Prévision séquentielle par mélange de prédicteurs.

Pierre Gaillard

Université Paris SUD

22 janvier 2013

Plan

I. Introduction à la prévision séquentielle à l'aide d'experts

Le cadre : prévision séquentielle à l'aide d'experts

On a une suite arbitraire à prévoir : $y_1, y_2, \ldots, y_T \in \mathcal{Y}$

On dispose d'un ensemble $E = \{1, \ldots, d\}$ d'experts.

À chaque instant t,

(1) Chaque expert propose une prévision $x_{i,t} \in \mathcal{X}$

(2) Le joueur leur attribue à chacun des pondérations $\widehat{p}_t = (\widehat{p}_{1,t}, \ldots, \widehat{p}_{d,t}) \in \Delta_d$ et prévoit

$$\widehat{y}_t = \sum_{i \in E} \widehat{p}_{i,t} x_{i,t} \in \mathcal{X}$$

(3) y_t est révélée et les erreurs de prévisions $\ell(\widehat{y}_t, y_t)$ et $\ell(x_{i,t}, y_t)$ peuvent être calculées.

Exemple, la prévision de consommation électrique à EDF

Objectif

Prévision de la consommation électrique à court terme.

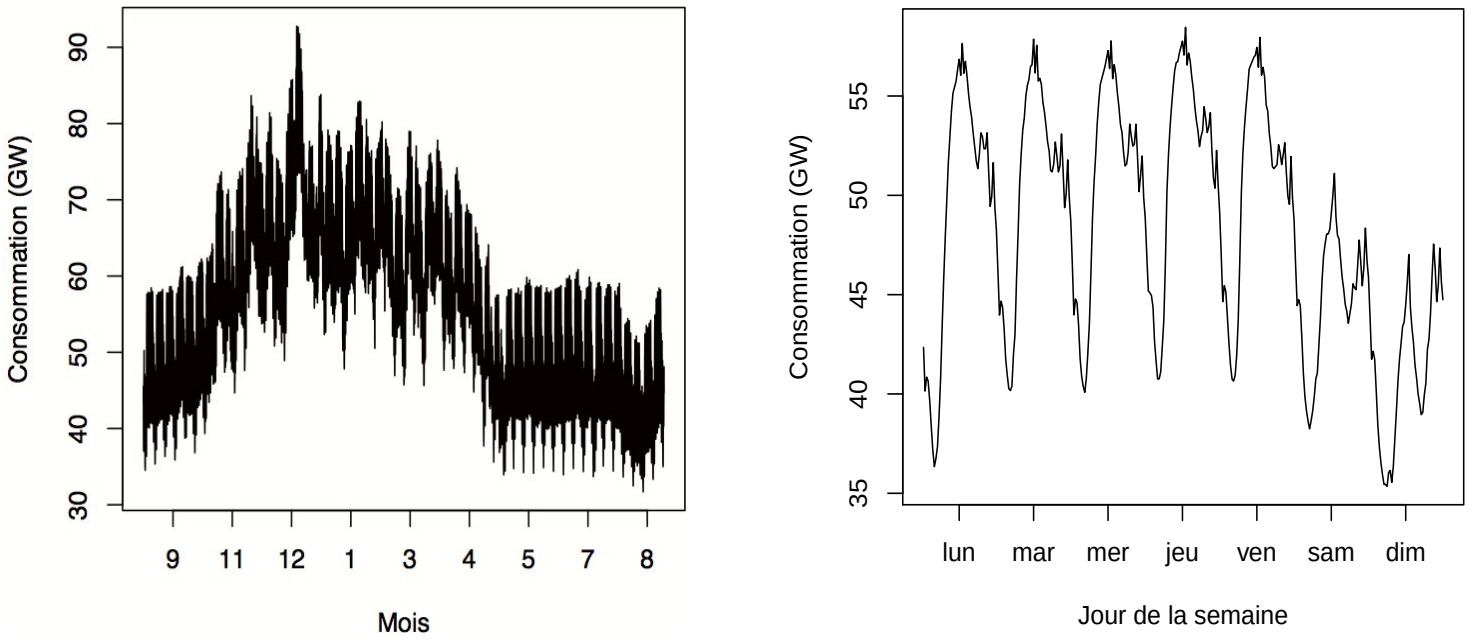


Figure : La consommation électrique en France au cours d'une année et d'une semaine.

Exemple, la prévision de consommation électrique à EDF

L'électricité se stocke difficilement. La prévision de consommation électrique représente donc une enjeu important.

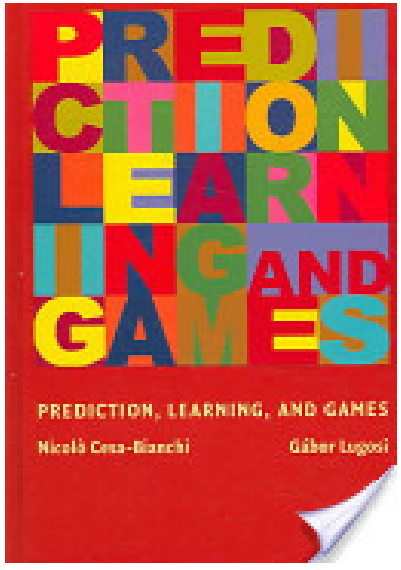
Parallèlement,

- EDF a créé de nombreuses méthodes de prévisions ces 20 dernières années.
- le paysage électrique français évolue rapidement.

$\Rightarrow$ EDF ne sait plus à laquelle se fier !

Références

Le livre incontournable.
Prediction Learning and Games, N. Cesa-Bianchi and G. Lugosi, 2006.



Application à EDF.

Forecasting the electricity consumption by aggregating specialized experts, M. Devaine and al, 2012.

Objectif

Rappel : À l'instant t, x_t sont les prévisions des experts, $\widehat{p}_t \in \Delta_d$ est le vecteur de poids choisi, $\ell(\widehat{p}_t^\top x_t, y_t)$ est notre perte instantanée et $\ell(x_{i,t}, y_t)$ celle de l'expert i.

Notre but est de minimiser notre perte cumulée

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

Difficulté

Impossible d'assurer une faible erreur cumulée $\widehat{L}_T$ de façon absolue. Si les experts sont tous mauvais, on l'est aussi.

