

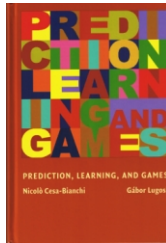
Prédiction séquentielle de suites individuelles

Gilles Stoltz

CNRS – École normale supérieure – HEC Paris



Au début de cet exposé, je voudrais saluer



Prediction, Learning, and Games de Nicolò Cesa-Bianchi et Gábor Lugosi



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Un statisticien accepte la mission de prédire une suite y_1, y_2, \dots d'observations vivant dans un ensemble \mathcal{Y} .

Ses prédictions $\hat{p}_1, \hat{p}_2, \dots$ sont formées dans un ensemble \mathcal{X} .

Les observations et prédictions (1) sont effectuées de manière **séquentielle** et (2) ne reposent sur **aucun modèle stochastique**.

(1) signifie qu'à chaque échéance, y_t est prédite sur le seul fondement du passé, $y_1^{t-1} = (y_1, \dots, y_{t-1})$.

La prédiction \hat{p}_t est formée avant que la vraie valeur y_t ne soit révélée au grand jour, et les deux sont ensuite comparées.

(2) explique pourquoi je prononcerai à peine le mot "probabilité".

Cependant, certaines techniques employées vous rappelleront des méthodes de statistique classique.



Pour que le problème ait un sens dans un cadre aussi général, on introduit des **experts**.



Pour que le problème ait un sens dans un cadre aussi général, on introduit des **experts**.



(Ce ne seront pas les mêmes experts que Gérard Biau !)

Pour que le problème ait un sens dans un cadre aussi général, on introduit des **experts** ; il leur sera fait référence par $j = 1, \dots, N$.

A chaque échéance, l'expert j procure une prédiction $f_{j,t} = f_{j,t}(y_1^{t-1}) \in \mathcal{X}$.

Le statisticien fonde maintenant ses prédictions \hat{p}_t sur les **observations passées** $y_1^{t-1} = (y_1, \dots, y_{t-1})$ et sur les **conseils passés et présents** des experts, $f_{j,s}$ pour $s = 1, \dots, t$.

L'objectif du statisticien est de prédire presque aussi bien que le meilleur expert.

Remarquez cependant que le meilleur expert ne saurait être déterminé que **rétrospectivement** tandis que le statisticien est assujéti à une contrainte de prédiction **séquentielle**.



On se limite à former des combinaisons **linéaires** ou **convexes** des conseils des experts (convexes, pour commencer).

On suppose donc que \mathcal{X} est convexe.

A chaque échéance, le statisticien forme une combinaison convexe des conseils des experts,

$$\hat{p}_t = \sum_{j=1}^N p_{j,t} f_{j,t}$$

où $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ est un élément du simplexe (de probabilité) d'ordre N .

On renforce l'objectif : le statisticien veut en fait prédire presque'aussi bien que la **meilleure combinaison convexe constante** des experts.



Pour quantifier mathématiquement la notion de meilleur expert, on introduit une **fonction de perte** $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$.

On définit les pertes cumulées du statisticien et des combinaisons convexes constantes $\mathbf{q} = (q_1, \dots, q_N)$ des experts par

$$\widehat{L}_n = \sum_{t=1}^n \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) \quad \text{et} \quad L_n(\mathbf{q}) = \sum_{t=1}^n \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right)$$

Le **regret** face à \mathbf{q} est la différence entre ces deux quantités,

$$R_n(\mathbf{q}) = \widehat{L}_n - L_n(\mathbf{q})$$

On veut construire des stratégies de prédiction telles que le **regret moyen** converge vers 0 pour toute suite d'observations y_1, y_2, \dots , uniformément en les \mathbf{q} , soit

$$\limsup \frac{1}{n} \max_{\mathbf{q}} R_n(\mathbf{q}) \leq 0$$



On résumé le cadre mathématique introduit par un **jeu répété**.

Paramètres : un ensemble convexe \mathcal{X} de prédictions, un ensemble \mathcal{Y} d'observations, une fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

A chaque échéance $t = 1, 2, \dots$,

- les experts procurent leurs conseils $f_{j,t}$
- le statisticien choisit une combinaison convexe $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ et prédit $\hat{p}_t = \sum_{j=1}^N p_{j,t} f_{j,t}$
- l'environnement choisit simultanément l'observation y_t
- y_t et \hat{p}_t (et donc les pertes respectives) sont révélées.

