

4.1 Description du jeu de données

Dans ce chapitre, les techniques d'agrégation détaillées précédemment (au chapitre 3) sont appliquées à la vitesse du vent modélisée et évaluée à dix mètres au-dessus du sol. Si la pression réduite au niveau de la mer était qualifiée de variable synoptique, la vitesse du vent possède un caractère plus local : des variations importantes peuvent exister entre des cellules pourtant voisines.

Le cadre spatial, le cadre temporel, et, *mutatis mutandis*, la nature de l'ensemble sont identiques à ceux du chapitre précédent. Nous soulignons les différences dans la nature des données et des résultats par rapport à la démarche précédente.
La carte de la moyenne temporelle de l'analyse (figure 4.1) fait apparaître deux types de zones de prévision : les zones maritimes et les zones terrestres. Dans les zones maritimes, les valeurs moyennes sont hautes relativement à celles des zones terrestres. Cela est dû aux faibles forces de frottement ou de friction à la surface maritime, comparées aux frottements terrestres engendrés par la présence de reliefs, de végétation, etc. La différence réside aussi dans la présence éventuelle de variations locales à petite échelle : celles-ci existent dans les zones terrestres, mais pas dans les zones maritimes, où, au contraire, les variations sont plus lisses et régulières. Il s'agit là encore d'une conséquence des frottements, et de la rareté probable de mesures au large des côtes qui conduisent à une analyse de qualité moindre.

Les figures 4.2 et 4.3 fournissent des représentations graphiques des variations de l'ensemble, des prévisions déterministes et de l'analyse. Ces vitesses de vent sont calculées dans une cellule au niveau de l'océan Atlantique, de coordonnées $(-13°, 37°)$, et sont relativement élevées comparativement aux valeurs en zone terrestre, avec une moyenne temporelle de $12{,}56\,\mathrm{m\,s^{-1}}$. Par rapport à l'analyse, la moyenne d'ensemble a un biais important de $3{,}43\,\mathrm{m\,s^{-1}}$. À la figure 4.3, une fois la valeur de la moyenne d'ensemble soustraite à chaque prévision protagoniste, l'analyse et la prévision déterministe de Météo France sont globalement en-dessous du zéro, traduisant un biais additif ou multiplicatif de l'ensemble par rapport à l'analyse.

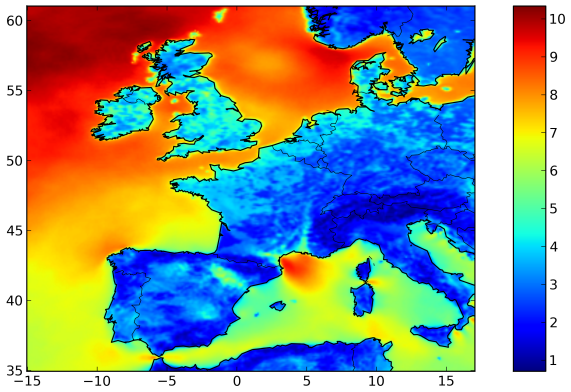

FIGURE 4.1 – Représentation graphique de la moyenne temporelle de l'analyse, dans le cas de la vitesse du vent à 10 mètres au-dessus du sol. Dans chaque cellule de coordonnées (i, j) est représentée la moyenne $\frac{1}{T}\sum_{t=1}^{T} y_{t,(i,j)}$. L'unité est le $\mathrm{m\,s^{-1}}$. La moyenne de ces valeurs est de $5{,}55\,\mathrm{m\,s^{-1}}$, et l'écart type de ces valeurs moyennes est de $2{,}62\,\mathrm{m\,s^{-1}}$.

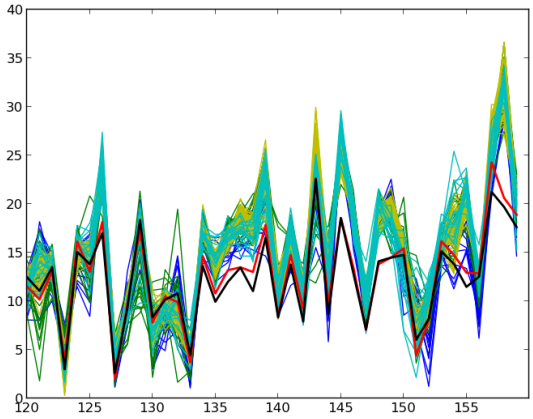

FIGURE 4.2 – Données de la vitesse du vent à 10 mètres au-dessus du sol dans une cellule typique (de coordonnées $(-13°, 37°)$), entre les échéances 320 (16 août 2012) et 360 (25 septembre 2012) de la période temporelle considérée. Un échantillon de 100 membres de l'ensemble parmi les 150 est représenté. En rouge, la prévision déterministe, $x^{\mathrm{det}}_{t,(i,j)}$; en noir, l'analyse, $y_{t,(i,j)}$; le reste des couleurs est dédié aux membres de l'ensemble, $x^{m}_{t,(i,j)}$.

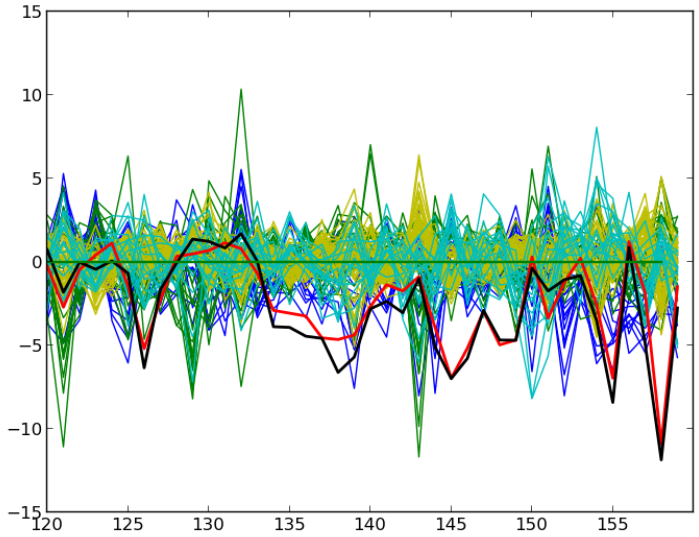

FIGURE 4.3 – Représentation graphique alternative des données de 4.2, ici centrées autour de la moyenne d'ensemble, $\overline{\boldsymbol{x}}_{t,(i,j)}$, représentée par la ligne horizontale en vert foncé et d'ordonnée 0, afin d'assurer une plus grande visibilité de la variabilité. Un échantillon de 100 membres de l'ensemble parmi les 150 est représenté. En rouge, la prévision déterministe centrée, $x^{\text{det}}_{t,(i,j)} - \overline{\boldsymbol{x}}_{t,(i,j)}$; en noir, l'analyse centrée, $y_{t,(i,j)} - \overline{\boldsymbol{x}}_{t,(i,j)}$; le reste des couleurs est dédié aux membres de l'ensemble, $x^{m}_{t,(i,j)} - \overline{\boldsymbol{x}}_{t,(i,j)}$.

4.2 Performance de l'ensemble, oracle et point de référence

Les performances de chacun des membres de l'ensemble, selon le score de la *RMSE*, sont représentées à la figure 4.4. Pour l'oracle convexe, le vecteur de poids constant est représenté à la figure 4.5 et pour l'oracle linéaire, à la figure 4.6. Les poids linéaires de l'oracle somment à une valeur qui est inférieure à un et cela peut être interprété, tout comme dans la partie 4.1, par le fait que l'ensemble a un biais multiplicatif supérieur à 1.

Les scores des oracles convexe et linéaire, reportés dans le tableau 4.1, en font des concurrents très difficiles à vaincre par une stratégie d'agrégation. Par ailleurs, ils ne constituent pas nécessairement les points de référence pertinents pour un météorologue. Ainsi, nous sélectionnons un point de référence qui constitue une prévision dont le score est bon et qui est susceptible d'être employée dans un cadre opérationnel. Au vu du tableau 4.1, rassemblant les scores *RMSE* de candidats raisonnables et des oracles, le rôle du point de référence est rempli, comme dans la partie précédente, par la prévision déterministe proposée par Météo France. Dans la suite de ce chapitre, les évaluations des différences relatives de *RMSE* ont lieu par rapport à la prévision déterministe de Météo France. Si MF correspond à la prévision déterministe de Météo France et A à la prévision issue d'un algorithme donné, La différence relative des gains est définie selon la formule

$$\Delta_{\%}(\mathrm{A}, \mathrm{MF}) = \frac{RMSE_{\mathrm{MF}} - RMSE_{\mathrm{A}}}{RMSE_{\mathrm{MF}}} .$$

Les scores sont reportés dans le tableau 4.1 pour un horizon de prévision de 6 heures.

Type de prévision	*RMSE* ($\mathrm{m\,s^{-1}}$)
Moyenne d'ensemble	3,09
Déterministe ECMWF	2,83
Déterministe Météo France	2,60
Oracle convexe	2,35
Oracle linéaire	1,61

TABLE 4.1 – *RMSE* : Score à un horizon de 6 heures avec une période d'entraînement de 100 jours.

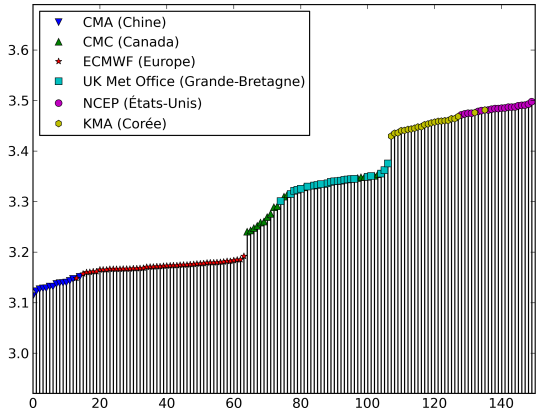

FIGURE 4.4 – Représentation graphique de la performance moyenne de chacun des membres de l'ensemble issus de TIGGE, triés par ordre croissant de *RMSE*, selon la formule $\frac{1}{Nx \times Ny} \sum_{(i,j) \in \text{Carte}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x^m_{(i,j),t} - y_{t,(i,j)})^2}$. Les prévisions sont réalisées à une échéance de 6 heures. Le symbole en haut de chacune des barres verticales indique l'origine (centre météorologique) du membre correspondant. Il s'agit de l'ensemble à 150 membres, ne comprenant pas les prévisions déterministes de Météo France ou d'ECMWF. Remarquons que ce sont les prévisions du centre européen qui réalisent les erreurs les plus basses.

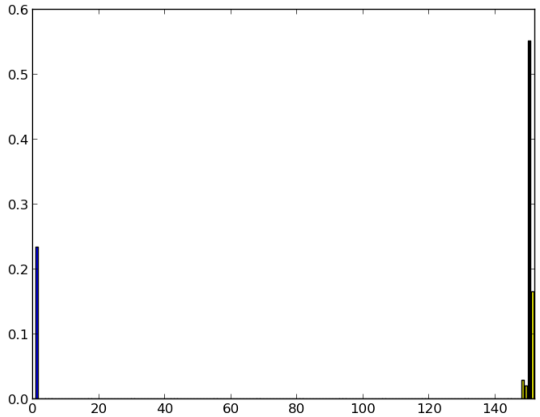

FIGURE 4.5 – Représentation graphique des poids de l'oracle convexe pour l'ensemble complet (152 membres). La meilleure combinaison convexe constante tire parti uniquement de quelques membres de l'ensemble : seuls les poids d'une dizaine de membres diffèrent de zéro à 10^{-6} près. Notons que ces poids sont bien tous compris entre 0 et 1 et leur somme est égale à 1 (à 10^{-11} près). Notons que les deux derniers poids sont affectés aux prévisions déterministes de l'ECMWF et de Météo France, dans cet ordre.

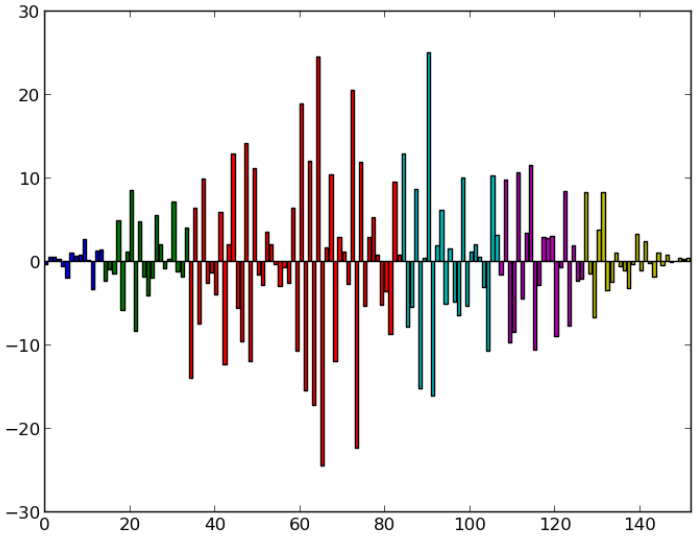

FIGURE 4.6 – Représentation graphique des poids de l'oracle linéaire pour l'ensemble complet (152 membres). La meilleure combinaison linéaire constante tire parti de tous les membres de l'ensemble : ils sont tous visiblement différents de zéro. Notons que ces poids varient dans $\mathbb{R}$. La somme de ces poids est égale à 0,89. Nous pouvons interpréter cet écart à la valeur 1 (qui est bien autorisé par la théorie) comme la conséquence d'un débiaisement de l'ensemble.

4.3 Résultats

4.3.1 Stratégie ridge sur une grille fixe

Nous souhaitons explorer les résultats de l'algorithme d'agrégation ridge escompté, mené indépendemment dans chaque cellule, avec un couple de paramètres optimal a posteriori, sélectionné sur une grille fixe. Le cadre est similaire à celui de la partie **??**, si ce n'est que la grille de paramètres change. Les performances de l'algorithme sur un échantillon de 1000 cellules pour la grille de paramètre (4.1) (régularisation, escompte),

$$\{0; 5; 10; 15; 20; 25; 30; 35; 40; 45; 50\} \times \{0{,}0; 0{,}1; 0{,}2; 0{,}3; 0{,}4; 0{,}5; 0{,}6; 0{,}7; 0{,}8; 0{,}9\}\,, \quad (4.1)$$

sont tout d'abord évaluées. Une fois le meilleur couple de paramètres rétrospectif sélectionné, $(\lambda; \gamma) = (25; 0{,}2)$, les prévisions agrégées sont calculées et leurs performances reportées dans les tableaux 4.2 et 4.3, et représentées sur la carte 4.7, les histogrammes 4.8, 4.9 et 4.10 et le diagramme en boîte 4.11.

