Prédiction séquentielle par agrégation d'experts

Jaouad Mourtada

Mémoire de Master 2 :
« Probabilités et Modèles Aléatoires »
Université Pierre-et-Marie-Curie

Sous la direction d'Odalric-ambrym Maillard
Inria Saclay & L.R.I.

Table des matières

Chapitre 1

Introduction

Dans ce mémoire, nous explorons le problème de la prédiction séquentielle à l'aide d'experts sous différentes formes, en vue d'appliquer ces méthodes à la prédiction de signaux non stationnaires. Dans cette optique, nous présentons des algorithmes d'agrégation robustes dans le pire des cas, et largement applicables, qui résolvent des problèmes de complexité croissante afin de fournir une grande souplesse dans leur usage.

1.1 Plan du mémoire

Dans le chapitre 2, nous introduisons le problème de la prédiction à l'aide d'experts sous ses différentes formulations, et nous définissons des stratégies compétitives face au meilleur expert, voire au meilleur poids. Des questions importantes, comme la calibration adaptative de la vitesse d'apprentissage, sont évoquées.

Le chapitre 3 introduit une variante plus ambitieuse du problème, où il s'agit d'être compétitif face à des *séquences* d'experts, ce qui est particulièrement adapté au cadre non stationnaire. Une sophistication de ce problème est introduite, qui consiste à identifier un petit groupe de bons experts entre lesquels alterner, en particulier lorsque le nombre d'experts est important. Nous présentons divers algorithmes, issus d'une littérature plutôt récente, qui traitent ces problèmes de façon satisfaisante avec une complexité raisonnable, ainsi que des interprétations de ces algorithmes.

Le chapitre 4, qui est une contribution de notre travail, vise à conférer encore davantage de flexibilité dans notre procédure agrégative, en autorisant le nombre d'experts à augmenter. Dans ce contexte, les mêmes sophistications que pour une classe constante d'experts peuvent être considérées, et nous montrons comment les stratégies et les résultats s'adaptent.

Les résultats faisant l'objet d'une contribution personnelle, qu'il s'agisse d'une borne légèrement améliorée ou d'un résultat nouveau, sont signalés par un astérisque (*). Généralement, les astérisques des chapitres 2 et 3 renvoient à la première catégorie de contribution, et ceux du chapitre 4 à la seconde.

1.2 Remerciements

Je tiens à exprimer ma sincère reconnaissance à Odalric-ambrym Maillard d'avoir accepté de m'encadrer et de m'avoir proposé ce sujet si riche et stimulant. Ses conseils et ses remarques m'ont été d'un grand renfort, et ses nombreuses suggestions ont été une source d'inspiration précieuse. Je remercie également Richard Combes d'avoir accepté de prendre part au jury.

Chapitre 2

Prédiction séquentielle à l'aide d'experts

Dans cette partie, nous introduisons un problème qui nous occupera tout au long de ce mémoire, celui de la *prédiction à l'aide d'experts* (*prediction with expert advice* en anglais). Ce cadre est l'une des façons possibles d'aborder le problème très général de la *prédiction séquentielle des suites individuelles* : étant donné un signal, dont on observe les valeurs successivement, il s'agit d'effectuer une prédiction sur la valeur suivante du signal. Bien entendu, à ce degré de généralité, le problème est mal posé : sans hypothèse particulière sur le signal, quel que soit l'algorithme de prédiction, il existe un signal que cette algorithme prédit très mal à chaque étape. Il est donc nécessaire de préciser le problème, afin qu'il devienne abordable.

Traditionnellement, les approches statistiques consistent à supposer que le signal est la réalisation d'un processus stochastique dont il s'agit d'estimer les caractéristiques, ce qui permet de construire des intervalles de confiance et des prédictions valides avec une forte probabilité. Cependant, ces approches offrent généralement peu voire pas de garanties lorsque les hypothèses sur le signal ne sont plus vérifiées, par exemple lorsque le modèle statistique n'est pas adapté, ou plus généralement lorsque le signal ne peut pas être modélisé de manière satisfaisante par un processus stochastique.

À l'inverse, le problème de la prédiction séquentielle de suites arbitraires a été étudié dans la littérature dite de *prédiction universelle* (cf. [MF98]), mobilisant des idées et méthodes issues de la théorie de l'information, de la théorie du codage universel, et de la compression de données. Ces méthodes, qui cherchent typiquement à prédire des signaux engendrés par n'importe quel programme fini, sont adaptées à l'inférence dans un cadre très général. Cependant, dans de nombreux problèmes concrets, il existe des méthodes particulières, simples ou plus sophistiquées, reposant par exemple sur des modèles statistiques, qui offrent de bonnes performances. En effet, le prévisionniste peut avoir une certaine connaissance *a priori* du système qu'il étudie ou des intuitions sur son fonctionnement ; il peut également disposer d'un certain nombre de modèles raisonnables, dont il soupçonne que l'un d'entre eux devrait décrire suffisamment bien le comportement du signal. Il peut dans ces conditions être profitable de restreindre le problème en intégrant cette connaissance a priori, et en ne cherchant pas nécessairement à se comparer à n'importe quel automate fini comme tendent à le faire les approches de prédiction universelle.

Dans le cadre de la prédiction à l'aide d'experts, le point de vue est différent. On dispose d'un certain nombre d'*experts* qui, à chaque étape, effectuent des prédictions sur le signal ; l'objectif consiste alors à combiner ces prévisions de manière à pouvoir garantir que l'on a *a posteriori* prédit presque aussi bien que le meilleur de ces experts (celui-ci étant inconnu *a priori*). Ainsi, si au moins l'un des experts disponibles prédit convenablement le signal, et si l'on parvient à définir une stratégie de combinaison des experts qui prédise presque aussi bien que le meilleur d'entre eux, alors on dispose de prédictions de bonne qualité. Cette décomposition du problème de la prédiction en deux sous-problèmes (choix d'une classe d'experts susceptible de contenir un bon expert d'une part, définition d'une stratégie d'agrégation de ces experts afin de prédire aussi bien que le meilleur d'entre eux d'autre part) est un analogue de la décomposition biais-variance, très classique en statistiques.

Notons que, dans ce cadre, nous cherchons à obtenir des garanties valables dans tous les cas, en ne faisant typiquement aucune hypothèse non seulement sur le mécanisme qui génère le signal, mais également sur la façon dont les experts forment leurs prédictions (algorithme d'apprentissage, prédiction fixe, prévisions d'un expert humain…) qui peuvent tous deux être absolument quelconques [1]. Ainsi, les algorithmes employés et les garanties sur leur performance sont valables dans des contextes très généraux, ce qui leur confère une forte robustesse et une large applicabilité.

Dans cette partie, après avoir défini formellement le problème de la prédiction séquentielle à l'aide d'experts (section 2.1), nous introduisons une stratégie d'agrégation efficace, l'*agrégation à poids exponentiels* (section 2.2), dont nous analysons les performances sous différentes hypothèses sur la fonction de perte (sous-sections 2.2.1 et 2.2.2), et montrons qu'elle généralise un procédé naturel, le *mélange bayésien* (sous-section 2.2.3). Nous présentons ensuite des raffinements de ces méthodes, qui permettent de se comparer non seulement au meilleur expert, mais également à la meilleure combinaison convexe d'experts (section 2.3). Nous abordons également le problème de la calibration du paramètre de l'algorithme d'agrégation (section 2.4), en montrant comment améliorer ses performances lorsque l'un des experts prédit bien (sous-section 2.4.1), puis en passant en revue des algorithmes récents qui s'adaptent à la difficulté du signal (sous-section 2.4.2). Enfin, nous discutons de l'optimalité des méthodes présentées dans cette partie, en mentionnant des bornes inférieures sur le regret dans le pire des cas de n'importe quelle stratégie de prédiction (section 2.5), puis nous concluons en décrivant des formulations alternatives du problème de la prédiction à l'aide d'experts (section 2.6).

Nous nous appuyons essentiellement sur l'ouvrage de référence [CBL06], sur l'article pionnier [Vov98] et sur l'article [Sto10], ainsi que sur d'autres références, mentionnées dans le cours du texte, sur des problèmes plus spécifiques.

1. Il arrive toutefois que les prévisions de certains experts ne soient pas quelconques, mais fixes ou tout au moins *simulables* (*i.e.* correspondent à des fonctions connues des observations passées). Cette connaissance supplémentaire donne plus de pouvoir à l'apprenant, en réduisant l'ensemble des prédictions possibles des experts (cf. [CBL06], partie 8). Dans l'essentiel de ce travail, en particulier dans cette partie, nous supposerons typiquement que les prédictions des experts sont quelconques ; nous verrons cependant (à partir de la section 3.1.2) que certains algorithmes plus sophistiqués peuvent s'interpréter en termes d'agrégation classique d'un grand nombre d'experts dits *structurés*, dont les prédictions sont reliées et s'obtiennent à partir de celles d'un petit nombre d'experts élémentaires.

2.1 Formulation du problème

Définissons le problème que nous allons étudier dans cette partie — dont nous introduirons par la suite des variantes et des raffinements —, celui de la *prédiction séquentielle à l'aide d'experts*. On se donne un espace $\mathscr{Y}$, auquel appartiennent les valeurs $y_1, y_2, \ldots$ du signal à prédire ; un espace $\mathscr{X}$ des prédictions, auquel appartiennent les prédictions successives $x_1, x_2, \ldots$ de l'apprenant (en anglais *learner*, ou parfois *forecaster* *i.e.* prévisionniste) ; et une fonction de perte $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}^+$, qui mesure la qualité des prédictions, en fonction des valeurs effectives du signal. On se donne enfin un nombre M d'experts $i = 1, \ldots, M$, qui effectuent des prévisions à chaque étape, à partir desquelles l'apprenant détermine sa propre prévision.

Nous cherchons à déterminer une stratégie pour l'apprenant garantissant que celui-ci prédit le signal (au sens de la perte cumulée) presque aussi bien que le meilleur des experts $i = 1, \ldots, M$, bien que ce dernier soit inconnu a priori. Puisque nous cherchons des stratégies valides quelles que soient les valeurs du signal et les prédictions des différents experts, le problème se formalise naturellement comme un jeu entre l'*apprenant* et l'*environnement*, qui décide des prévisions des experts ainsi que des valeurs du signal.

Problème 1 (Prédiction à l'aide d'experts)**.** Étant donnés l'espace $\mathscr{Y}$ du signal, celui des prédictions $\mathscr{X}$, la fonction de perte $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}^+$ et le nombre d'experts M, la prédiction se déroule comme suit : à chaque étape $t = 1, 2, \ldots$,

1. L'environnement décide des prévisions $x_{i,t} \in \mathscr{X}$ des experts $i = 1, \ldots, M$.
2. L'apprenant observe ces prévisions, puis effectue à son tour une prédiction $x_t \in \mathscr{X}$.
3. L'environnement détermine alors la valeur $y_t \in \mathscr{Y}$ du signal. Chaque expert $i = 1, \ldots, M$ subit alors une perte $\ell_{i,t} := \ell(x_{i,t}, y_t)$, tandis que l'apprenant souffre une perte $\ell_t := \ell(x_t, y_t)$.

En notant pour $T \geqslant 1$ $L_{i,T} = \sum_{t=1}^{T} \ell_{i,t}$ la perte cumulée de l'expert i jusqu'au temps T, et $L_T = \sum_{t=1}^{T} \ell_t$ celle de l'apprenant, on définit le *regret* par rapport à l'expert i comme étant $R_{i,T} := L_T - L_{i,T}$, qui correspond au regret que ressent l'apprenant de ne pas avoir suivi l'expert i du début jusqu'au temps T. L'objectif consiste alors à contrôler le regret uniformément sur tous les experts $i = 1, \ldots, M$, *i.e.* à minimiser le regret par rapport au meilleur expert :

$$R_T := \max_{1 \leqslant i \leqslant M} R_{i,T} = L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \,,$$

que l'on appellera tout simplement *regret*.

Nous chercherons donc à définir des stratégies pour l'apprenant qui garantissent un regret faible quelles que soient les valeurs du signal et les prévisions des différents experts, c'est-à-dire quelle que soit la stratégie de l'environnement. Par exemple, on peut tout au moins désirer que le regret par tour tende vers 0 avec l'horizon de temps, *i.e.* $R_T = o(T)$.

2.2 La stratégie d'agrégation à poids exponentiels

Nous allons définir dans cette section une stratégie simple mais fondamentale qui permet de traiter efficacement le problème de la prédiction à l'aide d'experts : l'*agrégation* *à poids exponentiels*. Cet algorithme offre de bonnes garanties théoriques (c'est-à-dire des bornes de regret satisfaisantes, que nous établirons tout au long de cette partie, et dont nous discuterons l'optimalité dans la section 2.5), et s'avère peu coûteux en temps comme en mémoire.

Nous verrons par ailleurs (cf. 2.2.3) que cet algorithme correspond à une généralisation d'un procédé classique de prédiction en statistiques, à savoir le *mélange bayésien*, ce qui explique que l'on qualifiera parfois ces méthodes de « bayésiennes ». En revanche, il convient de préciser que notre cadre d'analyse n'est pas bayésien (ni d'ailleurs fréquentiste), puisque nous cherchons des bornes de regret valables *quelles que soient* les valeurs du signal, et non des bornes en espérance ou en forte probabilité sous l'hypothèse que ce signal est la réalisation d'un processus stochastique.

2.2.1 Le cadre convexe borné

Dans cette sous-section, nous allons introduire la stratégie d'agrégation à poids exponentiels, et établir ses propriétés dans le cas d'une fonction de perte convexe et bornée, très étudié dans la littérature. Nous verrons en sous-section 2.2.2 que l'hypothèse d'une fonction de perte bornée peut être abandonnée, au prix d'une hypothèse de convexité plus forte qui permet en retour d'obtenir des bornes de regret bien meilleures.

Nous supposons donc ici (et essentiellement dans tout ce mémoire) que l'espace $\mathscr{X}$ des prédictions est une partie convexe d'un espace vectoriel, et que, pour tout $y \in \mathscr{Y}$, la fonction $\ell(\cdot, y) : \mathscr{X} \to \mathbf{R}^+$ est convexe. Nous supposons également, dans cette sous-section, que la fonction de perte ℓ est bornée, spécifiquement $0 \leqslant \ell \leqslant 1$.

Remarque 2.1. *Dans le cas où l'espace de prédiction $\mathscr{X}$ n'est pas convexe (par exemple lorsque $\mathscr{X}$ est un ensemble fini ou dénombrable), il est possible de se ramener à ce cas en effectuant à chaque étape t une prédiction* randomisée *selon un vecteur de probabilités $\boldsymbol{v}_t$, i.e. en suivant l'expert i avec probabilité $v_{i,t}$. On peut alors majorer l'espérance de la perte (en se ramenant à un problème de prédiction sur l'espace convexe $\mathscr{X}' = \mathscr{P}(\mathscr{X})$ des mesures de probabilité sur $\mathscr{X}$, avec pour fonction de perte convexe $\ell'(\pi, y) := \mathbb{E}_{X\sim\pi}[\ell(X, y)] = \int_X \ell(x, y)\pi(dx)$, et en associant à la prédiction $x_{i,t} \in \mathscr{X}$ la prédiction « aléatoire » $x'_{i,t} = \delta_{x_{i,t}} \in \mathscr{X}'$), et donc aussi la perte avec forte probabilité (cf. [CBL06], section 4 pour plus de détails).*

Stratégie d'agrégation à poids exponentiels. Nous allons adopter une stratégie d'*agrégation convexe*, qui à chaque étape t prédit selon une combinaison convexe des prédictions des différents experts :

$$x_t = \sum_{i=1}^{M} v_{i,t}\, x_{i,t} = \boldsymbol{v}_t \cdot \boldsymbol{x}_t \tag{2.1}$$

où $\boldsymbol{x}_t = (x_{i,t})_{1\leqslant i\leqslant M} \in \mathscr{X}^M$, et $\boldsymbol{v}_t = (v_{i,t})_{1\leqslant i\leqslant M} \in \mathscr{P}_M$ est un vecteur de probabilité. La stratégie est alors déterminée par la procédure de choix des poids des différents experts à chaque étape, en fonction des valeurs passées du signal et des prédictions. Un procédé commode — et assez général — de choix des poids, qui offre de bonnes garanties de regret, repose sur l'usage de *potentiels* (cf. [CBL06], section 2.1). Nous introduisons directement un cas particulier de cette famille d'algorithmes, l'*agrégation à poids exponentiels*, qui possède la propriété commode que les poids affectés ne dépendent que des pertes passées des différents experts, et pas des prédictions passées de l'apprenant. Outre sa simplicité, cet algorithme présente l'avantage d'offrir des bornes de regret essentiellement optimales.

L'idée de l'algorithme consiste à attribuer un poids important aux experts qui ont bien prédit jusqu'à présent (c'est-à-dire dont la perte cumulée $L_{i,t-1}$ est faible), et un faible poids aux experts dont la perte cumulée est élevée. Un cas extrême consisterait à prédire exactement comme le meilleur expert ; où, plus précisément, à donner un poids identique à tous les experts (s'il y en a plusieurs) dont le regret cumulé est minimal, et un poids nul à tous les autres. Cette stratégie, appelée *Follow the Leader* (FTL) dans la littérature, n'offre cependant pas de bonnes garanties dans le pire des cas [2], et s'avère peu efficace lorsque le meilleur expert change régulièrement.

L'agrégation à poids exponentiels correspond à une version « régularisée » de l'algorithme FTL, attribuant des poids proches à des experts dont la perte cumulée est voisine. Cette stratégie, dépendant d'un paramètre $\eta > 0$ appelé *vitesse d'apprentissage*, consiste à attribuer à l'expert i au temps t un poids :

$$v_{i,t} = \frac{e^{-\eta \, L_{i,t-1}}}{\sum_{j=1}^{M} e^{-\eta \, L_{j,t-1}}} \,. \tag{2.2}$$

Notons que, lorsque $\eta \to \infty$, les poids donnés par la formule (2.2) tendent vers les poids uniformes sur les meilleurs experts, *i.e.* vers les poids attribués par la stratégie FTL. À l'inverse, plus la valeur de η est faible, plus les poids sont proches de la loi uniforme sur les experts, et donc plus l'algorithme est conservateur.

Stratégie 2.1 (Agrégation à poids exponentiels)**.**

Paramètre : $\eta > 0$, vitesse d'apprentissage.

Initialisation : Les poids des experts sont initialisés par : $v_{i,1} = \frac{1}{M}$ pour $i = 1, \ldots, M$, *i.e.* $\boldsymbol{v}_1 := \frac{1}{M}\mathbf{1}$.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t = (v_{i,t})_{1 \leqslant i \leqslant M}$, déterminé à la fin de l'étape précédente. On observe alors les prédictions $x_{i,t} \in \mathscr{X}$ des experts, que l'on agrège selon les poids $v_{i,t}$, *i.e.* $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$ (cf. (2.1)).

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes ℓ_t et $\ell_{i,t}$ pour $i = 1, \ldots, M$.

Mise à jour des poids : Les poids sont alors mis à jour à partir des pertes, selon la formule :

$$v_{i,t+1} = \frac{v_{i,t} \, e^{-\eta \, \ell_{i,t}}}{\sum_{j=1}^{M} v_{j,t} \, e^{-\eta \, \ell_{j,t}}} \tag{2.3}$$

pour $i = 1, \ldots, M$.

2. Pour s'en convaincre, on peut par exemple considérer le cas suivant : $M = 2$, $\mathscr{X} = \mathscr{Y} = [0, 1]$, $\ell(x, y) = |x - y|$, l'expert 1 prédit successivement $0, 0, 0 \ldots$, l'expert 2 prédit $\frac{1}{2}, 1, 1, 1, \ldots$, tandis que le signal vaut $\frac{1}{2}, 0, 1, 0, 1, \ldots$. On vérifie alors que $L_{1,T} = \frac{1}{2} + \lfloor \frac{T-1}{2} \rfloor$, $L_{2,T} = \lfloor \frac{T}{2} \rfloor$ (de sorte que le meilleur expert alterne entre 1 et 2), et donc l'algorithme FTL donne $L_T = T + O(1)$, d'où un regret linéaire : $R_T = \frac{T}{2} + O(1)$.

Remarque 2.2. *Notons que les poids définis dans la stratégie 2.1 sont bien ceux indiqués par l'équation (2.2) : en effet, il est immédiat que les poids de l'équation (2.2) satisfont la formule de récurrence (2.3). En outre, il est à noter que les poids peuvent être mis à jour facilement, avec une complexité en temps et en espace de $O(M)$. Cet aspect computationnel n'est pas trivial : dans la suite de ce mémoire, nous rencontrerons des algorithmes qui requièrent a priori de stocker un grand nombre de paramètres, si bien qu'il s'agira parfois de déterminer des variantes qui, en plus d'avoir de bonnes garanties de regret, présentent une complexité acceptable.*

Analyse du regret. Nous allons à présent effectuer une analyse du regret de l'algorithme 2.1 d'agrégation à poids exponentiels. Énonçons d'emblée le résultat (cf. [CBL06]) :

Théorème 2.1. *Supposons que la fonction de perte est convexe en son premier argument, et à valeurs dans $[0,1]$. Alors, quelles que soient les valeurs du signal et les prédictions des experts, le regret de la stratégie 2.1 d'agrégation à poids exponentiels de paramètre η vérifie, pour tout $T \geqslant 1$:*

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant \frac{\log M}{\eta} + \frac{\eta T}{8}. \tag{2.4}$$

L'ingrédient principal de la preuve est le lemme de Hoeffding, qui permet d'établir l'inégalité de concentration éponyme pour les variables aléatoires bornées :

Lemme 2.2 (Hoeffding)**.** *Soit X une variable aléatoire réelle telle que $a \leqslant X \leqslant b$ p.s., où a et b sont deux réels. Pour tout $s \in \mathbf{R}$, l'inégalité suivante est vérifiée :*

$$\log \mathbb{E}\left[e^{sX}\right] \leqslant s\,\mathbb{E}[X] + \frac{(b-a)^2}{8} s^2. \tag{2.5}$$

Afin de ne pas encombrer le texte, nous repoussons la preuve de ce résultat à l'annexe A.1. Nous allons utiliser la conséquence[3] suivante du lemme de Hoeffding : pour tout vecteur de probabilités $(v_i)_{1 \leqslant i \leqslant M} \in \mathscr{P}_M$, et toutes pertes $\ell_1, \dots, \ell_M \in [0,1]$, on a

$$\sum_{i=1}^{M} v_i\, \ell_i \leqslant -\frac{1}{\eta} \log\left(\sum_{i=1}^{M} v_i\, e^{-\eta\, \ell_i}\right) + \frac{\eta}{8} \tag{2.6}$$

pour tout $\eta > 0$.

En outre, il est commode d'avoir recours aux poids non normalisés $w_{i,t} = e^{-\eta\, L_{i,t-1}}$, de sorte que $v_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^{M} w_{j,t}}$; on note également $W_t = \sum_{i=1}^{M} w_{i,t}$, qui forme une suite décroissante. Intuitivement, si W_{t+1} est proche de W_t, cela signifie que les experts ayant un poids important au temps t ont peu perdu de poids, *i.e.* que les experts sur lesquels l'apprenant s'est appuyé le plus afin de former sa prédiction ont eu une faible perte, ce qui suggère que l'apprenant a lui-même eu une faible perte. On peut donc minorer la perte de l'apprenant à partir des poids totaux W_t ; ces poids totaux peuvent à leur tour être minorés par le poids de chaque expert, donc par la perte cumulée de chaque expert.

3. Qui s'obtient en prenant $s = -\eta$, et X une variable aléatoire à valeurs dans $[0,1]$ de loi $\sum_{i=1}^{M} v_i \delta_{\ell_i}$.

Preuve du théorème 2.1. Pour tout $t = 1, \ldots, T$, on a par convexité de la fonction $\ell(\cdot, y_t)$: $\ell(x_t, y_t) = \ell(\sum_{i=1}^M v_{i,t}\, x_{i,t}, y_t) \leqslant \sum_{i=1}^M v_{i,t}\, \ell(x_{i,t}, y_t)$, *i.e.*

$$\ell_t \leqslant \sum_{i=1}^M v_{i,t}\, \ell_{i,t}\,. \tag{2.7}$$

L'inégalité (2.6) donne alors :

$$\sum_{i=1}^M v_{i,t}\, \ell_{i,t} \leqslant -\frac{1}{\eta} \log\left(\sum_{i=1}^M v_{i,t}\, e^{-\eta\, \ell_{i,t}}\right) + \frac{\eta}{8} = -\frac{1}{\eta} \log\left(\frac{\sum_{i=1}^M w_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{i=1}^M w_{i,t}}\right) + \frac{\eta}{8}$$

d'où :

$$\ell_t \leqslant -\frac{1}{\eta} \log \frac{W_{t+1}}{W_t} + \frac{\eta}{8} \tag{2.8}$$

ce qui, en sommant pour $t = 1, \ldots, T$, donne :

$$L_T \leqslant -\frac{1}{\eta} \log \underbrace{\frac{W_{T+1}}{W_1}}_{= \frac{1}{M} \sum_{j=1}^M e^{-\eta\, L_{j,T}}} + \frac{\eta\, T}{8} \leqslant -\frac{1}{\eta} \log\left(\frac{1}{M}\, e^{-\eta\, L_{i,T}}\right) + \frac{\eta\, T}{8} = L_{i,T} + \frac{1}{\eta} \log M + \frac{\eta\, T}{8}$$

pour tout $i = 1, \ldots, M$, ce qui est précisément l'inégalité annoncée. $\square$

Examinons à présent le théorème 2.1 : on a montré que la stratégie d'agrégation exponentielle de paramètre $\eta > 0$ avait, pour toute échéance $T \geqslant 1$, un regret R_T majoré par le terme $\frac{1}{\eta} \log M + \frac{\eta T}{8}$ (et ce quelles que soient les valeurs du signal et les prédictions des experts). Quel que soit η, ce terme est linéaire en T, ce qui suggère qu'aucune de ces stratégies n'est bonne pour tout T ; en particulier, aucune ne vérifie le critère $R_T = o(T)$.

En revanche, si l'on suppose que *l'horizon de temps T est fixé et connu à l'avance*, et donc que l'on ne cherche à contrôler le regret que jusqu'au temps T, alors il est possible de choisir le $\eta > 0$ optimal qui minimise ce terme de majoration :

Corollaire 2.3. *Reprenons les hypothèses du théorème 2.1. Pour tout $T \geqslant 1$, la stratégie d'agrégation à poids exponentiels de paramètre $\eta = \eta_T^* = \sqrt{(8 \log M)/T}$ admet un regret majoré de la façon suivante :*

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant \sqrt{\frac{T}{2} \log M} \tag{2.9}$$

quelles que soient les valeurs du signal et les prédictions des experts.

Remarque 2.3. *En fait, sous les hypothèses du corollaire 2.3, on contrôle non seulement le regret R_T, mais également les regrets R_t pour $t = 1, \ldots, T$: $R_t \leqslant \frac{1}{\eta_T^*} \log M + \frac{\eta_T^*\, t}{8} \leqslant \sqrt{(T/2) \log M}$.*

Remarque 2.4. *Notons que la borne de regret (2.9), si l'on fait abstraction du fait qu'elle requiert la connaissance a priori de l'horizon temps T (afin de choisir la vitesse d'apprentissage de manière optimale), est très intéressante. D'une part, la dépendance en T est sous-linéaire (de l'ordre de $\sqrt{T}$) : cela signifie que le regret par tour tend vers*

0 *lorsque la période de temps considérée tend vers l'infini. D'autre part, la dépendance en le nombre d'experts M est excellente : elle est en $\sqrt{\log M}$, quantité qui reste faible même lorsque le nombre d'experts est très élevé. Ainsi,* il est possible d'agréger un grand nombre d'experts tout en maintenant un regret maximal raisonnable *: il s'agit d'un fait remarquable sur l'agrégation d'experts.*

Variante « anytime » de l'algorithme. Nous allons à présent indiquer comment lever la limitation principale du corollaire 2.3, qui est de nécessiter la connaissance a priori de l'horizon de temps envisagé afin d'obtenir un regret contrôlé. En d'autres termes, il s'agit de définir une variante de l'algorithme 2.1 qui assure de bonnes garanties de regret — en particulier sous-linéaire en T — *quel que soit l'horizon de temps T envisagé* ; de telles stratégies (ainsi que les bornes qui leur sont attachées) sont dites « anytime » dans la littérature anglophone.

Il s'avère qu'une astuce simple, le « doubling trick » (cf. [CBL06]), permet de modifier la stratégie 2.1 afin d'obtenir une borne *anytime* presque aussi bonne que celle (à horizon de temps fixé) du corollaire 2.3, à un facteur près. L'idée consiste à subdiviser le temps en périodes dont la taille croît exponentiellement (les périodes $[\![2^p, 2^{p+1}-1]\!]$ avec $p \geqslant 0$, pour être précis). Au début de chaque nouvelle période, les poids sont réinitialisés en les poids uniformes, puis l'on applique la stratégie d'agrégation 2.1 pendant cette période, avec un paramètre $\eta > 0$ choisi comme dans 2.3 de manière optimale en fonction de la longueur de cette période.

Stratégie 2.2 (Stratégie de poids exponentiels avec le « doubling trick »)**.** À chaque échéance $t = 1, 2, \ldots$, on effectue les actions suivantes :

Choix des poids : Si $t = 2^p$ pour un certain $p \geqslant 0$, on pose $\eta_t = \sqrt{\frac{8 \log M}{2^p}}$ et $v_{i,t} = \frac{1}{M}$ pour tout $i = 1, \ldots, M$. Sinon, on pose $\eta_t = \eta_{t-1}$ et l'on met à jour les poids selon :

$$v_{i,t} = \frac{v_{i,t-1}\, e^{-\eta_t \ell_{i,t-1}}}{\sum_{j=1}^{M} v_{j,t-1}\, e^{-\eta_t \ell_{j,t-1}}} \tag{2.10}$$

Prédiction : On observe alors les prédictions $x_{i,t} \in \mathscr{X}$ des experts, que l'on agrège selon les poids $v_{i,t}$ (cf. (2.1)). La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes ℓ_t et $\ell_{i,t}$ pour $i = 1, \ldots, M$.

La stratégie 2.2, qui ne dépend d'aucun paramètre, offre de bonnes garanties de regret « anytime » :

Proposition 2.4. *Pour tout $T \geqslant 1$, quelles que soient les valeurs du signal et les prédictions successives des experts, la stratégie 2.2 admet un regret majoré de la façon suivante :*

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant \left(2 + \sqrt{2}\right) \sqrt{\frac{T}{2} \log M} \tag{2.11}$$

Démonstration. Soit $i \in \{1, \ldots, M\}$. Pour tout $q \geqslant 0$, pendant la période $[\![2^q, 2^{q+1}-1]\!]$, l'algorithme 2.2 prédit comme la stratégie 2.1, initialisée à l'instant 2^q, et avec $\eta =$

$\sqrt{\frac{8\log M}{2^q}}$. Par la remarque 2.3, pour tout $t = 2^q, \ldots, 2^{q+1}-1$, le regret cumulé de l'algorithme entre 2^q et t vérifie :

$$L_{2^q..t} - L_{i,2^q..t} = \sum_{s=2^q}^{t} (\ell_s - \ell_{i,s}) \leqslant \sqrt{\frac{2^q}{2}\log M} = \sqrt{2}^{\,q}\sqrt{\frac{\log M}{2}}\,. \tag{2.12}$$

Soit maintenant $T \geqslant 1$, et $p \geqslant 0$ tel que $2^p \leqslant T < 2^{p+1}$ (*i.e.* $p = \lfloor \log_2 T \rfloor$). En sommant l'inégalité (2.12) sur les périodes $[\![1, 2-1]\!], \ldots, [\![2^{p-1}, 2^p - 1]\!], [\![2^p, T]\!]$, il vient :

$$L_T - L_{i,T} \leqslant \sum_{q=0}^{p} \sqrt{2}^{\,q}\sqrt{\frac{\log M}{2}} = \frac{\sqrt{2}\sqrt{2^p}-1}{\sqrt{2}-1}\sqrt{\frac{\log M}{2}} \underset{(2^q \leqslant T)}{\leqslant} \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{\frac{T}{2}\log M}$$

ce qui est bien l'inégalité voulue puisque $\frac{\sqrt{2}}{\sqrt{2}-1} = 2 + \sqrt{2}$. □

Remarque 2.5. *La borne « anytime » (2.11) que nous avons obtenue pour l'algorithme 2.2 est presque aussi bonne que la borne de regret à horizon fixé (2.9) avec un choix optimal de la vitesse d'apprentissage, à une constante multiplicative près de $2+\sqrt{2} \simeq 3,414\ldots$*

Notons que la stratégie 2.2 réinitialise les poids de tous les experts à la fin de chaque période. Cela peut paraître sous-optimal, car ce procédé « efface » le retard accumulé par certains experts. Une approche plus directe, qui ne nécessite pas de partitionner le temps en périodes ni de réinitialiser les poids, consiste à faire varier le paramètre d'apprentissage η en fonction du temps. Plus précisément, à l'étape t, l'algorithme effectue une combinaison convexe des prédictions des différents experts, avec pour poids :

$$v_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^{M} w_{j,t}}\,, \qquad \text{avec} \qquad w_{i,t} = e^{-\eta_t\, L_{i,t-1}}\,. \tag{2.13}$$

Comme la borne de regret du théorème 2.1 à horizon de temps T fixé est minimisée pour $\eta = \eta_T^* = \sqrt{\frac{8\log M}{T}}$ (corollaire 2.3), un candidat naturel pour η_t est $\sqrt{\frac{8\log M}{t}}$, et l'analyse montre que ce choix est efficace.

Stratégie 2.3 (Agrégation avec une vitesse d'apprentissage variable)**.** À chaque échéance $t = 1, 2, \ldots$, l'algorithme prédit en utilisant une combinaison convexe des prédictions des experts : $x_t = \sum_{i=1}^{M} v_{i,t} x_{i,t}$, où les poids $v_{i,t}$ sont donnés par la formule (2.13), avec $\eta_t = \sqrt{\frac{8\log M}{t}}$.

La stratégie 2.3 fournit une nouvelle borne de regret « anytime », qui améliore la constante multiplicative $2+\sqrt{2}$ de la proposition 2.4, la réduisant essentiellement à 2 :

Théorème 2.5. *Sous les hypothèses du théorème 2.1, quelles que soient les valeurs successives du signal et les prédictions des experts, le regret de la stratégie 2.3 d'agrégation à vitesse d'apprentissage variable est majoré de la façon suivante :*

$$L_T - \min_{1\leqslant i\leqslant M} L_{i,T} \leqslant 2\sqrt{\frac{T}{2}\log M} + \sqrt{\frac{\log M}{8}} \tag{2.14}$$

Nous ne démontrons pas le théorème 2.5, que nous n'utiliserons pas dans les parties suivantes [4]. Sa preuve — que l'on peut trouver dans [CBL06] — est plus longue et délicate que celle de la proposition 2.4, la principale difficulté technique étant que le paramètre d'apprentissage η_t varie à chaque étape (tandis que dans la stratégie 2.1 celui-ci est fixe, et que dans la stratégie 2.2 il ne change que très rarement), ce qui fait que l'on ne dispose plus de termes de la forme $\log \frac{W_t}{W_{t-1}}$ comme dans (2.8) qui se télescopent.

Remarque 2.6. *Tout au long de cette section, nous avons supposé que la fonction de perte était à valeurs dans* $[0,1]$. *Nous avons utilisé cette hypothèse afin d'obtenir l'inégalité (2.6), qui résulte du lemme de Hoeffding 2.2. Si maintenant l'on suppose que la fonction de pertes est à valeurs dans* $[0,B]$ [5]*, le lemme 2.2 (ou un argument direct en considérant la fonction* ℓ/B*) montre que l'inégalité (2.4) de regret de la stratégie 2.1 d'agrégation exponentielle de paramètre* $\eta > 0$ *devient :*

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant \frac{\log M}{\eta} + \frac{\eta B^2 T}{8} . \tag{2.15}$$

Ainsi, la borne de regret à horizon de temps T *pour la stratégie d'agrégation exponentielle de paramètre* η *calibré de manière optimale* [6] *(corollaire 2.3) devient :*

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant B \sqrt{\frac{T}{2} \log M} . \tag{2.16}$$

Notons enfin que les inégalités (2.15) et (2.16) découlent de l'inégalité (2.6) qui est satisfaite (avec un facteur B^2 *devant le terme* $\frac{\eta}{8}$*) sous la seule hypothèse que les pertes* $\ell_{i,t}$ *soient dans* $[0,B]$*, et pas nécessairement que la fonction* ℓ *soit elle-même bornée. Ces garanties de regret restent donc valides sous cette hypothèse, même si la fonction* ℓ *n'est pas elle-même bornée. En revanche, si la fonction* ℓ *n'est pas bornée, il n'y a aucun moyen de s'assurer a priori que les pertes le seront, et donc cette observation n'apporte aucune garantie de regret dans le pire des cas.*

2.2.2 Le cadre exp-concave et mélangeable

Exp-concavité. Dans cette sous-section, nous ne supposons plus comme en 2.2.1 que la fonction de perte ℓ est bornée. En revanche, dans le cas général d'une fonction de perte convexe et non bornée, il est facile de se convaincre que l'on ne peut obtenir aucune garantie de regret valide dans le pire des cas [7]. On pourrait même être tenté de penser qu'il est toujours impossible d'obtenir une borne de regret dans le pire des cas si la fonction ℓ est non bornée (ou du moins si, pour tout $x \in \mathscr{X}$, la fonction $\ell(x,\cdot) : \mathscr{Y} \to \mathbf{R}$ est non bornée).

4. En effet, les cas qui nous intéresseront par la suite feront généralement intervenir des fonctions de perte exp-concaves (voir les sous-sections 2.2.2 et 2.2.3), pour lesquelles il est possible d'obtenir des bornes de regret bien meilleures que dans le cadre convexe borné de cette sous-section.

5. Cas auquel il est toujours possible se ramener, par translation, si ℓ est bornée.

6. En fonction de l'horizon de temps T **et** de la borne B, de sorte que $\eta = \frac{1}{B}\sqrt{\frac{8 \log M}{T}}$.

7. Par exemple, dans le cas de la fonction de perte $\mathbf{R} \times \mathbf{R} \to \mathbf{R}^+$, $(x,y) \mapsto |x-y|$, l'environnement peut garantir en un seul tour un regret arbitrairement élevé (disons, supérieur à $B > 0$ quelconque) par rapport au meilleur de deux experts en les faisant prédire respectivement $-B$ et B, et en choisissant $\pm 2B$ de signe opposé à la prédiction de l'apprenant.

Il s'avère en fait, et c'est en soi remarquable, qu'il est possible d'obtenir des bornes universelles de regret même dans ces conditions, pourvu seulement que la fonction ℓ vérifie une condition plus forte de convexité. Mieux encore : cette condition s'avère être bien plus efficace pour majorer le regret que l'hypothèse de pertes bornées, qui ne sert d'ailleurs qu'à s'y ramener de manière approximative.

En effet, comme nous l'avons observé dans la remarque 2.6, l'hypothèse d'une fonction de perte bornée ne sert qu'à pouvoir utiliser la majoration (2.6). Jointe à cette inégalité, la convexité de ℓ en son premier argument implique alors, pour tous $y \in \mathscr{Y}$, $(v_i)_{1 \leqslant i \leqslant M} \in \mathscr{P}_M$ et $(x_i)_{1 \leqslant i \leqslant M} \in \mathscr{X}^M$:

$$\ell \left(\sum_{i=1}^{M} v_i \, x_i, y \right) \leqslant -\frac{1}{\eta} \log \left(\sum_{i=1}^{M} v_i \, e^{-\eta \, \ell(x_i, y)} \right) + \frac{\eta}{8} \, . \tag{2.17}$$

Le premier terme de la somme du membre de droite, de la forme $g_\eta^{-1}\big(\sum_{i=1}^{M} v_i \, g_\eta(\ell_i) \big)$, peut être interprété comme une « moyenne » des $\ell_i = \ell(x_i, y)$ pondérés par les v_i, déformée par la fonction $g_\eta(x) = e^{-\eta x}$. L'inégalité de Jensen [8] montre que cette moyenne est décroissante en $\eta > 0$, allant de la moyenne arithmétique (lorsque $\eta \to 0$) au minimum des ℓ_i (lorsque $\eta \to \infty$). Elle peut donc être vue comme une version régularisée du minimum des ℓ_i, d'autant plus proche de ce minimum que le paramètre η est élevé.

La borne (2.17) est donc la somme de deux termes, l'un décroissant en η, l'autre croissant. Cela se retrouve dans la borne de regret du théorème 2.1, qui fait également apparaître un compromis entre un terme décroissant ($\frac{\log M}{\eta}$) et un terme croissant ($\frac{\eta T}{8}$) en η ; pour obtenir une borne de regret, on choisit alors η de manière optimale comme dans le corollaire 2.3. Dans cette sous-section, la situation est différente : nous faisons directement l'hypothèse que la fonction de perte ℓ vérifie l'inégalité

$$\ell \left(\sum_{i=1}^{M} v_i \, x_i, y \right) \leqslant -\frac{1}{\eta} \log \left(\sum_{i=1}^{M} v_i \, e^{-\eta \, \ell(x_i, y)} \right) \tag{2.18}$$

pour tous v_i, x_i, y ; c'est-à-dire l'inégalité (2.17), mais sans le terme $\eta/8$ croissant en η. Notons que l'inégalité (2.18) ne saurait être satisfaite pour tout η (contrairement à (2.17)), car alors (en faisant $\eta \to \infty$) la perte de la prédiction agrégée serait toujours inférieure à $\min_i \ell(x_i, y)$, propriété qui n'est satisfaite par aucune fonction de perte intéressante [9]. En outre, le terme de droite de (2.18), et donc aussi la borne du regret (que nous obtiendrons ci-dessous), sont décroissants en η, qu'il faut prendre aussi grand que possible, c'est-à-dire le plus grand η satisfaisant (2.18) : il n'y a donc plus de question de calibration de la vitesse d'apprentissage, qui est uniquement déterminée par la fonction de perte.

Enfin, observons que la condition (2.18) équivaut (en appliquant la fonction $x \mapsto e^{-\eta x}$ à gauche et à droite) à la concavité de la fonction $\exp(-\eta \, \ell(\cdot, y))$ pour tout $y \in \mathscr{Y}$. Ceci amène à la définition suivante :

8. Appliquée à la fonction $g_{\eta_1} \circ g_{\eta_2}^{-1}(x) = x^{\eta_1/\eta_2}$, concave sur $\mathbf{R}_+^*$ pour $0 < \eta_1 \leqslant \eta_2$.

9. Cette propriété implique en fait que, pour tout $y \in \mathscr{Y}$, la fonction $\ell(\cdot, y)$ est constante sur l'intérieur de $\mathscr{X}$.

