Le problème de l'apprentissage statistique

La plupart des problèmes étudiés dans cette thèse sont des variantes du problème général de l'apprentissage statistique. L'apprentissage statistique (Vapnik, 1998; Friedman et al., 2001; Bousquet et al., 2004; Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014) fournit un cadre commun aux problèmes de prédiction, où le but est de prédire certaines quantités inconnues à partir d'un jeu de données constitué d'observations de même nature. La formalisation de ce problème repose sur une modélisation aléatoire du phénomène étudié, qui permet de relier les observations disponibles aux réalisations non observées des quantités à prédire.

1.1.1 Formulation générale

Le problème de l'apprentissage statistique peut être formulé de façon générale comme suit. Soit $\mathcal{Z}$ un espace mesurable appelé *espace d'observations*, et $\mathcal{G}$ un espace mesurable dont les éléments sont appelés *prédicteurs*. L'adéquation entre prédicteurs et observations est quantifiée par une *fonction de perte* $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ mesurable, où $\ell(g, z)$ s'interprète comme "l'erreur" du prédicteur $g \in \mathcal{G}$ en l'observation $z \in \mathcal{Z}$. Soit P une loi de probabilité sur $\mathcal{Z}$, et Z une variable aléatoire de loi P. La qualité d'un prédicteur $g \in \mathcal{G}$ est mesurée par son *risque* $R(g)$, défini par[1] :

$$R(g) = R_P(g) = \mathbb{E}_{Z\sim P}[\ell(g, Z)]\,. \tag{1.1}$$

Un prédicteur $g \in \mathcal{G}$ est de bonne qualité lorsque son risque $R(g)$ est faible ; le prédicteur optimal $g^* = \arg\min_{g\in\mathcal{G}} R(g)$ est appelé *prédicteur de Bayes*. De manière cruciale, la loi P des observations est inconnue, de sorte que la fonction de risque $R : \mathcal{G} \to \mathbf{R}$ et son minimiseur g^* le sont également. L'objectif du problème consiste à construire, étant donné un *n-échantillon* $Z_1, \dots, Z_n$ indépendant et identiquement distribué (i.i.d.) de loi P, un prédicteur $\widehat{g}_n \in \mathcal{G}$ dépendant de $Z_1, \dots, Z_n$ de faible risque. Plus précisément, introduisons une classe $\mathcal{F} \subset \mathcal{G}$ de prédicteurs, appelée *classe de comparaison* ; on définit l'*excès de risque* relatif à la classe $\mathcal{F}$ d'un prédicteur $g \in \mathcal{G}$ par

$$\mathcal{E}(g) = \mathcal{E}_P(g; \mathcal{F}) := R(g) - \inf_{f\in\mathcal{F}} R(f)\,. \tag{1.2}$$

L'objectif est alors de produire un prédicteur $\widehat{g}_n$ (dans cette introduction, nous utiliserons de manière interchangeable les termes de *prédicteur*, d'*estimateur* et de *procédure* ou *algorithme d'apprentissage*) dont l'excès de risque $\mathcal{E}(\widehat{g}_n)$ est aussi faible que possible. L'excès de risque $\mathcal{E}(\widehat{g}_n)$ étant une variable aléatoire (en tant que fonction de l'échantillon $Z_1, \dots, Z_n$), il est possible de mesurer cette quantité par son espérance

$$\mathcal{E}_n(\widehat{g}_n; P, \mathcal{F}) = \mathbb{E}\big[\mathcal{E}_P(\widehat{g}_n; \mathcal{F})\big]\,, \tag{1.3}$$

ou par ses quantiles

$$\mathcal{E}_{n,\delta}(\widehat{g}_n; P, \mathcal{F}) = \inf\Big\{t \in \mathbf{R} : \mathbb{P}\big(\mathcal{E}_P(\widehat{g}_n; \mathcal{F}) \leqslant t\big) \geqslant 1 - \delta\Big\}\,, \tag{1.4}$$

où $1 - \delta \in (0, 1)$ est un *niveau de confiance*.

Afin d'illustrer les définitions générales précédentes, considérons les exemples suivants :

[1] Dans cette section, nous supposerons implicitement que les espérances considérées sont bien définies dans $\mathbf{R} \cup \{+\infty\}$; c'est notamment le cas lorsque la fonction de perte ℓ est à valeurs positives.

Exemple 1.1 (Estimation de l'espérance)**.** Soit $\mathcal{Z} = \mathbf{R}^d$ pour $d \geqslant 1$, $\mathcal{F} = \mathcal{G} = \mathbf{R}^d$, et $\ell(g, z) = \|g - z\|^2$ (où $\|\cdot\| = \|\cdot\|_2$ désigne la norme euclidienne sur $\mathbf{R}^d$). Supposons que P admette un moment d'ordre 2, c'est-à-dire que $\mathbb{E}[\|Z\|^2] < +\infty$. On a alors, pour tout $g \in \mathbf{R}^d$, $R(g) = \mathbb{E}[\|Z - g\|^2] = \|g - \mathbb{E}[Z]\|^2 + \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2]$, de sorte que $\arg\min_{f \in \mathcal{F}} R(f) = \mathbb{E}[Z]$ et $\mathcal{E}(g) = \|g - \mathbb{E}[Z]\|^2$. Ainsi, le problème d'apprentissage est ici équivalent à celui de l'estimation de la moyenne.

Exemple 1.2 (Estimation de densité)**.** Soit μ une mesure de référence sur l'espace mesurable $\mathcal{Z}$, et $\mathcal{G}$ l'ensemble des densités de probabilité sur $\mathcal{Z}$ par rapport à μ, c'est-à-dire des fonctions mesurables $g : \mathcal{Z} \to \mathbf{R}^+$ telles que $\int_{\mathcal{Z}} g \, \mathrm{d}\mu = 1$. Soit également $\mathcal{F} \subset \mathcal{G}$ une famille de densités sur $\mathcal{Z}$, appelée *modèle statistique*. La *perte logarithmique* (aussi appelée *perte de log-vraisemblance*) est définie par $\ell(g, z) := -\log g(z)$.

Lorsque la loi P admet une densité $p \in \mathcal{G}$, le risque $R(g)$ est minimisé par $g = p$, et coïncide (à une constante près) avec la *divergence de Kullback-Leibler* (ou *entropie relative*) entre p et g :

$$R(g) - R(p) = \mathbb{E}_{Z \sim P}\left[\log\left(\frac{p(Z)}{g(Z)}\right)\right] = \int_{\mathcal{Z}} p \log\left(\frac{p}{g}\right) \mathrm{d}\mu =: \mathrm{KL}(p, g) \geqslant 0\,. \tag{1.5}$$

Ainsi, le problème de l'apprentissage statistique avec perte logarithmique équivaut à celui de l'estimation de densité avec risque de Kullback-Leibler. Dans le cas particulier où $\mathcal{Z} = \mathbf{R}^d$, $\mathcal{F} = \mathcal{G} = \{\mathcal{N}(\theta, I_d) : \theta \in \mathbf{R}^d\}$ est le modèle (de translation) Gaussien et $\mu = (2\pi)^{-d/2}\mathrm{d}z$, ce problème équivaut à l'estimation de la moyenne (Exemple 1.1).

L'excès de risque $\mathcal{E}_n(\widehat{g}_n; P, \mathcal{F})$ d'une procédure $\widehat{g}_n$ par rapport à la classe $\mathcal{F}$ dépend de la loi P, qui est elle-même inconnue. Il est donc souhaitable d'obtenir des garanties valables sur un ensemble $\mathcal{P}$ de lois P sur $\mathcal{Z}$ aussi riche que possible, afin qu'il contienne la loi P du phénomène étudié. En particulier, il est possible de chercher des garanties *uniformes* sur la loi $P \in \mathcal{P}$. Cela conduit naturellement à considérer l'*excès de risque minimax*

$$\mathcal{E}_n^*(\ell, \mathcal{P}, \mathcal{F}) = \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \mathcal{E}(\widehat{g}_n; P, \mathcal{F}) = \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \left(\mathbb{E}[R(\widehat{g}_n)] - \inf_{f \in \mathcal{F}} R(f)\right), \tag{1.6}$$

où l'infimum porte sur tous les estimateurs $\widehat{g}_n$, comme mesure de la difficulté du problème d'apprentissage défini par $(\ell, \mathcal{P}, \mathcal{F})$ (Wald, 1949).

Concluons cette section par une remarque sur la définition du problème. Un estimateur $\widehat{g}_n$ est dit *propre* s'il prend ses valeurs dans la classe de comparaison $\mathcal{F}$, et *impropre* dans le cas contraire. Lorsque l'on se restreint aux estimateurs propres, c'est-à-dire lorsque $\mathcal{G} = \mathcal{F}$, on parle d'*apprentissage propre*. Nous verrons dans la Section 1.4 que la flexibilité additionnelle offerte par l'apprentissage impropre (non restreint) permet pour certains problèmes d'obtenir des garanties inaccessibles au moyen d'estimateurs propres (voir le Chapitre 7).

1.1.2 Apprentissage supervisé

Dans cette thèse, nous étudierons principalement des problèmes d'apprentissage dits *supervisés*. Il s'agit dans ce cas de chercher à prédire une variable de sortie Y, à partir d'une variable d'entrée X. Par exemple, dans un problème de reconnaissance d'objets, X peut encoder une image, et Y l'étiquette "l'image contient une voiture".

Formellement, en apprentissage supervisé, l'espace d'observation $\mathcal{Z}$ est un espace mesurable produit $\mathcal{X} \times \mathcal{Y}$. Les éléments de $\mathcal{X}$ sont appelées *entrées*, *caractéristiques*, *covariables* ou *variables prédictives*, tandis que les éléments de $\mathcal{Y}$ sont appelées *sorties*, *réponses* ou *étiquettes*. On se donne également un espace de prédictions $\widehat{\mathcal{Y}}$, auquel appartiennent les prédictions de la valeur y à partir de x, ainsi qu'une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$; $\ell(\widehat{y}, y)$ correspond à l'erreur de la prédiction $\widehat{y}$ étant donnée la valeur de y. Enfin, l'espace $\mathcal{G}$ des prédicteurs est dans ce cas l'ensemble des fonctions mesurables $g : \mathcal{X} \to \widehat{\mathcal{Y}}$, tandis que $\mathcal{F}$ est une sous-classe de prédicteurs. La fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ permet de définir naturellement une autre fonction de perte $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ (également notée ℓ par léger abus de notation) par $\ell(g, z) := \ell(g(x), y)$ pour $z = (x, y) \in \mathcal{Z}$ et $g \in \mathcal{G}$. Étant donnée une loi jointe P sur le couple $Z = (X, Y)$, le risque d'un prédicteur g s'écrit alors:

$$R(g) = R_P(g) = \mathbb{E}_{(X,Y)\sim P}\big[\ell(g(X), Y)\big]\,. \tag{1.7}$$

Dans ce cas, le prédicteur de Bayes $g^* : \mathcal{X} \to \mathcal{Y}$ s'écrit

$$g^*(x) = \operatorname*{arg\,min}_{\widehat{y}\in\widehat{\mathcal{Y}}} \mathbb{E}[\ell(\widehat{y}, Y)|X = x]\,, \tag{1.8}$$

ce qui justifie son appellation (au vu de l'espérance conditionnelle sur Y sachant X). Le risque $R(g^*)$ correspond à l'erreur "incompressible", due au caractère aléatoire de la réponse Y (y compris sachant X).

Considérons maintenant les exemples suivants, qui constituent des exemples classiques du problème de l'apprentissage supervisé. Ces problèmes sont définis par les espaces de sorties $\mathcal{Y}$ et de prédictions $\widehat{\mathcal{Y}}$, ainsi que la fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$. En pratique, le choix de la fonction de perte dépend de la nature des données et de l'objectif recherché.

Exemple 1.3 (Classification)**.** Lorsque l'on cherche à prédire une réponse (dite *discrète* ou *catégorielle*) appartenant à un ensemble fini $\mathcal{Y}$, il est naturel de considérer l'espace $\widehat{\mathcal{Y}} = \mathcal{Y}$ et la fonction de perte $\ell(\widehat{y}, y) = \mathbf{1}(\widehat{y} \neq y)$ pour $\widehat{y} \in \widehat{\mathcal{Y}}$ et $y \in \mathcal{Y}$, appelée *erreur de classification* (ou *perte* 0-1). Un prédicteur $g : \mathcal{X} \to \mathcal{Y}$ est alors appelé *classifieur*, et son risque $R(g) = \mathbb{P}(g(X) \neq Y)$ est simplement sa probabilité d'erreur. Le classifieur de Bayes (1.8) est alors donné par $g^*(x) = \arg\max_{\widehat{y}\in\mathcal{Y}} \mathbb{P}(Y = \widehat{y}|X = x)$, et le problème de la classification revient à déterminer la sortie la plus probable étant donnée l'entrée.

Exemple 1.4 (Régression)**.** Pour des données *quantitatives*, c'est-à-dire lorsque $\mathcal{Y} = \mathbf{R}$, la fonction de perte la plus courante est la *perte quadratique* $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ pour $\widehat{y}, y \in \mathbf{R}$. Le problème est alors parfois appelé *régression* ou problème des *moindres carrés*. Dans ce cas, un prédicteur $g : \mathcal{X} \to \mathbf{R}$ est une *fonction de régression*.

Le prédicteur de Bayes (1.8) est alors donné par l'*espérance conditionnelle* $g^*(x) := \mathbb{E}[Y|X = x]$ (dès lors que $\mathbb{E}[Y^2] < +\infty$), et est parfois appelé la *fonction de régression* de Y sachant X. De plus, pour tout $g : \mathcal{X} \to \mathbf{R}$, $R(g) - R(g^*) = \mathbb{E}[(g(X) - g^*(X))^2] = \|g - g^*\|^2_{L^2(P_X)}$, où P_X désigne la loi de X. Ainsi, le problème de la régression équivaut à celui de l'estimation de l'espérance conditionnelle de Y sachant X, sous la norme de $L^2(P_X)$ (qui est elle-même inconnue en pratique).

Exemple 1.5 (Estimation de densité conditionnelle)**.** Les deux problèmes précédents sont des exemples de prédiction *ponctuelle* où $\widehat{\mathcal{Y}} = \mathcal{Y}$, et où la prédiction $\widehat{y}$ est une valeur possible de la sortie y, qui doit s'en approcher selon une certaine métrique. Dans certaines situations, il

est souhaitable d'avoir de l'information sur l'incertitude attachée à la réalisation de la sortie. Dans ce cas, il est naturel de former une prédiction *probabiliste* de la réponse y, qui assigne des probabilités aux différentes valeurs possibles de cette variable.

Une fonction de perte possible est alors la perte *logarithmique*, décrite dans l'Exemple 1.2 dans le cas non conditionnel. Étant donné un espace mesurable $\mathcal{Y}$ muni d'une mesure de référence μ, $\widehat{\mathcal{Y}}$ est l'ensemble des densités de probabilité sur $\mathcal{Y}$ par rapport à μ, et la perte s'écrit $\ell(\widehat{y}, y) = -\log \widehat{y}(y)$. Un prédicteur $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ correspond alors à une *densité conditionnelle* de y sachant x, notée $g(y|x) := g(x)(y)$.

Le prédicteur de Bayes (1.8) est donné par la *densité conditionnelle* $g^*(\cdot|x) := \mathrm{d}P_{Y|X=x}/\mathrm{d}\mu$ de Y sachant X (lorsqu'elle existe), tandis que $R(g) - R(g^*)$ vaut $\mathbb{E}[\mathrm{KL}(g^*(\cdot|X), g(\cdot|X))]$ (qui coïncide avec la divergence de Kullback-Leibler entre les lois jointes sur $\mathcal{X} \times \mathcal{Y}$ induites par P_X et g^*, g respectivement). Ainsi, le problème de l'apprentissage supervisé équivaut ici à celui de l'estimation de densité conditionnelle.

Pour plus de détails sur le problème de la classification, outre les références générales indiquées précédemment, nous renvoyons à l'ouvrage de référence Devroye et al. (1996) ; pour une présentation plus succincte du sujet, nous renvoyons à Bousquet et al. (2004). Pour ce qui est de la régression (principalement non paramétrique), nous renvoyons à Györfi et al. (2002); Wasserman (2006); Tsybakov (2009) pour la régression.

1.1.3 Approches générative et discriminative

Comme nous l'avons vu dans la section précédente, le problème de l'apprentissage supervisé se formule de la façon suivante : étant donné un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. de loi P sur $\mathcal{X} \times \mathcal{Y}$, produire un prédicteur $\widehat{g}_n : \mathcal{X} \to \widehat{\mathcal{Y}}$ dont le risque $R(\widehat{g}_n)$ est faible. De plus, le risque minimal atteignable par un prédicteur $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ est le *risque de Bayes* $R(g^*)$, où g^* est le prédicteur de Bayes (1.8). Il est donc naturel de considérer comme objectif de contrôler l'excès de risque par rapport à g^*, c'est-à-dire la différence $R(\widehat{g}_n) - R(g^*)$ entre le prédicteur utilisé et le meilleur possible.

Résultats d'impossibilité. Cependant, comme le montre le résultat suivant (établi par Devroye, 1982), il n'est pas possible de s'approcher de la performance du prédicteur de Bayes avec un échantillon fini sans aucune hypothèse sur la loi P.

Théorème 1.1 (Devroye et al., 1996, Théorème 7.1)**.** *Considérons le problème de la classification binaire (Exemple 1.3), avec $\mathcal{X} = [0,1]$ et $\mathcal{Y} = \{0,1\}$. Pour tous $n \geqslant 1$, $\varepsilon > 0$ et tout classifieur $\widehat{g}_n$, il existe une loi jointe P sur $\mathcal{X} \times \mathcal{Y}$ telle que $R(g^*) = 0$ et $\mathbb{E}[R(\widehat{g}_n)] \geqslant 1/2 - \varepsilon$.*

Pour se convaincre du Théorème 1.1, on peut considérer l'exemple suivant. Supposons X uniforme sur $[0,1]$, et considérons $Y = g^*(X)$, où la fonction $g^* : [0,1] \to \{0,1\}$ prend une valeur constante $a_k \in \{0,1\}$ sur chacun des intervalles $[(k-1)/N, k/N]$, $k = 1, \ldots, N$ (avec $N \geqslant 1$). De plus, les $(a_k)_{1 \leqslant k \leqslant N}$ sont arbitraires (par exemples tirés selon une loi a priori uniforme sur $\{0,1\}^N$). Dans ce cas, un échantillon de taille n ne renseigne que sur (au plus) n des N valeurs de la suite (a_k), donc sur une fraction d'au plus n/N des valeurs de X ; pour les valeurs non observées, il n'est pas possible de faire mieux qu'une prédiction aléatoire. Par conséquent, si $N \gg n$, un échantillon de taille n n'apporte que très peu d'information sur la loi P de (X, Y) : on peut montrer qu'un classifieur $\widehat{g}_n$ ne peut atteindre un risque inférieur à

$(1 - n/N)/2$ dans le pire des cas (en considérant la moyenne de ce risque lorsque la suite (a_k) est elle-même tirée uniformément sur $\{0,1\}^N$).