On va donc tenter d'assurer une faible perte cumulée comparativement à celles des experts $L_{i,T} = \sum_{t=1}^{T} \ell(x_{i,t}, y_t)$.

La notion de regret, cas usuel

Rappel : À l'instant t, x_t sont les prévisions des experts, $\widehat{p}_t \in \Delta_d$ est le vecteur de poids choisi, $\ell(\widehat{p}_t^\top x_t, y_t)$ est notre perte instantanée. $\widehat{L}_T$ est notre perte cumulée et $L_{i,T}$ celle de l'expert i.

Notre perte cumulée s'écrit

$$\widehat{L}_T = \underbrace{\min_{i=1,\dots,d} L_{i,T}}_{\text{Erreur d'approximation}} + \underbrace{R_T}_{\text{Erreur d'estimation}}$$

Objectif du joueur

Avoir un regret moyen qui converge uniformément vers 0

$$\sup_{\ell_1,\dots,\ell_T \in [0,1]^d} \frac{R_T}{T} = 0\,.$$

C'est possible ! De nombreux algorithmes ont un regret optimal

$$R_T = O\left(\sqrt{T \ln d}\right).$$

Regret usuel (la meilleure combinaison convexe)

Rappel : À l'instant t, x_t sont les prévisions des experts, $\widehat{p}_t \in \Delta_d$ est le vecteur de poids choisi, $\ell(\widehat{p}_t^\top x_t, y_t)$ est notre perte instantanée. $\widehat{L}_T$ est notre perte cumulée et $L_{i,T}$ celle de l'expert i.

Notre perte cumulée s'écrit aussi

$$\widehat{L}_T = \underbrace{\inf_{q \in \Delta_d} \sum_{t=1}^{T} \ell(q^\top x_t, y_t)}_{\text{Erreur d'approximation}} + \underbrace{R_T}_{\text{Erreur d'estimation}}$$

Objectif du joueur

Avoir un regret moyen qui converge uniformément vers 0

$$\sup_{\ell_1, \dots, \ell_T \in [0,1]^d} \frac{R_T}{T} = 0\,.$$

C'est possible ! De nombreux algorithmes ont un regret optimal

$$R_T = O\left(\sqrt{T \ln d}\right).$$

Pour simplifier l'exposé : le gradient trick !

On suppose $\ell(\cdot, y)$ sont convexes pour tout $y \in \mathcal{Y}$. Alors,

$$\Psi_t \ : p \mapsto \ell(p^\top x_t, y_t)$$

sont convexes pour tout t.

Donc pour $\widehat{p}_t$ dans l'intérieur de Δ_d,

$$\begin{aligned} \ell(\widehat{y}_t, y_t) - \ell(q^\top x_t) &= \ell(\widehat{p}_t^\top x_t, y_t) - \ell(q^\top x_t) \\ &\leqslant \nabla \Psi_t(\widehat{p}_t)^\top (\widehat{p}_t - q) \\ &= \widehat{p}_t^\top \ell_t - q^\top \ell_t \end{aligned}$$

avec la notation $\ell_t = \nabla \Psi_t(\widehat{p}_t) \in \mathbb{R}^d$.

Le regret est donc borné par

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \inf_q \sum_{t=1}^{T} \ell(q^\top x_t, y_t) \leqslant \sum_{t=1}^{T} \widehat{p}_t^\top \ell_t - inf_q \sum_{t=1}^{T} q^\top \ell_t$$

Cadre simplifié

À chaque instant $t = 1, \dots, T$,

(1) Le joueur propose $\widehat{p}_t = (\widehat{p}_{1,t}, \dots, \widehat{p}_{d,t}) \in \Delta_d$

(2) La nature choisit un vecteur de pertes
$\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0,1]^d$

(3) Le joueur subit alors la perte linéaire $\widehat{\ell}_t = \widehat{p}_t^\top \ell_t$.

Objectif simplifié

Contrôler uniformément en $\ell_1^T = (\ell_1, \dots, \ell_t)$

$$R_T = \sum_{t=1}^{T} \widehat{p}^\top \ell_t - \min_{q \in \Delta_d} \sum_{t=1}^{T} q^\top \ell_t$$

II. Le cas d'un environnement stationnaire.

Algorithme de pondération par poids exponentiels

Paramètres : $\eta > 0$, vitesse d'apprentissage

Initialisation : $\widehat{p}_1 = (1/d, \ldots, 1/d)$

Pour chaque instant $t = 1, \ldots, T$,

(1) Prévoir $\widehat{p}_t$;

(2) Observer les pertes $\ell_t \in [0,1]^d$;

(3) [Mise à jour des poids selon les pertes] $\widehat{p}_{t+1}$ tq

$$\widehat{p}_{j,t+1} = \frac{\widehat{p}_{j,t}\, e^{-\eta\, \ell_{j,t}}}{\sum_{i=1}^{d} \widehat{p}_{i,t}\, e^{-\eta\, \ell_{i,t}}} \,.$$

Borne de regret : environnement stationnaire

Théorème

Pour η bien choisi en fonction T, pour toute suite $\ell_1, \dots, \ell_T$,

$$R_T = \widehat{L}_T - \min_{q \in \Delta_d} \sum_{t=1}^{T} q^\top \ell_t \leqslant \square \sqrt{T \ln d}$$

C'est optimal !

Idée de démonstration.

- Majoration des pertes instantanées $\widehat{\ell}_t = \widehat{p}_t^\top \ell_t$.
- Sommation sur t et télescopage.

