Un estimateur (presque) optimal dans le cas mal spécifié

Dans le Chapitre 7, nous introduisons une procédure générale pour l'estimation de densité (conditionnelle ou non), qui satisfait une borne générale d'excès de risque, valide dans le cas général mal spécifié. Dans cette section, nous décrivons brièvement cette procédure, ainsi que sa borne d'excès de risque, et des cas particuliers simples dans le cas non conditionnel. Dans les sections suivantes, nous étudierons cette procédure appliquée à deux modèles conditionnels classiques, à savoir le *modèle linéaire Gaussien* ainsi que le *modèle logistique*.

La procédure que nous introduisons, appelée *Sample Minmax Predictor* (SMP), est en fait valide pour l'apprentissage supervisé avec une fonction de perte générale. Elle apparaît naturellement comme la procédure minimisant une nouvelle borne d'excès de risque générale pour l'apprentissage supervisé. Nous reprenons les notations de la Section 1.1.2 ; en particulier, $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ désigne la fonction de perte, et $\mathcal{F}$ une classe de fonctions $\mathcal{X} \to \widehat{\mathcal{Y}}$. Pour $z = (x, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ et $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ un prédicteur, on note $\ell(g, z) := \ell(g(x), y)$.

Théorème 1.11 (Théorème 7.1, Chapitre 7)**.** *Soit $\widehat{g}_n$ un estimateur dépendant de l'échantillon i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$. Notons $Z_i = (X_i, Y_i)$, et pour tout $z \in \mathcal{Z}$*

$$\widehat{f}_n^{(z)} := \arg\min_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \ell(f, Z_i) + \ell(f, z) \right\}.$$

Alors, l'excès de risque $\mathcal{E}(\widehat{g}_n) := R(\widehat{g}_n) - \inf_{f \in \mathcal{F}} R(f)$ de $\widehat{g}_n$ satisfait :

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n)\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{g}_n(X), y) - \ell(\widehat{f}_n^{(X,y)}(X), y) \Big\} \Big] \tag{1.89}$$

où $Z = (X, Y) \sim P$ est indépendant de Z_1^n. De plus, la borne (1.89) est minimisée par le prédicteur

$$\widehat{g}_n^{\mathrm{SMP}}(x) = \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell(\widehat{f}_n^{(x,y)}(x), y) \Big\}, \tag{1.90}$$

que nous appelons SMP lorsqu'il est bien défini. Dans ce cas, la borne générale (1.89) s'écrit

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\inf_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell(\widehat{f}_n^{(X,y)}(X), y) \Big\} \Big]. \tag{1.91}$$

Le SMP admet également une variante régularisée, avec une borne d'excès de risque correspondante (voir l'énoncé exact du Théorème 7.1 du Chapitre 7). Dans le cas de la perte logarithmique, le SMP (1.90) et sa borne d'excès de risque (1.91) admettent une expression explicite.

Théorème 1.12 (Théorème 7.2, Chapitre 7)**.** *Dans le cas de l'estimation de densité conditionnelle, le SMP s'écrit*

$$\widehat{g}_n^{\mathrm{SMP}}(y|x) = \frac{\widehat{f}_n^{(x,y)}(y|x)}{\int_{\mathcal{Y}} \widehat{f}_n^{(x,y')}(y'|x) \mu(\mathrm{d}y')} \tag{1.92}$$

dès lors que le dénominateur de (1.92) est fini. De plus, son excès de risque est borné de la façon suivante :

$$\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})] \leqslant \mathbb{E}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_n^{(X,y)}(y|X) \mu(\mathrm{d}y) \Big) \Big]. \tag{1.93}$$

Notons que le SMP est en général un prédicteur *impropre*, tout comme les prédicteurs obtenus par (conversion online-to-batch de) mélange bayésien ou NML. Nous verrons que cet estimateur contourne les limitations inhérentes aux estimateurs propres (comme les approches fondées sur la conversion online-to-batch), et permet d'obtenir des bornes en $d/n + o(1/n)$ (contrairement à ces dernières).

L'expression (1.93) du SMP fait apparaître une intégrale comme constante de renormalisation. Pour les exemples que nous allons considérer, cette constante se calcule explicitement. Cependant, notons que, contrairement aux approches Bayésiennes où l'intégrale de la constante de renormalisation porte sur le paramètre $\theta \in \Theta$, celle-ci porte sur la réponse $y \in \mathcal{Y}$. Dans de nombreux exemples d'estimation de densité conditionnelle, l'espace des paramètres Θ est bien plus complexe que celui des sorties $\mathcal{Y}$. C'est notamment le cas pour le problème de la régression logistique, dont nous discuterons en Section 1.4.7, où $\mathcal{Y} = \{-1, 1\}$ tandis que $\Theta = \mathbf{R}^d$.

Exemple 1.16 (Modèle Gaussien, Proposition 7.2)**.** Dans le cas du modèle Gaussien $\mathcal{F} = \{\mathcal{N}(\theta, I_d) : \theta \in \mathbf{R}^d\}$, le SMP vaut $\widehat{g}_n^{\mathrm{SMP}} = \mathcal{N}(\bar{Z}_n, (1 + 1/n)^2 I_d)$, et sa borne d'excès de risque (1.93) vaut $d \log(1 + 1/n) \leqslant d/n$, quelle que soit la loi de Z telle que $\mathbb{E}[\|Z\|] < +\infty$.

À l'inverse, pour l'EMV et plus généralement tout estimateur propre $f_{\widehat{\theta}_n}$, une dépendance en la quantité $d_{\mathrm{eff}} = \mathbb{E}[\|Z\|^2]$ est inévitable : pour tout $t > 0$, il est impossible d'obtenir une borne uniformément meilleure que $t/(2n)$ sur la classe des lois de Z telles que $\mathbb{E}[\|Z\|^2] = t$ (Section 7.3.2).