Définition 2.6 (Exp-concavité)**.** On dit qu'une fonction $f : \mathscr{X} \to \mathbf{R}$ est $(\eta\text{-})$*exp-concave* si la fonction $\exp(-\eta\, f)$ est concave. On dit également qu'une fonction $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}$ est $(\eta\text{-})$*exp-concave* si, pour tout $y \in \mathscr{Y}$, la fonction $\ell(\cdot, y)$ l'est ; ceci équivaut à dire que ℓ vérifie l'inégalité (2.18) pour tous v_i, x_i et y.

Remarque 2.7. *L'exp-concavité implique la convexité*[10]*, mais la réciproque est bien entendu fausse, comme le montre l'exemple suivant.*

Exemple 2.1 (Valeur absolue, norme)**.** La fonction de perte convexe $\ell : [0,1]^2 \to \mathbf{R}$ définie par $\ell(x, y) = |x - y|$ n'est pas exp-concave. Plus généralement, si C est un convexe non trivial[11] d'un espace vectoriel normé $(E, \|\cdot\|)$, la fonction de distance $d : C^2 \to \mathbf{R}$, $d(x, y) = \|x - y\|$ n'est pas exp-concave.

Démonstration. Quel que soit $\eta > 0$, la fonction $x \mapsto \exp(-\eta\, \ell(0, x)) = e^{-\eta\, x}$ est strictement convexe sur $[0, 1]$, ce qui démontre la première assertion. Pour la seconde, il suffit de considérer un segment non trivial inclus dans le convexe C (sur lequel la norme correspond à la valeur absolue), et d'appliquer le premier point. □

Exemple 2.2 (Perte quadratique)**.** La fonction de perte quadratique $\ell(x, y) = (x - y)^2$ est η-exp-concave sur $[a, b]^2$ (avec $a, b \in \mathbf{R}$, $a < b$) si et seulement si $\eta \leqslant \frac{1}{2(b-a)^2}$.

Remarque 2.8. *En particulier, la perte quadratique sur un intervalle de* $\mathbf{R}$ *est exp-concave si et seulement si elle est bornée (i.e. si l'intervalle est borné). Cela signifie que, dès qu'on peut lui appliquer les résultats de la sous-section 2.2.1 sur les fonctions de pertes bornées, on peut également lui appliquer les résultats plus forts de cette sous-section sur les fonctions de pertes exp-concaves (par exemple, le corollaire 2.12 plus bas).*

Démonstration. Par translation puis par homothétie, il suffit de démontrer le résultat pour $[a, b] = [0, 1]$. Or la fonction ℓ est η-exp-concave sur $[0, 1]$ si et seulement si la fonction $\ell(\cdot, y) = (\cdot - y)^2$ est η-exp-concave sur $[0, 1]$ pour tout $y \in [0, 1]$, *i.e.* si et seulement si la fonction $f(x) = x^2$ est η-exp-concave sur $[-y, 1 - y]$ pour tout $y \in [0, 1]$, ce qui équivaut à dire que la fonction $f_\eta = \exp(-\eta\, f)$ est concave sur $[-1, 1]$. Or $f_\eta''(x) = -2\eta\,(1 - 2\eta x^2)\exp(-\eta\, x^2)$, qui est négative sur $[-1, 1]$ si et seulement si $\eta \leqslant \frac{1}{2}$. □

Exemple 2.3. Dans la sous-section 2.2.3, nous étudierons en détail un autre exemple fondamental fonction de perte exp-concave, la « self-information loss », aussi appelée *perte logarithmique.*

Avant d'établir les propriétés de l'agrégation à poids exponentiels pour une fonction de perte exp-concave, nous allons introduire des notions plus générales, la *mélangeabilité* et la (c, η)*-réalisabilité*. Ces notions permettent non seulement d'étendre les résultats de cette section à des fonctions de perte qui ne sont pas exp-concaves, mais également d'améliorer la vitesse d'apprentissage pour des fonctions de perte exp-concaves.

10. Ceci résulte du fait que le minimum « régularisé » est un inférieur à la moyenne arithmétique, comme nous l'avons observé plus haut (on peut aussi le montrer directement à partir de la convexité de la fonction $x \mapsto e^{-\eta\, x}$).

11. C'est-à-dire contenant au moins deux points distincts.

Mélangeabilité, réalisabilité. La propriété d'exp-concavité d'une fonction de perte signifie qu'étant donnés des prédictions $x_i \in \mathscr{X}$ des experts, auxquelles on affecte les poids v_i, la prédiction agrégée par combinaison convexe $x = \sum_{i=1}^M v_i\, x_i \in \mathscr{X}$ admet une perte majorée par le minimum régularisé des pertes des experts, et ce quelle que soit la valeur $y \in \mathscr{Y}$ du signal (équation (2.18)). Cependant, rien ne garantit que la combinaison convexe soit la meilleure façon d'agréger les prédictions pour toutes les fonctions de pertes, et effectivement ce n'est pas toujours le cas. Formalisons cela par la définition suivante, adaptée de l'article [Vov98] :

Définition 2.7 (Réalisabilité)**.** Soit $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}$ une fonction de perte, et $c, \eta > 0$.

1. On dit que la fonction ℓ est (c,η)-*réalisable* si elle vérifie la propriété suivante : pour tous $M \geqslant 1$, $(v_i)_{1 \leqslant i \leqslant M} \in \mathscr{P}_M$ et $(x_i)_{1 \leqslant i \leqslant M} \in \mathscr{X}^M$, il existe un élément $x \in \mathscr{X}$ tel que
$$\ell\,(x, y) \leqslant -\frac{c}{\eta} \log \left(\sum_{i=1}^M v_i\, e^{-\eta\, \ell(x_i, y)} \right) \tag{2.19}$$
pour tout $y \in \mathscr{Y}$.
2. Soit $\mathbf{pred} : \bigcup_{M \geqslant 1} (\mathscr{P}_M \times \mathscr{X}^M) \to \mathscr{X}$ une fonction de combinaison des prédictions. On dit que le couple $(\ell, \mathbf{pred})$ est (c,η)-*réalisable* si, pour tous $(v_i)_{1 \leqslant i \leqslant M} \in \mathscr{P}_M$ et $(x_i)_{1 \leqslant i \leqslant M} \in \mathscr{X}^M$, $x = \mathbf{pred}\,((v_i)_i, (x_i)_i)$ satisfait l'inégalité (2.19) pour tout $y \in \mathscr{Y}$.

Ainsi, ℓ est (c,η)-réalisable si et seulement s'il existe une fonction **pred** telle que $(\ell, \mathbf{pred})$ soit (c,η)-réalisable.

Remarque 2.9. *Aucune fonction de perte intéressante n'est (c,η)-réalisable si $c < 1$. En effet, une telle fonction vérifierait la propriété suivante : pour tout $x \in \mathscr{X}$, il existe un $x' \in \mathscr{X}$ tel que $\ell(x', y) \leqslant c\,\ell(x, y) \underset{(\mathrm{si}\,\ell(x,y)>0)}{<} \ell(x, y)$. Autrement dit, pour chaque prédiction, il existe une autre prédiction qui prédit (strictement) mieux quelle que soit la valeur du signal ! Dans ce qui suit, on suppose donc que $c \geqslant 1$; si $c = 1$, le membre de droite est le minimum régularisé des x_i, c'est-à-dire le même majorant que dans le cas exp-concave. Cela motive la définition suivante.*

Définition 2.8 (Mélangeabilité)**.** Soit $\eta > 0$. Une fonction de perte $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}$ est dite $(\eta$-$)$*mélangeable*[12] si elle est $(1,\eta)$-réalisable.

Ainsi, toute fonction de perte $(\eta$-$)$exp-concave est $(\eta$-$)$mélangeable, mais la réciproque est fausse. En fait, même si une fonction est exp-concave, elle peut être mélangeable pour une valeur de η plus élevée, comme le montre l'exemple 2.4 suivant.

Exemple 2.4 (Perte quadratique)**.** Soit $\ell : [0,1]^2 \to \mathbf{R}$ définie par $\ell(x,y) = (x-y)^2$. Nous avons vu (exemple 2.2) que ℓ est η-exp-concave si et seulement si $\eta \leqslant \frac{1}{2}$. Cependant, il est possible de montrer (voir [HKW98, KW99]) que ℓ est η-mélangeable si et seulement si $\eta \leqslant 2$. Plus précisément, soit **pred** la fonction de combinaison définie par :
$$\mathbf{pred}(\boldsymbol{v}, \boldsymbol{x}) = \frac{1}{2} \left\{ \ell_0^{-1} \biggl(-\frac{1}{2} \log \sum_{i=1}^M v_i\, e^{-2\,\ell_0(x_i)} \biggr) + \ell_1^{-1} \biggl(-\frac{1}{2} \log \sum_{i=1}^M v_i\, e^{-2\,\ell_1(x_i)} \biggr) \right\} \tag{2.20}$$

12. Dans la littérature anglophone, on parle de « mixable loss » ; cf. [CBL06].

où, pour $y \in \{0,1\}$, $\ell_y : [0,1] \to [0,1]$ est définie par $\ell_y(x) = \ell(x,y)$, et où $\ell_y^{-1} : [0,1] \to [0,1]$ est sa réciproque. Alors, le couple $(\ell, \mathbf{pred})$ est $(1,2)$-réalisable.

Remarque 2.10. *L'exemple précédent montre que, pour la fonction de perte quadratique sur un intervalle borné, l'usage d'une fonction d'une fonction d'agrégation plus sophistiquée permet d'obtenir une meilleure majoration de la perte : en effet, la borne de droite de l'équation (2.18) est décroissante en η, et la fonction **pred** de l'exemple 2.4 permet d'obtenir un paramètre d'apprentissage η quatre fois plus élevé. En revanche, la fonction de prédiction **pred** ainsi définie dépend explicitement de la borne sur les valeurs du signal*[13]*, contrairement à l'agrégation par combinaison convexe. Cette différence est toutefois mineure puisque, même dans le cas de l'agrégation convexe, le choix du paramètre η fait intervenir la connaissance explicite d'une borne sur les valeurs du signal.*

Remarque 2.11. *Dans notre cadre de majoration du regret dans le pire des cas, il n'est pas possible de s'affranchir de l'hypothèse d'un signal borné pour la perte quadratique*[14]*. En revanche, si l'on s'autorise à avoir une borne qui dépend de l'étendue des valeurs du signal observées (ce qui est inévitable dans le pire des cas), il est possible de procéder sans la connaissance a priori de bornes sur le signal, grâce à un algorithme adaptatif (cf. [CBMS07]). De manière générale, une technique efficace pour contrôler le regret consiste à se ramener au cas mélangeable en utilisant une inégalité de concentration (l'inégalité de Hoeffding 2.2 dans le cas borné, l'inégalité de Bernstein dans [CBMS07]) ; cf. la sous-section 2.4.2 pour plus de détails.*

Si l'exemple précédent se concentre sur la propriété de mélangeabilité, dont nous verrons en effet qu'elle est particulièrement commode pour majorer le regret, la notion de (c,η)-réalisabilité a elle aussi son intérêt pour $c > 1$. Nous en verrons une application dans la section 2.4.1, où il s'agira d'obtenir une meilleure borne de regret lorsque l'un des experts prédit bien. Afin de préciser cette discussion, posons la définition suivante :

Définition 2.9. Soit $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}^+$ une fonction de perte. Pour tout $\eta > 0$, on définit $c(\eta) = c_\ell(\eta) \in [0, +\infty]$ comme l'infimum des constantes $c \geqslant 0$ telles que ℓ soit (c,η)-réalisable.

Remarque 2.12. *Sous certaines conditions, par exemple si $\mathscr{X}$ est un espace topologique compact et si ℓ est continue en son premier argument, si $c_\ell(\eta) < \infty$ alors cette constante est atteinte, i.e. ℓ est $(c_\ell(\eta), \eta)$-réalisable.*

Exemple 2.5 (Perte en valeur absolue)**.** Soit $\ell : [0,1]^2 \to [0,1]$, $(x,y) \mapsto |x-y|$. Alors ℓ n'est pas mélangeable ; plus précisément, on a :

$$c_\ell(\eta) = \frac{\eta/2}{\log(2/(1+e^{-\eta}))} . \tag{2.21}$$

pour tout $\eta > 0$ (cf. [CBL06]).

13. Et également, en principe, d'une borne sur les prédictions des experts, mais ce n'est en fait pas nécessaire : en effet, on peut toujours projeter les prédictions des experts sur l'intervalle auquel appartient le signal, ce qui améliore ces prédictions.

14. Comme on le voit facilement par un argument identique à celui que nous avons fourni plus haut dans le cas de la valeur absolue.

Enfin, on dispose d'un résultat général de réalisabilité pour les fonctions convexes bornées, que nous utiliserons en section 2.4.1.

Exemple 2.6. Soit ℓ une fonction de perte convexe en son premier argument, à valeurs dans $[0,1]$. Pour tout $\eta > 0$, la fonction de combinaison convexe $\mathbf{pred}(\boldsymbol{v}, \boldsymbol{x}) = \boldsymbol{v} \cdot \boldsymbol{x}$ est $(\frac{\eta}{1-e^{-\eta}}, \eta)$-réalisable pour ℓ.

Démonstration. En faisant $s = -\eta$ dans le lemme A.2 (prouvé en annexe), on obtient l'inégalité suivante, valable pour toute variable aléatoire X à valeurs dans $[0,1]$ et tout $\eta > 0$:

$$\frac{1-e^{-\eta}}{\eta} \, \mathbb{E} X \leqslant -\frac{1}{\eta} \, \mathbb{E}\left[e^{-\eta X}\right] . \tag{2.22}$$

Soient maintenant $\boldsymbol{x} = (x_i)_{1 \leqslant i \leqslant M} \in \mathscr{X}^M$, $\boldsymbol{v} = (v_i)_{1 \leqslant i \leqslant M} \in \mathscr{P}_M$ et $y \in \mathscr{Y}$. En posant $x = \boldsymbol{v} \cdot \boldsymbol{x}$, on a par convexité de $\ell(\cdot, y)$: $\ell(x,y) \leqslant \sum_{i=1}^M \ell(x_i, y)$. En appliquant l'inégalité (2.22) à une variable X de loi $\sum_{i=1}^M v_i \delta_{\ell(x_i,y)}$, on en déduit l'inégalité :

$$\ell\,(x,y) \leqslant -\frac{\eta}{1-e^{-\eta}} \, \frac{1}{\eta} \log \left(\sum_{i=1}^M v_i \, e^{-\eta \, \ell(x_i, y)} \right) \tag{2.23}$$

comme annoncé. □

Algorithme d'agrégation. Nous allons maintenant décrire une adaptation de la stratégie d'agrégation à poids exponentiels 2.1, en remplaçant la combinaison convexe par une fonction de combinaison **pred** (c,η)-réalisable [15]. Outre l'introduction de la fonction **pred**, nous introduisons quelques variations et généralisations de l'algorithme 2.1 ; notons que les bornes de regret obtenues, et plus généralement toute borne pour les fonctions de perte exp-concaves, se transfèrent au cas des fonctions de pertes convexes bornées *via* l'inégalité (2.6).

Soit $M \geqslant 1$ un entier, ou $M = \infty$. Si $M < \infty$, les experts sont comme précédemment indexés par les entiers $i = 1, \ldots, M$, tandis que le cas $M = \infty$ permet de considérer une infinité dénombrable d'experts, indexés par les entiers $i \in \mathbf{N}^*$. Dans le cas d'une infinité d'experts, nous voyons qu'il n'est plus possible d'attribuer les mêmes poids initiaux à tous les experts (car alors ceux-ci ne seraient plus sommables) ; ceci suggère une généralisation de l'attribution initiale des poids, utile aussi lorsque $M < \infty$. Jusqu'à la fin de cette sous-section, on se fixe un couple $(\ell, \mathbf{pred})$ qui est (c,η)-réalisable.

On se donne donc une suite $(\omega_i)_{i=1}^M$ de poids initiaux strictement positifs, telle que $\sum_{i=1}^M \omega_i = 1$. À chaque étape $t \geqslant 1$, étant données les prédictions $\boldsymbol{x}_t = (x_{i,t})_{i=1}^M$ des experts, on prédit selon $x = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$, où les poids $\boldsymbol{v}_t$ sont définis en normalisant les poids $\boldsymbol{w}_t$ (c'est-à-dire en posant $\boldsymbol{v}_t = \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_1}$) définis par :

$$w_{i,t} = \omega_i \, e^{-\eta \, L_{i,t-1}} . \tag{2.24}$$

En résumé, on applique la stratégie suivante [16] :

15. [Vov98] désigne par le terme « aggregating algorithm » tout algorithme d'agrégation à poids exponentiels de paramètre η utilisé avec une fonction **pred** (c,η)-réalisable, avec $c = c_\ell(\eta)$ optimal.

16. Où l'on a utilisé une mise à jour récursive des poids $\boldsymbol{v}_t$, en utilisant la formule de récurrence qu'ils vérifient.

Stratégie 2.4 (Agrégation à poids exponentiels étendue)**.**

Paramètres : la fonction de combinaison **pred**, la vitesse d'apprentissage $\eta > 0$, et les poids initiaux des experts $(\omega_i)_{i=1}^M$.

Initialisation des poids : On pose $v_{i,1} = \omega_i$ pour tout i.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$.

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes $\ell_t = \ell(x_t, y_t)$ et $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout i.

Mise à jour des poids : Les poids sont alors mis à jour par :

$$v_{i,t+1} = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}} \,. \tag{2.25}$$

La borne regret de la stratégie 2.4 découle directement du lemme simple mais fondamental suivant :

Lemme 2.10. *Soient ℓ_t, $(\ell_{i,t})_{i=1}^M$ des pertes réelles, et $L_{i,t} = \sum_{s=1}^t \ell_{i,s}$ les pertes cumulées. Soient $w_{i,t}$ les poids (non normalisés) associés, définis par $w_{i,t} = \omega_i \exp(-\eta L_{i,t-1})$, où $(\omega_i)_{i=1}^M$ est un vecteur de probabilité à composantes strictement positives. On suppose que, pour tout $t \geqslant 1$, on a :*

$$\ell_t \leqslant -\frac{c}{\eta} \log \left(\frac{\sum_{i=1}^M w_{i,t} \exp(-\eta\, \ell_{i,t})}{\sum_{i=1}^M w_{i,t}} \right) . \tag{2.26}$$

Alors, en notant $L_T = \sum_{t=1}^T \ell_t$, on a pour tout T et tout i :

$$L_T \leqslant c L_{i,T} + \frac{c}{\eta} \log \left(\frac{1}{\omega_i} \right) . \tag{2.27}$$

Démonstration. Pour tout $t \geqslant 1$, soit $W_t = \sum_{i=1}^M w_{i,t}$; en particulier, $W_1 = 1$. L'inégalité (2.26) se réécrit alors en $\ell_t \leqslant -\frac{c}{\eta} \log \left(\frac{W_{t+1}}{W_t} \right)$, d'où, en sommant pour $t = 1, \ldots, T$:

$$L_T \leqslant -\frac{c}{\eta} \log W_{T+1} \,. \tag{2.28}$$

On conclut alors en observant que, pour tout i :

$$\log W_{T+1} = \log \left(\sum_{j=1}^M \omega_j \exp(-\eta L_{j,T}) \right) \geqslant \log \big(\omega_i \exp(-\eta L_{i,T}) \big) = -\eta L_{i,T} - \log \left(\frac{1}{\omega_i} \right) .$$

□

Pour appliquer le lemme 2.10 à l'étude de la stratégie 2.4, il suffit d'observer que, puisque $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$, et puisque $(\ell, \mathbf{pred})$ est (c, η)-réalisable, la perte de l'apprenant vérifie pour tout t :

$$\ell_t \leqslant -\frac{c}{\eta} \log \left(\sum_{i=1}^M v_{i,t} \exp(-\eta\, \ell_{i,t}) \right) , \tag{2.29}$$

c'est-à-dire précisément l'inégalité (2.26) compte tenu de la définition des poids $v_{i,t}$ (cf. (2.24)). On en déduit le théorème suivant :

Théorème 2.11. *Supposons que le couple* $(\ell, \boldsymbol{pred})$ *soit* (c, η)*-réalisable. Alors, quelles que soient les valeurs du signal et les prédictions des experts, le regret de l'algorithme d'agrégation 2.4 de paramètre d'apprentissage* η *et de poids initiaux* $(\omega_i)_{i=1}^M \in \mathscr{P}_M$ *vérifie, pour tout* $T \geqslant 1$ *et tout* i *:*

$$L_T \leqslant cL_{i,T} + \frac{c}{\eta} \log\left(\frac{1}{\omega_i}\right). \tag{2.30}$$

Le théorème 2.11 est particulièrement puissant dans le cadre mélangeable ($c = 1$), où il fournit une borne de regret explicite :

Corollaire 2.12. *Supposons que la fonction de perte est* η*-mélangeable, et que la fonction de combinaison* $\boldsymbol{pred}$ *est* $(1, \eta)$*-réalisable (ce qui est vérifié pour la combinaison convexe lorsque* ℓ *est* η*-exp-concave). Alors, quelles que soient les valeurs du signal et les prédictions des experts, le regret de l'algorithme d'agrégation 2.4 de paramètre d'apprentissage* η *et de poids initiaux* $(\omega_i)_{i=1}^M \in \mathscr{P}_M$ *vérifie, pour tout* $T \geqslant 1$ *et tout* i *:*

$$L_T - L_{i,T} \leqslant \frac{1}{\eta} \log\left(\frac{1}{\omega_i}\right). \tag{2.31}$$

En particulier, si M *est fini, le choix* $\omega_1 = \cdots = \omega_M = \frac{1}{M}$ *donne un regret maximal :*

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant \frac{1}{\eta} \log M. \tag{2.32}$$

Remarque 2.13. *Une propriété spectaculaire des bornes de regret* (2.31) *et* (2.32) *est que celles-ci* ne dépendent pas du temps T*. En d'autres termes,* dans le cadre mélangeable, le regret de la stratégie 2.4 par rapport à n'importe quel expert est borné au cours du temps ! *Il n'y a donc pas besoin dans ce cas de calibrer la vitesse d'apprentissage* η *au cours du temps, celle-ci devant être prise aussi grande que possible*[17] *afin de minimiser la borne* (2.31)*.*

Remarque 2.14. *Dans le cas particulier où* $\boldsymbol{pred}(\boldsymbol{v}, \boldsymbol{x}) = \boldsymbol{v} \cdot \boldsymbol{x}$ *est la fonction de combinaison convexe et où les poids initiaux sont uniformes* ($\omega_i = \frac{1}{M}$ *pour tout* $i = 1, \ldots, M$)*, l'algorithme d'agrégation 2.4 coïncide avec la stratégie 2.1 d'agrégation « simple » à poids exponentiels de paramètre* η*. Il résulte alors du corollaire 2.12 que, si la fonction de perte* ℓ *est* η*-exp-concave (pas nécessairement bornée), le regret de la stratégie 2.1 est majoré par* $\frac{1}{\eta} \log M$*, borne dont la dépendance temporelle est bien meilleure que celle des bornes en* $C\sqrt{T \log M}$ *obtenues dans le cas borné.*

Choix des poids pour une infinité d'experts. En plus d'être indépendante du temps, la borne (2.32) exhibe une bonne dépendance en le nombre M d'experts (en $\log M$), ce qui montre qu'il est possible d'agréger un grand nombre d'experts tout en gardant un regret contrôlé. Nous allons voir qu'il est en fait possible d'agréger une infinité d'experts, en maintenant pour tout $M \geqslant 1$ le regret par rapport aux M premiers experts majoré par un terme de l'ordre de $\frac{1}{\eta} \log M$ (c'est-à-dire la borne que l'on aurait obtenue en n'agrégeant que ces M experts). Ainsi, on peut ajouter presque « gratuitement » autant d'experts qu'on le souhaite !

17. C'est-à-dire égale à $\eta_m = \sup\{\eta > 0 \mid c_\ell(\eta) = 1\}$, avec une fonction de combinaison **pred** adaptée.

La preuve de cette assertion repose tout simplement sur la borne (2.31), avec un bon choix des poids initiaux $(\omega_i)_{i\geqslant 1}$. Ainsi, si les poids ω_i sont décroissants, l'inégalité (2.31) implique :

$$L_T - \min_{1\leqslant i\leqslant M} L_{i,T} \leqslant \frac{1}{\eta}\log\frac{1}{\omega_M} \tag{2.33}$$

pour tout $M \geqslant 1$ et tout $T \geqslant 1$. Nous allons discuter de différents choix possibles pour les poids ω_i, et des bornes qui s'y rapportent.

Exemple 2.7. Un premier choix simple mais tout-à-fait convenable des poids initiaux est donné par : $\omega_i = \frac{1}{i(i+1)}$ pour tout $i \geqslant 1$[18]. Le terme de droite de la borne (2.33) vaut alors :

$$\frac{1}{\eta}\big(\log M + \log(M+1)\big) \underset{M\to\infty}{=} \frac{2}{\eta}\log M + o(1)\,,$$

c'est-à-dire deux fois la borne obtenue en n'agrégeant que les M premiers experts.

Exemple 2.8. Pour $\alpha > 1$, on pose $\omega_i = \zeta(\alpha)^{-1}\, i^{-\alpha}$, où $\zeta(\alpha) = \sum_{i\geqslant 1} i^{-\alpha}$. La borne (2.33) devient alors : $\frac{\alpha}{\eta}\log M + \frac{\log\zeta(\alpha)}{\eta}$.

1. Si $\alpha = 2$, on a $\zeta(2) = \frac{\pi^2}{6}$, d'où $\omega_i = \frac{6}{\pi^2 i^2}$ et une borne de $\frac{2}{\eta}\log M + \frac{\log(\pi^2/6)}{\eta} \leqslant \frac{2}{\eta}\log M + \frac{1}{2\eta}$, c'est-à-dire du même ordre que dans l'exemple 2.7.
2. Dans le cas général, la borne est asymptotiquement de meilleure qualité lorsque α est proche de 1. En écrivant $\alpha = 1+\varepsilon$ avec $\varepsilon > 0$, la borne de regret devient, compte tenu de l'inégalité[19] $\zeta(1+\varepsilon) \leqslant 1 + \frac{1}{\varepsilon}$:

$$\frac{1+\varepsilon}{\eta}\log M + \frac{1}{\eta}\log\left(1+\frac{1}{\varepsilon}\right).$$

Exemple 2.9. Pour tout $i \geqslant 1$, on pose $\omega_i = \log 2\left(\frac{1}{\log(i+1)} - \frac{1}{\log(i+2)}\right)$. On a alors $\omega_M \underset{M\to\infty}{\sim} \frac{\log 2}{M\log^2(M)}$, la borne (2.33) vaut donc asymptotiquement :

$$\frac{1}{\eta}\log M + \frac{2}{\eta}\log\log M - \frac{\log\log 2}{\eta} + o(1) \underset{M\to\infty}{\sim} \frac{1}{\eta}\log M\,.$$

Cette borne est donc essentiellement équivalente à celle que l'on aurait obtenue en n'agrégeant que les M premiers experts, confirmant ce que l'on avait annoncé plus haut.

Remarque 2.15. *Pour conclure cette sous-section, revenons sur la borne* (2.31). *Elle montre que plus le poids initial attribué à l'expert i est élevé, plus le regret par rapport à cet expert est faible. Dans les exemples précédents, nous avons donné des choix génériques de ces poids (uniformes dans le cas fini, sommables mais à décroissance lente dans le cas infini) ; en revanche, si l'on soupçonne que l'un des experts sera meilleur que les autres, on a intérêt à lui accorder un poids important. Ces poids initiaux (ainsi que le procédé d'agrégation à poids exponentiels) ont en fait une interprétation naturelle, comme on va le voir dans la sous-section suivante.*

18. Ces poids sont bien normalisés car $\sum_{i\geqslant 1}\frac{1}{i(i+1)} = \sum_{i\geqslant 1}(\frac{1}{i} - \frac{1}{i+1}) = 1$.
19. En effet, $\zeta(1+\varepsilon) = \sum_{n\geqslant 1} n^{-(1+\varepsilon)} \leqslant 1 + \sum_{n\geqslant 2}\int_{n-1}^{n} t^{-(1+\varepsilon)}dt = 1 + \int_1^\infty t^{-\varepsilon-1}dt = 1 + \frac{1}{\varepsilon}$.

2.2.3 Un cas particulier fondamental : le mélange bayésien

Dans cette sous-section [20], nous présentons en détail un exemple fondamental de fonction de perte exp-concave (non bornée), la *perte logarithmique* ou « self-information loss ». Dans ce cadre, l'algorithme d'agrégation à poids exponentiels, ainsi qu'un certains nombre de quantités (pertes cumulées, poids initiaux ou mis à jour des experts...) s'interprètent naturellement en termes bayésiens, ce qui permet de retrouver rapidement les bornes de regret déjà obtenues.

Définition 2.13 (Perte logarithmique)**.** Soit $\mathscr{Y}$ un ensemble fini ou dénombrable, et $\mathscr{X} = \mathscr{P}(\mathscr{Y})$ l'ensemble [21] des mesures de probabilités sur $\mathscr{Y}$. La *perte logarithmique* est la fonction $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}^+$ définie par $\ell(p, y) = -\log p(y)$.

Ainsi, dans ce cadre, les prédictions sont des mesures de probabilité sur l'ensemble des valeurs possibles du signal, et la perte associée à une prédiction est d'autant plus élevée que cette prédiction attribuait une faible probabilité à la valeur du signal révélée.

Remarque 2.16. *La définition 2.13 s'étend naturellement au cas non discret, de la façon suivante. Supposons $\mathscr{Y}$ muni d'une tribu $\mathscr{B}$ et d'une mesure de référence λ sur $(\mathscr{Y}, \mathscr{B})$. Si $\mathscr{X}$ est une partie convexe de l'ensemble des fonctions mesurables $f : (\mathscr{Y}, \mathscr{B}) \to [0, +\infty]$ telles que $\int_{\mathscr{Y}} f d\lambda = 1$, on définit la perte logarithmique par : $\ell(f, y) = -\log f(y)$ pour tout $(f, y) \in \mathscr{X} \times \mathscr{Y}$ [22]. Pour simplifier les formulations, on se placera dans cette sous-section dans le cadre de la définition 2.13, mais tout ce qui sera dit des probabilités sur $\mathscr{Y}$ s'adapte immédiatement aux densités.*

Une propriété importante de la perte logarithmique, quoique triviale à vérifier, est son exp-concavité (avec $\eta = 1$). En effet, l'identité $e^{-\ell(p,y)} = p(y)$ implique immédiatement, pour tous $p_1, \dots, p_M \in \mathscr{X}$, $(\lambda_1, \dots, \lambda_M) \in \mathscr{P}_M$ et $y \in \mathscr{Y}$:

$$\ell\left(\sum_{i=1}^{M} \lambda_i\, p_i, y\right) = -\log \sum_{i=1}^{M} \lambda_i\, e^{-\ell(p_i, y)}\,; \tag{2.34}$$

de sorte que la perte logarithmique satisfait l'inégalité d'exp-concavité (2.18), qui est en fait une égalité dans ce cas. En particulier, on peut lui appliquer le corollaire 2.12, qui montre par exemple que la stratégie 2.1 admet dans ce cas un regret majoré par $\log M$. Nous allons retrouver directement ces bornes, après avoir donné une interprétation bayésienne de l'agrégation à poids exponentiels dans ce contexte.

Stratégies de prédiction et interprétation probabiliste. Commençons par quelques observations générales, valables pour une fonction de perte ℓ quelconque. Dans cette partie (et dans les deux suivantes), nous cherchons à établir des stratégies d'agrégation valides

20. Notre présentation s'inspire notamment du début du chapitre 9 de [CBL06]. Cependant, dans [CBL06] les auteurs se restreignent par commodité au cas d'experts « simulables », qui prédisent selon des stratégies fixes et connues *a priori* (cf. les explications ci-dessous) ; à l'inverse, notre discussion couvre le cas général où les stratégies de prédiction des experts sont quelconques et inconnues *a priori*.

21. Vu comme une partie convexe de l'espace vectoriel $\mathbf{R}^{\mathscr{Y}}$.

22. La définition 2.13 se retrouve comme un cas particulier de cette définition, en prenant (lorsque $\mathscr{Y}$ est dénombrable) pour $\mathscr{B}$ l'ensemble des parties de $\mathscr{Y}$, et pour λ la mesure de comptage $\sum_{y \in \mathscr{Y}} \delta_y$.

quelles que soient les valeurs du signal et les prédictions des experts, c'est-à-dire pour lesquelles la quantité suivante est contrôlée :

$$\sup_{(y_t)_{1\leqslant t\leqslant T}} \sup_{(x_{i,t})_{1\leqslant i\leqslant M, 1\leqslant t\leqslant T}} \big(L_T - \min_{1\leqslant i\leqslant M} L_{i,T}\big)\,. \tag{2.35}$$

Il s'avère que contrôler le regret $R_T = L_T - \min_{1\leqslant i\leqslant M} L_{i,T}$ par rapport à toutes les suites (y_t) et toutes les prédictions $(x_{i,t})$ des experts équivaut à contrôler le regret par rapport à toutes les suites (y_t) et toutes les *stratégies de prédiction* $\sigma_1, \ldots, \sigma_M$ des experts $i = 1, \ldots, M$:

Définition 2.14 (Stratégie de prédiction)**.** Une *stratégie de prédiction* σ est une famille $(\sigma_t)_{t\geqslant 1}$ de fonctions $\sigma_t : \mathscr{Y}^{t-1} \to \mathscr{X}$. On dit alors qu'un expert i suit la stratégie de prédiction σ_i si sa prédiction est $x_{i,t} = \sigma_{i,t}(y^{t-1})$ pour tout t.

On dit qu'une stratégie de prédiction σ est *statique* si, pour tout t, la fonction σ_t est constante ; en d'autres termes, si les prévisions de σ ne dépendent que de t, et pas de y^{t-1}.

Remarque 2.17. *La notion de stratégie de prédiction (ne dépendant que des valeurs passées du signal, et pouvant se rapporter aussi bien à l'apprenant qu'aux experts) ne doit pas être confondue avec celle de* stratégie d'agrégation *de l'apprenant, qui dépend non seulement des valeurs précédentes du signal, mais également des prédictions passées et courantes des experts. Les « stratégies » que nous avons considérées jusqu'ici (par exemple les stratégies 2.1 et 2.4) sont des stratégies d'agrégation.*

L'assertion qui précède la définition 2.14 se montre sans difficulté, en raisonnant dans les deux sens : étant donnée une borne de regret pour toutes les prédictions des experts, on en déduit une borne sur toutes les stratégies en l'appliquant aux prédictions $\sigma_{i,t}(y^{t-1})$; réciproquement, une borne de regret sur toutes les stratégies donne une borne de regret sur toutes les prédictions, en considérant les stratégies de prédiction statiques $\sigma_{i,t} \equiv x_{i,t}$.

Penchons-nous maintenant sur le cas où $\mathscr{X} = \mathscr{P}(\mathscr{Y})$, *i.e.* où les prédictions sont des probabilités sur les valeurs possibles du signal. Dans ce contexte, une stratégie de prédiction p associe, à tout t et tout $y^{t-1} \in \mathscr{Y}^{t-1}$, une mesure de probabilité $p_t(\,\cdot \mid y^{t-1})$ sur $\mathscr{Y}$. On lui associe, pour tout $t \geqslant 1$, la mesure de probabilité sur $\mathscr{Y}^t$ définie par :

$$p^t(y^t) = \prod_{s=1}^{t} p_s(y_s \mid y^{s-1})\,. \tag{2.36}$$

De plus, en sommant la relation de récurrence $p^t(y^t) = p^{t-1}(y^{t-1})\, p_t(y_t \mid y^{t-1})$ sur y_t, on voit que la suite de probabilités p^t est *cohérente*, au sens où la marginale de p^t sur les $t-1$ premières coordonnées est p^{t-1}. Par le théorème d'extension de Kolmogorov, il existe une unique mesure de probabilité sur $\mathscr{Y}^\infty$, encore notée p, dont la marginale sur les t premières coordonnées est p^t pour tout t. Réciproquement, étant donnée une mesure de probabilité $p \in \mathscr{P}(\mathscr{Y}^\infty)$, on obtient une stratégie de prédiction $(p_t)_{t\geqslant 1}$ donnée par les probabilités conditionnelles :

$$p_t(y_t \mid y^{t-1}) = \frac{p^t(y^t)}{p^{t-1}(y^{t-1})}\,, \tag{2.37}$$

où $p^t \in \mathscr{P}(\mathscr{Y}^t)$ désigne la marginale de p sur les t premières coordonnées, donnée par $p^t(y^t) = p\left(\{y^t\} \times \mathscr{Y}_{t+1}^{\infty}\right)$. Ainsi, *la donnée d'une stratégie de prédiction équivaut à celle d'une loi de probabilité sur les suites de valeurs du signal*[23].

Remarque 2.18. *Dans ce cadre, une stratégie de prédiction p est statique (définition 2.14) si et seulement si la probabilité associée est de la forme : $p^t(y^t) = p_1(y_1) \cdots p_t(y_t)$, i.e. si, sous p, les différentes valeurs du signal sont indépendantes*[24].

Via cette identification, la perte logarithmique cumulée admet elle-même une interprétation probabiliste. En effet, si $p \in \mathscr{P}(\mathscr{Y}^\infty)$ est une stratégie de prédiction, sa perte cumulée jusqu'au temps T, étant donnée la réalisation y^T du signal, vaut :

$$\sum_{t=1}^{T} - \log p_t(y_t \,|\, y^{t-1}) = - \log \prod_{t=1}^{T} p_t(y_t \,|\, y^{t-1}) = - \log p^T(y^T)\,, \tag{2.38}$$

qui est (l'opposé du) logarithme de la probabilité associée par p au signal, c'est-à-dire la log-vraisemblance.

Mélange de stratégies. Nous avons vu plus haut qu'une stratégie d'agrégation garantit un regret contrôlé quelles que soient les valeurs du signal et les prédictions des experts si et seulement si elle garantit un regret contrôlé quels que soient le signal et les stratégies de prédiction des experts.

Supposons pour commencer que les experts $i = 1, \dots, M$ ne suivent pas des stratégies quelconques, mais prédisent chacun selon une stratégie $p_i \in \mathscr{P}(\mathscr{Y}^\infty)$ connue à l'avance. Ce cas de figure survient dès lors que les experts correspondent à des algorithmes déterminés et connus, et non à des « boîtes noires » prédictives. Dans ce cas, l'apprenant ne cherche plus à contrôler le regret pour toutes les stratégies de prédictions possibles des experts, mais seulement pour ces stratégies de prédictions particulières.

Puisque les stratégies des experts sont fixées, leurs prédictions sont déterminées par les valeurs passées du signal. Ainsi, l'apprenant ne perd rien à utiliser une stratégie de prédiction $p \in \mathscr{P}(\mathscr{Y}^\infty)$, car toute stratégie d'agrégation prédit comme une stratégie de prédiction face à ces experts. Dans ce cas, d'après l'équation (2.38), le regret de l'apprenant par rapport à l'expert i au temps T est :

$$L_T - L_{i,T} = \log \frac{p_i^T(y^T)}{p^T(y^T)}\,. \tag{2.39}$$

Ainsi, majorer le regret revient à minorer le rapport $p^T(y^T)/p_i^T(y^T)$. Un choix simple, qui garantit un regret contrôlé à toute échéance et quel que soit le signal, consiste à prendre pour p un *mélange* des p_i :

Définition 2.15 (Mélange de stratégies)**.** Soit $(p_i)_{i=1}^M$ une famille de stratégies de prédiction de $\mathscr{P}(\mathscr{Y}^\infty)$, et $\boldsymbol{\omega} = (\omega_i)_{i=1}^M \in \mathscr{P}_M$ un vecteur de poids. La *stratégie de mélange des p_i selon $\boldsymbol{\omega}$* est par définition la stratégie de prédiction $p = \sum_{i=1}^M \omega_i \, p_i \in \mathscr{P}(\mathscr{Y}^\infty)$.

23. De même, on établit à partir des équations (2.36) et (2.37) une correspondance entre les stratégies de prédiction jusqu'au temps $t \geqslant 1$ et les lois de probabilité sur $\mathscr{Y}^t$.

24. Au sens où les variables aléatoires $Y_t : (\mathscr{Y}^\infty, \mathscr{F}^\infty, p) \to \mathscr{Y}$ ($\mathscr{F}^\infty$ désignant la tribu produit sur $\mathscr{Y}^\infty$) définies par $Y_t(y^\infty) = y_t$, $t \geqslant 1$, sont indépendantes.

Ainsi, lorsque $p = \sum_{i=1}^M \omega_i \, p_i$ est un mélange des p_i, on a pour tous $T \geqslant 1$, $i \in [\![1, M]\!]$, et $y^T \in \mathscr{Y}^T$: $p^T(y^T) = \sum_{j=1}^M \omega_j \, p_j^T(y^T) \geqslant \omega_i \, p_i^T(y^T)$, donc d'après l'équation (2.39)

$$L_T - L_{i,T} \leqslant \log \frac{1}{\omega_i} \, . \tag{2.40}$$

Nous pouvons également décrire explicitement la stratégie de prédiction associée à la mesure $p = \sum_{i=1}^M \omega_i \, p_i$. Pour tout $t \geqslant 1$ et tout $y^t \in \mathscr{Y}^t$, on a par définition (cf. (2.37)) :

$$p_t(y_t \,|\, y^{t-1}) = \frac{p^t(y^t)}{p^{t-1}(y^{t-1})} = \frac{\sum_{i=1}^M \omega_i \, p_i^t(y^t)}{\sum_{i=1} \omega_i \, p_i^{t-1}(y^{t-1})} = \frac{\sum_{i=1}^M \omega_i \, p_i^{t-1}(y^{t-1}) \, p_{i,t}(y_t \,|\, y^{t-1})}{\sum_{i=1} \omega_i \, p_i^{t-1}(y^{t-1})} \tag{2.41}$$

Remarque 2.19. *Ainsi, la prédiction $p_t(y_t \,|\, y^{t-1})$ de la stratégie de mélange est une combinaison convexe de celles des experts $p_{i,t}(y_t \,|\, y^{t-1})$. Nous allons interpréter les poids des différents experts dans l'équation (2.41) de deux manières, la première leur donnant un sens bayésien, et la seconde mettant en évidence le lien entre le mélange de stratégies de prédiction et la stratégie d'agrégation à poids exponentiels.*

Interprétation bayésienne du mélange. Commençons par donner une interprétation bayésienne du mélange de stratégie de prédictions. On définit la mesure de probabilité P sur l'espace mesurable $(\mathscr{M} \times \mathscr{Y}^\infty, \mathcal{P}(\mathscr{M}) \otimes \mathscr{F}^\infty)$ (où $\mathscr{M} = [\![1, M]\!]$ ou $\mathbf{N}^*$ est l'ensemble des experts, et $\mathscr{F}^\infty$ la tribu produit sur $\mathscr{Y}^\infty$) par, pour tout $i \in \mathscr{M}$ et $A \in \mathscr{F}$, $P(\{i\} \times A) = \omega_i \, p_i(A)$. La définition de P (que l'on appelle la *loi jointe bayésienne*) équivaut à se donner un *modèle bayésien*, de loi a priori $\boldsymbol{\omega} \in \mathscr{P}_M$ sur les experts i, et tel que la loi du signal conditionnellement à l'expert i soit $p_i \in \mathscr{P}(\mathscr{Y}^\infty)$.