□

Preuve

Lemme (Hoeffding)

Si X est une variable aléatoire avec $a \leqslant X \leqslant b$. Alors pour tout $s \in \mathbb{R}$,

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

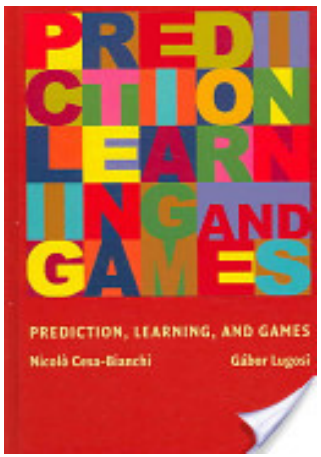
1. Majoration des pertes instantanées par Hoeffding.

$$\begin{aligned}
\widehat{\ell}_t &= \widehat{p}_t^\top \ell_t \leqslant -\frac{1}{\eta} \ln \left(\sum_{j=1}^{d} \widehat{p}_{j,t} e^{-\eta \ell_{j,t}} \right) + \frac{\eta}{8} \\
&= -\frac{1}{\eta} \ln \left(\frac{p_{i,t}}{p_{i,t+1}} e^{-\eta \ell_{i,t}} \right) + \frac{\eta}{8} \\
&= \ell_{i,t} + \frac{1}{\eta} \ln \frac{p_{i,t+1}}{p_{i,t}} + \frac{\eta}{8}
\end{aligned}$$

2. Sommation et télescopage.

$$\sum_{t=1}^{T} \widehat{\ell}_t - \ell_{i,t} \leqslant \frac{1}{\eta} \ln \frac{\cancel{p_{i,T+1}}}{p_{i,1}} + \frac{\eta T}{8} .$$

II. Le cas d'un environnement changeant.

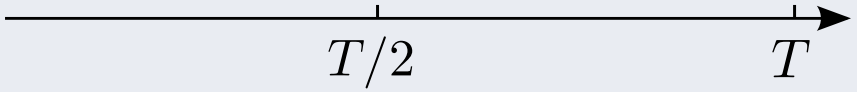


Le cas d'un environnement changeant

$$\widehat{L}_T = \underbrace{\min_{i=1,\dots,d} L_{i,T}}_{\text{Erreur d'approximation}} + \underbrace{R_T}_{\text{Erreur d'estimation}}$$

Exemple où $\min_i L_{i,T}$ est très mauvais.

$$\begin{aligned} \ell_{1,t} &: 1\,1\,1\,\dots 1\,1\,1\,0\,0\,0\,\dots 0\,0\,0 \\ \ell_{2,t} &: 0\,0\,0\,\dots 0\,0\,0\,1\,1\,1\,\dots 1\,1\,1 \end{aligned}$$



Remarque

On peut réduire fortement l'erreur d'approximation de $\widehat{L}_T$ en considérant plutôt :

$$\inf_{\substack{q_1,\dots,q_T \\ \text{tq}\dots}} \sum_{t=1}^{T} \ell(q_t^\top f_t, y_t)$$

Le cas d'un environnement changeant

Rappel : À l'instant t, x_t sont les prévisions des experts, $\widehat{p}_t \in \Delta_d$ est le vecteur de poids choisi, $\ell(\widehat{p}_t^\top x_t, y_t)$ est notre perte instantanée. $\widehat{L}_T$ est notre perte cumulée et $L_{i,T}$ celle de l'expert i.

$$\widehat{L}_T \quad = \quad \inf_{\substack{q_1,\dots,q_T \\ \text{tq}\dots}} \quad \sum_{t=1}^{T} \ell(q_t^\top f_t, y_t) \quad + \quad R_T$$

L'objectif serait toujours $R_T = o(T)$.

Ce n'est réalisable que si une condition est imposée sur $q_1, \dots q_t$.

L'ensemble des suites possibles ne doit pas être trop riche.

Conditions de régularité

Rappel : À l'instant t, x_t sont les prévisions des experts, $\widehat{p}_t \in \Delta_d$ est le vecteur de poids choisi, $\ell(\widehat{p}_t^\top x_t, y_t)$ est notre perte instantanée. $\widehat{L}_T$ est notre perte cumulée et $L_{i,T}$ celle de l'expert i.

$$\widehat{L}_T \quad = \quad \inf_{\substack{q_1,\dots,q_T \\ \text{tq}\dots}} \quad \sum_{t=1}^{T} \ell(q_t^\top f_t, y_t) \quad + \quad R_T$$

Conditions de régularité sur la suite $q_1, \dots, q_T$

- **Hard shifts** : $\#\left\{t \,\middle|\, q_t \neq q_{t-1}\right\} \leqslant m_0$
- **Soft shifts** : $\sum_{t=2}^{T} d_{TV}(q_t, q_{t-1}) \leqslant m_0$.

où $d_{TV}(x,y) = \sum_{i=1}^{d} (x_i - y_i)_+$.

Objectif simplifié après le gradient trick

Rappel : À l'instant t, $\widehat{p}_t \in \Delta_d$ est le vecteur de poids choisi, $\widehat{\ell}_t = \widehat{p}_t^\top \ell_t$ est notre perte instantanée et ℓ_t celles des experts.

Contrôler uniformément en $\ell_1, \ldots, \ell_T$,

$$R_T \quad = \quad \widehat{L}_T \quad - \quad \inf_{\substack{q_1, \ldots, q_T \\ \text{tq} \ldots}} \quad \sum_{t=1}^{T} q_t^\top \ell_t$$

Dans notre cas, la condition sur $q_1, \ldots, q_t$ sera par exemple :

$$\sum_{t=1}^{T} d_{TV}(q_t, q_{t-1}) \leqslant m_0$$

Algorithme Fixed-Share, Herbster & Warmuth 1998

Paramètres : $\eta > 0$, vitesse d'apprentissage et $0 < \alpha < 1$, proportion de mélange

Initialisation : $\widehat{p}_1 = v_1 = (1/d, \ldots, 1/d)$

Pour chaque instant $t = 1, \ldots, T$,

(1) Prévoir $\widehat{p}_t$;

(2) Observer les pertes $\ell_t \in [0,1]^d$;

(3) [Mise à jour des poids selon les pertes] v_{t+1} tq

$$v_{j,t+1} = \frac{\widehat{p}_{j,t}\, e^{-\eta\, \ell_{j,t}}}{\sum_{i=1}^{d} \widehat{p}_{i,t}\, e^{-\eta\, \ell_{i,t}}}$$

(4) [Mise à jour de partage] $\widehat{p}_{t+1} = (1-\alpha) v_{t+1} + \alpha \widehat{p}_1$.