Le Théorème 1.1 est un résultat de type *no free lunch*, qui affirme qu'il n'est pas possible d'obtenir de garantie non triviale sans hypothèse sur P. Notons cependant que dans le résultat précédent, la loi P mettant en défaut le classifieur $\widehat{g}_n$ dépend de la taille n de l'échantillon. Il existe une contrepartie positive au résultat négatif du Théorème 1.1, obtenue en fixant la loi P et en se plaçant dans un cadre asymptotique où la taille de l'échantillon n tend vers l'infini. Dans ce cas, la notion de garantie de risque à taille d'échantillon fixe est remplacée par la notion asymptotique de *consistance*.

Définition 1.1 (Consistance)**.** Une suite $(\widehat{g}_n)_{n\geqslant 1}$ de prédicteurs (où $\widehat{g}_n$ est fonction d'un échantillon de taille n) est dite *universellement consistante* (relativement à un ensemble $\mathcal{P}$ de lois sur $\mathcal{X} \times \mathcal{Y}$) si, pour toute loi $P \in \mathcal{P}$, on a $\mathbb{E}[R(\widehat{g}_n)] \to R(g^*)$ lorsque $n \to \infty$.

Signalons qu'il existe d'autres variantes de la consistance, obtenues en remplaçant la convergence en espérance dans la Définition 1.1 par la convergence en probabilité (qui est plus faible en général, mais équivalente pour des pertes bornées comme en classification), ou presque sûre (on parle alors de consistance *forte*, qui est plus forte pour des pertes bornées).

Le théorème suivant affirme qu'il existe une suite consistante de classifieurs sur $\mathbf{R}^d$.

Théorème 1.2 (Stone, 1977)**.** *Considérons le problème de la classification binaire avec $\mathcal{X} = \mathbf{R}^d$ et $\mathcal{Y} = \{0,1\}$. Pour $n \geqslant 1$ et $1 \leqslant k \leqslant n$, le classifieur des k-plus proches voisins $\widehat{g}_{n,k}$ est défini comme suit : pour $x \in \mathbf{R}^d$,*

$$\widehat{g}_{n,k}(x) = \mathbf{1}\left(\sum_{j=1}^{k} Y_{(i)} > k/2\right)$$

où $\big((X_{(i)}, Y_{(i)})\big)_{1\leqslant i\leqslant n}$ correspond à l'échantillon $\big((X_i, Y_i)\big)_{1\leqslant i\leqslant n}$ ordonné de sorte que $\|x - X_{(1)}\| \leqslant \ldots \leqslant \|x - X_{(n)}\|$. Alors, si $k_n \to \infty$ et $k_n/n \to 0$, la suite de classifieurs $\widehat{g}_{n,k_n}$ est universellement consistante sur l'ensemble $\mathcal{P}$ des mesures de probabilités sur $\mathbf{R}^d \times \{0,1\}$.

Notons qu'au vu du Théorème 1.1, il n'est pas possible d'obtenir de convergence *uniforme* sur la classe $\mathcal{P}$. Le Théorème 1.2 affirme néanmoins qu'il existe une suite $\widehat{g}_n$ de classifieurs consistante pour toute loi fixe P. Cependant, le résultat négatif suivant (Cover, 1968; Devroye, 1982) montre que la convergence peut être arbitrairement lente, même lorsque la loi P est fixée et lorsque $n \to \infty$.

Proposition 1.1 (Devroye et al., 1996, Théorème 7.2)**.** *Considérons le problème de la classification binaire avec $\mathcal{X} = \mathbf{R}$. Soit $(\varepsilon_n)_{n\geqslant 1}$ une suite décroissante de réels telle que $\varepsilon_1 \leqslant 1/16$ et $\varepsilon_n \to 0$ lorsque $n \to \infty$. Pour toute suite $\widehat{g}_n$ de classifieurs, il existe une loi P telle que $R(g^*) = 0$ et $\mathbb{E}[R(\widehat{g}_n)] \geqslant \varepsilon_n$ pour tout n.*

Approche générative. Il ressort de la discussion précédente qu'il n'est pas possible d'obtenir de garanties sur l'excès de risque $R(\widehat{g}_n) - R(g^*)$ d'un prédicteur $\widehat{g}_n$ sans hypothèse. Il est donc nécessaire d'introduire un *biais inductif* dans la procédure, c'est-à-dire de favoriser certaines distributions, ou certaines formes de dépendance entre la variable d'entrée X et la sortie Y.

Une première façon de le faire consiste à restreindre l'ensemble des lois P considérées, c'est-à-dire à faire une hypothèse de *modélisation* sur cette loi. Dans sa variante la plus forte,

ce type d'approche conduit à faire une hypothèse *paramétrique* sur P, c'est-à-dire à supposer que P appartient à un *modèle* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Dans le cas paramétrique, l'espace des paramètres Θ est un sous-ensemble (par exemple ouvert) de $\mathbf{R}^d$, et la mesure P_θ dépend de manière régulière de θ (au sens où $P_\theta = p_\theta \cdot \mu$ pour une mesure de domination μ fixe, et où la densité p_θ dépend de façon lisse de θ). Cette approche dite *générative* (ou de *modélisation*) a traditionnellement été la plus courante en statistiques (Breiman, 2001b).

L'avantage de l'approche générative est qu'en restreignant le problème, elle en rend possible une analyse fine. Le problème de l'apprentissage tombe alors dans le cadre de la théorie de la décision statistique (Wald, 1949; Lehmann and Casella, 1998; Berger, 1985), qui permet de considérer des notions précises d'optimalité. Il est alors en particulier possible de parler de procédures optimales dans le pire des cas, ou en moyenne selon une certaine loi *a priori* sur les valeurs possibles de P_θ. En outre, la théorie de l'estimation est bien comprise dans le cas paramétrique asymptotique, lorsque la taille de l'échantillon n tend vers l'infini (Ibragimov and Has'minskii, 1981; Le Cam, 1986; van der Vaart, 1998). L'approche générative est également pertinente pour traiter des problèmes de statistique inférentielle, allant au-delà de la stricte prédiction : on s'intéresse alors au paramètre θ en lui-même (ou à une propriété spécifique de la loi P_θ), plutôt que comme intermédiaire permettant d'effectuer des prédictions.

La principale limitation de l'approche générative est qu'elle repose sur l'hypothèse très forte que la loi P appartient à une famille spécifiée. Cette hypothèse n'a en pratique aucune raison d'être satisfaite ; en effet, le modèle ne constitue qu'une approximation de la loi P, cette dernière échappant au contrôle du statisticien. Ainsi, les résultats obtenus sous l'hypothèse de l'appartenance à un modèle ne donnent aucune garantie lorsque cette hypothèse n'est pas satisfaite.

Approche discriminative. Pour des problèmes de nature prédictive, il est en fait possible d'introduire un biais inductif tout en évitant des hypothèses fortement restrictives comme l'appartenance à un modèle paramétrique connu.

Pour s'en convaincre, commençons par considérer un modèle possible $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Définissons, pour tout $\theta \in \Theta$, f_θ le prédicteur de Bayes associé à la loi P_θ :

$$f_\theta(x) := \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathbb{E}_{(X,Y)\sim P_\theta}[\ell(\widehat{y}, Y)|X = x]\,,$$

ainsi que la classe $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ formée par ces prédicteurs. Si $P \in \mathcal{P}$, alors le prédicteur de Bayes appartient à la classe $\mathcal{F}$, et donc pour tout prédicteur g, l'excès de risque par rapport à g^* s'écrit

$$R(g) - R(g^*) = R(g) - \inf_{f \in \mathcal{F}} R(f)\,, \tag{1.9}$$

qui est tout simplement l'excès de risque par rapport à la classe $\mathcal{F}$. Notons à présent que ce dernier peut être défini même lorsque $P \notin \mathcal{F}$, sans référence au prédicteur de Bayes g^*. Il est alors possible de chercher à contrôler l'excès de risque par rapport à $\mathcal{F}$ plutôt que g^*, sans imposer d'hypothèse forte sur P ; intuitivement, cela est envisageable car la complexité de $\mathcal{F}$ est contrôlée, tandis que g^* peut être arbitrairement complexe en fonction de P.

Cette observation motive l'approche dite *discriminative*, où le biais inductif est introduit en restreignant la classe de comparaison $\mathcal{F}$ plutôt que l'ensemble $\mathcal{P}$ de lois considérées. L'objectif est alors de contrôler l'excès de risque par rapport à $\mathcal{F}$. Dans le cas général où $P \notin \{P_\theta : \theta \in$

$\Theta\}$, la relation (1.9) devient la *décomposition en erreurs d'approximation et d'estimation* :

$$R(g) - R(g^*) = \Big(R(g) - \inf_{f\in\mathcal{F}} R(f)\Big) + \Big(\inf_{f\in\mathcal{F}} R(f) - R(g^*)\Big)\,. \tag{1.10}$$

Le premier terme de la décomposition (1.10) est simplement l'excès de risque, aussi appelé *erreur d'estimation* ou *variance* ; lorsque g est un prédicteur $\widehat{g}_n$ construit à partir du jeu de données, ce terme est aléatoire (car dépendant des données). Le second terme de cette décomposition, appelé *erreur d'approximation* ou *biais*, est à l'inverse déterministe (indépendant des données) ; il mesure à quel point la classe $\mathcal{F}$ approche le prédicteur de Bayes g^* en termes de risque. L'approche discriminative sépare ainsi l'étude de l'erreur d'estimation de celle de l'erreur d'approximation.

Exemple 1.6 (Régression linéaire)**.** Considérons le problème de l'apprentissage supervisé avec $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \mathbf{R}$ et perte quadratique (Exemple 1.4). L'approche générative postule un modèle sur (X, Y) ; par exemple, que le vecteur aléatoire (X, Y) est Gaussien :

$$\mathcal{P} = \{\mathcal{N}(\mu_{XY}, \Sigma_{XY}) : \mu_{XY} \in \mathbf{R}^{d+1}, \Sigma_{XY} \in \mathbf{R}^{(d+1)\times(d+1)} \text{ positive}\}\,.$$

Dans ce cas, on a $Y = \langle \beta^*, X\rangle + \varepsilon$ pour un certain $\beta^* \in \mathbf{R}^d$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ $(\sigma^2 > 0)$ est indépendant de X. La fonction de régression s'écrit alors $g^*(x) = \langle \beta^*, x\rangle$, la classe des prédicteurs optimaux associée au modèle $\mathcal{P}$ est donc constituée des fonctions linéaires $\mathcal{F} = \{x \mapsto \langle \beta^*, x\rangle : \beta^* \in \mathbf{R}^d\}$.

À l'inverse, l'approche discriminative conduit à chercher à prédire Y à partir de X avec la même précision que la meilleure fonction linéaire de X, sans nécessairement supposer de forme particulière à la fonction de régression $x \mapsto \mathbb{E}[Y|X = x]$ ou à la loi de $\varepsilon = Y - \langle \beta^*, X\rangle$.

Pour conclure, il n'est en général pas possible d'obtenir de garanties d'excès de risque uniformes du type

$$\sup_{P\in\mathcal{P}} \Big(\mathbb{E}[R(\widehat{g}_n)] - \inf_{f\in\mathcal{F}} R(f)\Big) \tag{1.11}$$

pour un prédicteur $\widehat{g}_n$ calculé à partir d'un échantillon fini, lorsque ni l'ensemble $\mathcal{P}$ de lois considérées ni la classe $\mathcal{F}$ de comparaison ne sont restreintes (par exemple, dans le cas de la classification, lorsque $\mathcal{P}$ est l'ensemble des lois jointes sur $\mathbf{R}^d \times \{0, 1\}$, et $\mathcal{F}$ l'ensemble des fonctions mesurables $\mathbf{R}^d \to \{0, 1\}$). En d'autres termes, afin d'obtenir un excès de risque minimax (1.6) faible, il est nécessaire de restreindre soit l'ensemble $\mathcal{P}$ des lois considérées, soit la classe de comparaison $\mathcal{F}$.

L'approche générative consiste à restreindre la famille $\mathcal{P}$, tandis que l'approche discriminative consiste à restreindre la classe $\mathcal{F}$. La première permet parfois d'obtenir des résultats plus précis que la seconde, mais au prix d'hypothèses plus restrictives. Pour cette raison, l'approche discriminative est généralement privilégiée en apprentissage statistique. Il est cependant courant de combiner les deux approches, en imposant certaines restrictions à la loi P pour analyser plus finement le comportement du risque. En outre, l'approche purement générative est utile pour obtenir des *bornes inférieures* sur la difficulté du problème.

1.1.4 Apprentissage et processus empiriques

Dans cette section, nous passons brièvement en revue un paradigme classique permettant de traiter le problème de l'apprentissage (c'est-à-dire de contrôler l'excès de risque), celui de la

convergence uniforme des processus empiriques. Pour davantage de détails, nous renvoyons à Vapnik and Chervonenkis (1974); van der Vaart and Wellner (1996); van de Geer (1999); Bartlett and Mendelson (2002); Bartlett et al. (2005); Koltchinskii (2006); Boucheron et al. (2005); Koltchinskii (2011); Talagrand (2014); Massart (2007); Boucheron et al. (2013) entre autres références. Nous reprenons le formalisme général de l'apprentissage statistique introduit en Section 1.1.1, en nous restreignant au cas de l'apprentissage propre, pour lequel $\mathcal{F} = \mathcal{G}$.

Minimisation du risque empirique. L'approche sans doute la plus naturelle du problème de l'apprentissage statistique est la *minimisation du risque empirique* (Vapnik, 1998). Définissons, pour tout $f \in \mathcal{F}$, le *risque empirique*

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) \,. \tag{1.12}$$

Notons que $\widehat{R}_n(f)$ est aléatoire (en tant que fonction de $Z_1, \dots, Z_n$), et donc que le risque empirique peut être vu comme un *processus stochastique* $\big(\widehat{R}_n(f)\big)_{f \in \mathcal{F}}$ indexé par les prédicteurs $f \in \mathcal{F}$. Le *minimiseur du risque empirique* (en anglais *Empirical Risk Minimizer*, abrégé ERM) est par définition

$$\widehat{f}_n^{\mathrm{ERM}} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \widehat{R}_n(f) \,, \tag{1.13}$$

où l'on suppose l'existence d'un tel minimiseur (et où l'on fixe un choix de celui-ci lorsqu'il n'est pas unique). D'après la loi des grands nombres, pour tout $f \in \mathcal{F}$ le risque empirique $\widehat{R}_n(f)$ converge presque sûrement vers le risque $R(f)$ lorsque $n \to \infty$, dès lors que $\mathbb{E}[|\ell(f, Z)|] < +\infty$. Il est donc naturel d'espérer de $\widehat{f}_n^{\mathrm{ERM}}$, qui minimise l'approximation $\widehat{R}_n$ de R, qu'il minimise aussi approximativement R.

Borne en terme de l'erreur de généralisation. Pour mener à bien cet argument, la seule convergence *ponctuelle* (c'est-à-dire pour tout $f \in \mathcal{F}$) de $\widehat{R}_n$ vers R s'avère insuffisante. Pour s'en convaincre, on peut considérer l'exemple mentionné à la suite du Théorème 1.1, en prenant pour $\mathcal{F}$ l'ensemble des fonctions mesurables $[0,1] \to \{0,1\}$. Dans ce cas, par l'argument indiqué dans cet exemple, il n'est pas possible d'obtenir de garantie non triviale pour $\widehat{f}_n^{\mathrm{ERM}}$ (ou tout autre prédicteur) sans restriction sur f^*. Il est en revanche possible de contrôler l'excès de risque lorsque la convergence $\widehat{R}_n(f) \to R(f)$ a lieu *uniformément* sur $f \in \mathcal{F}$, comme le montre le théorème suivant.

Théorème 1.3 (Vapnik and Chervonenkis, 1974)**.** *L'excès de risque du minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$ satisfait :*

$$R(\widehat{f}_n^{\mathrm{ERM}}) - \inf_{f \in \mathcal{F}} R(f) \leqslant 2 \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \,. \tag{1.14}$$

Proof. Pour tout $f \in \mathcal{F}$, on a

$$R(\widehat{f}_n^{\mathrm{ERM}}) - R(f) = \big(R(\widehat{f}_n^{\mathrm{ERM}}) - \widehat{R}_n(\widehat{f}_n^{\mathrm{ERM}})\big) + \big(\widehat{R}_n(\widehat{f}_n^{\mathrm{ERM}}) - \widehat{R}_n(f)\big) + \big(\widehat{R}_n(f) - R(f)\big).$$

Le second terme du membre de droite est négatif par définition de $\widehat{f}_n^{\mathrm{ERM}}$, tandis que le premier et le troisième termes sont bornés par $\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$. La borne (1.14) en découle en considérant le supremum du membre de gauche sur $f \in \mathcal{F}$. □

Le Théorème 1.3 majore l'excès de risque en terme de la quantité $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$, appelée *erreur de généralisation*. Cette variable aléatoire est le supremum du processus centré $\big(\widehat{R}_n(f) - R(f)\big)_{f\in\mathcal{F}}$, appelé *processus empirique*. Il existe une riche théorie permettant de contrôler le supremum de processus empiriques (Dudley, 1984; Talagrand, 2014), qui joue un rôle important en statistiques (van der Vaart and Wellner, 1996; van de Geer, 1999; Koltchinskii, 2011). De manière générale, le supremum du processus empirique $(\widehat{R}_n(f) - R(f))_{f\in\mathcal{F}}$ peut être borné (en espérance, ou avec forte probabilité) en fonction de celui d'un autre processus appelé *processus de Rademacher*, au moyen d'une technique générale de *symétrisation* (Giné and Zinn, 1984; Bartlett and Mendelson, 2002; Massart, 2007; Koltchinskii, 2011; Boucheron et al., 2013).