Remarque 2.20. *Pour ne pas encombrer les notations, nous adaptons la convention légèrement abusive de désigner également par P les marginales de la loi P. Plus précisément, en posant les variables aléatoires*[25] *$I(i, y^\infty) = i$ et $Y^\infty(i, y^\infty) = y^\infty$, on note par exemple $P(i) = P(I = i) = P(\{i\} \times \mathscr{F}^\infty) = \omega_i$, $P(y^t) = P(Y^t = y^t)$ ou encore $P(y^t \,|\, i) = P(Y^t = y^t \,|\, I = i) = p_i^t(y^t)$.*

Dans ce cadre, *la stratégie de prédiction de mélange p définie précédemment correspond à la marginale de P sur $\mathscr{Y}^\infty$* (*i.e.* à la loi de Y^∞ sous P). En effet, pour tout $A \in \mathscr{F}^\infty$, on a $p(A) = \sum_{i=1}^M \omega_i \, p_i(A) = \sum_{i=1}^M P(\{i\} \times A) = P(\mathscr{M} \times A) = P(A)$.

Nous pouvons également interpréter l'équation (2.41) d'un point de vue bayésien. Tout d'abord, par l'observation précédente, il vient $p_t(y_t \,|\, y^{t-1}) = P(y_t \,|\, y^{t-1})$, tandis que $p_{i,t}(y_t \,|\, y^{t-1}) = P(y_t \,|\, i, y^{t-1})$ car la loi $P(\cdot \,|\, i) \in \mathscr{P}(\mathscr{Y}^\infty)$ est égale à p_i. Pour ce qui est des poids, on a pour tout i :

$$\frac{\omega_i \, p_i^{t-1}(y^{t-1})}{\sum_{j=1} \omega_j \, p_j^{t-1}(y^{t-1})} = \frac{P(i, y^{t-1})}{\sum_{j=1}^M P(j, y^{t-1})} = P(i \,|\, y^{t-1}) \, . \tag{2.42}$$

Ainsi, *le poids de l'expert i correspond à la probabilité a posteriori de cet expert conditionnellement aux valeurs passées du signal.* Compte tenu de ces égalités, l'équation (2.41) se réécrit sous la forme :

$$P(y_t \,|\, y^{t-1}) = \sum_{i=1}^M P(i \,|\, y^{t-1}) P(y_t \,|\, i, y^{t-1}) \, , \tag{2.43}$$

25. Définies sur l'espace de probabilités $(\mathscr{M} \times \mathscr{Y}^\infty, \mathcal{P}(\mathscr{M}) \otimes \mathscr{F}^\infty, P)$.

qui est une formule de Bayes pour la loi jointe P.

Lien avec l'agrégation à poids exponentiels. Nous avons montré que le mélange de stratégies, qui s'interprète naturellement de manière bayésienne, offre de bonnes garanties de regret. En revanche, tel qu'on l'a présenté, ce mélange ne semble être accessible que lorsque l'on connaît *a priori* les stratégies de prédiction des experts, puisqu'il est défini a partir de ces stratégies de prédiction.

Pour voir comment cette limitation peut être levée, revenons à des stratégies d'agrégation comme la stratégie 2.4, dont on sait déjà qu'elles offrent des garanties de regret quelles que soient le signal et les prédictions (donc les stratégies de prédiction) des experts.

Fait 2.16. *Soit $\boldsymbol{\omega} \in \mathscr{P}_M$ un vecteur de poids. Quelles que soient les stratégies de prédiction $p_i \in \mathscr{Y}^\infty$ des experts, ainsi que les valeurs du signal, la stratégie d'agrégation à poids exponentiels de paramètre $\eta = 1$ et de poids initiaux $\boldsymbol{\omega}$ prédit comme la stratégie de mélange $p = \sum_{i=1}^M \omega_i\, p_i$.*

Remarque 2.21. *Ainsi, il est possible de prédire comme le mélange des stratégies de prédictions des experts, même lorsque celles-ci sont inconnues a priori.*

En outre, la borne de regret (2.40) pour le mélange des stratégies permet de retrouver la borne de regret pour la stratégie d'agrégation à poids exponentiels (corollaire 2.12) dans le cas de la perte logarithmique.

Démonstration. À l'étape t, étant données les valeurs passées du signal $y^{t-1} \in \mathscr{Y}^{t-1}$ et les prédictions présentes $p_{i,t}(y_t \,|\, y^{t-1})$ des experts, la stratégie d'agrégation à poids exponentiels prédit $\hat{p}_t \in \mathscr{P}(\mathscr{Y})$, où :

$$\hat{p}_t(y_t) = \frac{\sum_{i=1}^M \omega_i\, e^{-L_{i,t-1}}\, p_{i,t}(y_t \,|\, y^{t-1})}{\sum_{i=1}^M \omega_i\, e^{-L_{i,t-1}}}\, ; \tag{2.44}$$

or, d'après l'équation (2.38), $e^{-L_{i,t-1}} = p_i^{t-1}(y^{t-1})$, donc d'après l'équation (2.41) il vient $\hat{p}_t(y_t) = p_t(y_t \,|\, y^{t-1})$, où $p = \sum_{i=1}^M \omega_i\, p_i \in \mathscr{P}(\mathscr{Y}^\infty)$. $\square$

Remarque 2.22. *Au regard de 2.16, il est possible de considérer la stratégie d'agrégation à poids exponentiels comme une extension du mélange bayésien au cas d'une fonction de perte quelconque. Ceci permet notamment de donner une interprétation bayésienne de cette stratégie, à partir de la formulation bayésienne du cas de la perte logarithmique.*

Ainsi, la prédiction x_t de l'apprenant selon l'algorithme d'agrégation correspond à la loi de Y_t conditionnellement à y^{t-1} (sous la loi jointe bayésienne P), soit $P(Y_t \,|\, y^{t-1}) \in \mathscr{P}(\mathscr{Y}^\infty)$, donnée par la formule de Bayes (2.43).

Les poids s'interprètent également de manière simple. Ainsi, les poids non normalisés $w_{i,t} = \omega_i\, e^{-\eta\, L_{i,t-1}}$ correspondent aux probabilités $P(i, y^{t-1})$; par conséquent, le poids total au temps t, soit $W_t = \sum_{i=1}^M w_{i,t}$, vaut $P(y^{t-1})$ (ce qui donne une autre façon de voir leur décroissance). En particulier, l'inégalité $L_T \leqslant -\frac{1}{\eta} \log W_{T+1}$ (établie dans la preuve du lemme 2.10, cf. (2.28)) devient une égalité. Quant aux poids normalisés $v_{i,t} = w_{i,t}/W_t$, ils correspondent aux probabilités postérieures $P(i \,|\, y^{t-1})$; la formule (2.3) de mise à jour de ces poids s'écrit alors :

$$P(i \,|\, y^t) = P(i \,|\, y^{t-1})\, \frac{P(y_t \,|\, i, y^{t-1})}{P(y_t \,|\, y^{t-1})}\, . \tag{2.45}$$

2.3 Regret par rapport à une combinaison convexe d'experts

Nous avons vu jusqu'à présent comment la stratégie d'agrégation à poids exponentiels permet de se comparer au meilleur expert, avec un regret contrôlé. Un objectif plus général consiste à se comparer à la meilleure combinaison convexe d'experts. Nous verrons que l'on peut donner deux sens différents à cet objectif, le second plus ambitieux que le premier. Pour le premier, la stratégie d'agrégation 2.4 convient, tandis que pour le second il faut avoir recours à des stratégies différentes, un peu plus sophistiquées.

2.3.1 Entropie relative et perte logarithmique

Commençons par définir une quantité importante issue de la théorie de l'information (cf. [CT06]), l'*entropie relative*, qui mesure la différence entre deux mesures de probabilités.

Une façon naturelle d'introduire cette quantité est de partir de la fonction de perte logarithmique $\ell(p, y) = -\log p(y)$ (sur un espace $\mathscr{Y}$ discret) de la sous-section précédente. Dans ce cadre, on doit prédire une loi de probabilité sur le signal, puis la valeur de ce signal est révélée. Il est alors naturel, lorsque le signal est la réalisation d'une variable aléatoire Y de loi $p \in \mathscr{P}(\mathscr{Y})$, de s'intéresser à l'espérance de la perte obtenue en prédisant $q \in \mathscr{P}(\mathscr{Y})$, soit :

$$\ell(q, p) = -\mathbb{E}_{Y \sim p} \log q(Y) = \sum_{y \in \mathscr{Y}} p(y) \log \frac{1}{q(y)} . \tag{2.46}$$

Intuitivement, on aimerait que l'espérance de cette perte soit minimisée lorsque l'on prédit la « vraie » loi du signal, soit $q = p$. Dans cette veine, introduisons les définitions suivantes :

Définition 2.17 (Entropie, entropie relative)**.** Soient $p, q \in \mathscr{P}(\mathscr{Y})$.

1. L'*entropie de* p, soit $\mathrm{Ent}(p) \in \mathbf{R}^+$, est l'espérance de la perte $\ell(p, p) = -\mathbb{E}_{Y \sim p} \log p(Y)$ obtenue en prédisant selon la loi du signal.
2. L'*entropie relative*, ou *divergence de Kullback-Leibler*, entre p et q est l'espérance du regret obtenu en prédisant q plutôt que la loi p du signal, soit :

$$\Delta(p \,\|\, q) = \ell(q, p) - \ell(p, p) = \mathbb{E}_{Y \sim p} \log \frac{p(Y)}{q(Y)} = \sum_{y \in \mathscr{Y}} p(y) \log \frac{p(y)}{q(y)} . \tag{2.47}$$

Remarque 2.23. *De même que pour la perte logarithmique (remarque 2.16), la définition de l'entropie et de l'entropie relative s'étendent au cas de mesures à densité par rapport à une mesure de référence. De plus, l'entropie relative ne dépend pas du choix de la mesure de référence, puisque la formule (2.47) se réécrit*[26] *:* $\Delta(P \,\|\, Q) = \mathbb{E}_{Y \sim P} \log \frac{dP}{dQ}(Y)$.

Mentionnons quelques faits élémentaires sur l'entropie et l'entropie relative (cf. [CT06]) :

26. Avec la convention que $\Delta(P \,\|\, Q) = +\infty$ lorsque P n'est pas absolument continue par rapport à Q, *i.e.* lorsque Q attribue une masse nulle à un événement de P-probabilité positive.

Propriétés 2.18. *On a les propriétés suivantes :*

1. *Si $\mathscr{Y}$ est discret, l'entropie est positive, et nulle uniquement pour les mesures de Dirac.*
2. *Pour toute probabilité $\boldsymbol{u} \in \mathscr{P}_M$ avec $M \geqslant 1$, on a $\mathrm{Ent}(u) \leqslant \log M$ avec égalité si et seulement si $\boldsymbol{u} = \frac{1}{M}\mathbf{1}$ est la mesure uniforme sur $[\![1, M]\!]$.*
3. *Pour toutes lois $p, q \in \mathscr{P}(\mathscr{Y})$, on a $\Delta(p \,\|\, q) \geqslant 0$, avec égalité si et seulement si $p = q$.*
4. *Pour tout $\boldsymbol{u} \in \mathscr{P}_M$, on a $\Delta(u \,\|\, \frac{1}{M}\mathbf{1}) \leqslant \log M$.*

Démonstration. Le premier point est évident. Le point crucial est le troisième, qui reste valide dans le cas général de deux mesures P, Q sur un même espace mesurable.

Dans ce cas, on suppose sans perte de généralité P absolument continue par rapport à Q, puis on pose $A = \{x : \frac{dP}{dQ}(x) > 0\}$ et $\frac{dQ}{dP} = \frac{1}{dP/dQ}$ sur A (de sorte que $Q = \frac{dQ}{dP} P$ sur A). Comme $P(A) = 1$, l'inégalité de Jensen donne :

$$\Delta(P \,\|\, Q) = -\int_A \log \frac{dQ}{dP}(x) P(dx) \geqslant -\log \int_A \frac{dQ}{dP}(x) P(dx) = -\log Q(A) \geqslant 0\,.$$

Pour que l'égalité ait lieu, il faut d'une part que $Q(A) = 1$ et, d'autre part, par stricte concavité du log, que $\frac{dQ}{dP}$ soit P-presque sûrement constante sur A. Comme on a aussi $P(A) = 1$, ceci implique $P = Q$.

Enfin, l'identité $\Delta(\boldsymbol{u} \,\|\, \frac{1}{M}\mathbf{1}) = \log M - \mathrm{Ent}(\boldsymbol{u})$ montre le quatrième et le deuxième point, en lui appliquant respectivement le premier et le troisième point. □

Remarque 2.24. *D'après la seconde propriété ci-dessus, lorsque l'on cherche à prédire une variable $Y \sim p$ par une loi q, la perte en espérance $\ell(q, p) = -\mathbb{E}_{Y \sim p} \log q(Y) = \mathrm{Ent}(p) + \Delta(p \,\|\, q)$ est toujours supérieure à $\mathrm{Ent}(p)$, avec égalité si q est la loi p du signal. Cela permet d'interpréter l'entropie comme la difficulté intrinsèque à prévoir le signal, du fait de son incertitude a priori ; quant à l'entropie relative, elle s'interprète comme l'espérance du regret d'avoir prédit q plutôt que la loi p du signal.*

Nous utiliserons également le lemme suivant, dont la preuve est immédiate :

Lemme 2.19. *Pour tous $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathscr{P}_M$, on a :*

$$\Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}) = \Delta(\boldsymbol{u} \,\|\, \boldsymbol{w}) + \sum_{i=1}^{M} u_i \log \frac{w_i}{v_i}\,. \tag{2.48}$$

En particulier, si $\boldsymbol{v} \geqslant \beta \boldsymbol{w}$ avec $\beta > 0$, on a $\Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}) \leqslant \Delta(\boldsymbol{u} \,\|\, \boldsymbol{w}) + \log \frac{1}{\beta}$.

2.3.2 Comparaison aux combinaisons convexes des pertes

Une première façon de se comparer à n'importe quelle combinaison convexe des experts consiste à majorer, pour tout $\boldsymbol{u} \in \mathscr{P}_M$, la quantité :

$$L_T - \sum_{i=1}^{M} u_i L_{i,T} = \sum_{i=1}^{M} u_i (L_T - L_{i,T})\,, \tag{2.49}$$

c'est-à-dire n'importe quelle combinaison convexe des regrets. Par commodité, nous supposons la fonction de perte η-mélangeable ; le cas convexe borné s'y ramène par l'inégalité (2.6).

Remarque 2.25. *La combinaison convexe des pertes $\sum_{i=1}^M u_i L_{i,T}$ à laquelle on se compare admet une interprétation naturelle : il s'agit de l'espérance de la perte cumulée obtenue en choisissant à chaque étape de manière aléatoire un expert selon la loi de probabilité $\boldsymbol{u} \in \mathscr{P}_M$.*

Naturellement, les borne du corollaire 2.12 pour le regret par rapport au meilleur expert implique directement une borne sur n'importe quelle combinaison convexe des regrets. Par exemple, si M est fini, l'algorithme d'agrégation 2.4 avec des poids initiaux uniformes ($\boldsymbol{\omega} = \frac{1}{M}\mathbf{1}$) donne (cf. corollaire 2.12) un regret $L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} \leqslant \frac{1}{\eta} \log M$. Comme $\inf_{\boldsymbol{u} \in \mathscr{P}_M} \left(\sum_{i=1}^M u_i L_{i,T} \right) = \min_{1 \leqslant i \leqslant M} L_{i,T}$, le regret (2.49) est également majoré par $\frac{1}{\eta} \log M$. Plus généralement, pour M fini ou infini, soit $\boldsymbol{u} \in \mathscr{P}_M$. Par le corollaire 2.12, le regret de la stratégie 2.4 de poids initiaux $\boldsymbol{\omega} \in \mathscr{P}_M$ vérifie $L_T - L_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\omega_i}$ pour tout i, d'où :

$$\sum_{i=1}^M u_i (L_T - L_{i,T}) \leqslant \sum_{i=1}^M u_i \log \frac{1}{\omega_i} \, . \tag{2.50}$$

Comme nous allons le voir, il est possible d'obtenir mieux que la borne (2.50), en particulier lorsque la mesure $\boldsymbol{u}$ est proche de la distribution de poids initiale $\boldsymbol{\omega}$ de l'algorithme.

Commençons par le lemme utile suivant (il s'agit essentiellement du lemme 4 de [BW02]), qui relie de manière fine le regret à l'évolution des poids. Afin d'appliquer ce lemme à des stratégies d'agrégation différentes de l'algorithme 2.4 — comme nous le ferons dans le chapitre 3 — introduisons une notation spécifique. Si $\boldsymbol{v}_t \in \mathscr{P}_M$ est le vecteur de poids au temps t, et si les pertes au temps t sont les $\ell_{i,t} \in \mathbf{R}$, le vecteur de poids *postérieur* $\boldsymbol{v}_t^m \in \mathscr{P}_M$ est défini par :

$$v_{i,t}^m = \frac{v_{i,t} \, e^{-\eta \, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t} \, e^{-\eta \, \ell_{j,t}}} \, , \tag{2.51}$$

qui correspond simplement à la mise à jour des poids (2.25) pour la stratégie d'agrégation à poids exponentiels 2.4, et s'interprète d'un point de vue bayésien comme la mise à jour des poids conditionnellement à l'observation du signal au temps t (équation (2.45)).

Lemme 2.20. *Soit $\boldsymbol{v}_t \in \mathscr{P}_M$ un vecteur de poids, et $\ell_{i,t} \in \mathbf{R}$ les pertes des experts. Supposons que ℓ_t vérifie l'inégalité :*

$$\ell_t \leqslant -\frac{1}{\eta} \log \sum_{i=1}^M v_{i,t} \, e^{-\eta \, \ell_{i,t}} \, . \tag{2.52}$$

Alors, en notant $\boldsymbol{v}_t^m \in \mathscr{P}_M$ les poids postérieurs définis par (2.51), on a, pour tout $\boldsymbol{u} \in \mathscr{P}_M$:

$$\ell_t - \sum_{i=1}^M u_i \, \ell_{i,t} \leqslant \frac{1}{\eta} \sum_{i=1}^M u_i \log \frac{v_{i,t}^m}{v_{i,t}} = \frac{1}{\eta} \big(\Delta(\boldsymbol{u} \parallel \boldsymbol{v}_t) - \Delta(\boldsymbol{u} \parallel \boldsymbol{v}_t^m) \big) \, . \tag{2.53}$$

Remarque 2.26. *Via l'inégalité* (2.6)*, le lemme* 2.20 *implique le lemme 1 de [CBGLS12], qui sert à l'étude du cas convexe borné : pour tous poids* $\boldsymbol{v}_t, \boldsymbol{u} \in \mathscr{P}_M$ *et tout vecteur de pertes* $\boldsymbol{\ell}_t \in [0,1]^M$*, on a*

$$\sum_{i=1}^{M} v_{i,t}\, \ell_{i,t} - \sum_{i=1}^{M} u_i\, \ell_{i,t} \leqslant \frac{1}{\eta} \sum_{i=1}^{M} u_i \log \frac{v_{i,t}^m}{v_{i,t}} + \frac{\eta}{8}\,. \tag{2.54}$$

Preuve du lemme 2.20*.* Pour tout i, on a :

$$\begin{aligned} -\frac{1}{\eta} \log \sum_{i=1}^{M} v_{i,t}\, e^{-\eta\, \ell_{i,t}} - \ell_{i,t} &= -\frac{1}{\eta} \log \sum_{i=1}^{M} v_{i,t}\, e^{-\eta\, \ell_{i,t}} + \frac{1}{\eta} \log v_{i,t}\, e^{-\eta\, \ell_{i,t}} - \frac{1}{\eta} \log v_{i,t} \\ &= \frac{1}{\eta} \left(\log v_{i,t}^m - \log v_{i,t} \right). \end{aligned}$$

L'inégalité (2.52) implique donc : $\ell_t - \ell_{i,t} \leqslant \frac{1}{\eta} \log \frac{v_{i,t}^m}{v_{i,t}}$ pour tout i. L'inégalité (2.53) s'en déduit en faisant une combinaison convexe selon $\boldsymbol{u}$ de ces inégalités. Quant à l'égalité de l'équation (2.53), elle résulte du lemme 2.19. □

Venons-en à l'analyse de la stratégie 2.4 d'agrégation à poids exponentiels dans le cas mélangeable. Si la fonction de perte ℓ est η-mélangeable avec une fonction de combinaison **pred**, la perte de la prédiction $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$ satisfait l'inégalité (2.52). On peut donc appliquer le lemme 2.20 ; compte-tenu de ce que de plus $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t^m$, il vient alors : $\ell_t - \sum_{i=1}^{M} u_i\, \ell_{i,t} \leqslant \frac{1}{\eta}\big(\Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}_t) - \Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}_{t+1})\big)$ pour tout t. En sommant pour $t = 1, \ldots, T$, et en notant que $\Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}_{T+1}) \geqslant 0$, on obtient la borne de regret (cf. par exemple [BW02]) :

Proposition 2.21. *Si la fonction de perte* ℓ *est* η*-mélangeable, l'algorithme d'agrégation* 2.4 *de poids initial* $\boldsymbol{v}_1 = \boldsymbol{\omega} \in \mathscr{P}_M$ *garantit la borne suivante : pour tout* $\boldsymbol{u} \in \mathscr{P}_M$ *et tout* $T \geqslant 1$*,*

$$L_T - \sum_{i=1}^{M} u_i\, L_{i,T} \leqslant \frac{1}{\eta}\, \Delta(\boldsymbol{u} \,\|\, \boldsymbol{\omega})\,. \tag{2.55}$$

Remarque 2.27. *En prenant* $\boldsymbol{u} = \delta_i$ *dans la proposition* 2.21*, on retrouve la borne du corollaire* 2.12*.*

Remarque 2.28. *La borne* (2.55) *améliore la borne de regret « naïve »* (2.50)*, en remplaçant* $\sum_{i=1}^{M} u_i \log \frac{1}{\omega_i}$ *par* $\sum_{i=1}^{M} u_i \log \frac{u_i}{\omega_i}$*. Lorsque* $\boldsymbol{u} = \boldsymbol{\omega}$*, l'amélioration est particulièrement sensible, puisque la borne* (2.55) *devient nulle. Autre exemple : si* $\boldsymbol{\omega}$ *est la mesure uniforme sur* $[\![1, M]\!]$*, alors pour toute mesure* $\boldsymbol{u}$ *uniforme sur* $N \leqslant M$ *experts, la borne* (2.55) *donne un majorant en* $\frac{1}{\eta} \log \frac{M}{N}$*, ce qui améliore la borne* $\frac{1}{\eta} \log M$ *donnée par* (2.50)*.*

Notons enfin que la proposition 2.21 s'adapte facilement au cadre convexe borné :

Proposition 2.22. *Supposons la fonction de perte* ℓ *convexe et à valeurs dans* $[0,1]$*. Alors, l'algorithme d'agrégation à poids exponentiels* 2.4 *de poids initial* $\boldsymbol{v}_1 = \boldsymbol{\omega} \in \mathscr{P}_M$ *garantit la borne suivante : pour tout* $\boldsymbol{u} \in \mathscr{P}_M$ *et tout* $T \geqslant 1$*,*

$$L_T - \sum_{i=1}^{M} u_i\, L_{i,T} \leqslant \frac{1}{\eta}\, \Delta(\boldsymbol{u} \,\|\, \boldsymbol{\omega}) + \frac{\eta\, T}{8}\,. \tag{2.56}$$

2.3.3 Comparaison au meilleur choix constant de poids

Il est possible de donner un second sens, plus ambitieux, à l'objectif de se comparer au meilleur choix constant de poids. Au lieu de se comparer à la quantité $\sum_{i=1}^{M} u_i \, L_{i,t} = \sum_{t=1}^{T} \boldsymbol{u} \cdot \boldsymbol{\ell}_t$ (ce qui revient essentiellement à se comparer au meilleur expert, même si l'on a pu obtenir des bornes plus fines), on peut chercher à se comparer à la perte :

$$L_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \; ; \qquad \ell_t(\boldsymbol{u}) = \ell(\boldsymbol{u} \cdot \boldsymbol{x}_t, y_t) \tag{2.57}$$

qui correspond à la perte obtenue en combinant à chaque étape les prédictions des experts selon le vecteur de poids $\boldsymbol{u} \in \mathscr{P}_M$.

Remarque 2.29. *L'interprétation de la quantité (2.57) est à contraster avec celle de la combinaison linéaire des pertes, qui correspond à l'espérance de la perte en choisissant l'un des experts de manière aléatoire selon la probabilité* $\mathbf{u}$ *(remarque 2.25).*

Remarque 2.30. *Notons que, par convexité de la fonction de perte en son premier argument, on a* $\ell_t(\boldsymbol{u}) \leqslant \boldsymbol{u} \cdot \boldsymbol{\ell}_t = \sum_{i=1}^{M} u_i \, \ell_{i,t}$ *et donc* $L_T(\boldsymbol{u}) \leqslant \sum_{i=1}^{M} u_i \, L_{i,T}$*. Ainsi, la tâche présente est plus difficile, d'autant que l'on ne peut pas se ramener à une comparaison au meilleur expert puisqu'en général :*

$$\inf_{\boldsymbol{u} \in \mathscr{P}_M} L_T(\boldsymbol{u}) < \min_{1 \leqslant i \leqslant M} L_{i,T} \, , \tag{2.58}$$

la perte du meilleur poids étant parfois bien meilleure que celle du meilleur expert[27].

L'algorithme EG. La stratégie d'agrégation à poids exponentiels 2.1 n'étant compétitive que face au meilleur expert, nous allons définir un nouvel algorithme, qui utilise davantage la fonction de perte de ℓ. Notre analyse suit celle de [Sto10].

Dans ce qui suit, en plus de supposer ℓ convexe en son premier argument, nous faisons l'hypothèse qu'elle admet en tout point un *sous-gradient*, c'est-à-dire un élément $\nabla \ell(x, y) = \nabla_x \ell(x, y) \in \mathbf{R}^M$ tel que, pour tout $x' \in \mathscr{X}$:

$$\ell(x, y) - \ell(x', y) \leqslant \nabla \ell(x, y) \cdot (x - x') \, . \tag{2.59}$$

L'existence d'un sous-gradient est garantie si $\ell(\cdot, y)$ est différentiable pour tout $y \in \mathscr{Y}$, et il est alors unique : il s'agit du gradient en x de $\ell(\cdot, y)$. En outre, par un résultat élémentaire d'analyse convexe, un sous-gradient existe également si $\ell(\cdot, y)$ se prolonge en une fonction convexe sur un voisinage convexe de $\mathscr{X}$.

Supposons maintenant qu'à l'étape t, l'apprenant prédise selon $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$, avec $\boldsymbol{v}_t \in \mathscr{P}_M$. L'inégalité (2.59) donne alors, pour tout $\boldsymbol{u} \in \mathscr{P}_M$,

$$\ell(x_t, y_t) - \ell(\boldsymbol{u} \cdot \boldsymbol{x}_t, y_t) \leqslant \nabla \ell(x_t, y_t) \cdot (\boldsymbol{v}_t \cdot \boldsymbol{x}_t - \boldsymbol{u} \cdot \boldsymbol{x}_t) = \nabla \ell(x_t, y_t) \cdot \sum_{i=1}^{M} (v_{i,t} - u_i) \, x_{i,t} \tag{2.60}$$

27. Pour un cas extrême, considérons la perte logarithmique sur $\mathscr{Y} = \{0, 1\}$ avec deux experts, le premier prédisant toujours δ_0 et le second δ_1. Si le signal contient au moins un 0 et un 1, chacun des experts a une perte infinie, tandis que les combinaisons convexes strictes de ces experts ont des pertes finies. Il est dans ce contexte inutile de se comparer au meilleur expert, tandis qu'il est bien plus intéressant de contrôler le regret par rapport au meilleur poids.

c'est-à-dire, en notant $\widetilde{\ell}_{i,t} = \nabla \ell(x_t, y_t) \cdot x_{i,t} \in \mathbf{R}$ les *pseudo-pertes* :

$$\ell_t - \ell_t(\boldsymbol{u}) \leqslant \sum_{i=1}^{M} v_{i,t}\, \widetilde{\ell}_{i,t} - \sum_{i=1}^{M} u_i\, \widetilde{\ell}_{i,t} \tag{2.61}$$

Si les pseudo-pertes $\widetilde{\ell}_{i,t}$ sont bornées, on aimerait majorer le terme de droite de l'inégalité (2.61) en appliquant l'inégalité (2.54). Cela suggère de mettre à jour les poids $v_{i,t}$ en fonction des pseudo-pertes plutôt que des pertes des experts. On obtient alors l'algorithme suivant, dit *algorithme EG* (pour « exponentiated gradient »), analysé dans [KW97] :

Stratégie 2.5 (Algorithme EG)**.**

Paramètres : $\eta > 0$ la vitesse d'apprentissage, et $\boldsymbol{\omega} \in \mathscr{P}_M$ les poids initiaux ;

Initialisation des poids : On pose $\boldsymbol{v}_1 = \boldsymbol{\omega}$.

Prédiction : À l'étape $t = 1, 2, \dots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$.

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, ce qui permet de calculer les pseudo-pertes $\widetilde{\ell}_{i,t} = \nabla \ell(x_t, y_t) \cdot x_{i,t}$ pour tout i.

Mise à jour des poids : Les poids sont alors mis à jour par :

$$v_{i,t+1} = \frac{v_{i,t}\, e^{-\eta \widetilde{\ell}_{i,t}}}{\sum_{j=1}^{M} v_{j,t}\, e^{-\eta \widetilde{\ell}_{j,t}}} \,. \tag{2.62}$$

En d'autres termes, l'algorithme EG prédit à chaque étape par combinaison convexe $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$ des prédictions des experts, où :

$$v_{i,t} = \frac{\omega_i \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(x_s, y_s) \cdot x_{i,s}\right)}{\sum_{j=1}^{M} \omega_j \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(x_s, y_s) \cdot x_{j,s}\right)} \,. \tag{2.63}$$

Le poids de l'expert i est donc faible si [28] $\sum_{s=1}^{t-1} \nabla \ell(x_s, y_s) \cdot (x_{i,s} - x_s)$ est élevé, c'est-à-dire si le fait de modifier légèrement les prévisions de l'apprenant en direction de celles de l'expert i aurait augmenté la perte cumulée. L'algorithme EG garantit alors la borne de regret suivante, que nous exprimons d'abord sous une forme générale :

Proposition* 2.23. *Supposons que, pour tout t, les pseudo-pertes de l'algorithme 2.5 satisfont :* $\sup_{i,j} |\widetilde{\ell}_{i,t} - \widetilde{\ell}_{j,t}| \leqslant C_t$*, avec $C_t > 0$. Alors, en notant $\overline{C}_T = \sqrt{\frac{1}{T}\sum_{t=1}^{T} C_t^2}$, l'algorithme EG de poids initial $\boldsymbol{\omega}$ et de vitesse d'apprentissage $\eta > 0$ garantit, pour tout $\boldsymbol{u} \in \mathscr{P}_M$:*

$$L_T - L_T(\boldsymbol{u}) \leqslant \frac{1}{\eta} \Delta(\boldsymbol{u} \,\|\, \boldsymbol{\omega}) + \frac{\eta\, \overline{C}_T^2\, T}{8} \,. \tag{2.64}$$

Notons que la proposition 2.23 ne nécessite pas d'hypothèse particulière sur la fonction de perte, hormis sa convexité et l'existence de sous-gradients. En revanche, nous avons besoin d'une borne sur les pseudo-pertes, que l'on ne peut garantir dans le pire des cas

28. En retranchant le terme $\sum_{s=1}^{t-1} \nabla \ell(x_s, y_s) \cdot x_s$ ne dépendant pas de i.

sans hypothèse particulière sur ℓ. Nous avons donné une hypothèse relativement faible, quoique l'on puisse faire un peu mieux au prix de quelques efforts supplémentaires [29].

Démonstration. Tout d'abord, par l'inégalité (2.61), on a $\ell_t - \ell_t(\boldsymbol{u}) \leqslant \sum_{i=1}^M (v_{i,t} - u_i)\,\widetilde{\ell}_{i,t}$. De plus, comme par hypothèse les $\widetilde{\ell}_{i,t}$ appartiennent à un intervalle de longueur C_t, et par définition de la mise à jour (2.62), l'inégalité (2.54) donne :

$$\sum_{i=1}^M (v_{i,t} - u_i)\,\widetilde{\ell}_{i,t} \leqslant \frac{1}{\eta}\sum_{i=1}^M u_i \log\frac{v_{i,t+1}}{v_{i,t}} + \frac{\eta\, C_t^2}{8}\,. \tag{2.65}$$

On obtient alors la borne désirée en sommant sur $t = 1, \dots, T$ et en utilisant (2.48). $\square$

Dans le cas où M est fini, en prenant dans la proposition 2.23 les poids initiaux uniformes $\boldsymbol{\omega} = \frac{1}{M}\mathbf{1}$ et en simplifiant l'hypothèse de la borne sur les pseudo-pertes, on obtient le corollaire suivant (théorème 2 de [Sto10]) :

Corollaire 2.24. *Supposons que les pseudo-pertes sont bornées, à valeurs dans l'intervalle $[-C, C]$. Alors, pour tout $\eta > 0$, la stratégie 2.5 de poids initiaux uniformes garantit :*

$$L_T - \inf_{\boldsymbol{u}\in\mathscr{P}_M} L_T(\boldsymbol{u}) \leqslant \frac{\log M}{\eta} + \frac{\eta\, C^2\, T}{2}\,. \tag{2.66}$$

En particulier, en prenant $\eta = \eta^*_{C,T} = \frac{1}{C}\sqrt{(2\log M)/T}$, la borne précédente devient $C\sqrt{2T\log M}$.

Remarque 2.31. *La proposition 2.23 et son corollaire 2.24 nécessitent une borne sur les pseudo-pertes. Une telle borne peut être garantie si la fonction de perte vérifie certaines hypothèses. Par exemple, $\mathscr{X}$ est une partie de diamètre $D > 0$ d'un espace vectoriel normé $(E, \|\cdot\|)$, et si*[30] *$\|\nabla\ell_x(x, y)\| \leqslant G$, i.e. ℓ est G-lipschitzienne en son premier argument, il vient $|\widetilde{\ell}_{i,t} - \widetilde{\ell}_{j,t}| \leqslant GD$ pour tous i, j, t, ce qui permet d'obtenir la borne de regret $R_T \leqslant GD\sqrt{\frac{T}{2}\log M}$ pour l'algorithme EG avec une calibration optimale de η.*

Remarque 2.32. *La calibration optimale de η dans le corollaire 2.24 se fait en fonction de l'horizon de temps T. Si l'on désire une borne valable pour tout T, il est possible d'avoir recours au « doubling trick » (cf. 2.2), en réinitialisant les poids et en recalibrant η à chaque puissance de 2. On montre alors, en procédant comme dans la preuve de la proposition 2.4, que cela conduit à la même borne multipliée par $2+\sqrt{2}$, soit pour tout T : $R_T \leqslant C\big(2 + \sqrt{2}\big)\sqrt{2T\log M}$.*

Ainsi, l'algorithme EG 2.5 est une alternative à la stratégie d'agrégation à poids exponentiels 2.1, facile et rapide à mettre en œuvre (la mise à jour ne nécessite que le calcul du gradient puis une normalisation, et se fait donc en $O(M)$, comme pour la stratégie 2.1) et qui, sous certaines conditions sur la fonction ℓ (gradient borné par exemple [31]), est compétitive face au meilleur choix constant de poids.

29. Par exemple, en majorant en fonction des variances des pseudo-pertes (pondérées par les $v_{i,t}$), au moyen d'une inégalité de concentration du second ordre de type Bernstein, à la manière de [CBMS07].
30. Pour la norme duale sur les formes linéaires.
31. Notons que les hypothèses de la remarque 2.31 sont satisfaites par les fonctions de perte en valeur absolue et quadratique sur $[0, 1]$, avec $D = 1$ et respectivement $G = 1$ et $G = 2$. En revanche, elles sont pas vérifiées par la perte logarithmique.

Agrégation à poids exponentiels sur le simplexe. Nous allons maintenant présenter une autre stratégie, dont nous allons montrer qu'elle offre des garanties de regret dans les contextes qui nous ont principalement intéressés jusqu'ici, à savoir le cadre exp-concave et le cadre convexe borné. Contrairement à l'algorithme EG, l'analyse ne nécessite pas d'hypothèse supplémentaire sur les données ou sur la fonction de perte, les résultats sont donc applicables sans restriction à la perte logarithmique.

Le point de vue de l'algorithme consiste à considérer chaque poids $\boldsymbol{u} \in \mathscr{P}_M$ comme un expert, dont la perte à chaque étape est $\ell_t(\boldsymbol{u})$. Ainsi, majorer la quantité $L_T - \inf_{\boldsymbol{u}\in\mathscr{P}_M} L_T(\boldsymbol{u})$ revient à contrôler le regret par rapport à tous ces « experts ». Il s'agit alors de procéder par agrégation à poids exponentiels, en partant d'une mesure sur $\mathscr{P}_M$, et en la mettant à jour en fonction des pertes de chacun des poids. Ce procédé peut être interprété très naturellement d'un point de vue bayésien [32] : en prenant pour nouvel ensemble $\mathscr{M}$ des experts l'espace $\mathscr{P}_M$, cela revient à se donner une loi a priori sur $\mathscr{M}$ et à prédire selon la loi $P(Y_t \,|\, y^{t-1})$, où P désigne la loi jointe bayésienne [33].

Remarque 2.33. *Il existe une différence importante entre l'agrégation d'un nombre fini ou dénombrable d'experts et celle d'un ensemble « continu » non dénombrable d'experts. Dans le premier cas, que l'on a étudié jusqu'ici, les prédictions de chacun des experts peuvent être absolument quelconques, et le regret par rapport à un expert est contrôlé par son poids initial.*

Dans le second cas, il n'est plus possible de procéder ainsi, car chaque expert a un poids nul. Il faut donc de la structure sur les prédictions des experts continus, notamment que des experts proches aient des pertes proches (hypothèse qui est satisfaite dans le cas présent). On procède alors en minorant le poids d'un ensemble des experts qui prédisent presque aussi bien que le meilleur expert, typiquement un voisinage de cet expert.

Dans ce qui suit, nous supposons M fini, et nous notons par commodité Δ le simplexe $\mathscr{P}_M \subset \mathbf{R}^M$, qui est compact et convexe. Notons que, pour tout t, la fonction $\ell_t : \Delta \to \mathbf{R}$ est convexe, et exp-concave (resp. bornée) si ℓ l'est.

L'algorithme d'agrégation sur le simplexe procède comme suit. Pour tout $\boldsymbol{u} \in \mathscr{P}_M$, on pose $w_t(\boldsymbol{u}) = \exp(-\eta\, L_{t-1}(\boldsymbol{u}))$, ainsi que :

$$v_t(\boldsymbol{u}) = \frac{w_t(\boldsymbol{u})}{\int_\Delta w_t(\boldsymbol{u}')d\boldsymbol{u}'}\,; \tag{2.67}$$

$$W_t = \int_\Delta w_t(\boldsymbol{u})d\boldsymbol{u}\,. \tag{2.68}$$

À l'étape t, on prédit en combinant selon la loi $v_t(\boldsymbol{u})d\boldsymbol{u}$ les prédictions des « experts » $\boldsymbol{u} \in \Delta$, celles-ci s'obtenant à partir de celles des « vrais » experts $i = 1, \ldots, M$ par $\boldsymbol{u}\cdot\boldsymbol{x}_t = \sum_{i=1}^M u_i\, x_{i,t}$. De manière équivalente, on prédit $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$, où le vecteur de poids $\boldsymbol{v}_t \in \Delta$ est donné par :

$$\boldsymbol{v}_t = \int_\Delta v_t(\boldsymbol{u})\boldsymbol{u}\, d\boldsymbol{u} = \mathbb{E}_{\boldsymbol{U}_t \sim v_t(\cdot)}[\boldsymbol{U}_t]\,. \tag{2.69}$$

32. Dans le cas de la perte logarithmique, et selon l'équivalence détaillée dans la sous-section 2.2.3.

33. Comme dans le cas d'un espace $\mathscr{M}$ discret, la stratégie d'agrégation à poids exponentiels sur le simplexe permet de prédire comme le *mélange* des stratégies de prédictions associées aux éléments $\boldsymbol{u} \in \mathscr{P}_M$, sans nécessiter la connaissance *a priori* de ces stratégies de prédiction (qui équivaut à la connaissance des stratégies de prédiction des experts de base $i = 1, \ldots, M$).

En d'autres termes, on suit la stratégie suivante :

Stratégie 2.6 (Agrégation à poids exponentiels sur le simplexe)**.**

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

Initialisation de la densité : On pose $v_1(\boldsymbol{u}) = \frac{1}{\mathrm{vol}(\Delta)}$, où $\mathrm{vol}(\Delta) = \int_\Delta d\boldsymbol{u} = \frac{1}{M!}$. On pose également $\boldsymbol{v}_1 = \mathbb{E}_{\boldsymbol{U}\sim v_1(\cdot)}[\boldsymbol{U}] = \frac{1}{M}\mathbf{1}$.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$.

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, ce qui permet de calculer les pertes $\ell_t(\boldsymbol{u}) = \ell(\boldsymbol{u} \cdot \boldsymbol{x}_t, y_t)$ pour tout $\boldsymbol{u} \in \Delta$.

Mise à jour de la densité : La densité est alors mise à jour par, pour tout $\boldsymbol{u} \in \Delta$:

$$v_{t+1}(\boldsymbol{u}) = \frac{v_t(\boldsymbol{u}) \exp(-\eta\, \ell_t(\boldsymbol{u}))}{\int_\Delta v_t(\boldsymbol{u}') \exp(-\eta\, \ell_t(\boldsymbol{u}'))\, d\boldsymbol{u}'} . \tag{2.70}$$

Calcul des poids : On en déduit enfin le poids $\boldsymbol{v}_{t+1}$ en prenant l'espérance de la loi de densité v_{t+1}, selon (2.69).