On veut minimiser le regret face à tous les \mathbf{q} ,

$$R_n(\mathbf{q}) = \hat{L}_n - L_n(\mathbf{q}) = \sum_{t=1}^n \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) - \sum_{t=1}^n \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right)$$



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



On considère ici un cadre **méta-statistique**.

Des modélisations stochastiques des observations, les procédures statistiques en dérivant (et le choix de certains paramètres pour ces dernières) conduisent à différents experts.

On dispose alors en pratique de nombreux experts et on est confronté à un problème de **sélection**.

Il se trouve que l'**agrégation** obtient souvent de meilleures performances.

Ici, le problème est séquentiel par nature et il s'agit de réaliser une **agrégation séquentielle** des experts.

De plus, le critère d'évaluation retenu (le regret par rapport à toutes les suites d'observations possibles) requiert que les stratégies soient **robustes**.



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Des exemples d'application des suites individuelles sont
l'**investissement dans le marché boursier**, la prédiction de pics
d'ozone, la prédiction de consommation électrique.



Pour le problème de l'investissement, on considère m valeurs boursières. Une **prédiction** est une allocation $P = (P^1, \dots, P^m)$ des capitaux entre les m valeurs (donnée par un élément du **simplexe** \mathcal{X} d'ordre m).

Les **experts** sont des stratégies élémentaires d'investissement (on peut penser à N courtiers nous recommandant chaque jour une allocation respective $P_{j,t}$).

Les **observations** sont données par les facteurs multiplicatifs d'évolution y_t^i de chaque valeur i entre ses prix à l'ouverture et à la fermeture de la bourse au jour t ; alors, $y_t = (y_t^1, \dots, y_t^m)$ et $\mathcal{Y} = (\mathbb{R}_+)^N$.

Au jour t , l'allocation P voit ses capitaux multipliés par le facteur

$$P \cdot y_t = \sum_{i=1}^m P^i y_t^i$$



Résumé du cadre :

L'investissement séquentiel dans m valeurs boursières est formalisé en considérant des prédictions dans le simplexe d'ordre m , en recourant à N courtiers procurant des recommandations $P_{j,t}$, et en ayant affaire à des observations $y_t = (y_t^1, \dots, y_t^m)$ décrivant l'évolution du marché.

Pour rendre les choses additives et tenir compte du fait qu'on obtient ici des paiements plutôt que des pertes, on définit la **fonction de perte**

$$\ell(P, y_t) = -\ln P \cdot y_t$$

On veut construire de manière séquentielle des allocations $\hat{P}_1, \hat{P}_2, \dots$ telles que

$$\ln \prod_{t=1}^n \hat{P}_t \cdot y_t \quad \text{soit proche de} \quad \max_{\mathbf{q}} \ln \prod_{t=1}^n \left(\sum_{j=1}^N q_j P_{j,t} \right) \cdot y_t$$

La différence (le **regret**) peut être un $o(n)$, même un $O(m \ln n)$.



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Le jeu de données typique a été introduit ici par Cover ('91) et consiste en 22 années (du début des années 60 jusqu'au milieu des années 80) de facteurs d'évolution quotidiens de 36 valeurs boursières.

Les **experts** considérés dans les études empiriques sont toujours, à ma connaissance, les m **valeurs boursières** sous-jacentes.

Ces experts n'étant pas très sophistiqués, il n'est pas surprenant que les techniques d'agrégation par suites individuelles ne mènent pas à des performances formidables...

... mais plutôt en une multiplication par 20 des capitaux misés, ce qui est facile à obtenir en vue de l'inflation galopante et du biais de survie des valeurs boursières.



Le jeu de données typique a été introduit ici par Cover ('91) et consiste en 22 années (du début des années 60 jusqu'au milieu des années 80) de facteurs d'évolution quotidiens de 36 valeurs boursières.

Les **experts** considérés dans les études empiriques sont toujours, à ma connaissance, les m **valeurs boursières** sous-jacentes.

Ces experts n'étant pas très sophistiqués, il n'est pas surprenant que les techniques d'agrégation par suites individuelles ne mènent pas à des performances formidables...

Pire, la plupart du temps, les performances de techniques d'agrégation elles sophistiquées sont **très proches** de la stratégie naïve qui consiste à rééquilibrer chaque matin ses capitaux vers une distribution uniforme entre les valeurs boursières.