Théorème 1.4 (Giné and Zinn, 1984)**.** *Soit $\varepsilon_1, \dots, \varepsilon_n$ des variables aléatoire indépendantes entre elles et de $Z_1, \dots, Z_n$, telles que $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$. Alors, on a*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\Big] \leqslant 2\,\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\Big|\sum_{i=1}^{n} \varepsilon_i\, \ell(f, Z_i)\Big|\Big]\,. \tag{1.15}$$

La borne (1.15) est essentiellement optimale, car une borne dans le sens inverse est valable (au recentrage de la classe près), voir par exemple (Koltchinskii, 2011, Théorème 2.1). Le terme de droite de l'inégalité (1.15) est appelé *complexité de Rademacher* de la classe

$$\ell \circ \mathcal{F} := \ell(\mathcal{F}, \cdot) = \{\ell(f, \cdot) : f \in \mathcal{F}\}\,.$$

Cette quantité mesure la "richesse" de la classe $\mathcal{F}$, en un sens dépendant de la loi P (à travers les $Z_i \sim P$ dans (1.15)). Lorsque les $\ell(f, \cdot)$ sont bornées par une constante R et lorsque la classe $\mathcal{F}$ est finie, alors il existe une constante absolue $C > 0$ telle que $\mathrm{Rad}_n(\mathcal{F}, P) \leqslant CR\sqrt{\log|\mathcal{F}|/n}$ (Boucheron et al., 2013). De plus, pour des classes $\mathcal{F}$ de "dimension" d (en un sens précis, par exemple la dimension de Vapnik-Chervonenkis, Vapnik 1998 dans le cas de classes de fonctions à valeurs dans $\{0, 1\}$), la complexité de Rademacher satisfait $\mathrm{Rad}_n(\ell \circ \mathcal{F}, P) \leqslant C\sqrt{d/n}$. Enfin, il est aussi possible de contrôler la complexité de Rademacher de classes définies par des conditions de normes (indépendamment de la dimension), pour des fonctions de pertes Lipschitz, en utilisant une inégalité de contraction pour les complexités de Rademacher (Ledoux and Talagrand, 2013).

De manière générale, la complexité de Rademacher $\mathrm{Rad}_n(\mathcal{F}, P)$ (1.15) peut être contrôlée à partir de techniques générales pour l'étude du supremum de processus (sous-) Gaussiens. Ces bornes dépendent de la structure métrique de la classe $\mathcal{F}$ sous la distance $L^2(P)$. La principale technique pour majorer de telles quantités est celle du *chaînage*, dont l'usage remonte à Kolmogorov, exploitée par Dudley (1967) et raffinée par Fernique (1975) et Talagrand, qui consiste à décomposer chaque fonction comme une "chaîne" d'approximations à différents niveaux, puis à contrôler la différence entre niveaux successifs par une majoration du supremum de processus finis sous-Gaussiens par une borne d'union (Talagrand, 2014; Dudley, 1999; Ledoux and Talagrand, 2013; Massart, 2007; Vershynin, 2018). L'usage de cette technique est crucial pour obtenir des bornes optimales sur les complexités de Rademacher de classes "riches", non paramétriques.

Des Théorèmes 1.3 et 1.4, il ressort que l'excès de risque du minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$ peut être contrôlé en fonction de l'erreur de généralisation (1.14), elle-même majorée par la complexité de Rademacher $\mathrm{Rad}_n(\mathcal{F}, P)$. Dans le cas de classes $\mathcal{F}$ de faible complexité (comme les classes finies ou de "dimension" finie), cela implique une borne d'excès de risque

d'ordre $O(1/\sqrt{n})$. Cette vitesse de convergence dite *lente* est optimale dans le pire des cas (c'est-à-dire sans hypothèse particulière sur P) pour certaines fonctions de perte, comme par exemple l'erreur de classification (Exemple 1.3) par un résultat de Devroye and Lugosi (1995).

Localisation. La vitesse lente en $O(1/\sqrt{n})$ s'avère en revanche sous-optimale pour d'autres fonctions de pertes, comme la perte quadratique utilisée en régression (Exemple 1.4) et la perte logarithmique utilisée en estimation de densité (Exemple 1.2). De plus, même dans le cas de la classification, des vitesses plus rapides sont possibles sous certaines hypothèses sur la loi P. Dans ce cas, l'approche consistant à borner l'excès de risque en fonction de l'erreur de généralisation (Théorème 1.3) est fondamentalement sous-optimale : en effet, même dans le cas d'une classe $\mathcal{F} = \{f_0\}$ à un élément, l'erreur de généralisation est d'ordre $O(1/\sqrt{n})$, car l'écart-type de $\widehat{R}_n(f_0)$ est $\sqrt{\mathrm{Var}(\ell(f_0, Z))/n}$.

Commençons par illustrer la possibilité de vitesses rapides par un exemple simple.

Exemple 1.7. On considère le problème de l'estimation de la moyenne (Exemple 1.1) dans le cas où $d = 1$, et l'on pose $\sigma^2 := \mathrm{Var}(Z)$ et $f^* = \mathbb{E}[Z]$ où $Z \sim P$. Le minimiseur du risque empirique est dans ce cas $\widehat{f}_n^{\mathrm{ERM}} = n^{-1} \sum_{i=1}^n Z_i$, et son excès de risque vaut (Exemple 1.1):

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}_n^{\mathrm{ERM}})\big] = \mathbb{E}\big[(\widehat{f}_n^{\mathrm{ERM}} - f^*)^2\big] = \frac{\sigma^2}{n}\,.$$

Au-delà de cet exemple, il est possible d'obtenir des vitesses rapides sous des conditions plus générales sur la fonction de perte et la distribution P. Dans le cas de l'Exemple 1.7, une propriété importante est que *les éléments $f \in \mathcal{F}$ dont l'excès de risque est faible ont une perte fortement corrélée à celle de f^**. En effet, pour tout $f \in \mathcal{F}$, $\ell(f, Z) - \ell(f^*, Z) = (f - f^*)(f + f^* - 2Z)$, de sorte que $\mathrm{Var}(\ell(f, Z) - \ell(f^*, Z)) = 4\sigma^2(f - f^*)^2 = 4\sigma^2\mathcal{E}(f)$.

Ce type de propriété est appelé *hypothèse de marge*. Une hypothèse de cette nature a été introduite par Mammen and Tsybakov (1999); Tsybakov (2004) dans le cas de la classification. La condition de marge et ses conséquences ont également été étudiées par Massart and Nédélec (2006). Nous considérons ici l'hypothèse suivante, introduite par Bartlett and Mendelson (2006) sous le nom de *condition de Bernstein*, et valable pour une fonction de perte générale.

Définition 1.2. Soient $\beta \in (0, 1]$ et $B \geqslant 1$. On dit que la classe $\mathcal{F}$ satisfait la (β, B)-*condition de Bernstein* sous la fonction de perte $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbf{R}$ et la loi P si, en notant $f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}[\ell(f, Z)]$, on a pour tout $f \in \mathcal{F}$,

$$\mathbb{E}\big[\big(\ell(f, Z) - \ell(f^*, Z)\big)^2\big] \leqslant B \cdot \mathbb{E}\big[\ell(f, Z) - \ell(f^*, Z)\big]^\beta\,. \tag{1.16}$$

Remarque 1.1. L'hypothèse de Bernstein est de nature *générative*, en ce sens qu'il s'agit d'une hypothèse sur la loi P de Z (Section 1.1.3).

Puisque $\mathbb{E}\big[(\ell(f, Z) - \ell(f^*, Z))^2\big] \geqslant \mathrm{Var}\big(\ell(f, Z) - \ell(f^*, Z)\big)$, la condition de Bernstein (1.16) implique que $\mathrm{Var}\big(\ell(f, Z) - \ell(f^*, Z)\big) \leqslant B\mathcal{E}(f)^\beta$. Cela signifie précisément que la perte $\ell(f, Z)$ d'éléments $f \in \mathcal{F}$ de faible excès de risque est corrélée à celle de f^*.

Sous une hypothèse de marge de type (1.16) — ou, plus généralement, étant donnée une borne sur $\mathrm{Var}(\ell(f, Z) - \ell(f^*, Z))$ en termes de $\mathcal{E}(f)$ — il est possible d'obtenir des bornes améliorées d'excès de risque. En effet, $\widehat{f}_n^{\mathrm{ERM}}$ peut s'écrire :

$$\widehat{f}_n^{\mathrm{ERM}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \big(\ell(f, Z_i) - \ell(f^*, Z_i)\big)\,; \tag{1.17}$$

la borne d'excès de risque en terme de l'erreur de généralisation implique alors que, avec forte probabilité, $\mathcal{E}(\widehat{f}_n^{\mathrm{ERM}}) \leqslant \delta_1$, où δ_1 dépend de la complexité de Rademacher de la classe $\big(\ell(f,\cdot) - \ell(f^*,\cdot)\big)_{f\in\mathcal{F}}$, et donc de la quantité $\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} \mathbb{E}\big[(\ell(f,Z)-\ell(f^*,Z))^2\big]$ (Talagrand, 1996). Dans ce cas, on a $\widehat{f}_n^{\mathrm{ERM}} \in \mathcal{F}(\delta_1) := \{f \in \mathcal{F} : \mathcal{E}(f) \leqslant \delta_1\}$. Il est alors possible de répéter l'argument précédent, mais en obtenant une nouvelle borne car l'hypothèse (1.16) assure que $\sigma^2(\mathcal{F}(\delta_1)) \leqslant B\delta_1^{\beta}$. En répétant ce processus et en tenant compte des termes supplémentaires qui apparaissent, il est possible d'obtenir une borne améliorée tenant compte de la complexité locale de la classe (c'est-à-dire de celle de sous-classes de la forme $\mathcal{F}(\delta)$). Cette approche fondée sur l'étude de *complexités de Rademacher locales* est due à Koltchinskii (2001, 2006); Bartlett et al. (2005).

Exemple 1.8 (Classe finie)**.** Dans le cas d'une classe $\mathcal{F}$ finie, sous l'hypothèse de Bernstein (1.16) (ainsi qu'une hypothèse de pertes bornées), l'excès de risque $\mathbb{E}[\mathcal{E}(\widehat{f}_n^{\mathrm{ERM}})]$ est majoré par

$$C\left[\Big(\frac{B\log|\mathcal{F}|}{n}\Big)^{1/(2-\beta)} + \frac{\log|\mathcal{F}|}{n}\right]$$

(Boucheron et al., 2005), qui constitue une vitesse améliorée en $O(1/n^{2-\beta})$, allant de $O(1/\sqrt{n})$ (on parle alors de *vitesse lente*) à $O(1/n)$ (vitesse rapide).

Ainsi, l'approche fondée sur les complexités de Rademacher localisées combine l'idée de la convergence uniforme de processus empiriques et celle de localisation. Cette approche est très générale, et fournit des bornes précises sur l'excès de risque du minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$, qui sont optimales dans de nombreuses situations (Koltchinskii, 2006). En particulier, la notion de complexité de Rademacher (et ses variantes localisées) fournit un contrôle précis de la complexité de la classe, en un sens dépendant de la loi P. Mentionnons également que, dans le cas de la régression avec perte quadratique (Exemple 1.4), il existe une variante "ajustée" de la complexité de Rademacher, qui permet d'obtenir des vitesses rapides sans avoir à recourir explicitement aux complexités de Rademacher localisées (Liang et al., 2015).

Le principal inconvénient de l'approche en termes de processus empiriques est sa complexité technique : elle requiert en effet une machinerie sophistiquée permettant le contrôle de déviations de processus empiriques. Cela conduit notamment à des bornes faisant apparaître des constantes souvent sous-optimales, qui ne permettent pas de comparer ou de calibrer précisément des procédures en pratique.

1.1.5 Apprentissage et optimisation stochastique

Nous évoquons maintenant un autre point de vue sur le problème d'apprentissage, celui de l'*optimisation stochastique* (Robbins and Monro, 1951; Polyak and Juditsky, 1992; Nemirovski and Yudin, 1983; Benveniste et al., 1990; Kushner and Yin, 2003; Nesterov, 2004; Boyd and Vandenberghe, 2004; Bach and Moulines, 2013; Bubeck, 2015). Considérons le problème d'apprentissage général décrit en Section 1.1.1, en supposant de plus que $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, où Θ est une partie convexe de $\mathbf{R}^d$. Pour tout $z \in \mathcal{Z}$, on définit la fonction $\ell_z : \Theta \to \mathbf{R}$ par

$$\ell_z(\theta) := \ell(f_\theta, z)\,,$$

de sorte que $R(\theta) = \mathbb{E}[\ell_Z(\theta)]$. Pour $i = 1,\dots,n$, on pose également $\ell_i := \ell_{Z_i}$, de sorte que $\ell_1,\dots,\ell_n$ sont des fonctions aléatoires i.i.d. de même loi que ℓ_Z. En posant $\theta^* = \arg\min R$, le but est de produire un élément $\widehat{\theta}_n \in \Theta$ (dépendant de $\ell_1,\dots,\ell_n$) dont l'excès de risque

$\mathcal{E}(\widehat{\theta}_n) = R(\widehat{\theta}_n) - R(\theta^*)$ est faible. Ainsi posé, ce problème (dit de l'*optimisation stochastique*) est équivalent à celui de l'apprentissage. Cependant, dans cette formulation, le prédicteur f_θ (qui correspond à une fonction $\mathcal{X} \to \widehat{\mathcal{Y}}$ en apprentissage supervisé) est maintenant vu comme un *paramètre* $\theta \in \Theta$, tandis que l'observation Z est représentée par la *fonction* ℓ_Z. Cette différence de formulation conduit à un point de vue complémentaire sur ce problème.

D'une part, dans le cas de l'optimisation stochastique, les hypothèses sont formulées en termes de propriétés analytiques des fonctions ℓ_z ou du risque R. En particulier, la convexité (et ses variantes renforcées) joue un rôle clé dans cette théorie (Boyd and Vandenberghe, 2004; Nesterov, 2004; Bubeck, 2015). Cela tient notamment au fait qu'il est possible de réduire l'étude de problèmes convexes à celle de problèmes *linéaires*. En effet, si la fonction ℓ_z est convexe et différentiable, alors pour tous $\theta, \theta^* \in \Theta$, l'excès de perte est contrôlé par le gradient $\nabla \ell_z(\theta)$:

$$\ell_z(\theta) - \ell_z(\theta^*) \leqslant \langle \nabla \ell_z(\theta), \theta - \theta^* \rangle \, ;$$

nous renvoyons à la Section 1.2.1 pour une conséquence précise de cette inégalité dans le cas de l'optimisation séquentielle. Du point de vue de l'*optimisation*, la convexité garantit que tout minimum local est un minimum global, et permet d'établir la convergence globale d'algorithmes d'optimisation. Du point de vue *statistique*, la *forte convexité* (voir la Section 1.6.3 pour une définition) implique une condition de marge (voir l'Exemple 1.9 ci-dessous), et permet donc d'obtenir des vitesses rapides pour l'excès de risque.

Exemple 1.9 (Forte convexité et condition de Bernstein)**.** Soit Θ une partie convexe de $\mathbf{R}^d$. Supposons que, pour tout $z \in \mathcal{Z}$, la fonction $\ell_z : \Theta \to \mathbf{R}$ est L-Lipschitz (par rapport à la norme euclidienne $\|\cdot\|$ sur $\mathbf{R}^d$) avec $L > 0$. Supposons également que la fonction de risque $R : \Theta \to \mathbf{R}$ est λ-*fortement convexe*, au sens de la Définition 1.8. Supposons enfin que le risque R est minimisé par un élément θ^* intérieur à Θ. Alors, pour tout $\theta \in \Theta$,

$$\mathbb{E}\big[\big(\ell(\theta, Z) - \ell(\theta^*, Z)\big)^2\big] \leqslant L^2 \|\theta - \theta^*\|^2 \leqslant \frac{2L^2}{\lambda}\big(R(\theta) - R(\theta^*)\big)$$

où l'on a utilisé l'inégalité (1.125) Section 1.6.3. Ceci signifie que la condition de Bernstein (Définition 1.2) est satisfaite avec $\beta = 1$ et $B = 2L^2/\lambda$.

Une autre particularité de l'approche "optimisation stochastique" est qu'elle porte souvent sur des procédures *explicites*, c'est-à-dire des algorithmes d'optimisation calculables à partir de certaines quantités associées aux fonctions ℓ_i (par exemple, leurs gradients), là où le point de vue "processus empiriques" conduit plus naturellement à des estimateurs définis implicitement (tel le minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$). L'algorithme typique pour les problèmes d'optimisation stochastique, en particulier lorsque le nombre d'observations n et la dimension d sont élevés, est la *descente de gradient stochastique* (en anglais *stochastic gradient descent*, abrégé SGD). La Proposition 1.2 suivante (Nemirovski and Yudin, 1983; Nemirovski et al., 2009; Bubeck, 2015) décrit une variante projetée et "en ligne" de cet algorithme, et énonce une borne d'excès de risque pour sa variante *moyennée* (Polyak and Juditsky, 1992; Ruppert, 1988).

Proposition 1.2. *Supposons que Θ est une partie convexe fermée de $\mathbf{R}^d$ et que, pour tout $z \in \mathcal{Z}$, la fonction $\ell_z : \Theta \to \mathbf{R}$ est convexe, différentiable et L-Lipschitz. Considérons l'algorithme de* descente de gradient stochastique projetée :

- $\widehat{\theta}_1 := \theta_1 \in \Theta$ *fixe ;*

- *pour tout* $i = 1, \dots, n$, *étant donné* $\widehat{\theta}_i$, *on définit :* $\widehat{\theta}_{i+1} := \mathrm{proj}_\Theta(\widehat{\theta}_i - \eta \nabla \ell_{Z_i}(\widehat{\theta}_i))$, *où* $\eta > 0$ *et* $\mathrm{proj}_\Theta(x) := \arg\min_{\theta \in \Theta} \|x - \theta\|$ *pour* $x \in \mathbf{R}^d$.