Remarque 2.34. *Notons que, d'un point de vue pratique, la mise en œuvre de la stratégie 2.6 est plus délicate que celle des stratégies précédentes. Dans certains cas, la densité (ou tout au moins son espérance) admet une expression exacte simple, ce qui permet de la déterminer facilement. Dans les autres cas, on peut procéder soit en calculant et en mettant à jour les valeurs de la densité sur une grille, soit en gardant en mémoire toutes les valeurs précédentes du signal y_s et des prédictions $x_{i,s}$ des experts, afin de disposer d'une expression exacte de $v_t(\cdot)$ que l'on intègre de manière numérique (ce qui implique un espace mémoire et un temps de mise à jour en $O(MT)$).*

Dans ce qui suit, nous laissons de côté ces questions numériques afin de nous concentrer sur l'analyse des performances de la stratégie 2.6.

Commençons par traiter le cas où ℓ est η-exp-concave. Notre approche suit de près celle de [HAK07], quoique la borne obtenue ici soit légèrement meilleure. L'idée de la preuve consiste à considérer un petit simplexe Δ_λ, obtenu en contractant le grand simplexe Δ, et centré autour du poids optimal $\boldsymbol{u}^*$. Ce simplexe doit être suffisamment petit pour que la perte maximale sur cet ensemble soit proche de la perte optimale $L_T(\boldsymbol{u}^*)$, et suffisamment grand pour avoir un poids non négligeable.

Théorème* 2.25. *Supposons la fonction de perte ℓ η-exp-concave. Alors, quelles que soient les valeurs du signal et les prédictions des experts $i = 1, \ldots, M$, la stratégie 2.6 d'agrégation à poids exponentiels sur le simplexe de paramètre η garantit la borne de regret suivante :*

$$L_T - \inf_{\boldsymbol{u}\in\mathscr{P}_M} L_T(\boldsymbol{u}) \leqslant \frac{M}{\eta}\left\{1 + \log\left(1 + \frac{T}{M}\right)\right\} . \tag{2.71}$$

Remarque 2.35. *Comparons la borne (2.71) de regret face au meilleur poids à la borne de regret face au meilleur expert dans le cas exp-concave, soit $\frac{1}{\eta}\log M$. La présente borne, en $M \log \frac{T}{M}$, exhibe une moins bonne dépendance en le nombre M d'experts (linéaire au* *lieu de logarithmique), en plus de dépendre de T (certes de manière logarithmique). Cette différence vient du fait que la borne est valable* dans le pire des cas *et pour toute fonction exp-concave ; dans ce contexte, la perte du meilleur poids peut être dramatiquement meilleure que celle du meilleur expert, comme on l'a observé dans la remarque 2.30. Il s'avère en fait qu'il n'est pas possible de faire mieux que ces dépendances en M et T dans le cas général.*

Démonstration. Commençons par majorer la perte de la stratégie 2.6 à partir des poids totaux W_t, comme dans le cas discret. Pour tout t, la fonction ℓ_t est η-exp-concave sur le simplexe (car ℓ est exp-concave) ; comme $\boldsymbol{v}_t = \int_\Delta v_t(\boldsymbol{u})\boldsymbol{u}d\boldsymbol{u}$ (cf. (2.69)), ceci implique :

$$\exp(-\eta\,\ell_t(\boldsymbol{v}_t)) \geqslant \int_\Delta v_t(\boldsymbol{u})\exp(-\eta\,\ell_t(\boldsymbol{u}))d\boldsymbol{u} = \frac{\int_\Delta w_t(\boldsymbol{u})\exp(-\eta\,\ell_t(\boldsymbol{u}))\,d\boldsymbol{u}}{\int_\Delta w_t(\boldsymbol{u}')d\boldsymbol{u}'}\,,$$

c'est-à-dire $\ell_t(\boldsymbol{v}_t) \leqslant -\frac{1}{\eta}\log\frac{W_{t+1}}{W_t}$. En sommant pour $t = 1,\ldots,T$, il vient :

$$L_T \leqslant -\frac{1}{\eta}\log\frac{W_{T+1}}{W_1} = -\frac{1}{\eta}\log\left\{\frac{1}{\mathrm{vol}(\Delta)}\int_\Delta e^{-\eta\,L_T(\boldsymbol{u})}d\boldsymbol{u}\right\}. \tag{2.72}$$

Soit maintenant $\boldsymbol{u}^* \in \Delta$ quelconque[34] ; pour $\lambda \in]0,1[$, on note $\Delta_\lambda = (1-\lambda)\boldsymbol{u}^* + \lambda\Delta$. Si $\boldsymbol{u} = (1-\lambda)\boldsymbol{u}^* + \lambda\boldsymbol{u}' \in \Delta_\lambda$, on a par exp-concavité de ℓ (donc de ℓ_t) :

$$e^{-\eta\,\ell_t((1-\lambda)\boldsymbol{u}^*+\lambda\boldsymbol{u}')} \geqslant (1-\lambda)e^{-\eta\,\ell_t(\boldsymbol{u}^*)} + \lambda e^{-\eta\,\ell_t(\boldsymbol{u}')}\,,$$

et donc : $e^{-\eta\,\ell_t(\boldsymbol{u})} \geqslant (1-\lambda)e^{-\eta\,\ell_t(\boldsymbol{u}^*)}$. En faisant le produit pour $t = 1,\ldots,T$, ceci implique, pour tout $\boldsymbol{u} \in \Delta_\lambda$:

$$e^{-\eta\,L_T(\boldsymbol{u})} \geqslant (1-\lambda)^T\,e^{-\eta\,L_T(\boldsymbol{u}^*)}\,. \tag{2.73}$$

Il vient alors :

$$\int_\Delta e^{-\eta\,L_T(\boldsymbol{u})}d\boldsymbol{u} \geqslant \int_{\Delta_\lambda} e^{-\eta\,L_T(\boldsymbol{u})}d\boldsymbol{u} \geqslant \mathrm{vol}(\Delta_\lambda)\,(1-\lambda)^T\,e^{-\eta\,L_T(\boldsymbol{u}^*)} \tag{2.74}$$

Or, pour tout $\lambda \in]0,1[$, Δ_λ s'obtient à partir de Δ par dilatation de paramètre λ (suivie d'une translation), donc $\mathrm{vol}(\Delta_\lambda) = \lambda^{M-1}\,\mathrm{vol}(\Delta) \geqslant \lambda^M\,\mathrm{vol}(\Delta)$. En injectant l'inégalité (2.74) dans la majoration (2.72), on a donc :

$$L_T \leqslant -\frac{1}{\eta}\log\left(\lambda^M(1-\lambda)^T e^{-\eta\,L_T(\boldsymbol{u}^*)}\right) = L_T(\boldsymbol{u}^*) - \frac{M}{\eta}\log\lambda - \frac{T}{\eta}\log(1-\lambda). \tag{2.75}$$

La borne (2.71) s'obtient alors en optimisant la borne (2.75) sur λ, c'est-à-dire en prenant $\lambda = \frac{M}{T+M}$ (cf. le lemme 2.26 plus bas). □

Dans la preuve, nous avons fait appel au lemme élémentaire suivant, que nous énonçons séparément car nous y ferons régulièrement appel par la suite :

34. Si ℓ est continue en son premier argument, par compacité de Δ on peut directement prendre $\boldsymbol{u}^*$ tel que $L_T(\boldsymbol{u}^*) = \inf_{\boldsymbol{u}\in\Delta} L_T(\boldsymbol{u})$; sinon, l'argument reste le même, puisque l'on va majorer $L_T - L_T(\boldsymbol{u}^*)$ pour tout $\boldsymbol{u}^* \in \Delta$.

Lemme 2.26. *Soient $A, B > 0$. Le minimum de la fonction $f :]0,1[\to \mathbf{R}$ définie par $f(\lambda) = -A\log\lambda - B\log(1-\lambda)$ est atteint pour $\lambda = \frac{A}{A+B}$, et vaut :*

$$(A+B)\,h\left(\frac{A}{A+B}\right) \leqslant A\log\frac{A+B}{A} + A\,, \tag{2.76}$$

où $h(x) = -x\log x - (1-x)\log(1-x)$ est la fonction d'entropie binaire[35] *sur $[0,1]$.*

Démonstration. La valeur de λ et celle du minimum se déduisent d'une analyse de fonction élémentaire. Quant à l'inégalité (2.76), elle découle de la majoration $B\log(\frac{A+B}{B}) = B\log(1+\frac{A}{B}) \leqslant A$ (car $\log(1+x) \leqslant x$ pour tout $x > -1$). □

Venons-en maintenant à l'étude des performances de la stratégie 2.6 lorsque la fonction de perte est convexe et bornée. L'analyse procède essentiellement comme dans le cas exp-concave, à quelques légères modifications près. Le théorème 2.27 ci-dessous correspond au théorème 3.3 de [CBL06], à ceci près que dans [CBL06] la fonction de perte est supposée bornée *et* exp-concave.

Théorème* 2.27. *Supposons la fonction de perte ℓ convexe et à valeurs dans $[0,1]$. Alors, quelles que soient les valeurs du signal et les prédictions des experts $i = 1, \ldots, M$, la stratégie 2.6 de paramètre $\eta > 0$ garantit la borne de regret suivante par rapport au meilleur poids :*

$$L_T - \inf_{\boldsymbol{u}\in\mathscr{P}_M} L_T(\boldsymbol{u}) \leqslant \frac{M}{\eta}\log\frac{e\,\eta\,T}{M} + \frac{\eta\,T}{8}\,. \tag{2.77}$$

En particulier, le choix de $\eta = \eta_T^* = 2\sqrt{M/T}$ conduit à la majoration[36]

$$L_T - \inf_{\boldsymbol{u}\in\mathscr{P}_M} L_T(\boldsymbol{u}) \leqslant \frac{\sqrt{MT}}{4}\log\left(\frac{4e^3T}{M}\right). \tag{2.78}$$

Démonstration. La preuve procède comme celle du théorème 2.25, à ceci près qu'il faut remplacer les inégalités (2.72) et (2.73), qui requièrent la propriété d'exp-concavité. Pour remplacer la majoration (2.72), on utilise comme d'habitude le lemme 2.2, qui donne :

$$L_T \leqslant -\frac{1}{\eta}\log\frac{W_{T+1}}{W_1} + \frac{\eta\,T}{8} = -\frac{1}{\eta}\log\left\{\frac{1}{\mathrm{vol}(\Delta)}\int_\Delta e^{-\eta\,L_T(\boldsymbol{u})}d\boldsymbol{u}\right\} + \frac{\eta\,T}{8}\,. \tag{2.79}$$

Quant à l'inégalité (2.73), on la remplace en notant que, par convexité de ℓ, il vient $\ell_t((1-\lambda)\boldsymbol{u}^* + \lambda\boldsymbol{u}') \leqslant (1-\lambda)\ell_t(\boldsymbol{u}^*) + \lambda\ell_t(\boldsymbol{u}') \leqslant \ell_t(\boldsymbol{u}^*) + \lambda$ (car $0 \leqslant \ell \leqslant 1$) d'où :

$$L_T(\boldsymbol{u}) \leqslant L_T(\boldsymbol{u}^*) + \lambda T \tag{2.80}$$

pour tout $\boldsymbol{u} \in \Delta_\lambda = (1-\lambda)\boldsymbol{u}^* + \lambda\Delta$. En raisonnant comme précédemment, et en utilisant les inégalités (2.79) et (2.80), on obtient la majoration $L_T - L_T(\boldsymbol{u}^*) \leqslant -\frac{M}{\eta}\log\lambda + \lambda T + \frac{\eta T}{8}$ que l'on optimise en prenant $\lambda = \frac{M}{\eta T}$, ce qui aboutit à la borne (2.77). □

35. Qui donne l'entropie d'une loi de Bernoulli de paramètre x.

36. Ce choix de η donne une majoration relativement simple. Le choix alternatif $\eta = 2\sqrt{\frac{\log(T/M)}{T/M}}$ induit une borne légèrement meilleure mais plus complexe, équivalente lorsque $T/M \to \infty$ à $\frac{1}{2}\sqrt{MT\log\frac{T}{M}}$.

Ainsi, la stratégie 2.6 d'agrégation sur le simplexe offre des garanties de regret universelles, à la fois dans le cadre exp-concave et dans le cadre convexe et borné. Cependant, si la fonction de perte satisfait les conditions de la remarque 2.31 (*i.e.* $\mathscr{X}$ est borné et ℓ admet un sous-gradient borné), l'algorithme EG (stratégie 2.5) lui est préférable, du fait de sa mise en œuvre plus simple ainsi que de la meilleure dépendance en M de sa borne de regret. Ces deux algorithmes s'étendent naturellement au cadre de l'*optimisation convexe en ligne*, dont nous discuterons brièvement en sous-section 2.6.2.

2.4 Calibration de la vitesse d'apprentissage

Dans cette section, nous allons discuter de la calibration du paramètre d'apprentissage η de l'algorithme d'agrégation à poids exponentiels dans le cadre convexe borné, dont le besoin se fait parfois ressentir en raison du caractère trop conservateur de cette stratégie.

En sous-section 2.2.1, nous avons montré que l'usage de la stratégie d'agrégation à poids exponentiels, avec une vitesse d'apprentissage η calibrée en fonction de l'horizon de temps T envisagé, permet d'obtenir un regret majoré par $\sqrt{(T/2)\log M}$. Comme on le verra dans la section 2.5 (proposition 2.32), cette borne ne peut pas être améliorée dans le pire des cas pour une fonction convexe bornée quelconque. En revanche, et c'est l'objet de cette section, il existe des stratégies dites *adaptatives*, qui calibrent η en fonction des données, afin d'obtenir des bornes de regret améliorées dans des cas favorables, tout en conservant les garanties dans le pire des cas de la calibration « conservatrice » de η du corollaire 2.3.

2.4.1 Le cas où le meilleur expert prédit bien

Dans cette sous-section, nous allons voir qu'il est possible d'obtenir de meilleures bornes de regret lorsque l'on sait à l'avance que le meilleur expert aura une faible perte. Nous verrons même dans un second temps qu'il est possible d'obtenir cette amélioration dans le cas où le meilleur expert prédit bien sans aucune information *a priori* sur les pertes des experts.

Pour illustrer le phénomène en jeu, considérons un cas extrême, tiré de [CBL06]. On cherche à prédire une suite de bits (*i.e.* $\mathscr{X} = \mathscr{Y} = \{0, 1\}$, et $\ell(x, y) = \mathbf{1}_{\{x \neq y\}}$) à l'aide de M experts. Notons au passage que, l'espace $\mathscr{X}$ des prédictions n'étant pas convexe, aucune des stratégies précédentes (reposant toutes sur des combinaisons convexes des prédictions des experts) ne s'applique ; on peut en fait montrer que, même lorsque $M = 2$, le regret dans le pire des cas de n'importe quel algorithme est linéaire [37]. Supposons maintenant que l'apprenant sache *a priori* que le meilleur expert ne fait aucune erreur ; dans ce cas, il peut prédire selon la stratégie *Follow the Leader* (FTL), qui revient ici à prédire comme la majorité des experts qui ne se sont jamais trompés. À chaque fois que l'expert fait une erreur de prédiction, au moins la moitié des experts « sans fautes » s'est trompée, ce qui réduit au moins de moitié le nombre de tels experts. Comme par hypothèse il y a au

37. Il suffit pour cela de considérer deux experts, le premier prédisant toujours 0 et le second toujours 1 ; alors, étant données les prédictions $x_1, \dots, x_T \in \{0, 1\}$ de l'apprenant, le signal $y_t = 1 - x_t$ donne $L_T = T$, mais $\min_{i=1,2} L_{i,T} \leqslant \frac{T}{2}$ (car $L_{1,T} + L_{2,T} = T$), donc $R_T \geqslant \frac{T}{2}$.

moins un expert qui ne se trompe jamais, le nombre L_T d'erreurs de l'apprenant vérifie : $M/2^{L_T} \leqslant 1$, *i.e.* $L_T \leqslant \lfloor \log_2 M \rfloor$ pour tout $T \geqslant 1$.

L'exemple précédent suggère que, lorsque l'on sait à l'avance que l'un des experts prédira bien, il est avantageux d'apprendre plus agressivement, *i.e.* de choisir une vitesse d'apprentissage η plus élevée que celle suggérée par la calibration conservatrice dans le pire des cas du corollaire 2.3. Le principal outil de la preuve est le résultat suivant :

Proposition 2.28. *Supposons la fonction de perte convexe et à valeurs dans* $[0,1]$. *Alors, quelles que soient les valeurs du signal et les prédictions des experts, le regret de la stratégie 2.1 de paramètre* $\eta > 0$ *vérifie :*

$$L_T \leqslant \frac{\eta L_T^* + \log M}{1 - e^{-\eta}}, \tag{2.81}$$

où $L_T^* = \min_{i=1,\dots,M} L_{i,T}$.

Démonstration. Comme nous l'avons montré dans l'exemple 2.6 (en utilisant l'inégalité de concentration A.2), si ℓ est convexe et à valeurs dans $[0,1]$, la fonction de combinaison convexe est $(\frac{\eta}{1-e^{-\eta}}, \eta)$-réalisable pour ℓ. On en déduit l'inégalité (2.81) en appliquant le théorème 2.11. □

On déduit de la proposition 2.28 la borne de regret suivante, lorsque l'on connaît *a priori* une borne sur la perte du meilleur expert, et que l'on calibre la vitesse d'apprentissage η en fonction de cette borne (cf. [CBL06], corollaire 2.4) :

Corollaire 2.29. *Supposons que la perte* L_T^* *du meilleur expert au temps* $T \geqslant 1$ *est inférieure à* $L_0 > 0$. *Alors, sous les hypothèses de la proposition 2.28, le regret de la stratégie 2.1 de paramètre* $\eta = \eta_{L_0}^* = \log(1 + \sqrt{(2 \log M)/L_0})$ *est majoré de la façon suivante :*

$$L_T - L_T^* \leqslant \sqrt{2 L_0 \log M} + \log M\,. \tag{2.82}$$

Démonstration. Par la proposition 2.28, on a : $L_T - L_T^* \leqslant (\frac{\eta}{1-e^{-\eta}} - 1) L_T^* + \frac{\log M}{1-e^{-\eta}}$, donc par l'hypothèse $L_T^* \leqslant L_0$ et en utilisant l'inégalité $\eta \leqslant \operatorname{sh}(\eta) = \frac{e^{\eta}-e^{-\eta}}{2}$:

$$L_T - L_T^* \leqslant \left(\frac{e^{\eta} - e^{-\eta}}{2(1-e^{-\eta})} - 1 \right) L_0 + \frac{\log M}{1 - e^{-\eta}} = \frac{e^{\eta} - 1}{2} L_0 + \frac{e^{\eta}}{e^{\eta} - 1} \log M\,,$$

ce qui, en remplaçant par la valeur de η, donne précisément l'inégalité (2.82). □

Remarque 2.36. *Arrêtons-nous un instant sur la borne (2.82). En ignorant le terme* $\log M$, *qui est inévitable (voir la proposition 2.31 plus bas), le terme principal de ce majorant est en* $\sqrt{L_0 \log M}$, *à contraster avec la perte en* $\sqrt{T \log M}$ *pour la calibration dans le pire des cas (corollaire 2.12). Cette borne ne dépend pas de* T, *mais seulement (d'une borne sur) la perte* L_T^* *du meilleur expert au temps* T. *Il s'agit d'une amélioration remarquable : en effet, si* $L_T^* \ll T$, *cette borne est nettement inférieure à la borne dans le pire des cas ; en outre, comme on a toujours* $L_T^* \leqslant T$, *le choix pessimiste* $L_0 = T$ *donne une borne qui n'est jamais pire (à une constante près) que la borne du corollaire 2.12.*

Une limitation importante de l'approche du corollaire 2.29 est qu'elle nécessite la connaissance *a priori* d'une borne sur la perte du meilleur expert. Il est néanmoins possible de s'affranchir de cette limitation, en procédant de manière analogue à ce que l'on a fait pour obtenir des bornes « anytime » dans la sous-section 2.2.1.

L'idée est d'utiliser une version du « doubling trick », mais définie en termes des pertes. Plus précisément, à chaque fois que la perte du meilleur expert dépasse une puissance de 2, on réinitialise les compteurs des pertes des experts ainsi que leurs poids, puis l'on calibre η en fonction de la perte L_0 valant la puissance de 2 suivante.

Stratégie 2.7 (Agrégation à poids exponentiels avec le « doubling trick » sur les pertes)**.** La stratégie, qui ne dépend d'aucun paramètre, procède comme suit :

Initialisation : On pose $q_1 = 0$, $\eta = \eta^*(1)$ [38], $L'_{i,0} = 0$ et $v_{i,1} = \frac{1}{M}$ pour $i = 1, \ldots, M$.

Pour tout $t = 1, 2, \ldots$ on effectue les actions suivantes à l'étape t :

Prédiction : On observe les prédictions $x_{i,t} \in \mathscr{X}$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot x_t$. La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes ℓ_t et $\ell_{i,t}$ pour $i = 1, \ldots, M$.

Mise à jour : Si $L'_{i,t-1} + \ell_{i,t} > 2^{q_t}$ pour tout $i = 1, \ldots, M$, on pose $q_{t+1} = q_t + 1$, $\eta_{t+1} = \eta^*(2^{q_{t+1}})$, ainsi que $L'_{i,t+1} = 0$ et $v_{i,t} = \frac{1}{M}$ pour tout $i = 1, \ldots, M$. Sinon, on pose $q_{t+1} = q_t$, $\eta_{t+1} = \eta_t$, $L'_{i,t} = L'_{i,t-1} + \ell_{i,t}$ et :

$$v_{i,t+1} = \frac{v_{i,t}\, e^{-\eta_{t+1}\ell_{i,t}}}{\sum_{j=1}^{M} v_{j,t}\, e^{-\eta_{t+1}\ell_{j,t}}} = \frac{e^{-\eta_{t+1}L'_{i,t}}}{\sum_{j=1}^{M} e^{-\eta_{t+1}L'_{j,t}}} . \tag{2.83}$$

En majorant la perte sur chaque période entre deux réinitialisations par le corollaire 2.29 (en effet, par construction, la perte cumulée du meilleur expert sur la période q est majorée par $2^q + 1$), et en procédant comme dans la preuve de la proposition 2.4, on obtient la borne suivante sur le regret de la stratégie 2.7 : pour tout $T \geqslant 1$,

$$L_T - L_T^* \leqslant (2+\sqrt{2})\sqrt{2L_T^* \log M} + \log_2(1 + L_T^*) \log M + C \log M , \tag{2.84}$$

C étant une constante absolue déterminée par l'analyse. Notons que cette borne de terme dominant $(2+\sqrt{2})\sqrt{2L_T^* \log M}$, qui a été obtenue sans aucune connaissance a priori, est à la fois comparable dans le pire des cas aux bornes de la sous-section 2.2.3, tout en étant bien meilleure lorsque l'un des experts prédit bien.

Remarque 2.37. *Il est possible d'améliorer la constante $(2+\sqrt{2})\sqrt{2} \simeq 4.83$ du facteur dominant de la borne (2.84), en ayant recours à une variation de la stratégie 2.7. En effet, il est montré dans [CBFH+97] (théorème 4.6.3) que le choix d'une perte « pallier » pour la période q de $a\,c^q$ (au lieu de 2^q), avec $c = \frac{3+\sqrt{5}}{2}$ et a suffisamment grand, permet d'obtenir une borne de regret en $(3.34 + o(1))\sqrt{L_T^* \log M}$.*

Remarque 2.38. *Comme dans la section 2.2.1, une autre façon de procéder consiste à considérer une vitesse d'apprentissage η_t variant à chaque étape, plutôt que constante sur des périodes de temps. Dans ce cas, les poids à l'étape t sont donnés par la formule :*

$$v_{i,t} = \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^{M} e^{-\eta_t L_{j,t-1}}} . \tag{2.85}$$

38. Où l'on note, pour tout $L_0 > 0$, $\eta^*(L_0)$ la constante définie dans le corollaire 2.29.

Le choix du paramètre $\eta_t = \log(1 + \sqrt{(2\log M)/L^*_{t-1}})$ conduit alors (cf. [CBL06]) à une majoration de la forme :

$$L_T - L^*_T \leqslant 2\sqrt{2L^*_T \log M} + c\log M\,. \tag{2.86}$$

2.4.2 Calibration adaptative de la vitesse d'apprentissage

Plus généralement, il existe des stratégies plus évoluées de calibration adaptative de la vitesse d'apprentissage en fonction des données. Dans cette sous-section, nous effectuons un rapide survol des idées et méthodes dans ce domaine. Pour une discussion plus approfondie, on se référera par exemple à l'un des articles [dRvEGK14, KvEG14].

Dans de nombreux contextes, les méthodes dans le pire des cas de la section 2.2.1, voire même les approches de la sous-section 2.4.1 précédente qui s'adaptent à la performance du meilleur expert, conduisent à des choix trop conservateurs de la vitesse d'apprentissage. À l'inverse, la stratégie FTL, qui consiste à prédire comme le meilleur expert (ce qui revient à prendre $\eta = \infty$), donne de bons résultats dans de nombreux cas raisonnablement favorables, bien que ses performances dans le pire des cas soient désastreuses [39]. Cela invite à formuler des stratégies plus évoluées, qui choisissent lorsque les données le permettent des vitesses d'apprentissage plus agressives.

D'un point de vue technique, les bornes de regret reposent sur des majorations de la quantité $\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t = \sum_{i=1}^M v_{i,t}\,\ell_{i,t}$ (qui domine clairement la perte de la prédiction $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$ dans le cas convexe) en fonction de la quantité [40] $m_t^\eta = -\frac{1}{\eta}\log\sum_{i=1}^M v_{i,t}\,e^{-\eta\,\ell_{i,t}}$. Dans l'analyse du pire des cas (sous-section 2.2.1), nous nous sommes appuyés sur la borne $\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t \leqslant m_t^\eta + \frac{\eta}{8}$, qui découle du lemme de Hoeffding 2.2, et qui mène à la calibration conservatrice $\eta = \sqrt{(8\log M)/T}$ (cf. corollaire 2.3). Dans le cas où le meilleur expert admet une faible perte L^*_T, on a utilisé la borne $\boldsymbol{v}_t \cdot \ell_t \leqslant \frac{\eta}{1-e^{-\eta}}\,m_t^\eta$, qui résulte du lemme A.2, et conduit à la calibration plus agressive $\eta = \log(1 + \sqrt{(2\log M)/L^*_T}) \approx \sqrt{(2\log M)/L^*_T}$ (cf. corollaire 2.29). Dans ces deux cas, lorsque la quantité qui dirige l'optimisation de la vitesse d'apprentissage (l'horizon de temps ou la perte du meilleur expert) est inconnue, il est possible d'avoir recours au « doubling trick », qui consiste à anticiper une valeur de cet objectif et à calibrer η en fonction de cette valeur, puis à doubler cet objectif lorsqu'il est atteint (ce qui arrive suffisamment rarement). Comme nous l'avons signalé, il est également possible d'avoir recours à une vitesse d'apprentissage η_t qui varie à chaque étape, en fonction de la valeur présente de la quantité d'intérêt.

C'est cette approche que suit l'article [CBMS07], qui borne la différence $\delta_t^{\eta_t} = \boldsymbol{v}_t \cdot \ell_t - m_t^{\eta_t}$ en fonction de la variance V'_t de la variable aléatoire valant $\ell_{i,t}$ avec probabilité $v_{i,t}$, au moyen d'une inégalité de concentration de second ordre de type Bernstein. En notant $V_t = \sum_{s=1}^t V'_t$, l'optimisation par rapport à cette variance conduit au choix d'une vitesse variable $\eta_t = c\sqrt{\log M/V_{t-1}}$, et à une borne de regret de la forme :

$$L_T - L^*_T \leqslant 4\sqrt{V_T \log M} + C\log M\,. \tag{2.87}$$

Cette borne de regret est très forte, particulièrement lorsque les poids tendent à se concentrer sur un seul expert qui prédit mieux que les autres, car alors la variance V_T devient

39. Comme on l'a observé en sous-section 2.2.1.

40. Qui apparaît comme une majoration de la perte dans le cas exp-concave.

faible. Cette borne implique également la majoration suivante, plus lisible quoique généralement plus faible :

$$L_T - L_T^* \leqslant 4\sqrt{\frac{L_T^*(T - L_T^*)}{T} \log M} + C \log M \,. \tag{2.88}$$

En particulier, ce choix plus agressif de η_t conduit toujours à de meilleures bornes (éventuellement à un facteur constant près) que la calibration en fonction du temps ou de la perte du meilleur expert.

Dans l'article [dRvEGK14], l'idée de calibrer la vitesse d'apprentissage à partir d'une borne sur la quantité $\delta_t^{\eta_t}$ est poussée jusqu'à l'extrême. En effet, la stratégie AdaHedge consiste à optimiser le paramètre η_t directement à partir de la quantité $\Delta_{t-1} = \sum_{s=1}^{t-1} \delta_s^{\eta_s}$. Plus précisément, le regret admet la décomposition suivante, en notant $M_t = \sum_{s=1}^{t} m_s^{\eta_s}$:

$$R_T \leqslant (M_T - L_T^*) + \Delta_T \,, \tag{2.89}$$

le terme $M_T - L_T^*$ étant décroissant en les η_t (lorsque $\eta_t = \eta$ est constant, ce terme vaut $-\frac{1}{\eta} \log \sum_{i=1}^{M} e^{-\eta L_{i,T}} \leqslant L_T^* + \frac{1}{\eta} \log M$), tandis que le terme Δ_T est croissant en les η_t. L'algorithme AdaHedge consiste alors à choisir η_t de manière à égaliser approximativement ces deux termes, en prenant $\eta_t = \frac{\log M}{\Delta_{t-1}}$. Cela conduit à une analyse relativement limpide, qui mène aux bornes (2.88) et (2.89) mais avec une constante améliorée (2 au lieu de 4).

La stratégie AdaHedge effectue donc une calibration agressive de la vitesse d'apprentissage, tout en étant robuste aux signaux adversariaux. Cependant, même cette stratégie peut mener à des choix trop conservateurs de η, dans certaines situations où l'algorithme FTL prédit bien mieux. Pour pallier cette faiblesse, [dRvEGK14] introduit un autre algorithme, appelé FlipFlop, qui alterne entre des choix de η_t valant ∞ (ce qui correspond à FTL) et des valeurs plus sûres, proches de celles de l'algorithme AdaHedge. Cette approche aboutit à une stratégie qui est à la fois compétitive (à un facteur constant près) avec la stratégie FTL et avec la stratégie AdaHedge.

Mentionnons enfin que des stratégies plus récentes, notamment la stratégie LLR (pour « Learning the Learning Rate ») de [KvEG14], permettent d'obtenir des bornes de regret compétitives non seulement avec la stratégie robuste AdaHedge et la stratégie agressive FTL, mais aussi avec tout un intervalle de calibrations intermédiaires de la vitesse d'apprentissage.

2.5 Optimalité des bornes de regret

Dans la section 2.2, nous avons défini la stratégie d'agrégation à poids exponentiels dans les cadres convexe borné, exp-concave et réalisable, dans lesquels nous avons obtenu des garanties de regret indépendantes des valeurs du signal et des prédictions des experts. Dans cette section, nous mentionnons sans preuve un certain nombre de résultats qui montrent que ces bornes de regret sont en fait optimales dans le pire des cas.

Afin de formuler des bornes inférieures sur la performance de n'importe quel algorithme, introduisons la définition suivante :

Définition 2.30 (Regret minimax)**.** Soit $\ell : \mathscr{X} \times \mathscr{Y} \to \mathbf{R}$ une fonction de perte, et $M, T \geqslant 1$. Le *regret minimax* à horizon de temps T avec M experts est le regret minimal dans le pire des cas de n'importe quelle stratégie d'agrégation, soit :

$$V_T^{(M)} = \inf_{S \in \mathscr{S}_T^{(M)}} \sup_{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) \in (\mathscr{X}^M)^T} \sup_{(y_1, \ldots, y_T) \in \mathscr{Y}^T} \left(L_T(S) - \min_{1 \leqslant i \leqslant M} L_{i,T} \right), \tag{2.90}$$

où $\mathscr{S}_T^{(M)} = \prod_{t=1}^T \mathscr{X}^{(\mathscr{Y}^{t-1} \times \mathscr{X}^{Mt})}$ désigne l'ensemble des stratégies d'agrégation de M experts $S = (S_t : \mathscr{Y}^{t-1} \times \mathscr{X}^{Mt} \to \mathscr{X})_{1 \leqslant t \leqslant T}$ à horizon de temps T. De manière équivalente, le regret peut être défini par l'équation :

$$V_T^{(M)} = \left\langle\!\!\left\langle \sup_{\boldsymbol{x}_t \in \mathscr{X}^M} \inf_{x_t \in \mathscr{X}} \sup_{y_t \in \mathscr{Y}} \right\rangle\!\!\right\rangle_{t=1}^T \left(\sum_{t=1}^T \ell(x_t, y_t) - \min_{1 \leqslant i \leqslant M} \sum_{i=1}^M \ell(x_{i,t}, y_t) \right), \tag{2.91}$$

où $\langle\!\langle \cdot \rangle\!\rangle_{t=1}^T$ désigne l'application successive des opérateurs de $t = 1$ à T (de gauche à droite).

Remarque 2.39. *Ainsi, une borne supérieure $\overline{R}$ sur le regret minimax signifie qu'il existe une stratégie d'agrégation dont le regret est inférieur à $\overline{R}$ quelles que soient les valeurs du signal et les prédictions des experts. À l'inverse, une borne inférieure $\underline{R}$ sur le regret minimax signifie que, pour toute stratégie d'agrégation, il existe une suite de prédictions des experts et de valeurs du signal pour laquelle le regret de cette stratégie dépasse $\underline{R}$.*

Les résultats de la section 2.2 fournissent des bornes supérieures sur le regret minimax dans les cas convexe borné et exp-concave/réalisable, ainsi que des stratégies qui réalisent ces bornes. Nous allons maintenant donner des bornes inférieures correspondantes.

Commençons par un résultat général (cf. [CBL06], théorème 3.6), valable pour n'importe quelle fonction de perte, qui capture la dépendance minimale du regret en le nombre M d'experts :

Proposition 2.31. *Soit ℓ une fonction de perte quelconque. Pour tout $M \geqslant 2$ et tout $T \geqslant 1$, on a : $V_{T\lfloor \log_2 M \rfloor}^{(M)} \geqslant \lfloor \log_2 M \rfloor V_T^{(M)}$.*

Remarque 2.40. *Notons que, lorsque la fonction ℓ est η-mélangeable, le corollaire 2.12 donne une borne de regret dans le pire des cas de $\frac{1}{\eta} \log M$. La borne inférieure de la proposition 2.31 est du même ordre de grandeur, quoiqu'elle ne fournisse pas de constante.*

Venons-en maintenant au cas des fonctions convexes bornées. Dans ce contexte, nous avons obtenu (cf. corollaire 2.3) la borne de regret $\sqrt{\frac{T}{2} \log M}$ pour la stratégie d'agrégation à poids exponentiels 2.1 avec un paramètre d'apprentissage calibré en fonction de T. Il s'avère que, pour la fonction de perte en valeur absolue, cette borne est optimale, tant dans l'ordre de grandeur de la dépendance en T que dans la constante multiplicative :

Proposition 2.32. *Considérons la fonction de perte ℓ définie par $\ell(x, y) = |x - y|$ pour $x \in [0, 1]$ et $y \in \{0, 1\}$. Pour tout $M \geqslant 2$, on a :*

$$\liminf_{T \to \infty} \frac{V_T^{(M)}}{\sqrt{(T/2) \log M}} \geqslant 1\,. \tag{2.92}$$

Pour prouver cette proposition (cf. [CBL06], théorème 3.7), l'idée consiste à minorer dans la définition (2.90) de $V_T^{(M)}$ le sup sur $y^T \in \mathscr{Y}^T$ par une espérance sur $Y^T = (Y_1, \ldots, Y_T)$, où les Y_t sont des variables aléatoires de Bernoulli i.i.d. de paramètre $\frac{1}{2}$. La dépendance en $\sqrt{T}$ apparaît alors à cause du théorème central limite, que l'on utilise après quelques transformations. Les quantités manipulées par la preuve, appelées *moyennes de Rademacher*, jouent un rôle important dans la détermination du regret minimax, que ce soit face à des experts génériques ou à une famille particulière d'experts (cf. [CBL06, RS15]).

Enfin, nous avons également traité le cas des fonctions exp-concaves et mélangeables dans la sous-section 2.2.2. Le théorème 2.11 affirme que, pour une fonction de perte (c, η)-réalisable (définition 2.7), l'algorithme d'agrégation 2.4 (avec une fonction de combinaison **pred** idoine) garantit une perte $L_T \leqslant cL_T^* + \frac{c}{\eta} \log M$. En particulier, on peut prendre $c = c(\eta)$ minimale (définition 2.9). Les constantes multiplicatives respectives $c(\eta)$ et $\frac{c(\eta)}{\eta}$ de L_T^* et $\log M$ sont en fait optimales, comme l'affirme le théorème suivant (cf. [Vov98]) :

Théorème 2.33. *Fixons $M \geqslant 2$ et $T \geqslant 1$. Pour tout $(a, b) \in \mathbf{R}_+^2$, on définit le jeu $\mathscr{G}(a, b)$ suivant entre l'apprenant et l'environnement : les deux joueurs agissent successivement comme dans le problème 1 jusqu'au temps T. Si $L_T \leqslant c \min_{1 \leqslant i \leqslant M} L_{i,T} + a \log M$, l'apprenant est déclaré vainqueur ; sinon, l'environnement l'emporte. On a alors les résultats suivants*[41] *:*

1. *Pour tout couple $(a, b) \in \mathbf{R}_+^2$, le jeu $\mathscr{G}(a, b)$ est déterminé : il existe une stratégie gagnante soit pour l'apprenant, soit pour l'environnement, qui l'emporte quelle que soit la stratégie de l'adversaire.*
2. *L'apprenant admet une stratégie gagnante pour le jeu $\mathscr{G}(a, b)$ si et seulement s'il existe $\eta \geqslant 0$ tel que $c(\eta) \leqslant c$ et $c(\eta)/\eta \leqslant a$.*

2.6 Formulations alternatives de la prédiction séquentielle

Dans cette section, nous montrons que les méthodes que nous avons développées pour résoudre le problème 1 de la prédiction à l'aide d'experts, ainsi que les preuves correspondantes, se formulent naturellement dans des cadres plus simples et plus généraux.

En effet, toutes les stratégies que nous avons décrites procèdent par combinaison (en général convexe) des prédictions des experts. Plus précisément, ces stratégies consistent à choisir un vecteur de poids $\boldsymbol{v}_t$ au début de l'étape t, à partir des valeurs précédentes du signal et des prédictions des experts, puis à combiner les prédictions présentes des experts selon ces poids, en prenant typiquement $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$. Les problèmes que nous allons décrire, le *problème de Hedge* et l'*optimisation convexe en ligne*, font abstraction de l'espace $\mathscr{X}$ des prédictions (puisque celles-ci sont toujours utilisées de la même manière), et mettent l'accent sur le choix du vecteur de poids à chaque étape.

41. Valides sous des conditions très générales sur $\mathscr{X}$, $\mathscr{Y}$ et ℓ, satisfaites dans tous les exemples de ce texte.

2.6.1 Le problème de Hedge

Dans le cadre d'une fonction de perte convexe et bornée, l'approche pour majorer le regret par rapport au meilleur expert part [42] de l'inégalité de convexité :

$$\ell(\boldsymbol{v}_t \cdot \boldsymbol{x}_t, y_t) \leqslant \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t \tag{2.93}$$

où $\ell_t = (\ell_{i,t})_{1 \leqslant i \leqslant M} = (\ell(x_{i,t}, y_t))_{1 \leqslant i \leqslant M} \in [0,1]^M$ désigne le vecteur des pertes des experts, révélé après coup (une fois que la valeur y_t du signal a été révélée). La preuve consiste alors à majorer la quantité $\sum_{t=1}^T \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^T \ell_{i,t}$. De plus, dans la stratégie 2.1 d'agrégation à poids exponentiels, les poids $\boldsymbol{v}_t \in \mathscr{P}_M$ choisis ne dépendent que des pertes précédentes $(\ell_{i,s})_{1 \leqslant i \leqslant M,\, 1 \leqslant s \leqslant t-1}$, et pas des $x_{i,s}$ ou des y_s (équation (2.3)). Cela suggère la formulation suivante, appelée *problème de Hedge* ou *optimisation linéaire en ligne* :

Problème 2 (Hedge)**.** Le problème se présente comme un jeu entre l'apprenant et l'environnement. À chaque étape $t = 1, \ldots, T$, les actions suivantes ont lieu :

1. L'apprenant choisit un vecteur de poids $\boldsymbol{v}_t \in \mathscr{P}_M$.
2. L'environnement détermine ensuite un vecteur de pertes $\boldsymbol{\ell}_t \in [0,1]^M$.

Le but de l'apprenant consiste alors à minimiser le *regret* :

$$R_T = \sum_{t=1}^T \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^T \ell_{i,t} \,. \tag{2.94}$$

Dans la preuve de la borne du théorème 2.1 pour le choix des poids (2.2), on a majoré la quantité (2.94) en ne faisant appel qu'au caractère borné des $\ell_{i,t}$. Cette borne se transpose donc au cadre de Hedge :

Proposition 2.34. *Si $\eta > 0$, le choix des poids définis par $\boldsymbol{v}_1 = \frac{1}{M}\mathbf{1}$ et, pour tout t :*

$$v_{i,t} = \frac{v_{i,t-1}\, e^{-\eta\, \ell_{i,t-1}}}{\sum_{j=1}^M v_{j,t-1}\, e^{-\eta\, \ell_{j,t-1}}} \tag{2.95}$$

conduit à la borne de regret

$$\sum_{t=1}^T \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^T \ell_{i,t} \leqslant \frac{1}{\eta} \log M + \frac{\eta T}{8} \,. \tag{2.96}$$

Les autres bornes de regret dans le cadre convexe borné se transposent également au cadre de Hedge, avec une preuve rigoureusement identique. Par exemple, la majoration de la proposition 2.22 devient [43], pour tout $\boldsymbol{u} \in \mathscr{P}_M$ et tout $\boldsymbol{v}_1 \in \mathscr{P}_M$:

$$\sum_{t=1}^T \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \sum_{t=1}^T \boldsymbol{u} \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} \Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}_1) + \frac{\eta T}{8} \,, \tag{2.97}$$

lorsque les poids $\boldsymbol{v}_t$ sont mis à jour selon l'équation (2.95).

De plus, toute la discussion de la section 2.4 sur la calibration de η en fonction des données se formule naturellement dans ce contexte, comme le font d'ailleurs généralement les articles cités [CBMS07, dRvEGK14, KvEG14].

42. Voir en effet la preuve du théorème 2.1, qui commence par appliquer l'inégalité (2.7).
43. Comme on le montre en appliquant directement l'inégalité (2.54) indiquée dans la remarque 2.26.