Commençons l'analyse de ces résultats par deux remarques immédiates. L'observation de la carte de différence relative 4.7, montre qu'il existe une variabilité locale plus grande des gains que dans le cas de la pression réduite au niveau de la mer à la figure 3.10. De plus, la bimodalité des histogrammes s'explique simplement via la distinction nette entre les deux types de zones, terrestres et marines, déjà relevée dans le paragraphe 4.1.

Ces résultats sont novateurs par rapport à ceux du chapitre précédent à deux égards. D'une part, le tableau 4.2, montre que, contrairement à ce qui se passait dans le cas de la pression réduite au niveau de la mer, en **??**, ici, l'oracle convexe est battu. D'autre part, la carte 4.7, nous permet d'observer que les différences relatives de *RMSE* sont positives, dans une écrasante majorité des cellules. Autrement dit, les prévisions agrégées par la stratégie ridge escomptée améliorent quasiment systématiquement les prévisions déterministes de Météo France. D'ailleurs, le minimum de différence relative de *RMSE* est de $-0{,}80\%$, ce qui est à mettre en regard du maximum, lui de $53{,}86\%$. Dans le cas de la vitesse du vent, l'emploi de la stratégie ridge apparaît être un gain quasiment à coup sûr face à la prévision de référence et à être un compétiteur solide face aux oracles, les références à battre selon la théorie.

L'étude de l'influence des paramètres, non reportée ici, indique une sensibilité infime des performances par rapport au paramètre d'escompte ce qui simplifie l'adaptation locale et temporelle de l'algorithme. Les performances en *RMSE* des prévisions sont donc considérablement améliorées de manière fiable et robuste.

À un horizon de prévision de 6 heures, la *RMSE* est relevée dans le tableau 4.2 et le biais dans le tableau 4.5.

Type de prévision	*RMSE* ($\mathrm{m\,s^{-1}}$)	Gain relatif (%)
Moyenne d'ensemble	3,09	−18,75
Déterministe	2,60	0,00
Oracle convexe	2,35	9,76
Oracle linéaire	1,61	38,10
Agrégation ridge (avec paramètres optimaux rétrospectifs)	2,32	10,93

TABLE 4.2 – *RMSE* : scores et gains relatifs par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres optimisés au préalable : $(\lambda;\,\gamma) = (25; 0{,}20)$. La période d'entraînement est de 100 jours.

Type de prévision	Biais ($\mathrm{m\,s^{-1}}$)	Gain relatif (%)
Moyenne d'ensemble	2,29	−29,22
Déterministe	1,77	0,00
Oracle convexe	1,61	8,91
Oracle linéaire	1,17	33,92
Agrégation ridge (avec paramètres optimaux rétrospectifs)	1,57	11,25

TABLE 4.3 – Biais : scores et gains relatifs par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres optimisés au préalable : $(\lambda;\,\gamma) = (25; 0{,}20)$. La période d'entraînement est de 100 jours.

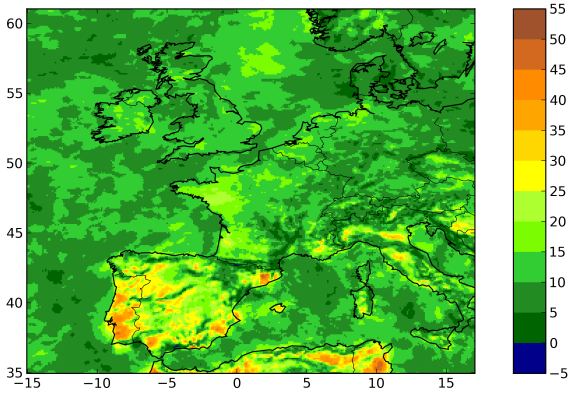

FIGURE 4.7 – Représentation graphique des différences relatives en *RMSE* entre l'agrégé avec paramètres optimaux rétrospectifs et le déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i, j) est représentée la différence relative : $\left(RMSE^{\mathrm{det}}_{(i,j)} - RMSE^{\mathrm{agr}}_{(i,j)}\right)/RMSE^{\mathrm{det}}_{(i,j)}$ où $r^{\mathrm{agr}}_{(i,j)}$ (respectivement $r^{\mathrm{det}}_{(i,j)}$) est la *RMSE* moyenne de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

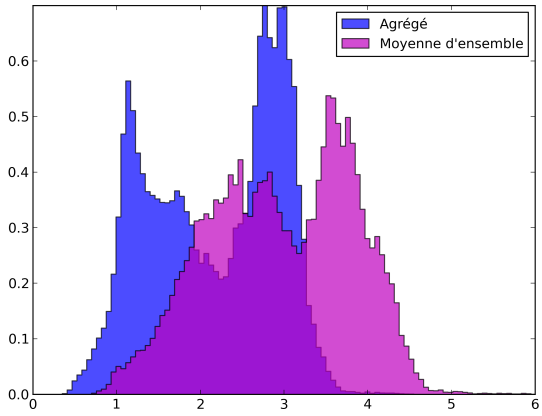

FIGURE 4.8 – Histogramme des *RMSE* de l'agrégé (en bleu) et de la moyenne d'ensemble (en magenta). Les individus représentés sont les *RMSE* pour l'agrégé $r^{\text{agr}}_{(i,j)}$ et les *RMSE* pour la moyenne d'ensemble $r^{\text{ens}}_{(i,j)}$ des cellules (i, j) de la carte dans son intégralité. Les prévisions sont réalisées à une échéance de 6 heures, pour des paramètres optimaux de la régression ridge. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). La *RMSE* est prise par rapport à l'analyse Météo France. Le poids initial est une mesure de Dirac sur le déterministe Météo France.

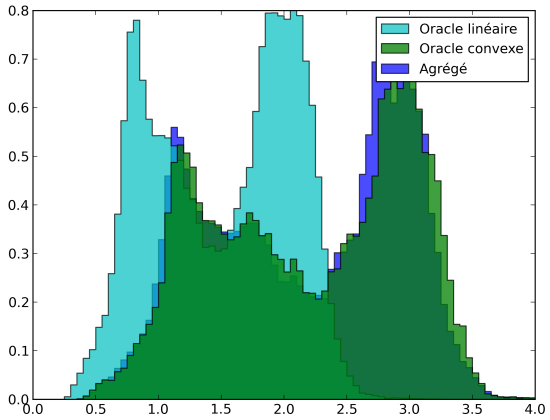

FIGURE 4.9 – Histogramme des *RMSE* de l'oracle linéaire (en cyan), de l'oracle convexe (en vert) et de l'agrégé (en bleu). Les individus représentés sont les *RMSE* pour l'oracle linéaire $r^{\text{lin}}_{(i,j)}$, les *RMSE* pour l'oracle convexe $r^{\text{cvx}}_{(i,j)}$ et les *RMSE* pour l'agrégé $r^{\text{agr}}_{(i,j)}$ des cellules (i, j) de la carte dans son intégralité. Les prévisions sont réalisées à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). La *RMSE* est prise par rapport à l'analyse Météo France. Le poids initial est une mesure de Dirac sur le déterministe Météo France.

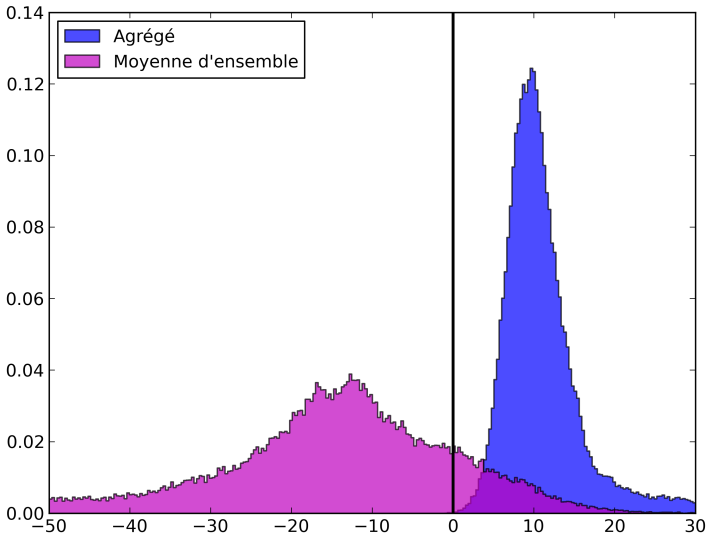

FIGURE 4.10 – Histogramme de la différence relative par rapport au déterministe de la *RMSE* de la moyenne d'ensemble (en magenta) et de l'agrégé (en bleu). Les individus représentés sont les différences relatives en *RMSE* pour l'agrégé $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ et les différences relatives en *RMSE* pour la moyenne d'ensemble $\left(r^{\text{det}}_{(i,j)} - r^{\text{ens}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$ et $r^{\text{ens}}_{(i,j)}$) est la *RMSE* moyen de la prévision agrégée (respectivement déterministe et de la moyenne d'ensemble) dans la cellule (i,j). . Les prévisions sont à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). La *RMSE* est prise par rapport à l'analyse Météo France. Le poids initial est une mesure de Dirac sur le déterministe Météo France.

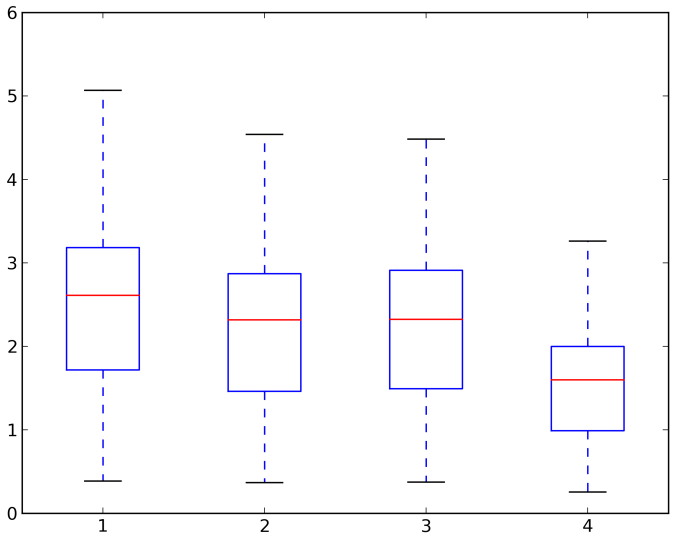

FIGURE 4.11 – Diagramme en boîte des *RMSE* à horizon de 6 heures entre l'analyse Météo France et respectivement, le déterministe (1), l'agrégé issu de la régression ridge (2), l'oracle convexe (3) et l'oracle linéaire (4). Les individus représentés sont les *RMSE* pour l'agrégé $r^{\text{agr}}_{(i,j)}$, les *RMSE* pour la moyenne d'ensemble $r^{\text{ens}}_{(i,j)}$, les *RMSE* pour l'oracle convexe $r^{\text{cvx}}_{(i,j)}$ et les *RMSE* pour l'oracle linéaire $r^{\text{lin}}_{(i,j)}$ des cellules (i, j) de la carte dans son intégralité. À un individu plus bas correspond une *RMSE* plus basse donc de meilleures performances. Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est une mesure de Dirac sur le déterministe Météo France.

4.3.2 Stratégie avec optimisation locale et temporelle des paramètres

Comme dans le chapitre précédent, nous cherchons, en dernier lieu de cette analyse, à nous rapprocher d'une situation plus réaliste, ce qui consiste à employer une stratégie d'agrégation avec optimisation locale et temporelle des paramètres. Celle-ci correspond à l'algorithme 7 du chapitre précédent. Étant donné que l'agrégation est très peu sensible aux variations du paramètre d'escompte γ, la grille (4.2) (régularisation, escompte) employée est construite en conséquence :

$$\{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\} \times \{0.2\}\,. \tag{4.2}$$

Dans le cas du vent, l'estimation des performances des différents algorithmes et prévisions de référence dépend plus fortement du nombre de cellules aléatoires que dans le cas de la pression réduite au niveau de la mer. Grâce à la grille de paramètres dont la taille est relativement réduite, il est possible de conduire une estimation avec un plus grand nombre de cellules aléatoires que dans le chapitre précédent 3. Cela pallie partiellement l'erreur liée à un faible nombre de cellules qui entâche la fiabilité des résultats et des performances affichées. Ainsi, avec un nombre de cellules fixées à 10000, les scores sont rassemblés dans les tableaux 4.4 pour la *RMSE* et 4.5 pour le biais.

Type de prévision	*RMSE* ($\mathrm{m\,s^{-1}}$)	Différence relative (%)
Moyenne d'ensemble	2,32	−46,96
Déterministe	1,58	0,00
Oracle convexe	1,43	9,45
Oracle linéaire	1,38	12,45
Agrégation ridge (paramètres optimaux rétrospectifs)	1,43	8,95
Agrégation ridge (adaptation locale en ligne sur une grille de paramètres)	1,47	6,48

TABLE 4.4 – Stratégie partiellement automatique. *RMSE* : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres pouvant varier tous les $t_{\text{bascule}} = 30$ jours sur la grille (4.2). 10000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

4.4 Conclusion

Par rapport à l'étude de la pression réduite au niveau de la mer, la principale différence réside dans le caractère local (à l'opposé de synoptique) ainsi que bimodal des données de vitesse du vent. Là encore, les résultats sont intéressants : la régression ridge avec paramètres optimaux rétrospectifs améliore les résultats de prévision opérationnelle de 8,95% (partie 4.3.1) tandis que la méthode complètement automatique, avec adaptation locale en ligne sur une grille de paramètres, donne une une différence relative de 6,48%

Type de prévision	Biais ($\mathrm{m\,s^{-1}}$)	Différence relative (%)
Moyenne d'ensemble	1,73	−66,61
Déterministe	1,04	0,00
Oracle convexe	0,97	6,54
Oracle linéaire	0,95	8,45
Agrégation ridge (paramètres optimaux rétrospectifs)	0,96	7,34
Agrégation ridge (adaptation locale en ligne sur une grille de paramètres)	0,97	6,44

TABLE 4.5 – Stratégie partiellement automatique. Biais : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision agrégée à un horizon de 6 heures pour des paramètres pouvant varier tous les $t_{\text{bascule}} = 30$ jours sur la grille (4.2). 10000 cellules sont sélectionnées aléatoirement. Le poids initial est une mesure de Dirac sur le déterministe Météo France. La période d'entraînement est de 100 jours.

(partie 4.3.2). Tout comme dans le cas de la pression réduite au niveau de la mer, ces résultats sont encourageants.