Soit $B > 0$ *; posons* $\eta = B/(L\sqrt{n})$, *et notons* $\Theta_B := \{\theta \in \Theta : \|\theta - \theta_1\| \leqslant B\}$. *Alors, la moyenne des itérés* $\bar{\theta}_n := \frac{1}{n+1} \sum_{i=1}^{n+1} \widehat{\theta}_i$ *satisfait la borne d'excès de risque suivante :*

$$\mathbb{E}[R(\bar{\theta}_n)] - \inf_{\theta \in \Theta_B} R(\theta) \leqslant \frac{BL}{\sqrt{n+1}}\,. \tag{1.18}$$

Proof. Ce résultat est une conséquence de la conversion online-to-batch (Proposition 1.3) décrite dans la Section 1.2.2 ci-dessous, et de la borne de regret de la descente de gradient en ligne (Proposition 1.11, Section 1.6.4). □

Exemple 1.10 (Prédiction linéaire)**.** Soit $\mathcal{Y}$ un espace mesurable, et $\widetilde{\ell} : \mathbf{R} \times \mathcal{Y} \to \mathbf{R}$ telle que $\widetilde{\ell}(\cdot, y)$ est convexe et C-Lipschitz pour tout $y \in \mathcal{Y}$. Cette condition inclut notamment, lorsque $\mathcal{Y} = \{-1, 1\}$, la *perte logistique* $\widetilde{\ell}(\widehat{y}, y) = \log(1 + e^{-y\widehat{y}})$ (explorée plus en détail dans les Sections 1.4.6 et 1.4.7) et la *perte Hinge* $\widetilde{\ell}(\widehat{y}, y) = \max(-y\widehat{y}, 0)$, avec $C = 1$; la perte *quadratique* bornée $\widetilde{\ell} : [-A, A]^2 \to \mathbf{R}$, $\widetilde{\ell}(\widehat{y}, y) = (\widehat{y} - y)^2$, satisfait également cette condition avec $C = 2A$. Soient également $\Theta = \mathbf{R}^d$ et $\mathcal{Z} = \mathcal{X} \times \mathbf{R}$, où $\mathcal{X} = \{x \in \mathbf{R}^d : \|x\| \leqslant R\}$ pour un certain $R > 0$. Alors, pour tout $z = (x, y) \in \mathcal{Z}$, la fonction $\ell_z(\theta) := \widetilde{\ell}(\langle \theta, x \rangle, y)$ est convexe et CR-Lipschitz ; la borne (1.18) d'excès de risque est donc de $CBR/\sqrt{n+1}$.

La Proposition 1.2 permet d'illustrer les points forts de l'approche en termes d'optimisation stochastique : d'une part, la garantie porte sur un estimateur explicite ; d'autre part, la preuve de la borne d'excès de risque est plus simple que celle fondée sur la convergence uniforme des processus empiriques. En effet, dans le cas particulier de la prédiction linéaire (Exemple 1.10), il est possible d'établir une borne similaire pour le minimiseur du risque empirique restreint à la boule de $\mathbf{R}^d$ de rayon B, par un argument de convergence uniforme reposant sur une inégalité (non triviale) de *contraction* pour les complexités de Rademacher, voir Koltchinskii (2011). Il s'avère de plus qu'il existe des exemples de problèmes d'optimisation stochastique convexe Lipschitz (satisfaisant les hypothèses de la Proposition 1.2) pour lesquels la convergence uniforme du risque empirique n'a pas lieu[2], et où l'excès de risque de l'ERM ne converge pas vers 0 (Shalev-Shwartz et al., 2010).

En outre, il est possible avec cette approche d'obtenir des vitesses rapides de manière plus directe qu'en ayant recours aux complexités de Rademacher localisées (Koltchinskii, 2006; Bartlett et al., 2005) ; une des façons de le faire est de se ramener à la variante séquentielle du problème, décrite dans la Section 1.2, voir par exemple Hazan et al. (2007); Audibert (2009); Lacoste-Julien et al. (2012) (la dernière référence utilisant une variante pondérée du problème séquentiel décrit en Section 1.2).

En contrepartie, cette approche conduit généralement à contrôler le risque par des quantités moins fines que celles apparaissant dans l'analyse en termes de processus empiriques. En effet, les bornes génériques d'excès de risque sous l'hypothèse de régularité Lipschitz (telle la Proposition 1.2) et/ou de forte convexité conduisent typiquement à des bornes dépendant du diamètre du paramètre de comparaison ou de constantes de courbure (forte convexité) uniformes. Or, le diamètre d'une classe est une mesure de complexité moins fine (en particulier, moins adaptative aux propriétés de la loi de Z) que la complexité de Rademacher. Par

[2]Plus précisément, ces contre-exemples sont en dimension infinie, où $\mathbf{R}^d$ est remplacé par un espace de Hilbert. En dimension finie, cela correspond au fait qu'ERM ne satisfait pas de borne de risque générale indépendante de la dimension d (par exemple, en $BL/\sqrt{n}$) en optimisation stochastique convexe Lipschitz.

ailleurs, la forte convexité de la perte en le paramètre θ n'est pas satisfaite pour les problèmes d'apprentissage classiques (telle la régression linéaire, voir la Section 1.3). Si la forte convexité du risque (qui apparaît dans l'Exemple 1.9) est nettement moins restrictive et satisfaite dans certains problèmes "bien conditionnés" en dimension finie (voir la Section 1.4.6), la constante de forte convexité admet généralement une dépendance implicite en la dimension du problème, et est très faible dans un contexte "non paramétrique" de grande dimension (Bach and Moulines, 2013). Il est cependant possible d'obtenir des garanties plus fines, avec une meilleure dépendance en la loi de Z (notamment au spectre de la Hessienne du risque) et sans hypothèse de forte convexité, dans le cas des problèmes quadratiques (Bach and Moulines, 2013; Dieuleveut and Bach, 2016), ou sous des hypothèses plus générales de régularité (Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019) ; nous revenons sur cette question dans la Section 1.4.6 sur la régression logistique.

1.1.6 Le point de vue de la stabilité

Il existe une autre technique générale pour établir des bornes d'excès de risque, reposant sur la notion de *stabilité*. De manière informelle, un prédicteur $\widehat{g}_n$ est dit "stable" s'il n'est pas sensible à de petites perturbations de l'échantillon $Z_1, \dots, Z_n$, et en particulier s'il n'est pas trop influencé par des observations individuelles Z_i.

La notion de stabilité et son lien avec l'erreur de généralisation ont été introduits par Bousquet and Elisseeff (2002), bien que des arguments de même nature remontent à Devroye and Wagner (1979a,b) et Kearns and Ron (1999). Le lien entre la notion de stabilité et les bornes de risque et d'erreur de généralisation a également été étudié par Rakhlin et al. (2005); Sridharan et al. (2009); Shalev-Shwartz et al. (2010). Nous nous plaçons ici dans le cadre de l'apprentissage statistique général introduit en Section 1.1.1.

Définition 1.3 (Stabilité par substitution)**.** Soit $\widehat{g}_n = \widehat{g}_n(Z_1, \dots, Z_n)$ un prédicteur formé à partir de $Z_1, \dots, Z_n$. Pour tous $z \in \mathcal{Z}$ et $i = 1, \dots, n$, notons $\widehat{g}_n^{[Z_i, z]}$ le prédicteur obtenu sur l'échantillon $(Z_1, \dots, Z_{i-1}, z, Z_{i+1}, \dots, Z_n)$ où Z_i a été remplacé par z.

Pour tout $\varepsilon > 0$, on dit que $\widehat{g}_n$ est *ε-stable par substitution (en moyenne)* si, en notant Z une réalisation indépendante de même loi P que $Z_1, \dots, Z_n$, on a

$$\mathbb{E}\big[\ell(\widehat{g}_n, Z) - \ell(\widehat{g}_n^{[Z_n, Z]}, Z)\big] \leqslant \varepsilon\,. \tag{1.19}$$

On dit également que $\widehat{g}_n$ est *uniformément ε-stable par substitution* si, pour tous $Z_1, \dots, Z_n$ et $z \in \mathcal{Z}$, $\ell(\widehat{g}_n, z) - \ell(\widehat{g}_n^{[Z_n, z]}, z) \leqslant \varepsilon$.

La stabilité par substitution uniforme implique la stabilité par substitution en moyenne. La stabilité permet de contrôler l'espérance de l'erreur de généralisation, c'est-à-dire de la différence $R(\widehat{g}_n) - \widehat{R}_n(\widehat{g}_n)$. Dans ce qui suit, nous supposons que $\widehat{g}_n$ dépend de façon symétrique de $Z_1, \dots, Z_n$, c'est-à-dire que sa valeur est inchangée par permutation de Z_i et Z_j, $i \neq j$.

Il existe en fait plusieurs notions de stabilité, en fonction de la notion de "distance" ainsi que du type de "perturbation" de l'échantillon considérées. Par exemple, si $\mathcal{G}$ est un espace vectoriel normé (de norme $\|\cdot\|$), il est en principe possible de définir la stabilité en fonction de la quantité $\|\widehat{g}_n^{[Z_n, Z]} - \widehat{g}_n\|$, ou de considérer la stabilité de l'estimateur à l'ajout (ou la suppression) d'un échantillon (plutôt que le remplacement). Dans la Définition 1.3, nous avons considéré la notion de stabilité de la perte par substitution, car elle conduit naturellement à des bornes d'excès de risque, comme le montrent les résultats suivants.

Lemme 1.1. *Si $\widehat{g}_n$ est un prédicteur symétrique en les observations, on a*

$$\mathbb{E}\big[R(\widehat{g}_n) - \widehat{R}_n(\widehat{g}_n)\big] = \mathbb{E}\big[\ell(\widehat{g}_n, Z) - \ell(\widehat{g}_n^{[Z_n,Z]}, Z)\big]\,. \tag{1.20}$$

En particulier, si $\widehat{g}_n$ est ε-stable par substitution, cette quantité est d'au plus ε.

Proof. Tout d'abord, Z étant indépendant de $Z_1, \dots, Z_n$ (et donc de $\widehat{g}_n$), on a $\mathbb{E}[\ell(\widehat{g}_n, Z)] = \mathbb{E}[\mathbb{E}[\ell(\widehat{g}_n, Z)|\widehat{g}_n]] = \mathbb{E}[R(\widehat{g}_n)]$.

En outre, la loi jointe de $(Z_1, \dots, Z_n, Z)$ est invariante par permutation des variables. Or, permuter Z_n et Z change $\ell(\widehat{g}_n^{[Z_n,Z]}, Z)$ en $\ell(\widehat{g}_n, Z_n)$, donc $\mathbb{E}[\ell(\widehat{g}_n^{[Z_n,Z]}, Z)] = \mathbb{E}[\ell(\widehat{g}_n, Z_n)]$; de même, échanger Z_i et Z_n change $\ell(\widehat{g}_n, Z_n)$ en $\ell(\widehat{g}_n, Z_i)$ (par symétrie de $\widehat{g}_n$ en les observations), donc $\mathbb{E}[\ell(\widehat{g}_n, Z_n)] = \mathbb{E}[\ell(\widehat{g}_n, Z_i)]$ pour $i = 1, \dots, n$. Ainsi

$$\mathbb{E}[\ell(\widehat{g}_n^{[Z_n,Z]}, Z)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(\widehat{g}_n, Z_i)] = \mathbb{E}[\widehat{R}_n(\widehat{g}_n)]\,. \qquad \square$$

Du Lemme 1.1 découle en particulier la borne de risque suivante sur des estimateurs régularisés en optimisation stochastique. Ce résultat est une variation mineure de résultats de Shalev-Shwartz et al. (2010); Sridharan et al. (2009), avec la différence que ces derniers considèrent l'estimateur régularisé (1.21) ci-dessous restreint à la boule de $\mathbf{R}^d$ de rayon $B > 0$.

Corollaire 1.1. *Supposons que $\Theta = \mathbf{R}^d$ et que pour tout $z \in \mathcal{Z}$, la fonction $\theta \mapsto \ell(\theta, z) := \ell(f_\theta, z)$ est convexe et L-Lipschitz. Définissons, pour tout $\lambda > 0$,*

$$\widehat{\theta}_{\lambda,n} := \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i) + \frac{\lambda}{2}\|\theta\|^2 \right\}. \tag{1.21}$$

Alors, la borne d'excès de risque suivante est valide :

$$\mathbb{E}[R(\widehat{\theta}_{\lambda,n})] - \inf_{\theta \in \mathbf{R}^d} \left\{ R(\theta) + \frac{\lambda}{2}\|\theta\|^2 \right\} \leqslant \frac{4L^2}{\lambda n}\,. \tag{1.22}$$

En particulier, pour tout $B > 0$, le choix de $\lambda = 2\sqrt{2} \cdot L/(B\sqrt{n})$ conduit à :

$$\mathbb{E}[R(\widehat{\theta}_{\lambda,n})] - \inf_{\|\theta\| \leqslant B} R(\theta) \leqslant \frac{2\sqrt{2}BL}{\sqrt{n}}\,.$$

Proof. Soit $\widehat{R}_{\lambda,n}(\theta) := \widehat{R}_n(\theta) + \lambda\|\theta\|^2/2$ le risque empirique pénalisé. On a

$$\widehat{\theta}_{\lambda,n}^{[Z_n,Z]} = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left\{ \frac{1}{n}\left[\sum_{i=1}^{n-1} \ell(\theta, Z_i) + \ell(\theta, Z)\right] + \frac{\lambda}{2}\|\theta\|^2 \right\} = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \Big\{ \widehat{R}_{\lambda,n}(\theta) + \big(\ell(\theta, Z) - \ell(\theta, Z_n)\big) \Big\}.$$

Or, la fonction $\widehat{R}_{\lambda,n}$ est λ-fortement convexe (en tant que somme de la fonction λ-fortement convexe $\theta \mapsto \lambda\|\theta\|^2/2$ et de la fonction convexe $\widehat{R}_n$, par convexité de $\ell(\cdot, z)$) ; de plus, la fonction $\theta \mapsto \ell(\theta, Z) - \ell(\theta, Z_n)$ est $2L$-Lipschitz. Par le Lemme 1.3 de l'annexe technique (Section 1.6.3), on en déduit que

$$\ell(\widehat{\theta}_{\lambda,n}, Z) - \ell(\widehat{\theta}_{\lambda,n}^{[Z_n,Z]}, Z) \leqslant \frac{4L^2}{\lambda}\,,$$

c'est-à-dire que $\widehat{\theta}_{\lambda,n}$ est $(4L^2)/\lambda$-uniformément stable par substitution. Par le Lemme 1.1, on a donc $\mathbb{E}[R(\widehat{\theta}_n) - \widehat{R}_n(\widehat{\theta}_{\lambda,n})] \leqslant 4L^2/\lambda$. La borne (1.22) s'en déduit en notant que, pour tout $\theta \in \mathbf{R}^d$,

$$\mathbb{E}[\widehat{R}_n(\widehat{\theta}_{\lambda,n})] \leqslant \mathbb{E}[\widehat{R}_{\lambda,n}(\widehat{\theta}_{\lambda,n})] \leqslant \mathbb{E}[\widehat{R}_{\lambda,n}(\theta)] = R(\theta) + \frac{\lambda}{2}\|\theta\|^2 . \qquad \square$$

La stabilité est un paradigme complémentaire à celui de la convergence des processus empiriques (Section 1.1.4). Cette notion sous-tend les bornes obtenues pour le problème de l'apprentissage séquentiel décrit dans la Section 1.2, et par conséquent la preuve de la Proposition 1.2, voir également la Section 1.2.5. De plus, l'existence d'estimateurs stables minimisant approximativement le risque empirique caractérise la possibilité d'estimateurs consistants dans un cadre général, même lorsque la convergence uniforme du risque empirique vers le risque n'a pas lieu, comme par exemple pour le problème d'optimisation stochastique considéré dans la Proposition 1.2 et le Corollaire 1.1 (Shalev-Shwartz et al., 2010). Dans le Chapitre 7 (voir également la Section 1.4.4 de cette introduction), nous introduisons un raffinement des bornes d'excès de risque en termes de stabilité de la perte, qui conduit dans le cas de l'estimation de densité (Exemples 1.2 et 1.5), et en particulier de la régression logistique, à une procédure admettant des garanties d'excès de risque améliorées.

1.1.7 Décomposition biais-variance et approximation quadratique locale

Considérons le cas de la régression avec perte quadratique (Exemple 1.4). Étant donné un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ (où X prend ses valeurs dans l'espace mesurable $\mathcal{X}$ et Y dans $\mathbf{R}$), le but est de produire un estimateur $\widehat{g}_n$ de la fonction $g^*(x) = \mathbb{E}[Y|X = x]$ de régression de Y sachant X (on suppose ici que $\mathbb{E}[Y^2] < +\infty$). Dans ce cas, l'excès de risque de $\widehat{g}_n$ par rapport à g^* satisfait la *décomposition biais-variance* suivante :

$$\mathbb{E}[R(\widehat{g}_n)] - R(g^*) = \mathbb{E}\big[(\widehat{g}_n(X) - g^*(X))^2\big] = \mathbb{E}\big[(\widehat{g}_n(X) - \bar{g}_n(X))^2\big] + \mathbb{E}\big[(\bar{g}_n(X) - g^*(X))^2\big]$$

où $\bar{g}_n(x) := \mathbb{E}[\widehat{g}_n(x)]$ pour tout $x \in \mathcal{X}$; le premier terme du membre de droite de l'équation ci-dessus correspond à la *variance* (il est égal à $\mathbb{E}[\mathrm{Var}(\widehat{g}_n(X)|X)]$), tandis que le second est par définition le *biais* (voir également la Section 1.5.2, qui traite des estimateurs "ensemblistes").

La décomposition biais-variance simplifie considérablement l'étude de la régression avec perte quadratique, en permettant des calculs explicites (Györfi et al., 2002). Elle est cependant spécifique à cette fonction de perte, à l'inverse des approches fondées sur la convergence de processus empiriques (Section 1.1.4), la stabilité (Section 1.1.6) ou la conversion "online-to-batch" (décrite dans la Section 1.2.2 ci-dessous). Il est en revanche possible d'étendre les résultats obtenus pour la perte quadratique à des fonctions de pertes "lisses" au moyen d'approximations quadratiques locales du risque. Cela est possible lorsque l'on dispose d'un contrôle de l'erreur de cette approximation quadratique ; la notion d'*auto-concordance*, c'est-à-dire un contrôle de la dérivée troisième en fonction de la dérivée seconde, permet d'étendre l'analyse fine des problèmes quadratiques à d'autres cas (Bach, 2010; Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019). Nous reviendrons sur cette approche dans la Section 1.4.6 dédiée à la régression logistique.