2.6.2 L'optimisation convexe en ligne

Il arrive cependant que la fonction de perte ℓ satisfasse des conditions de convexité plus fortes ; dans ce cas, il se peut que la majoration (2.93), qui n'exploite pas ces propriétés supplémentaires, ne soit pas assez fine. C'est notamment le cas lorsque la fonction ℓ est exp-concave (voir la sous-section 2.2.2), et pas nécessairement bornée. Dans ces conditions, il est avantageux de remplacer la majoration (2.93) par l'inégalité plus forte :

$$\ell(\boldsymbol{v}_t \cdot \boldsymbol{x}_t, y_t) \leqslant -\frac{1}{\eta} \log \left(\boldsymbol{v}_t \cdot e^{-\eta \boldsymbol{\ell}_t} \right) . \tag{2.98}$$

Nous sommes donc amenés à considérer un cadre plus riche que le problème de Hedge, qui puisse prendre en compte les propriétés de la fonction de perte, tout en mettant l'accent sur le choix des poids plutôt que sur les prédictions $x \in \mathscr{X}$.

L'idée repose sur l'observation suivante : si, au temps t, l'apprenant choisit de combiner les prédictions des experts selon $\boldsymbol{v}_t \in \mathscr{P}_M$, sa perte sera $\ell_t(\boldsymbol{v}_t)$, où la fonction $\ell_t : \mathscr{P}_M \to \mathbf{R}$ est définie par :

$$\ell_t(\boldsymbol{v}_t) = \ell(\boldsymbol{v}_t \cdot \boldsymbol{x}_t, y_t) . \tag{2.99}$$

Comme nous l'avons déjà noté en sous-section 2.3.3, la fonction ℓ_t hérite des propriétés de la fonction de perte : elle est convexe (resp. exp-concave, bornée) sur $\mathscr{P}_M$ si ℓ l'est (en son premier argument $x \in \mathscr{X}$). De plus, la fonction ℓ_t est révélée après le choix des poids $\boldsymbol{v}_t$, au moment où y_t est déterminé. Le regret de l'apprenant s'écrit alors :

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{v}_t) - \min_{1 \leqslant i \leqslant M} \ell_t(\delta_i) . \tag{2.100}$$

Cela suggère la reformulation suivante du problème de la prédiction à l'aide d'experts, où l'on a éliminé l'espace $\mathscr{X}$ des prédictions et celui $\mathscr{Y}$ du signal, l'*optimisation convexe en ligne*[44] :

Problème 3 (Optimisation convexe en ligne, version faible)**.** Le problème s'énonce comme un jeu entre l'apprenant et l'environnement. À chaque étape $t = 1, \ldots, T$, les actions suivantes ont lieu :

1. L'apprenant choisit un vecteur de poids $\boldsymbol{v}_t \in \mathscr{P}_M$.
2. L'environnement détermine ensuite une fonction convexe $\ell_t \in \mathscr{L}$, où $\mathscr{L}$ est un ensemble de fonctions convexes $\mathscr{P}_M \to \mathbf{R}$.

Le but de l'apprenant consiste alors à minimiser la quantité[45]

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{v}_t) - \sum_{t=1}^{T} \boldsymbol{u} \cdot \boldsymbol{\ell}_t . \tag{2.101}$$

pour tout $\boldsymbol{u} \in \mathscr{P}_M$, où l'on a noté $\boldsymbol{\ell}_t = (\ell_t(\delta_i))_{1 \leqslant i \leqslant M}$.

44. Le problème 3 correspond à une version plus faible de l'optimisation convexe en ligne sur le simplexe ; nous décrivons la version « forte » de ce problème dans ce qui suit (problème 4).

45. Notons que ce dernier objectif équivaut à contrôler le regret (2.100).

Il est alors possible de définir la stratégie d'agrégation à poids exponentiels de paramètre $\eta > 0$ dans ce cadre, par la relation de récurrence (2.95). Si l'on prend pour $\mathscr{L}$ l'ensemble des fonctions convexes bornées, les bornes de la sous-section précédente s'appliquent car $\ell_t(\boldsymbol{v}_t) \leqslant \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t$ par convexité, avec $\boldsymbol{\ell}_t \in [0,1]^M$. Si maintenant $\mathscr{L}$ est un ensemble de fonctions η-exp-concaves, la majoration (2.98) permet d'adapter toutes les bornes démontrées dans la sous-section 2.2.2, avec des preuves identiques. En outre, en appliquant le lemme 2.20, on montre comme en 2.3.2 la borne suivante pour la stratégie d'agrégation à poids exponentiels :

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{v}_t) - \sum_{t=1}^{T} \boldsymbol{u} \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} \Delta(\boldsymbol{u} \,\|\, \boldsymbol{v}_1)\,, \tag{2.102}$$

pour tous $\boldsymbol{u}, \boldsymbol{v}_1 \in \mathscr{P}_M$. En particulier, pour $\boldsymbol{v}_1 = \frac{1}{M}\mathbf{1}$, cette borne est inférieure à $\frac{1}{\eta} \log M$.

Maintenant, comme on l'a observé en section 2.3, un objectif nettement plus ambitieux (en particulier dans le cadre exp-concave) consiste à se comparer à la perte du meilleur choix constant de poids $\boldsymbol{u} \in \mathscr{P}_M$, c'est-à-dire à contrôler la quantité $\sum_{t=1}^{T} \big(\ell_t(\boldsymbol{v}_t) - \ell_t(\boldsymbol{u})\big)$; c'est précisément ce que l'on a réalisé en 2.3.3 dans le cas de la prédiction à l'aide d'experts. Dans la formulation présente, ce problème porte le nom d'*optimisation convexe en ligne* ; pour une référence sur le sujet, on peut par exemple consulter [Haz16].

Problème 4 (Optimisation convexe en ligne)**.** Soit Δ une partie convexe et compacte d'un espace affine [46], et $\mathscr{L}$ un ensemble de fonctions convexes $\Delta \to \mathbf{R}$. À chaque étape $t = 1, \dots, T$, l'apprenant et l'environnement agissent comme suit :

1. L'apprenant choisit un élément $\boldsymbol{v}_t \in \Delta$.
2. L'environnement détermine ensuite une fonction convexe $\ell_t \in \mathscr{L}$.

Le but de l'apprenant consiste alors à contrôler la quantité suivante, appelée *regret* :

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{v}_t) - \inf_{\boldsymbol{u} \in \Delta} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})\,. \tag{2.103}$$

Pour la complétude du propos, nous mentionnons brièvement trois algorithmes d'optimisation convexe en ligne qui, sous certaines conditions sur Δ et sur les fonctions $\ell_t \in \mathscr{L}$, garantissent des bornes de regret logarithmiques en T. Les deux premiers (l'algorithme EG et l'algorithme d'agrégation à poids exponentiels sur Δ) ont déjà été introduits dans le cas de la prédiction à l'aide d'experts (en sous-section 2.3.3), et les preuves se transposent *mutatis mutandis* dans ce cadre.

L'algorithme EG. Pour cet algorithme, on suppose que Δ est le simplexe $\mathscr{P}_M$. Supposons que toutes les fonctions $\ell \in \mathscr{L}$ admettent un sous-gradient en tout point $\boldsymbol{v} \in \Delta$. Alors, si $\widetilde{\boldsymbol{\ell}}_t \in \nabla \ell_t(\boldsymbol{v}_t)$ est un sous-gradient de ℓ_t en $\boldsymbol{v}_t$, on a pour tout $\boldsymbol{u} \in \Delta$, par convexité de ℓ_t :

$$\ell_t(\boldsymbol{v}_t) - \ell_t(\boldsymbol{u}) \leqslant (\boldsymbol{v}_t - \boldsymbol{u}) \cdot \widetilde{\boldsymbol{\ell}}_t\,. \tag{2.104}$$

46. On peut typiquement prendre pour Δ le simplexe $\mathscr{P}_M$, mais le problème et les méthodes présentées restent valides pour un convexe compact quelconque.

En particulier, si l'on suppose que toutes les fonctions de $\mathscr{L}$ admettent des sous-gradients bornés, disons dans $[-B, B]^M$, la majoration (2.104) permet de se ramener au problème de Hedge. Il suffit alors de prendre $\boldsymbol{v}_t$ donné par l'algorithme de Hedge en fonction des pseudo-pertes $\widetilde{\boldsymbol{\ell}}_t$, soit :

$$v_{i,t} = \frac{v_{i,t-1} \exp(-\eta\, \widetilde{\ell}_{i,t-1})}{\sum_{j=1}^{M} v_{j,t-1} \exp(-\eta\, \widetilde{\ell}_{j,t-1})} \tag{2.105}$$

avec $\boldsymbol{v}_1 = \frac{1}{M}$ et η calibré de façon optimale en fonction de B et de T (cf. la remarque 2.6), ce qui donne une borne de regret de $B\sqrt{2T \log M}$. Pour voir que cette stratégie correspond bien à l'algorithme 2.5 étudié en sous-section 2.3.3, il suffit de noter que, si $\ell_t(\boldsymbol{v}) = \ell(\boldsymbol{v} \cdot \boldsymbol{x}_t, y_t)$, alors les pseudo-pertes $\widetilde{\ell}_{i,t} = \nabla \ell(x_t, y_t) \cdot x_{i,t}$ (avec $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$) de l'algorithme 2.5 définissent bien un sous-gradient $\widetilde{\boldsymbol{\ell}}_t \in \nabla \ell_t(\boldsymbol{v}_t)$.

Agrégation à poids exponentiels sur Δ. On suppose maintenant que Δ est un convexe compact de $\mathbf{R}^d$, que l'on peut toujours supposer d'intérieur non vide quitte à se restreindre au sous-espace affine minimal qui le contient. Pour la mesure de Lebesgue sur cet espace, notée $d\boldsymbol{u}$, on a donc $0 < \mathrm{vol}(\Delta) := \int_\Delta d\boldsymbol{u} < +\infty$. On définit alors une suite de densités de probabilité $v_t : \Delta \to \mathbf{R}^+$, en posant $v_1(\boldsymbol{u}) = \frac{1}{\mathrm{vol}(\Delta)}$ et, pour tout $t \geqslant 2$:

$$v_t(\boldsymbol{u}) = \frac{v_{t-1}(\boldsymbol{u})\, e^{-\eta\, \ell_{t-1}(\boldsymbol{u})}}{\int_\Delta v_{t-1}(\boldsymbol{u}')\, e^{-\eta\, \ell_{t-1}(\boldsymbol{u}')} d\boldsymbol{u}'} = \frac{e^{-\eta\, L_{t-1}(\boldsymbol{u})}}{\int_\Delta e^{-\eta\, L_{t-1}(\boldsymbol{u}')} d\boldsymbol{u}'} \,. \tag{2.106}$$

À l'étape t, disposant des pertes jusqu'au temps $t-1$ et donc de la densité v_t, on prédit alors l'élément $\boldsymbol{v}_t \in \Delta$ défini par :

$$\boldsymbol{v}_t = \int_\Delta v_t(\boldsymbol{u})\, \boldsymbol{u}\, d\boldsymbol{u} = \mathbb{E}_{\boldsymbol{U}_t \sim v_t(\cdot)}[\boldsymbol{U}_t] = \frac{\mathbb{E}\big[\boldsymbol{U} e^{-\eta\, L_{t-1}(\boldsymbol{U})}\big]}{\mathbb{E}\big[e^{-\eta\, L_{t-1}(\boldsymbol{U})}\big]} \,, \quad \boldsymbol{U} \sim v_1(\cdot) \,. \tag{2.107}$$

Dans le cas du simplexe, il s'agit précisément de la stratégie 2.6 formulée dans le cadre de l'optimisation convexe en ligne. Les preuves des théorèmes 2.25 et 2.27, qui n'utilisent que les propriétés des fonctions ℓ_t, restent valides dans ce cadre, et permettent d'établir les résultats suivants :

Proposition* 2.35. *Supposons que les éléments de $\mathscr{L}$ soient tous des fonctions η-exp-concaves. Alors, la stratégie d'agrégation à poids exponentiels sur Δ de paramètre η garantit la borne de regret suivante*[47] *pour tout $T \geqslant 1$:*

$$L_T - \inf_{\boldsymbol{u} \in \Delta} L_T(\boldsymbol{u}) \leqslant \frac{d}{\eta} \left\{ 1 + \log\left(1 + \frac{T}{d}\right) \right\} . \tag{2.108}$$

Proposition* 2.36. *Supposons que les éléments de $\mathscr{L}$ soient à valeurs dans $[0,1]$. Alors, la stratégie d'agrégation à poids exponentiels sur Δ de paramètre $\eta = 2\sqrt{d/T}$ garantit la borne de regret suivante :*

$$L_T - \inf_{\boldsymbol{u} \in \Delta} L_T(\boldsymbol{u}) \leqslant \frac{\sqrt{dT}}{4} \log\left(\frac{4e^3 T}{d}\right) . \tag{2.109}$$

47. Ce résultat a été démontré dans l'article [HAK07] (sous la forme $R_T \leqslant \frac{d}{\eta}\{1 + \log(1+T)\}$), qui s'intéresse à divers algorithmes en optimisation convexe en ligne admettant, sous certaines conditions, des bornes de regret logarithmiques en T.

Descente de gradient en ligne. Mentionnons enfin un dernier algorithme, introduit dans [Zin03], qui est une version en ligne de l'algorithme de descente de gradient stochastique. L'algorithme, paramétré par une suite de pas $\eta_1, \eta_2, \ldots$, consiste à commencer par un point de départ quelconque $\boldsymbol{v}_1 \in \Delta$, puis à choisir à l'étape t le point :

$$\boldsymbol{v}_t = \Pi\left(\boldsymbol{v}_{t-1} - \eta_t \nabla \ell_{t-1}(\boldsymbol{v}_{t-1})\right) , \tag{2.110}$$

où, pour $\boldsymbol{x} \in \mathbf{R}^d$, $\Pi(\boldsymbol{x}) = \mathrm{argmin}_{\boldsymbol{u}\in\Delta} \|\boldsymbol{x} - \boldsymbol{u}\|$ est la projection sur Δ[48]. On dispose alors de la borne de regret logarithmique suivante (cf. [HAK07], théorème 1), lorsque ℓ est supposée lipschitzienne et fortement convexe :

Proposition 2.37. *Supposons que toutes les fonctions $\ell_t \in \mathscr{L}$ admettent un gradient borné $\|\nabla \ell_t\| \leqslant G$, et une Hessienne minorée[49] : $\nabla^2 \ell_t \geqslant H I_d$, avec $G, H > 0$. Alors, l'algorithme de descente de gradient en ligne, avec le choix des pas $\eta_t = \frac{1}{Ht}$, garantit la borne de regret suivante pour tout $T \geqslant 1$:*

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{v}_t) - \inf_{\boldsymbol{u}\in\Delta} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leqslant \frac{G^2}{2H}(1 + \log T) \,. \tag{2.111}$$

Remarque 2.41. *Récapitulons les bornes de regret obtenues via les différents algorithmes proposés.*

La seule borne disponible dans le cas général exp-concave (qui inclut la perte quadratique, bornée, ainsi que la perte logarithmique, non bornée) est la proposition 2.35 pour l'agrégation à poids exponentiels sur Δ, qui fournit une borne logarithmique en T.

Dans le cas général convexe borné, la seule borne qui s'applique est celle de la proposition 2.36 (en $\sqrt{T}$), elle aussi obtenue pour l'agrégation à poids exponentiels sur Δ. Si l'on est de plus dans un cadre lipschitz[50] et sur le simplexe, alors l'algorithme EG, dont le regret admet également une dépendance temporelle en $\sqrt{T}$ mais une dépendance bien meilleure (logarithmique) en la dimension, est préférable.

Enfin, dans le cas lipschitz et fortement convexe[51] (qui ne couvre plus que le cas de la perte quadratique parmi les exemples mentionnés), l'algorithme de descente de gradient en ligne donne une borne logarithmique en T, ce qui le rend attractif par sa simplicité.

Remarque 2.42. *Même du point de vue de la prédiction à l'aide d'experts, le cas général où Δ n'est pas nécessairement le simplexe $\mathscr{P}_M$ des probabilités sur les experts présente un intérêt. En effet, les résultats de l'optimisation convexe en ligne peuvent également être appliqués à l'espace $\Delta = \mathscr{X}$, que l'on suppose toujours convexe et qui est souvent compact, avec $\mathscr{L} = \{\ell(\cdot, y) \mid y \in \mathscr{Y}\}$. Ceci permet de se comparer à la meilleure prédiction constante, ce qui est plus robuste si tous les experts $i = 1, \ldots, M$ prédisent très mal, bien qu'il soit plus intéressant de se comparer à la meilleure combinaison des experts lorsque l'on s'attend à ce que certains prédisent suffisamment bien.*

48. Ce minimum existe par compacité, et est unique par convexité de Δ.
49. Au sens des inégalités entre matrices symétriques, *i.e.* $A \leqslant B$ si $B - A$ est symétrique positive.
50. Ce qui est le cas pour les fonctions de perte quadratique et en valeur absolue sur un intervalle borné, et donc pour les fonctions de perte ℓ_t associées.
51. Ce qui implique aussi l'exp-concavité avec $\eta = \frac{G^2}{D}$, comme on le voit en se ramenant à la dimension 1 puis en procédant directement.

Chapitre 3

Regret par rapport à la meilleure séquence d'experts

Dans le chapitre précédent, nous avons étudié un certain nombre de stratégies compétitives avec le meilleur expert, voire la meilleure combinaison des experts. Cet objectif peut néanmoins sembler trop restrictif. En effet, nous ne nous comparons qu'au meilleur choix *constant* d'un expert (ou de poids sur ces experts), ce qui n'est intéressant que lorsque l'un des experts prédit très bien, mieux que tous les autres tout au long du processus de prédiction. Pourtant, il peut arriver que l'expert le plus performant varie au cours du temps, de sorte qu'aucun expert n'a une perte cumulée très faible. Dans ces conditions, il est bien plus intéressant de se comparer à des séquences d'experts. S'il n'est pas possible de se comparer à la meilleure séquence quelconque d'experts (qui peut changer d'expert à chaque échéance) en raison de la nature adverse du problème, nous verrons que des stratégies bien conçues garantissent un regret contrôlé face à la meilleure séquence d'experts qui ne change pas trop souvent.

Nous nous intéresserons dans un second temps à un raffinement de ce problème, proposé par Freund dans l'article [FS97]. L'idée est la suivante : lorsque l'on dispose d'un grand nombre d'experts, il arrive souvent qu'un petit nombre d'entre eux se distingue en prédisant mieux que les autres, et que le meilleur expert sur différentes périodes de temps oscille entre ces quelques experts. Dans cette situation, plutôt que de chercher à se comparer à n'importe quelle séquence d'experts ne changeant pas trop souvent, on aimerait pouvoir se comparer à la meilleure séquence parmi un petit groupe d'experts (ce groupe étant inconnu a priori, de même que les instants des changements d'experts). Là encore, nous verrons qu'il existe un algorithme *efficace* qui exhibe une borne de regret essentiellement optimale pour ce problème.

Un leitmotiv qui traverse ce chapitre est l'interprétation de nombreux algorithmes *ad hoc* en termes d'agrégation à poids exponentiels classique d'un grand nombre d'experts *structurés*, dont les prédictions sont liées et s'obtiennent à partir de celles des experts « de base ». L'avantage d'une telle reformulation, outre le gain en intelligibilité et la justification *a posteriori* de l'algorithme, est de fournir directement une borne de regret, à partir du poids initial attribué à l'expert structuré en question. Cependant, la construction et l'agrégation à poids exponentiels « naïve » d'experts structurés n'est pas à elle seule satisfaisante. En effet, si les bornes de regret du chapitre 2 montrent qu'il est possible d'agréger un grand nombre d'experts en maintenant un regret contrôlé (du fait de la dépendance *logarithmique* du regret en le nombre d'experts), la complexité algorithmique est quant à elle *linéaire* en le nombre d'experts, ce qui la rend prohibitive lorsque le nombre d'experts est trop élevé. Il est donc crucial de montrer dans chaque cas que l'agrégation naïve des experts structurés se simplifie, donnant lieu à un algorithme dont les besoins en mémoire ainsi que le temps de mise à jour sont bien moindres.

Dans la section 3.1, nous traitons le problème de la comparaison à la meilleure séquence d'experts, en nous appuyant sur l'approche fondatrice de Herbster et Warmuth [HW98]. Nous étudions ensuite en section 3.2 le problème de Freund, dont nous montrons qu'il admet une solution élégante, quoique difficile à interpréter, due à Bousquet et Warmuth [BW02]. Suivant [CBGLS12], nous revisitons ces algorithmes en section 3.3, en montrant qu'ils permettent en fait de se comparer à des suites de poids variant de façon continue [1], ainsi que de contrôler des notions plus générales de regret. Enfin, nous terminons ce chapitre en décrivant l'agrégation de *spécialistes* (section 3.4), qui fournit une interprétation « bayésienne » de l'algorithme *ad hoc* de Bousquet et Warmuth (cf. [KAW12]) ; cette approche suggère une modification de cette stratégie, proposée par [KAW12] dans le cadre bayésien de la perte logarithmique, que nous étendons au cas d'une perte quelconque. Le cadre de spécialistes permet également de faire une transition vers le chapitre suivant, qui s'appuie essentiellement sur ces idées.

3.1 Comparaison à la meilleure séquence d'experts

Dans cette section, nous étudions comme annoncé le problème consistant à se comparer à la meilleure *séquence* d'experts, plutôt qu'au meilleur choix constant d'un expert. Ce problème, appelé dans la littérature « tracking the best expert », a été traité par l'article pionnier [HW98], sur lequel nous nous appuyons.

Après avoir formulé explicitement le problème, nous introduisons l'algorithme *Fixed Share* (FS) de Herbster et Warmuth, pour lequel nous établissons une borne de regret de manière directe (sous-section 3.1.1). Nous voyons ensuite (sous-section 3.1.2) comment cet algorithme peut s'interpréter en termes d'agrégation classique d'experts « structurés » obtenus à partir des experts de base, comme l'a montré [Vov99]. Nous décrivons enfin en sous-section 3.1.3 un raffinement de l'algorithme FS, l'algorithme *Variable Share* (cf. [HW98]), qui s'applique dans le cas de pertes bornées.

3.1.1 Formulation du problème, algorithme *Fixed Share* (FS) de Herbster et Warmuth

Commençons par décrire formellement le problème. Le cadre est le même que celui du problème 1, la différence étant qu'au lieu de chercher à se comparer au meilleur expert constant, c'est-à-dire à contrôler le regret « statique » suivant :

$$L_T - \min_{1 \leqslant i \leqslant M} L_{i,T} = \sum_{t=1}^{T} \ell_t - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^{T} \ell_{i,t} \,, \tag{3.1}$$

1. Nous proposons également une variation des méthodes de l'article [CBGLS12], qui permet d'obtenir simultanément les avantages des approches de [BW02] et de [CBGLS12].

nous cherchons cette fois-ci à nous comparer à la meilleure *séquence* $i^T = (i_1, \dots, i_T)$ d'experts, c'est-à-dire à contrôler le regret « avec ruptures » (ou « shifting regret » dans la littérature anglophone) :

$$L_T - \inf_{i^T} L_T(i^T) = \sum_{t=1}^{T} \ell_t - \inf_{i^T} \sum_{t=1}^{T} \ell_{i_t,t} \,. \tag{3.2}$$

Pour que l'on ait une chance de pouvoir contrôler la quantité (3.2), il faut restreindre l'ensemble des suites $i^T \in [\![1, M]\!]^T$ auxquelles on se compare. Nous nous limitons donc aux séquences i^T qui ne changent pas trop souvent, *i.e.* dont le nombre de ruptures :

$$k(i^T) = \{1 < t \leqslant T \mid i_t \neq i_{t-1}\} \tag{3.3}$$

est limité.

Remarque 3.1. *Notons que majorer le regret* (3.2) *pour toute suite* i^T *dont le nombre de ruptures* $k(i^T)$ *est majoré par* $k_0 \geqslant 0$ *revient à se comparer au meilleur découpage de* $[\![1, T]\!]$ *en* $k_0 + 1$ *périodes de temps continues, où l'on choisit sur chaque segment le meilleur expert sur cette période.*

Les algorithmes du chapitre 2, notamment la stratégie d'agrégation à poids exponentiels 2.1 et sa variante 2.4, ne sont pas compétitifs face aux suites d'experts avec ruptures. En effet, considérons la situation suivante : on se donne deux experts, ainsi que deux périodes de temps. Sur le premier segment, le premier expert prédit bien le signal tandis que le second le prédit très mal ; pendant la seconde période, les rôles sont inversés. Alors, à la fin de la première période, l'algorithme d'agrégation attribue au second expert un poids arbitrairement faible (si sa perte était arbitrairement mauvaise), ce qui fait qu'il continue à suivre le premier expert pendant l'essentiel de la seconde période, sur laquelle il n'est donc pas compétitif face au « bon » expert, le second.

L'idée de l'algorithme *Fixed Share*, introduit par Herbster et Warmuth [HW98], consiste à garantir à chaque expert un poids minimal $\alpha > 0$, ce qui permet à des experts mauvais dans le passé mais qui commencent à mieux prédire que les autres de rattraper plus rapidement leur retard, et donc à l'algorithme de s'adapter rapidement aux changements du meilleur expert. Cela s'obtient de la façon suivante : après la mise à jour classique des poids en fonction des pertes

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^{M} v_{j,t}\, e^{-\eta\, \ell_{j,t}}} \tag{3.4}$$

(cf. (2.51)), chaque expert partage une fraction α de son poids $v_{i,t}^m$ (d'où le nom de l'algorithme), qui redistribuée équitablement entre tous les experts, soit [2] :

$$v_{i,t+1} = (1-\alpha)\, v_{i,t}^m + \frac{\alpha}{M} \,. \tag{3.5}$$

Cela conduit à l'algorithme suivant :

Stratégie 3.1 (Fixed Share)**.** L'algorithme *Fixed Share* (FS) est défini comme suit :

2. En notant que chaque expert reçoit $\frac{1}{M}\sum_{i=1}^{M}(\alpha v_{i,t}^m) = \frac{\alpha}{M}$.

Paramètres : $\eta > 0$ la vitesse d'apprentissage, et $0 < \alpha < 1$ la fraction partagée.

Initialisation : On pose $\boldsymbol{v}_1 = \frac{1}{M}$.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$, ou plus généralement $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$.

Calcul des postérieurs : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, ce qui permet de calculer les pertes $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i = 1, \ldots, M$. On en déduit alors les poids postérieurs :

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}} \,. \tag{3.6}$$

Partage des poids : Les poids sont enfin mis à jour par la formule (3.5), c'est-à-dire : $\boldsymbol{v}_{t+1} = (1-\alpha)\, \boldsymbol{v}_t^m + \alpha\, \frac{1}{M}\mathbf{1}$.

Venons-en à l'analyse du regret de l'algorithme 3.1 face à la meilleure séquence d'experts avec k ruptures (3.2). Nous suivons l'approche de l'article [BW02], qui permet en fait de se comparer à toute suite de poids $\boldsymbol{u}^T = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ avec k ruptures. Si $\boldsymbol{u}^T$ est une telle suite, on note $1 < t_1 < \cdots < t_k \leqslant T$ les instants des changements (tels que $\boldsymbol{u}_{t_j} \neq \boldsymbol{u}_{t_j - 1}$) ; on pose aussi par commodité $t_0 = 1$ et $t_{k+1} = T + 1$.

Le résultat principal, dont la preuve repose sur les lemmes 2.19 et 2.20 de la section 2.3, est la proposition suivante :

Proposition 3.1. *Supposons la fonction de perte η-mélangeable*[3]*. Alors, l'algorithme FS (stratégie 3.1) de paramètres α et η vérifie la borne de regret suivante : pour toute suite de poids $\boldsymbol{u}^T$ avec k ruptures,*

$$\sum_{t=1}^T \ell_t - \sum_{t=1}^T \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} \sum_{j=0}^k \Delta\Big(\boldsymbol{u}_{t_j} \Big\| \frac{1}{M}\mathbf{1}\Big) + \frac{1}{\eta} k \log\frac{1}{\alpha} + \frac{1}{\eta}(T-k-1)\log\frac{1}{1-\alpha} \tag{3.7}$$

$$\leqslant \frac{1}{\eta}(k+1)\log M + \frac{1}{\eta} k \log\frac{1}{\alpha} + \frac{1}{\eta}(T-k-1)\log\frac{1}{1-\alpha} \tag{3.8}$$

Démonstration. Par η-mélangeabilité de ℓ, et par le choix $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$, on a pour tout t : $\ell_t \leqslant -\frac{1}{\eta}\log \sum_{i=1}^M v_{i,t} e^{-\eta\, \ell_{i,t}}$. Par le lemme 2.20 avec $\boldsymbol{u} = \boldsymbol{u}_t$, ceci implique :

$$\ell_t - \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta}\big(\Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_t) - \Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_t^m)\big) \tag{3.9}$$

Si $t_j < t < t_{j+1}$ avec $j = 0, \ldots, k$, on majore le terme de droite en notant que $\boldsymbol{v}_t \geqslant (1-\alpha)\boldsymbol{v}_{t-1}^m$, donc par le lemme 2.19, $\Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_t) \leqslant \Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_{t-1}^m) + \log\frac{1}{1-\alpha}$. En sommant

3. À partir de maintenant, nous nous plaçons par commodité dans le cadre exp-concave (ou mélangeable si **pred** n'est pas la combinaison convexe), qui est celui qui nous intéresse le plus. Tous ces résultats s'adaptent bien entendu au cas convexe borné par le lemme de Hoeffding 2.2 ; cela pose cependant à chaque fois la question de la calibration du paramètre η, que nous préférons éviter maintenant que nous introduisons des paramètres supplémentaires.

sur $t_j \leqslant t < t_{j+1}$, il vient alors (en notant que $\boldsymbol{u}_t = \boldsymbol{u}_{t_j}$ par définition des t_j) :

$$\begin{aligned}\sum_{t_j \leqslant t < t_{j+1}} (\ell_t - \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t) &\leqslant \frac{1}{\eta}\big(\Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}_{t_j}) - \Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}^m_{t_j})\big) + \frac{1}{\eta} \sum_{t_j < t < t_{j+1}} \big(\Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}^m_{t-1}) - \Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}^m_t)\big) \\ &\quad + \frac{1}{\eta}(t_{j+1} - t_j - 1) \log \frac{1}{1-\alpha} \\ &= \frac{1}{\eta}\big(\Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}_{t_j}) - \Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}^m_{t_{j+1}-1})\big) + \frac{1}{\eta}(t_{j+1} - t_j - 1) \log \frac{1}{1-\alpha}\end{aligned}$$

On obtient alors l'inégalité (3.7) en remarquant que $\boldsymbol{v}_{t_0} = \boldsymbol{v}_1 = \frac{1}{M}\mathbf{1}$ et que, pour $j = 1, \ldots, k$, $\boldsymbol{v}_{t_j} \geqslant \alpha \frac{1}{M}\mathbf{1}$ donc $\Delta(\boldsymbol{u}_{t_j} \,\|\, \boldsymbol{v}_{t_j}) \leqslant \Delta(\boldsymbol{u}_{t_j} \,\|\, \frac{1}{M}\mathbf{1}) + \log \frac{1}{\alpha}$, puis en sommant sur $j = 0, \ldots, k$. L'inégalité (3.8) s'en déduit par la propriété 2.18. $\square$

Remarque 3.2. *La majoration (3.7), qui est à la borne (3.8) ce que la proposition 2.21 est au théorème 2.11, lui est préférable lorsque les $\boldsymbol{u}_{t_j}$ donnent suffisamment de poids à plusieurs experts.*

Notons que la borne (3.8) de la proposition 3.1 dépend du paramètre α de l'algorithme FS. Si l'on cherche à obtenir une borne intéressante du regret face aux suites de poids dont le nombre de ruptures est contrôlé, on est amené à calibrer α. Pour ce faire, il suffit d'appliquer le lemme 2.26.

Corollaire 3.2. *Reprenons les hypothèses de la proposition 3.1. Alors, pour toute suite de poids $\boldsymbol{u}^T$ avec au plus k_0 ruptures, l'algorithme FS avec $\alpha = \frac{k_0}{T-1}$ donne une perte majorée de la façon suivante*[4] *:*

$$\sum_{t=1}^T \ell_t - \sum_{t=1}^T \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta}(k_0+1)\log M + \frac{1}{\eta}(T-1)\, h\Big(\frac{k_0}{T-1}\Big) \tag{3.10}$$

$$\leqslant \frac{1}{\eta}(k_0+1)\log M + \frac{1}{\eta}k_0 \log\frac{T-1}{k_0} + \frac{1}{\eta}k_0\,. \tag{3.11}$$

Remarque 3.3. *Ainsi, l'algorithme FS correctement calibré permet de se comparer à la meilleure suite d'experts avec au plus k_0 changements, avec le regret indiqué dans le corollaire 3.2. De plus, la mise à jour est très simple et se fait en temps $O(M)$, tandis que l'algorithme ne nécessite que de stocker les M poids courants et les pertes courantes (voir la mise à jour de l'algorithme 3.1), d'où une complexité temporelle de $O(M)$ par étape et une complexité en mémoire de $O(M)$.*

Il s'avère que la borne de regret obtenue par l'algorithme FS est essentiellement optimale. En effet, le nombre de façons de choisir k instants de ruptures entre 2 et T, ainsi qu'un expert sur chaque segment, est $\binom{T-1}{k}M^{k+1}$; or, le regret minimal étant logarithmique en le nombre d'experts auxquels on se compare (cf. le théorème 2.33 avec $c(\eta) = 1$ dans le cas mélangeable), on ne doit pas espérer mieux que le regret $\frac{1}{\eta}\log\binom{T-1}{k}M^{k+1} \approx \frac{1}{\eta}(k+1)\log M + \frac{1}{\eta}k\log\frac{T-1}{k}$.

Notons enfin que cette dernière borne de regret peut être obtenue par un algorithme « direct », en appliquant la stratégie d'agrégation à poids exponentiels 2.4 à toutes les

4. Où $h(p) = -p\log p - (1-p)\log(1-p)$ désigne la fonction d'entropie binaire (cf. le lemme 2.26).

séquences avec au plus k changements. Cette stratégie est néanmoins impraticable car elle nécessite de maintenir $\binom{T-1}{k} M^{k+1} \gg M$ poids, ce qui implique une complexité en mémoire et en temps par étape du même ordre. L'algorithme FS fait aussi bien, avec une complexité moindre.

3.1.2 Interprétation bayésienne de l'algorithme FS

Nous allons maintenant montrer que l'algorithme FS, que nous avons introduit de manière *ad hoc*, peut en fait s'interpréter comme une agrégation à poids exponentiels classique sur les séquences d'experts, ce qui permet de mieux comprendre l'algorithme, de retrouver ses bornes de regret et surtout de suggérer comment il peut être modifié. Dans ce qui suit, on suppose la fonction de perte η-exp-concave, et l'on utilise la fonction **pred** de combinaison convexe.

Agrégation à poids exponentiels de suites d'experts. Puisque nous cherchons à nous comparer aux suites d'experts, une façon naturelle de procéder est de considérer des experts « structurés » associés aux suites d'experts $(i_1, i_2, \dots)$. Cela conduit à la définition suivante :

Définition 3.3. Soient $i = 1, \dots, M$ des experts de base. On leur associe des experts *structurés* dont les prédictions s'obtiennent à partir de celles des experts de base, dans les deux cadres suivants :

1. Lorsque l'horizon de temps T est fixé à l'avance, les experts correspondent aux suites $i^T = (i_1, \dots, i_T)$. Les prédictions de ces experts sont alors données par, pour tout $t \in [\![1, T]\!]$, $x_t(i^T) = x_{i_t,t}$.
2. Lorsqu'aucun horizon de temps n'est précisé, les experts correspondent aux suites infinies $i^\infty = (i_1, i_2, \dots)$ d'experts ; à nouveau, leurs prédictions sont données par $x_t(i^\infty) = x_{i_t,t}$ pour tout $t \geqslant 1$.

La définition 3.3 implique immédiatement le fait élémentaire mais important qui suit :

Fait 3.4. *Soit i^T (resp. i^∞) une suite d'experts, comme dans la définition 3.3. Alors, pour tout $t \in [\![1, T]\!]$ (resp. pour tout $t \geqslant 1$), la perte cumulée $L_t(i^T)$ (resp. $L_t(i^\infty)$) ne dépend que de i^t, et est notée $L_t(i^t)$.*

Remarque 3.4. *Notons que, dans le cas à horizon de temps T fini, le nombre d'experts structurés est fini lorsque M est lui-même fini*[5]*, égal à M^T, et dénombrable sinon, tandis que dans le cas sans horizon de temps fixé il y a une infinité non dénombrable d'experts.*

Cela suggère que, même pour M fini et T fixé, il n'est pas raisonnable d'agréger les experts structurés selon des poids uniformes (ce qui n'a aucun sens dans le cas « any-time »), puisque la borne de regret de l'algorithme d'agrégation 2.4 deviendrait alors $\frac{1}{\eta} \log M^T = \frac{T}{\eta} \log M$, qui est linéaire en T. Il n'y a rien de surprenant à cela, puisque nous avons déjà fait remarquer qu'il est impossible de se comparer à une séquence quelconque d'experts, et qu'il faut donc se limiter aux suites d'experts dont le nombre de ruptures est limité.

5. Hypothèse que nous ferons implicitement dès que l'on parlera des poids uniformes sur ces experts, mais qui peut être relaxée sans modification lorsque l'on considère des poids généraux ω_i sur les experts i, exactement comme dans le chapitre 2.

La remarque précédente conduit naturellement à introduire des poids initiaux non uniformes sur les suites d'experts, si l'on cherche à leur appliquer l'algorithme d'agrégation 2.4. Rappelons que dans la reformulation bayésienne de cet algorithme dans le cas de la perte logarithmique (sous-section 2.2.3), ces poids correspondent à une loi a priori sur les experts.

Définition 3.5 (Poids sur les suites d'experts)**.** Les poids sur les experts structurés se définissent comme une mesure de probabilité sur les séquences d'experts, ce qui donne dans les deux cas de la définition 3.3 :

1. Lorsque l'horizon de temps T est fixé, les poids correspondent à des réels $\omega(i^T) \geqslant 0$ tels que $\sum_{i^T} \omega(i^T) = 1$. On note également, pour tout $t = 1, \dots, T$, $\omega(i^t)$ la loi marginale $\sum_{i_{t+1}^T} \omega(i^t, i_{t+1}^T)$, et $\omega_i = \omega(i)$.
2. Dans le cas sans horizon de temps fini, les poids correspondent donc à une mesure μ sur l'espace mesurable[6] $(\mathscr{M}^\infty, \mathscr{B}^\infty)$. On note alors, pour tout $t \geqslant 1$ et toute séquence finie i^t, $\omega_\mu(i^t) = \mu(\{i^t\} \times \mathscr{Y}_{t+1}^\infty) = \mathbb{P}(I^t = i^t)$, où $I^\infty : (\Omega, \mathscr{A}, \mathbb{P}) \to (\mathscr{M}^\infty, \mathscr{B}^\infty)$ est une variable aléatoire de loi μ.

Remarque 3.5. *En particulier, une loi a priori sur les suites infinies d'experts définit pour tout $T \geqslant 1$ des poids initiaux $\omega_\mu(i^T)$ sur les séquences i^T de longueur T.*

La stratégie d'agrégation à poids exponentiels se généralise naturellement au cas où l'ensemble des experts est non dénombrable, comme nous l'avons vu dans le cas du simplexe dans la section 2.3.3. Dans le cas de l'espace $(\mathscr{M}^\infty, \mathscr{B}^\infty)$, muni d'une mesure initiale μ, la prédiction à l'étape t se fait en moyennant les prédictions des experts selon la mesure ν_t définie par :

$$\nu_t(d\, i^\infty) = \frac{e^{-\eta\, L_{t-1}(i^\infty)}}{\int_{\mathscr{M}^\infty} e^{-\eta\, L_{t-1}(j^\infty)} \mu(d\, j^\infty)}\, \mu(d\, i^\infty)\,. \tag{3.12}$$

En particulier, pour tout $T \geqslant t$ et tout i^T, en notant $v_t(i^T)$ le poids attribué par ν_t à la suite d'experts i^T et $\omega_\mu(i^T)$ le poids initial associé à μ, le fait 3.4 montre que l'on a $L_{t-1}(i^\infty) = L_{t-1}(i^{t-1}) = L_{t-1}(i^T)$, d'où :

$$v_t(i^T) = \frac{\omega_\mu(i^T)\, e^{-\eta\, L_{t-1}(i^T)}}{\sum_{j^T} \omega_\mu(j^T)\, e^{-\eta\, L_{t-1}(j^T)}}\,,$$

ce qui correspond précisément aux poids de la stratégie d'agrégation à poids exponentiels 2.4 appliquée aux experts i^T avec les poids initiaux $\omega_\mu(i^T)$. Ainsi, via la correspondance de la remarque 3.5 précédente, *la stratégie d'agrégation à poids exponentiels sur des suites d'experts infinies avec la loi initiale μ prédit jusqu'au temps T comme l'agrégation à poids exponentiels des suites d'experts i^T de longueur T de poids initiaux $\omega_\mu(i^T)$.*

Principe général. Dans le contexte présent, et plus généralement lorsque l'on étudie un algorithme qui agrège un grand nombre d'experts structurés obtenus à partir d'experts « de base », l'analyse procède en deux étapes. D'une part, on établit une borne de regret pour cette stratégie, en utilisant des résultats généraux sur l'agrégation à poids exponentiels (par exemple le théorème 2.11). Dans un second temps, on montre que l'algorithme

6. Où $\mathscr{M}$ est l'ensemble des experts de base, soit $[\![1, M]\!]$ ou $\mathbf{N}^*$, et $\mathscr{B}$ l'ensemble des parties de $\mathscr{M}$.

se simplifie, grâce à la structure particulière qui relie les prédictions des experts, en un algorithme de complexité raisonnable.

Dans ce qui suit, on se donne une mesure de probabilité μ sur $(\mathscr{M}^\infty, \mathscr{B}^\infty)$. D'après la discussion précédente, pour tout $T \geqslant 1$, la stratégie d'agrégation à poids exponentiels des suites infinies d'experts, de mesure initiale μ, coïncide jusqu'au temps T avec la stratégie 2.4 appliquée aux experts i^T de poids initiaux $\omega_\mu(i^T)$.