5 Agrégation de fonctions de répartition

Nous présentons dans ce chapitre les enjeux et les outils de la prévision probabiliste avant de mettre en pratique deux algorithmes sur les jeux de données décrits précédemment. La partie 5.1 motive l'utilisation de prévisions probabilistes et expose l'état de l'art dans ce domaine et la partie 5.2 présente des scores probabilistes historiques et populaires. Les algorithmes utilisés sont ensuite décrits dans la partie 5.3 avant que la partie 5.4 ne détaille les résultats empiriques de ceux-ci sur les jeux de données des chapitres 3 et 4.

Sommaire

5.1 Introduction des prévisions probabilistes et incertitudes

5.1.1 Notations

Le tableau suivant liste les principales notations utilisées pour une échéance temporelle s fixée.

$\boldsymbol{x}_s$	Vecteur d'état de la prévision, à valeur dans $\mathbb{R}^M$
y_s	Observation scalaire
$\boldsymbol{p}_s$	Vecteur de poids
$\boldsymbol{p}$	Vecteur constant de poids
$\boldsymbol{\delta}_m$	Vecteur dont la $m^{\text{ème}}$ composante vaut 1, nul ailleurs
M	Taille de l'ensemble : nombre de membres
$\mathcal{P}$	Simplexe de $\mathbb{R}^M$
H_σ	Fonction échelon de Heaviside de seuil σ : $\mathbb{1}_{[\sigma,\infty[}$

5.1.2 Utilité et motivation des prévisions probabilistes

Dans les chapitres précédents, nous avons appliqué la théorie des suites arbitraires à la prévision de valeurs uniques, et avons détaillé les résultats empiriques des algorithmes associés. Nous nous intéressons dans ce chapitre à la prévision d'une fonction de répartition complète, toujours à l'aide des outils issus des suites individuelles. Dans ce cadre, le prévisionniste cherche à fournir, non plus une valeur unique qui se rapprochera le plus possible de l'observation (ou de l'analyse), mais plutôt une fonction de répartition de cette observation à venir.

Pourquoi s'intéresser à ce cadre de prévision probabiliste ? En réalité, une manière naturelle de prévoir une certaine grandeur est de donner une fonction de répartition. En effet, si l'on nous demande quelle sera la vitesse du vent demain à six heures en un lieu précis, nous répondrons plus volontiers « très probablement, c'est-à-dire, avec 90% de chance, entre 6 et 7 ms^{-1} » que « à coup sûr et précisément 6,73 ms^{-1} ». En effet, la seconde proposition semble excessivement assurée. Alternativement, la première pourrait se préciser sous la forme d'une distribution complète de probabilité plutôt que sous celle d'un intervalle de confiance avec son niveau associé. Si cette première réponse semble plus réaliste, cela est aussi dû implicitement aux incertitudes inhérentes à toute prévision. Celles-ci peuvent être liées aux modèles d'évolution, aux erreurs de représentativités (les grandeurs sont moyennées dans chaque cellule et diffèrent des valeurs en chaque point), ou encore aux incertitudes concernant les conditions initiales, éventuellement liées à des phénomènes chaotiques. Prendre en compte ces incertitudes est d'ailleurs l'objectif originel de la génération des ensembles qui constituent la matière première des algorithmes d'agrégation employés dans ces travaux. Ces incertitudes ainsi que la volonté pragmatique d'éviter une assurance irréaliste dans les prévisions, constituent des arguments majeurs pour raisonner en termes probabilistes. Bien sûr, dans de nombreux cas courants, un prévisionniste fournira une valeur unique s'il ne peut pas faire autrement ou s'il souhaite simplifier volontairement son propos, ce qui est fréquent, par exemple dans un bulletin météorologique. Mais les prévisions gagnent souvent à s'enrichir d'un volet probabiliste. Par exemple, les organisateurs d'un événement en plein

air ou des décideurs politiques ont tout intérêt à posséder le maximum d'informations sur l'advenue ou non de pluies et la distribution de leur intensité éventuelle plutôt que la donnée d'une simple valeur de précipitations moyennes. Une information de qualité et détaillée est au fondement d'une aide à la décision efficace. Par ailleurs, une quantification des incertitudes peut être utilisée dans les étapes suivantes d'une prévision : la distribution prévue pour la vitesse du vent peut ainsi être réutilisée en entrée dans les modèles de prévision d'une autre grandeur, par exemple des modèles de dispersion de polluants.

Pour simplifier, cette application de la théorie des suites arbitraires à la prévision probabiliste sera dénommée agrégation de fonctions de répartition. Pour raisonner dans ce nouveau cadre, il est nécessaire de définir une fonction de perte évaluant la précision de prévisions probabilistes par rapport à des observations, ce que nous faisons dans la partie 5.2.4. Dès lors, il devient possible de déterminer la prévision de référence par rapport à laquelle les prévisions agrégées se compareront, en partie 5.4.1, puis d'adapter et d'évaluer certains algorithmes classiques ou plus récents à la prévision probabiliste en parties 5.3.1 et 5.3.3. Mais commençons tout d'abord par un état de l'art de ce domaine.

5.1.3 État de l'art

Dans cette partie, nous inventorions différentes manières de combiner un jeu de prévisions en une prévision probabiliste. Pour réaliser ces combinaisons, plusieurs approches existent selon la nature du score choisi, le type de données à prévoir (si elles admettent une densité par exemple, de loi stationnaire ou ergodique, ou encore s'il s'agit de tirages indépendants et identiquement distribués) et enfin, selon que l'on propose des prévisions probabilistes en partant de lois de probabilité ou de fonctions de répartition. Cependant, chaque fonction de perte considérée est liée à la formulation du problème et conditionne les algorithmes disponibles.

Concernant les scores, Dawid [Daw08] propose un passage en revue chronologique de leur développement dans le domaine des prédictions probabilistes. Le score de Brier [Bri50] a été étudié entre autres par Good [Goo52]. Il a été suivi par le « Ranked Probability Score » [Eps69] et peu après par sa version continue, le « Continuous Ranked Probability Score ». La partie 5.2 offre une introduction progressive et détaillée de ces trois fonctions de score. Il existe d'autres fonctions de score intéressantes, par exemple le score dérivé de la perte logarithmique [GA11 ; GR07].

Certaines propriétés mathématiques souhaitables pour les scores ont été définies et étudiées et permettent de les comparer entre eux. Ainsi, Bröcker et Smith [BS07b] définissent et discutent les caractères (*strictement*) *propre* et *local* d'un score. Le caractère propre, dont Brier [Bri50] possédait déjà une intuition, est une propriété désormais communément considérée comme indispensable, sur laquelle nous revenons à la partie 5.2. L'article de Fricker, Ferro et Stephenson [FFS13] définit la *justesse* d'un score comme la capacité d'un score à favoriser un ensemble de prévisions ayant les

mêmes caractéristiques que la distribution des observations. Ferro [Fer14] montre que le score de Brier, le *RPS* et le *CRPS* ne sont pas *justes* mais que l'on peut proposer des versions ajustées qui le sont. Enfin, sur le plan de l'interprétation des résultats des méthodes, Bröcker et Smith [BS07a] étudient le diagramme de fiabilité, outil populaire de diagnostic des prévisions probabilistes et le rendent plus robuste par une méthode de rééchantillonnage qui permet l'ajout de barres de cohérence et l'évaluation à vue de la fiabilité.

En dehors des diverses fonctions de perte et de leurs caractéristiques, d'autres paradigmes existent qui se traduisent par une variété d'approches conceptuelles et algorithmiques. Par exemple, le « *kernel dressing* » est une méthode qui cherche à fournir une densité de probabilité à partir de prévisions ponctuelles en combinant des fonctions gaussiennes centrées sur ces points. Il s'agit ensuite de calculer les paramètres optimaux de ces fonctions. C'est la méthode qui ressemble le plus à l'application des suites arbitraires que nous allons décrire. Populaires, les méthodes de « Bayesian model averaging » combinent quant à elles des densités de probabilités dans un cadre bayésien [Hoe+99]. Si la combinaison sous forme de moyennes arithmétiques pondérées est la plus répandue, Geweke et Amisano [GA11] proposent une combinaison dérivée de moyennes géométriques pondérées et emploient une fonction de score logarithmique pour évaluer des prédicteurs agrégés. Pour une comparaison de ces différentes méthodes, on pourra se reporter à Clemen et Winkler [CW99].

Les suites arbitraires constituent une approche innovante dans la prévision probabiliste, qui s'émancipe de la distinction habituelle fréquentiste / bayésien. De plus, elles permettent de faire un minimum d'hypothèses sur la fonction objectif à prévoir (hypothèse raisonnable de bornitude, mais pas d'hypothèses sur les lois) et promettent à la fois des garanties de convergence et de vitesse. Il est possible de construire une suite de prévisions calibrées (exactement ou à un ε près), ce qui a un intérêt pour la prévision météorologique (Foster et Vohra [FV98] et Cesa-Bianchi et Lugosi [CL06], partie 4.5). Par ailleurs, ces suites arbitraires trouvent un nouvel écho dans le cadre probabiliste chez Vovk et Zhdanov [VZ09].

5.2 Scores probabilistes

5.2.1 Introduction

Nous présentons le « *continuous ranked probability score* » (*CRPS*), un score populaire en sciences environnementales. Ce score joue le rôle de fonction de perte dans le cadre de l'agrégation de fonctions de répartition. Afin de faciliter l'appréhension du *CRPS*, nous présentons d'abord les scores plus simples et dont le *CRPS* est dérivé. La partie 5.2.2 présente le score de Brier, permettant d'évaluer des prévisions binaires. Le « Ranked Probability Score », (*RPS*), généralise le score de Brier à plus de deux catégories dans la partie 5.2.3. Enfin, la version continue du *RPS*, le *CRPS*, est présentée dans la partie 5.2.4. Cette démarche permet d'introduire intuitivement et progressive-

ment des versions plus abstraites et générales des scores.

Les trois fonctions de score décrites sont convexes par rapport à leurs variables d'entrée. Le cadre d'analyse des suites individuelles est alors bien connu et les techniques d'analyse convexe s'appliquent. Par ailleurs, les trois scores proposés sont des scores strictement propres. Une fonction de score S dont les deux variables d'entrée sont des fonctions de répartition est un score dit *strictement propre* si $S(Q, P) \geqslant S(Q, Q)$ avec égalité si et seulement si $P = Q$. Ainsi, une fonction de score ayant un argument candidat et un argument objectif est dite une fonction de score propre si elle est minimisée (dans le cas général, optimisée) uniquement lorsque l'argument candidat atteint l'argument objectif. Pour une preuve concernant le caractère strictement propre du *CRPS*, on pourra par exemple se référer à Gneiting et Raftery [GR07].

La divergence de Kullback-Leibler, issue de la théorie de l'information, est un autre score connu. Néanmoins, cette divergence prend en argument deux densités de probabilités, avec, pour être définie, une condition d'absolue continuité de l'une par rapport à l'autre. Ce n'est pas le cas du *CRPS* qui permet de comparer des fonctions de répartition plus générales et est ainsi plus approprié au cadre de prévisions de fonction de répartition qui nous intéresse dans ce chapitre.

5.2.2 Score de Brier

Supposons que nous souhaitions prévoir, dans une cellule donnée, un événement dont le résultat est de nature binaire, $y \in \mathcal{Y} = \{0, 1\}$. Par exemple, $(y = 0)$ peut encoder l'événement « il ne pleut pas » et $(y = 1)$ l'événement « il pleut ». Le score de Brier est une mesure de la précision entre les prévisions données sous la forme de densités discrètes de probabilité et les occurrences successives de ces événements binaires. Si, à chaque échéance $s \in \{1, \dots, t\}$, le prévisionniste fournit la probabilité $p_s \in [0, 1]$ que l'événement se réalise $(y_s = 1)$, alors le score de Brier sur la période $\{1, \dots, t\}$ s'écrit :

$$BS_t = \frac{1}{t} \sum_{s=1}^{t} (p_s - y_s)^2 \,.$$

Autrement dit, le score de Brier est l'erreur en moyenne quadratique de la prévision probabiliste par rapport à l'observation. Compris entre 0 et 1, il est égal à 0 dans le cas d'une prévision parfaite et à 1 dans le pire des cas (prévision déterministe systématiquement fausse). Il s'agit, de plus, d'une fonction convexe des prévisions p_s.

Nous découpons l'espace des prévisions $[0, 1]$ en une partition uniforme de K intervalles

$$[0, 1] = \bigcup_{k=1}^{K} I_k \,,$$

et notons, $\forall k \in \{1, \dots, K\}$,

- $S_k = \{s \in \{1, \dots, t\} : p_s \in I_k\}$, l'ensemble des indices temporels auxquels les prévisions appartiennent à I_k ;

- $n_k = |\{s \in \{1, \dots, t\} : p_s \in I_k\}|$ le nombre de prévisions appartenant à I_k ;
- $\overline{p}_k = \frac{1}{n_k} \sum_{s \in S_k} p_s$, la moyenne des prévisions de I_k ;
- $\overline{y}_k = \frac{1}{n_k} \sum_{s \in S_k} y_s$, la moyenne des observations sachant que $p_s \in I_k$;
- et $\overline{y} = \frac{1}{t} \sum_{s=1}^{t} y_s$ la moyenne globale des observations.

Le score de Brier peut alors être décomposé en une somme de trois termes (voir à ce sujet Murphy [Mur73]) :

$$BS_t = \underbrace{\frac{1}{t} \sum_{k=1}^{K} n_k (\overline{p}_k - \overline{y}_k)^2}_{\text{calibration}} - \underbrace{\frac{1}{t} \sum_{k=1}^{K} n_k (\overline{y}_k - \overline{y})^2}_{\text{résolution}} + \underbrace{\overline{y}\,(1 - \overline{y})}_{\text{incertitude}} . \tag{5.1}$$

Dans cette décomposition, $\frac{1}{t} \sum_{k=1}^{K} n_k (\overline{p}_k - \overline{y}_k)^2$ est le terme de calibration (aussi appelé terme de fiabilité), $\frac{1}{t} \sum_{k=1}^{K} n_k (\overline{y}_k - \overline{y})^2$ est le terme de résolution et $\overline{y}\,(1 - \overline{y})$, le terme d'incertitude. Tous ces termes sont positifs, compris entre 0 et 1 et le terme d'incertitude est indépendant de la prévision. Le score de Brier est une fonction décroissante de la performance d'une prévision : le prévisionniste cherche donc à minimiser le terme de calibration et à maximiser le terme de résolution.