En outre, le regret minimax (1.85) par rapport à la classe $\mathcal{F}$ est infini (Exemple 1.15), il n'est donc pas possible d'obtenir de garantie d'excès de risque uniforme sur $\mathcal{F}$ par conversion online-to-batch.

Exemple 1.17 (Modèle multinomial, Proposition 7.1)**.** Considérons le cas où $\mathcal{Z}$ est fini, de cardinal d, et considérons le modèle multinomial $\mathcal{F} = \mathcal{P}(\mathcal{Z})$ (qui est toujours bien spécifié). Dans ce cas, le SMP correspond à l'*estimateur de Laplace*, donné par $\widehat{g}_n^{\mathrm{SMP}}(z) = (N_n(z) + 1)/(n + d)$, où $d = |\mathcal{Z}|$ et où $N_n(z)$ est le nombre d'occurrences de z parmi $Z_1, \ldots, Z_n$. De plus, sa borne d'excès de risque (1.93) est de $\log[(n + d)/(n + 1)] \leqslant (d - 1)/n$.

L'excès de risque de l'EMV est quant à lui infini avec probabilité positive (Section 7.3.1), donc aussi en espérance. Enfin, le regret minimax étant d'ordre $(d - 1)(\log n)/2 + O(1)$ (le nombre de paramètres est ici $d-1$), la conversion online-to-batch ne peut fournir qu'une borne en $\Theta(d \log(n)/n)$.

1.4.5 Application au modèle linéaire Gaussien (Chapitre 7 et Section 8.1)

Considérons à présent les espaces $\mathcal{X} = \mathbf{R}^d$ et $\mathcal{Y} = \mathbf{R}$, $\mu = (2\pi)^{-1/2}\mathrm{d}y$ et la famille de densités conditionnelles $\mathcal{F} = \{f_\beta : \beta \in \mathbf{R}^d\}$, où $f_\beta(\cdot|x) = \mathcal{N}(\langle \beta, x\rangle, 1)$. Ainsi, pour tous $\beta \in \mathbf{R}^d$ et $(x, y) \in \mathbf{R}^d \times \mathbf{R}$,

$$\ell(\beta, (x, y)) = \frac{1}{2}(\langle \beta, x\rangle - y)^2 .$$

Lorsque l'on se restreint aux prédicteurs propres, de la forme $\widehat{f}_n = f_{\widehat{\beta}_n}$, le problème est donc équivalent à celui des moindres carrés mentionné à la Section 1.3 et étudié dans le Chapitre 6. Cependant, le problème est de nature différente, puisqu'il s'agit d'effectuer une prédiction *probabiliste* de la réponse, c'est-à-dire d'estimer la loi conditionnelle de Y sachant X et non son espérance conditionnelle (Exemple 1.5). La possibilité d'utiliser des estimateurs impropres

permet à nouveau d'obtenir des garanties améliorées pour ce problème, en particulier dans le cas mal spécifié.

Borne d'excès de risque uniforme. Commençons par considérer le SMP (1.92) appliqué à la classe $\mathcal{F}$, ainsi que la borne d'excès de risque correspondante.

Théorème 1.13 (Théorème 7.4, Chapitre 7)**.** *Dans le cas du modèle linéaire Gaussien, le SMP vaut* $\widehat{g}_n^{\mathrm{SMP}}(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle, (1 + \langle (n\widehat{\Sigma}_n)^{-1}x, x\rangle)^2)$*. De plus, si* P_X *est non dégénérée (au sens discuté en Section 1.3.3) et si* $\mathbb{E}[Y^2] < +\infty$*, sa borne d'excès de risque* (1.93) *vaut*

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \mathbb{E}\big[\log(1 + \langle (n\widehat{\Sigma}_n)^{-1}X, X\rangle)\big] \leqslant \mathbb{E}\Big[\log\Big(1 + \frac{1}{n}\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\Big)\Big]\,. \tag{1.94}$$

De plus, dans le cas bien spécifié, on a $\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})] = \mathbb{E}\big[\log(1+\langle (n\widehat{\Sigma}_n)^{-1}X, X\rangle)\big] - d/(2(n+1))$.

La borne (1.94) montre que l'excès de risque $\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})]$ est d'au plus $\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)]/n$, c'est-à-dire au plus deux fois celui de l'EMV dans le cas bien spécifié[15]. Par le Théorème 1.9, il en découle que sous les Hypothèses 1.1 et 1.2 sur la loi P_X, l'excès de risque du SMP vérifie

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \frac{d}{n}\Big(1 + 8C'\frac{\kappa d}{n}\Big) = \frac{d}{n} + O\Big(\Big(\frac{d}{n}\Big)^2\Big)\,,$$

uniformément sur toutes les lois de Y sachant X telles que $\mathbb{E}[Y^2] < +\infty$ (Corollaire 7.1, Chapitre 7). À l'inverse, le risque de l'EMV dans le cas mal spécifié est de

$$\frac{1}{2n}\mathbb{E}\big[(Y - \langle \beta^*, X\rangle)^2 \|\Sigma^{-1/2}X\|^2\big] + O\Big(\Big(\frac{d}{n}\Big)^{3/2}\Big)$$

sous des hypothèses convenables sur P_X (Proposition 1.6). Cette vitesse dépend de l'erreur d'approximation du modèle linéaire, soit $\mathbb{E}[Y|X] - \langle \beta^*, X\rangle$, ainsi que de la variance conditionnelle $\mathrm{Var}(Y|X)$, puisque $\mathbb{E}[(Y - \langle \beta^*, X\rangle)^2|X] = (\mathbb{E}[Y|X] - \langle \beta^*, X\rangle)^2 + \mathrm{Var}(Y|X)$. Plus généralement, sur la classe des lois de Y sachant X telles que $\mathbb{E}[(Y - \langle \beta^*, X\rangle)^2|X] \leqslant \sigma^2$, l'excès de risque minimax parmi les estimateurs *propres* est d'au moins $\sigma^2 d/(2n)$ (par le Théorème 1.8 de la Section 1.3.2 sur la régression linéaire). Cela tient au fait que le SMP quantifie mieux l'incertitude sur la valeur de Y sachant X que tout estimateur propre ; le SMP exploite aussi implicitement la "courbure" (mélangeabilité) globale de la perte logarithmique, qui peut être nettement supérieure à celle de la perte restreinte au modèle.