Par le théorème 2.11, cet algorithme garantit donc la borne de regret suivante : pour tout $i^T = (i_1, \ldots, i_T)$,

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} = L_T - L_T(i^T) \leqslant \frac{1}{\eta} \log \frac{1}{\omega_\mu(i^T)} \,. \tag{3.13}$$

ce qui donne la borne de regret générale en fonction de μ. De plus, pour tout $t \geqslant 1$, cet algorithme prédit[7], en notant $x_{i,t}$ les prédictions des experts de base $i = 1, \ldots, M$:

$$\frac{\sum_{i^t} \omega_\mu(i^t)\, e^{-\eta\, L_{t-1}(i^{t-1})}\, x_t(i^t)}{\sum_{i^t} \omega_\mu(i^t)\, e^{-\eta\, L_{t-1}(i^{t-1})}} = \frac{\sum_{i=1}^{M} \sum_{i^{t-1}} \omega_\mu(i^{t-1}, i)\, e^{-\eta\, L_{t-1}(i^{t-1})}\, x_{i,t}}{\sum_{i^{t-1}} \omega_\mu(i^{t-1})\, e^{-\eta\, L_{t-1}(i^{t-1})}} = \boldsymbol{v}_t \cdot \boldsymbol{x}_t \,,$$

où les poids $\boldsymbol{v}_t = (v_{i,t})_{1 \leqslant i \leqslant M}$ sont définis par :

$$v_{i,t} = \frac{\sum_{i^{t-1}} \omega_\mu(i^{t-1})\, e^{-\eta\, L_{t-1}(i^{t-1})}\, \omega_\mu(i \,|\, i^{t-1})}{\sum_{i^{t-1}} \omega_\mu(i^{t-1})\, e^{-\eta\, L_{t-1}(i^{t-1})}} \,. \tag{3.14}$$

Notons que, dans la formulation bayésienne de l'agrégation à poids exponentiels pour la perte logarithmique (cf. sous-section 2.2.3), l'équation (3.14) se réécrit sous la forme suivante :

$$\mathbb{P}(I_t = i \,|\, Y^{t-1} = y^{t-1}) = \sum_{i^{t-1}} \mathbb{P}(I^{t-1} = i^{t-1} \,|\, Y^{t-1} = y^{t-1})\, \mathbb{P}(I_t = i \,|\, I^{t-1} = i^{t-1}) \tag{3.15}$$

où la loi du couple (I^∞, Y^∞) est définie par : $I^\infty \sim \mu$, et

$$\mathbb{P}(Y^t = y^t \,|\, I^\infty) = p_{I^\infty}(y^t) = \prod_{s=1}^{t} p_{I_s,s}(y_s \,|\, y^{s-1}) \,.$$

Le cas où μ est une loi de Markov. Pour spécifier notre algorithme, nous devons choisir une mesure de probabilité μ sur $(\mathscr{M}^\infty, \mathscr{B}^\infty)$. Cette loi doit être déterminée de manière à satisfaire un double objectif : d'une part, garantir une borne de regret de bonne qualité face aux séquences auxquelles on désire se comparer ; d'autre part, conférer à l'algorithme une implémentation efficace.

Pour le second objectif, en s'inspirant du cas bayésien, on aimerait simplifier la formule (3.15), qui porte sur M^{t-1} termes. Une manière particulièrement efficace de remplir cet objectif est d'imposer la condition $I_t \perp\!\!\!\perp I^{t-2} \,|\, I_{t-1}$ (*i.e.* I^∞ est un processus de Markov

7. En utilisant à nouveau le fait qu'il prédit jusqu'au temps t comme l'algorithme d'agrégation des i^t avec pour poids initiaux les $\omega_\mu(i^t)$.

d'ordre 1), car alors $\mathbb{P}(I_t = i \,|\, I^{t-1} = i^{t-1}) = \mathbb{P}(I_t = i \,|\, I_{t-1} = i_{t-1})$, et la somme sur i^{t-2} se simplifie en la relation de récurrence :

$$\mathbb{P}(I_t = i \,|\, Y^{t-1} = y^{t-1}) = \sum_{i_{t-1}} \mathbb{P}(I_{t-1} = i_{t-1} \,|\, Y^{t-1} = y^{t-1}) \, \mathbb{P}(I_t = i \,|\, I_{t-1} = i_{t-1}) \,. \tag{3.16}$$

Quant à la première condition, elle impose de donner un poids important aux suites qui présentent peu de changements, donc de choisir une loi qui favorise la cohérence temporelle (par exemple, une faible probabilité de changer d'expert d'une échéance à la suivante).

Les observations précédentes suggèrent de prendre pour μ une loi de Markov d'ordre 1, de loi initiale $\boldsymbol{\omega} \in \mathscr{P}_M$ (c'est-à-dire $\mathbb{P}(I_1 = i) = \omega_i$) et de probabilités de transition $\mathbb{P}(I_t = i \,|\, I_{t-1} = j) = \theta_t(i \,|\, j)$. Avec ce choix, le regret de l'algorithme (cf. (3.13)) devient :

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \log \frac{1}{\omega_{i_1}} + \frac{1}{\eta} \sum_{t=2}^{T} \log \frac{1}{\theta_t(i_t \,|\, i_{t-1})} \,. \tag{3.17}$$

En outre, le numérateur de (3.14), noté $w_{i,t}$, devient :

$$\begin{aligned} w_{i,t} &= \sum_{i^{t-1}} \omega_\mu(i^{t-1}) \, e^{-\eta \, L_{t-1}(i^{t-1})} \, \theta_t(i \,|\, i_{t-1}) \\ &= \sum_{i_{t-1}} \underbrace{\sum_{i^{t-2}} \omega_\mu(i^{t-2}) \, e^{-\eta \, L_{t-1}(i^{t-2})} \, \omega_\mu(i_{t-1} \,|\, i^{t-2})}_{w_{i_{t-1},t-1}} \, e^{-\eta \, \ell_{i_{t-1},t-1}} \, \theta_t(i \,|\, i_{t-1}) \,, \end{aligned}$$

soit, en notant $j = i_{t-1}$ et en divisant par la somme sur i :

$$v_{i,t} = \frac{\sum_{j=1}^M \theta_t(i \,|\, j) \, v_{j,t-1} e^{-\eta \, \ell_{j,t-1}}}{\sum_{j=1}^M v_{j,t-1} e^{-\eta \, \ell_{j,t-1}}} = \sum_{j=1}^{M} \theta_t(i \,|\, j) \, v_{j,t-1}^m \,. \tag{3.18}$$

Ainsi, la mise à jour des poids $\boldsymbol{v}_t$ des experts de base se fait en temps et en espace $O(M)$, ceux-ci ne dépendant que des poids précédents $\boldsymbol{v}_{t-1}$, des pertes $\boldsymbol{\ell}_{t-1}$ encourues à l'échéance précédente ainsi que des probabilités de transition de la loi μ. Pour résumer, on a obtenu une simplification et une borne de regret de l'algorithme d'agrégation de séquences d'experts avec une loi a priori markovienne, qui résultent respectivement de (3.17) et de (3.18) :

Stratégie 3.2. L'algorithme d'agrégation de suites d'experts avec une mesure initiale de Markov procède comme suit :

- Paramètres : la vitesse d'apprentissage $\eta > 0$, un vecteur de poids $\boldsymbol{\omega} \in \mathscr{P}_M$ et pour tout $t \geqslant 2$, une matrice stochastique $\boldsymbol{\theta}_t = \big(\theta_t(i \,|\, j)\big)_{1 \leqslant i,j \leqslant M}$.
- Initialisation : On pose $\boldsymbol{v}_1 = \boldsymbol{\omega}$.
- Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$.

Calcul des postérieurs : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, ce qui permet de calculer les pertes $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i = 1, \dots, M$. On en déduit alors les poids postérieurs :

$$v^m_{i,t} = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}} .$$

Partage des poids : Les poids sont enfin mis à jour par la formule (3.18), c'est-à-dire : $\boldsymbol{v}_{t+1} = \boldsymbol{\theta}_{t+1}\, \boldsymbol{v}^m_t$.

Proposition* 3.6. *Supposons la fonction de perte η-exp-concave. Alors, pour tout $T \geqslant 1$ et toute séquence $(i_1, \dots, i_T)$ d'experts, la stratégie 3.2 de paramètres η, $\boldsymbol{\omega}$ et $\boldsymbol{\theta}_t$ garantit la borne de regret suivante :*

$$\sum_{t=1}^T \ell_t - \sum_{t=1}^T \ell_{i_t,t} \leqslant \frac{1}{\eta} \log \frac{1}{\omega_{i_1}} + \frac{1}{\eta} \sum_{t=2}^T \log \frac{1}{\theta_t(i_t \,|\, i_{t-1})} . \tag{3.19}$$

Pour obtenir un algorithme (et une majoration) concrets, il reste à choisir les matrices de transition $\boldsymbol{\theta}_t$ pour $t \geqslant 2$. Notons que la matrice $\boldsymbol{\theta}_t$ n'est utilisée qu'à la fin de l'étape $t-1$ afin de calculer le poids $\boldsymbol{v}_t$, ce qui autorise à ne choisir cette matrice qu'à ce moment-là, en fonction des pertes précédentes [8].

Nous allons ici proposer un choix très simple de matrice de transition, qui permet de retrouver l'algorithme FS. Comme nous l'avons observé, afin de se comparer aux suites exhibant peu de ruptures, il faut que la loi μ favorise la cohérence temporelle, donc qu'à chaque étape le changement d'expert soit peu probable. Considérons alors la matrice de transition *homogène* suivante :

$$\theta_t(i \,|\, j) = (1-\alpha) \mathbf{1}_{i=j} + \alpha\, \omega_i ; \tag{3.20}$$

i.e., avec probabilité $1-\alpha$ l'expert I_t est égal à l'expert I_{t-1}, et avec probabilité α il est tiré au hasard parmi les M experts selon le vecteur de probabilité initial $\boldsymbol{\omega} \in \mathscr{P}_M$. Dans ce cas, la mise à jour (3.18) s'écrit :

$$\boldsymbol{v}_t = (1-\alpha)\, \boldsymbol{v}^m_{t-1} + \alpha\, \boldsymbol{\omega} , \tag{3.21}$$

ce qui correspond à la mise à jour de l'algorithme FS lorsque $\boldsymbol{\omega} = \frac{1}{M}\mathbf{1}$. De plus, la borne de regret (3.19) devient, en notant $k = k(i^T)$ le nombre de changements de i^T, et $1 = t_0 < t_1 < \dots < t_k \leqslant T$ les instants de rupture :

$$\begin{aligned}
\sum_{t=1}^T \ell_t - \sum_{t=1}^T \ell_{i_t,t} &\leqslant \frac{1}{\eta} \log \frac{1}{\omega_{i_1}} + \frac{1}{\eta} \sum_{j=1}^k \log \frac{1}{\alpha\, \omega_{i_{t_j}}} + \frac{1}{\eta} \sum_{t \neq t_j} \log \frac{1}{(1-\alpha) + \alpha\, \omega_{i_t}} \\
&\leqslant \frac{1}{\eta} \sum_{j=0}^k \log \frac{1}{\omega_{i_{t_j}}} + \frac{1}{\eta} k \log \frac{1}{\alpha} + \frac{1}{\eta} (T-k-1) \log \frac{1}{1-\alpha} .
\end{aligned}$$

Fait 3.7. *Ainsi, la stratégie d'agrégation à poids exponentiels sur les séquences d'experts avec pour mesure a priori μ la loi de Markov homogène d'ordre 1, de probabilité initiale $\frac{1}{M}\mathbf{1}$ et de loi de transition (3.20) coïncide avec la stratégie FS, et la borne de regret générale (3.17) permet de retrouver la borne de la proposition 3.1 obtenue par une analyse directe.*

8. C'est notamment ce que nous ferons dans la sous-section 3.1.3 suivante.

Autres choix de mesure μ. Le fait d'avoir réinterprété la stratégie FS en termes bayésiens permet de gagner en flexibilité, puisqu'on dispose désormais d'un moyen générique de construction d'algorithmes compétitifs face aux séquences d'experts. Il suffit en effet de se donner une loi μ sur $(\mathscr{M}^\infty, \mathscr{B}^\infty)$, qui aboutit à la formule de mise à jour (3.14) (que l'on doit simplifier à partir de la forme de μ), et à la borne de regret (3.13).

Nous reviendrons dans la suite de ce mémoire à ce principe : en section **??**, le choix d'une loi μ idoine permettra d'obtenir un algorithme et des bornes adaptés au problème que nous traiterons. Pour le moment, contentons-nous d'indiquer une application simple, due à Bousquet [Bou03].

Comme nous l'avons noté dans la sous-section précédente, l'algorithme FS dépend d'un paramètre α qu'il faut calibrer en fonction de l'objectif, par exemple $\alpha = \frac{k_0}{T-1}$ pour obtenir la meilleur borne de regret au temps T face aux suites d'experts avec au plus k_0 changements (corollaire 3.2). Il est alors naturel de chercher à s'affranchir de ce paramètre à calibrer, en faisant en sorte que l'algorithme apprenne au cours de la prédiction le meilleur choix de α, et admette un regret contrôlé face à celui-ci.

Pour ce faire, l'idée consiste à introduire une loi a priori $q(\alpha)\, d\alpha$ sur les valeurs du paramètre α. On prend alors pour loi $\mu = \mu_q$ sur $(\mathscr{Y}^\infty, \mathscr{B}^\infty)$ la moyenne des lois μ_α définies dans la discussion précédente [9], soit pour tout $i^T \in \mathscr{M}^T$:

$$\mu(i^T) = \int_0^1 \mu_\alpha(i^T) q(\alpha) d\alpha \,. \tag{3.22}$$

Bousquet montre dans [Bou03] que les choix $q_U(\alpha) = 1$ (la loi uniforme sur $[0,1]$) et $q_J(\alpha) = \frac{1}{\pi\sqrt{\alpha(1-\alpha)}}$ (la distribution de Jeffreys pour le modèle de Bernoulli, *i.e.* la loi $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$) conduisent à des algorithmes dont la mise à jour se fait en temps $O(MT)$ (donc l'algorithme se met en œuvre jusqu'à l'étape T en temps $O(MT^2)$). De plus, ces choix conduisent aux bornes de regret suivantes :

Proposition 3.8. *Supposons la fonction de perte η-exp-concave. Alors, pour tout $T \geqslant 1$, tout $k \in [\![1, T]\!]$ et toute suite d'experts $i^T = (i_1, \ldots, i_T)$ avec k ruptures, les stratégies d'agrégation à poids exponentiels des séquences d'experts de paramètre η et de mesures initiales $\mu_U = \mu_{q_U}$ et $\mu_J = \mu_{q_J}$ vérifient respectivement :*

$$\sum_{t=1}^T \ell_t - \sum_{t=1}^T \ell_{i_t} \leqslant \frac{1}{\eta}(k_0+1)\log M + \frac{1}{\eta}(T-1)\, h\Big(\frac{k}{T-1}\Big) + \frac{1}{\eta}\log(T+1)\,, \text{ et} \tag{3.23}$$

$$\sum_{t=1}^T \ell_t - \sum_{t=1}^T \ell_{i_t} \leqslant \frac{1}{\eta}(k_0+1)\log M + \frac{1}{\eta}(T-1)\, h\Big(\frac{k}{T-1}\Big) + \frac{1}{2\eta}\log T + \frac{1}{\eta}\log 2\,; \tag{3.24}$$

c'est-à-dire à des bornes excédant celle obtenue pour une calibration optimale de α en fonction de k et T (corollaire 3.2) de respectivement $\frac{1}{\eta}\log(T+1)$ et $\frac{1}{2\eta}\log T + \frac{1}{\eta}\log 2$.

9. Qui sont markoviennes homogènes d'ordre 1, de loi initiale $\frac{1}{M}\mathbf{1}$ et de probabilités de transition données par la formule (3.20), et conduisent aux prédictions de l'algorithme FS de paramètre α.

3.1.3 Variante *Variable Share* de l'algorithme FS

Dans la section 2.4.1, nous avons vu que, dans le cas de pertes bornées, la calibration du paramètre η en fonction de la perte du meilleur expert permet d'obtenir des bornes plus fines lorsque celle-ci est faible. Le même genre de raffinement est possible pour le paramètre α lorsque l'on se compare à la meilleure séquence d'experts.

Dans cette sous-section, on suppose la fonction de pertes à valeurs dans $[0,1]$. Nous allons introduire dans ce contexte une variante de l'algorithme FS, l'algorithme *Variable Share* (VS), également introduit et analysé par Herbster et Warmuth dans [HW98]. L'idée est la suivante : au lieu de partager une fraction fixe α de son poids avec les autres experts, chaque expert cède une fraction de son poids d'autant plus faible que sa perte courante est faible. Plus précisément, la stratégie se décline de la façon suivante :

Stratégie 3.3 (Variable Share)**.** L'algorithme *Variable Share* (VS) procède comme suit :

Paramètres : $\eta > 0$ et $0 < \alpha < 1$.

Initialisation : On pose $\boldsymbol{v}_1 = \frac{1}{M}\mathbf{1}$.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$, ou plus généralement $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$.

Calcul des postérieurs : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, ce qui permet de calculer les pertes $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i = 1, \ldots, M$, ainsi que les poids postérieurs :

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,.$$

Partage des poids : Les poids sont enfin mis à jour selon la formule suivante :

$$v_{i,t+1} = (1-\alpha)^{\ell_{i,t}}\, v_{i,t}^m + \frac{1}{M}\sum_{j=1}^{M}\left(1-(1-\alpha)^{\ell_{j,t}}\right) v_{j,t}^m\,. \tag{3.25}$$

Remarque 3.6. *Ainsi, d'après l'équation* (3.25)*, si l'expert i subit au temps t la perte maximale $\ell_{i,t} = 1$, il partagera la fraction α de son poids ; si au contraire sa perte est nulle, il ne cédera aucune fraction de son poids.*

Remarque 3.7. *L'algorithme VS peut en fait être vu comme un algorithme d'agrégation des suites d'experts. En effet, nous avons déjà signalé que dans l'algorithme 3.2 d'agrégation de suites d'experts avec une loi initiale de Markov, la matrice de transition $\boldsymbol{\theta}_{t+1}$ n'est utilisée que pour le calcul du poids $\boldsymbol{v}_{t+1}$, et peut donc être choisie en fonction des pertes jusqu'au temps t. En particulier, le choix*

$$\theta_{t+1}(i\,|\,j) = (1-\alpha)^{\ell_{j,t}}\, \mathbf{1}_{i=j} + \frac{1-(1-\alpha)^{\ell_{j,t}}}{M} \tag{3.26}$$

conduit précisément à l'algorithme VS.

L'algorithme VS admet la borne de regret suivante, démontrée par [HW98] puis retrouvée dans [Vov99] par une approche légèrement différente, que nous suivons ici.

Théorème 3.9. *Supposons la fonction de perte η-exp-concave et à valeurs dans* $[0,1]$*. Alors, pour tout $T \geqslant 1$ et toute séquence $i^T = (i_1, \dots, i^T)$ avec $k = k(i^T)$ ruptures, l'algorithme VS de paramètres η et α satisfait la borne de regret suivante :*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} \leqslant \frac{1}{\eta}(k+1)\log M + \frac{1}{\eta} k \log \frac{1}{\alpha} + \frac{1}{\eta}\left(\sum_{t=1}^{T} \ell_{i_t,t}\right) \log \frac{1}{1-\alpha} + k\,. \tag{3.27}$$

Remarque 3.8. *Ainsi, en notant $L^*_{T,k}$ la perte cumulée de la meilleure séquence de T experts avec k changements, la borne (3.27) est similaire à celle (3.8) de l'algorithme FS, à ceci près que le terme $\frac{1}{\eta}(T-k-1)\log\frac{1}{1-\alpha}$ a été remplacé par $\frac{1}{\eta}L^*_{T,k}\log\frac{1}{1-\alpha} + k$, ce qui est une amélioration sensible lorsque $L^*_{T,k} \ll T$.*

Démonstration. Au vu de la remarque 3.7, on pourrait être tenté d'appliquer directement la proposition 3.6 pour majorer le regret par rapport à n'importe quelle séquence à partir de son poids initial. Ce n'est en fait pas une bonne idée, car la loi définie par l'équation (3.26) attribue un poids nul aux suites i^T telles qu'il existe t tel que $\ell_{i_t,t} = 0$ et $i_{t+1} \neq i_t$. On peut cependant contourner cette difficulté en utilisant le lemme ci-contre :

Lemme 3.10. *Soit $\mathscr{E}$ un ensemble fini ou dénombrable d'experts, et $\boldsymbol{\omega} \in \mathscr{P}(\mathscr{E})$. Si la fonction de perte est η-mélangeable, l'algorithme d'agrégation 2.4 de poids initiaux $\boldsymbol{\omega}$ garantit la perte suivante : pour tout $L \in \mathbf{R}$,*

$$L_T \leqslant L + \frac{1}{\eta} \log \frac{1}{\sum_{e : L_{e,T} \leqslant L} \omega_e}\,. \tag{3.28}$$

Preuve du lemme 3.10. La preuve du théorème 2.11 montre que $L_T \leqslant -\frac{1}{\eta}\log W_{T+1}$, où :

$$W_{T+1} = \sum_{e \in \mathscr{E}} \omega_e\, e^{-\eta\, L_{e,T}} \geqslant \sum_{e : L_e \leqslant L} \omega_e\, e^{-\eta\, L}\,,$$

ce qui donne la majoration annoncée. □

Soit maintenant μ la mesure de Markov sur $(\mathscr{M}^\infty, \mathscr{B}^\infty)$ de loi initiale $\frac{1}{M}\mathbf{1}$ et de probabilités de transitions données par l'équation (3.26). Soit également I^∞ une variable aléatoire de loi μ. En particulier, conditionnellement à $I_t = i_t$, avec probabilité $(1-\alpha)^{\ell_{i,t}}$ I_{t+1} vaut i_t, et avec probabilité $1-(1-\alpha)^{\ell_{i,t}}$ I_{t+1} est tirée au hasard uniformément sur les M experts.

Fixons $T \geqslant 1$, et i^T une suite d'experts avec k changements $1 < t_1 < \cdots < t_k \leqslant T$ (on note aussi $t_0 = 1$ et $t_{k+1} = T$). Nous allons montrer :

$$\mathbb{P}(L_T(I^T) \leqslant L_T + 2k) \geqslant \frac{1}{M}\left(\frac{\alpha}{M}\right)^k (1-\alpha)^{L_T(i^T)}\,, \tag{3.29}$$

ce qui implique, par le lemme 3.10 et la remarque 3.7, la borne (3.27), à ceci près que le dernier terme k est remplacé par[10] $2k$.

Pour établir (3.29), on va en fait minorer par le terme de droite par la probabilité que $I_1 = I_{t_1} = i_1$, qu'à chaque rupture t_j, $j = 1, \dots, k$ de la séquence i^T, I_t rejoigne i_t

10. Pour obtenir le terme k au lieu de $2k$, une analyse et une majoration un peu plus fines sont nécessaires, cf. [HW98, Vov99].

avant de subir une perte de t_j à t supérieure à 2, puis reste égal à i_t jusqu'au changement suivant. En effet, ces conditions garantissent que $L_T(I^T) \leqslant L_T(i^T) + 2k$.

Tout d'abord, la probabilité que I_t commence à i_1 et y reste jusqu'à $t_1 - 1$ est :

$$\mathbb{P}(I_1 = i_1)\,\mathbb{P}(I_2 = i_1 \,|\, I_1 = i_1) \cdots \mathbb{P}(I_{t_1-1} = i_1 \,|\, I_{t_1-2} = i_1) \geqslant \frac{1}{M}(1-\alpha)^{\ell_{i_1,1}} \ldots (1-\alpha)^{\ell_{i_{t-2},1}}$$

soit, en notant $L_{t..t'}$ la perte de i^T sur $[\![t, t']\!]$, $\frac{1}{M}(1-\alpha)^{L_{1..t_1-2}}$. De plus, pour toute étape j, conditionnellement au fait que le processus I_t se comporte comme prescrit jusqu'au temps $t_j - 1$ (en particulier $I_{t_j-1} = i_{t_{j-1}}$), la probabilité qu'il rejoigne i_t avant de subir une perte de 2 puis reste égal à i_t jusqu'au prochain changement se majore de la façon suivante. La probabilité qu'il n'effectue pas de changement avant du subir une perte cumulée de 2 est, en notant $i = i_{t_{j-1}}$ et $t'_j = \inf\{t \geqslant t_j \mid \sum_{t'=t_j-1}^{t} \ell_{i,t'} \geqslant 2\}$, majorée ainsi :

$$\begin{aligned}
\mathbb{P}(I_{t_j} = \ldots I_{t'_j} = i \,|\, I_{t_{j-1}} = i) &\leqslant (1-\alpha)^{\ell_{i,t_j-1}} \cdots (1-\alpha)^{\ell_{i,t'_j-1}} \\
&= (1-\alpha)^{\sum_{t=t_j}^{t'_j-1} \ell_{i,t}} \\
&\leqslant 1 - \alpha\,,
\end{aligned}$$

où l'on a utilisé pour la dernière inégalité le fait que $\sum_{t=t_j}^{t'_j-1} \ell_{i,t} = \sum_{t=t_j}^{t'_j} \ell_{i,t} - \ell_{i,t'_j} \geqslant 2-1 = 1$ par définition de t'_j et car les pertes sont dans $[0,1]$. Ainsi, la probabilité que I effectue un changement (en tirant le nouvel expert au hasard uniformément parmi tous les experts) avant du subir une perte cumulée de 2 à partir de $t_j - 1$ est minorée par α ; de plus, conditionnellement au fait que le changement se fait au temps t, la probabilité que $I_t = i_t$ est égale à $\frac{1}{M}$, et, conditionnellement à cela, la probabilité qu'il y reste jusqu'à l'instant $t_{j'} - 1$ précédent la prochaine rupture est minorée par $(1-\alpha)^{L_{t..t_{j'}-1}}$. En effectuant le produit sur les ruptures successives t_j de i^T, on obtient la borne (3.29). □

3.2 Comparaison à la meilleure séquence parmi un petit nombre d'experts

Nous nous intéressons à présent à un problème plus délicat que celui de la comparaison à la meilleure séquence d'experts (avec peu de ruptures). On suppose que l'on dispose d'un grand nombre d'experts de base, dont un petit nombre prédit mieux que les autres, et tel que le meilleur expert au cours du temps oscille entre ces quelques experts. Dans ces conditions, plutôt que de chercher à se comparer à des séquences d'experts quelconques, il est plus intéressant de définir une stratégie compétitive face à la meilleure séquence d'expert oscillant entre un petit nombre d'experts.

Dans la sous-section 3.2.1, nous formalisons ce problème, en discutons la complexité et exhibons un algorithme « direct », mais déraisonnablement coûteux d'un point de vue algorithmique, traitant ce problème. Dans la sous-section 3.2.2, nous présentons une stratégie efficace, l'algorithme *Mixing Past Posteriors* (MPP) introduit par Bousquet et Warmuth [BW02], qui résout ce problème.

3.2.1 Le problème de Freund

Nous nous intéressons au problème, posé par Freund et traité dans [BW02], de se comparer non pas à des séquences quelconques d'experts, mais à des suites qui alternent entre un petit nombre d'experts (en anglais, des « sparse sequences »). Cet objectif est particulièrement pertinent lorsqu'on dispose d'un grand nombre d'experts de base, mais dont on peut soupçonner qu'un petit nombre d'entre eux surpassera les autres. En d'autres termes, le problème de la section 3.1 et les stratégies correspondantes, qui conduisent à se garantir contre des suites d'experts quelconques, ne rendent pas compte du fait que des experts qui ont bien prédit dans le passé tendent souvent à mieux prévoir dans le futur que des experts qui ont toujours mal prédit.

Pour formaliser, nous cherchons à définir une stratégie qui permette de contrôler le regret $L_T - L_T(i^T)$ face aux suites i^T dont le nombre de ruptures $k(i^T)$ (cf. (3.3)) ainsi que le nombre d'experts différents

$$n(i^T) = \left|\left\{i_t \mid 1 \leqslant t \leqslant T\right\}\right| \tag{3.30}$$

sont contrôlés. Bien entendu, les méthodes de la section précédente s'appliquent, mais nous aimerions obtenir de meilleures bornes de regret lorsque $n(i^T)$ est faible. En effet, dans cette section, il faut penser que le nombre total M d'experts de base est grand, et donc que le terme de regret en $k \log M$ de l'algorithme FS est trop élevé. En résumé, on cherche à traiter le problème suivant :

Problème 5 (« Sparse shifting regret »)**.** Il s'agit de construire une stratégie d'agrégation d'experts *efficace* d'un point de vue algorithmique qui permette de contrôler le regret :

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i,t} \tag{3.31}$$

uniformément sur toutes les suites i^T d'experts telles que $k(i^T) \leqslant k$ et $n(i^T) \leqslant n$, où $n \ll k$, $n \ll M$ et $k \ll T$, avec une borne proche de la borne optimale pour ce problème.

Pour saisir la difficulté du problème, notons que l'algorithme doit se comparer non seulement au meilleur découpage en périodes de temps et au meilleur choix d'expert pour chaque période (comme dans la section 3.1), mais aussi au meilleur groupe de n experts auxquels appartiennent les experts choisis à chaque période. Comme l'ont montré [BW02], le problème de déterminer a posteriori, étant données les pertes $(\ell_{i,t})_{1 \leqslant i \leqslant M, 1 \leqslant t \leqslant T}$ des experts, le meilleur choix d'une classe de n experts, de $k+1$ segments temporels et d'un expert parmi les n retenus par segment, est NP-difficile.

Puisque le nombre de configurations auxquelles on se compare est égal à $\binom{M}{n}\binom{T-1}{k}n^{k+1}$, le regret de l'algorithme 2.4 d'agrégation à poids exponentiels de toutes les configurations considérées donne un regret de :

$$\frac{1}{\eta}\log\binom{M}{n}\binom{T-1}{k}n^{k+1} \approx \frac{1}{\eta} n \log\frac{M}{n} + \frac{1}{\eta} k \log\frac{T-1}{k} + \frac{1}{\eta}(k+1)\log n \tag{3.32}$$

$$\approx \frac{1}{\eta} n \log\frac{M}{n} + \frac{1}{\eta} k \log\frac{nT}{k} \tag{3.33}$$

ce qui constituera donc notre objectif en termes de regret. Bien entendu, comme pour l'agrégation de toutes les séquences d'experts avec au plus k ruptures dans la section 3.1, cet algorithme est impraticable, à cause de sa complexité prohibitive (qui exige de stocker et de mettre à jour à chaque étape autant de poids que d'experts composites).

Un algorithme « direct », dû à Freund et mentionné dans [BW02], conduit essentiellement à la borne optimale (3.32) avec un coût algorithmique moindre. L'algorithme consiste à considérer, pour tout sous-ensemble de n experts parmi les M experts de base, un expert obtenu en appliquant l'algorithme FS sur ces n experts [11], puis à agréger les $\binom{M}{n}$ experts ainsi définis par l'agrégation classique 2.4 (sans changements). Par le théorème 2.11 et le corollaire 3.2, cette stratégie conduit à la borne de regret :

$$\left(\frac{1}{\eta}(k+1)\log n + \frac{1}{\eta} k \log \frac{T-1}{k} + \frac{1}{\eta} k \right) + \frac{1}{\eta} \underbrace{\log \binom{M}{n}}_{\approx n \log \frac{M}{n}}, \tag{3.34}$$

ce qui est la borne (3.32) au terme $\frac{1}{\eta}k$ près. Cependant, cet algorithme est lui aussi impraticable, car il nécessite de stocker et de mettre à jour à chaque étape $n\binom{M}{n}$ poids (ce qui est prohibitif lorsque M est élevé, ce qui est précisément le cadre de notre analyse). En outre, cet algorithme nécessite de spécifier n et k à l'avance, et n'offre aucune garantie pour des valeurs plus élevées de ces paramètres.

3.2.2 Algorithme *Mixing Past Posteriors* (MPP) de Bousquet et Warmuth

Comme annoncé, nous en venons dans cette sous-section à décrire un algorithme efficace qui résout le problème 5 de Freund : l'algorithme *Mixing Past Posteriors* (MPP) de Bousquet et Warmuth [BW02].

L'algorithme MPP. Pour comprendre l'idée sous-jacente, revenons un instant à l'algorithme FS. Celui-ci consiste, pour tout $t \geqslant 1$, à combiner les prédictions des experts à partir du vecteur de poids $\boldsymbol{v}_t$, celui-ci étant mis à jour par la formule : $\boldsymbol{v}_{t+1} = (1-\alpha)\boldsymbol{v}_t^m + \alpha\frac{1}{M}\mathbf{1}$. Par rapport à la mise à jour classique $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t^m$ de l'algorithme d'agrégation à poids exponentiels, le terme $\frac{\alpha}{M}\mathbf{1}$ garantit à chaque expert un poids minimal à tout instant, ce qui lui permet de rattraper plus rapidement son retard par rapport au meilleur expert dans le passé s'il se met à prédire mieux que les autres ; cela permet alors d'être compétitif face à toute séquence d'experts.

Dans notre cadre, nous aimerions nous comparer aux séquences qui alternent entre un petit nombre d'experts. Par conséquent, plutôt que de distribuer un poids commun à tous les experts, nous aimerions donner un poids plus important aux experts qui ont bien prédit dans le passé, pour permettre à ceux-ci de récupérer plus rapidement du poids s'ils se remettent à mieux prédire que les autres.

11. Avec α calibré de façon optimale en fonction de k et T, soit $\alpha = \frac{k}{T-1}$ (cf. corollaire 3.2), ou en utilisant la stratégie alternative de la proposition 3.8 qui agrège sur α.

L'algorithme MPP procède selon cette logique, en choisissant comme vecteur de poids $\boldsymbol{v}_{t+1}$ une combinaison convexe des « postérieurs » précédents (d'où son nom), selon des coefficients prédéterminés :

$$\boldsymbol{v}_{t+1} = \sum_{q=0}^{t} \beta_{t+1}(q)\, \boldsymbol{v}_q^m\,, \tag{3.35}$$

où l'on note par convention $\boldsymbol{v}_0^m = \frac{1}{M}\mathbf{1}$. Ainsi, lorsqu'un expert a bien prédit par le passé, certains de ses postérieurs passés sont élevés, et donc par l'équation (3.35) il bénéficie d'un poids minimal qui lui permet de récupérer plus vite.

Remarque 3.9. *Notons que l'algorithme MPP est plus difficile à interpréter en termes d'agrégation bayésienne que l'agrégation à poids exponentiels classique ou que l'algorithme FS. En particulier, l'appellation « postérieurs » se rapportant aux poids $\boldsymbol{v}_t^m$ est ici purement conventionnelle, tandis que ces poids pouvaient s'interpréter comme des probabilités postérieures en un sens précis pour l'agrégation à poids exponentiels et l'algorithme VS*[12].

Dans cette section, nous présentons l'algorithme MPP comme une stratégie ad hoc, *mais nous verrons en section 3.4 qu'il admet lui aussi une interprétation bayésienne en termes de* spécialistes, *révélée dans [KAW12].*

Stratégie 3.4 (Mixing Past Posteriors)**.** L'algorithme *Mixing Past Posteriors* (MPP) est défini comme suit :

Paramètres : $\eta > 0$ la vitesse d'apprentissage, et pour tout $t \geqslant 1$, une mesure de probabilité β_t sur $[\![0, t-1]\!]$.

Initialisation : On pose $\boldsymbol{v}_1 = \boldsymbol{v}_0^m = \frac{1}{M}$.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{v}_t$, déterminé à l'étape précédente. On observe alors les prédictions $\boldsymbol{x}_t \in \mathscr{X}^M$ des experts, que l'on combine en $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$, ou plus généralement $x_t = \mathbf{pred}(\boldsymbol{v}_t, \boldsymbol{x}_t)$.

Calcul des postérieurs : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, ce qui permet de calculer les pertes $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i = 1, \ldots, M$. On en déduit alors les poids postérieurs :

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,.$$

Mélange des postérieurs : Les poids sont enfin mis à jour par la formule (3.35), c'est-à-dire : $\boldsymbol{v}_{t+1} = \sum_{q=0}^t \beta_{t+1}(q)\, \boldsymbol{v}_q^m$.

La stratégie 3.4 peut s'instancier de différentes manières, en fonctions des choix des coefficients de mélange $\beta_{t+1}(q)$, $q = 0, \ldots, t$ (cf. [BW02]) :

Agrégation classique : Le choix $\beta_{t+1}(t) = 1$ et $\beta_{t+1}(q) = 0$ pour $q = 0, \ldots, t-1$ conduit à l'algorithme d'agrégation classique 2.4.

Fixed Share : Nous considérerons des choix de coefficients qui attribuent une part fixe au dernier expert, soit $\beta_{t+1}(t) = 1 - \alpha$ et donc $\sum_{q=0}^{t-1} \beta_{t+1}(q) = \alpha$ pour tout t.

- **(To Start Vector)** Lorsque l'on prend de plus $\beta_{t+1}(0) = \alpha$, on aboutit à l'algorithme FS (stratégie 3.1).

12. Où $v_{i,t}^m$ correspondait, dans les reformulations bayésiennes respectives de ces algorithmes pour la perte logarithmique, aux probabilités à posteriori $\mathbb{P}(I = i \,|\, Y^t = y^t)$ et $\mathbb{P}(I_t = i \,|\, Y^t = y^t)$.

- **(To Uniform Past)** Dans ce cas, on attribue un poids identique à tous les postérieurs du passé, soit $\beta_{t+1}(q) = \frac{\alpha}{t}$ pour $q = 0, \ldots, t-1$.
- **(To Decaying Past)** Cela correspond au choix $\beta_{t+1}(q) = \alpha \frac{1}{(t-q)^\gamma} Z_t^{-1}$, avec $\gamma > 0$ et $Z_t = \sum_{q=1}^{t} \frac{1}{(t-q)^\gamma}$ la constante de normalisation.

Borne de regret générale. Venons-en à l'analyse de regret de l'algorithme MPP pour des coefficients β_t généraux tels que $\beta_{t+1}(q) > 0$ pour tout $q = 0, \ldots, t$, comme les deux derniers choix (ce qui exclut les algorithmes d'agrégation classique et FS, déjà étudiés). Le résultat principal est le suivant (cf. [BW02], théorème 7), et peut comme la proposition 3.1 (qui repose sur le même principe) s'énoncer en termes de suites de poids :

Théorème 3.11. *Supposons la fonction de perte η-mélangeable. Pour toute suite de poids $\boldsymbol{u}^T = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ avec k ruptures et parmi les poids $\{\widetilde{\boldsymbol{u}}_j \mid j = 1, \ldots, n\}$, l'algorithme MPP (stratégie 3.4) satisfait la borne de regret suivante :*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} \sum_{j=1}^{n} \Delta\Big(\widetilde{\boldsymbol{u}}_j \Big\| \frac{1}{M}\mathbf{1}\Big) + \frac{1}{\eta} \sum_{t=1}^{T} \log \frac{1}{\beta_t(q_t)} \tag{3.36}$$

$$\leqslant \frac{1}{\eta} n \log M + \frac{1}{\eta} \sum_{t=1}^{T} \log \frac{1}{\beta_t(q_t)}\,, \tag{3.37}$$

où q_t est le dernier instant $q < t$ tel que $\boldsymbol{u}_q = \boldsymbol{u}_t$, ou 0 si $\boldsymbol{u}_t$ apparaît pour la première fois.

Démonstration. Comme dans la preuve de la proposition 3.1, on a par le choix de x_t et par le lemme 2.20 : $\ell_t - \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta}\big(\Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_t) - \Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_t^m)\big)$. De plus, le choix $\boldsymbol{v}_t = \sum_{q=0}^{t-1} \beta_t(q) \boldsymbol{v}_q^m$ implique $\boldsymbol{v}_t \geqslant \beta_t(q_t)\, \boldsymbol{v}_q^m$, donc par l'inégalité précédente et le lemme 2.19 on a :

$$\ell_t - \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta}\big(\Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_{q_t}^m) - \Delta(\boldsymbol{u}_t \,\|\, \boldsymbol{v}_t^m)\big) + \frac{1}{\eta} \log \frac{1}{\beta_t(q_t)}\,. \tag{3.38}$$

En sommant (3.38) sur l'ensemble des $t = 1, \ldots, T$ tels que $\boldsymbol{u}_t = \widetilde{\boldsymbol{u}}_j$, il vient par simplification télescopique, et par le fait que $q_t = 0$ pour la première apparition t de $\widetilde{\boldsymbol{u}}_j$:

$$\sum_{t:\boldsymbol{u}_t = \widetilde{\boldsymbol{u}}_j} (\ell_t - \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t) \leqslant \frac{1}{\eta}\big(\Delta(\widetilde{\boldsymbol{u}}_j \,\|\, \boldsymbol{v}_0^m) - \Delta(\widetilde{\boldsymbol{u}}_j \,\|\, \boldsymbol{v}_{l_j}^m)\big) + \frac{1}{\eta} \sum_{t:\boldsymbol{u}_t = \widetilde{\boldsymbol{u}}_j} \log \frac{1}{\beta_t(q_t)} \tag{3.39}$$

où l_j est l'instant de la dernière apparition de $\widetilde{\boldsymbol{u}}_j$. L'inégalité (3.36) s'en déduit en sommant sur $j = 1, \ldots, n$ et en notant que $\boldsymbol{v}_0^m = \frac{1}{M}\mathbf{1}$. □

Nous allons maintenant spécifier cette borne de regret générale pour les deux choix de coefficients évoqués plus haut, respectivement *Fixed Share to Uniform Past* (FSUP) et *Fixed Share to Decaying Past* (FSDP), suivant [BW02]. Le premier choix conduit à une borne de regret satisfaisante, quoique légèrement sous-optimale, ainsi qu'à une complexité en temps et en mémoire remarquablement faible ; le second admet de meilleures garanties de regret, mais au prix d'une complexité algorithmique plus importante.

Le cas *Fixed Share to Uniform Past.* Pour ce choix de coefficients, on a $\beta_t(t-1) = 1-\alpha$ et $\beta_t(q) = \frac{\alpha}{t-1}$ pour $0 \leqslant q < t-1$. La borne de regret du théorème 3.11 devient alors :

Corollaire 3.12. *L'algorithme MPP avec le choix de coefficients Fixed Share to Uniform Past de paramètre α admet la borne de regret suivante : pour toute suite de poids $\boldsymbol{u}^T$ avec k ruptures et parmi n poids distincts, on a*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} n \log M + \frac{1}{\eta} k \log \frac{1}{\alpha} + \frac{1}{\eta}(T-k-1) \log \frac{1}{1-\alpha} + \frac{1}{\eta} k \log(T-1)\,. \quad (3.40)$$

En particulier, le choix $\alpha = \frac{k}{T-1}$ conduit à la majoration :

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} n \log M + \frac{1}{\eta} k \log \frac{T-1}{k} + \frac{1}{\eta} k + \frac{1}{\eta} k \log(T-1)\,. \quad (3.41)$$

Démonstration. On applique la majoration (3.37) du théorème 3.11, de sorte qu'il suffit de majorer $\sum_{t=1}^{T} \log \frac{1}{\beta_t(q_t)}$. On a $\beta_1(q_1) = \beta_1(0) = 1$ et, si $t \geqslant 2$ n'est pas un instant de rupture, on a $\beta_t(q_t) = 1-\alpha$. Si enfin $t = t_j$, $j = 1, \ldots, k$, est un instant de rupture, on a $\beta_t(q_t) = \frac{\alpha}{t-1}$, donc $\log \frac{1}{\beta_t(q_t)} = \log \frac{1}{\alpha} + \log(t-1) \leqslant \log \frac{1}{\alpha} + \log(T-1)$. En sommant ces inégalités et en les injectant dans (3.37), on obtient l'inégalité (3.40) comme voulu. □

Remarque 3.10. *La borne de regret (3.41) est relativement satisfaisante. En effet, comparée à l'objectif :*

$$\frac{1}{\eta} n \log \frac{M}{n} + \frac{1}{\eta} k \log \frac{nT}{k}\,, \quad (3.42)$$

la borne que nous avons obtenue admet pour terme supplémentaire principal $\frac{1}{\eta} k \log T$, qui est raisonnable sans être toutefois négligeable, puisqu'il domine en fait le terme principal en T, qui est $\frac{1}{\eta} \log \frac{nT}{k}$. Nous verrons qu'un choix plus judicieux des coefficients conduit à une meilleure majoration.