Borne de regret : forme générale

Pour une suite $u_1^T = u_1, \dots, u_T \in \mathbb{R}_+^d$, on note

$$m(u_1^T) = \sum_{t=2}^{T} d_{TV}(u_t, u_{t-1}) \,.$$

Théorème

Pour tout horizon $T \geqslant 1$, toute suite de pertes $\ell_1, \dots, \ell_T \in [0,1]^d$ et toute suite $u_1^T = u_1, \dots, u_T \in \mathbb{R}_+^d$,

$$\sum_{t=1}^{T} \|u_t\|_1 \, \widehat{p}_t^{\top} \ell_t - \sum_{t=1}^{T} u_t^{\top} \ell_t \leqslant \frac{\|u_1\|_1 \ln d}{\eta} + \frac{\eta}{8} \sum_{t=1}^{T} \|u_t\|_1 + \frac{m(u_1^T)}{\eta} \ln \frac{d}{\alpha} + \frac{\sum_{t=2}^{T} \|u_t\|_1 - m(u_1^T)}{\eta} \ln \frac{1}{1-\alpha} \,.$$

Borne de regret : comparaison à $q_1, \ldots, q_t$ variant peu

On fixe m_0. On se restreint aux $u_t = q_t \in \Delta_d$ tels que $m(q_1^T) \leqslant m_0$.

Théorème

Pour η et α bien choisis en fonction de m_0 et T, pour tout $\ell_1, \ldots, \ell_T$,

$$R_T = \widehat{L}_T - \inf_{q_1^T,\ m(q_1^T) \leqslant m_0} \sum_{t=1}^{T} q_t^\top \ell_t \leqslant \square \sqrt{m_0 T \ln\left(\frac{dT}{m_0}\right)}$$

En particulier, $R_T \ll T$ dès que $m_0 \ll T$.

Preuve du théorème

En deux temps :

- borne sur les pertes instantanées $\widehat{p}_t^\top \ell_t$;
- sommation de ces bornes et télescopage.

Lemma (Borne sur le regret instantané)

Pour tout $t \geqslant 1$ et tout vecteur convexe $q_t \in \Delta_d$, l'Algorithme (1) satisfait

$$\widehat{p}_t^\top \ell_t \leqslant q_t^\top \ell_t + \frac{1}{\eta} \sum_{i=1}^{d} q_{i,t} \ln \frac{v_{i,t+1}}{\widehat{p}_{i,t}} + \frac{\eta}{8} .$$

III. Application à la prévision de consommation électrique.

Application à la prévision de consommation électrique

Qu'est-ce que la consommation électrique ?

- C'est un processus temporel
- Dépend de nombreuses variables contextuelles :
  - météo : température, nébulosité, vent, ...
  - calendaires : type de jour, position dans l'année, ...
  - temporelles : consommation de la veille, ...

La prévision de consommation électrique à EDF

Objectif

Prévision de la consommation électrique à court terme.

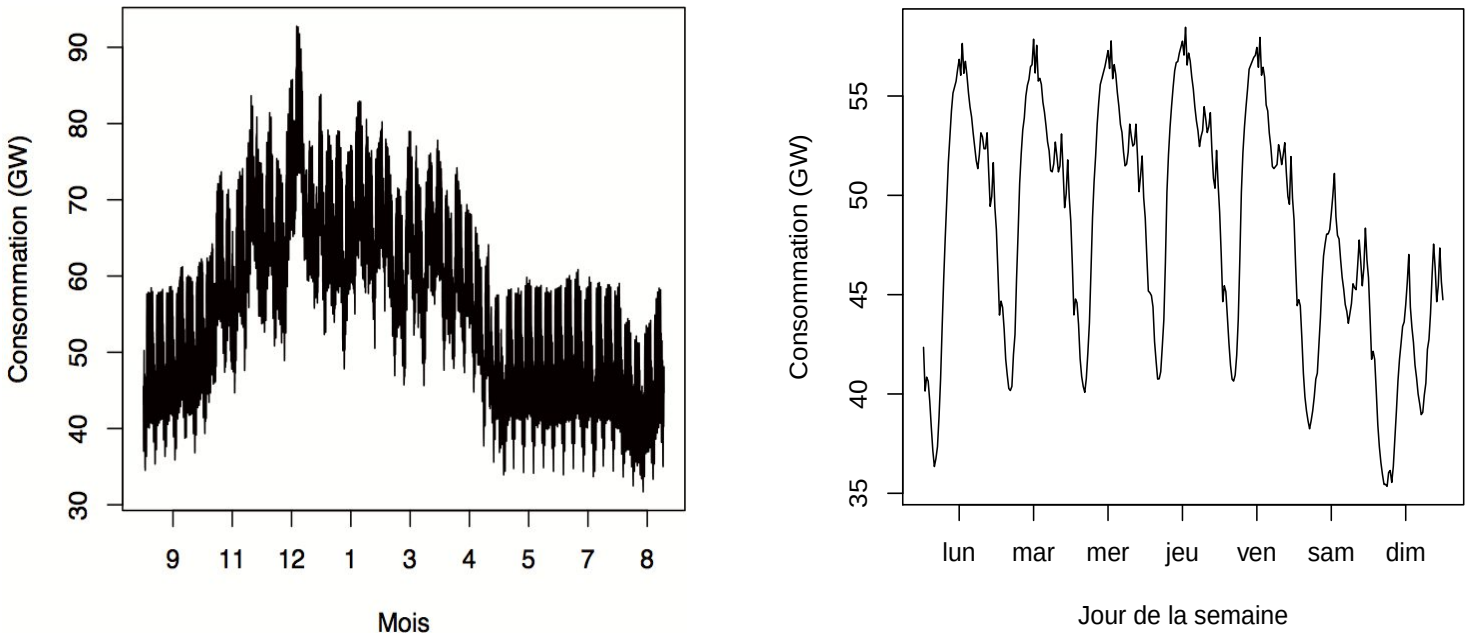


Figure : La consommation électrique en France au cours d'une année et d'une semaine.

Les experts

Plusieurs modèles de prévision existent déjà (merci au groupe R39), ils ont permis de créer 24 experts :

- paramétriques : Eventail
- semi-paramétriques : modèle GAM
- non paramétrique : modèle fonctionnel de similarités

Sleeping. Ils sont spécialisés (jours fériés, grand froid,...). Chaque jour tous les experts ne proposent pas des prévisions. Certain sont en sleeping.

À qui faire confiance ?