1.2 Prédiction séquentielle de suites arbitraires

Dans la Section 1.1.3, nous avons distingué les points de vue génératif (reposant sur l'hypothèse que la loi P appartient à un modèle $\mathcal{P}$) et discriminatif (qui ne fait pas cette hypothèse, mais considère l'excès de risque par rapport à une classe restreinte $\mathcal{F}$) du problème de l'apprentissage statistique. Dans cette section, nous considérons un problème voisin de l'apprentissage statistique, l'*apprentissage séquentiel* (aussi appelé *apprentissage en ligne*, ou *online learning*), pour lequel il est possible de se passer de l'hypothèse de stochasticité des observations, c'est-à-dire de l'existence d'une loi P telle que $Z_1, \dots, Z_n$ soient des variables i.i.d. de loi P.

L'ouvrage de référence sur l'apprentissage séquentiel est Cesa-Bianchi and Lugosi (2006) ; ce compte-rendu complet couvre notamment la théorie de la prédiction séquentielle avec perte logarithmique (dont Merhav and Feder, 1998 fournit une présentation spécifique), qui est l'une des sources de cette théorie, ainsi que les liens avec la théorie des jeux (Blackwell, 1956; Von Neumann and Morgenstern, 1947) et l'optimisation. Pour une approche complémentaire, nous renvoyons également aux ouvrages plus récents Shalev-Shwartz (2012); Hazan (2016), qui mettent en avant le lien avec l'optimisation convexe. Enfin, le problème de l'apprentissage séquentiel admet une variante à information partielle (incluant le problème dit des "bandits"), dont nous ne traiterons pas ici, présentée dans les ouvrages Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2019). Des présentations plus succinctes de la prédiction séquentielle et des problèmes de bandits figurent également dans les articles introductifs Stoltz (2010); Faure et al. (2015).

La prédiction séquentielle fait l'objet de la Partie II de ce manuscrit. Après une introduction au problème (Section 1.2.1), une discussion du lien avec l'apprentissage statistique (Section 1.2.1) et quelques résultats classiques sur l'agrégation à poids exponentiels (Sections 1.2.3 à 1.2.6), nous présentons nos contributions dans les Sections 1.2.7 (Chapitre 4) et 1.2.8 (Chapitre 5).

1.2.1 Apprentissage séquentiel

Dans cette section, nous introduisons le formalisme de l'apprentissage séquentiel, et détaillons quelques formulations de ce problème.

Apprentissage séquentiel général. Nous adoptons ici les mêmes notations que dans la Section 1.1.1. Dans le cas de l'apprentissage séquentiel, étant donnée une fonction de perte $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$, l'objectif est de prédire une à une des observations $z_1, \dots, z_n \in \mathcal{Z}$. Plus précisément, le problème se formule comme un jeu entre un Agent (qui cherche à prédire les observations) et un Environnement (qui génère celles-ci). À chaque instant $t = 1, \dots, n$:

- l'Agent choisit un prédicteur $\widehat{g}_t \in \mathcal{G}$, en se fondant sur les observations précédentes $z_1, \dots, z_{t-1}$;
- l'Environnement révèle la valeur z_t de l'observation au temps t (qui peut dépendre de $\widehat{g}_1, \dots, \widehat{g}_t$). L'Agent subit alors une perte de $\ell(\widehat{g}_t, z_t)$.

L'objectif de l'Agent est d'obtenir une perte cumulée faible. Il n'est pas difficile de voir que cela n'est pas possible sans hypothèse sur la suite $z_1, \dots, z_n$: en effet, à chaque instant t, l'Environnement peut choisir z_t en fonction de $\widehat{g}_t$, de sorte que $\ell(\widehat{g}_t, z_t)$ soit élevé. En d'autres termes, quelle que soit la stratégie de l'Agent, il existe une suite d'observations pour laquelle

la stratégie mène à une erreur élevée. Comme dans le cas de l'apprentissage statistique, cette difficulté est levée en adoptant une approche discriminative (Section 1.1.3), consistant à restreindre la classe de comparaison plutôt que la suite $z_1, \dots, z_n$. Ainsi, étant donnée une sous-classe $\mathcal{F} \subset \mathcal{G}$, l'objectif est de déterminer une stratégie de l'Agent telle que le *regret*

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t) \tag{1.23}$$

soit contrôlé indépendamment de (ou sous de faibles hypothèses sur) la suite d'observations $z_1, \dots, z_n$. À l'instar de l'excès de risque, il est possible de définir l'excès de risque minimax. Étant donné un ensemble $\mathcal{S} \subset \mathcal{Z}^n$ de suites $(z_1, \dots, z_n)$ d'observations et une classe $\mathcal{F}$ de prédicteurs de référence, le *regret minimax* est par définition:

$$\mathsf{Reg}_n^*(\ell, \mathcal{S}, \mathcal{F}) := \inf_{g_1, \dots, g_n} \sup_{(z_1, \dots, z_n) \in \mathcal{S}} \sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t) \tag{1.24}$$

où $(g_1, \dots, g_n)$ est une *stratégie de prédiction*, c'est-à-dire que g_t est une fonction de $\mathcal{Z}^{t-1}$ vers $\mathcal{G}$, et où $\widehat{g}_t := g_t(z_1, \dots, z_{t-1})$. Le biais inductif est alors introduit à travers le choix de la classe $\mathcal{F}$, plutôt que la suite $\mathcal{S}$ d'observations ; il est alors typique de choisir $\mathcal{S} = \mathcal{Z}^n$.

La notion de regret constitue l'analogue séquentiel de celle d'excès de risque. Comme indiqué au début de cette section, il n'est pas nécessaire de supposer ici que la suite $z_1, \dots, z_n$ est formée de n variables i.i.d. Au contraire, l'objectif est d'obtenir des garanties de regret (1.23) valables pour toute suite $z_1, \dots, z_n \in \mathcal{Z}$; on parle alors de garanties valables pour des *suites arbitraires* (*individual sequences* en anglais). Notons que dans le cas de l'apprentissage statistique, l'hypothèse d'observations aléatoires suivant une loi P est nécessaire à la formulation du problème, puisqu'elle permet de définir le risque (erreur moyenne sur une population non observée). Dans le cas de l'apprentissage séquentiel, l'erreur est mesurée sur la suite d'observations $z_1, \dots, z_n$ elle-même (chaque prédicteur $\widehat{g}_t$ étant choisi avant d'observer z_t), ce qui permet de se passer de l'hypothèse de stochasticité.

Nous passerons en revue de manière détaillée deux exemples du problème de l'apprentissage séquentiel : l'agrégation d'experts (Sections 1.2.3 à 1.2.8) et l'estimation de densité en ligne (Section 1.4.3). Nous concluons cette présentation en évoquant la variante *supervisée* de l'apprentissage séquentiel, ainsi que l'*optimisation convexe en ligne*.

Apprentissage séquentiel supervisé. Soient $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$ trois ensembles, et $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ une fonction de perte. Le problème de l'apprentissage séquentiel *supervisé* est défini par la classe $\mathcal{G}$ des fonctions $\mathcal{X} \to \widehat{\mathcal{Y}}$, l'espace $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, la fonction de perte $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ définie par $\ell(g, (x, y)) = \ell(g(x), y)$, ainsi qu'une classe de fonctions $\mathcal{F} \subset \mathcal{G}$. Dans ce cas, choisir une fonction $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ (avant de voir (x, y)) et recevoir la perte $\ell(g, (x, y))$ étant donné (x, y) équivaut à voir x, puis choisir $\widehat{y} = g(x)$ étant donné x, puis subir la perte $\ell(\widehat{y}, y)$ étant donné y. Ainsi, le problème admet la formulation équivalente suivante: pour tout $t \geqslant 1$,

- l'Environnement révèle la valeur $x_t \in \mathcal{X}$ de la variable prédictive ;
- l'Agent effectue une prédiction $\widehat{y}_t \in \widehat{\mathcal{Y}}$ (dépendant de x_t et des observations passées) ;
- l'Environnement révèle la valeur y_t de la réponse au temps t (pouvant dépendre de $x_t, \widehat{y}_t$ ainsi que des observations et prédictions passées). L'Agent subit alors une perte $\ell(\widehat{y}_t, z_t)$.

Le regret s'écrit alors :

$$\mathsf{Reg}_n = \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\,.$$

Optimisation en ligne. Tout comme l'apprentissage statistique, l'optimisation stochastique admet une variante séquentielle, appelée *optimisation (convexe) en ligne* (Zinkevich, 2003; Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012; Hazan, 2016). Le problème se formule de la façon suivante : étant donné un domaine de contrainte $\Theta \subset \mathbf{R}^d$ et une classe $\mathscr{C}$ de fonctions $\Theta \to \mathbf{R}$ (par exemple, si Θ est convexe, les fonctions convexes et L-Lipschitz sur Θ), à chaque instant $t = 1, \dots, n$,

- L'Agent choisit un élément $\widehat{\theta}_t \in \Theta$ (dépendant de $\ell_1, \dots, \ell_{t-1}$) ;
- L'Environnement révèle une fonction $\ell_t \in \mathscr{C}$ (qui peut dépendre de $\widehat{\theta}_1, \dots, \widehat{\theta}_t$). L'Agent subit alors la perte $\ell_t(\widehat{\theta}_t)$.

Le but est alors de déterminer une stratégie de choix de $\widehat{\theta}_1, \dots, \widehat{\theta}_n$ pour laquelle le *regret*

$$\mathsf{Reg}_n = \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \ell_t(\theta) \tag{1.25}$$

est contrôlé indépendamment de la suite $\ell_1, \dots, \ell_n$ de fonctions. De la même manière que dans le cas statistique, l'optimisation en ligne est équivalente à l'apprentissage séquentiel *propre*, c'est-à-dire avec la restriction $\mathcal{F} = \mathcal{G}$.

Tout comme en optimisation stochastique, la notion de convexité joue un rôle clé en optimisation en ligne. Cela tient au fait que le regret dans le cas convexe peut être majoré par le regret dans le cas linéaire. En effet, si Θ et $\ell_t : \Theta \to \mathbf{R}$ sont convexes et si ℓ_t est différentiable[3], alors pour tout $\theta \in \Theta$,

$$\ell_t(\widehat{\theta}_t) - \ell_t(\theta) \leqslant \langle \nabla \ell_t(\widehat{\theta}_t), \widehat{\theta}_t - \theta \rangle\,, \tag{1.26}$$

de sorte qu'en posant $h_t := \nabla \ell_t(\widehat{\theta}_t)$, le regret sur la suite de fonctions (ℓ_t) est majoré par celui sur la suite de fonctions linéaires $\langle h_t, \cdot \rangle$:

$$\mathsf{Reg}_n = \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \ell_t(\theta) \leqslant \sum_{t=1}^{n} \langle h_t, \widehat{\theta}_t \rangle - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \langle h_t, \theta \rangle\,. \tag{1.27}$$

Ceci montre que, du point de vue de l'optimisation en ligne, les fonctions linéaires sont les fonctions convexes les plus "difficiles", et qu'il est possible de transférer des résultats pour les fonctions linéaires aux fonctions convexes générales.

Agrégation d'experts. L'agrégation d'experts (ou *prédiction séquentielle à l'aide d'experts*, en anglais *prediction with expert advice*) est une formulation alternative du problème de la prédiction séquentielle, qui sera utilisée dans la Partie II de cette thèse. Dans ce cas, on se donne une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$, où $\widehat{\mathcal{Y}}$ est l'espace des prédictions et $\mathcal{Y}$ l'espace du *signal*, ainsi qu'un ensemble abstrait Θ d'*experts*, c'est-à-dire de sources de prédictions. Dans ce qui suit, on pourra supposer Θ fini. À chaque étape $t = 1, \dots, n$:

[3] Il est possible de se passer de cette hypothèse si $\widehat{\theta}_t$ est intérieur à Θ, en considérant un *sous-gradient* $h_t \in \partial \ell_t(\widehat{\theta}_t)$ satisfaisant par définition l'inégalité voulue (Boyd and Vandenberghe, 2004).

- l'Environnement révèle les prédictions $(\widehat{y}_{\theta,t})_{\theta\in\Theta} \in \widehat{\mathcal{Y}}^{\Theta}$ des experts $\theta \in \Theta$;
- l'Agent détermine sa propre prédiction $\widehat{y}_t \in \widehat{\mathcal{Y}}$;
- l'Environnement choisit la valeur $y_t \in \mathcal{Y}$ du signal.

Le but est alors de déterminer une stratégie de l'Agent garantissant un regret

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} \ell(\widehat{y}_{\theta,t}, y_t)$$

contrôlé. Notons que, dans ce cadre, les prédictions $\widehat{y}_{\theta,t}$ des experts sont arbitraires, et peuvent être considérées comme des "boîtes noires". Cela confère une grande flexibilité ; par exemple, les experts $\theta \in \Theta$ peuvent eux-mêmes correspondre à des stratégies de prédiction séquentielles, dont les prédictions dépendent des observations précédentes.

L'apprentissage séquentiel général (décrit dans le premier paragraphe de cette section) peut être vu comme un cas particulier de ce problème, en posant $\widehat{\mathcal{Y}} := \mathcal{G}$, $\mathcal{Y} := \mathcal{Z}$ (avec identification des fonctions de pertes), et enfin en posant $\widehat{y}_{\theta,t} := f_\theta$ pour tous $\theta \in \Theta$ et $t \geqslant 1$. L'intérêt de la formulation précédente est toutefois de faire le lien avec le problème général de l'apprentissage statistique, à travers des notations similaires et la conversion "online to batch" (voir la Section 1.2.2).

Il est possible d'envisager l'apprentissage séquentiel supervisé comme une variante de l'agrégation d'experts de manière plus directe. En reprenant les notations du cas supervisé (les espaces $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$, la fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ et la classe $\mathcal{F}$ de fonctions $\mathcal{X} \to \mathcal{Y}$), notons $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. Ce problème se réduit à l'agrégation d'experts avec la même fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$, en posant $\widehat{y}_{\theta,t} := f_\theta(x_t)$ pour tous $t = 1, \dots, n$ et $\theta \in \Theta$.

Enfin, dans les Sections 1.2.4 et 1.2.5, nous verrons que l'étude de l'apprentissage séquentiel général peut se ramener, sous certaines hypothèses sur la fonction de perte ℓ, à celle d'une variante de l'apprentissage séquentiel général décrit en début de section.

1.2.2 Conversion "online to batch"

Il existe un lien général entre l'apprentissage séquentiel et l'apprentissage statistique : il est en effet possible de convertir toute garantie de regret pour un algorithme de prédiction séquentielle en une borne d'excès de risque pour un certain prédicteur. Ce procédé appelé *conversion "online to batch"* (Littlestone, 1989; Cesa-Bianchi et al., 2004) a été introduit dans le cas de l'estimation de densité par Barron (1987); Catoni (1997); Yang (2000). Dans ce qui suit, nous reprenons les notations de la Section 1.1.1.

Proposition 1.3 (Conversion "online to batch"). *Soit $\widehat{g}_1, \dots, \widehat{g}_{n+1}$ des prédicteurs, tels que $\widehat{g}_t$ dépende de $Z_1, \dots, Z_{t-1}$. Supposons que $\mathcal{G}$ est un espace convexe*[4]*, et que la fonction $g \mapsto \ell(g, z)$ est convexe pour $z \in \mathcal{Z}$. Soit P une mesure de probabilité sur $\mathcal{Z}$, et $Z_1, \dots, Z_{n+1}$ des variables i.i.d. de loi P. Définissons le prédicteur moyenné $\bar{g}_n$, dépendant de $Z_1, \dots, Z_n$, par*

$$\bar{g}_n := \frac{1}{n+1} \sum_{t=1}^{n+1} \widehat{g}_t \,. \tag{1.28}$$

[4] C'est-à-dire une partie convexe (mesurable) d'un espace vectoriel réel (mesurable).

Alors, en notant Reg_{n+1} *le regret* (1.23) *de* $\widehat{g}_1, \dots, \widehat{g}_{n+1}$ *par rapport à la classe* $\mathcal{F}$ *sur la suite* $Z_1, \dots, Z_{n+1}$, $\bar{g}_n$ *satisfait la borne d'excès de risque suivante:*

$$\mathbb{E}\big[R(\bar{g}_n)\big] - \inf_{f \in \mathcal{F}} R(f) \leqslant \frac{1}{n+1} \mathbb{E}\big[\mathsf{Reg}_{n+1}\big] \,. \tag{1.29}$$

Proof. Pour tout $t = 1, \dots, n+1$, $\widehat{g}_t$ (qui est une fonction de $Z_1, \dots, Z_{t-1}$) est indépendant de Z_t, de sorte que $\mathbb{E}[\ell(\widehat{g}_t, Z_t)] = \mathbb{E}[\mathbb{E}[\ell(\widehat{g}_t, Z_t)|\widehat{g}_t]] = \mathbb{E}[R(\widehat{g}_t)]$. Ainsi, pour tout $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}\big[R(\bar{g}_n)\big] - R(f) &\leqslant \frac{1}{n+1} \sum_{t=1}^{n+1} \big(\mathbb{E}[R(\widehat{g}_t)] - R(f)\big) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{t=1}^{n+1} \big(\ell(\widehat{g}_t, Z_t) - \ell(f, Z_t)\big)\right] \leqslant \frac{\mathbb{E}[\mathsf{Reg}_{n+1}]}{n+1}
\end{aligned}$$

où la première inégalité découle de la convexité de ℓ (et donc de R). L'inégalité (1.29) s'en déduit en considérant le supremum sur $f \in \mathcal{F}$. □

Remarque 1.2 (Randomisation). L'hypothèse de la Proposition 1.3 que $\mathcal{G}$ est convexe et que ℓ est convexe en son premier argument n'est pas restrictive : en effet, il est toujours possible de s'y ramener en considérant des prédicteurs randomisés. Cela revient à considérer la classe $\mathcal{G}'$ des mesures de probabilités sur $\mathcal{G}$ (dont $\mathcal{G}$ s'identifie à un sous-ensemble, en associant à $g \in \mathcal{G}$ la mesure de Dirac δ_g), ainsi que la fonction de perte $\ell'(\rho, z) = \mathbb{E}_{g\sim\rho}[\ell(g, z)]$ pour $\rho \in \mathcal{G}'$ et $z \in \mathcal{Z}$, qui est linéaire (donc convexe) en ρ et satisfait $\ell'(\delta_g, z) = \ell(g, z)$. La conversion online-to-batch (1.28) revient alors à tirer $\bar{g}_n$ uniformément parmi $\widehat{g}_1, \dots, \widehat{g}_{n+1}$.