Présentons une prévision probabiliste naïve. La *prévision climatologique* propose une probabilité constante de réalisation de $(y = 1)$, égale à la probabilité empirique moyenne de cet événement (que l'on suppose connue sur la période considérée) : $\forall s \in \{1, \dots, t\}$, $p_s = \overline{y}$. Autrement dit, s'il pleut 20% du temps, alors, la prévision climatologique affirme chaque jour qu'il y a 20% de chances qu'il pleuve. Dans ce cas, le terme de résolution vaut exactement 0 et le score de Brier se ramène à la somme des termes de calibration et d'incertitude. Puisque réaliser cette prévision climatologique est relativement simple sous des hypothèses peu contraignantes, l'objectif du prévisionniste est de construire une prévision plus originale que cette dernière dont le score de Brier soit inférieur à celui de la prévision climatologique. La prévision plus fine qu'il propose est nécessairement davantage variable en temps.

Le terme de calibration $\frac{1}{t} \sum_{k=1}^{K} n_k (\overline{p}_k - \overline{y}_k)^2$ est minimal lorsqu'il y a adéquation parfaite entre les moyennes des prévisions $\{p_s, s \in S_k\}$ et les moyennes des observations conditionnelles $\{y_s, s \in S_k\}$ correspondantes.
Le terme de résolution $\frac{1}{t} \sum_{k=1}^{K} n_k (\overline{y}_k - \overline{y})^2$ traduit l'écart des observations conditionnelles de $\{y_s, s \in S_k\}$ par rapport à la moyenne des observations $\overline{y}$. Chaque observation conditionnelle aux prévisions distinctes de la moyenne des observations enrichit la prévision en l'éloignant de la prévision climatologique.
Le terme d'incertitude $\overline{y}\,(1 - \overline{y})$ est minimal lorsque $\overline{y} = 0$ ou $\overline{y} = 1$ et est maximal lorsque $\overline{y} = 1/2$. Il s'agit de la variance empirique des observations. Plus l'on s'éloigne d'observations constantes, systématiquement égales à 0 ou 1, plus ce terme d'incertitude augmente et contribue à augmenter le score de Brier : les observations sont de nature imprévisible. Cette contribution traduit une difficulté intrinsèque de la tâche,

indépendante des prévisions $(p_s)_{s\in\{1,\dots,t\}}$.

Pour conclure, le score de Brier quantifie deux aspects d'une prévision et rassemble ces informations en un unique score synthétique. En effet, la décomposition 5.1 montre qu'une prévision de qualité au sens du score de Brier est simultanément calibrée (précise dans chaque intervalle de discrétisation I_k) et raffinée (assez éloignée de la prévision climatologique par sa variabilité temporelle). Ces exigences de variabilité globale et de précision locale sont intuitivement souhaitables. Doublées du caractère convexe et de score strictement propre, elles font de cette fonction un choix pertinent pour l'évaluation de prévisions probabilistes.

5.2.3 *Ranked probability score*

Plutôt qu'un événement binaire, supposons désormais que le prévisionniste cherche à prévoir un événement pouvant prendre un nombre K fini d'états. Par exemple, dans l'exemple précédent de la prévision de pluie, avec $K = 3$ pour fixer les idées, les observations possibles sont $y \in \mathcal{Y} = \{1, 2, 3\}$, avec, dans ce cas, $(y = 1)$ qui encode l'événement « il ne pleut pas », $(y = 2)$ l'événement « il pleut peu (entre 0 et 6 mm h^{-1}) » et $(y = 3)$ l'événement « il pleut beaucoup (plus de 6 mm h^{-1}) ». À chaque échéance $s \in \{1, \dots, t\}$, le prévisionniste fournit une distribution de probabilité $(p_{1,s}, \dots, p_{K,s})$ ce qui permet d'en déduire les valeurs de la fonction de répartition associée $(r^p_{1,s}, \dots, r^p_{K,s})$, où l'exposant p signifie qu'il s'agit d'une prévision. Simultanément, l'observation est révélée et permet de construire une fonction de répartition empirique (qui correspond à une fonction échelon discrétisée dont le saut a lieu à la valeur y_s observée) : $(r^o_{1,t}, \dots, r^o_{K,t})$, où l'exposant o signifie qu'il s'agit d'une observation.
Le « Ranked Probability Score » mesure la différence quadratique entre les fonctions de répartition discrètes prévue et observée selon la formule

$$RPS_t = \frac{1}{t}\sum_{s=1}^{t}\frac{1}{K-1}\sum_{k=1}^{K}(r^p_{k,s} - r^o_{k,s})^2\,. \tag{5.2}$$

Il s'agit bien d'une généralisation à plusieurs classes du score de Brier. On vérifie que le RPS pour $K = 2$ correspond au score de Brier : $K-1 = 2-1 = 1$ et $\forall s$, $(r^p_{2,s}-r^o_{2,s})^2 = (1-1)^2 = 0$, i.e. le second terme de différence entre les densités de probabilité est systématiquement égal à 0 et n'apparaît donc jamais ce qui autorise l'écriture du score de Brier à partir de densités de probabilités. Cette formulation permet la décomposition en somme de termes dont l'interprétation est éclairante.

5.2.4 *Continuous ranked probability score*

Généraliser le score de Brier de 2 à K observations conduit à la construction du RPS. Puis, généraliser ce RPS de K observations à un continuum conduit à la construction du $CRPS$. Le score du $CRPS$ permet de comparer deux fonctions de répartition définies sur un intervalle. Formellement, passer du discret au continu revient à remplacer la somme sur les catégories k d'observation par une intégrale sur l'intervalle des valeurs

considérées.

Ainsi, l'événement à prévoir peut désormais prendre ses valeurs dans un intervalle, par exemple, $\mathcal{Y} = \mathbb{R}_+$ pour la vitesse de vent. À chaque échéance $s \in \{1, \dots, t\}$, le prévisionniste fournit une fonction de répartition empirique complète F_s^p. Simultanément, l'observation est révélée et permet de construire la fonction de répartition empirique associée $F_s^o(z)$: une fonction de Heaviside dont le saut a lieu à la valeur y_s observée.

Le « *continuous ranked probability score* » mesure la différence quadratique entre les fonctions de répartition prévue et observée selon la formule

$$CRPS_t = \frac{1}{t} \sum_{s=1}^{t} \int_{-\infty}^{\infty} \left(F_s^p(z) - F_s^o(z)\right)^2 dz \,. \tag{5.3}$$

Cette fonction admet, elle aussi, une décomposition et une interprétation similaire à celle du score de Brier, dérivée et décrite dans Hersbach [Her00].

Dans toute la suite, nous considérons aussi la valeur instantanée à chaque échéance de cette fonction, c'est-à-dire que nous définissons la fonction $CRPS_s^i$ par

$$CRPS_s^i = \int_{-\infty}^{\infty} \left(F_s^p(z) - F_s^o(z)\right)^2 dz \,, \tag{5.4}$$

avec bien sûr,

$$CRPS_t = \frac{1}{t} \sum_{s=1}^{t} CRPS_s^i \,. \tag{5.5}$$

Notons que dans le cadre des suites arbitraires, les fonctions de perte sont instantanées et on considère les pertes cumulées, non pas les pertes moyennes normalisées par $1/t$: $\sum_{s=1}^{t} CRPS_s^i$. Remarquons que ce score du *CRPS* instantané est proche du critère de Cramér-von Mises en statistique, employé dans les tests d'adéquation entre deux fonctions de répartition, avec la différence notable que la mesure d'intégration de ce critère est $dF_s^o(z)$ et non pas la mesure de Lebesgue dz.

La fonction de répartition associée à la prévision probabiliste est une combinaison linéaire de fonctions de Heaviside $H_{x_{m,s}}$, dont les seuils sont les membres de l'ensemble (triés) $x_{m,s}$ et dont les poids convexes $p_{m,s}$ sont calculés à partir d'un des algorithmes de la partie 5.3. Dans ce cadre d'agrégation de fonctions de répartition, les fonctions de Heaviside $H_{x_{m,s}}$ jouent donc désormais le rôle d'experts, en lieu et place de $x_{m,s}$. Pour une échéance s fixée et dans une cellule donnée, les fonctions de répartition associées à l'observation et à la prévision probabiliste s'expriment donc respectivement

$$F_s^o(z) = H_{y_s}(z) \tag{5.6}$$

et

$$F_s^p(z) = \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z) \,. \tag{5.7}$$

Le *CRPS* instantané s'exprime alors

$$CRPS_s^i(\boldsymbol{p}_s) = \int_{-\infty}^{\infty} \left(H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z) \right)^2 dz \,, \tag{5.8}$$

où nous soulignons la dépendance de ce score au vecteur de poids $\boldsymbol{p}_s$.
Puisque les poids de prévision sont convexes, la fonction de répartition des prévisions $\sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$ atteint nécessairement la valeur 1 au-delà du dernier membre de l'ensemble $x_{M,s}$. Si l'on suppose que les analyses et les prévisions de l'ensemble appartiennent toutes à un segment, i.e. il existe un minorant γ et un majorant Γ tels que,

$$y_s \in [\gamma, \Gamma] \qquad \forall s \in \{1, \ldots, T\}$$

et

$$x_{m,s} \in [\gamma, \Gamma] \qquad \forall (m,s) \in \{1, \ldots, M\} \times \{1, \ldots, T\} \,,$$

alors la différence $H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$ est nulle sur la réunion d'intervalles $]-\infty, \gamma[\cup]\Gamma, \infty[$. Ainsi l'intégrale sur $]-\infty, \infty[$ se résume à une intégrale sur le segment $[\gamma, \Gamma]$. Le *CRPS* instantané peut donc se réécrire

$$CRPS_s^i(\boldsymbol{p}_s) = \int_{\gamma}^{\Gamma} \left(H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z) \right)^2 dz \,. \tag{5.9}$$

Explicitons la valeur de ce *CRPS* et montrons qu'elle est indépendante des bornes γ et Γ. En employant les relations suivantes :

$$H_x H_y = H_{\max(x,\, y)} \qquad \forall x, y \in [\gamma, \Gamma]^2 \,,$$

et

$$\int_{\gamma}^{\Gamma} H_x(z) dz = \Gamma - x \,,$$

on développe et on réduit le carré, avant d'intégrer. Le *CRPS* instantanné, $CRPS_s^i(\boldsymbol{p}_s)$ se réécrit alors :

$$\begin{aligned}
&\int_{\gamma}^{\Gamma} \left(H_{y_s}(z) - 2 \sum_{m=1}^{M} p_{m,s} H_{y_s}(z) H_{x_{m,s}}(z) + \sum_{k,m=1}^{M} p_{k,s} p_{m,s} H_{x_{k,s}}(z) H_{x_{m,s}}(z) \right) dz \\
&\quad = (\Gamma - y_s) - 2 \sum_{m=1}^{M} p_{m,s} (\Gamma - \max(x_{m,s}, y_s)) + \sum_{k,m=1}^{M} p_{k,s} p_{m,s} (\Gamma - \max(x_{k,s}, x_{mk,s})) \\
&\quad = -y_s + 2 \sum_{m=1}^{M} p_{m,s} \max(x_{m,s}, y_s) - \sum_{k,m=1}^{M} p_{k,s} p_{m,s} \max(x_{k,s}, x_{m,s}) \,.
\end{aligned} \tag{5.10}$$

La dernière formule indique que le *CRPS* instantané ne dépend pas des bornes γ et Γ : peu importent les majorants choisis, tant qu'ils conviennent. De plus, cette formule est celle qui est implémentée en pratique pour les calculs de *CRPS*.

5.2.5 Présentation didactique du *CRPS*

La figure 5.1 propose une représentation de deux prévisions probabilistes distinctes et de leur évaluation dans un cas simple. Supposons que l'on se place dans une cellule donnée à un pas de temps fixé, que les prévisions et les observations sont à valeur dans $[0, 1]$ et que l'ensemble fournit les prévisions $\{0, 0.1, \ldots, 0.9, 1\}$. Dans le premier cas (en violet, à gauche), les poids affectés aux membres de l'ensemble sont uniformes, ce qui est visible par l'aspect régulier de la courbe dont toutes les marches sont de même hauteur. Dans le second cas (en rose, à droite), ils sont distincts les uns des autres, tout en restant convexes. Dans chaque graphique, la partie colorée (violette ou rose) correspond à la différence entre la fonction de répartition de l'observation $H_{y_s}(z)$ et de la fonction de répartition de la prévision d'ensemble probabiliste $\sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$: $H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$. Dans chaque cas, le carré de cette différence est représenté par la fonction en jaune, et le *CRPS* est alors l'aire du graphique de cette fonction. Au vu de ces exemples, le *CRPS* dans le cas non-uniforme est plus faible car un accroissement relatif des poids autour de la valeur à prévoir y_s contribue à diminuer cette aire en jaune par rapport à une prévision uniforme.

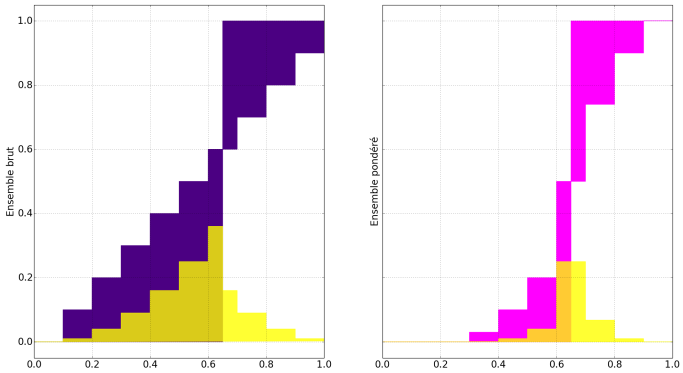

FIGURE 5.1 – Représentation schématique de deux prévisions probabilistes et visualisation de leur *CRPS* respectif. L'aire colorée (violette ou rose) correspond à la fonction $H_{y_s}(z) - \sum_{m=1}^{M} p_{m,s} H_{x_{m,s}}(z)$, avec, à droite, des poids uniformes convexes, et à gauche, des poids non uniformes. La surface superposée en jaune dans chaque figure correspond au carré de la valeur précédente. Son aire est donc la valeur du $CRPS_s^i$.

Notons que la modélisation des experts sous forme de fonctions échelons est un choix arbitraire. Nous pourrions proposer des fonctions plus régulières, afin de traduire davantage l'incertitude que nous avons sur les prévisions. Par exemple, il pourrait sembler plus pertinent de choisir des fonctions logistiques tronquées, modélisant approximativement une éventuelle loi de probabilité gaussienne sous-jacente, ou bien une fonction rampe. Néanmoins, si ces choix satisfont une envie de réalisme, ils n'apportent aucun accroissement de performances. En effet, un simple schéma permet de se convaincre que l'aire délimitée par la différence de deux fonctions échelons est nécessairement inférieure à l'aire délimitée par une fonction échelon et une fonction logistique. Dès lors, conserver les fonctions échelons simplifie les calculs et les implémentations informatiques et n'en-

traîne aucune dégradation de la précision des algorithmes d'agrégation. En revanche, traduire mathématiquement l'incertitude sur les observations apporte un surcroît de réalisme et de précision dans le score du *CRPS*. Cependant, c'est une tâche ardue que nous n'explorons pas plus avant.