Le sleeping. Des experts spécialisés.

Comment adapter les méthodes de mélange aux sleeping ?

- les poids ne sont mis à jour que pour les experts actifs :

$$v_{j,t+1} = \begin{cases} \widehat{p}_{j,t} & \text{si } j \text{ est inactif au tour } t. \\ \widehat{p}_{j,t} e^{-\eta(\ell_{j,t} - \widehat{\ell}_t)} & \text{si } j \text{ est actif.} \end{cases}$$

$$\widehat{p}_{j,t+1} = v_{j,t+1} / \sum_{i=1}^{d} v_{i,t+1}$$

- On prévoit le mélange sur les experts actifs seulement :

$$\widehat{y}_t = \sum_{i=1}^{d} \widehat{p}_{i,t} y_{i,t} \mathbb{1}_{\{i \text{ est actif}\}} / \sum_{i=1}^{d} \widehat{p}_{i,t} \mathbb{1}_{\{i \text{ est actif}\}} .$$

Toute la théorie précédente s'adapte.

Adaptation opérationnelle

Problème pratique : retour d'information sur la consommation réelle qu'une fois par jour (tous les h instants).

Objectif : Prédire simultanément les consommations des h prochains instants (la prochaine journée).

Méthode : ne faire évoluer dans les poids attribués à chaque expert, que ce qui ne dépend pas de leur perte et de la consommation réelle.

Les bornes théoriques sont conservées !

Calibration automatique des paramètres d'apprentissage

Objectif : adapter automatiquement et en ligne les paramètres des algorithmes précédents

Méthode : considérer une grille de paramètres potentiels que l'on construit au fur et à mesure :

- `Initialiser` : $\Lambda =$ valeur théorique optimale du paramètre
- À chaque instant t
  - Choisir dans Λ le paramètre donnant pour l'instant la meilleure performance
  - Si il est sur un bord, agrandir Λ de façon exponentielle

Quelle perte considérer ?

- La perte carrée définie pour tout $x, y \in \mathbb{R}_+$ par

$$\ell(x, y) = (x - y)^2 .$$

Les performances d'un algorithme $\mathcal{A}$ sont alors mesurées par

$$\mathrm{rmse}_T(\mathcal{A}) = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\widehat{y}_t - y)^2} .$$

- La perte absolue définie pour tout $x, y \in \mathbb{R}_+$ par

$$\ell(x, y) = |x - y| .$$

- Le pourcentage d'erreur absolue défini pour tout $x, y \in \mathbb{R}_+$ par

$$\ell(x, y) = \frac{|x - y|}{y} .$$

Le jeu de données

Nombre de jours D	320
Intervalles de temps	30 minutes
Nombre d'instants T	15 360 $(= 320 \times 48)$
Nombre d'experts N	24 $(= 15 + 8 + 1)$
Unité	GW
Médiane des y_t	56.33
Borne B sur les y_t	92.76

Table : Quelques caractéristiques des observations y_t (consommations demi-horaire) du jeu de données considéré.

Les experts

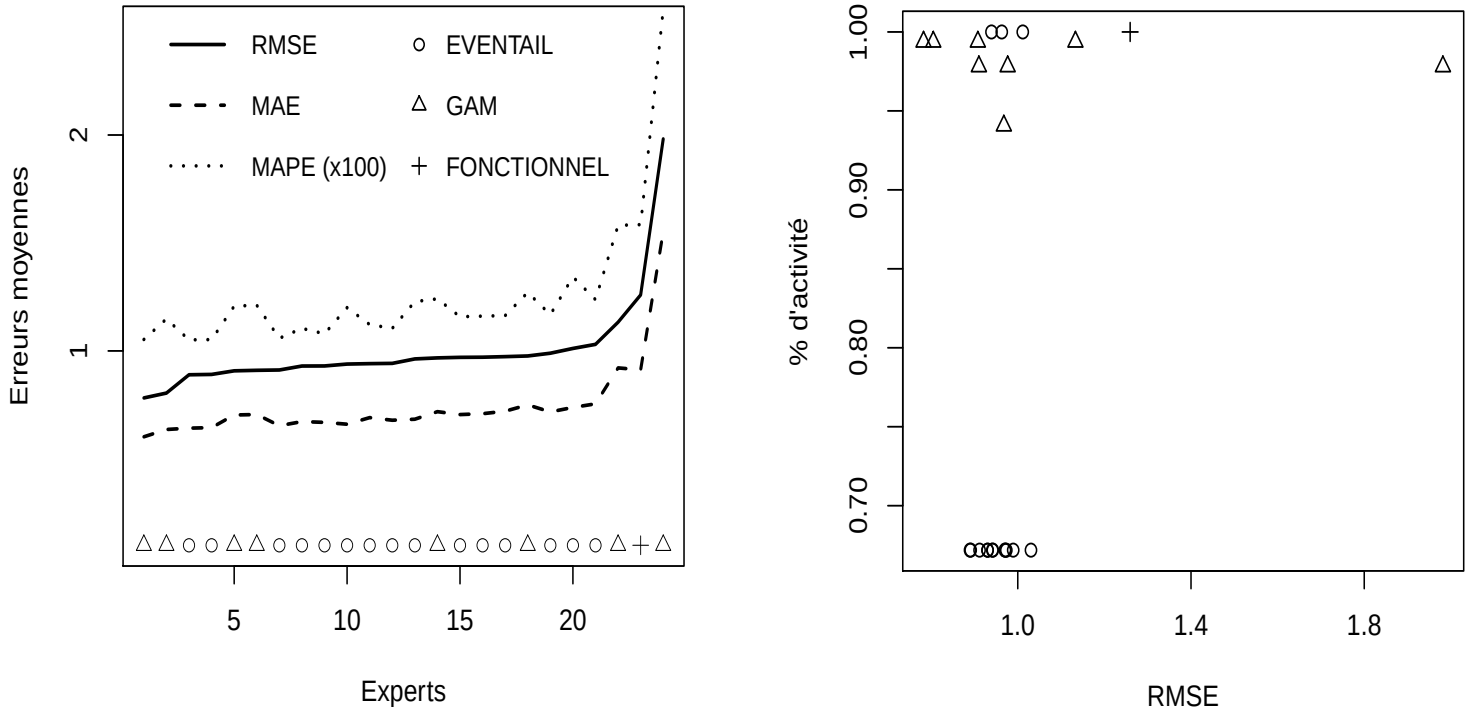


Figure : [Figure gauche] Erreurs moyennes et rmses des experts (axe-y) triés selon leur rmses (axe-x). [Figure droite] Frequences d'activité des experts (axe-y) selon leur rmses (axe-x).