Ces performances pratiques décevantes illustrent que le **choix des experts** est tout à fait **crucial**.

Ma suggestion serait d'essayer l'agrégation de m **vraies stratégies d'investissement** fondamentales, avec les techniques générales que je vais présenter dans leurs détails mathématiques dans la troisième partie de cet exposé.

Si l'un(e) d'entre vous est intéressé(e), qu'il (elle) me fasse signe !



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Voici le fruit d'une collaboration avec **Vivien Mallet (INRIA)**, publiée par *Journal of Geophysical Research*.

On veut prédire, jour après jour, les hauteurs des pics d'ozone du lendemain (ou les concentrations horaires, heure après heure).

On dispose d'un réseau de stations météorologiques à travers l'Europe pour les relever (tout se passe pendant l'été 2001).

On construit tout d'abord **48 prédicteurs fondamentaux** en choisissant pour **chacun** d'entre eux **un modèle**, défini par une formulation physico-chimique (parmi plusieurs possibles), un schéma numérique (parmi plusieurs possibles) de résolution approchée des EDPs en jeu, et un jeu de données d'entrée.



Voici le fruit d'une collaboration avec **Vivien Mallet (INRIA)**, publiée par *Journal of Geophysical Research*.

On veut prédire, jour après jour, les hauteurs des pics d'ozone du lendemain (ou les concentrations horaires, heure après heure).

On dispose d'un réseau de stations météorologiques à travers l'Europe pour les relever (tout se passe pendant l'été 2001).

On construit tout d'abord **48 prédicteurs fondamentaux** en choisissant pour **chacun** d'entre eux **un modèle**.

Au lieu de devoir se fier à un prédicteur plutôt qu'un autre en le **sélectionnant**, on recourt à une procédure plus gloutonne qui les considère tous et les **agrège** séquentiellement.



On dispose d'un **réseau** \mathcal{S} de stations à travers l'Europe et chaque modèle $j = 1, \dots, 48$ procure une prédiction $f_{j,t}^s$ pour le pic à la station s et au jour t , qui est ensuite comparée au pic réalisé y_t^s .

Le statisticien détermine chaque jour une unique combinaison convexe $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ à utiliser en **toutes les stations** pour agréger les prédictions (et obtenir ainsi un champ de prévisions).

Les écarts sont mesurés en perte quadratique moyenne, ce qui revient à considérer la **fonction de perte**

$$\ell(\mathbf{p}_t, (y_t^s)_{s \in \mathcal{S}_t}) = \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} p_{j,t} f_{j,t}^s - y_t^s \right)^2$$

où \mathcal{S}_t est le sous-ensemble des stations actives au jour t .

La définition s'étend au cas des **combinaisons linéaires** \mathbf{u}_t (qui permettent par exemple de réduire le biais des modèles).



Les figures ci-dessous montrent que **tous** les experts sont utiles et apportent de l'information.

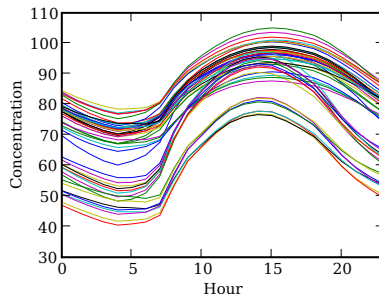
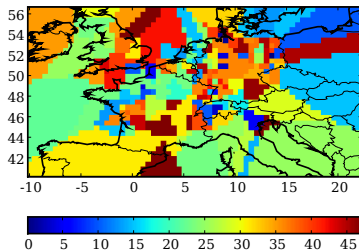


FIG.: **A gauche** : Coloration de l'Europe en fonction de l'indice du meilleur expert local. **A droite** : Profils moyens de prédiction sur une journée (moyennes spatiales et temporelles, en $\mu\text{g}/\text{m}^3$).

Les **erreurs cumulées** de la méthode d'agrégation et de la combinaison linéaire constante induite par \mathbf{u} valent respectivement

$$\widehat{L}_n = \sum_{t=1}^n \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} u_{j,t} f_{j,t}^s - y_t^s \right)^2$$

$$\text{et } L_n(\mathbf{u}) = \sum_{t=1}^n \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} u_j f_{j,t}^s - y_t^s \right)^2$$

où \mathcal{S}_t est le sous-ensemble des stations actives au jour t .