5.3 Algorithmes employés dans le cadre de l'agrégation de fonctions de répartition

Le *CRPS*, l'unique score envisagé dans toute les implémentations et applications pratiques de ce chapitre, est convexe par rapport au vecteur de poids $\boldsymbol{p}$. Cela permet d'envisager l'utilisation de tous les outils conceptuels des suites arbitraires avec pertes convexes, en particulier l'algorithme des poids exponentiels (partie 5.3.1). Nous présentons succintement en partie 5.3.3 l' algorithme récent ML-poly dont la borne est du second ordre. Comme indiqué dans l'introduction (chapitre 1), l'algorithme des poids exponentiels est capable de fournir des garanties de regret uniquement par rapport au meilleur membre de l'ensemble. C'est aussi le cas de l'algorithme ML-poly. Il est néanmoins possible de se ramener à des garanties contre la meilleure combinaison convexe constante de membres de l'ensemble avec l'aide de « l'astuce du gradient », qui consiste à remplacer les pertes par les sous-gradients des pertes. Les deux algorithmes peuvent tous deux utiliser cette substitution. Nous l'explicitons et l'appliquons dans la partie 5.3.2, après le rappel de la stratégie exponentielle dans la partie 5.3.1.

5.3.1 Pondération par poids exponentiels

La démonstration des garanties théoriques de l'algorithme des poids exponentiels est donnée dans le chapitre introductif (chapitre 1). Rappelons ici la stratégie $\mathcal{E}_\eta$ de pondération par poids exponentiels des pertes cumulées mise en œuvre dans l'algorithme 8 dans le cas où $y_s \in [\gamma, \Gamma]$,$\forall s \in \{1, \dots, T\}$ et $x_{m,s} \in [\gamma, \Gamma]$,$\forall (m,s) \in \{1, \dots, M\} \times \{1, \dots, T\}$.

Paramètre : on choisit la *vitesse d'apprentissage* $\eta > 0$.

Initialisation : $\boldsymbol{p}_1$ est le vecteur de poids uniforme, $p_{m,1} = 1/M$, pour $m = 1, \dots, M$.

À chaque échéance $t = 1, 2, \dots$, le vecteur des poids $\boldsymbol{p}_{t+1}$ est défini pour $m = 1, \dots, M$ par

$$p_{m,t+1} = \frac{\mathrm{e}^{-\eta L_t^m}}{\sum_{m=1}^M \mathrm{e}^{-\eta L_t^m}} = \frac{\mathrm{e}^{-\eta \sum_{s=1}^t \int_\gamma^\Gamma \left(H_{y_s}(z) - H_{x_{m,s}}(z)\right)^2 dz}}{\sum_{m=1}^M \mathrm{e}^{-\eta \sum_{s=1}^t \int_\gamma^\Gamma \left(H_{y_s}(z) - H_{x_{m,s}}(z)\right)^2 dz}} .$$

Algorithm 8: Pondération par poids exponentiels.

Les garanties théoriques découlent alors du chapitre 1 et sont fournies par le théorème 5.

Théorème 5. *Le $CRPS_s^i(\boldsymbol{p}_s)$ est à valeur dans $[0, \Gamma - \gamma]$ et est convexe par rapport à $\boldsymbol{p}_s$. Alors pour tout $\eta > 0$,*

$$\begin{aligned}\sup\left\{ \sum_{t=1}^{T} CRPS_t^i(\boldsymbol{p}_t) - \inf_{m=1,\dots,M} \sum_{t=1}^{T} CRPS_t^i(\boldsymbol{\delta}_m) \right\} \\ = \sup\left\{ \sum_{t=1}^{T} \int_\gamma^\Gamma \left(H_{y_t}(z) - \sum_{m=1}^{M} p_{m,t} H_{x_{m,t}}(z) \right)^2 dz \right. \\ \left. - \min_{m=1,\dots,M} \sum_{t=1}^{T} \int_\gamma^\Gamma \left(H_{y_t}(z) - H_{x_{m,t}}(z) \right)^2 dz \right\} \\ \leqslant \frac{\ln M}{\eta} + \eta \frac{(\Gamma - \gamma)^2}{8} T \,.\end{aligned} \tag{5.11}$$

où le supremum porte sur toutes les suites possibles d'observations y_s et de prévisions des experts $x_{m,s}$. En particulier, le choix du meilleur paramètre d'apprentissage a posteriori, $\eta^\star = (1/(\Gamma - \gamma))\sqrt{(8 \ln M)/T}$, conduit à la majoration

$$\sup\left\{ \sum_{t=1}^{T} CRPS_t^i(\boldsymbol{p}_t) - \inf_{m=1,\dots,M} CRPS_s^i(\boldsymbol{\delta}_m) \right\} \leqslant (\Gamma - \gamma)\sqrt{\frac{T}{2} \ln M} \,.$$

5.3.2 Linéarisation de la perte par passage aux sous-gradients

La stratégie des poids exponentiels ne s'approche que de la performance du meilleur expert et non pas nécessairement de celle de la meilleure combinaison convexe. Pour résoudre ce problème, l'astuce du gradient (« *gradient trick* ») consiste à remplacer dans la formule de mise à jour des poids la perte par une formule faisant intervenir son gradient. Dans cette partie, nous modifions la stratégie précédente $\mathcal{E}_\eta$ de pondération exponentielle pour obtenir la stratégie $\mathcal{E}_\eta^{\text{grad}}$ de pondération exponentielle des sous-gradients des pertes cumulées.

Afin de simplifier le raisonnement et les notations, plaçons-nous à une échéance fixée et définissons les espaces fonctionnels $\mathcal{X}$ et $\mathcal{Y}$ comme :

$$\mathcal{X} = \left\{ \sum_{m=1}^{M} p_m H_{x_m},\, \boldsymbol{p} \in \mathcal{P}, \boldsymbol{x} \in [\gamma, \Gamma]^M \right\},$$

et

$$\mathcal{Y} = \left\{ H_y, y \in [\gamma, \Gamma] \right\}.$$

Pour toute fonction $H_y \in \mathcal{Y}$, la fonctionnelle à valeurs dans $\mathbb{R}$, qui à tout $\boldsymbol{p}$ du simplexe

$\mathcal{P}$ associe

$$\begin{aligned} CRPS^i(\boldsymbol{p}) &= \int_\gamma^\Gamma \left(H_y(z) - \sum_{m=1}^M p_m H_{x_m}(z) \right)^2 dz \\ &= -y_s + 2\sum_{m=1}^M p_{m,s} \max(x_{m,s}, y_s) - \sum_{k,m=1}^M p_{k,s} p_{m,s} \max(x_{k,s}, x_{m,s}) \,, \end{aligned} \tag{5.12}$$

est convexe et différentiable par rapport à $\boldsymbol{p}$.

Nous explicitons ici le calcul de la différentielle de $CRPS^i(\boldsymbol{p})$ et renvoyons à la partie 1 et à l'article de Stoltz [Sto10] pour l'utilisation de l'inégalité des pentes et la suite de calculs conduisant aux garanties dans le cas convexe. Soit $\sum_{m=1}^M p_m H_{x_m} \in \mathcal{X}$ et $H_y \in \mathcal{Y}$, l'expression de la différentielle $\partial_{\boldsymbol{p}} CRPS^i$, se déduit directement de l'équation (5.10). La $m^{\text{ème}}$ composante de la différentielle du $CRPS$ instantané $\tilde{\ell}_m$ s'écrit en effet

$$\tilde{\ell}_m = \left(\partial_{\boldsymbol{p}} CRPS^i \right)_m = 2 \left\{ \max(x_m, y) - \sum_{k=1}^M p_k \max(x_m, x_k) \right\} . \tag{5.13}$$

La stratégie $\mathcal{E}_\eta^{\text{grad}}$ de l'algorithme 9 est alors obtenue en remplaçant dans la règle de mise à jour des poids de la stratégie classique $\mathcal{E}_\eta$, les termes

$$\int_\gamma^\Gamma (H_y(z) - H_{x_m}(z))^2 \, dz$$

par des termes

$$\tilde{\ell}_m = 2 \left\{ \max(x_m,\, y) - \sum_{k=1}^M p_k \max(x_m,\, x_k) \right\} .$$

La majoration du regret par rapport à la meilleure combinaison convexe de l'algorithme 9 est alors donnée dans le théorème 6.

Théorème 6. *Les pseudo-pertes sont bornées car pour $m = 1, \ldots, M$,*

$$\left\{ \max(x_{m,t},\, y_t) - \sum_{k=1}^M p_{k,t} \max(x_{m,t},\, x_{k,t}) \right\} \in [\gamma - \Gamma,\ \Gamma - \gamma] \,.$$

Alors pour tout $\eta > 0$,

$$\begin{aligned} \sup\Bigg\{ \sum_{t=1}^T CRPS_t^i(\boldsymbol{p}_t) - \inf_{\boldsymbol{q}\in\mathcal{P}} \sum_{t=1}^T CRPS_t^i(\boldsymbol{q}) \Bigg\} & \\ = \ \sup\Bigg\{ \sum_{t=1}^T \int_\gamma^\Gamma \left(H_{y_t}(z) - \sum_{m=1}^M p_{m,t} H_{x_{m,t}}(z) \right)^2 dz & \\ - \inf_{\boldsymbol{q}\in\mathcal{P}} \sum_{t=1}^T \int_\gamma^\Gamma \left(H_{y_t}(z) - \sum_{m=1}^M q_m H_{x_{m,t}}(z) \right)^2 dz \Bigg\} & \\ \leqslant \frac{\ln M}{\eta} + \eta \frac{(\Gamma-\gamma)^2}{2} T & \end{aligned} \tag{5.14}$$

Paramètre : on choisit la *vitesse d'apprentissage* $\eta > 0$.

Initialisation : $\boldsymbol{p}_1$ est le vecteur de poids uniforme, $p_{m,1} = 1/M$, pour $m = 1, \dots, M$.

À chaque échéance $t = 1, 2, \dots$, le vecteur des poids $\boldsymbol{p}_{t+1}$ est défini pour $m = 1, \dots, M$ par

$$p_{m,t+1} = \frac{\mathrm{e}^{-2\eta \sum_{s=1}^{t} \left\{ \max(x_{m,s}, y_s) - \sum_{k=1}^{M} p_{k,s} \max(x_{m,s}, x_{k,s}) \right\}}}{\sum_{m=1}^{M} \mathrm{e}^{-2\eta \sum_{s=1}^{t} \left\{ \max(x_{m,s}, y_s) - \sum_{k=1}^{M} p_{k,s} \max(x_{m,s}, x_{k,s}) \right\}}} .$$

Algorithm 9: Pondération par poids exponentiels des sous-gradients des pertes.

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

$$\sup \left\{ \sum_{t=1}^{T} CRPS_t^i(\boldsymbol{p}_t) - \inf_{\boldsymbol{q} \in \mathcal{P}} \sum_{t=1}^{T} CRPS_t^i(\boldsymbol{q}) \right\} \leqslant (\Gamma - \gamma) \sqrt{\frac{T \ln M}{2}} .$$

5.3.3 Algorithme ML-poly

L'algorithme ML-poly de Gaillard, Stoltz et van Erven [GSv14] induit une borne théorique du second ordre. La version que nous présentons ici est complètement adaptative par rapport au vecteur de paramètres d'apprentissage $\boldsymbol{\eta}$. Tout comme dans le cas des poids exponentiels, il existe deux versions : une version avec calibration séquentielle théorique du paramètre d'apprentissage et une version employant une méthode de grille. Dans le cas de ML-poly, soulignons que la calibration séquentielle théorique conduit à un algorithme ayant de bons résultats pratiques (cf. Gaillard [Gai15]), ce qui n'est pas habituel dans le cas des poids exponentiels. Cela nous incite à retenir et employer cette version qui nous affranchit de la calibration des paramètres. Sur le plan théorique, une originalité de ML-poly est qu'il dérive d'un potentiel polynomial, comme cela est décrit dans Cesa-Bianchi et Lugosi [CL06]. L'algorithme ML-poly est le suivant :

Initialisation : $\boldsymbol{R}_0 = (0, \ldots, 0)$ est le vecteur de regret initial.
À chaque échéance $t = 1, 2, \ldots$

1. on calcule les composantes du vecteur de paramètres d'apprentissage $\boldsymbol{\eta}_{t-1}$ en suivant la règle de décision :

$$\eta_{m,t-1} = \frac{1}{1 + \sum_{s=1}^{t-1} (\widehat{\ell}_s - \ell_{m,s})^2}$$

2. on crée le vecteur de poids $\boldsymbol{p}_t$ définit composante par composante par

$$p_{m,t} = \eta_{m,t-1}(R_{m,t-1})_+ / \boldsymbol{\eta}_{t-1}(\boldsymbol{R}_{t-1})_+$$

   où $\boldsymbol{x}_+$ représente le vecteur constitué des parties positives de chaque composante de $\boldsymbol{x}$;

3. on observe le vecteur de perte $\boldsymbol{\ell}_t$, de composantes

$$\ell_{m,t} = \int_\gamma^\Gamma \left(H_{y_t}(z) - H_{x_{m,t}}(z)\right)^2 dz$$

   et on calcule $\widehat{\ell}_t = \sum_{m=1}^M p_{m,t}\, \ell_{m,t}$;

4. on met à jour le regret pour chaque expert $R_{m,t} = R_{m,t-1} + \widehat{\ell}_t - \ell_{m,t}$.

Algorithm 10: Moyenne pondérée via un potentiel polynomial avec plusieurs paramètres d'apprentissage (ML-Poly).

Enfin, les garanties théoriques pour cet algorithme sont données par le théorème 7.

Théorème 7. *Pour toute séquence de vecteur de perte $\boldsymbol{\ell}_t \in [0, \Gamma - \gamma]^M$, la perte cumulée de l'algorithme 10 est majorée selon la borne*

$$\sum_{t=1}^{T} \widehat{\ell}_t \leqslant \min_{1 \leqslant m \leqslant M} \left\{ \sum_{t=1}^{T} \ell_{m,t} + \mathcal{O}\left\{ \sqrt{M \ln(T) \left(\sum_{t=1}^{T} (\widehat{\ell}_t - \ell_{m,t})^2 \right)} \right\} \right\} . \tag{5.15}$$

Nous pouvons adapter et appliquer aisément l'astuce du gradient d'une manière similaire à celle des poids exponentiels dans le cas de ML-poly. Nous nommons *ML-poly-grad* cette version avec sous-gradient des pertes. Pour l'obtenir, il suffit de remplacer $\ell_{m,t}$ par $\widetilde{\ell}_{m,t}$, où $\widetilde{\ell}_{m,t}$ correspond à (5.13) en remplaçant les x_m par les $x_{m,t}$ et les les y par les y_t correspondants.