Ainsi, l'excès de risque minimax $\mathcal{E}_n^*(\ell, \mathcal{P}, \mathcal{F})$ (1.6) par rapport à une classe $\mathcal{F}$ avec $\mathcal{P} = \mathcal{P}(\mathcal{Z})$ est majoré par le regret minimax $\mathsf{Reg}_{n+1}^*(\ell, \mathcal{Z}^{n+1}, \mathcal{F})/(n+1)$. Nous discuterons plus en détail les avantages et les limites de la réduction de l'apprentissage statistique à l'apprentissage séquentiel dans le cas de l'estimation de densité en ligne (Section 1.4.3).

1.2.3 Agrégation à poids exponentiels

Dans cette section, nous décrivons une stratégie fondamentale de prédiction séquentielle, à savoir l'*agrégation à poids exponentiels* (Vovk, 1990, 1998; Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006). Cette procédure (et son analyse de regret) généralise les stratégies séquentielles de prédiction par *mélange Bayésien* utilisées dans le cas de la perte logarithmique (Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006).

Nous nous plaçons dans le cas du problème de l'agrégation d'experts, formulé à la fin de la Section 1.2.1. Pour tout $t \geqslant 1$, on note

$$\widehat{L}_t := \sum_{s=1}^{t} \ell(\widehat{y}_s, y_s), \qquad L_{\theta,t} := \sum_{s=1}^{t} \ell(\widehat{y}_{\theta,s}, y_s) \,,$$

les pertes cumulées respectives de l'algorithme de prédiction et de l'expert $\theta \in \Theta$.

Limite de la minimisation du risque empirique. La stratégie la plus naturelle consiste à prédire, à chaque étape $t \geqslant 1$, comme l'expert $\widehat{\theta}_{t-1} \in \Theta$ dont la perte cumulée sur les observations précédentes est la plus faible :

$$\widehat{y}_t := \widehat{y}_{\widehat{\theta}_{t-1},t}\,, \qquad \widehat{\theta}_{t-1} := \operatorname*{arg\,min}_{\theta\in\Theta} L_{\theta,t-1}\,. \tag{1.30}$$

Cette stratégie correspond à (la variante séquentielle de) la minimisation du risque empirique, considérée dans le cas statistique en Section 1.1.4. Cependant, à l'inverse du cas statistique, cette approche n'admet aucune garantie non triviale de regret valide pour des suites arbitraires, comme le montre l'exemple suivant.

Exemple 1.11 ("Inconsistance" d'ERM pour des suites arbitraires)**.** Considérons la perte quadratique $\ell(\widehat{y}, y) = (y - \widehat{y})^2$ sur $\widehat{\mathcal{Y}} = \mathcal{Y} = [0, 1]$, ainsi que la classe à deux experts $\Theta = \{1, 2\}$, avec $\widehat{y}_{1,t} = 0$ et $\widehat{y}_{2,t} = 1$ pour tout $t \geqslant 1$. Considérons la suite $(y_t)_{t\geqslant 1}$ donnée par $y_1 = 1/3$, $y_{2k} = 1$ et $y_{2k+1} = 0$ ($k \geqslant 1$). Alors, pour tout $k \geqslant 1$, la stratégie (1.30) prédit $\widehat{y}_{2k} = \widehat{y}_{1,2k} = 1$ au temps $2k$, et subit donc une perte $(\widehat{y}_{2k} - y_{2k})^2 = 1$; de même, au temps $2k + 1$, $(\widehat{y}_{2k+1} - y_{2k+1})^2 = (\widehat{y}_{2,2k+1} - y_{2k+1})^2 = 1$. Ainsi (quel que soit le choix de $\widehat{y}_1$), on a $\widehat{L}_n \geqslant n - 1$ pour tout $n \geqslant 1$, tandis que $\max(L_{1,n}, L_{2,n}) \leqslant n/2 + 1$, de sorte que le regret de la stratégie (1.30) est supérieur à $n/2 - 2$. Le regret est linéaire, le regret moyen ne tend donc pas vers 0.

Dans l'Exemple (1.30), la stratégie (1.30) est mise en défaut même pour une classe Θ de faible complexité ($|\Theta| = 2$). À l'inverse, dans le cas statistique, un argument de convergence uniforme (Section 1.1.4) montre que l'excès de risque d'ERM est d'au plus $O(1/\sqrt{n})$.

Agrégation à poids exponentiels. L'Exemple (1.30) met en évidence la fragilité de la minimisation du risque empirique pour des suites arbitraires, due à l'instabilité de ses prédictions. La stratégie d'agrégation à poids exponentiels, décrite ci-dessous, corrige ce défaut en stabilisant les prédictions.

Dans ce qui suit, supposons que l'ensemble Θ est un espace mesurable. Soit π une mesure de probabilité sur Θ, appelée *loi a priori*, et $\eta > 0$ un paramètre appelé *paramètre d'apprentissage* (ou *température inverse*). L'*agrégation à poids exponentiels* (APE), ou *algorithme Hedge* (Vovk, 1998; Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), de paramètre $\eta > 0$ est par définition la stratégie

$$\widehat{y}_t := \int_\Theta \widehat{y}_{\theta,t}\, v_t(\mathrm{d}\theta)\,, \quad \text{où} \qquad \frac{\mathrm{d}v_t}{\mathrm{d}\pi}(\theta) := \frac{e^{-\eta L_{\theta,t-1}}}{\int_\Theta e^{-\eta L_{\theta',t-1}}\pi(\mathrm{d}\theta')}\,. \tag{1.31}$$

Le paramètre d'apprentissage η quantifie l'attache aux données de l'algorithme : pour $\eta \to 0$, l'APE ne dépend pas des données, et effectue la moyenne selon π des prédictions des experts ; pour $\eta \to \infty$, l'APE se réduit à la minimisation du risque empirique[5]. La mesure de probabilité v_t sur Θ (qui donne les "poids" des différents experts $\theta \in \Theta$) est parfois appelée *postérieur* (par analogie avec le postérieur Bayésien, voir la Section 1.4.3).

Remarque 1.3. Dans ce qui suit, nous supposons donnée une notion d'espérance de variables aléatoires à valeurs dans $\widehat{\mathcal{Y}}$ satisfaisant l'inégalité de Jensen pour les fonctions convexes considérées. Cette propriété est par exemple satisfaite lorsque $\widehat{\mathcal{Y}}$ est l'espace des mesures de

[5]Ou, plus précisément, à la moyenne selon π des experts θ de perte cumulée minimale.

probabilité sur $\mathcal{Y}$ (auquel cas l'espérance est le mélange de distributions) et $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est la perte logarithmique, ou lorsque Θ est fini (car les espérances se réduisent à des combinaisons convexes finies).

Afin d'éviter les questions d'intégrabilité et de continuité, on pourra supposer l'ensemble des experts Θ fini. Dans ce cas, la mesure de probabilité $v \in \mathcal{P}(\Theta)$ s'identifie au vecteur $(v_\theta)_{\theta\in\Theta}$ avec $v_\theta := v(\{\theta\})$, et l'intégrale de $h : \Theta \to \mathbf{R}$, $h(\theta) = h_\theta$ vaut $\int_\Theta h\mathrm{d}v = \sum_{\theta\in\Theta} v_\theta h_\theta$. Nous utilisons cependant des notations "fonctionnelles" génériques pour indiquer que les bornes ne dépendent pas explicitement du nombre $|\Theta|$ d'experts et s'étendent au cas de classes infinies, sous réserve d'intégrabilité.

1.2.4 Le cas des pertes exp-concaves

Dans cette section, nous analysons le comportement de l'APE pour une classe de fonctions de perte, dites "exp-concaves".

Concavité exponentielle. La concavité exponentielle est une propriété de courbure de la fonction de perte, analogue à celle de forte convexité.

Définition 1.4 (Concavité exponentielle)**.** Supposons que $\widehat{\mathcal{Y}}$ est un espace convexe, et soit $\eta > 0$. Une fonction $f : \widehat{\mathcal{Y}} \to \mathbf{R}$ est dite *η-exp-concave* si $\exp(-\eta f) : \widehat{\mathcal{Y}} \to \mathbf{R}$ est concave. De même, une fonction $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est dite *η-exp-concave* si $\ell(\cdot, y)$ l'est pour tout $y \in \mathcal{Y}$.

Une fonction exp-concave est convexe (par composition avec la fonction convexe décroissante $-\eta^{-1}\log$). Une fonction η-exp-concave est également η'-exp-concave pour tout $\eta' < \eta$ (par concavité de $x \mapsto x^{\eta'/\eta}$). De plus, une fonction $f : \Omega \to \mathbf{R}$ deux fois différentiable avec Ω un ouvert convexe de $\mathbf{R}^d$ est η-exp-concave si et seulement si

$$0 \succcurlyeq \nabla^2 \exp(-\eta f) = -\eta \exp(-\eta f)\left[\nabla^2 f - \eta(\nabla f)(\nabla f)^\top\right],$$

c'est-à-dire si et seulement si $\nabla^2 f \succcurlyeq \eta(\nabla f)(\nabla f)^\top$. La concavité exponentielle est donc une propriété de courbure, qui stipule que la Hessienne est minorée dans la direction du gradient.

Exemple 1.12 (Apprentissage supervisé)**.** Considérons le cas de l'apprentissage séquentiel supervisé (Section 1.2.1). Si $\widehat{\mathcal{Y}}$ est convexe et si $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est η-exp-concave, alors la fonction $\ell : (g, (x, y)) \mapsto \ell(g(x), y)$ est η-exp-concave.

Exemple 1.13 (Pertes classiques)**.** Considérons les fonctions de perte suivantes :

- la perte logarithmique (Exemple 1.2) est 1-exp-concave, car $\exp(-\ell(f, z)) = f(z)$ est linéaire en f.
- la perte quadratique $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ sur $\widehat{\mathcal{Y}} \times \mathcal{Y} = [-B, B]^2$ est $1/(8B^2)$-exp-concave.
- la perte absolue $\ell(\widehat{y}, y) = |\widehat{y} - y|$ sur $[0, 1]^2$ est convexe mais n'est pas exp-concave.

Remarque 1.4 (Mélangeabilité). Signalons également une généralisation de la convexité exponentielle, la *mélangeabilité* introduite par Vovk (Vovk, 1998; Haussler et al., 1998; Cesa-Bianchi and Lugosi, 2006). Une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est dite η-mélangeable si, pour tous $M \geqslant 1$, $\widehat{y}_1, \ldots, \widehat{y}_M \in \mathcal{G}$ et $v_1, \ldots, v_M \geqslant 0$ tels que $\sum_{i=1}^M v_i = 1$, il existe un élément $\widehat{y} \in \mathcal{G}$ tel que, pour tout $y \in \mathcal{Y}$,

$$\exp(-\eta\ell(\widehat{y}, y)) \geqslant \sum_{i=1}^M v_i \exp(-\eta\ell(\widehat{y}_i, y)).$$

Une perte η-exp-concave est η-mélangeable, en prenant pour $\widehat{y}$ la combinaison $\sum_{i=1}^{M} v_i \widehat{y}_i$. La notion de mélangeabilité est donc une extension de la convexité exponentielle, qui est indépendante de la paramétrisation et de l'éventuelle structure convexe de $\widehat{\mathcal{Y}}$. La perte logarithmique étant η-mélangeable si et seulement si $\eta \leqslant 1$, la notion de mélangeabilité n'apporte pas de gain pour cette perte. En revanche, la perte quadratique sur $[-B, B]$ est $1/(2B^2)$-mélangeable (Haussler et al., 1998; Vovk, 1998) ; la mélangeabilité permet donc un gain d'un facteur 4 dans le paramètre de "courbure" η, en ayant recours à une combinaison $\widehat{g}$ différente de la combinaison convexe (Vovk, 1990). En revanche, cette combinaison dépend de la borne supposée connue B sur les valeurs de $y, \widehat{y}$, ce qui la rend plus difficilement applicable en pratique.

Dans la plupart des exemples usuels, la notion de mélangeabilité coïncide avec celle de concavité exponentielle dans une paramétrisation bien choisie. Pour cette raison, nous nous restreignons dans ce texte à la seconde notion.

Remarque 1.5 (Extensions stochastiques). Tout comme la forte convexité, la concavité exponentielle permet d'obtenir des vitesses améliorées. Il existe également, dans le cas statistique, des extensions "stochastiques" des notions de concavité exponentielle et de mélangeabilité, dépendant de la loi P et permettant des vitesses rapides même dans le cas de la classification (Juditsky et al., 2008; Audibert, 2009; van Erven et al., 2015). Ces conditions sont des hypothèses de marge intimement liées à la condition de Bernstein (Définition 1.2) ; nous renvoyons à van Erven et al. (2015) pour plus de détails.

Réduction à la perte de mélange. Si la fonction de perte ℓ est η-exp-concave, alors pour toute mesure de probabilité $v \in \mathcal{P}(\Theta)$, toutes prédictions $(\widehat{y}_\theta)_{\theta\in\Theta}$ des experts et tout $y \in \mathcal{Y}$,

$$\ell\Big(\int_\Theta \widehat{y}_\theta \, v(\mathrm{d}\theta), y\Big) \leqslant -\frac{1}{\eta} \log\Big(\int_\Theta e^{-\eta\ell(\widehat{y}_\theta, y)} v(\mathrm{d}\theta)\Big) . \tag{1.32}$$

Cette inégalité montre qu'il est possible, dans le cas de stratégies utilisant une combinaison convexe des prédictions des experts, de réduire l'agrégation d'experts à un problème d'apprentissage séquentiel général (Section 1.2.1), défini par la fonction de perte suivante.

Définition 1.5 (Perte de mélange)**.** Soit $\mathcal{G} = \mathcal{P}(\Theta)$ et $\mathcal{Z}$ l'espace des fonctions mesurables $\Theta \to \mathbf{R}$. La *perte de mélange* est la fonction de perte $\ell_{\mathrm{mix}} : \mathcal{G} \times \mathcal{Z} \to \mathbf{R} \cup \{-\infty\}$ (à valeurs finies si $h \geqslant 0$ ou si Θ est fini) définie par

$$\ell_{\mathrm{mix}}(v, h) := -\log\Big(\int_\Theta e^{-h(\theta)} v(\mathrm{d}\theta)\Big) . \tag{1.33}$$

Pour tout $t \geqslant 1$, soit $h_t : \Theta \to \mathbf{R}$ (*i.e.*, $h_t \in \mathcal{Z}$) la fonction définie par $h_t(\theta) := \eta \cdot \ell(\widehat{y}_{\theta,t}, y_t)$. Il résulte de l'inégalité (1.32) que, si $\widehat{y}_t = \int_\Theta \widehat{y}_{\theta,t} v_t(\mathrm{d}\theta)$, alors

$$\ell(\widehat{y}_t, y_t) \leqslant \frac{1}{\eta} \ell_{\mathrm{mix}}(v_t, h_t) .$$

Puisqu'en outre $\ell(\widehat{y}_{\theta,t}, y_t) = h_t(\theta)/\eta = \ell_{\mathrm{mix}}(\delta_\theta, h_t)/\eta$ pour tout $\theta \in \Theta$, on a

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \sum_{t=1}^{n} \ell(\widehat{y}_{\theta,t}, y_t) \leqslant \frac{1}{\eta}\Bigg(\sum_{t=1}^{n} \ell_{\mathrm{mix}}(v_t, h_t) - \sum_{t=1}^{n} \underbrace{\ell_{\mathrm{mix}}(\delta_\theta, h_t)}_{h_t(\theta)}\Bigg) .$$

Ainsi, le regret pour l'agrégation d'experts est majoré par (η^{-1} fois) le regret pour la perte $\ell_{\mathrm{mix}} : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ par rapport à la classe $\mathcal{F} = \{\delta_\theta : \theta \in \Theta\}$.

Regret de l'APE. Nous considérons maintenant l'APE appliquée à la perte de mélange. Avec la définition de h_t précédente, le postérieur v_t de l'APE (1.31) s'écrit

$$v_t = \frac{\exp\big(-\sum_{s=1}^{t-1} h_s\big)}{\int_\Theta \exp\big(-\sum_{s=1}^{t-1} h_s\big)\mathrm{d}\pi} \cdot \pi\,,$$

soit :

$$v_1 = \pi\,, \qquad v_{t+1} = \frac{\exp(-h_t)}{\int_\Theta \exp(-h_t)\mathrm{d}v_t} \cdot v_t\,. \tag{1.34}$$

Dans ce qui suit, nous notons $\mathrm{KL}(\rho,\pi) := \int_\Theta \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}\mathrm{d}\rho$ la *divergence de Kullback-Leibler* de ρ par rapport à π (voir la Section 1.6).

Proposition 1.4 (Vovk, 1998; Littlestone and Warmuth, 1994)**.** *Considérons la procédure d'APE donnée par (1.34). Alors, pour toute suite $h_1,\dots,h_n \in \mathcal{Z}$ et toute mesure de probabilité $\rho \in \mathcal{P}(\Theta)$, on a*

$$\sum_{t=1}^n \ell_{\mathrm{mix}}(v_t,h_t) - \int_\Theta \Big(\sum_{t=1}^n h_t(\theta)\Big)\rho(\mathrm{d}\theta) \leqslant \mathrm{KL}(\rho,\pi)\,. \tag{1.35}$$

Proof. Cette proposition est une conséquence de la formule variationnelle de Donsker-Varadhan (Théorème 1.18 de l'annexe technique, Section 1.6). En effet, en combinant l'équation (1.34) avec l'identité (1.122) (avec $f = -h_t$, $\rho = v_t$), il vient pour tous $t = 1,\dots,n$ et $\rho \in \mathcal{P}(\Theta)$,

$$\ell_{\mathrm{mix}}(v_t,h_t) - \int_\Theta h_t(\theta)\rho(\mathrm{d}\theta) = \mathrm{KL}(\rho,v_t) - \mathrm{KL}(\rho,v_{t+1})\,.$$

La borne (1.35) s'obtient alors en sommant sur $t = 1,\dots,n$, en simplifiant la somme télescopique et en utilisant le fait que $v_1 = \pi$ et $\mathrm{KL}(\rho,v_{n+1}) \geqslant 0$. □

De la Proposition 1.4 et de la réduction précédente découlent le résultat suivant.