Performances de références

Stratégies et oracles de référence	rmse (MW)
Meilleur expert	782
Meilleure combinaison convexe	658
Meilleure combinaison linéaire	625
Mélange uniforme	724
Meilleur expert composé $m = 13$	629
$m = 50$	534
$m = T - 1 = 15\,359$	223

Les algorithmes de mélange

Deux algorithmes de mélange :

- $\mathcal{E}_\eta$ – Mélange par poids exponentiels.
- $\mathcal{F}_{\eta,\alpha}$ – Fixed-Share.

On peut chacun les appliquer

- directement aux pertes subies par les experts :
  $\ell_{i,t} = \ell(y_{i,t}, y_t) = (y_{i,t} - y_t)^2$.
- aux pertes-gradient : $\Psi_t(p) = \ell(p_t^\top x_t, y_t)$.
  $\tilde{\ell}_{i,t} = \nabla\Psi_t(\widehat{p}_t)^\top(\widehat{p}_t - q) = 2(\widehat{y}_t - y_t)(\widehat{p}_t - q)$.
  Dans ce cas, on note les algorithmes respectivement $\tilde{\mathcal{E}}_\eta$ et $\tilde{\mathcal{F}}_{\eta,\alpha}$

Performances des algorithmes de mélange pour des valeurs fixées des paramètres

Algorithme de mélange	rmse (MW)	Gains (MW)
$\mathcal{E}_\eta$	718	64
$\mathcal{F}_{\eta,\alpha}$	632	150
$\tilde{\mathcal{E}}_\eta$	629	29
$\tilde{\mathcal{F}}_{\eta,\alpha}$	599	59

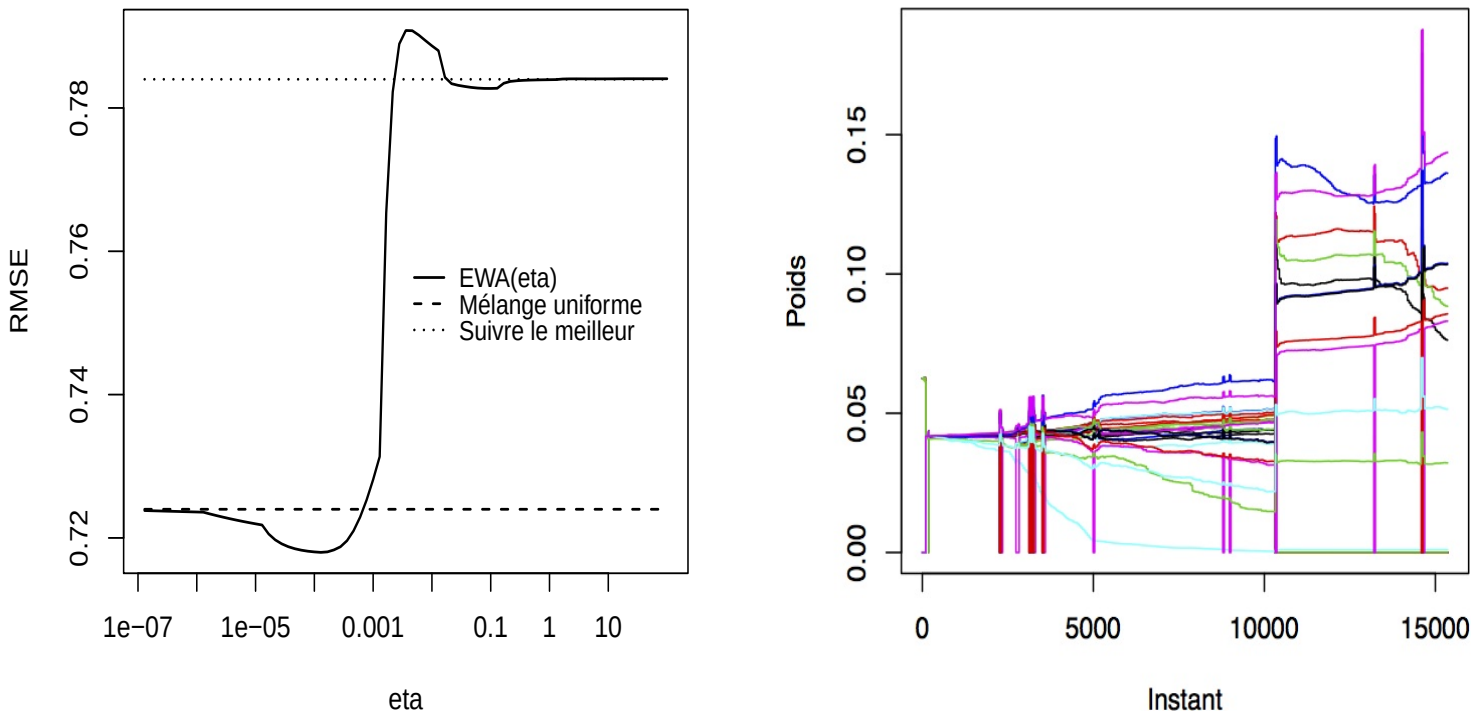
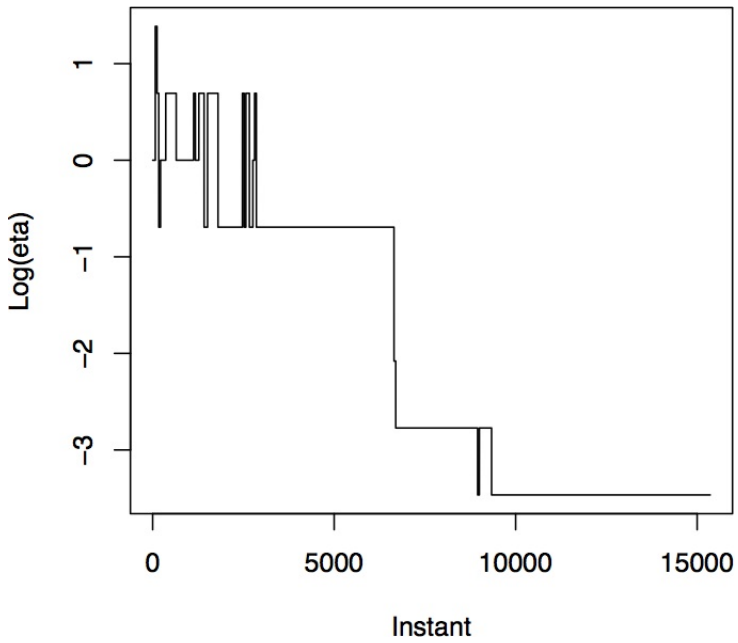