Remarque 1.6 (Cas bien spécifié). Il est possible de montrer que la première borne de (1.94) vaut précisément le double de l'excès de risque minimax dans le cas *bien spécifié* (voir la Section 8.1 ainsi que la fin de cette section). Ainsi, la performance du SMP dans le cas mal spécifié est proche de la performance optimale atteignable même dans le cas bien spécifié, quelle que soit la loi des variables X. Cela montre notamment que l'excès de risque minimax dans le cas mal spécifié vaut au plus deux fois celui du cas bien spécifié.

Remarque 1.7 (Lien avec le levier). La première borne de (1.94) peut s'exprimer en fonction de la loi du levier statistique $\widehat{\ell}_{n+1}$ (voir la Section 1.3.2) d'un point X_{n+1} parmi $X_1, \ldots, X_{n+1}$:

$$\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})] \leqslant -\mathbb{E}[\log(1 - \widehat{\ell}_{n+1})]\,.$$

[15] En raison du facteur 1/2 devant la perte quadratique, le risque de $\widehat{\beta}_n^{\mathrm{LS}}$ dans le cas bien spécifié où $Y|X \sim \mathcal{N}(\langle \beta^*, X\rangle, 1)$ est $\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)]/(2n)$.

Ainsi, le levier caractérise l'excès de risque minimax (et celui du SMP) pour le problème de l'estimation de densité conditionnelle, comme pour celui de la régression (Section 1.3). L'interprétation est la même : un levier $\widehat{\ell}_{n+1}$ "déséquilibré" (non concentré autour de d/n) signifie que le prédicteur optimal β^* est mal estimé dans certaines directions, et que $\widehat{\beta}_n^{\mathrm{LS}}$ est déterminé par un plus petit nombre d'observations, ce qui accroît sa variabilité.

SMP régularisé et bornes non uniformes. Nous nous intéressons à présent à une variante régularisée du SMP (nous renvoyons au Chapitre 7 pour une définition générale du SMP régularisé), obtenue en considérant la pénalité $\beta \mapsto \lambda\|\beta\|^2/2$ pour un certain $\lambda > 0$. Cet estimateur satisfait une borne d'excès de risque *non uniforme* sur la classe $\mathcal{F}$, dépendant de la norme $\|\beta\|$ du paramètre de comparaison. Cette procédure est utile dans le cas où des bornes uniformes satisfaisantes ne sont pas possibles, c'est-à-dire lorsque (1) la loi P_X ne satisfait pas l'Hypothèse 1.1 (Section 1.3.2) de régularité, garantissant un excès de risque en d/n pour le SMP non régularisé; ou (2) la dimension d est élevée et supérieure à n, c'est-à-dire dans un cadre *non paramétrique*.

Pour ce qui est du premier cas, le résultat suivant montre qu'il est possible d'obtenir une borne de risque en dimension finie sans hypothèse de petite boule (Hypothèse 1.1), au prix d'une (modeste) dépendance en la norme $\|\beta\|$.

Proposition 1.8 (Proposition 7.3, Chapitre 7)**.** *Pour tout $\lambda > 0$, le SMP pénalisé s'écrit $\widehat{g}_{\lambda,n}(\cdot|x) = \mathcal{N}(\widetilde{\mu}_\lambda(x), \widetilde{\sigma}^2_\lambda(x))$, où $\widetilde{\mu}_\lambda(x), \widetilde{\sigma}^2_\lambda(x)$ sont indiqués dans la Proposition 7.3. Supposons que $\mathbb{E}[Y^2] < +\infty$ et $\|X\| \leqslant R$ presque sûrement. Alors, pour tout $B > 0$, le choix $\lambda = d/(B^2(n+1))$ conduit à :*

$$\mathbb{E}\big[R(\widehat{g}_{\lambda,n})\big] - \inf_{\|\beta\|\leqslant B} R(\beta) \leqslant \frac{5d\log\big(2 + BR/\sqrt{d}\big)}{n+1} = O\left(\frac{d}{n}\log\left(2 + \frac{BR}{\sqrt{d}}\right)\right). \tag{1.95}$$

En particulier, lorsque $R = O(\sqrt{d})$ et $\|\beta\| = O(1)$ (ce qui correspond au cas où X est approximativement "isotrope", c'est-à-dire où Σ est bien conditionnée[16], et où la "force du signal" $\|\beta\|_\Sigma$ est bornée), cette borne est d'ordre $O(d/n)$. La borne (1.95) est un raffinement de la garantie obtenue par conversion online-to-batch. En effet, Kakade and Ng (2005) montrent que la stratégie de mélange Bayésien $(\widehat{f}_{\nu,t})_{1\leqslant t\leqslant n+1}$ sur $\mathcal{F}$, avec pour loi a priori sur β donnée par $\pi = \mathcal{N}(0, \nu^2)$, satisfait la borne de regret

$$\sum_{t=1}^{n+1} \ell(\widehat{f}_{\nu,t}, (X_t, Y_t)) - \inf_{\|\beta\|\leqslant B} \sum_{t=1}^{n+1} \ell(f_\beta, (X_t, Y_t)) \leqslant \frac{B^2}{2\nu^2} + \frac{d}{2}\log\left(1 + \frac{\nu^2 R^2(n+1)}{d}\right)$$

dès lors que $\|X_t\| \leqslant R$ pour tout t. Par conversion online-to-batch (Proposition 1.3), on en déduit que, sous les hypothèses de la Proposition 1.8, l'estimateur de mélange $\bar{f}_{\nu,n} := (n+1)^{-1}\sum_{t=1}^{n+1}\widehat{f}_{\nu,t}$ avec $\nu = B/\sqrt{d}$ satisfait

$$\mathbb{E}\big[R(\bar{f}_{\nu,n})\big] - \inf_{\|\beta\|\leqslant B} R(\beta) \leqslant \frac{d(1 + \log(1 + B^2R^2(n+1)/d^2))}{2(n+1)} = O\left(\frac{d}{n}\log\left(2 + \frac{B^2R^2}{d}\frac{n}{d}\right)\right).$$