Le principal attrait du choix de coefficients uniformes sur le passé est qu'ils conduisent à un algorithme particulièrement simple, avec une complexité en mémoire et un temps de mise à jour en $O(M)$. En effet, en notant pour tout $t \geqslant 1$ $\overline{\boldsymbol{v}}_t = \frac{1}{t} \sum_{q=0}^{t-1} \boldsymbol{v}_t^m$ la moyenne des postérieurs passés, la mise à jour (3.35) devient :

$$\boldsymbol{v}_{t+1} = (1-\alpha)\boldsymbol{v}_t^m + \alpha \overline{\boldsymbol{v}}_t \quad (3.43)$$

ce qui, compte tenu du fait que la moyenne $\overline{\boldsymbol{v}}_t$ admet elle-même la mise à jour en ligne

$$\overline{\boldsymbol{v}}_{t+1} = \frac{t}{t+1} \overline{\boldsymbol{v}}_t + \frac{1}{t+1} \boldsymbol{v}_t^m\,, \quad (3.44)$$

conduit à l'implémentation efficace suivante, qui ne nécessite de disposer à l'étape t que des poids $\boldsymbol{v}_t$ et $\overline{\boldsymbol{v}}_t$:

1. Prédire selon $\boldsymbol{v}_t$.
2. Après l'observation des pertes, calculer le postérieur $\boldsymbol{v}_t^m$ à partir de $\boldsymbol{v}_t$.
3. Mettre à jour $\boldsymbol{v}_{t+1}$ selon (3.43).
4. Mettre à jour $\overline{\boldsymbol{v}}_{t+1}$ selon (3.44).

Le cas *Fixed Share to Decaying Past.* Venons-en à la borne de regret pour un meilleur choix des coefficients de mélange (cf. [BW02]). On considère ici le choix des coefficients *Fixed Share to Decaying Past*, en particulier le cas $\gamma = 1$.

Corollaire 3.13. *L'algorithme MPP avec le choix de coefficients de mélange Fixed Share to Decaying Past avec $\gamma = 1$ satisfait la borne de regret suivante : pour toute suite de poids $\boldsymbol{u}^T$ avec k ruptures et parmi n poids distincts, on a*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} n \log M + \frac{1}{\eta} k \log \frac{1}{\alpha} + \frac{1}{\eta}(T-k-1) \log \frac{1}{1-\alpha} + \frac{1}{\eta} k \log \frac{nT}{k} + \frac{1}{\eta} k \log(1 + \log T)\,.$$

Remarque 3.11. *Le choix alternatif $\gamma > 1$ conduit à une borne similaire, mais en remplaçant $\frac{1}{\eta} k \log \frac{nT}{k} + \frac{1}{\eta} k \log(1+\log T)$ par $\frac{1}{\eta}\gamma k \log \frac{nT}{k} + \frac{1}{\eta} k \log\left(\frac{\gamma}{\gamma-1}\right)$.*

Démonstration. On applique à nouveau le théorème 3.11. Pour tout $t \geqslant 2$ qui n'est pas un instant de rupture, on a à nouveau $\beta_t(q_t) = \beta_t(t-1) = 1-\alpha$; de plus, $\beta_1(q_1) = 1$. Si maintenant $t = t_j$, $j = 1, \ldots, k$ est un instant de rupture, il vient : $\beta_{t_j}(q_{t_j}) = Z_t^{-1}(t_j - 1 - q_{t_j})^{-1}$, d'où $\frac{1}{\eta} \log \frac{1}{\beta_{t_j}(q_{t_j})} = \frac{1}{\eta} \log Z_{t_j} + \frac{1}{\eta} \log(t_j - 1 - q_{t_j})$. Or on a :

$$Z_{t_j} = \sum_{q=1}^{t_j-1} \frac{1}{q} \leqslant 1 + \int_1^{t_j-1} \frac{dx}{x} = 1 + \log(t_j - 1) \leqslant 1 + \log T\,.$$

Il reste donc à majorer $\sum_{j=1}^{k} \log(t_j - 1 - q_{t_j})$. Par concavité du logarithme, il vient :

$$\sum_{j=1}^{k} \log(t_j - 1 - q_{t_j}) \leqslant k \log \left(\frac{1}{k} \sum_{j=1}^{k} (t_j - 1 - q_{t_j}) \right)$$

Enfin, en notant $\widetilde{\boldsymbol{u}}_p$, $p = 1, \ldots, n$ les poids distincts pris par la suite $\boldsymbol{u}^T$, on a par simplification télescopique :

$$\sum_{j=1}^{k} (t_j - 1 - q_{t_j}) = \sum_{p=1}^{n} \sum_{j:\boldsymbol{u}_{t_j} = \widetilde{\boldsymbol{u}}_p} (t_j - 1 - q_{t_j}) \leqslant nT\,,$$

ce qui donne la borne annoncée en substituant dans l'équation précédente, puis en appliquant le théorème 3.11. □

Remarque 3.12. *Comme précédemment, le choix $\alpha = \frac{k}{T-1}$ dans la majoration du corollaire 3.13 conduit à la borne de regret suivante :*

$$\frac{1}{\eta} n \log M + \frac{1}{\eta} k \log \frac{T-1}{k} + \frac{1}{\eta} k + \frac{1}{\eta} k \log \frac{nT}{k} + \frac{1}{\eta} k \log(1 + \log T)\,, \tag{3.45}$$

qui est majorée par

$$\frac{1}{\eta} n \log M + \frac{1}{\eta} 2k \log \frac{T}{k} + \frac{1}{\eta} k \log n + \frac{1}{\eta} k \big(1 + \log(1 + \log T)\big)\,. \tag{3.46}$$

Si l'on fait abstraction du dernier terme doublement logarithmique en T, cette borne est essentiellement la borne optimale $\frac{1}{\eta} n \log \frac{M}{n} + \frac{1}{\eta} k \log \frac{nT}{k}$, à ceci près que le terme en $\log \frac{T}{k}$ est doublé[13]*. Cette borne est donc très satisfaisante, et meilleure que celle obtenue pour le mélange avec des coefficients uniformes sur le passé.*

Ainsi, le choix de coefficients décroissants dans le passé conduit à une borne de regret de meilleure qualité. En revanche, on ne dispose plus dans ce cas d'une simplification de l'algorithme MPP, ce qui implique que le calcul de $\boldsymbol{v}_{t+1} = \sum_{q=0}^{t} \boldsymbol{v}_q^m$ nécessite de stocker les postérieurs passés $\boldsymbol{v}_q^m$. Par conséquent, l'algorithme MPP jusqu'au temps T nécessite un espace mémoire en $O(MT)$, et en temps de mise à jour par étape de $O(MT)$. Cette complexité est sensiblement meilleure que celle de l'algorithme « direct » de Freund, mais peut elle-même devenir trop importante lorsque T est très grand.

Pour pallier cette limitation, [BW02] propose (cf. annexe C de l'article) une version approximative de cet algorithme, qui donne une borne de regret proche, avec une complexité en mémoire et en temps de mise à jour par étape de $O(M \log T)$. En section 3.3 (plus précisément en sous-section 3.3.3), nous présenterons suivant [CBGLS12] un choix alternatif de mise à jour des poids, qui conduit à la fois à une mise à jour en $O(M)$ et à une borne de regret légèrement plus fine, mais au prix d'un paramètre supplémentaire à calibrer de manière optimale en fonction de l'objectif[14]. Nous verrons également en section 3.4 qu'une interprétation « bayésienne » de l'algorithme MPP, due à [KAW12], suggère une modification de cette algorithme qui offre à la fois des garanties de regret légèrement meilleures et une complexité en $O(M)$ par étape.

3.3 Compléments sur les algorithmes FS et MPP

Dans cette section, nous présentons des résultats issus de [CBGLS12], qui montrent que les algorithmes FS et MPP offrent en fait des garanties de regret plus générales que celles que l'on a obtenues dans les sections 3.1 et 3.2.

3.3.1 Une mesure généralisée du regret

Dans ce chapitre, nous nous sommes intéressés au problème de concevoir des algorithmes capables de s'adapter aux changements du meilleur expert. La façon de formaliser cet objectif que nous avons retenue consiste à garantir un faible regret face aux séquences d'experts, c'est-à-dire à contrôler la quantité :

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \tag{3.47}$$

pour toute suite de poids $\boldsymbol{u}^T = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)$, $\boldsymbol{u}_t \in \mathscr{P}_M$, dont le nombre de ruptures

$$k(\boldsymbol{u}^T) = \left| \{ t \in [\![2, T]\!] \mid \boldsymbol{u}_t \neq \boldsymbol{u}_{t-1} \} \right| \tag{3.48}$$

13. Nous donnerons une explication de ce phénomène en section 3.4.

14. Nous suggérons toutefois en remarque 3.17 comment pallier cette limitation, de manière à bénéficier de la simplicité et des performances de l'algorithme sans avoir à introduire de paramètre supplémentaire.

est contrôlé. Dans la littérature, cette quantité est appelée « shifting regret », ou « regret avec changements ».

Une formulation alternative de l'objectif d'adaptativité aux changements a été proposée par [HS09]. Il s'agit alors de définir des algorithmes qui garantissent un faible regret sur chaque période de temps, c'est-à-dire qui contrôlent la quantité :

$$R_T^{(\tau_0)} = \sup_{[\![r,s]\!] \subset [\![1,T]\!],\; s-r+1 \leqslant \tau_0} \sum_{t=r}^{s} \ell_t - \min_{i=1,\dots M} \sum_{t=r}^{s} \ell_{i,t} = \sup_{[\![r,s]\!]} \left(\sum_{t=r}^{s} \ell_t - \inf_{\boldsymbol{u} \in \mathscr{P}_M} \sum_{t=r}^{s} \boldsymbol{u} \cdot \boldsymbol{\ell}_t \right) \quad (3.49)$$

appelée « regret τ_0-adaptatif », pour tout τ_0. Les deux notions sont bien évidemment liées : par exemple, toute borne du regret adaptatif donne une borne sur le « shifting regret ».

Dans l'article [CBGLS12], les auteurs proposent la quantité suivante, appelée « generalized shifing regret » (« regret avec changements généralisé ») qui étend à la fois le regret avec changements et le regret adaptatif, ainsi que d'autres quantités liées [15]. Ce regret est défini par la formule suivante :

$$\sum_{t=1}^{T} \|\boldsymbol{u}_t\| \, \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \quad (3.50)$$

où les poids $\boldsymbol{u}_t \in \mathbf{R}_+^M$ ne sont pas nécessairement normalisés. En particulier, le regret avec changements (3.47) s'obtient en prenant des poids normalisés ($\|\boldsymbol{u}_t\| = 1$), et le regret adaptatif en prenant $\boldsymbol{u}_t = \boldsymbol{u}$ sur $[\![r, s]\!]$ et 0 ailleurs. Dans ce contexte, la quantité qui mesure la longueur de la période de temps considérée est $\sum_{t=1}^{T} \|u_t\|$ (qui vaut T pour le regret avec changements, et τ_0 pour le regret τ_0-adaptatif).

L'analyse de [CBGLS12] montre des bornes de regret dans le cadre unifié du regret avec changements généralisé, donc valables tant pour pour le regret avec changements que pour le regret adaptatif, pour les algorithmes FS et MPP. De plus, dans chacun de ces cas, les bornes sont obtenues pour des notions plus « souples » de régularité de la suite $\boldsymbol{u}^T$ que le nombre de ruptures ou le nombre de valeurs différentes empruntées.

3.3.2 Le cas de l'algorithme FS

Dans le cas des suites de poids avec changements, la mesure de régularité que nous avons adoptée jusqu'ici est le nombre de ruptures $k(\boldsymbol{u}^T)$ défini par (3.48). Cette quantité, conçue par analogie avec les suites d'experts i^T où l'on passe d'un expert à un autre, est pensée pour prendre en comptes des changements brusques (« hard shifts ») des poids, ce qui est particulièrement pertinent lorsque l'environnement ou la qualité des experts évolue soudainement.

Une mesure de régularité plus intrinsèque aux poids, et adaptée au cas où l'environnement évolue de manière continue, s'exprime en fonction de la variation des poids au

15. Comme par exemple le regret avec escompte, ou « discounted regret », qui pondère les pertes selon des poids décroissants.

cours du temps, de manière à prendre en compte le cas où ces poids varient lentement et continûment (« soft shifts »). Dans cette veine, on pose alors :

$$\widetilde{k}(\boldsymbol{u}^T) = \sum_{t=1}^{T} D_{\mathrm{TV}}(\boldsymbol{u}_t, \boldsymbol{u}_{t-1}) \,, \tag{3.51}$$

où $D_{\mathrm{TV}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i : x_i \geqslant y_i} (x_i - y_i)$ désigne la *distance en variation totale* sur $\mathbf{R}_+^M$ (qui n'est pas symétrique sur $\mathbf{R}_+^M$), qui est toujours majorée par $\|\boldsymbol{x} - \boldsymbol{y}\|_1$ et vaut $\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_1$ si $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{P}_M$. De plus, si $\boldsymbol{u}^T$ est une suite de poids normalisés, on a toujours $\widetilde{k}(\boldsymbol{u}^T) \leqslant k(\boldsymbol{u}_T)$, avec égalité si $\boldsymbol{u}^T$ est une suite d'experts [16] i^T.

L'analyse de [CBGLS12] montre que l'algorithme FS (stratégie 3.1) offre des garanties de regret dans le cadre général du regret avec changements généralisé (c'est-à-dire pour des poids $\boldsymbol{u}_t$ non nécessairement normalisés), et valables tant pour des « soft shifts » que des « hard shifts » (car exprimées en termes de la quantité régularisée $\widetilde{k}(\boldsymbol{u}^T)$).

Proposition 3.14. *Supposons la fonction de perte η-exp-concave. Alors, l'algorithme FS de paramètres η et α garantit la borne de regret suivante : pour toute suite de poids $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \in \mathbf{R}_+^M$*

$$\sum_{t=1}^{T} \|\boldsymbol{u}_t\| \, \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{\|\boldsymbol{u}\|_1 + \widetilde{k}(\boldsymbol{u}^T)}{\eta} \log M + \frac{\widetilde{k}(\boldsymbol{u}^T)}{\eta} \log \frac{1}{\alpha} + \frac{\sum_{t=2}^{T} \|\boldsymbol{u}_t\|_1 - \widetilde{k}(\boldsymbol{u}^T)}{\eta} \log \frac{1}{1-\alpha} \,. \tag{3.52}$$

L'idée de la preuve de la proposition 3.14 est d'utiliser, comme nous l'avons fait dans ce chapitre, le lemme 2.20, qui fournit la majoration suivante :

$$\ell_t - \frac{\boldsymbol{u}_t}{\|\boldsymbol{u}_t\|_1} \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} \sum_{i=1}^{M} \frac{u_{i,t}}{\|\boldsymbol{u}_t\|_1} \log \frac{v_{i,t}^m}{v_{i,t}} \,,$$

mais cette fois-ci sans réécrire le membre de droite comme $\frac{1}{\eta}\big(\Delta(\frac{\boldsymbol{u}_t}{\|\boldsymbol{u}_t\|_1} \,\|\, \boldsymbol{v}_t) - \Delta(\frac{\boldsymbol{u}_t}{\|\boldsymbol{u}_t\|_1} \,\|\, \boldsymbol{v}_t^m)\big)$. En multipliant par $\|\boldsymbol{u}_t\|_1$ il vient :

$$\|\boldsymbol{u}_t\|_1 \, \ell_t - \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} \sum_{i=1}^{M} u_{i,t} \log \frac{v_{i,t}^m}{v_{i,t}} \,. \tag{3.53}$$

La preuve consiste alors à sommer l'inégalité (3.53), en faisant apparaître la quantité $D_{\mathrm{TV}}(\boldsymbol{u}_t, \boldsymbol{u}_{t-1})$ et en utilisant les inégalités $\boldsymbol{v}_t \geqslant (1-\alpha)\, \boldsymbol{v}_{t-1}^m$ et $\boldsymbol{v}_t \geqslant \frac{\alpha}{M}\mathbf{1}$ comme nous l'avons fait précédemment.

Remarque 3.13. *Dans le cas du « shifting regret », i.e. $\|\boldsymbol{u}_t\|_1 = 1$ pour tout t, la proposition 3.14 permet de retrouver la proposition 3.1, car $\widetilde{k}(\boldsymbol{u}^T) \leqslant k(\boldsymbol{u}^T)$. De plus, la proposition 3.14 étend cette borne au cas des suites de poids $\boldsymbol{u}_t \in \mathscr{P}_M$ variant continûment.*

Dans le cas du regret adaptatif, on déduit de la proposition 3.14 le résultat suivant :

16. C'est-à-dire de masses de Dirac associées à des experts.

Corollaire 3.15. *Sous les hypothèses de la proposition 3.14, l'algorithme FS garantit la borne de regret adaptatif suivante : pour tout $T \geqslant 1$ et $1 \leqslant \tau_0 \leqslant T$,*

$$R_T^{(\tau_0)} \leqslant \frac{1}{\eta} \log M + \frac{1}{\eta} \log \frac{1}{\alpha} + \frac{\tau_0 - 1}{\eta} \log \frac{1}{1-\alpha} \,. \tag{3.54}$$

En particulier, le choix $\alpha = 1/\tau_0$ conduit à la majoration

$$R_T^{(\tau_0)} \leqslant \frac{1}{\eta} \log M + \frac{\tau_0}{\eta} h\Big(\frac{1}{\tau_0}\Big) \leqslant \frac{\log(eM\tau_0)}{\eta} \,. \tag{3.55}$$

Démonstration. Soit $[\![r,s]\!] \subset [\![1,T]\!]$ avec $s+1-r \leqslant \tau_0$, et $\boldsymbol{u} \in \mathscr{P}_M$. On pose $\boldsymbol{u}_t = \boldsymbol{u}$ si $r \leqslant t \leqslant s$, et $\boldsymbol{u}_t = 0$ sinon. Si $r = 1$, on a $\|u_1\|_1 = 1$ et $\widetilde{k}(\boldsymbol{u}^T) = 0$; sinon, on a $\|\boldsymbol{u}_1\|_1 = 0$ et $\widetilde{k}(\boldsymbol{u}^T) = 1$. En particulier, on a $\|\boldsymbol{u}\|_1 + \widetilde{k}(\boldsymbol{u}^T) = 1$, $\widetilde{k}(\boldsymbol{u}^T) \leqslant 1$ et $\sum_{t=2}^T \|\boldsymbol{u}_t\|_1 - \widetilde{k}(\boldsymbol{u}^T) = \sum_{t=1}^T \|\boldsymbol{u}_t\|_1 - (\|\boldsymbol{u}_t\|_1 + \widetilde{k}(\boldsymbol{u}^T)) = \tau_0 - 1$. En injectant ces trois (in)égalités dans la borne de la proposition 3.14, on obtient la borne (3.54) annoncée. La majoration (3.55) s'en déduit comme d'habitude par le lemme 2.26. □

3.3.3 Le cas de l'algorithme MPP

Il est également possible d'étendre les résultats de la section 3.2 sur les suites d'experts qui alternent entre un petit nombre d'experts. À cet effet, nous introduisons la quantité $\widetilde{n}(\boldsymbol{u}^T)$ suivante, qui est une mesure plus souple du nombre $n(\boldsymbol{u}^T)$ de poids distincts pris par la suite $\boldsymbol{u}^T$:

$$\widetilde{n}(\boldsymbol{u}^T) = \sum_{i=1}^M \max_{1 \leqslant t \leqslant T} u_{i,t} \,. \tag{3.56}$$

La quantité $\widetilde{n}(\boldsymbol{u}^T)$ satisfait l'inégalité [17] $\widetilde{n}(\boldsymbol{u}^T) \leqslant n(\boldsymbol{u}^T)$, avec égalité si $\boldsymbol{u}^T$ est une suite i^T d'experts. De plus, si tous les poids de la suite $\boldsymbol{u}^T$ sont à support dans un sous-ensemble d'experts $S \subset [\![1,M]\!]$ de cardinal s, alors [18] $\widetilde{n}(\boldsymbol{u}^T) \leqslant s$. Ainsi, la quantité $\widetilde{n}(\boldsymbol{u}^T)$ généralise la notion de suite de poids ne prenant que peu de valeurs.

Nous allons voir que l'algorithme MPP, ou plus exactement une variante proposée par [CBGLS12] que nous allons décrire ci-dessous, permet de se comparer aux suites de poids qui sont *régulières* (au sens où $\widetilde{k}(\boldsymbol{u}^T)$ est contrôlé, comme dans la sous-section 3.3.2) et *parcimonieuses*, au sens où $\widetilde{n}(\boldsymbol{u}^T)$ est contrôlé.

Dans l'algorithme MPP (stratégie 3.4), la mise à jour des poids se fait en prenant une combinaison convexe des postérieurs passés, soit $\boldsymbol{v}_{t+1} = \sum_{q=0}^t \beta_{t+1}(q) \boldsymbol{v}_q^m$ (équation (3.35)). Plus généralement, on peut considérer le cas où $\boldsymbol{v}_{t+1} = \psi_{t+1}[(\boldsymbol{v}_q^m)_{0 \leqslant q \leqslant t}]$ est une fonction quelconque des postérieurs précédents. En pratique, comme dans la section 3.2, nous nous intéressons spécifiquement aux mises à jour de type *Fixed Share* qui attribuent au moins une part fixe $1-\alpha$ au dernier postérieur :

$$\boldsymbol{v}_{t+1} = (1-\alpha)\, \boldsymbol{v}_t^m + \frac{\alpha}{Z_t} \widehat{\boldsymbol{w}}_t \,, \tag{3.57}$$

17. En effet, si $n = n(\boldsymbol{u}^T)$ et si $\widetilde{\boldsymbol{u}}_p$, $p = 1, \ldots, n$ désignent les différents poids pris par la suite $\boldsymbol{u}^T$, on a : $\widetilde{n}(\boldsymbol{u}^T) = \sum_{i=1}^M \max_{1 \leqslant p \leqslant n} \widetilde{u}_{i,p} \leqslant \sum_{i=1}^M \sum_{p=1}^n \widetilde{u}_{i,p} = n$.
18. En notant que $\widetilde{n}(\boldsymbol{u}^T) = \sum_{i \in S} \max_{1 \leqslant t \leqslant T} u_{i,t} \leqslant \sum_{i \in S} 1 = s$.

où $\boldsymbol{w}_t \in \mathbf{R}_+^M$ est une fonction des $(\boldsymbol{v}_q^m)_{0 \leqslant q \leqslant t}$, et $Z_t = \sum_{i=1}^M \widehat{w}_{i,t}$ est la constante de normalisation. On dispose alors du théorème suivant ([CBGLS12], théorème 3) :

Théorème 3.16. *Supposons la fonction de perte η-exp-concave. Considérons la variante de l'algorithme MPP, obtenue en utilisant la mise à jour (3.57) ; on suppose de plus qu'il existe une constante $C \geqslant 1$ telle que les poids $\widehat{w}_{i,t}$ satisfassent les conditions suivantes : pour tout $i = 1, \ldots, M$, et tout $t = 1, \ldots, T$,*

$$v_{i,t}^m \leqslant \widehat{w}_{i,t} \leqslant 1 \qquad \text{et} \qquad C\, \widehat{w}_{i,t+1} \geqslant \widehat{w}_{i,t}\,. \tag{3.58}$$

Alors, pour toute suite de poids $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \in \mathbf{R}_+^M$, il vient :

$$\sum_{t=1}^T \|\boldsymbol{u}_t\|\, \ell_t - \sum_{t=1}^T \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{\widetilde{n}(\boldsymbol{u}^T)}{\eta} \log M + \frac{\widetilde{n}(\boldsymbol{u}^T) T \log C}{\eta} \\ + \frac{\widetilde{k}(\boldsymbol{u}^T)}{\eta} \log \frac{\max_{t \leqslant T} Z_t}{\alpha} + \frac{\sum_{t=2}^T \|\boldsymbol{u}_t\|_1 - \widetilde{k}(\boldsymbol{u}^T)}{\eta} \log \frac{1}{1-\alpha}\,.$$

Comme pour la proposition 3.14, la preuve du théorème 3.16 part de l'inégalité (3.53).

Remarque 3.14. *En fait, la preuve de [CBGLS12] du théorème 3.16 montre que, si l'on relaxe la seconde condition de (3.58) en $C_t\, \widehat{w}_{i,t+1} \geqslant \widehat{w}_{i,t}$ avec $C_t \geqslant 1$, i.e. si l'on autorise la constante C à dépendre du temps, la borne du théorème reste valide en remplaçant le terme $T \log C$ par $\sum_{t=1}^T \log C_t$. Ce gain en flexibilité peut être utile, comme nous l'expliquons dans la remarque 3.17.*

Nous allons maintenant, toujours suivant [CBGLS12], présenter deux choix possibles pour les poids $\widehat{\boldsymbol{w}}_t$, et déduire du théorème général 3.16 les bornes de regret correspondantes [19]. Comme nous allons le voir, ces deux choix mènent à des algorithmes efficaces.

Premier choix. Un choix élémentaire consiste à prendre, pour tout $i = 1, \ldots, M$, $\widehat{w}_{i,t} = \max_{1 \leqslant s \leqslant t} v_{i,s}^m$. Dans ce cas, les conditions (3.58) sont vérifiées pour $C = 1$. De plus, on a pour tout $t = 1, \ldots, T$, $Z_t = \sum_{i=1}^M \widehat{w}_{i,t} \leqslant \sum_{i=1}^M \sum_{s=1}^t v_{i,s}^m = t \leqslant T$. En appliquant le théorème 3.16, on en déduit directement le résultat suivant :

Corollaire 3.17. *Supposons la fonction de perte η-exp-concave. On considère la variante de l'algorithme MPP définie par la mise à jour (3.57), avec $\widehat{w}_{i,t} = \max_{1 \leqslant s \leqslant t} v_{i,s}^m$. Alors, pour tout $T \geqslant 1$ et toute suite de poids $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \in \mathscr{P}_M$ telle que $\widetilde{k}(\boldsymbol{u}^T) \leqslant k_0$ et $\widetilde{n}(\boldsymbol{u}^T) \leqslant n_0$, on a la majoration suivante :*

$$\sum_{t=1}^T \ell_t - \sum_{t=1}^T \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \leqslant \frac{1}{\eta} n_0 \log M + \frac{1}{\eta} k_0 \log \frac{1}{\alpha} + \frac{1}{\eta}(T - k_0 - 1) \log \frac{1}{1-\alpha} + \frac{1}{\eta} k_0 \log T\,. \tag{3.59}$$

Remarque 3.15. *Cette borne de regret est en fait exactement la même que la borne de regret de l'algorithme MPP de la section 3.2, avec le choix de coefficients de mélange Fixed Share to Uniform Past (corollaire 3.12).*

19. En corrigeant au passage une erreur mineure dans l'article [CBGLS12] pour le second cas, en fait sans grande conséquence puisque l'on retrouve une borne similaire.

Un des attraits du choix de $\widehat{\boldsymbol{w}}_t$ du corollaire 3.17 est qu'il conduit à un algorithme rapide. En effet, la relation de récurrence $\widehat{w}_{i,t+1} = \max(\widehat{w}_{i,t}, v_{i,t+1}^m)$ montre que l'algorithme peut être implémenté avec un espace de stockage en $O(M)$, ainsi qu'un temps de mise à jour par étape en $O(M)$. Cet algorithme présente donc les mêmes caractéristiques que l'algorithme MPP (stratégie 3.4), avec des coefficients uniformes sur le passé : une borne de regret identique, ainsi qu'une complexité équivalente.

Second choix. Une alternative au premier choix consiste à introduire un paramètre supplémentaire $\gamma \geqslant 0$, et à prendre $\widehat{w}_{i,t} = \max_{1 \leqslant s \leqslant t} e^{\gamma(s-t)} v_{i,s}^m$; ce choix satisfait bien les conditions (3.58), avec $C = e^\gamma$. Notons que l'on retrouve le choix précédent lorsque $\gamma = 0$, mais nous allons ici calibrer γ de manière optimale en fonction des paramètres du problème. On a alors, pour tout $\gamma > 0$:

$$Z_t \leqslant \sum_{i=1}^{M} \sum_{s=1}^{t} e^{\gamma(s-t)} v_{i,s}^m \leqslant \sum_{\tau \geqslant 0} e^{-\gamma\tau} = \frac{1}{1 - e^{-\gamma}} .$$

Notons qu'à ce stade [CBGLS12] commet l'erreur mineure de majorer $\frac{1}{1-e^{-\gamma}}$ par $\frac{1}{\gamma}$, alors que c'est l'inégalité inverse qui a lieu ; comme on va le voir, cela n'affecte pas sensiblement la borne obtenue. La borne de regret du théorème 3.16 devient alors, pour toute suite de poids $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T \in \mathscr{P}_M$ avec $\widetilde{k}(\boldsymbol{u}^T) \leqslant k_0$ et $\widetilde{n}(\boldsymbol{u}^T) \leqslant n_0$:

$$\frac{n_0}{\eta} \log M + \frac{n_0 T \gamma}{\eta} + \frac{k_0}{\eta} \log \frac{1/(1 - e^{-\gamma})}{\alpha} + \frac{T - 1 - k_0}{\eta} \log \frac{1}{1 - \alpha} \tag{3.60}$$

Pour calibrer γ en fonction de T, k_0 et n_0 de manière à minimiser cette borne, on utilise le résultat suivant, qui se montre par une simple étude de fonction :

Lemme* 3.18. *Le minimum sur $\mathbf{R}_+^*$ de la fonction $f(\gamma) = n_0 T\gamma + k_0 \log \frac{1}{1-e^{-\gamma}}$ est atteint pour $\gamma = \log\left(1 + \frac{k_0}{n_0 T}\right)$, et vaut :*

$$n_0 T \log\left(1 + \frac{k_0}{n_0 T}\right) + k_0 \log\left(1 + \frac{n_0 T}{k_0}\right) \leqslant k_0 \left\{ 1 + \log\left(1 + \frac{n_0 T}{k_0}\right) \right\} . \tag{3.61}$$

Le lemme 3.18, combiné à la borne (3.60), implique immédiatement le résultat suivant :

Corollaire* 3.19. *Supposons la fonction de perte η-exp-concave. On considère la variante de l'algorithme MPP définie par la mise à jour (3.57) avec $\widehat{w}_{i,t} = \max_{1 \leqslant s \leqslant t} e^{\gamma_0(s-t)} v_{i,s}^m$, où $\gamma_0 = \log\left(1 + \frac{k_0}{n_0 T}\right)$. Alors, pour tout $T \geqslant 1$ et toute suite de poids $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T \in \mathscr{P}_M$ telle que $\widetilde{k}(\boldsymbol{u}^T) \leqslant k_0$ et $\widetilde{n}(\boldsymbol{u}^T) \leqslant n_0$, on dispose de la majoration*

$$\begin{aligned} \sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \boldsymbol{u}_t \cdot \boldsymbol{\ell}_t \; &\leqslant \frac{1}{\eta} n_0 \log M + \frac{1}{\eta} k_0 \left\{ 1 + \log\left(1 + \frac{n_0 T}{k_0}\right) \right\} \\ &\quad + \frac{1}{\eta} k_0 \log \frac{1}{\alpha} + \frac{1}{\eta} (T - k_0 - 1) \log \frac{1}{1 - \alpha} . \end{aligned} \tag{3.62}$$

Remarque 3.16. *La borne de regret (3.62) du corollaire 3.19 est de très bonne qualité : en effet, il s'agit essentiellement de la borne de regret de l'algorithme MPP avec les* *coefficients Fixed Share to Decaying Past (corollaire 3.13), avec le terme* $\frac{1}{\eta}k\log(1+\log T)$ *en moins. De plus, pour une calibration adéquate de* α*, la borne* (3.62) *devient :*

$$\frac{1}{\eta}n_0\log M+\frac{1}{\eta}k_0\Big\{1+\log\Big(1+\frac{n_0 T}{k_0}\Big)\Big\}+\frac{1}{\eta}Th\Big(\frac{k_0}{T-1}\Big)\approx\frac{1}{\eta}n_0\log M+\frac{1}{\eta}2k_0\log\frac{T}{k_0}+\frac{1}{\eta}k_0\log n_0 \tag{3.63}$$

qui est la borne idéale recherchée, à ceci près que le terme $\frac{1}{\eta}k_0\log\frac{T}{k_0}$ *est à nouveau doublé, phénomène que nous expliquerons en section suivante.*

Outre sa borne de regret de très bonne qualité, le principal avantage de l'algorithme du corollaire 3.19 est sa faible complexité algorithmique. En effet, la relation de récurrence $\widehat{w}_{i,t+1}=\max(e^{-\gamma}\widehat{w}_{i,t},v^m_{i,t+1})$ montre que l'algorithme peut être implémenté avec un espace de stockage et un temps de mise à jour par étape en $O(M)$. Cet algorithme combine donc une borne de très bonne qualité et une implémentation rapide. Le prix à payer est l'introduction d'un nouveau paramètre, qu'il faut calibrer *a priori* de manière optimale par rapport à l'objectif.

Remarque 3.17. *Une façon possible de dépasser cette contrainte serait de considérer un paramètre* γ_t *dépendant du temps, et d'utiliser les poids définis par* $\widehat{w}_{i,t+1}=\max(e^{-\gamma_t}\widehat{w}_{i,t},v^m_{i,t+1})$*. On conserve alors la complexité en* $O(M)$ *de l'algorithme, et un choix idoine de la suite* γ_t *(éventuellement calibrée de manière adaptative en fonction des données) devrait permettre d'obtenir des bornes de regret satisfaisantes, en utilisant la remarque 3.14 (avec* $C_t=e^{\gamma_t}$*). Nous n'avons cependant pas suffisamment exploré cette piste pour pouvoir présenter des résultats probants dans cette direction.*

3.4 Le cadre de spécialistes

Dans cette section, nous introduisons une variante du problème de la prédiction à l'aide d'experts, introduite dans la littérature de l'apprentissage en ligne (cf. [FSSW97]) sous le terme de prédiction à partir de *spécialistes* (sous-section 3.4.1). Suivant [KAW12], nous montrons en sous-section 3.4.2 que ce cadre permet de donner une interprétation bayésienne à l'algorithme MPP introduit en section 3.2. Ceci suggère en retour une modification de l'algorithme MPP, proposée par [KAW12], qui présente une borne de regret fine tout en étant peu coûteuse d'un point de vue algorithmique.

L'article [KAW12] se plaçant dans le cas particulier de la perte logarithmique, c'est-à-dire dans le cadre bayésien, nous reformulons ses algorithmes et résultats en termes d'agrégation à poids exponentiels pour une fonction de perte quelconque [20].

3.4.1 Agrégation de spécialistes

Jusqu'ici, nous nous sommes intéressés à la prédiction à l'aide d'experts qui, à chaque étape, forment une prévision sur la prochaine valeur du signal. Une variante de ce problème consiste à considérer des experts qui se *spécialisent*, c'est-à-dire qui n'émettent des prédictions que pendant certaines périodes, et qui s'abstiennent le reste du temps. Cette

20. Que l'on suppose exp-concave comme dans tout ce chapitre, afin d'éviter d'avoir à discuter de la calibration du paramètre d'apprentissage, déjà traitée dans le chapitre 2.

complication du problème de la prédiction à l'aide d'experts, introduite par [FSSW97] sous le terme de *prédiction à l'aide de spécialistes*, est pertinente dans de nombreux contextes pratiques, où l'on dispose de modèles qui ne sont utilisés que dans certains contextes (par exemple, des modèles météorologiques saisonniers).

Problème 6 (Prédiction à l'aide de spécialistes)**.** Le problème procède comme le problème 1 de la prédiction à l'aide d'experts, à ceci près qu'à chaque étape $t \geqslant 1$, seul un sous-ensemble A_t de l'ensemble des *spécialistes* $[\![1, M]\!]$ (ou $\mathbf{N}^\star$ lorsque $M = \infty$) effectue des prédictions. Les spécialistes $i \in A_t$ sont dits *actifs*, et les autres *inactifs*.

L'objectif consiste alors à définir une stratégie pour l'apprenant qui garantisse une majoration du *regret* par rapport à chaque spécialiste i, celui-ci étant défini par :

$$R_{i,T} = \sum_{t \leqslant T \,:\, i \in A_t} (\ell_t - \ell_{i,t}) \,. \tag{3.64}$$

Remarque 3.18. *Dans ce problème, on ne suppose pas que l'apprenant dispose de la connaissance a priori des spécialistes actifs à chaque période, et les stratégies que nous allons définir n'utilisent pas cette connaissance.*

Pour obtenir des stratégies efficaces d'agrégation de spécialistes, l'idée fondamentale, introduite dans [CV09], consiste à transformer les spécialistes en experts, en attribuant aux spécialistes inactifs $i \notin A_t$ la même prédiction que l'apprenant, soit $x_{i,t} = x_t$. Alors, on peut prendre comme stratégie pour l'apprenant une stratégie d'agrégation des experts ainsi obtenus. Ainsi formulée, cette définition est circulaire ; on vérifie en fait que les conditions $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$ et $x_{i,t} = x_t$ pour $i \notin A_t$ équivalent à, pour tout $i \notin A_t$:

$$x_{i,t} = x_t = \frac{\sum_{j \in A_t} v_{j,t}\, x_{j,t}}{\sum_{j \in A_t} v_{j,t}} \,. \tag{3.65}$$

Cette idée conduit à la stratégie suivante, qui est l'analogue direct de l'algorithme d'agrégation 2.4. Notre formulation se fait en termes des poids non normalisés $w_{j,t}$, ce qui est plus commode dans le cas de spécialistes (car il faut de toutes façons renormaliser les poids à cause des spécialistes inactifs).

Stratégie 3.5 (Agrégation à poids exponentiels de spécialistes)**.**

Paramètres : la vitesse d'apprentissage $\eta > 0$, les poids initiaux des spécialistes $\boldsymbol{\omega} \in \mathscr{P}_M$, et éventuellement la fonction de combinaison **pred**.

Initialisation des poids : On pose $\boldsymbol{w}_1 = \boldsymbol{\omega}$.

Prédiction : À l'étape $t = 1, 2, \ldots$, on dispose du vecteur de poids $\boldsymbol{w}_t$, déterminé à l'étape précédente. On observe alors les prédictions $x_{i,t} \in \mathscr{X}$ des spécialistes *actifs* $i \in A_t$, que l'on combine en :

$$x_t = \frac{\sum_{i \in A_t} w_{i,t}\, x_{i,t}}{\sum_{i \in A_t} w_{i,t}} \,, \tag{3.66}$$

ou plus généralement $x_t = \mathbf{pred}(\boldsymbol{v}_t', \boldsymbol{x}_t)$, où $v_{i,t}' = \mathbf{1}_{i \in A_t} w_{i,t} / (\sum_{i \in A_t} w_{i,t})$.

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes $\ell_t = \ell(x_t, y_t)$ et $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i \in A_t$.

Mise à jour des poids : Les poids sont alors mis à jour par : $w_{i,t+1} = w_{i,t}\, e^{-\eta\, \ell_{i,t}}$ si $i \in A_t$, et $w_{i,t+1} = w_{i,t}\, e^{-\eta\, \ell_t}$ sinon.

La stratégie 3.5 admet alors la borne de regret suivante (cf. [FSSW97], théorème 1) :

Théorème 3.20. *Supposons la fonction de perte η-mélangeable. Alors, quelles que soient les valeurs du signal, les spécialistes actifs A_t et leurs prédictions, la stratégie 3.5 d'agrégation à poids exponentiels des spécialistes satisfait la borne de regret suivante :*

$$\sum_{t \leqslant T \,:\, i \in A_t} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log \frac{1}{\omega_i} \tag{3.67}$$

pour tout $T \geqslant 1$ et tout spécialiste i. Plus généralement, pour tout $\boldsymbol{u} \in \mathscr{P}_M$, il vient :

$$\sum_{i=1}^{M} u_i \sum_{t \leqslant T \,:\, i \in A_t} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \Delta(\boldsymbol{u} \,\|\, \boldsymbol{\omega})\,. \tag{3.68}$$

Démonstration. Pour tout t et tout $i \notin A_t$, on définit $x_{i,t}$ comme étant le membre de droite de l'équation (3.67). En particulier, si $i \notin A_t$, on a $\ell_{i,t} = \ell_t$, et la mise à jour de la stratégie 3.5 s'écrit : $w_{i,t+1} = w_{i,t}\, e^{-\eta\, \ell_{i,t}}$ pour tout i. De plus, le choix de x_t implique :

$$\ell_t \leqslant \frac{\sum_{i \in A_t} w_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{i \in A_t} w_{i,t}} \qquad \text{et donc} \qquad \ell_t \leqslant \frac{\sum_{i=1}^{M} w_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{i=1}^{M} w_{i,t}}$$

compte tenu de ce que $\ell_{i,t} = \ell_t$ si $i \notin A_t$. Les inégalités désirées (3.67) et (3.68) s'en déduisent classiquement, en utilisant respectivement les lemmes 2.10 et 2.20. □

Concluons cette sous-section par un petit fait général sur les poids des spécialistes sous l'algorithme 3.5. Considérons les poids normalisés $\boldsymbol{v}_t \in \mathscr{P}_M$, définis par $\boldsymbol{v}_t = \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_1}$. En particulier, on a $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t^m$, où :

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^{M} v_{j,t}\, e^{-\eta\, \ell_{j,t}}} \tag{3.69}$$

avec la convention $\ell_{i,t} = \ell_t$ pour tout $i \notin A_t$. La prédiction de l'algorithme 3.5 est alors [21]

$$x_t = \frac{\sum_{i \in A_t} w_{i,t}\, x_{i,t}}{\sum_{i \in A_t} w_{i,t}} = \frac{\sum_{i \in A_t} v_{i,t}\, x_{i,t}}{\sum_{i \in A_t} v_{i,t}}\,. \tag{3.70}$$

Fait 3.21. *Si la fonction de perte est η-exp-concave, alors $v_{i,t+1} = v_{i,t}^m \geqslant v_{i,t}$ pour tout $i \notin A_t$. De plus, pour la fonction de perte logarithmique, cette inégalité est une égalité.*

Démonstration. En effet, si $i \notin A_t$, $\ell_{i,t} = \ell_t$ et donc $v_{i,t}^m = v_{i,t} e^{-\eta\, \ell_t} / (\sum_{j=1}^{M} v_{j,t} e^{-\eta\, \ell_{j,t}})$. L'inégalité désirée découle alors du fait que $e^{-\eta\, \ell_t} \geqslant \sum_{j=1}^{M} v_{j,t} e^{-\eta\, \ell_{j,t}}$ par exp-concavité de la fonction de perte (inégalité qui est une égalité pour la perte logarithmique). □

21. Dans le cas de la combinaison convexe, pour une fonction de perte exp-concave, auquel nous nous restreignons à partir de maintenant par commodité. Tout ce qui suit reste néanmoins valide dans le cas d'une fonction mélangeable avec une fonction de combinaison **pred** quelconque, avec quelques modifications d'écriture.