Figure : [Figure gauche] Performance de l'algorithme $\mathcal{E}_\eta$ en fonction de la valeur de son paramètre d'apprentissage η. [Figure droite] Évolution des poids accordés à chaque expert par l'algorithme $\mathcal{E}_\eta$.

Performances des algorithmes adaptatifs

Algorithme de mélange	rmse (MW)
$\mathcal{E}_\eta$	724
$\tilde{\mathcal{E}}_\eta$	637
$\mathcal{F}_{\eta,\alpha}$	639
$\tilde{\mathcal{F}}_{\eta,\alpha}$	623



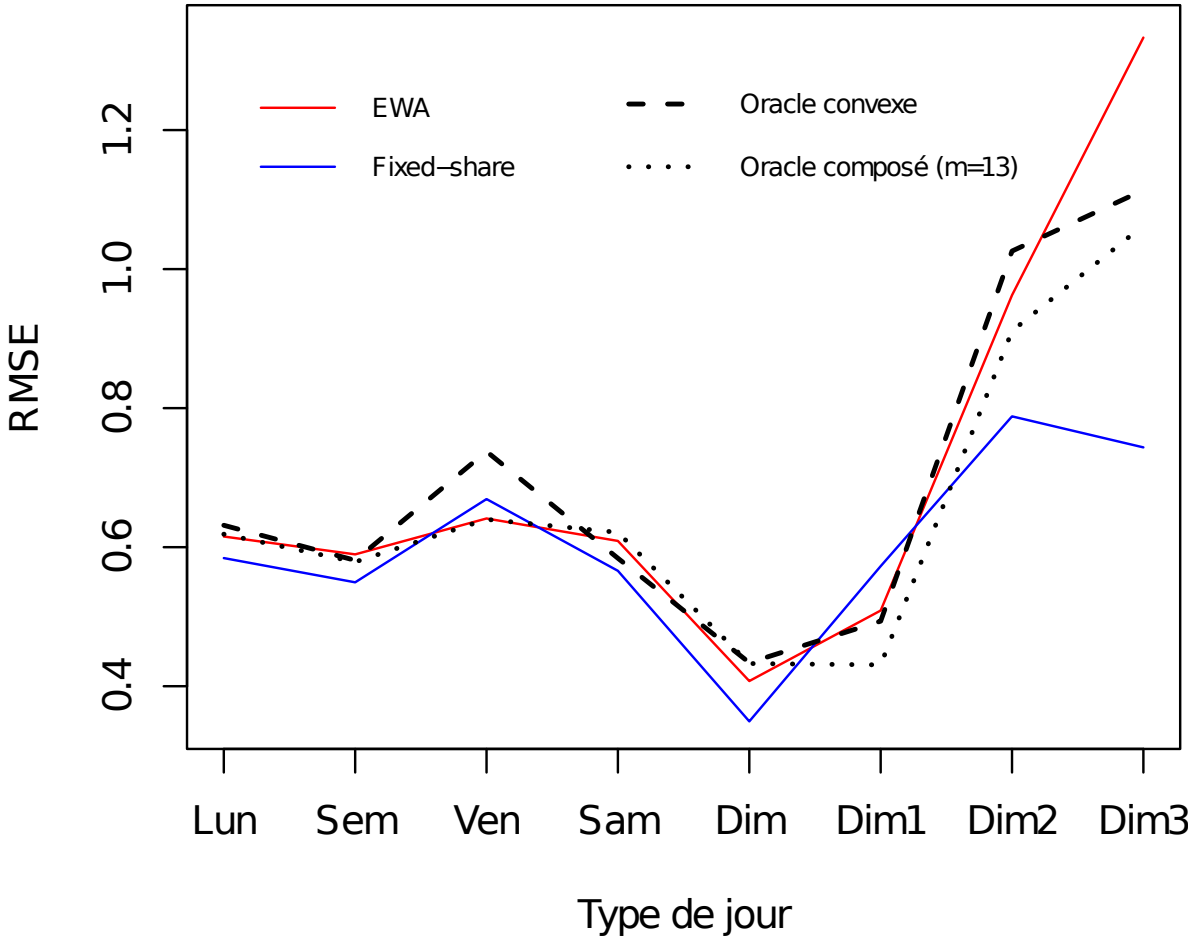


Figure : Erreur moyenne des différents algorithmes et de deux oracles selon le type de jour.