[16] Au sens où $C^{-1}I_d \preccurlyeq \Sigma \preccurlyeq CI_d$, avec $C = O(1)$. En pratique, le bon conditionnement peut s'obtenir en normalisant les X_i par la matrice de covariance empirique (éventuellement calculée sur une fraction séparée du jeu de données), c'est-à-dire par une étape de *pré-conditionnement*.

Cette borne est du même type que celle (1.95) du SMP pénalisé, mais avec un facteur additionnel en $\log(n/d)$. Par exemple, si $R = O(\sqrt{d})$ et $B = O(1)$, la borne précédente est en $O(d\log(n/d)/n)$, correspondant à une vitesse asymptotique en $O(d\log(n)/n)$ pour $n \gg d$.

Par ailleurs, il est également possible d'obtenir des bornes pour le SMP régularisé lorsque $d \geqslant n$. En effet, le Théorème 7.5 du Chapitre 7 montre que, sous les hypothèses de la Proposition 1.8, le SMP pénalisé $\widetilde{f}_{\lambda,n}$ satisfait, pour tout $\beta \in \mathbf{R}^d$,

$$\mathbb{E}[R(\widehat{g}_{\lambda,n})] - R(\beta) \leqslant 1.25 \cdot \frac{\mathrm{Tr}[(\Sigma + \lambda I_d)^{-1}\Sigma]}{n+1} + \frac{\lambda\|\beta\|^2}{2}\,.$$

Dans cette borne, la dimension d est remplacée par la quantité $\mathrm{Tr}((\Sigma + \lambda I_d)^{-1}\Sigma)$, correspondant aux *degrés de liberté* de l'estimateur Ridge (Wahba, 1990; Friedman et al., 2001; Wasserman, 2006). À nouveau, cette borne est valable uniformément sur les lois P telles que $\mathbb{E}[Y^2] < +\infty$.

Cas bien spécifié : sous-optimalité de l'EMV en grande dimension (Section 8.1). Considérons à présent le cas bien spécifié, où $Y|X \sim \mathcal{N}(\langle\beta^*, X\rangle, 1)$ pour un certain $\beta^* \in \mathbf{R}^d$, c'est-à-dire où $P \in \mathcal{P} := \mathcal{P}_{\mathrm{Gauss}}(P_X, 1)$ (cf. (1.55)). Supposons également P_X non dégénérée (Définition 6.1). Dans ce cas, comme souligné plus haut ainsi que dans la Section 1.3, l'excès de risque minimax parmi les estimateurs *propres* vaut

$$\frac{1}{2n}\mathbb{E}\big[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\big]\,,$$

et est atteint par l'EMV $f_{\widehat{\beta}_n^{\mathrm{LS}}}$. En outre, comme mentionné plus haut (voir le Théorème 8.2 de la Section 8.1), l'excès de risque minimax (sans restriction, donc en autorisant les estimateurs impropres) est

$$\frac{1}{2}\mathbb{E}\big[\log\big(1 + \langle(n\widehat{\Sigma}_n)^{-1}X, X\rangle\big)\big]\,,$$

atteint par l'estimateur impropre $\widehat{g}_n(\cdot|x) := \mathcal{N}(\langle\widehat{\beta}_n^{\mathrm{LS}}, \cdot\rangle, 1 + \langle(n\widehat{\Sigma}_n)^{-1}x, x\rangle)$. Le premier risque est supérieur à $d/(2(n-d+1))$ pour toute loi P_X, et vaut $d/(2(n-d-1))$ lorsque $X \sim \mathcal{N}(0,\Sigma)$. Par un argument similaire, on montre que le second risque est supérieur à $-\log(1-d/(n+1))/2$ pour toute loi P_X, et inférieur à $-\log(1-d/(n-1))/2$ si $P_X = \mathcal{N}(0,\Sigma)$.

Considérons maintenant le régime asymptotique de *grande dimension*, où $d, n \to \infty$ avec $d/n \to \gamma \in (0,1)$. Dans ce cadre, la performance optimale atteignable par un estimateur propre est de $\gamma/(2(1-\gamma))$ (avec égalité dans le cas où $P_X \sim \mathcal{N}(0,\Sigma)$). Cependant, le risque optimal atteint par l'estimateur impropre $\widehat{g}_n$ est (dans le cas Gaussien) de

$$-\frac{1}{2}\log(1-\gamma) = \frac{1}{2}\log\Big(1 + \frac{\gamma}{1-\gamma}\Big) < \frac{\gamma}{2(1-\gamma)}\,.$$

Ainsi, même dans le cas bien spécifié où P appartient à la classe $\mathcal{F}$, les estimateurs propres (restreints à $\mathcal{F}$) sont sous-optimaux lorsque la dimension d est relativement élevée (de l'ordre de n). Ceci contraste avec le cadre asymptotique classique (évoqué en Section 1.4.2), où d est fixé et $n \to \infty$, pour lequel l'EMV est asymptotiquement optimal. Intuitivement, le prédicteur $\widehat{g}_n$ admet un meilleur risque que l'EMV car il quantifie mieux l'incertitude sur la loi P (et donc sur ses réalisations futures) que l'EMV.