5.4 Résultats empiriques

Les jeux de données sont en tout point identiques à ceux décrits dans les parties 3.2.1 et 4.1. La différence majeure du cadre d'agrégation de fonctions de répartition réside évidemment dans les briques de base de l'agrégation de fonctions de répartition qui ne sont plus ici les prévisions ponctuelles, y_m et $x_{m,s}$, mais les fonctions de Heaviside associées H_{y_m} et $H_{x_{m,s}}$. Dans une première partie 5.4.1, nous déterminons la prévision de référence face à laquelle comparer les performances des différents algorithmes. Puis, la partie 5.4.2 détaille les résultats dans le cas de la pression réduite au niveau de la mer et la partie 5.4.2, dans celui de la vitesse du vent.

5.4.1 Prévision de référence

Dans les chapitres précédents, nous commencions par fournir les oracles convexes et linéaires (voir les parties 3.2.2 et 4.2) pour quantifier les performances asymptotiques potentielles des algorithmes. Nous omettons la partie correspondante ici, faute d'avoir eu le temps de calculer ces oracles. Similairement à la partie 3.2.2, nous cherchons donc à sélectionner une prévision qui sert de référence pour l'évaluation des prévisions issues des algorithmes. Trois choix sont en lice pour la détermination de ce point de référence :

- la fonction de répartition avec poids uniforme sur l'ensemble comprenant les 150 membres issus de TIGGE ;
- la fonction de répartition uniforme sur l'ensemble constitué des prévisions du centre européen (ECMWF),
- la fonction de Heaviside associée à la prévision déterministe du centre européen.

Ces fonctions de répartition semblent raisonnables car elles correspondent aux versions probabilistes des prévisions ponctuelles de moyenne d'ensemble et de prévision déterministe (voir la partie 3.2.2). Les scores correspondants sont reportés dans le tableau 5.1. Dans les deux cas, la prévision de l'échelon déterministe de Météo France obtient un

CRPS	Pression réduite au niveau de la mer (Pa)	Vitesse du vent ($\mathrm{m\,s^{-1}}$)
Ensemble TIGGE	27,98	1,15
Ensemble ECMWF	32,77	1,41
Échelon déterministe	23,38	0,84

TABLE 5.1 – Comparaison des valeurs de *CRPS* entre trois types de prévisions probabilistes dans le cas de la pression réduite au niveau de la mer et de la norme de la vitesse du vent. Les prévisions sont effectuées à un horizon de 6 heures. Ces scores sont évalués sur l'intégralité du domaine et de la période.

CRPS inférieur aux deux autres fonctions et signale cette prévision comme la plus performante.

Dans la suite de ce chapitre, les évaluations des différences relatives de *CRPS* ont donc lieu par rapport à l'échelon déterministe de Météo France. Si MF correspond à cet échelon déterministe de Météo France et A à une fonction de répartition issue d'un algorithme A donné, La différence relative des gains est définie selon la formule

$$\Delta_{\%}(\mathrm{A}, \mathrm{MF}) = \frac{CRPS_{\mathrm{MF}} - CRPS_{\mathrm{A}}}{CRPS_{\mathrm{MF}}}\,.$$

5.4.2 Résultats pour la pression réduite au niveau de la mer

Pour fournir une prévision compétitive de pression réduite au niveau de la mer, nous appliquons l'algorithme des poids exponentiels des sous-gradients des pertes 9 et ML-polygrad. L'unique étape préliminaire consiste à déterminer le paramètre d'apprentissage η de l'algorithme 9. Nous exécutons l'algorithme pour $\eta \in \{10^k, k \in \{-4, -3, -2, -1, 0\}\}$. Les bornes de cet intervalle conduisent à des différences relatives de *CRPS* négatives ou bien impliquent des problèmes numériques (pour $\eta = 1$). Les scores sont reportés dans le tableau 5.2.

η	*CRPS* (Pa)	différence relative (%)
0.0001	23,44	−2,98
0.001	16,69	26,68
0.01	15,35	32,55
0.1	21,73	4,54

TABLE 5.2 – Performance du *CRPS* en fonction au paramètre η de la prévision de pression réduite au niveau de la mer. Scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision probabiliste agrégée à un horizon de 6 heures. Le poids initial est uniforme. La période d'entraînement est de 100 jours et 100 cellules sont sélectionnées aléatoirement pour évaluer ces scores.

Une fois le bon ordre de grandeur trouvé, ici 0.01, nous réalisons une recherche manuelle du paramètre autour de cet ordre de grandeur. Le meilleur paramètre, avec 1000

cellules sélectionnées aléatoirement est alors $\eta = 0.005$, conduisant à un *CRPS* de 15,19 Pa, soit à une différence relative de *CRPS* de 33,27 %. Nous retenons donc ce paramètre pour toute la suite de l'étude de la pression.

Les résultats de performance des algorithmes d'agrégation pour la prévision d'incertitude dans le cas de la prévision de pression réduite au niveau de la mer sont reportés au tableau 5.3.

Type de prévision	*CRPS* (Pa)	différence relative (%)
Ensemble TIGGE	27,98	−19,67
Ensemble ECMWF	32,77	−40,15
Échelon déterministe	23,38	0,00
Poids exponentiels (avec paramètre optimal a posteriori)	15,58	33,36
ML-poly-grad	19,27	17,60

TABLE 5.3 – *CRPS* : Scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision probabiliste agrégée à un horizon de 6 heures. Le poids initial est uniforme chaque ensemble considéré. La période d'entraînement est de 100 jours. Ces scores sont évalués sur l'intégralité de la carte.

Lorsque le paramètre η est calibré a posteriori, l'algorithme $\mathcal{E}_\eta^{\text{grad}}$ obtient d'excellents résultats avec 33,36% de différence relative de *CRPS* par rapport à l'échelon déterministe de Météo France. Soulignons qu'il s'agit d'une version non-automatique de l'algorithme. Ces bons résultats sont aussi visibles graphiquement via les figures 5.3 et 5.4 pour $\mathcal{E}_\eta^{\text{grad}}$.

L'algorithme ML-poly-grad, obtient de très bons résultats avec 17,60% de différence relative de *CRPS* par rapport à l'échelon déterministe de Météo France. Les figures 5.5 et 5.6 soulignent ces bons résultats. Notons que l'algorithme est calibré séquentiellement, ce qui empêche de le comparer à $\mathcal{E}_\eta^{\text{grad}}$ qui est calibré a posteriori.

Les figures de résultat 5.3 et 5.5 font apparaître peu de variations brusques et locales. Nous retrouvons bien que la pression réduite au niveau de la mer constitue une variable synoptique.

La figure 5.2 montre plusieurs fonctions de répartition prévues à un instant fixe dans une cellule donnée ainsi qu'une représentation du *CRPS* associé.

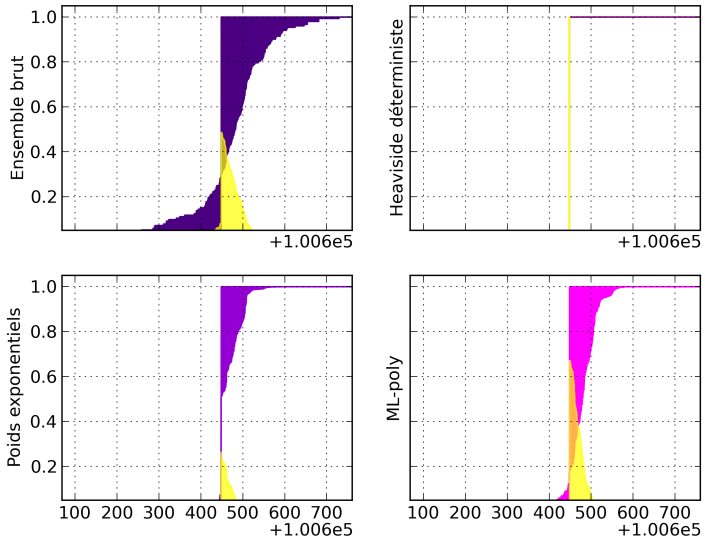

FIGURE 5.2 – Représentation de plusieurs prévisions probabilistes et visualisation de leurs *CRPS* respectifs, dans une cellule typique (de coordonnées $(-13°, 37°)$), à l'échéance 320 (16 août 2012). Les fonctions de répartition de haut en bas, de gauche à droite sont la pondération uniforme de l'ensemble, l'échelon déterministe, $\mathcal{E}_\eta^{\text{grad}}$ et ML-poly-grad.

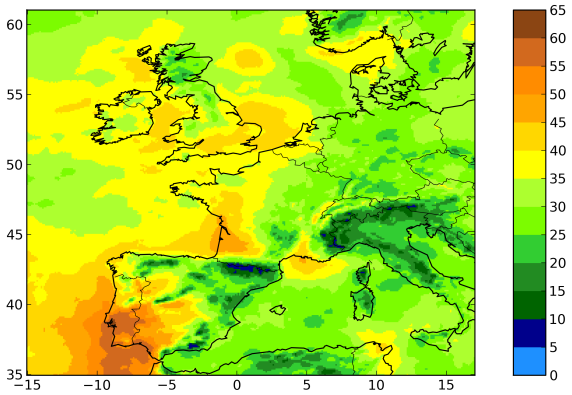

FIGURE 5.3 – Représentation graphique des différences relatives du *CRPS* entre la fonction de répartition agrégée par les poids exponentiels $\mathcal{E}_\eta^{\text{grad}}$ avec paramètre optimal rétrospectif et l'échelon déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i, j) est représentée la différence relative : $\left(r_{(i,j)}^{\text{det}} - r_{(i,j)}^{\text{agr}}\right)/r_{(i,j)}^{\text{det}}$ où $r_{(i,j)}^{\text{agr}}$ (respectivement $r_{(i,j)}^{\text{det}}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Les poids initiaux sont uniformes. La période d'entraînement est de 100 jours.

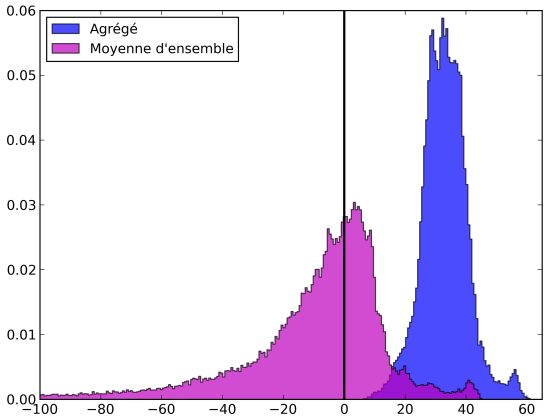

FIGURE 5.4 – Histogramme de la différence relative par rapport à l'échelon déterministe du *CRPS* entre la fonction de répartition uniforme issue de l'ensemble TIGGE (en magenta) et de l'agrégé de ML-poly-grad (en bleu). Les individus représentés sont les différences relatives du *CRPS* pour l'agrégé $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ et les différences relatives du *CRPS* pour la moyenne d'ensemble $\left(r^{\text{det}}_{(i,j)} - r^{\text{ens}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$ et $r^{\text{ens}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe et de la moyenne d'ensemble) dans la cellule (i, j). Les prévisions sont à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). Le *CRPS* est pris par rapport à l'échelon d'analyse Météo France. Les poids initiaux sont uniformes.

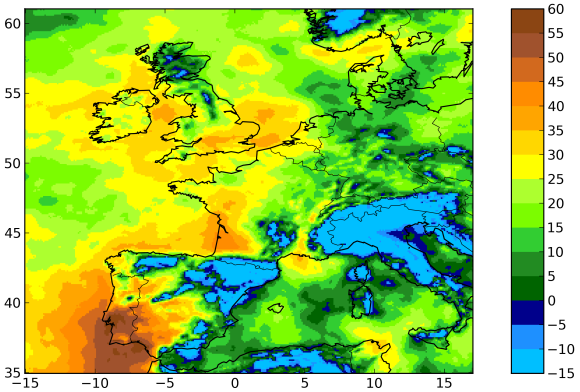

FIGURE 5.5 – Représentation graphique des différences relatives du *CRPS* entre la fonction de répartition agrégée par ML-poly-grad et l'échelon déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i, j) est représentée la différence relative : $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). L'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est uniforme. La période d'entraînement est de 100 jours.

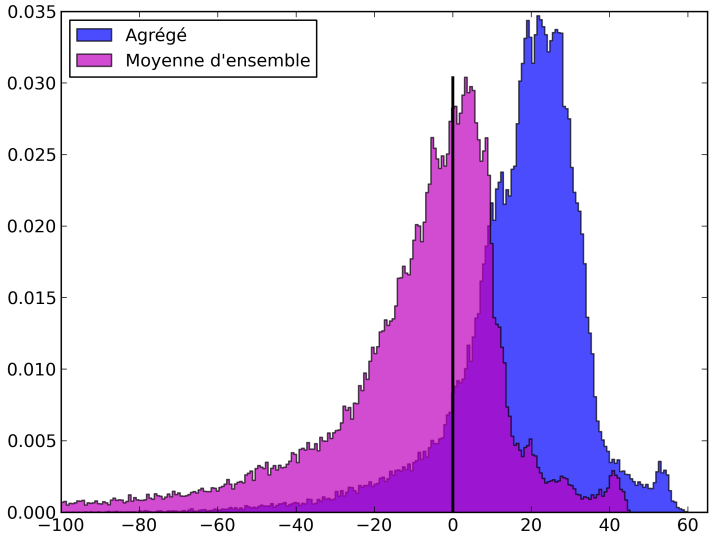

FIGURE 5.6 – Histogramme de la différence relative par rapport à l'échelon déterministe du *CRPS* entre la fonction de répartition uniforme issue de l'ensemble TIGGE (en magenta) et de l'agrégé ML-poly-grad (en bleu). Les individus représentés sont les différences relatives du *CRPS* pour l'agrégé $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ et les différences relatives du *CRPS* pour la moyenne d'ensemble $\left(r^{\text{det}}_{(i,j)} - r^{\text{ens}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$ et $r^{\text{ens}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe et de la moyenne d'ensemble) dans la cellule (i, j). Les prévisions sont à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). Le *CRPS* est pris par rapport à l'échelon d'analyse Météo France. Le poids initial est uniforme.