Corollaire 1.2 (Regret de l'APE : cas exp-concave)**.** *Si la fonction ℓ est η-exp-concave, alors l'APE (1.31) de paramètre η satisfait, pour tout $\rho \in \mathcal{P}(\Theta)$:*

$$\sum_{t=1}^n \ell(\widehat{y}_t,y_t) - \int_\Theta \Big(\sum_{t=1}^n \ell(\widehat{y}_{\theta,t},y_t)\Big)\rho(\mathrm{d}\theta) \leqslant \frac{\mathrm{KL}(\rho,\pi)}{\eta} \tag{1.36}$$

quelles que soient les suites $y_1,\dots y_n \in \mathcal{Y}$ et $(\widehat{y}_{\theta,t})_{\theta\in\Theta,1\leqslant t\leqslant n}$. En particulier, si Θ est fini (ou dénombrable), pour tout $\theta \in \Theta$,

$$\sum_{t=1}^n \ell(\widehat{y}_t,y_t) - \sum_{t=1}^n \ell(\widehat{y}_{\theta,t},y_t) \leqslant \frac{\log(\pi_\theta^{-1})}{\eta}\,; \tag{1.37}$$

si $\Theta = \{1,\dots,M\}$ et π est la loi uniforme sur Θ, alors

$$\sum_{t=1}^n \ell(\widehat{y}_t,y_t) - \min_{1\leqslant i\leqslant M} \sum_{t=1}^n \ell(\widehat{y}_{i,t},y_t) \leqslant \frac{\log M}{\eta}\,. \tag{1.38}$$

La borne (1.36) contrôle la différence entre la perte de l'APE et la perte des experts moyennée selon une certaine loi ρ, en fonction de la "complexité" $\mathrm{KL}(\rho, \pi)$. Notons que cette borne est valide simultanément sur toutes les lois ρ, en particulier pour tout $K = \mathrm{KL}(\rho, \pi)$, pour une même procédure. Bien qu'elle implique les bornes (1.37) et (1.38) dans le cas fini, cette borne ne dépend pas explicitement du nombre $|\Theta|$ d'experts, ni même de la "complexité" de la classe Θ des experts. Il est alors possible d'optimiser le borne supérieure (1.36) sur ρ, par un argument de dualité (Théorème 1.18 de la Section 1.6.2) :

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) \leqslant \inf_{\rho \in \mathcal{P}(\Theta)} \left\{ \int_\Theta L_{\theta,n} \, \rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho, \pi)}{\eta} \right\} = -\frac{1}{\eta} \log \left(\int_\Theta e^{-\eta L_{\theta,n}} \, \pi(\mathrm{d}\theta) \right).$$

La quantité du membre de droite dépend de la loi de la perte $L_{\theta,n}$ selon π. En particulier, pour toute partie $E \subset \Theta$ telle que $\pi(E) > 0$, le choix de $\rho = \pi(E)^{-1} \mathbf{1}(\theta \in E) \cdot \pi$ donne

$$\widehat{L}_t - \sup_{\theta \in E} L_{\theta,n} \leqslant \frac{\log \pi(E)^{-1}}{\eta} ;$$

ainsi, pour tout $\varepsilon \in (0, 1)$, le regret de l'APE par rapport à la fraction ε (au sens de la loi π sur Θ) des meilleurs experts est d'au plus $\log(1/\varepsilon)/\eta$. Cette borne est d'autant plus petite que la loi a priori π attribue une probabilité importante à des paramètres $\theta \in \Theta$ de faible perte. Notons enfin qu'un regret constant (1.38) en $O((\log M)/\eta)$ correspond à un regret moyen en $O((\log M)/(\eta n))$, c'est-à-dire à une vitesse rapide.

1.2.5 Pertes (convexes) bornées et problème de Hedge

Dans cette section, nous ne supposons plus que la perte ℓ est η-exp-concave. Nous supposons cependant que $0 \leqslant \ell \leqslant B$ pour un certain $B > 0$.

Réduction à la perte linéaire. Supposons la fonction $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$ convexe en son premier argument. Si $\widehat{y}_t = \int_\Theta \widehat{y}_{\theta,t} v_t(\mathrm{d}\theta)$, alors l'inégalité de Jensen implique que

$$\ell(\widehat{y}_t, y_t) \leqslant \int_\Theta \ell(\widehat{y}_{\theta,t}, y_t) v_t(\mathrm{d}\theta) = \langle v_t, h_t \rangle ,$$

où $h_t : \Theta \to [0, B]$ est la fonction définie par $h_t(\theta) = \ell(\widehat{y}_{\theta,t}, y_t)$ et où $\langle v, h \rangle := \int_\Theta h \mathrm{d}v$. À l'instar du cas exp-concave (Section 1.2.4), l'agrégation d'experts se ramène alors au problème d'apprentissage séquentiel avec $\mathcal{G} = \mathcal{P}(\Theta)$, $\mathcal{Z}$ l'espace des fonctions mesurables $\Theta \to [0, B]$, ℓ la fonction de perte *linéaire* $(v, h) \mapsto \langle v, h \rangle$ et $\mathcal{F} = \{\delta_\theta : \theta \in \Theta\}$. Ce problème d'apprentissage est appelé *problème de Hedge* (Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), et coïncide avec l'optimisation linéaire en ligne sur le simplexe $\mathcal{P}(\Theta)$.

Si $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$ n'est pas convexe, il est possible de se ramener à ce cas en considérant des prédictions randomisées (Remarque 1.2). En effet, si $\widehat{y}_t = \widehat{y}_{\widehat{\theta}_t,t}$, où $\widehat{\theta}_t \sim v_t$ (conditionnellement aux tirages précédents), alors $\mathbb{E}[\ell(\widehat{y}_t, y_t)] = \langle v_t, h_t \rangle$ dès lors que y_t ne dépend pas de $\widehat{\theta}_t$ (mais peut dépendre de v_t ainsi que de $\widehat{\theta}_1, \ldots, \widehat{\theta}_{t-1}$), ce qui est notamment le cas si la suite $y_1, \ldots, y_n$ est déterministe. La réduction au problème de Hedge permet donc de contrôler le regret moyen (sur le tirage de $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$), et des inégalités de concentration sur les martingales impliquent que le regret est proche de son espérance avec forte probabilité (Cesa-Bianchi and Lugosi, 2006).

Regret de l'APE pour le problème de Hedge. Avec les notations précédentes, l'APE de paramètre $\eta > 0$ pour l'agrégation d'experts (1.31) coïncide avec la stratégie suivante pour le problème linéarisé :

$$v_t = \frac{\exp\big(-\eta \sum_{s=1}^{t-1} h_s\big)}{\int_\Theta \exp\big(-\eta \sum_{s=1}^{t-1} h_s\big) \mathrm{d}\pi} \cdot \pi \,. \tag{1.39}$$

Cet algorithme admet la garantie de regret suivante (Freund and Schapire, 1997; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006) :

Proposition 1.5 (Regret de l'algorithme Hedge)**.** *Pour tous $\eta, B > 0$ et $n \geqslant 1$, toute suite de fonctions $h_1, \dots, h_n : \Theta \to [0, B]$ et toute loi $\rho \in \mathcal{P}(\Theta)$, l'algorithme Hedge (1.39) de paramètre $\eta > 0$ satisfait :*

$$\sum_{t=1}^n \langle v_t, h_t \rangle - \sum_{t=1}^n \langle \rho, h_t \rangle \leqslant \frac{\mathrm{KL}(\rho, \pi)}{\eta} + \frac{\eta B^2 n}{8} \,. \tag{1.40}$$

En particulier, pour tout $K > 0$, le choix de $\eta = B^{-1}\sqrt{8K/n}$ conduit à une borne de regret de $B\sqrt{nK/2}$ par rapport à la classe $\mathcal{F}_K := \{\rho \in \mathcal{P}(\Theta) : \mathrm{KL}(\rho, \pi) \leqslant K\}$. Si $\Theta = \{1, \dots, M\}$ est fini et π est la mesure uniforme sur Θ, le choix $\eta = B^{-1}\sqrt{8(\log M)/n}$ donne :

$$\sum_{t=1}^n \langle v_t, h_t \rangle - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^n h_{i,t} \leqslant B\sqrt{\frac{n \log M}{2}} \,. \tag{1.41}$$

Proof. Pour $t = 1, \dots, n$, posons $h'_t := \eta \cdot h_t$. L'inégalité de Hoeffding (Lemme 1.2, Section 1.6) avec $\lambda = -\eta B$, appliquée à la variable $h_t(\widehat{\theta}_t)/B$, $\widehat{\theta}_t \sim v_t$, à valeurs dans $[0, 1]$, donne

$$\langle v_t, h_t \rangle \leqslant -\frac{1}{\eta} \log \int_\Theta e^{-\eta\, h_t(\theta)} v_t(\mathrm{d}\theta) + \frac{\eta B^2}{8} = \frac{1}{\eta} \ell_{\mathrm{mix}}(v_t, h'_t) + \frac{\eta B^2}{8} \,.$$

La borne de regret (1.40) s'obtient alors en sommant sur $t = 1, \dots, n$, en notant que l'algorithme Hedge (1.39) coïncide avec l'APE sur la suite $h'_1, \dots, h'_n$, et en utilisant la Proposition 1.4. Les assertions restantes s'en déduisent en optimisant le choix de η et en prenant $\rho = \delta_i$ dans le cas où $\Theta = \{1, \dots, M\}$. □

La borne (1.41) correspond à un regret moyen d'au plus $O(\sqrt{(\log M)/n})$. On retrouve ainsi la même vitesse (lente) qu'ERM pour l'apprentissage statistique sur des classes finies et avec perte bornée (Section 1.1.4). La borne de regret de $O(\sqrt{n \log M})$ s'avère optimale au sens minimax dans le cas de classes finies : pour tout algorithme pour le problème de Hedge, il existe une suite de fonctions $h_t : \{1, \dots, M\} \to [0, 1]$, $1 \leqslant t \leqslant n$, pour laquelle le regret de l'algorithme est d'au moins $\Theta(\sqrt{n \log M})$ (Cesa-Bianchi and Lugosi, 2006). Cela découle aussi du fait que, par conversion online-to-batch (Proposition 1.3), le regret minimax (moyen) est supérieur à l'excès de risque minimax pour la classification avec classes finies à M classifieurs, qui est de $\Theta(\sqrt{(\log M)/n})$.

1.2.6 Algorithmes adaptatifs pour le problème de Hedge

Dans cette section, nous considérons le problème de Hedge de la section précédente, en supposant de plus que $\Theta = \{1, \dots, M\}$ et $B = 1$. Ainsi, les fonctions h_t ($t = 1, \dots, M$) s'identifient

à des vecteurs de perte dans $[0,1]^M$, tandis que les lois de probabilités $v_t \in \mathcal{P}(\Theta)$ s'identifient aux vecteurs $(v_{i,t})_{1\leqslant i\leqslant M} \in \mathbf{R}_+^M$ tels que $\sum_{i=1}^M v_{i,t} = 1$. Pour $t = 1, \dots, n$, nous notons

$$H_{i,t} := \sum_{s=1}^{t} h_{i,s}\,, \qquad \widehat{H}_t := \sum_{s=1}^{t} \langle v_t, h_t\rangle$$

les pertes cumulées au temps t de l'expert i et de l'algorithme, respectivement. Nous considérons également des bornes de regret uniformes par rapport aux vecteurs $\rho \in \mathcal{P}(\Theta)$, c'est-à-dire par rapport aux *experts* $i = 1, \dots, M$, telles que la borne (1.41).

Borne minimax, horizon de temps. Par la Proposition 1.5, l'algorithme Hedge avec loi à priori π uniforme, soit

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

pour tous $t = 1, \dots, M$ et $1 \leqslant i \leqslant M$, et de paramètre $\eta = c\sqrt{(\log M)/n}$ (où $c, C > 0$ désignent des constantes numériques), admet la borne de regret

$$\mathsf{Reg}_n := \widehat{H}_t - \min_{1\leqslant i\leqslant M} H_{i,t} = \sum_{t=1}^{n} \langle v_t, h_t\rangle - \min_{1\leqslant i\leqslant M} \sum_{t=1}^{n} h_{i,t} \leqslant C\sqrt{n\log M}\,.$$

(Freund and Schapire, 1997; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006), ce qui correspond au regret minimax (Cesa-Bianchi and Lugosi, 2006). Notons que le choix du paramètre $\eta \asymp \sqrt{(\log M)/n}$ dépend de l'horizon de temps n considéré. Il peut être souhaitable d'obtenir la borne de regret minimax de $O(\sqrt{n\log M})$ *simultanément* pour tout $n \geqslant 1$, pour un même algorithme.

Une façon générique d'obtenir une telle garantie est la *technique du doublement* (*doubling trick* en anglais), qui consiste à utiliser l'algorithme Hedge de paramètre $\eta_p = c\sqrt{(\log M)/2^p}$ sur les intervalles de temps "géométriques" $\{2^p, \dots, 2^{p+1} - 1\}$ (pour $p \geqslant 0$), en réinitialisant les poids à la fin de chaque intervalle (Cesa-Bianchi et al., 1997; Cesa-Bianchi and Lugosi, 2006). On obtient alors, pour tout $n \geqslant 1$, en notant p l'entier tel que $2^p \leqslant n < 2^{p+1}$, un regret d'au plus

$$C\sum_{k=0}^{p} \sqrt{2^k \log M} \leqslant \frac{C}{1-1/\sqrt{2}}\sqrt{2^p \log M} = O(\sqrt{n\log M})\,.$$

Une autre façon d'obtenir cette garantie est d'utiliser un paramètre d'apprentissage η_t variable, c'est-à-dire de choisir à l'instant $t \geqslant 1$,

$$v_{i,t} = \frac{e^{-\eta_t H_{i,t-1}}}{\sum_{j=1}^M e^{-\eta_t H_{j,t-1}}}\,, \tag{1.42}$$

où η_t dépend de t et éventuellement de $h_1, \dots, h_{t-1} \in [0,1]^M$. En effet, pour toute suite décroissante $\eta_1, \eta_2, \dots$ de paramètres, le regret de la stratégie (1.42) satisfait :

$$\mathsf{Reg}_n \leqslant \frac{\log M}{\eta_n} + \frac{1}{8}\sum_{t=1}^{n} \eta_t$$

(Chernov and Zhdanov, 2010). En particulier, le choix $\eta_t = c\sqrt{(\log M)/t}$ $(t \geqslant 1)$ conduit à une borne de regret en $C\sqrt{n\log M}$ pour tout $n \geqslant 1$.

Dépendance en la complexité : bornes de "quantiles". Considérons le cas où le nombre d'experts M est élevé. Il est alors souhaitable d'obtenir des bornes avec une faible dépendance en M, quitte à considérer une classe de comparaison moins "complexe" que l'ensemble des experts $\{1,\dots,M\}$. La Proposition 1.5 affirme que l'algorithme Hedge de paramètre $\eta = c\sqrt{K/n}$ admet une borne de regret d'au plus $C\sqrt{K/n}$ par rapport à tout mélange des experts selon une loi ρ sur $\{1,\dots,M\}$ telle que $\mathrm{KL}(\rho,\pi)$. Comme montré dans la discussion suivant le Corollaire 1.2, en choisissant $K = \log(1/\varepsilon)$ pour $\varepsilon \in (0,1)$, ceci implique une borne de regret de $\sqrt{\log(1/\varepsilon)/n}$ par rapport au $\lceil \varepsilon M \rceil$-meilleur expert, indépendamment de M.

Cependant, cette borne est atteinte en choisissant η en fonction de K (ou, de manière équivalente, en fonction de ε), et n'est donc valable que pour cette valeur de ε. À l'inverse, le Corollaire 1.2 montre que l'APE atteint dans le cas exp-concave la borne de regret $C\log(1/\varepsilon)/n$ simultanément pour tout $\varepsilon \in (0,1)$, par un choix de η indépendant pas de ε. Dans le cas de pertes linéaires considéré ici, il découle de la Proposition 1.5 que le choix de $\eta = c/\sqrt{n}$ conduit à une borne de regret de $C \cdot \mathrm{KL}(\rho,\pi)/\sqrt{n}$ pour tout ρ, c'est-à-dire de $C \cdot \log(1/\varepsilon)/\sqrt{n}$. Cette borne admet une dépendance sous-optimale en ε. Un axe de recherche consiste à obtenir des algorithmes adaptatifs à la "complexité" ε, c'est-à-dire un regret de $C\sqrt{\log(1/\varepsilon)/n}$ pour tout $\varepsilon > 0$ (Chaudhuri et al., 2009; Chernov and Vovk, 2010; Luo and Schapire, 2014, 2015; Koolen and van Erven, 2015) ; ces bornes sont appelées "bornes de quantiles", car elles contrôlent le regret par rapport aux quantiles de niveau $\varepsilon \in (0,1)$ des experts (ordonnés par leur perte).

Au-delà du minimax : algorithmes adaptatifs. L'attrait des bornes minimax est leur robustesse : elles sont en effet valables pour toute suite $h_1,\dots,h_n \in [0,1]^M$ de vecteurs de pertes, et en particulier sans hypothèse de stochasticité. Cependant, la vitesse en $O(\sqrt{n\log M})$, bien qu'optimale dans le pire des cas, peut s'avérer trop pessimiste dans certaines situations. Une littérature importante est dédiée à la conception d'algorithmes *adaptatifs*, qui combinent le regret minimax en $O(\sqrt{n\log M})$ avec des garanties améliorées lorsque la suite $h_1,\dots,h_n$ présente certaines régularités (Cesa-Bianchi et al., 1997; Auer et al., 2002; Cesa-Bianchi et al., 2007; de Rooij et al., 2014; Gaillard et al., 2014; Koolen et al., 2014; Sani et al., 2014; Koolen and van Erven, 2015; Luo and Schapire, 2015; Wintenberger, 2017).

Une première situation où il est possible d'obtenir des garanties améliorées de regret correspond au cas où la perte moyenne $H_n^*/n = \min_{1\leqslant i\leqslant n} H_{i,n}/n$ du meilleur expert est faible. Les bornes de regret dites du *premier ordre* (Cesa-Bianchi et al., 1997; Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006), de la forme $\mathsf{Reg}_n \leqslant C(\sqrt{H_n^*\log M} + \log M)$, permettent d'exploiter cette régularité : elles impliquent un regret d'au plus $O(\sqrt{n\log M})$ dans le pire des cas, mais améliorent cette garantie lorsque $H_n^* \ll n$. Cette garantie est satisfaite par l'algorithme Hedge (1.42) avec $\eta_t = \eta = c\sqrt{(\log M)/(1\vee H_n^*)}$ lorsque H_n^* est connu a priori (Cesa-Bianchi and Lugosi, 2006) ; cette quantité étant généralement inconnue, il est possible d'obtenir la garantie précédente pour tout $H_n^* \in [0,n]$ par la technique du doublement (Cesa-Bianchi et al., 1997), ou en prenant $\eta_t = c\sqrt{(\log M)/(1\vee H_{t-1}^*)}$ (Auer et al., 2002).