Les **erreurs quadratiques moyennes** associées sont données par

$$\widehat{r}_n = \sqrt{\frac{\widehat{L}_n}{\sum_{t=1}^n |\mathcal{S}_t|}} \quad \text{et} \quad r_n(\mathbf{u}) = \sqrt{\frac{L_n(\mathbf{u})}{\sum_{t=1}^n |\mathcal{S}_t|}}$$



Moyenne	M. fondamental	M. convexe	M. linéaire	Prescient
24.41	22.43	21.45	19.24	11.99

Ci-dessus, les erreurs quadratiques moyennes (en $\mu\text{g}/\text{m}^3$)

- de la **moyenne** des prédictions des 48 modèles, i.e., $r_n((1/48, \dots, 1/48))$,
- du **meilleur** modèle **fondamental** parmi $j = 1, \dots, 48$,
- de la **meilleure** combinaison **convexe** \mathbf{q} des 48 modèles, i.e., $\min_{\mathbf{q}} r_n(\mathbf{q})$,
- de la **meilleure** combinaison **linéaire** \mathbf{u} (parmi tous les vecteurs de \mathbb{R}^{48}) des 48 modèles, i.e., $\min_{\mathbf{u}} r_n(\mathbf{u})$,
- du prédicteur **prescient** qui aurait connaissance des y_t^s avant de former sa prédiction et ne serait contraint que par l'obligation de choisir une combinaison linéaire des prédictions des modèles.



Nous avons mis en œuvre environ 20 méthodes d'agrégation différentes et nous concentrons ici sur deux familles qui ont obtenu de bons résultats, EG et la régression ridge (et leurs variantes).

EG est l'abréviation d'exponentielle des gradients. Cette méthode forme des combinaisons **convexes** dont les composantes sont données par une pondération exponentielle des sommes des composantes des gradients des pertes passées.

Son regret moyen par rapport à l'ensemble des combinaisons convexes constantes est plus petit que $1/\sqrt{n}$.

La **régression ridge** est une méthode d'estimation classique en perte quadratique et qui utilise la meilleure **combinaison linéaire** pénalisée sur les données passées (pénalisation en terme de norme ℓ^2).

Son regret moyen par rapport à toute combinaison linéaire constante est plus petite qu'une quantité de l'ordre de $(\ln n)/n$.



Les versions **fenêtrées** n'utilisent qu'un nombre fixe des plus récentes pertes passées, pour ensuite pondérer exponentiellement leurs gradients (EG) ou calculer sur elles seulement une meilleure combinaison linéaire pénalisée (régression ridge).

L'**escompte** multiplie chaque perte passée par un facteur d'autant plus petit que ce passé est lointain.

EG	EG fenêtré	EG esc.	Ridge	Ridge fenêtrée	Ridge esc.
21.47	21.37	21.31	20.77	20.03	19.45

La **meilleure** combinaison **convexe** constante est battue et la version escomptée de la régression ridge a des performances très proches de celles de la **meilleure** combinaison **linéaire** constante.

Moyenne	M. fondamental	M. convexe	M. linéaire	Prescient
24.41	22.43	21.45	19.24	11.99



Les méthodes d'agrégation séquentielle ne se concentrent **pas** sur un seul expert.

Les poids attribués aux modèles peuvent changer rapidement et de manière significative au cours du temps.

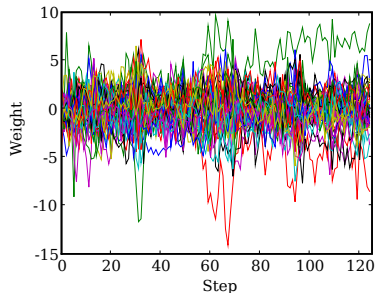
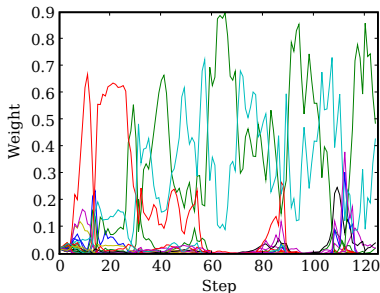


FIG.: Poids produits au cours du temps par (à gauche) EG et la version escomptée de la régression ridge (à droite).

- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Voici une allusion rapide à la thèse de **Yannig Goude (EDF & Université Paris-Sud)**, soutenue en janvier 2008.