3.4.2 Les « sleeping experts » : une interprétation bayésienne de l'algorithme MPP

Dans la section 3.2, nous avions introduit l'algorithme MPP de Bousquet et Warmuth (stratégie 3.4) comme un algorithme *ad hoc*, présentant des bornes de regret remarquables. Dans cette sous-section, nous présentons suivant [KAW12] une classe de *spécialistes* structurés, obtenus à partir des experts de base, permettant de traiter le problème de Freund. L'agrégation de ces spécialistes structurés permet de retomber sur l'algorithme MPP, ce qui donne une interprétation « bayésienne » à ce dernier [22].

Sleeping experts. Pour commencer, revenons sur l'interprétation bayésienne de l'algorithme FS, détaillée en sous-section 3.1.2. Nous y avons montré que l'algorithme FS peut être vu comme l'agrégation à poids exponentiels des experts « structurés » associés aux suites d'experts, avec une loi initiale sur ces suites favorisant la cohérence temporelle (en l'occurrence, une loi de Markov homogène avec de faibles probabilités de changements).

Pour voir en quoi cette approche n'est pas adaptée au problème du « sparse shifting regret », considérons l'exemple suivant, avec trois experts et trois périodes de temps successives. Durant la première période, l'expert 1 prédit bien, tandis que les experts 2 et 3 prédisent de manière médiocre ; pendant la deuxième période, c'est l'expert 2 qui prédit bien, tandis que les experts 1 et 3 prédisent mal. Ainsi, au début de la troisième période, le poids est concentré sur la séquence valant 1 sur la première période, et 2 sur la seconde [23]. Mais alors, en vertu de la propriété de Markov, cette suite aura autant tendance à transitionner vers 3 (qui a toujours mal prédit) que vers 1 (qui a bien prédit à une certaine période) ; dit autrement, la bonne performance de cette suite profitera autant au poids de l'expert 1 (qui y a contribué) qu'à celui de l'expert 3 (qui n'y a pas contribué). Cela tient pour partie au fait que la structure utilisée, les séquences d'experts, tend à perdre la trace des performances *individuelles* des experts.

Nous avons donc besoin d'une structure qui traduise les performances individuelles des experts (afin de pouvoir sélectionner un petit groupe de « bons » experts entre lesquels alterner), tout en autorisant un expert alternant des périodes très bonnes et des périodes très mauvaises à être retenu (afin de pouvoir l'employer lorsqu'il est très bon). C'est précisément ce à quoi correspond la structure de « sleeping expert » proposée par [KAW12] :

Définition 3.22 (Sleeping expert)**.** Soit $\mathcal{X}^\infty = \{0,1\}^{\mathbf{N}^*}$ l'ensemble des suites à valeurs dans $\{0,1\}$, muni de la tribu produit $\mathscr{G}^\infty$. Un élément $\chi^\infty \in \mathcal{X}^\infty$ est appelé un *circadien* (ou *cycle de sommeil*).

Étant donné l'ensemble $\mathscr{M}$ des experts de base, un *sleeping expert* est un couple (i, χ^∞), où i est un expert et χ^∞ un circadien. On définit une structure de spécialistes

22. Dans l'article [KAW12], l'approche se restreint au cas de la perte logarithmique, ce qui permet de se placer dans un cadre *littéralement* bayésien. Dans cette sous-section, nous étendons les définitions et les preuves au cas d'une fonction de perte exp-concave générale. Cette généralisation — bien peu coûteuse — a pour principal avantage de permettre de bien distinguer ce qui relève des poids initiaux sur les spécialistes structurés, de ce qui a trait aux prédictions des experts (éléments hétérogènes que la présentation bayésienne regroupe en une seule loi jointe).

23. Et éventuellement sur quelques séquences proches, qui font la transition de 1 à 2 un peu plus tôt ou plus tard.

sur l'ensemble $\mathscr{M} \times \mathcal{X}^\infty$ des sleeping experts, de la façon suivante [24] : (i, χ^∞) est actif au temps t si et seulement si $\chi_t = 1$ (on dit alors que le sleeping expert est *éveillé*), et dans ce cas sa prédiction est celle $x_{i,t}$ de l'expert de base i au temps t.

Remarque 3.19. *Le fait de considérer des spécialistes associés aux sleeping experts permet précisément de sélectionner les experts qui sont bons par intermittence. En effet, si un expert prédit bien pendant certaines périodes et médiocrement le reste du temps, le sleeping expert éveillé pendant les bonnes phases et endormi pendant les mauvaises prédira toujours bien le signal, et aura donc un poids élevé.*

Comme pour l'agrégation d'experts (voir la sous-section 3.1.2), il est possible d'étendre la stratégie 3.5 d'agrégation de spécialistes à un ensemble de spécialistes non dénombrable. Dans le cas des sleeping experts, cette stratégie s'énonce comme suit. On se donne une mesure de probabilité *a priori* μ sur l'espace $(\mathscr{M} \times \mathcal{X}^\infty, \mathscr{B} \otimes \mathscr{G}^\infty)$ des spécialistes. À l'étape t, étant donnés l'ensemble $A_t \subset \mathscr{M} \times \mathcal{X}^\infty$ des spécialistes actifs et leurs prédictions [25] $x_t(i, \chi^\infty)$, on prédit suivant :

$$x_t = \frac{\int_{A_t} x_t(i, \chi^\infty)\, e^{-\eta\, L_{t-1}(i,\chi^\infty)}\, d\mu(i, \chi^\infty)}{\int_{A_t} e^{-\eta\, L_{t-1}(i,\chi^\infty)}\, d\mu(i, \chi^\infty)} = \frac{\mathbb{E}[\, x_t(I, \mathrm{X}^\infty)\, \mathbf{1}_{(I,\mathrm{X}^\infty) \in A_t} e^{-\eta\, L_{t-1}(I,\mathrm{X}^\infty)}]}{\mathbb{E}[\mathbf{1}_{(I,\mathrm{X}^\infty) \in A_t} e^{-\eta\, L_{t-1}(I,\mathrm{X}^\infty)}]} \tag{3.71}$$

où (I, X^∞) est une variable aléatoire de loi μ. Or, les prédictions de (i, χ^∞) jusqu'au temps t ne dépendent que de (i, χ^t), donc comme dans la section 3.1.2, la stratégie d'agrégation des spécialistes (i, χ^∞) avec pour loi initiale μ prédit jusqu'au temps t comme la stratégie d'agrégation des spécialistes (i, χ^t) avec pour poids initiaux les $\omega_\mu(i, \chi^t) = \mu\big(\{(i, \chi^t)\} \times \mathcal{X}^\infty_{t+1}\big) = \mathbb{P}[(I, \mathrm{X}^t) = (i, \chi^t)]$.

Dans ce qui suit, on se donne une loi a priori μ sur l'espace $(\mathscr{M} \times \mathcal{X}^\infty, \mathscr{B} \otimes \mathscr{G}^\infty)$ des sleeping experts, dont on suppose qu'elle est une mesure produit, *i.e.* que si $(I, \mathrm{X}^\infty) \sim \mu$, alors I et X^∞ sont indépendantes. On note également $\boldsymbol{\omega} \in \mathscr{P}_M$ la marginale de μ sur $\mathscr{M}$; si M est fini, on prendra typiquement $\boldsymbol{\omega} = \frac{1}{M}\mathbf{1}$. L'indépendance s'écrit alors, pour tous $t \geqslant 1$, $\chi^t \in \mathcal{X}^t$ et tout expert $i \in \mathscr{M}$: $\omega_\mu(i, \chi^t) = \omega_i\, \omega_\mu(\chi^t)$.

Regret et application au problème de Freund. Nous allons maintenant voir comment l'agrégation des spécialistes associés aux « sleeping experts » permet d'obtenir des bornes de regret face aux séquences alternant entre un petit nombre d'expert.

Soit $T \geqslant 1$, et $i^T = (i_1, \dots, i_T)$ une suite d'experts appartenant à l'ensemble $\{\widetilde{i}_1, \dots, \widetilde{i}_n\}$. Pour $p = 1, \dots, n$, on définit $\chi_p^T \in \mathcal{X}^T$ par $\chi_{p,t} = 1$ si et seulement si $i_t = \widetilde{i}_p$. Ainsi, les circadiens $\chi_1^T, \dots, \chi_n^T$ partitionnent la période $[\![1, T]\!]$, en fonction de la valeur prise par la suite i^T. Pour tout $p \geqslant 1$, le regret de l'apprenant par rapport au spécialiste associé au sleeping expert $(\widetilde{i}_p, \chi_p^T)$ vaut alors :

$$\sum_{t \leqslant T\,:\, (\widetilde{i}_p, \chi_p^T) \in A_t} (\ell_t - \ell_t(\widetilde{i}_p, \chi_p)) = \sum_{t \leqslant T\,:\, i_t = \widetilde{i}_p} (\ell_t - \ell_{\widetilde{i}_p, t}) \tag{3.72}$$

24. En effet, une structure de spécialistes est déterminée par l'ensemble des spécialistes actifs à chaque étape, et par les prédictions de ceux-ci.

25. Dans le cas général de l'agrégation de spécialistes, il faut supposer l'ensemble A_t des spécialistes et les prédictions de ceux-ci mesurables ; dans le cas des spécialistes associés aux sleeping experts, ces conditions sont automatiquement vérifiées en vertu de la définition 3.22.

En sommant sur $p = 1, \ldots, n$, on obtient :

$$\sum_{t=1}^{T} (\ell_t - \ell_{i_t,t}) = \sum_{p=1}^{n} \sum_{t \leqslant T \,:\, (\widetilde{i}_p, \chi_p^T) \in A_t} (\ell_t - \ell_t(\widetilde{i}_p, \chi_p)) \,; \tag{3.73}$$

en d'autres termes, le regret par rapport à la suite d'experts i^T est égal à la somme des regrets par rapport aux (spécialistes associés aux) sleeping experts $(\widetilde{i}_p, \chi_p)$, ce qui établit le lien entre le problème de Freund et les sleeping experts.

Soit maintenant μ une loi a priori sur les sleeping experts. Comme on l'a vu, l'agrégation à poids exponentiels des spécialistes associés aux sleeping experts (i, χ^∞) de loi initiale μ coïncide jusqu'au temps T avec la stratégie 3.5 d'agrégation des sleeping experts (i, χ^T), de poids initiaux $\omega_\mu(i, \chi^T)$. Le théorème 3.20 permet alors de majorer le regret (3.73), de deux manières différentes. La première consiste à majorer le regret par rapport à chacun des spécialistes $(\widetilde{i}_p, \chi_p)$ en utilisant la borne (3.67), puis à sommer sur $p = 1, \ldots, n$, ce qui donne :

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \sum_{p=1}^{n} \log \frac{1}{\omega_\mu(\widetilde{i}_p, \chi_p^T)} \,. \tag{3.74}$$

La seconde, plus fine, consiste à réécrire (3.73) sous la forme :

$$L_T - L_T(i^T) = n \sum_{(i,\chi^T)} u(i, \chi^T) \sum_{t \leqslant T \,:\, (i,\chi^T) \in A_t} (\ell_t - \ell_t(i, \chi^T)) \,, \tag{3.75}$$

où $u = \big(u(i, \chi^T)\big)_{i,\chi^T}$ est la mesure de probabilité sur l'ensemble des spécialistes (i, χ^T) qui est uniforme sur le sous-ensemble $\{(\widetilde{i}_p, \chi_p^T) \mid 1 \leqslant p \leqslant n\}$. La borne (3.68) du théorème 3.20 implique alors :

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} n \, \Delta(u(i, \chi^T) \,\|\, \omega_\mu(i, \chi^T)) = \frac{1}{\eta} \sum_{p=1}^{n} \log \left(\frac{1/n}{\omega_\mu(\widetilde{i}_p, \chi^T)} \right) , \tag{3.76}$$

d'où un gain d'un terme additif de $n \log n$ par rapport à la borne (3.74). Pour voir comment s'incarne cette borne, écrivons $\omega_\mu(i, \chi^T) = \omega_i \, \omega_\mu(\chi^T)$ (par hypothèse d'indépendance sous μ), et supposons de plus M fini et $\omega_i = \frac{1}{M}$. La borne (3.76) s'écrit alors :

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} n \log \frac{M}{n} + \frac{1}{\eta} \sum_{p=1}^{n} \log \frac{1}{\omega_\mu(\chi^T)} \,. \tag{3.77}$$

Notons que cette borne remplace le terme $\frac{1}{\eta} n \log M$, présent dans toutes les bornes que l'on a obtenues pour l'algorithme MPP, par le terme plus fin $\frac{1}{\eta} n \log \frac{M}{n}$.

L'algorithme MPP. Considérons des coefficients $\beta_t(\cdot)$ comme dans la définition 3.4 de l'algorithme MPP. On définit la loi a priori μ sur $(\mathscr{M} \times \mathcal{X}^\infty, \mathscr{B} \otimes \mathscr{G}^\infty)$ de la façon suivante. La mesure est une mesure produit, soit $\mu(i, \chi^\infty) = \frac{1}{M} \mu(\chi^\infty)$. De plus, la marginale de la mesure μ sur $(\mathscr{X}^\infty, \mathscr{G}^\infty)$ est à support dans l'ensemble des suites χ^∞ qui ne valent 1 qu'à un nombre fini d'instants ; en notant $s_1 < \cdots < s_J$ ces instants, $\mu(\chi^\infty)$ est défini par :

$$\mu(\chi^\infty) = \frac{1}{s_J(s_J+1)} \prod_{j=1}^{J} \beta_{s_j}(s_{j-1}) \,. \tag{3.78}$$

En particulier, sous cette loi, $\beta_t(s)$ correspond à la probabilité que le dernier instant d'éveil soit s, conditionnellement au fait que l'on est éveillé au temps t. On montre alors (cf. [KAW12], théorème 2) que l'algorithme d'agrégation des sleeping experts avec pour mesure initiale μ se simplifie en l'algorithme MPP, ce qui fournit une interprétation bayésienne à celui-ci.

Remarque 3.20. *Le doublement du terme $\frac{1}{\eta} k \log \frac{T}{k}$ dans la borne de regret de l'algorithme MPP par rapport à la borne optimale, que nous avons constaté sans pouvoir l'expliquer jusqu'ici, s'explique dans notre cadre. En effet, le terme $k \log \frac{T}{k}$ correspond dans la borne optimale au nombre de bits nécessaires à encoder les $\binom{T}{k}$ instants de rupture. Dans notre cadre, chacun de ces instants intervient* deux fois*, en tant que mise en veille de l'expert précédent et en tant qu'éveil du nouvel expert ; le terme supplémentaire $k \log \frac{T}{k}$ encouru correspond alors au* coût de la coordination *entre les différents spécialistes.*

Une amélioration de l'algorithme MPP. À présent que nous avons donné une interprétation de l'algorithme MPP en termes d'agrégation de sleeping experts avec une loi a priori bien choisie, il est naturel de se demander si un choix plus judicieux de la loi a priori conduit à un algorithme plus simple.

Comme dans le cas de l'agrégation des suites d'experts, comme nous l'avons vu en section 3.1.2, il s'avère que le choix pour μ d'une loi de Markov (homogène ou non) conduit à un algorithme qui se simplifie remarquablement. On aboutit alors (cf. [KAW12]) à un algorithme qui ne nécessite de stocker que $2M$ poids (un poids « éveillé » et un poids « endormi » pour chaque expert), et dont la mise à jour se fait en temps $O(M)$. De plus, une calibration des probabilités de transition en fonction de T, k et n conduit à une borne de regret optimale, au terme supplémentaire $\frac{1}{\eta} k \log \frac{T}{k}$ près.

Chapitre 4

Prédiction à l'aide d'un nombre croissant d'experts

Dans le chapitre 3, nous avons complexifié le problème de la comparaison au meilleur expert, étudié en détail au chapitre 2, en cherchant à se comparer à la meilleure séquence d'experts (sous-section 3.1), voire à la meilleure séquence parmi un petit nombre d'experts (sous-section 3.2). Pour ces deux problèmes, nous avons présenté des algorithmes efficaces aux bornes de regret essentiellement optimales, au moins modulo une bonne calibration des paramètres. Ces algorithmes, qui permettent de s'adapter aux situations où le meilleur expert varie au cours du temps, fournissent donc des procédures d'agrégation remarquablement flexibles.

Cependant, dans ces raffinements successifs, il s'agit de se comparer à un expert variable, mais appartenant toujours à une classe d'experts fixe. Pour obtenir davantage de flexibilité, nous nous intéressons dans ce chapitre à un autre raffinement du problème, où l'ensemble des experts lui-même s'enrichit au cours du temps. Bien sûr, les méthodes des chapitres 2 et 3 sont valables lorsque l'on utilise d'emblée une infinité d'experts (cas $M = \infty$), et l'on pourrait se demander s'il n'est pas plus simple de considérer directement l'ensemble infini de tous les experts que nous souhaiterions utiliser. Cette approche est en fait insuffisante, pour deux raisons.

Tout d'abord, il existe en pratique des experts qui n'émettent des prédictions qu'à partir d'un certain moment (par exemple, les algorithmes d'apprentissage qui cherchent à prédire la prochaine valeur du signal à partir des $k \geqslant 1$ valeurs précédentes). Par ailleurs, les stratégies d'agrégation d'une infinité d'experts quelconques ne sont en fait pas implémentables en pratique, puisqu'elles nécessitent de manipuler une infinité de prédictions d'experts, qu'il n'est possible de combiner que lorsqu'elles admettent une forme particulière. Nous cherchons ainsi à définir des stratégies d'agrégation d'un nombre croissant d'experts, à la fois simples à mettre en œuvre et avec de bonnes garanties de regret.

Ce chapitre est une contribution originale de notre mémoire, quoique les méthodes utilisées s'appuient sur les stratégies et les résultats des chapitres précédents, tirés d'articles existants.

4.1 Bornes de regret pour la prédiction avec un nombre croissant d'experts

Dans cette section, après avoir introduit le problème (sous-section 4.1.1), nous montrons comment l'agrégation de spécialistes, introduite en section 3.4, fournit une stratégie naturelle lorsque l'horizon de temps est fixé (sous-section 4.1.2). Nous montrons ensuite que la structure particulière du problème permet en fait de simuler l'agrégation d'une infinité de spécialistes (sous-section 4.1.3). Enfin, une observation élémentaire sur cet algorithme permet de généraliser les bornes de regret au cas où les poids initiaux $(\omega_i)_{i\geqslant 1}$ sur les experts ne sont pas sommables (sous-section 4.1.4).

4.1.1 Cadre du problème

Donnons-nous une suite croissante $(M_t)_{t\geqslant 1}$, correspondant au nombre d'experts présents à chaque échéance. À l'étape t, les experts $i = 1, \ldots, M_t$ émettent chacun une prédiction $x_{i,t} \in \mathscr{X}$. En particulier, à l'étape t, le nombre de nouveaux experts est $m_t = M_t - M_{t-1}$, correspondant aux entrants $i = M_{t-1} + 1, \ldots, M_t$. Pour tout $i \geqslant 1$, on note également $t_i = \inf\{t \geqslant 1 \mid i \leqslant M_t\}$ le temps d'entrée de l'expert i. Nous cherchons alors à traiter le problème suivant :

Problème 7. Définir une stratégie d'agrégation des experts en nombre croissant, qui permette de contrôler le regret par rapport à chaque expert i donné par :

$$R_{i,T} = \sum_{t=t_i}^{T} (\ell_t - \ell_{i,t}) \,. \tag{4.1}$$

Remarque 4.1. *Nous aimerions définir une stratégie qui ne nécessite pas de connaître à l'avance la suite M_t, afin de pouvoir gagner en souplesse et éventuellement de pouvoir le nombre d'experts ajoutés à chaque étape en fonction des données.*

À notre connaissance, le problème de l'agrégation d'un nombre croissant d'experts quelconques n'est pas traité tel quel dans la littérature. Dans l'article [GLM99], dédié à la prédiction des processus stationnaires ergodiques, les auteurs ont recours à une forme de « doubling trick », où de nouveaux experts sont ajoutés et les poids sont réinitialisés à chaque puissance de 2. Cela n'est cependant pas adapté à notre problème, où de nouveaux experts sont ajoutés à chaque étape.

4.1.2 Analyse à horizon de temps T fixé

Pour commencer, supposons T fixé et $M = M_T$ connu à l'avance (nous relaxerons ces hypothèses par la suite). Le problème 7 est alors un cas particulier du problème 6 d'agrégation des spécialistes $i = 1, \ldots, M$, avec $A_t = [\![1, M_t]\!]$ pour $t = 1, \ldots, T$. Nous pouvons alors directement appliquer la stratégie 3.5 d'agrégation des spécialistes, à partir d'un vecteur de poids initiaux $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M) \in \mathscr{P}_M$. Notons également que, dans ce cas particulier, le spécialiste i est inactif jusqu'au temps t_i, on peut donc n'introduire son poids qu'à cet instant, où il vaut $w_{i,t_i} = \omega_i \, e^{-\eta \, L_{t_i - 1}}$ (puisque les poids des spécialistes non inactifs sont mis à jour à partir des pertes de l'apprenant). Compte tenu de cette observation, l'algorithme 3.5 d'agrégation des spécialistes s'écrit alors :

Stratégie 4.1 (Agrégation d'un nombre croissant d'experts à horizon T).

Paramètres : la vitesse d'apprentissage $\eta > 0$, les poids initiaux des spécialistes $\boldsymbol{\omega} \in \mathscr{P}_M$, et éventuellement la fonction de combinaison **pred**.

Initialisation des poids : On pose $w_{i,1} = \omega_i$ pour $i = 1, \dots, M_1$.

Prédiction : À l'étape $t = 1, \dots, T$, on dispose des poids $w_{i,t}$ ($1 \leqslant i \leqslant M_t$), déterminés à l'étape précédente. On observe alors les prédictions $x_{i,t} \in \mathscr{X}$ des experts $i = 1, \dots, M_t$, que l'on combine en :

$$x_t = \frac{\sum_{i=1}^{M_t} w_{i,t}\, x_{i,t}}{\sum_{i=1}^{M_t} w_{i,t}}, \tag{4.2}$$

ou plus généralement $x_t = \mathbf{pred}(\boldsymbol{v}'_t, \boldsymbol{x}_t)$, où $v'_{i,t} = \mathbf{1}_{i \leqslant M_t} w_{i,t} / (\sum_{i \leqslant M_t} w_{i,t})$.

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes $\ell_t = \ell(x_t, y_t)$ et $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i = 1, \dots, M_t$.

Mise à jour des poids : Les poids sont alors mis à jour par : $w_{i,t+1} = w_{i,t}\, e^{-\eta\, \ell_{i,t}}$ pour $1 \leqslant i \leqslant M_t$. De plus, on introduit les poids $w_{i,t+1} = \omega_i\, e^{-\eta\, L_t}$ pour $M_t + 1 \leqslant i \leqslant M_{t+1}$.

De la borne de regret générale pour l'agrégation de spécialistes (théorème 3.20), on déduit le résultat suivant :

Proposition* 4.1. *Supposons la fonction de perte η-mélangeable. Quelles que soient les valeurs successives du signal et les prédictions des experts, la stratégie 4.1 avec les poids initiaux $\omega_1, \dots, \omega_{M_T}$ vérifie la borne de regret suivante : pour tout $i = 1, \dots, M_T$*

$$\sum_{t=t_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log \frac{1}{\omega_i}. \tag{4.3}$$

En particulier, si $\omega_i = \frac{1}{M_T}$ pour $i = 1, \dots, M_T$, la borne précédente devient $\frac{1}{\eta} \log M_T$.

Remarque 4.2. *Notons que le résultat précédent est à horizon de temps T fixé, en particulier parce que le nombre de spécialistes M_T qu'agrège la stratégie 4.1 dépend de T. Nous allons voir dans ce qui suit qu'il est en fait possible de définir des stratégies « anytime ».*

4.1.3 Stratégie « anytime » et agrégation d'une infinité de spécialistes

Pour obtenir une stratégie « anytime », l'idée naturelle consiste à agréger l'infinité des spécialistes $i \geqslant 1$ (avec $A_t = \{1, \dots, M_t\}$ pour tout $t \geqslant 1$) plutôt que les spécialistes $i = 1, \dots, M_T$. En général, l'agrégation d'une infinité d'experts (ou de spécialistes) n'est pas implémentable avec des ressources finies, comme on l'a observé plus haut. Cependant, la structure particulière des growing experts [1] permet d'aboutir à un algorithme « fini », en introduisant le poids de chaque spécialiste i à son temps d'entrée t_i (comme on l'a fait dans la stratégie 4.1). Cela aboutit à l'algorithme suivant :

1. Spécifiquement, le fait qu'à chaque étape seul un nombre fini de spécialistes est actif.

Stratégie 4.2 (Agrégation « anytime » d'un nombre croissant d'experts).

Paramètres : la vitesse d'apprentissage $\eta > 0$, les poids initiaux $\omega_i > 0$ ($i \geqslant 1$), et éventuellement la fonction de combinaison **pred**.

Initialisation des poids : On pose $w_{i,1} = \omega_i$ pour $i = 1, \dots, M_1$.

Prédiction : À l'étape $t = 1, \dots, T$, on dispose des poids $w_{i,t}$ ($1 \leqslant i \leqslant M_t$), déterminés à l'étape précédente. On observe alors les prédictions $x_{i,t} \in \mathscr{X}$ des experts $i = 1, \dots, M_t$, que l'on combine en :

$$x_t = \frac{\sum_{i=1}^{M_t} w_{i,t}\, x_{i,t}}{\sum_{i=1}^{M_t} w_{i,t}}\,, \tag{4.4}$$

ou plus généralement $x_t = \mathbf{pred}(\boldsymbol{v}_t', \boldsymbol{x}_t)$, où $v_{i,t}' = \mathbf{1}_{i \leqslant M_t} w_{i,t} / (\sum_{i \leqslant M_t} w_{i,t})$.

Observation du signal : La valeur du signal $y_t \in \mathscr{Y}$ est alors révélée, et induit les pertes $\ell_t = \ell(x_t, y_t)$ et $\ell_{i,t} = \ell(x_{i,t}, y_t)$ pour tout $i = 1, \dots, M_t$.

Mise à jour des poids : Les poids sont alors mis à jour par : $w_{i,t+1} = w_{i,t}\, e^{-\eta\, \ell_{i,t}}$ pour $1 \leqslant i \leqslant M_t$. De plus, on introduit les poids $w_{i,t+1} = \omega_i\, e^{-\eta\, L_t}$ pour $M_t + 1 \leqslant i \leqslant M_{t+1}$.

Remarque 4.3. *Comme on le voit, il s'agit essentiellement de la stratégie 4.1, à ceci près que l'on ne s'arrête plus arbitrairement à T et $M = M_T$, et que l'on prend ici $M = \infty$ et donc des poids initiaux $\boldsymbol{\omega} \in \mathscr{P}_\infty$.*

À nouveau, la borne de regret du théorème 3.20 pour l'agrégation de spécialistes fournit la borne de regret suivante (qui est cette fois « anytime ») :

Proposition* 4.2. *Supposons la fonction de perte η-mélangeable ; soit $\boldsymbol{\omega} \in \mathscr{P}_\infty$ un vecteur de probabilité sur $\mathbf{N}^\star$. Pour tout $T \geqslant 1$, la stratégie 4.2 de poids initiaux ω_i garantit la borne de regret suivante :*

$$\sum_{t=t_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log \frac{1}{\omega_i}\,. \tag{4.5}$$

Voyons à présent comment s'instancie la borne (4.5) en fonction du choix des poids ω_i. Comme $\boldsymbol{\omega} \in \mathscr{P}_\infty$, il n'est plus possible de prendre des poids uniformes.

Exemple 4.1. Une première façon de choisir les poids ω_i consiste à prendre un loi de probabilité $\boldsymbol{\omega}$ sur $\mathbf{N}^*$ déterminée à l'avance, et qui décroisse suffisamment lentement pour être « approximativement uniforme ». À cet effet, on peut notamment avoir recours aux poids donnés par les exemples 2.7, 2.8 et 2.9. Une propriété remarquable de cette façon de choisir les poids est que la borne obtenue (4.5) ne dépend alors pas du temps. Par exemple, pour le choix de l'exemple 2.9, on obtient [2] :

$$\sum_{t=t_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log(i+1) + \frac{2}{\eta} \log\log(i+2) + \frac{4}{3\eta}\,. \tag{4.6}$$

En revanche, ce choix de poids présente l'inconvénient de distinguer artificiellement des experts introduits au même moment en fonction de leur indice, ce qui est peu satisfaisant.

2. En observant que $\omega_i = \log 2 \left(\frac{1}{\log(i+1)} - \frac{1}{\log(i+2)}\right) = \log 2\, \frac{\log(1+1/(i+1))}{\log(i+1)\log(i+2)} \geqslant \log 2\, \frac{\log(3/2)\cdot 1/(i+1)}{\log^2(i+2)}$ par concavité de log et comme $1/(i+1) \leqslant 1/2$, puis en notant que $-\log(\log 2 \cdot \log(3/2)) \simeq 1.27 \leqslant 4/3$.

Exemple 4.2. Il est plus naturel d'avoir des poids initiaux ω_i ne dépendant que du temps d'entrée t_i, de manière à traiter de manière identique tous les experts arrivant au même moment. Cela est en fait possible même lorsque le nombre d'experts présent à chaque étape (*i.e.* les valeurs des M_t, donc des t_i) n'est pas connu à priori. En effet, dans la stratégie 4.2, le poids ω_i n'est utilisé qu'à l'échéance t_i, il peut donc être défini à ce moment-là par : $\omega_i = \frac{1}{m_{t_i}}\nu_{t_i}$, où la suite ν_t est définie à l'avance et vérifie $\sum_{t=1}^{\infty} \nu_t = 1$. On a alors bien $\sum_{i=1}^{\infty} \omega_i = 1$, et la borne (4.5) devient :

$$\sum_{t=t_i}^{T}(\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta}\log m_{t_i} + \frac{1}{\eta}\log\frac{1}{\nu_{t_i}}\,, \tag{4.7}$$

ce qui avec un choix de ν_t décroissant lentement (par exemple $\nu_t = \log 2\,(\frac{1}{\log(t+1)} - \frac{1}{\log(t+2)})$), donne une borne de l'ordre de $\frac{1}{\eta}\log(t_i m_{t_i})$.

4.1.4 Cas général : extension à des poids non sommables

Nous allons maintenant voir qu'il est possible de généraliser les résultats de la section précédente au cas de poids initiaux $\omega_i > 0$ non nécessairement sommables. L'observation fondamentale est la suivante : *la stratégie 4.2 est définie pour des poids $\omega_i > 0$ quelconques, et coïncide*[3] *jusqu'au temps $T \geqslant 1$ avec la stratégie 4.1 de poids initiaux (normalisés) $\omega_1/(\sum_{i=1}^{M_T}\omega_i), \dots, \omega_{M_T}/(\sum_{i=1}^{M_T}\omega_i)$.*

De cette observation, et de la proposition 4.1, on déduit le théorème général suivant, qui étend les propositions 4.1 et 4.2 :

Théorème* 4.3. *Supposons la fonction de perte η-mélangeable. On se donne une suite de poids $\omega_i > 0$ quelconques. Pour tout $T \geqslant 1$, la stratégie 4.2 de poids initiaux ω_i garantit la borne de regret suivante : pour tout $i \geqslant 1$,*

$$\sum_{t=t_i}^{T}(\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta}\log\left(\frac{1}{\omega_i}\sum_{j=1}^{M_T}\omega_j\right). \tag{4.8}$$

Pour apprécier cette borne, voyons ce qu'elle donne pour différents choix de poids ω_i.

Exemple 4.3. Une première façon de choisir les poids ω_i, indépendamment de la connaissance de la suite M_t, est de leur attribuer une valeur déterminée à l'avance.

- Pour le choix $\omega_i = 1$ pour tout $i \geqslant 1$, la borne de regret (4.8) devient : $\frac{1}{\eta}\log M_T$. Il s'agit donc d'une version « anytime » de la borne de regret de la proposition 4.1 avec $\omega_i = \frac{1}{M_T}$.
- À l'inverse, en prenant les ω_i sommables (par exemple de somme 1), le théorème 4.3 permet de retrouver la borne (4.5). En particulier, le choix des poids de l'exemple 2.9 conduit à la borne (4.6).

3. En effet, multiplier tous les ω_i par un même facteur ne modifie pas les prédictions de la stratégie 4.2, car alors tous les $w_{i,t}$ sont multipliés par cette constante, et donc x_t reste inchangé par (4.4).

Exemple 4.4. Comme on l'a observé précédemment, il est plus naturel d'avoir recours à des poids ω_i ne dépendant que de l'instant d'entrée, soit $\omega_i = \upsilon_{t_i} = \frac{1}{m_{t_i}}\nu_{t_i}$ (m_{t_i} étant connu à l'instant t_i, donc à l'entrée de l'expert i). La borne du théorème 4.3 s'écrit alors :

$$\sum_{t=t_i}^{T}(\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta}\log\left(\frac{1}{\upsilon_{t_i}}\sum_{t=1}^{T} m_t\upsilon_t\right) = \frac{1}{\eta}\log\left(\frac{m_{t_i}}{\nu_{t_i}}\sum_{t=1}^{T}\nu_t\right) \tag{4.9}$$

Dans ce cas, deux choix sont possibles : on donne une valeur déterminée à l'avance soit à la suite υ_t (poids attribué à chaque expert entrant à l'instant t), soit à la suite ν_t (poids total attribué à l'ensemble des experts arrivant au temps t). La première option [4] permet d'éviter de pénaliser les experts ajoutés au même instant qu'un grand nombre d'experts, mais la seconde conduit à des bornes de regret plus simples à expliciter :

- Le cas où l'on prend pour ν_t une suite sommable a déjà été discuté dans l'exemple 4.2 ; comme on l'a observé, le choix d'une suite ν_t décroissant lentement conduit à une borne de l'ordre de $\frac{1}{\eta}\log m_{t_i} + \frac{1}{\eta}\log t_i$ (avec un terme supplémentaire en $2\log\log t_i$ plus une constante si ν_t est la suite de l'exemple 2.9).
- Si $\nu_t = 1$ pour tout $t \geqslant 1$, la borne (4.9) devient $\frac{1}{\eta}\log m_{t_i} + \frac{1}{\eta}\log T$. Plus généralement, si $0 \leqslant \alpha < 1$, le choix $\nu_t = t^{-\alpha}$ conduit à la borne $\frac{1}{\eta}\log m_{t_i} + \frac{1}{\eta}\alpha\log t_i + \frac{1}{\eta}(1-\alpha)\log T$.
- Un bon compromis entre simplicité et qualité de la borne de regret est atteint en prenant $\nu_t = \frac{1}{t}$, car alors on obtient [5] une borne de $\frac{1}{\eta}\log m_{t_i} + \frac{1}{\eta}\log t_i + \frac{1}{\eta}\log\log(eT)$.

4. Dont un cas particulier notable est le choix $\omega_i = \upsilon_{t_i} = 1$ pour tout i.

5. Via l'inégalité $\sum_{t=1}^{T}\frac{1}{t} \leqslant 1 + \log T$, déjà rencontrée dans ce texte, et qui se montre par une comparaison somme-intégrale.

Annexes

Annexe A

Inégalités de concentration

A.1 Inégalité de Hoeffding

En section 2.2.1, nous avons fait usage de l'inégalité suivante, qui permet de ramener l'étude des fonctions de pertes convexes bornées à celle des fonctions de pertes exp-concaves :

Lemme A.1 (Hoeffding)**.** *Soit X une variable aléatoire réelle telle que $a \leqslant X \leqslant b$ p.s., où a et b sont deux réels. Pour tout $s \in \mathbf{R}$, l'inégalité suivante est vérifiée :*

$$\log \mathbb{E}\big[e^{sX}\big] \leqslant s\,\mathbb{E}[X] + \frac{(b-a)^2}{8} s^2 \,. \tag{A.1}$$

Démonstration. Définissons la fonction $\psi(s) = \log \mathbb{E}\big[e^{sX}\big]$. Par convergence dominée (X étant bornée), ψ est indéfiniment dérivable sur $\mathbf{R}$. De plus, pour tout $s \in \mathbf{R}$, une dérivation sous le signe intégral donne

$$\psi''(s) = \frac{\mathbb{E}\big[X^2 e^{sX}\big]}{\mathbb{E}\big[e^{sX}\big]} - \left(\frac{\mathbb{E}\big[X e^{sX}\big]}{\mathbb{E}\big[e^{sX}\big]}\right)^2$$

ce qui correspond à la variance de X sous la probabilité $\widetilde{\mathbb{P}}_s = \frac{e^{sX}}{\mathbb{E}[e^{sX}]} \cdot \mathbb{P}$. Or $a \leqslant X \leqslant b$ $\mathbb{P}$-p.s., donc aussi $\widetilde{\mathbb{P}}_s$-p.s., ce qui implique[1] que cette variance est majorée par $\frac{(b-a)^2}{4}$. L'inégalité $\psi''(s) \leqslant \frac{(b-a)^2}{4}$, compte tenu de ce que $\psi(0) = 0$ et $\psi(0)' = \mathbb{E}[X]$, implique alors la majoration (A.1) par l'inégalité de Taylor. □

A.2 Une inégalité « poissonienne »

Lemme A.2. *Soit X une variable aléatoire à valeurs dans $[0,1]$. Pour tout $s \in \mathbf{R}$, on a :*

$$\log \mathbb{E}\left[e^{sX}\right] \leqslant (e^s - 1)\,\mathbb{E}\,X \,. \tag{A.2}$$

1. En se ramenant par une transformation affine à $[0,1]$, puis en remarquant si $0 \leqslant X \leqslant 1$, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \leqslant \mathbb{E}[X] - \mathbb{E}[X]^2 = \mathbb{E}[X](1-\mathbb{E}[X]) \leqslant \frac{1}{4}$.

Démonstration. Par convexité de la fonction $x \mapsto e^{sx}$, on a pour tout $x \in [0,1]$: $e^{sx} \leqslant (1-x) + x\, e^s = 1 + (e^s - 1)\, x$, d'où :

$$\log \mathbb{E}\left[e^{sX}\right] \leqslant \log\big(1 + (e^s - 1)\, \mathbb{E}\, X\big) \leqslant (e^s - 1)\, \mathbb{E}\, X\,,$$

où l'on a utilisé le fait que $\log(1+y) \leqslant y$ pour tout $y > -1$. $\square$

Bibliographie

[Bou03] Olivier Bousquet : A note on parameter tuning for on-line shifting algorithms. Rapport technique, Max Planck Institute for Biological Cybernetics, 2003.

[BW02] Olivier Bousquet et Manfred K. Warmuth : Tracking a small set of experts by mixing past posteriors. *The Journal of Machine Learning Research*, 3:363–396, 2002.

[CBFH+97] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire et Manfred K. Warmuth : How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

[CBGLS12] Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi et Gilles Stoltz : Mirror descent meets fixed share (and feels no regret). *In Advances in Neural Information Processing Systems 25*, pages 980–988. Curran Associates, Inc., 2012.

[CBL06] Nicolò Cesa-Bianchi et Gábor Lugosi : *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, New York (USA), 2006.

[CBMS07] Nicolò Cesa-Bianchi, Yishay Mansour et Gilles Stoltz : Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

[CT06] Thomas M. Cover et Joy A. Thomas : *Elements of Information Theory, 2nd Edition*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, USA, 2006.

[CV09] Alexey Chernov et Vladimir Vovk : Prediction with expert evaluators' advice. *In Proceedings of the 20th international conference on Algorithmic learning theory*, ALT '09, pages 8–22, Berlin, Heidelberg, 2009. Springer-Verlag.

[dRvEGK14] Steven de Rooij, Tim van Erven, Peter Grünwald et Wouter M. Koolen : Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.

[FS97] Yoav Freund et Robert E. Schapire : A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[FSSW97] Yoav Freund, Robert E. Schapire, Yoram Singer et Manfred K. Warmuth : Using and combining predictors that specialize. *In Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, pages 334–343, New York, USA, 1997. ACM.

[GLM99] László Györfi, Gábor Lugosi et Gustáv Morvai : A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650, 1999.

[HAK07] Elad Hazan, Amit Agarwal et Satyen Kale : Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[Haz16] Elad Hazan : Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[HKW98] D. Haussler, J. Kivinen et M. K. Warmuth : Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, Sep 1998.

[HS09] Elad Hazan et Comandur Seshadhri : Efficient learning algorithms for changing environments. *In Proceedings of the 26th annual international conference on machine learning*, ICML '09, pages 393–400, 2009.

[HW98] Mark Herbster et Manfred K. Warmuth : Tracking the best expert. *Machine Learning*, 32(2):151–178, août 1998.

[KAW12] Wouter M. Koolen, Dmitry Adamskiy et Manfred K. Warmuth : Putting bayes to sleep. *In Advances in Neural Information Processing Systems 25*, pages 135–143. Curran Associates, Inc., 2012.

[KvEG14] Wouter M. Koolen, Tim van Erven et Peter Grünwald : Learning the learning rate for prediction with expert advice. *In Advances in Neural Information Processing Systems 27*, pages 2294–2302. Curran Associates, Inc., 2014.

[KW97] Jyrki Kivinen et Manfred K. Warmuth : Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[KW99] Jyrki Kivinen et Manfred K. Warmuth : Averaging expert predictions. *In Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT 99)*, pages 153–167. Springer, 1999.

[MF98] Neri Merhav et Meir Feder : Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[RS15] Alexander Rakhlin et Karthik Sridharan : Online nonparametric regression with general loss functions. *Preprint Arxiv*, 2015.

[Sto10] Gilles Stoltz : Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique*, 151(2):66–106, 2010.

[Vov98] Vladimir Vovk : A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[Vov99] Vladimir Vovk : Derandomizing stochastic prediction strategies. *Machine Learning*, 35(3):247–282, 1999.

[Zin03] Martin Zinkevich : Online convex programming and generalized infinitesimal gradient ascent. *In Proceedings of the 20th International Conference on Machine Learning, ICML 2003*, pages 928–936, 2003.