5.4.3 Résultats pour la vitesse du vent

Nous renouvelons la démarche de recherche du paramètre η avec des modalités identiques à la partie 5.4.2. Nous exécutons l'algorithme pour $\eta \in \{10^k, k \in \{-3, -2, -1, 0, 1\}\}$. Les résultats sont reportés dans le tableau 5.4.

η	*CRPS* ($\mathrm{m\,s^{-1}}$)	différence relative (%)
0.001	1,30	−20,00
0.01	0,94	12,55
0.1	0,80	25,46
1	1,01	6,68
10	1,34	−24,05

TABLE 5.4 – Performance du *CRPS* en fonction du paramètre η : scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision probabiliste agrégée à un horizon de 6 heures. Le poids initial est uniforme chaque ensemble considéré. La période d'entraînement est de 100 jours. 100 cellules sont sélectionnées aléatoirement pour évaluer ces scores.

Une fois le bon ordre de grandeur trouvé, ici 0.1, nous réalisons une recherche manuelle du paramètre autour de cet ordre de grandeur. Le meilleur paramètre, avec 1000 cellules sélectionnées aléatoirement est alors $\eta = 0.09$, conduisant à un *CRPS* de $0{,}77\mathrm{m\,s^{-1}}$, soit à une différence de *CRPS* relative de 25,49%. Nous retenons donc ce paramètre pour toute la suite de l'étude.

Type de prévision	*CRPS* ($\mathrm{m\,s^{-1}}$)	différence relative (%)
Ensemble TIGGE	1,15	−36,57
Ensemble ECMWF	1,41	−66,85
Échelon déterministe	0,84	0,00
Poids exponentiels (avec paramètre optimal a posteriori)	0,64	24,12
ML-poly-grad	0,73	13,07

TABLE 5.5 – *CRPS* : Scores et différences relatives par rapport à la prévision déterministe de Météo France pour une prévision probabiliste agrégée à un horizon de 6 heures. Le poids initial est uniforme chaque ensemble considéré. La période d'entraînement est de 100 jours. 1000 cellules sont sélectionnées aléatoirement pour évaluer ces scores.

L'analyse est similaire au cas de la pression. Lorsque le paramètre η est calibré a posteriori, l'algorithme $\mathcal{E}_\eta^{\mathrm{grad}}$ obtient d'excellents résultats avec 24,12% de différence relative de *CRPS* par rapport à l'échelon déterministe de Météo France. Soulignons qu'il s'agit d'une version non-automatique de l'algorithme. Ces bons résultats sont aussi visibles graphiquement via les figures 5.3 et 5.4 pour $\mathcal{E}_\eta^{\mathrm{grad}}$.

L'algorithme ML-poly-grad, obtient de très bons résultats avec 13,07% de différence relative de *CRPS* par rapport à l'échelon déterministe de Météo France. Les figures 5.5 et 5.6 soulignent ces bons résultats. Notons que l'algorithme est calibré séquentielle-

ment, ce qui empêche de le comparer à $\mathcal{E}_\eta^{\text{grad}}$ qui est calibré a posteriori.

Au contraire de la pression réduite au niveau de la mer, les cartes de résultat 5.8 et 5.10 montrent des variations rapides en espace et brusques qui sont bien reliées au caractère local de la variable vitesse du vent.

La figure 5.2 montre plusieurs fonctions de répartition prévues à un instant fixe dans une cellule donnée ainsi qu'une représentation du *CRPS* associé.

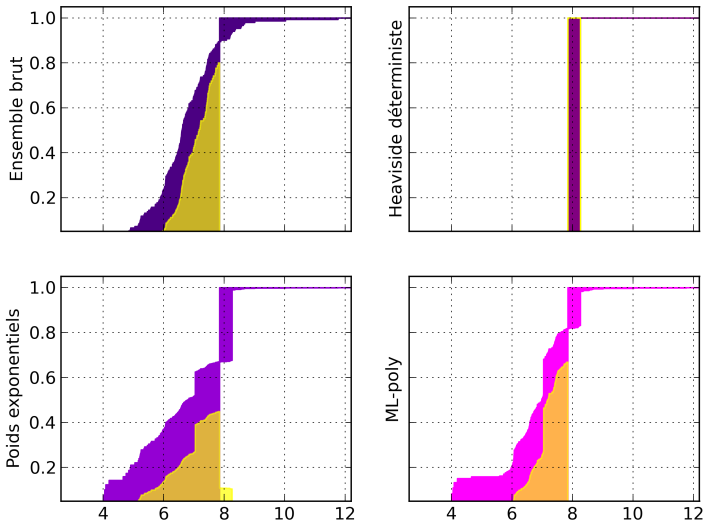

FIGURE 5.7 – Représentation de plusieurs prévisions probabilistes et visualisation de leurs *CRPS* respectifs, dans une cellule typique (de coordonnées $(-13°, 37°)$), à l'échéance 320 (16 août 2012). Les fonctions de répartition de haut en bas, de gauche à droite sont la pondération uniforme de l'ensemble, l'échelon déterministe, $\mathcal{E}_\eta^{\text{grad}}$ et ML-poly-grad.

5.5 Conclusion

Après avoir passé en revue l'état de l'art des scores probabilistes, nous montrons comment les généralisations du score de Brier vers des versions plus continues conservent les propriétés mathématiques souhaitables pour de tels scores. Ces raisonnements nous amènent au *CRPS*, score probabiliste prenant des fonctions de répartition en argument que nous employons dans la suite du chapitre. Pour ce score, nous choisissons comme prédicteurs élémentaires les fonctions échelons : partant des mêmes ensembles de prévisions ponctuelles que dans les chapitres 3 et 4, nous construisons les fonctions de répartition de Heaviside dont les seuils correspondent aux simulations d'ensemble, aux prévisions déterministes et aux analyses.

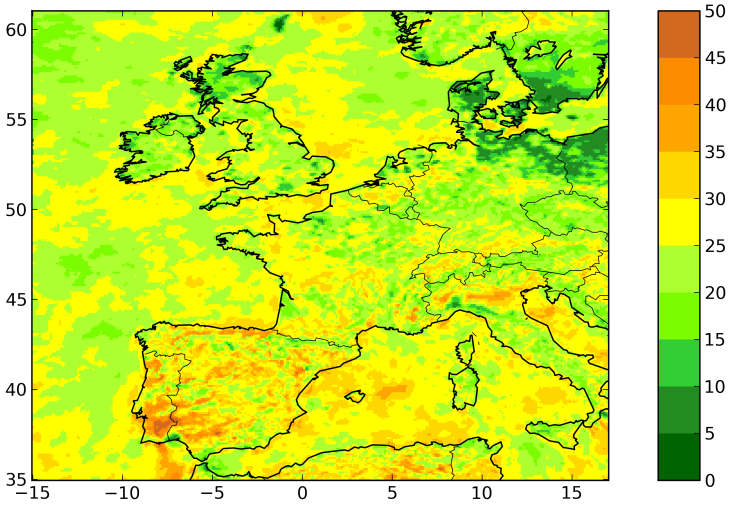

FIGURE 5.8 – Représentation graphique des différences relatives du *CRPS* entre la fonction de répartition agrégée par les poids exponentiels $\mathcal{E}_\eta^{\text{grad}}$ avec paramètre optimal rétrospectif et l'échelon déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i,j) est représentée la différence relative : $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe) dans la cellule (i,j). Les paramètres sont optimaux et fixés à l'avance, l'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est uniforme. La période d'entraînement est de 100 jours.

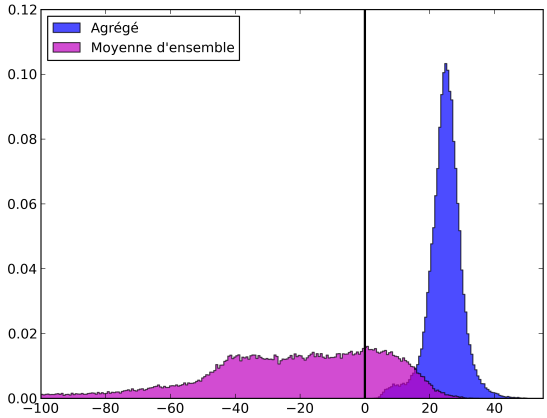

FIGURE 5.9 – Histogramme de la différence relative par rapport à l'échelon déterministe du *CRPS* entre la fonction de répartition uniforme issue de l'ensemble TIGGE (en magenta) et de l'agrégé de ML-poly-grad (en bleu). Les individus représentés sont les différences relatives du *CRPS* pour l'agrégé $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ et les différences relatives du *CRPS* pour la moyenne d'ensemble $\left(r^{\text{det}}_{(i,j)} - r^{\text{ens}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$ et $r^{\text{ens}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe et de la moyenne d'ensemble) dans la cellule (i, j). Les prévisions sont à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). Le *CRPS* est pris par rapport à l'échelon d'analyse Météo France. Le poids initial est uniforme.

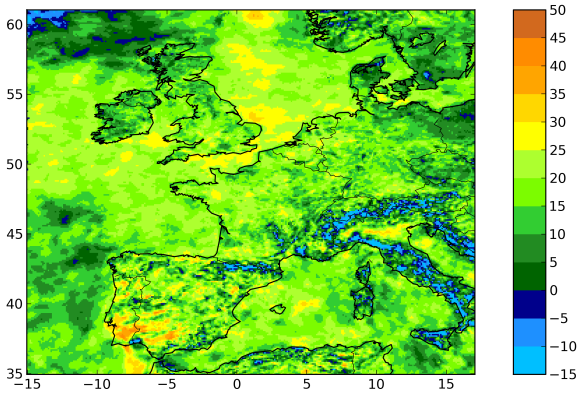

FIGURE 5.10 – Représentation graphique des différences relatives du *CRPS* entre la fonction de répartition agrégée par ML-poly-grad et l'échelon déterministe Météo France à un horizon de 6 heures. Dans chaque cellule (i, j) est représentée la différence relative : $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe) dans la cellule (i, j). L'ensemble compte 152 membres (les déterministes d'ECMWF et de Météo France sont compris). Le poids initial est uniforme. La période d'entraînement est de 100 jours.

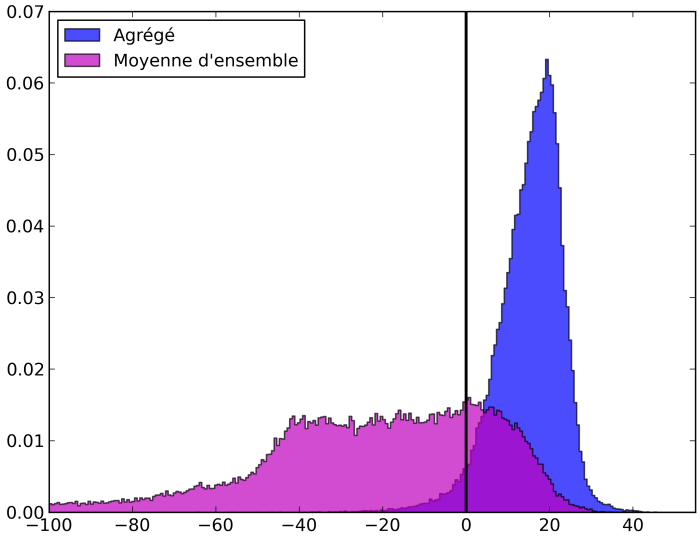

FIGURE 5.11 – Histogramme de la différence relative par rapport à l'échelon déterministe du *CRPS* entre la fonction de répartition uniforme issue de l'ensemble TIGGE (en magenta) et de l'agrégé ML-poly-grad (en bleu). Les individus représentés sont les différences relatives du *CRPS* pour l'agrégé $\left(r^{\text{det}}_{(i,j)} - r^{\text{agr}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ et les différences relatives du *CRPS* pour la moyenne d'ensemble $\left(r^{\text{det}}_{(i,j)} - r^{\text{ens}}_{(i,j)}\right)/r^{\text{det}}_{(i,j)}$ où $r^{\text{agr}}_{(i,j)}$ (respectivement $r^{\text{det}}_{(i,j)}$ et $r^{\text{ens}}_{(i,j)}$) est le *CRPS* moyen de la prévision agrégée (respectivement déterministe et de la moyenne d'ensemble) dans la cellule (i, j). Les prévisions sont à une échéance de 6 heures, pour des paramètres optimaux. L'ensemble fait 152 membres (le déterministe Météo France et le déterministe ECMWF sont inclus). Le *CRPS* est pris par rapport à l'échelon d'analyse Météo France. Le poids initial est uniforme.

Vu que le *CRPS* est une perte convexe et afin d'obtenir une prévision ayant un sens physique, nous employons des algorithmes convexes d'agrégation. Nous précisons ainsi deux algorithmes : celui des poids exponentiels des sous-gradients de perte (*Exponentiated Gradient*) et ML-poly que nous employons ensuite en pratique. Concernant la pression réduite au niveau de la mer, nous obtenons une différence relative de CRPS de 33,36% pour *EG* et 17,60% pour ML-poly (partie 5.4.2) ; et pour la vitesse du vent, une différence relative de CRPS de 24,12% et 13,07% respectivement (partie 5.4.3). En pratique, cette méthode d'agrégation est donc prometteuse d'une manière comparable aux cas de prévisions ponctuelles précédemment décrites. Elle permet, avec un minimum d'hypothèses et des garanties robustes, de mettre à profit les données existantes pour proposer une prévision déterministe performante.

6 Conclusion

Dans cette thèse, notre intérêt s'est porté sur des problèmes de prévision tour après tour et sur les stratégies automatiques de prévision associées. Dans le cadre de la prévision séquentielle, il est possible d'obtenir des garanties mathématiques robustes (valables dans des cas de figure très généraux) pour les algorithmes de prévision, ce que nous exposons dans un premier temps. Puis nous appliquons ces algorithmes à la prévision concrète de grandeurs météorologiques. Enfin, nous nous sommes intéressés aux déclinaisons théoriques et pratiques dans un cadre de prévision de fonctions de répartition.

Dans le chapitre 2, nous avons étudié la prévision en ligne de processus ergodiques stationnaires. Pour ce faire, nous avons considéré un ensemble de prédictions de séquences individuelles et proposé un arbre de régression déterministe dont les performances sont asymptotiquement aussi bonnes que le meilleur des prédicteurs L-Lipschitz. Nous avons montré comment le regret obtenu conduit à une optimalité asymptotique par rapport aux classes de processus stationnaires ergodiques.

Dans le cas de l'arbre de régression déterministe, certains cas de figure généraux n'ont pas été traités. Les observations et prévisions appartiennent par hypothèse à un ensemble $[0, 1]$ et ce segment a pu être généralisé à tout segment borné quelconque. Mais, nous avons laissé de côté le cas non-borné qui rendrait l'algorithme plus général encore.