1.4.6 Régression logistique

Nous considérons à présent un autre problème classique d'estimation de densité conditionnelle, à savoir la *régression logistique* (Berkson, 1944; McCullagh and Nelder, 1989; van der Vaart, 1998). Ce problème correspond au cas où la réponse Y est binaire, ou plus généralement catégorielle.

Définitions et notations. Ici, $\mathcal{Y} = \{-1, 1\}$ est binaire, et $\mu = \delta_1 + \delta_{-1}$ est la mesure de comptage sur $\mathcal{Y}$. De plus, $\mathcal{X} = \mathbf{R}^d$, et $\mathcal{F}$ est le *modèle logistique* $\{f_\beta : \beta \in \mathbf{R}^d\}$, où

$$f_\beta(1|x) = 1 - f_\beta(-1|x) = \sigma(\langle \beta, x \rangle) \tag{1.96}$$

pour tous $\beta, x \in \mathbf{R}^d$, avec $\sigma : \mathbf{R} \to (0, 1)$ la *fonction sigmoïde* $\sigma(u) = e^u/(1 + e^u)$. Puisque $\sigma(-u) = 1 - \sigma(u)$, on a $f_\beta(y|x) = \sigma(y\langle \beta, x \rangle)$ pour tout $y \in \{-1, 1\}$. Ainsi, en notant $\ell(u) = \log(1 + e^u)$ pour $u \in \mathbf{R}$, on a pour tous $\beta, x \in \mathbf{R}^d$ et $y \in \{-1, 1\}$,

$$\ell(f_\beta, (x, y)) = \log(1 + e^{-y\langle \beta, x \rangle}) = \ell(-y\langle \beta, x \rangle) = \ell(\langle \beta, z \rangle)$$

où $z := -yx$. L'EMV est par définition, en notant $Z_i = -Y_i X_i$,

$$\widehat{\beta}_n^{\text{EMV}} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \widehat{R}_n(\beta) := \frac{1}{n} \sum_{i=1}^n \ell(\langle \beta, Z_i \rangle) \right\},$$

dès lors que ce minimiseur existe. Tout d'abord, le risque empirique $\widehat{R}_n(\beta)$ est convexe en β, car $u \mapsto \ell(u)$ est convexe. Considérons alors les cas de figure suivants :

- Le jeu de données est *(linéairement) séparé (au sens strict)*, s'il existe $\theta \in \mathbf{R}^d$ tel que $\langle \theta, Z_i \rangle < 0$ pour $i = 1, \dots, n$; cela revient à dire que $\langle \theta, X_i \rangle < 0$ si $Y_i = 1$, et $\langle \theta, X_i \rangle > 0$ si $Y_i = -1$, c'est-à-dire que l'hyperplan $\{x \in \mathbf{R}^d : \langle \theta, x \rangle = 0\}$ sépare les classes $\{X_i : Y_i = 1\}$ et $\{X_i : Y_i = -1\}$. Dans ce cas, $\widehat{R}_n(t\theta) \to 0$ lorsque $t \to +\infty$; comme par ailleurs $\widehat{R}_n > 0$, $\widehat{R}_n$ n'admet pas de minimum sur $\mathbf{R}^d$. On peut alors étendre la classe $\mathcal{F}$ en y ajoutant les densités conditionnelles f^θ ($\theta \in \mathbf{R}^d$, $\|\theta\| = 1$), où $f^\theta(y|x)$ vaut 1 si $y\langle \theta, x \rangle > 0$, 0 si $y\langle \theta, x \rangle < 0$ et $1/2$ si $y\langle \theta, x \rangle = 0$. Les minimiseurs du risque empirique sont alors donnés par les hyperplans séparateurs ; un tel hyperplan n'est pas unique.

- Le jeu de donné est dit *(strictement) non séparé* si pour tout $\theta \in \mathbf{R}^d \setminus \{0\}$, il existe un i tel que $\langle \theta, Z_i \rangle > 0$. Dans ce cas, la fonction $\beta \mapsto \widehat{R}_n(\beta)$ diverge lorsque $\|\beta\| \to \infty$; par continuité, elle admet donc un minimum dans $\mathbf{R}^d$. Enfin, les Z_i engendrent linéairement $\mathbf{R}^d$ (sinon il existerait un $\theta \neq 0$ tel que $\langle \theta, Z_i \rangle = 0$ pour tout i), ce qui implique que $\widehat{R}_n$ est strictement convexe et donc que son minimiseur $\widehat{\beta}_n^{\text{EMV}}$ est unique.

En revanche, pour tout $\lambda > 0$, l'estimateur pénalisé (de type Ridge)

$$\widehat{\beta}_{\lambda,n} = \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\langle \beta, Z_i \rangle) + \frac{\lambda}{2} \|\beta\|^2 \right\} \tag{1.97}$$

est bien défini de manière unique, par λ-forte convexité du risque pénalisé.

Garanties pour la régression logistique. Il découle de résultats classiques sur l'EMV que, sous des hypothèses modestes sur la loi jointe de (X, Y) (non séparation, moments contrôlés de $\|X\|$), $\widehat{\beta}_n^{\mathrm{EMV}}$ est consistant : $\widehat{\beta}_n^{\mathrm{EMV}} \to \beta^* = \arg\min_{\beta \in \mathbf{R}^d} R(\beta)$ en probabilité, et asymptotiquement normal (van der Vaart, 1998), au sens où $\sqrt{n}(\widehat{\beta}_n^{\mathrm{EMV}} - \beta^*)$ converge en loi vers $\mathcal{N}(0, H^{-1}GH^{-1})$, avec les notations de la Section 1.4.2. La constante $d_{\mathrm{eff}} = \mathrm{Tr}(H^{-1/2}GH^{-1/2})$, qui caractérise l'excès de risque asymptotique de l'EMV (voir la Section 1.4.2) peut être significativement plus élevée que d. En effet, si $\|X\| \leqslant R$ presque sûrement et $\|\beta^*\| \leqslant B$, d_{eff} peut être aussi élevée que de^{BR} (Bach and Moulines, 2013). De plus, l'excès de risque en espérance de l'EMV (avec choix arbitraire de l'hyperplan séparateur dans le cas séparé) est typiquement infini, par exemple lorsque $\mathbb{P}(Y = 1|X) \in (0, 1)$ presque sûrement : en effet, dans ce cas le jeu de données est séparé avec probabilité positive, l'EMV est alors un hyperplan séparateur qui admet une erreur infinie dans la région de $\mathbf{R}^d \times \{-1, 1\}$ à laquelle il attribue une probabilité nulle. Plus généralement, l'EMV tend à produire des prédictions trop confiantes, c'est-à-dire proches de 0 ou 1 (Sur and Candès, 2019).