Vous disposez du système de prédiction **Eventail**, qui bénéficie de décennies d'expérience, et qui requiert la donnée de différents paramètres (le temps, les périodes de vacances, mais aussi des paramètres d'utilisateur).

Il existe des jeux de paramètres efficaces **la plupart du temps** mais dont les performances sont désastreuses dans des circonstances **atypiques**. Pour ces dernières, d'autres jeux de paramètres sont cependant efficaces.

Le traitement statistique serait de **détecter les ruptures**.

L'approche prise par les suites individuelles est de se comparer non pas seulement à la meilleure combinaison convexe constante, mais à la meilleure suite de combinaisons convexes avec **k changements** au plus dans les n pas de prédiction.



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Soit un ensemble convexe de prédictions \mathcal{X} , l'espace des observations \mathcal{Y} et une fonction de perte convexe $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$.

L'**algorithme EG** choisit successivement des combinaisons convexes $\mathbf{p}_1, \mathbf{p}_2, \dots$ des prédictions des experts et ses pertes sont données par

$$\tilde{\ell}_t(\mathbf{p}_t) = \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right)$$

Chaque \mathbf{p}_t ne dépend que des (**gradients** des) pertes passées,

$$\mathbf{p}_1 = (1/N, \dots, 1/N)$$

et la j -ième composante de \mathbf{p}_t est définie, pour $t \geq 2$, par une pondération **exponentielle**,

$$p_{j,t} = \frac{\exp \left(-\eta \sum_{s=1}^{t-1} \left(\nabla \tilde{\ell}_s(\mathbf{p}_s) \right)_j \right)}{\sum_{i=1}^N \exp \left(-\eta \sum_{s=1}^{t-1} \left(\nabla \tilde{\ell}_s(\mathbf{p}_s) \right)_i \right)}$$



Theorem

Le regret de EG face à toute combinaison convexe constante \mathbf{q} est **uniformément** borné (en \mathbf{q} et en les suites y_1, y_2, \dots) selon

$$\sum_{t=1}^n \ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) - \sum_{t=1}^n \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right) \leq \frac{\ln N}{\eta} + \frac{\eta n}{2} B^2$$

où B est une borne sur les gradients, $\|\nabla \tilde{\ell}_t\|_{\infty} \leq B$ pour tout t .

Deux éléments de démonstration : par **convexité**,

$$\ell \left(\sum_{j=1}^N p_{j,t} f_{j,t}, y_t \right) - \ell \left(\sum_{j=1}^N q_j f_{j,t}, y_t \right) \leq \nabla \tilde{\ell}_t(\mathbf{p}_t) \cdot (\mathbf{p}_t - \mathbf{q})$$

et l'analyse de ce majorant (linéaire en \mathbf{q}) repose sur le lemme de **Hoeffding**.



Il faut **calibrer** η : le regret est plus petit que

$$\frac{\ln N}{\eta} + \frac{\eta n}{2} B^2 = \square B \sqrt{n \ln N}$$

avec le choix (optimal pour la théorie) $\eta = \square (1/B) \sqrt{(\ln N)/n}$.

Mais n et/ou B peuvent être inconnus ; une version **adaptive** de EG est donnée par les mêmes formules où l'on remplace simplement le paramètre fixe η par une suite adaptative (η_t) ,

$$\eta_{t+1} = \frac{1}{\max_{s \leq t} \|\nabla \tilde{\ell}_s\|_\infty} \sqrt{\frac{\ln N}{t}}$$

La borne sur son **regret** est **similaire**.

Bibliographie : EG a été introduit par Vovk '90, Littlestone et Warmuth '94, Cesa-Bianchi '99 et des versions adaptatives ont été étudiées par Auer, Cesa-Bianchi et Gentile '02, Cesa-Bianchi, Mansour et Stoltz '07.



La version **fenêtrée** repose sur une largeur de fenêtre T et produit les combinaisons convexes données par

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1, t-T\}}^{t-1} \left(\nabla \tilde{\ell}_s(\mathbf{p}_s)\right)_j\right)}{\sum_{i=1}^N \exp\left(-\eta \sum_{s=\max\{1, t-T\}}^{t-1} \left(\nabla \tilde{\ell}_s(\mathbf{p}_s)\right)_i\right)}$$

La version **escomptée** utilise une suite décroissante (β_s) pour former

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1 + \beta_{t-s}) \left(\nabla \tilde{\ell}_s(\mathbf{p}_s)\right)_j\right)}{\sum_{i=1}^N \exp\left(-\eta_t \sum_{s=1}^{t-1} (1 + \beta_{t-s}) \left(\nabla \tilde{\ell}_s(\mathbf{p}_s)\right)_i\right)}$$

On peut exhiber une borne théorique sur le regret de la version escomptée, dépendant de (β_s) .