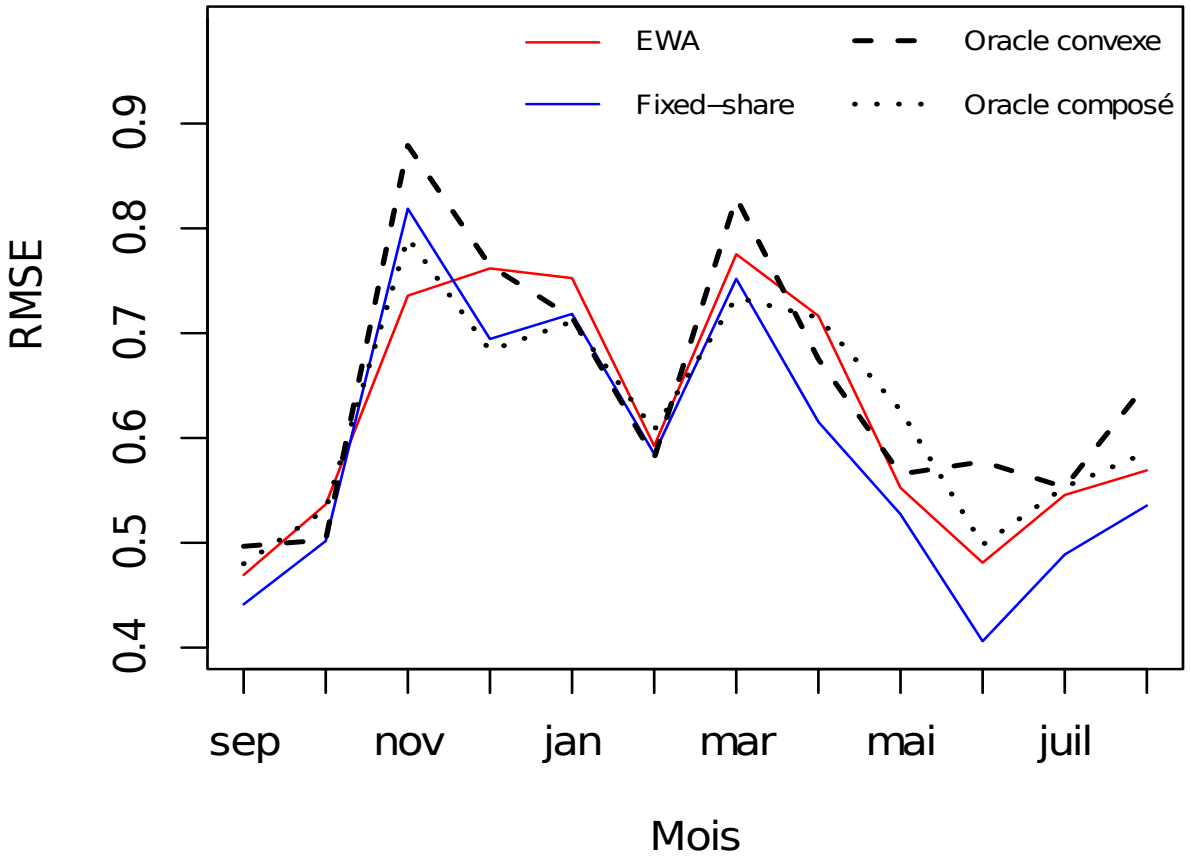


Figure : Erreur moyenne des différents algorithmes et de deux oracles en fonction du mois de l'année.

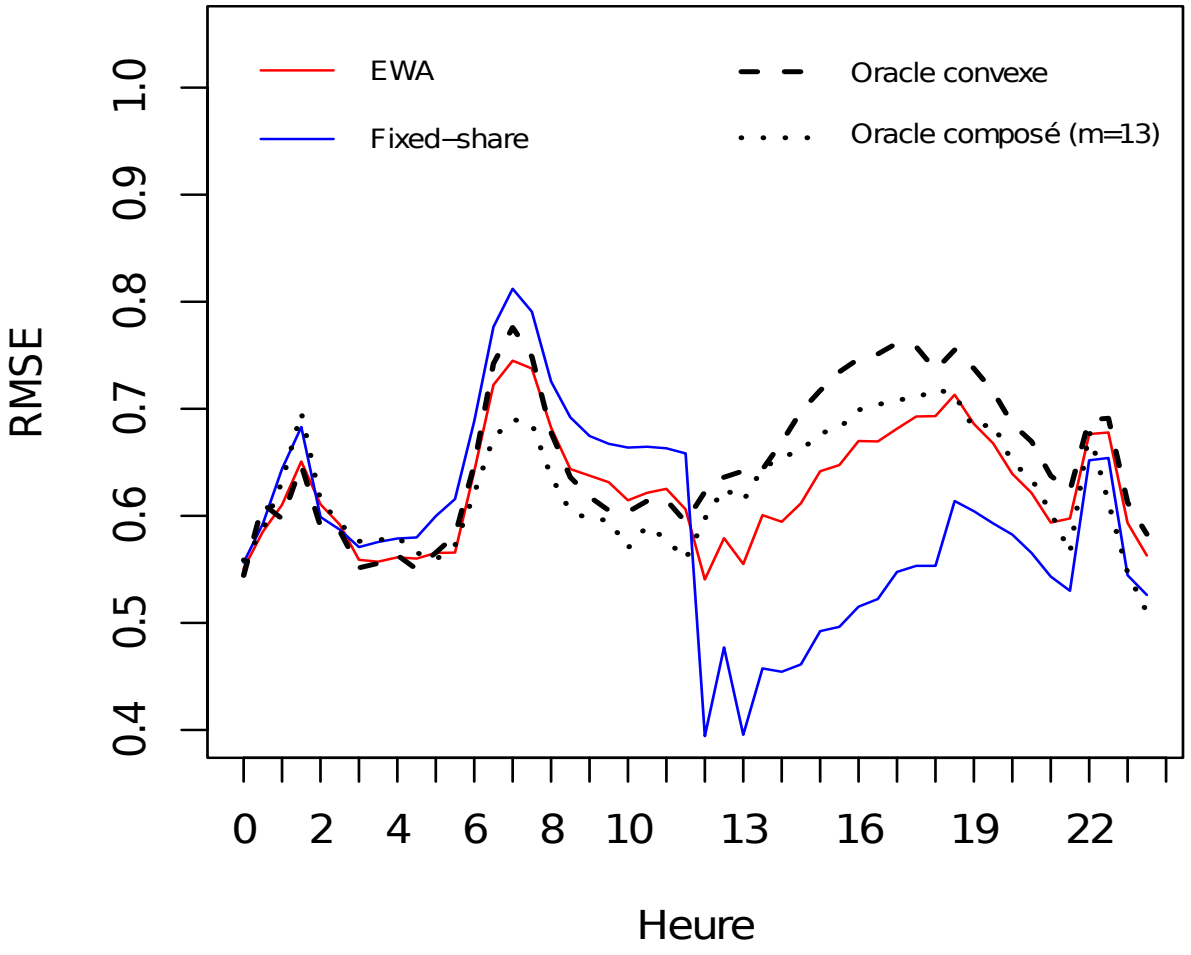


Figure : Erreur moyenne des différents algorithmes et de deux oracles en fonction de l'heure de la journée.