Supposons dans ce qui suit que $\|X\| \leqslant R$ presque sûrement, et considérons des bornes d'excès de risque par rapport à la classe $\mathcal{F}_B := \{f_\beta : \beta \in \mathbf{R}^d, \|\beta\| \leqslant B\}$. Ce problème peut être envisagé comme un problème d'optimisation stochastique (Section 1.1.5). Tout d'abord, la perte $\ell(\beta, Z)$ est convexe en β et R-Lipschitz car $\|Z\| \leqslant R$. Cela implique qu'un excès de risque en $O(BR/\sqrt{n})$ est atteignable, de plusieurs façons :

- par minimisation du risque empirique sur $\mathcal{F}_B$ (par un argument de convergence uniforme de processus empiriques, voir la Section 1.1.4, et une inégalité de contraction sur les complexités de Rademacher, Ledoux and Talagrand, 2013) ;
- par l'estimateur pénalisé $\widehat{\beta}_{\lambda,n}$ (1.97) avec $\lambda = R/(B\sqrt{n})$ (par le Corollaire 1.1) ;
- par descente de gradient stochastique projetée sur $\mathcal{F}_B$ et moyennée (par la Proposition 1.2) ;
- par agrégation à poids exponentiels sur β (suivi d'une moyenne/conversion online to batch), avec une loi a priori sur β uniforme sur la boule $\{\beta \in \mathbf{R}^d : \|\beta\| \leqslant B\}$, par la Proposition 1.5 (dans ce cas, la borne obtenue est en fait de $BR\sqrt{(\log n)/n}$).

De plus, on vérifie que la fonction de perte $\ell(\beta, Z)$ est également e^{-BR}-exp-concave sur la boule de norme B. Ceci implique qu'une vitesse en $O(de^{BR}\log(n)/n)$ est possible, par conversion online-to-batch à partir des poids exponentiels avec $\eta = e^{-BR}$, ou de l'algorithme *Online Newton Step* (Hazan et al., 2007; Mahdavi et al., 2015). Une vitesse en $O(de^{BR}/n)$ peut également être obtenue par minimisation du risque empirique, pénalisée (Koren and Levy, 2015) ou non (Gonen and Shalev-Shwartz, 2018), ou d'autres procédures (Mehta, 2017).

De ce qui précède, il ressort qu'il est possible d'obtenir une borne d'excès de risque en

$$O\Big(\min\Big(\frac{BR}{\sqrt{n}}, \frac{de^{BR}}{n}\Big)\Big). \tag{1.98}$$

La première vitesse ci-dessus est une vitesse lente ; la seconde est une vitesse rapide (il s'agit de la vitesse asymptotique de l'EMV lorsque d est fixe et $n \to \infty$, pour une loi P choisie telle que $d_{\mathrm{eff}} \asymp de^{BR}$), mais avec une dépendance exponentielle prohibitive en BR (qui est typiquement d'ordre $\sqrt{d}$). Il s'avère en fait, par un résultat de Hazan et al. (2014), qu'il n'est pas possible d'améliorer la borne (1.98) pour un estimateur *propre* sans hypothèse supplémentaire sur P :

pour tout estimateur $\widehat{\beta}_n$, il existe une loi P pour laquelle l'excès de risque $\mathbb{E}[\mathcal{E}(f_{\widehat{\beta}_n})]$ par rapport à $\mathcal{F}_B$ est d'au moins (1.98). Pour des valeurs typiques de BR, d, n, on a $n = O(e^{BR})$ et la vitesse lente dans (1.98) est la meilleure. Ainsi, pour $B = \|\beta^*\| = O(1)$ et $R = O(\sqrt{d})$ (ce qui correspond au cas de "dimension finie" et "bien conditionné", voir la Section 1.4.5), cette vitesse est de $O(\sqrt{d/n})$.

Afin de contourner cette borne inférieure dans le pire des cas, il est naturel de faire des hypothèses supplémentaires, afin d'obtenir des bornes avec une dépendance explicite en certaines quantités dépendant de la loi P. Dans le pire des cas, ces bornes mènent inévitablement à la vitesse lente (1.98), mais pour certaines lois plus favorables (telles que $d_{\text{eff}} \ll de^{BR}$) la borne obtenue peut être bien meilleure. Pour mener à bien ce type d'analyse, la seule convexité (qui conduit à une vitesse lente) ou la courbure *globale* (qui induit une dépendance exponentielle en BR) ne sont pas suffisamment précises. Intuitivement, la difficulté du problème est déterminée, au moins asymptotiquement, par les propriétés *locales* de la fonction de perte $\ell(\beta, z)$ et du risque $R(\beta)$, pour β proche de l'optimum β^*. En particulier, pour $\beta \approx \beta^*$, le risque est approximativement quadratique : $R(\beta) - R(\beta^*) \approx \frac{1}{2}\|\beta - \beta^*\|_H^2$, où $H = \nabla^2 R(\beta^*)$. Afin d'obtenir des garanties non-asymptotiques fines en termes de quantités locales telles que H et d_{eff}, un contrôle global de la qualité de l'approximation quadratique locale de R est nécessaire.