Pour la prédiction de la qualité de l'air, nous avons utilisé des **escomptes assez forts**, $\beta_s = 100/s^2$.



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



En s'inspirant de Cover '91 et Blum et Kalai '97, on peut définir, pour les pertes **exp-concaves**, un mélange uniforme adaptatif sur le simplexe.

S'il existe $\eta > 0$ tel que pour tous y_t et $f_{j,t}$,

$$\mathbf{p} \mapsto \exp(-\eta \ell_t(\mathbf{p}))$$

est **concave**, alors agréger selon

$$\mathbf{p}_t = \frac{\int \mathbf{p} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(\mathbf{p})\right) d\mu(\mathbf{p})}{\int \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(\mathbf{p})\right) d\mu(\mathbf{p})}$$

(où μ est la mesure uniforme sur le simplexe) assure que le regret est plus petit que $\square (N \ln n)/\eta$, ce qui est d'un ordre de grandeur bien plus petit que \sqrt{n} .

Cette hypothèse est vérifiée dans le cas de l'investissement boursier et lorsque ℓ est la perte quadratique.



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



La régression ridge a été introduite dans les années 70 par Hoerl et Kennard et intensivement étudiée depuis dans un cadre stochastique.

Vovk '01 et Azoury et Warmuth '01 en proposent une analyse pour des **suites individuelles**.

Formellement, en perte quadratique, la régression ridge choisit des combinaisons **linéaires** \mathbf{u}_t des prédictions des experts données, à l'échéance $t \geq 2$, par un critère de moindres carrés pénalisés,

$$\mathbf{u}_t = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^N} \left[\lambda \|\mathbf{u}\|_2^2 + \sum_{s=1}^{t-1} \left(y_s - \sum_{j=1}^N u_j f_{j,s} \right)^2 \right]$$

Elle peut être mise en œuvre efficacement de manière séquentielle et assure que son regret est $O(\ln n)$.



Une propriété tout à fait sympathique de la régression ridge est qu'elle semble **débiaiser** automatiquement les experts.

On peut en effet la faire tourner sur un seul expert (proposant les prédictions $f_{j,t}^s$) et faire ainsi presque aussi bien que le meilleur des experts, indexés chacun par γ , proposant les prédictions $\gamma f_{j,t}^s$.

S'il y a un facteur de **biais multiplicatif** à-peu-près constant $1/\gamma$, il est donc corrigé.

Sur les données d'ozone, cela donne les erreurs quadratiques moyennes suivantes, par exemple sur le **meilleur** et le **moins bon** modèle :

Sans Ridge	Avec Ridge	Sans Ridge	Avec Ridge
35.79	24.78	22.43	21.66



- 1 Agrégation séquentielle de prédicteurs
 - Cadre mathématique
 - La philosophie sous-jacente à ce cadre
 - Un premier exemple (théorique) : Investissement boursier
- 2 Applications à des données réelles
 - Investissement boursier
 - Prédiction de la qualité de l'air
 - Prédiction de la consommation électrique
- 3 Deux familles d'algorithmes d'agrégation séquentielle
 - Exponentielle des gradients
 - Une pondération exponentielle sans gradients
 - La régression ridge
- 4 Travaux récents et perspectives



Dans ce qui suit, je vais passer rapidement en revue les travaux que Sébastien Gerchinovitz a effectués ces derniers mois à l'occasion de son stage de M2 avec Vivien Mallet et moi-même.

Ils portaient sur

- une **meilleure calibration** des paramètres d'apprentissage,
- la **prédiction lacunaire** : la sélection séquentielle d'un sous-ensemble de modèles pour la prédiction,
- l'étude de plus **longues durées** et/ou avec **plus d'experts** (prédictions horaires, prédictions sur un an, utilisation de 100 experts).

Le dernier point indiquera l'intérêt du transfert de ces techniques vers les systèmes de prédiction d'EDF.



J'ai fait allusion à la **calibration automatique** de l'exponentielle des gradients.