La borne du premier ordre permet d'obtenir un regret constant en $O(\log M)$ dans le cas où $H_n^* = O(1)$, c'est-à-dire lorsque le meilleur expert admet une perte négligeable. Cependant, lorsque le meilleur expert admet une perte moyenne non négligeable (ce qui est typiquement le cas pour les problèmes d'apprentissage), la borne du premier ordre n'améliore pas la vitesse du regret, mais au mieux la constante du terme dominant. Cette garantie apporte donc une adaptativité limitée aux propriétés de la suite $h_1,\dots,h_n$: elle n'exploite pas d'autres formes de régularité qui devraient permettre d'obtenir un regret amélioré, par exemple le fait que

l'un des experts prédise mieux que les autres (sans pour autant avoir une perte négligeable). Un second type de garantie, qui raffine la borne du premier ordre, est donné par les bornes du *second ordre* (Cesa-Bianchi et al., 2007; de Rooij et al., 2014; Gaillard et al., 2014; Koolen and van Erven, 2015). Celles-ci dépendent de quantités du second ordre, comme la somme au cours du temps des variances des pertes $(h_{i,t})_{1\leqslant i\leqslant M}$ selon la loi de probabilité $v_t = (v_{i,t})_{1\leqslant i\leqslant M}$ (Cesa-Bianchi et al., 2007; de Rooij et al., 2014), ou des quantités liées (Gaillard et al., 2014; Koolen and van Erven, 2015). Ces bornes sont atteintes par des algorithmes plus sophistiqués, qui calibrent η_t en fonction de ces quantités du second ordre (Cesa-Bianchi et al., 2007; de Rooij et al., 2014), ou reposent sur d'autres choix de poids que les poids exponentiels (1.42), dans le cas de Gaillard et al. (2014); Wintenberger (2017); Koolen and van Erven (2015). Le comportement typique de ces algorithmes est le suivant :

- si la suite $h_1, \dots, h_n$ est "difficile", c'est-à-dire choisie de manière adverse comme dans le pire des cas, alors l'algorithme adoptera un comportement conservateur, similaire à celui de l'algorithme Hedge : les poids ne seront pas excessivement concentrés, en raison d'un niveau non négligeable de régularisation. Dans le cas d'algorithmes à poids exponentiels de type (1.42), cela signifie que $\eta_t \lesssim \sqrt{(\log M)/t}$;
- si au contraire la suite $h_1, \dots, h_n$ est "favorable", c'est-à-dire présente certaines régularités exploitées par l'algorithme, alors celui-ci adoptera un comportement plus "agressif", et aura tendance à éliminer plus rapidement les experts sous-optimaux et à se concentrer sur le meilleur expert. Dans le cas d'un algorithme à poids exponentiels (1.42), cela revient à choisir un paramètre η_t plus élevé, par exemple $\eta_t \gtrsim c$, et donc à se rapprocher de la minimisation du risque empirique.

Les bornes du second ordre impliquent celles du premier ordre (qui impliquent elles-mêmes la borne minimax "d'ordre 0" en $O(\sqrt{n \log M})$), mais les améliorent sensiblement dans certains cas. Par exemple, si l'un des experts domine les autres, alors les poids de l'algorithme vont avoir tendance à se concentrer sur le meilleur expert, ce qui réduira la variance des pertes ; ceci conduit en retour l'algorithme à éliminer les experts sous-optimaux de manière plus agressive, ce qui réduit la variance des pertes d'autant plus vite.

Un type naturel de régularité considéré dans la littérature est le cas *stochastique*, où les vecteurs de pertes $h_1, \dots, h_n \in [0,1]^M$ sont des variables aléatoires i.i.d.[6] (van Erven et al., 2011; Gaillard et al., 2014; Luo and Schapire, 2015). Cela couvre notamment le cas où $h_{i,t} = \ell(f_i, Z_t)$ pour $1 \leqslant i \leqslant M$, avec $f_1, \dots, f_M \in \mathcal{F}$ et $Z_1, \dots, Z_n$ sont des variables i.i.d. de même loi P. Supposons dans ce cas que le meilleur expert $i^* = \arg\min_{1\leqslant i\leqslant n} \mathbb{E}[h_{i,1}]$ est unique. On note alors

$$\Delta := \arg\min_{i \neq i^*} \mathbb{E}[h_{i,t} - h_{i^*,t}] > 0\,. \tag{1.43}$$

Le paramètre Δ, qui correspond à l'écart entre le meilleur expert et les autres, est une mesure de la difficulté du problème : si Δ est suffisamment élevé, alors le meilleur expert aura tendance à dominer rapidement, et sera donc clairement distinguable. Dans ce cas, Gaillard et al. (2014) montre qu'une borne de regret du second ordre implique un regret d'au plus $O((\log M)/\Delta)$; cette garantie améliore la borne minimax de $O(\sqrt{n \log M})$ dès lors que $\Delta \gtrsim \sqrt{(\log M)/n}$, et correspond à un regret constant. La même borne du second ordre (combinée à une garantie de type "quantile") est également atteinte par une procédure de Koolen and van Erven (2015) ;

[6] Notons que l'on suppose ici l'indépendance des pertes $h_{i,t}$ entre les différents instants $t \geqslant 1$, mais pas nécessairement entre les différents experts $1 \leqslant i \leqslant M$ à un même instant.

par ailleurs, un algorithme proposé par Luo and Schapire (2015), qui admet une garantie de regret plus faible qu'une garantie de second ordre, combine également le regret minimax en $O(\sqrt{n \log M})$ avec la garantie améliorée de $O((\log M)/\Delta)$.

Plus généralement, Koolen et al. (2016) ont montré que, si les pertes $(h_{i,t})_{1 \leqslant i \leqslant M}$ satisfont la condition de Bernstein de paramètres (β, B), avec $B > 0$ et $\beta \in [0,1]$ (Définition 1.2), c'est-à-dire $\mathbb{E}[(h_{i,t} - h_{i^*,t})^2] \leqslant B \cdot \mathbb{E}[h_{i,t} - h_{i^*,t}]^\beta$ pour tout i, alors le regret par rapport à i^* satisfait

$$\mathbb{E}[\widehat{H}_n - H_{i^*,n}] \leqslant C\big((B \log M)^{\frac{1}{2-\beta}} n^{\frac{1-\beta}{2-\beta}} + \log M\big)\,, \tag{1.44}$$

avec une borne correspondante en forte probabilité (voir aussi la Proposition 4.4 du Chapitre 4 pour une preuve élémentaire de cette borne en espérance, avec dépendance en B).

1.2.7 Optimalité de l'algorithme Hedge dans le cas stochastique (Chapitre 4)

Des résultats de la section précédente, il découle que :

- l'algorithme de Hedge (1.42), avec paramètre d'apprentissage constant $\eta_t = \sqrt{(\log M)/n}$ ou variable $\eta_t = \sqrt{(\log M)/t}$, ou avec la technique du doublement, admet un regret d'au plus $O(\sqrt{n \log M})$, qui correspond au regret minimax ;
- des stratégies plus sophistiquées admettent des garanties adaptatives plus fines (comme les bornes de second ordre), qui impliquent une borne de regret améliorée en $O((\log M)/\Delta)$ dans le cas stochastique (avec Δ défini par (1.43)), en plus de la borne de $O(\sqrt{n \log M})$ dans le pire des cas. Ces algorithmes adaptatifs se comportent typiquement comme Hedge avec un paramètre d'apprentissage conservateur $\eta_t \asymp \sqrt{(\log M)/t}$ pour une suite "adverse", mais se montrent plus agressifs (proches d'ERM) sur des suites favorables (de Rooij et al., 2014).

Les résultats précédents portent sur des *bornes supérieures* sur le regret des différents algorithmes considérés. De telles garanties ne permettent cependant pas à elles seules de conclure quant aux performances et avantages respectifs des différents algorithmes, le regret pouvant être plus faible que ce que ces bornes supérieures suggèrent.

Dans le Chapitre 4, nous étudions le comportement de l'algorithme Hedge dans le cas stochastique. Tout d'abord, nous montrons que Hedge avec un paramètre d'apprentissage décroissant $\eta_t = c\sqrt{(\log M)/t}$, calibré pour le pire des cas, est également adaptatif au cas stochastique des données, où il atteint la même garantie de $O((\log M)/\Delta)$.

Théorème 1.5 (Théorème 4.1, Chapitre 4)**.** *Si les vecteurs de perte $h_1, \dots, h_n$ sont i.i.d., alors en notant Δ le paramètre (1.43), l'algorithme Hedge avec $\eta_t = 2\sqrt{(\log M)/t}$ admet un regret*[7] *d'au plus*

$$\mathbb{E}[\widehat{H}_n - H_{i^*,n}] \leqslant \frac{4 \log M + 25}{\Delta}\,. \tag{1.45}$$

Une borne similaire est également valable sous des hypothèses plus générales, et en forte probabilité (Corollaire 4.1). En outre, la borne de regret (1.45) admet la dépendance optimale en M et Δ : pour tous $M \geqslant 2, \Delta \in (0,1)$ et toute stratégie, il existe une loi de h_1 pour laquelle le regret de cette stratégie est d'au moins $O((\log M)/\Delta)$ (Proposition 4.2). L'algorithme de

[7] La quantité apparaissant dans la borne (1.45) n'est pas tout-à-fait le regret, mais le regret par rapport à i^*. Ces quantités sont toutefois proches, et il est possible d'obtenir des bornes de regret (voir la Remarque 4.3 et le Corollaire 4.1 du Chapitre 4).

Hedge avec le paramètre $\eta_t = c\sqrt{(\log M)/t}$ calibré pour le pire des cas est donc adaptatif à la difficulté du problème, pour tout $\Delta \in (0, 1)$.

À l'inverse, nous établissons que cette adaptativité au cas stochastique n'est pas partagée par les variantes proches de Hedge, également minimax, obtenues avec le paramètre constant $\eta_t = c\sqrt{(\log M)/n}$ ou par la technique du doublement. En effet, la Proposition 4.3 montre que ces algorithmes exhibent un regret en $\Theta(\sqrt{n \log M})$ du pire des cas même pour des problème stochastiques "faciles" avec $\Delta \simeq 1$. En particulier, le choix d'un paramètre d'apprentissage variable conduit à un algorithme plus adaptatif que la technique du doublement, laquelle n'est adéquate que dans le pire des cas. De plus, le paramètre d'apprentissage variable est préférable au paramètre constant, même lorsque l'horizon de temps n est connu.

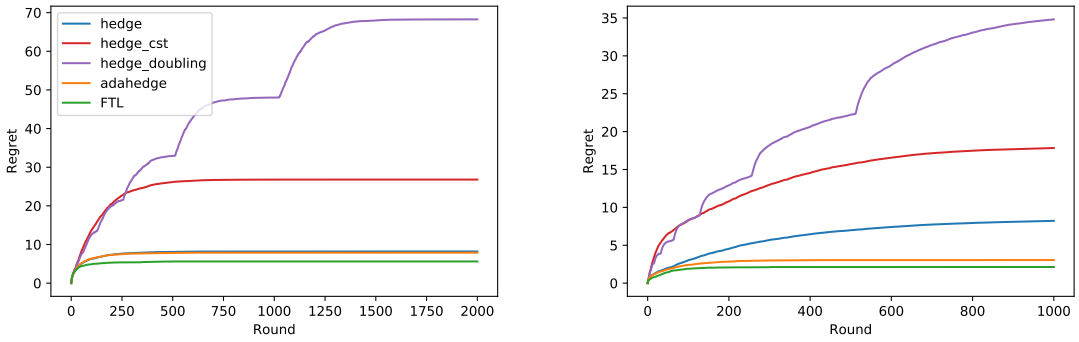
Le Théorème 1.5 montre que l'algorithme de Hedge atteint la même borne de regret dans le cas stochastique que les algorithmes adaptatifs mentionnés dans la Section 1.2.6 (Gaillard et al., 2014; Luo and Schapire, 2015). Il est donc naturel de se demander si ceux-ci apportent un gain dans le cas stochastique. Il s'avère que les algorithmes du second ordre ont bien un avantage sur la variante de Hedge avec paramètre variable. Pour le voir, il est nécessaire de considérer une notion de régularité plus fine que l'écart Δ, à savoir la condition de Bernstein (Définition 1.2). En effet, la borne (1.44) montre qu'une garantie du second ordre implique un regret d'au plus $O(B \log M)$ sur des pertes stochastiques satisfaisant la condition de Bernstein $(1, B)$ avec $B \geqslant 1$. Notons que si $\Delta > 0$, alors cette condition est satisfaite pour $B \leqslant 1/\Delta$. Cependant, cette condition est sensiblement moins restrictive, et peut être satisfaite avec $B = O(1)$ même pour des valeurs arbitrairement faibles de Δ ; nous renvoyons à l'Exemple 4.2 Chapitre 4 pour davantage de détails.

À l'inverse, l'algorithme de Hedge avec paramètre $\eta_t = c\sqrt{(\log M)/t}$ n'est pas adaptatif à la condition de Bernstein. En effet, le Théorème 4.3 implique qu'il existe une loi satisfaisant la condition de Bernstein avec $B = O(1)$, mais pour laquelle cet algorithme admet un regret de $\Theta(\sqrt{n \log M})$. Plus généralement, il s'avère que l'algorithme Hedge avec $\eta_t = c\sqrt{(\log M)/t}$ ne peut pas s'adapter à d'autres régularités que l'écart Δ : quelle que soit la loi de h_1, le regret de cet algorithme pour $n \geqslant C/\Delta^2$ est d'au moins

$$\frac{c}{(\log M)^2 \Delta} .$$

(Théorème 4.4). Ceci caractérise (à un facteur $\log^3 M$ près, en fonction du nombre d'experts presque optimaux) le regret de cet algorithme sur *tout* problème stochastique.

L'avantage d'un paramètre d'apprentissage η_t adaptatif, plus élevé que le paramètre en $c\sqrt{(\log M)/t}$ du pire des cas, sur certains problèmes stochastiques peut se comprendre de la façon suivante. Considérons une loi de h_t avec B peu élevé, mais Δ faible (donc $1/\Delta$ élevé). Le paramètre d'apprentissage en $c\sqrt{(\log M)/t}$ est suffisamment élevé pour éliminer les "mauvais" experts i (tels que $\Delta_i := \mathbb{E}[h_{i,t} - h_{i^*,t}]$ est suffisamment grand) après un nombre d'étapes optimal (en $O((\log M)/\Delta_i^2)$). Cependant, une fois ces experts écartés, les experts presque optimaux (avec $\Delta_i \simeq \Delta$) ne seront éliminés que très tard (après $O((\log M)/\Delta^2)$ étapes). En revanche, la condition de Bernstein implique que les pertes de ces experts sont fortement corrélées à celles de l'expert optimal i^* (voir la Section 1.1.4), de sorte que le niveau de bruit dans leurs pertes relatives $h_{i,t} - h_{i^*,t}$ est faible. Les algorithmes du second ordre utiliseront alors un paramètre d'apprentissage plus élevé, de sorte que ces experts presque optimaux seront éliminés plus tôt et contribueront donc moins au regret.



(a) (b)

Figure 1.1: Regret d'algorithmes d'agrégation d'experts sur deux problèmes stochastiques. Sont évalués les algorithmes Hedge avec paramètre décroissant (`hedge`), constant (`hedge_cst`) et technique du doublement (`hedge_doubling`), l'algorithme adaptatif du second ordre `adahedge` (de Rooij et al., 2014), et la minimisation du risque empirique (`FTL`). (a) Problème stochastique avec un écart ($M = 20$, $\Delta = 0.1$) ; (b) Problème stochastique avec Δ faible, mais satisfaisant la condition de Bernstein ($M = 10$, $\Delta = 0.04$, $B \leqslant 4$).

1.2.8 Prédiction à l'aide d'une classe croissante d'experts (Chapitre 5)

Dans le Chapitre 5, nous étudions une variante du problème de l'agrégation d'experts (Section 1.2.1), dans le cas où la classe d'experts $\Theta_t := \{1, \dots, M_t\}$ s'enrichit au cours du temps : à chaque instant $t \geqslant 1$, $m_t := M_t - M_{t-1} \geqslant 0$ nouveaux experts sont disponibles. Ce problème est motivé par des situations pratiques, où il peut être souhaitable d'incorporer de nouvelles sources de prédictions (comme de nouveaux algorithmes d'apprentissage ou des prédicteurs utilisant de nouvelles variables).

Dans ce contexte, nous considérons plusieurs notions de regret, qui diffèrent par le choix du comparateur : (1) chaque expert, depuis l'instant de son introduction, ce qui revient à se comparer aux suites d'experts ne transitionnant que vers de nouveaux experts ; (2) les suites arbitraires d'experts avec un nombre de changements (soit vers un nouvel expert, soit vers un expert déjà introduit) contrôlé ; (3) les suites d'experts à support "parcimonieux", qui prennent leurs valeurs dans un petit sous-ensemble de "bons" experts.

Dans chaque cas, en étendant les stratégies existantes dans le cas d'une classe fixe d'experts (Vovk, 1998; Herbster and Warmuth, 1998; Vovk, 1999; Bousquet and Warmuth, 2002; Koolen et al., 2012), nous proposons des algorithmes efficaces (avec une complexité linéaire en le temps et le nombre d'experts) admettant des garanties de regret optimales. Ces extensions au cas d'un nombre croissant d'experts utilisent la notion de *spécialistes*, c'est-à-dire d'experts pouvant s'abstenir à certains instants (Freund et al., 1997; Chernov and Vovk, 2009), des (reformulations génériques de) stratégies efficaces équivalentes à l'agrégation de *suites d'experts* avec pour loi a priori une chaîne de Markov (Herbster and Warmuth, 1998; Vovk, 1999; Koolen and de Rooij, 2013; Bousquet and Warmuth, 2002; Koolen et al., 2012), ainsi que la combinaison de ces techniques.