Dans le chapitre 3, nous avons présenté le jeu de données météorologiques de pression réduite au niveau de la mer en détaillant les origines et les spécificités des simulations et des analyses. Cela nous a permis d'exposer puis de mettre en pratique un plan d'étude de ces simulations. À cet effet, différentes versions de l'algorithme d'agrégation ridge ont été décrites (partie 3.1.6) puis implémentées (partie 3.2.3) par ordre croissant d'automatisation. En effet, en opérationnel, l'algorithme, si l'on veut qu'il soit pérenne, doit être le plus autonome possible et se fonder le moins possible sur la calibration manuelle des paramètres par un statisticien.

Rappelons les résultats : la régression ridge avec paramètres optimaux rétrospectifs améliore les résultats de prévision opérationnelle de 17,35% (partie 3.2.3) tandis que la méthode complètement automatique, avec adaptation locale en ligne sur une grille de paramètres, nécessairement plus coûteuse en temps de calcul, conduit à une différence relative de 17,69% (partie 3.2.3). L'un des intérêts des algorithmes d'agrégation est qu'ils accroissent la précision des prévisions quelles que soient les simulations sous-jacentes.

Dans le chapitre 4, la vitesse du vent apparaît comme une variable locale (à l'opposé de synoptique) et bimodale. Là encore, les résultats sont intéressants : la régression ridge avec paramètres optimaux rétrospectifs améliore les résultats de prévision opérationnelle de 8,95% (partie 4.3.1) tandis que la méthode complètement automatique, avec adaptation locale en ligne sur une grille de paramètres, donne une différence relative de 6,48% (partie 4.3.2).

Ces applications de la théorie des suites individuelles permettent de montrer la portée et les performances des algorithmes. Le travail mené permet de valider la pertinence de ces méthodes pour un emploi opérationnel.

Dans le chapitre 5, après avoir passé en revue l'état de l'art des scores probabilistes, nous montrons comment les généralisations du score de Brier vers des versions plus continues conservent les propriétés mathématiques souhaitables pour de tels scores. Ces raisonnements nous amènent au *CRPS*, score probabiliste prenant des fonctions de répartition en argument que nous employons dans la suite du chapitre. Pour ce score, nous choisissons comme prédicteurs élémentaires les fonctions échelons : partant des mêmes ensembles de prévisions ponctuelles que dans les chapitres 3 et 4, nous construisons les fonctions de répartition de Heaviside dont les seuils correspondent aux simulations d'ensemble, aux prévisions déterministes et aux analyses.

Vu que le *CRPS* est une perte convexe et afin d'obtenir une prévision ayant un sens physique, nous employons des algorithmes convexes d'agrégation. Nous précisons ainsi deux algorithmes : celui des poids exponentiels des sous-gradients de perte (*Exponentiated Gradient*) et ML-poly que nous employons ensuite en pratique. Concernant la pression réduite au niveau de la mer, nous obtenons une différence relative de CRPS de 33,36% pour *EG* et 17,60% pour ML-poly (partie 5.4.2) ; et pour la vitesse du vent, une différence relative de CRPS de 24,12% et 13,07% respectivement (partie 5.4.3). En pratique, cette méthode d'agrégation est donc prometteuse d'une manière comparable aux cas de prévisions ponctuelles précédemment décrites. Elle permet, avec un minimum d'hypothèses et des garanties robustes, de mettre à profit les données existantes pour proposer une prévision déterministe performante.

Les travaux réalisés sur les données concrètes montrent que les algorithmes d'agrégation séquentielle sont performants pour combiner les membres de l'ensemble au sens du *CRPS*. Une perspective possible serait donc là encore d'appliquer les méthodes décrites dans cette thèse à la prévision opérationnelle de fonctions de répartition. Par ailleurs, en considérant que la fonction de répartition de chaque observation y_s est une fonction de Heaviside H_{y_s}, nous ignorons sciemment l'incertitude sur les observations, dont certaines estimations sont disponibles. Proposer des versions alternatives de fonctions de répartition (fonction rampe ou fonction de répartition de loi normale sous-jacente par exemple) améliorerait la modélisation générale du problème de prévision et par conséquent la précision des prévisions probabilistes issues des algorithmes.

Bibliographie

[Alg94] Algoet, P. H. "The strong law of large numbers for sequential decisions under uncertainty". In : *IEEE Transactions on Information Theory* 40.3 (1994), p. 609–633.

[Bia+10] Biau, G., Bleakley, K., Györfi, L. et Ottucsák, G. "Nonparametric sequential prediction of time series". In : *Journal of Nonparametric Statistics* 22.3 (2010), p. 297–317.

[BP11] Biau, G. et Patra, B. "Sequential Quantile Prediction of Time Series." In : *IEEE Transactions on Information Theory* 57.3 (2011), p. 1664–1674.

[Bos96] Bosq, D. *Nonparametric statistics for stochastic processes : estimation and prediction*. Lecture notes in statistics. New York : Springer, 1996.

[Bre57] Breiman, L. "The individual ergodic theorem of information theory". In : *Annals of Mathematical Statistics* 31 (1957), p. 809–811.

[Bre+84] Breiman, L., Friedman, J., Stone, C. J. et Olshen, R. A. *Classification and Regression Trees*. Sous la dir. de no. Belmont, CA : Wadsworth International Group, 1984.

[Bri50] Brier, G. W. "Verification of forecasts expressed in terms of probability". In : *Monthly Weather Review* 78.1 (1950), p. 1–3. doi : http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[BS07a] Bröcker, J. et Smith, L. A. "Increasing the reliability of reliability diagrams". In : *Weather and forecasting* 22.3 (2007), p. 651–661.

[BS07b] Bröcker, J. et Smith, L. A. "Scoring probabilistic forecasts : The importance of being proper". In : *Weather and Forecasting* 22.2 (2007), p. 382–388.

[BD91] Brockwell, P. J. et Davis, R. A. *Time series : theory and methods*. Springer Series in Statistics. New York : Springer, 1991.

[CMR05] Cappé, O., Moulines, E. et Rydén, T. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005, p. xviii+652. isbn : 978-0387-40264-2 ; 0-387-40264-0.

[Ces99] Cesa-Bianchi, N. "Analysis of Two Gradient-Based Algorithms for On-Line Regression". In : *Journal of Computer and System Sciences* 59.3 (1999), p. 392–411. issn : 0022-0000. doi : http://dx.doi.org/10.1006/jcss.1999.1635. url : http://www.sciencedirect.com/science/article/pii/S0022000099916355.

[CL06] Cesa-Bianchi, N. et Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

[Cho65] Chow, Y. S. "Local convergence of martingales and the law of large numbers". In : *Annals of Mathematical Statistics* 36 (1965), p. 552–558.

[CW99] Clemen, R. T. et Winkler, R. L. "Combining probability distributions from experts in risk analysis". In : *Risk analysis* 19.2 (1999), p. 187–203.

[Daw08] Dawid, A. "Comments on : Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds". In : *TEST* 17.2 (2008), p. 243–244. issn : 1133-0686. doi : 10.1007/s11749-008-0118-6. url : http://dx.doi.org/10.1007/s11749-008-0118-6.

[Dev+13] Devaine, M., Gaillard, P., Goude, Y. et Stoltz, G. "Forecasting electricity consumption by aggregating specialized experts - A review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions". In : *Machine Learning* 90.2 (2013), p. 231–260.

[Eps69] Epstein, E. S. "A Scoring System for Probability Forecasts of Ranked Categories". In : *Journal of Applied Meteorology* 8.6 (1969), p. 985–987. doi : http://dx.doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

[Fer14] Ferro, C. "Fair scores for ensemble forecasts". In : *Quarterly Journal of the Royal Meteorological Society* 140.683 (2014), p. 1917–1923.

[FV98] Foster, D. et Vohra, R. "Asymptotic calibration". In : *Biometrika* 85 (1998), p. 379–390.

[Fre+97] Freund, Y., Schapire, R. E., Singer, Y. et Warmuth, M. K. "Using and combining predictors that specialize". In : *Proceedings of STOC*. 1997, p. 334–343.

[FFS13] Fricker, T. E., Ferro, C. A. T. et Stephenson, D. B. "Three recommendations for evaluating climate predictions". In : *Meteorological Applications* 20.2 (2013), p. 246–255. issn : 1469-8080. doi : 10.1002/met.1409. url : http://dx.doi.org/10.1002/met.1409.

[Gai15] Gaillard, P. "Contributions à l'agrégation séquentielle robuste d'experts : travaux sur l'erreur d'approximation et la prévision en loi. Applications à la prévision pour les marchés de l'énergie." Thèse de doct. Université Paris-Sud 11, 2015.

[GSv14] Gaillard, P., Stoltz, G. et van Erven, T. "A Second-order Bound with Excess Losses". In : *Proceedings of COLT*. 2014.

[GA11] Geweke, J. et Amisano, G. "Optimal prediction pools". In : *Journal of Econometrics* 164.1 (2011), p. 130–141.

[GR07] Gneiting, T. et Raftery, A. E. "Strictly proper scoring rules, prediction, and estimation". In : *Journal of the American Statistical Association* 102.477 (2007), p. 359–378.

[Goo52] Good, I. J. “Rational Decisions”. English. In : *Journal of the Royal Statistical Society. Series B (Methodological)* 14.1 (1952), pages. ISSN : 00359246. URL : http://www.jstor.org/stable/2984087.

[Gyö+89] Györfi, L., Härdle, W., Sarda, P. et Vieu, P. *Nonparametric curve estimation from time series*. Sous la dir. de Györfi, L. Lecture notes in statistics 60. Berlin : Springer-Verlag, 1989.

[GLF01] Györfi, L., Lugosi, G. et Fargas, R. T. *Strategies for Sequential Prediction of Stationary Time Series*. 2001.

[GO07] Gyorfi, L. et Ottucsak, G. “Sequential Prediction of Unbounded Stationary Time Series”. In : *Information Theory, IEEE Transactions on* 53.5 (2007), p. 1866–1872.

[Her00] Hersbach, H. “Decomposition of the continuous ranked probability score for ensemble prediction systems”. In : *Weather and Forecasting* 15.5 (2000), p. 559–570.

[Hoe+99] Hoeting, J. A., Madigan, D., Raftery, A. E. et Volinsky, C. T. “Bayesian Model Averaging : A Tutorial”. In : *STATISTICAL SCIENCE* 14.4 (1999), p. 382–417.

[KW97] Kivinen, J. et Warmuth, M. K. “Exponentiated Gradient Versus Gradient Descent for Linear Predictors”. In : *Information and Computation* 132.1 (1997), p. 1–63.

[LW94] Littlestone, N. et Warmuth, M. K. “The Weighted Majority Algorithm”. In : *Inf. Comput.* 108.2 (fév. 1994), p. 212–261. ISSN : 0890-5401. DOI : 10.1006/inco.1994.1009. URL : http://dx.doi.org/10.1006/inco.1994.1009.

[MF98] Merhav, N. et Feder, M. “Universal Prediction.” In : *IEEE Transactions on Information Theory* 44.6 (1998), p. 2124–2147.

[MW11] Morvai, G. et Weiss, B. “Nonparametric sequential prediction for stationary processes.” In : *Ann. Probab.* 39.3 (2011), p. 1137–1160.

[Mur73] Murphy, A. “A new vector partition of the probability score”. In : *Journal of applied Meteorology* 12.4 (1973), p. 595–600. DOI : 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

[RS14] Rakhlin, A. et Sridharan, K. “Online Nonparametric Regression”. In : *Proceedings of COLT*. 2014.

[Ric05] Richardson, D. “The THORPEX interactive grand global ensemble (TIGGE)”. In : *Geophysical Research Abstracts*. T. 7. 2005, p. 02815.

[Roo+14] Rooij, S. de, van Erven, T., Grünwald, P. D. et Koolen, W. M. “Follow the Leader If You Can, Hedge If You Must”. In : *Journal of Machine Learning Research* 15 (2014), p. 1281–1316.

[Rud91] Rudin, W. *Functional Analysis*. McGraw-Hill Science, 1991.

[Sej+13] Sejdinovic, D., Sriperumbudur, B., Gretton, A. et Fukumizu, K. "Equivalence of distance-based and RKHS-based statistics in hypothesis testing". In : *Ann. Statist.* 41.5 (oct. 2013), p. 2263–2291. DOI : 10.1214/13-AOS1140. URL : http://dx.doi.org/10.1214/13-AOS1140.

[Sto10] Stoltz, G. "Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique". In : *Journal de la Société Française de Statistique* 151.2 (2010), p. 66–106. URL : https://hal-hec.archives-ouvertes.fr/hal-00637060.

[SL07] Stoltz, G. et Lugosi, G. "Learning correlated equilibria in games with compact sets of strategies". In : *Games and Economic Behavior* 59 (2007), p. 187–208.

[SR05a] Szekely, G. J. et Rizzo, M. L. "Hierarchical Clustering via Joint Between-Within Distances : Extending Ward's Minimum Variance Method". In : *Journal of Classification* 22.2 (2005), p. 151–183. URL : http://EconPapers.repec.org/RePEc:spr:jclass:v:22:y:2005:i:2:p:151-183.

[SR05b] Székely, G. J. et Rizzo, M. L. "A new test for multivariate normality". In : *Journal of Multivariate Analysis* 93.1 (2005), p. 58–80. ISSN : 0047-259X. DOI : http://dx.doi.org/10.1016/j.jmva.2003.12.002. URL : http://www.sciencedirect.com/science/article/pii/S0047259X03002124.

[SR13] Székely, G. J. et Rizzo, M. L. "Energy statistics : A class of statistics based on distances". In : *Journal of Statistical Planning and Inference* 143.8 (2013), p. 1249–1272. ISSN : 0378-3758. DOI : http://dx.doi.org/10.1016/j.jspi.2013.03.018. URL : http://www.sciencedirect.com/science/article/pii/S0378375813000633.

[Vov90] Vovk, V. G. "Aggregating Strategies." In : *Proceedings of COLT*. 1990, p. 371–386.

[Vov05] Vovk, V. "On-line regression competitive with reproducing kernel Hilbert spaces". In : *CoRR* abs/cs/0511058 (2005).

[Vov06] Vovk, V. "Metric entropy in competitive on-line prediction". In : *CoRR* abs/cs/0609045 (2006).

[VZ09] Vovk, V. et Zhdanov, F. "Prediction With Expert Advice For The Brier Game". In : *J. Mach. Learn. Res.* 10 (déc. 2009), p. 2445–2471. ISSN : 1532-4435. URL : http://dl.acm.org/citation.cfm?id=1577069.1755868.