Mais cette dernière est trop précautionneuse en pratique, même si les bornes théoriques correspondantes ne sont modifiées que d'un facteur multiplicatif et pas dans leurs ordres de grandeur.

On veut ici une méthode **plus efficace**.



On rappelle que l'exponentielle des gradients prédit, pour $t \geq 2$, avec \mathbf{p}_t défini, composante j par composante j selon

$$p_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla \tilde{\ell}_s(\mathbf{p}_s)\right)_j\right)}{\sum_{i=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla \tilde{\ell}_s(\mathbf{p}_s)\right)_i\right)}$$

L'idée ici est de faire **varier** η en fonction de t et considérer pour η_t le meilleur paramètre η sur les échéances $1, \dots, t-0$,

$$\eta_t = \operatorname{argmin}_{\eta > 0} \sum_{s=1}^{t-1} \ell_s(\mathbf{p}_s(\eta)) .$$

On utilise alors $\mathbf{p}_t(\eta_t)$ pour la prédiction au jour t .



On peut définir de manière similaire une calibration automatique de Ridge. Sur les données d'ozone :

Meilleure convexe	21.45
EG avec meilleur η	21.47
EG avec (η_t)	21.80
Meilleure linéaire	19.24
Ridge avec meilleur λ	20.77
Ridge avec (λ_t)	20.81

Le “meilleur” paramètre désigne le paramètre constant η ou λ , choisi de manière **rétrospective**, qui aurait donné les meilleurs résultats en termes d'erreur quadratique.

Il n'y a pas encore de **borne théorique** pour cette méthode de calibration, mais nous y travaillons !



Pour obtenir des combinaisons linéaires ou convexes n'utilisant qu'un nombre restreint de modèles, on peut **seuiller** les combinaisons proposées (pour EG) ou changer le type de **pénalité** (pour Ridge).

La méthode LASSO (Tibshirani, '96) choisit des combinaisons **linéaires** \mathbf{u}_t des prédictions des experts données, à l'échéance $t \geq 2$, par un critère de moindres carrés pénalisés en **norme ℓ^1** ,

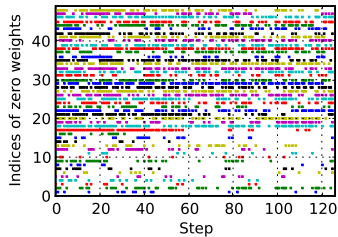
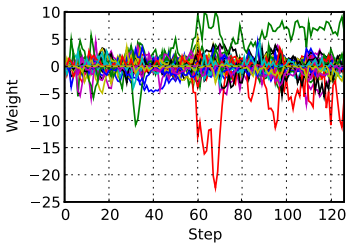
$$\mathbf{u}_t = \underset{\mathbf{u} \in \mathbb{R}^N}{\operatorname{argmin}} \left[\lambda \|\mathbf{u}\|_1 + \sum_{s=1}^{t-1} \left(y_s - \sum_{j=1}^N u_j f_{j,s} \right)^2 \right]$$

Les combinaisons qui en résultent ont en général de nombreux coefficients nuls (et sont dites lacunaires).



Une version escomptée de LASSO conduit ainsi à une très **forte sélection** parmi les modèles (une vingtaine est éliminée sur les données d'ozone).

Ridge esc.	LASSO esc.	M. linéaire
19.45	19.31	19.24



D'autres résultats sont disponibles dans un rapport technique récent (et on peut les évoquer rapidement ensemble lors des questions) :

- Sur un ensemble de **100 modèles** et pendant **un an**, le gain de performance relative des méthodes présentées par rapport à la meilleure combinaison convexe ou linéaire est encore meilleur !
- La **prédiction horaire**, en faisant tourner 24 méthodes d'agrégation en parallèle est possible (et améliore également la performance relative).
- Pour la prédiction des **événements extrêmes**, comme le dépassement de seuils d'alerte par exemple, on peut voir que Ridge est plus performante que le meilleur modèle.

Malheureusement, il faut savoir s'**arrêter**...



Conclusion :

La prédiction de suites individuelles est un cadre **méta-statistique** où il s'agit d'agréger séquentiellement les prédictions de méthodes fondamentales, par exemple, celles de différentes méthodes statistiques.

J'espère avoir réussi ce matin ma mission d'évangéliste en vous montrant (à la suite de Yannig) le **potentiel** et la **flexibilité** de ces techniques d'agrégation séquentielle...

Le début d'une longue romance entre nous ?