Pour cela, une propriété utile de la fonction de perte logistique $\ell(\beta, z)$ est la *(pseudo-)auto-concordance*, introduite par Bach (2010) et exploitée par Bach (2014); Bach and Moulines (2013); Ostrovskii and Bach (2018); Marteau-Ferey et al. (2019) afin d'analyser la régression logistique (ainsi que d'autres problèmes d'optimisation stochastique) de manière non asymptotique. La notion d'*auto-concordance* (Nesterov and Nemirovskii, 1994), c'est-à-dire une majoration de la dérivée tierce d'une fonction en fonction de la puissance $3/2$ de sa dérivée seconde, est utilisée dans le cadre de l'analyse d'algorithmes d'optimisation du second ordre tels que la méthode de Newton (Nesterov and Nemirovskii, 1994; Boyd and Vandenberghe, 2004). La notion de *(pseudo-)auto-concordance* introduite par Bach (2010) correspond également à un contrôle de la dérivée troisième en fonction de la dérivée seconde, mais sans l'exposant $3/2$. Dans le cas de la perte logistique, cette propriété s'écrit, pour tout $u \in \mathbf{R}$,

$$|\ell'''(u)| = |\sigma(u)\big(1 - \sigma(u)\big)\big(1 - 2\sigma(u)\big)| \leqslant \sigma(u)\big(1 - \sigma(u)\big) = \ell''(u)\,.$$

À un haut niveau, l'auto-concordance est préférable à une borne uniforme sur la dérivée troisième, car elle permet de contrôler les variations *relatives* de la Hessienne

$$H(\beta) := \nabla^2 R(\beta) = \mathbb{E}[\ell''(\langle \beta, Z\rangle) ZZ^\top]\,.$$

En utilisant la propriété d'auto-concordance, Bach (2010) obtient une vitesse rapide pour la régression logistique avec pénalisation Ridge (1.97), dans le cas d'un design déterministe et d'un modèle logistique bien spécifié.

Dans le cas général mal spécifié et avec un design déterministe, supposons l'existence de $\beta^* = \arg\min_{\beta \in \mathbf{R}^d} R(\beta)$, et considérons les matrices suivantes :

$$\begin{aligned}\Sigma &= \mathbb{E}[XX^\top] = \mathbb{E}[ZZ^\top] \\ G(\beta) &= \mathbb{E}[\nabla \ell(\beta, Z) \nabla \ell(\beta, Z)^\top] = \mathbb{E}[\sigma(\langle \beta, Z\rangle)^2 ZZ^\top]\,,\end{aligned}$$

ainsi que $G = G(\beta^*)$, $H = H(\beta^*)$. L'excès de risque asymptotique de l'EMV est caractérisé par la *dimension effective* $d_{\text{eff}} = \operatorname{Tr}(H^{-1/2} G H^{-1/2})$ (voir la Section 1.4.2). Une autre quantité importante (Bach and Moulines, 2013) est la norme d'opérateur $\rho = \lambda_{\max}(H^{-1/2}\Sigma H^{-1/2})$,

c'est-à-dire la plus petite constante telle que $\Sigma \preccurlyeq \rho H$. Puisque $\ell'' = \sigma(1-\sigma) \leqslant 1/4$, on a nécessairement $\rho \geqslant 4$. De plus, si $B = \|\beta^*\|$ et $\|X\| \leqslant R$ presque sûrement, alors $\ell''(\langle \beta^*, X\rangle) \gtrsim e^{-BR}$, et donc $\rho \lesssim e^{BR}$; dans le pire des cas, $\rho \asymp e^{BR}$, mais ρ peut être bien plus faible en pratique (Bach and Moulines, 2013; Ostrovskii and Bach, 2018).

Bach (2014) considère l'algorithme de descente de gradient stochastique, avec un pas de $C/(R^2\sqrt{n})$ et avec moyenne des itérés, et établit une borne d'excès de risque de

$$\mathbb{E}\big[R(\widehat{\beta}_n)\big] - R(\beta^*) \lesssim \frac{R^2}{\mu n}\big(B^4R^4 + 1\big)\,, \tag{1.99}$$

avec $B = \|\beta^*\|$, et où μ désigne la plus petite valeur propre de H, c'est-à-dire la constante de forte convexité *locale* du risque au voisinage de β^*.

La borne (1.98) est sensible au conditionnement des données, c'est-à-dire au caractère isotrope (ou non) de la matrice de covariance Σ. Pour le voir, définissons $\lambda = \lambda_{\min}(\Sigma)$ la plus petite valeur propre de Σ. Alors

$$\lambda = \lambda_{\min}(\Sigma) \leqslant \frac{1}{d}\mathrm{Tr}(\Sigma) = \frac{\mathbb{E}[\|X\|^2]}{d} \leqslant \frac{R^2}{d}\,,$$

de sorte que

$$\frac{R^2}{\lambda} \geqslant d\,.$$

Par ailleurs, l'inégalité $H \preccurlyeq \Sigma/4$ implique que $\mu \leqslant \lambda/4$, et donc $R^2/\mu \geqslant 4R^2/\lambda \geqslant 4d$. De plus, pour les problèmes "mal conditionnés" où X est anisotrope, c'est-à-dire ici $\mathrm{Tr}(\Sigma)/\lambda_{\min}(\Sigma) \gg d$, on a $R^2/\mu \geqslant 4R^2/\lambda \gg d$. Enfin, notons que $\mu \geqslant \lambda/\rho$, et qu'il est possible d'avoir égalité. Dans le cas où $\mu \asymp \lambda/\rho$, on obtient $R^2/\mu \asymp \rho R^2/\lambda \gtrsim \rho d$, avec égalité (à une constante près) dans le cas bien conditionné. Ainsi, du fait de sa dépendance implicite en $\lambda = \lambda_{\min}(\Sigma)$, la borne (1.98) est principalement adaptée au cas paramétrique (avec $d \ll n$) bien conditionné. Dans ce cas, avec les ordres de grandeur $R^2/\mu = O(\rho d)$ et $BR = O(\sqrt{d})$, la borne (1.98) donne la vitesse $O(\rho \cdot d^3/n)$, avec une dépendance sous-optimale en d. Toutefois, en choisissant un pas différent (dépendant de B), il est possible de remplacer le terme $O(B^4R^4 + 1)$ par $O(B^2R^2)$, ce qui donne dans ce cas une vitesse en $O(\rho \cdot d^2/n)$.

À partir d'un algorithme et d'une analyse raffinées, Bach and Moulines (2013) montrent qu'il est possible d'éviter la dépendance en λ dans une borne de risque. Plus précisément, Bach and Moulines (2013) proposent un algorithme d'optimisation stochastique en deux étapes qui admet la garantie suivante : pour $n \gtrsim B^4R^4$,

$$\mathbb{E}[R(\widehat{\beta}_n)] - R(\beta^*) \lesssim \frac{\kappa^{3/2}\rho^3 d}{n}(B^4R^4 + 1)\,, \tag{1.100}$$

où κ est une borne sur la kurtosis des marginales unidimensionnelles des variables $[\ell''(\langle \beta, Z\rangle)]^{1/2}Z$ pour $\beta \in \mathbf{R}^d$. Cette borne correspond essentiellement à la précédente (1.99) dans le cas bien conditionné, mais cette fois-ci sans dépendance en λ lorsque $\lambda \ll R^2/d$. En particulier, en dimension finie avec $BR = O(\sqrt{d})$, cette borne mène à un excès de risque en $O(\rho^3 d^3/n)$ pour $n \gtrsim d^2$. Plus récemment, au moyen d'une analyse fine exploitant également l'auto-concordance, Ostrovskii and Bach (2018) ont obtenu des garanties améliorées pour l'estimateur du maximum de vraisemblance $\widehat{\beta}_n^{\mathrm{EMV}}$. En supposant que les vecteurs aléatoires "décorrélés" $\Sigma^{-1/2}Z$, $G^{-1/2}\ell'(\langle \beta^*, Z\rangle)Z$ et $H^{-1/2}[\ell''(\langle \beta, Z\rangle)]^{1/2}Z$ (pour β proche de β^*) sont

sous-Gaussiens, Ostrovskii and Bach (2018) montrent que, pour $n \gtrsim \max(\rho d_{\text{eff}}, d \log d)$, l'excès de risque est borné par

$$R(\widehat{\beta}_n^{\text{EMV}}) - R(\beta^*) \lesssim \frac{d_{\text{eff}}}{n}$$

avec forte probabilité, ce qui correspond à une variante non asymptotique du risque asymptotique de $\widehat{\beta}_n^{\text{EMV}}$. Cette borne améliore la précédente (1.100), tant du point de vue du risque que de la valeur de n requise. Notons que cette borne dépend des normes sous-Gaussiennes des vecteurs cités plus haut, qui dépendent elles-mêmes implicitement de la loi P, en particulier de la loi P_X, de β^* et de la loi de Y sachant X. Dans le cas où $X \sim \mathcal{N}(0, \Sigma)$ (par invariance de l'EMV par transformation linéaire, il est possible de supposer $\Sigma = I_d$), Ostrovskii and Bach (2018) montrent que ces normes peuvent être bornées en fonction de $B = \|\beta^*\|_\Sigma = \|\beta^*\|$, avec dans le cas de $G^{-1/2}\ell'(\langle \beta^*, Z\rangle)Z$ l'hypothèse supplémentaire que le modèle est bien spécifié. Enfin, Marteau-Ferey et al. (2019) considèrent le cas *non paramétrique* bien spécifié, et obtiennent des bornes non asymptotiques où la dimension d est remplacée par les degrés de libertés $\operatorname{Tr}[(H + \lambda I_d)^{-1} H]$ de H, et qui dépendent de la décroissance des coefficients de β^* dans une base de vecteurs propres de H. Ces bornes correspondent aux vitesses non paramétriques obtenues par l'estimateur Ridge dans le cas des moindres carrés (Caponnetto and De Vito, 2007).

Rappelons que dans le cas général où l'on suppose seulement que $\|X\| \leqslant R$, les bornes précédentes exhibent toutes une dépendance inévitable en e^{BR} dans le pire des cas, en raison de la borne inférieure (1.98) de Hazan et al. (2014) pour les estimateurs propres. Il est cependant possible de contourner cette borne inférieure sans hypothèse supplémentaire, en ayant recours à des prédicteurs *impropres* (Foster et al., 2018), tels que les estimateurs par mélange Bayésien (Section 1.4.3). Plus précisément, Kakade and Ng (2005) et Foster et al. (2018) ont montré que la stratégie de prédiction en ligne par mélange Bayésien, avec pour lois a priori respectives $\mathcal{N}(0, B^2/d)$ et la loi uniforme sur la boule de norme B, admettent une borne de regret en

$$O\Big(d \log\Big(2 + \frac{BRn}{d}\Big)\Big)$$

dès lors que $\|X_t\| \leqslant R$ pour tout t. Par convergence online-to-batch (Proposition 1.3), Foster et al. (2018) en déduit une borne d'excès de risque de

$$\mathbb{E}[R(\bar{f}_n)] - \inf_{\|\beta\| \leqslant B} R(\beta) = O\Big(\frac{d}{n} \log\Big(2 + \frac{BRn}{d}\Big)\Big) \tag{1.101}$$

pour la moyenne $\bar{f}_n$ des itérés. Signalons au passage que Foster et al. (2018) proposent également un autre estimateur $\widehat{f}_{n,\delta}$ pour lequel ils énoncent une borne d'excès de risque avec forte probabilité $1 - \delta$ (Théorème 10). Cette borne est néanmoins erronée : en effet, la preuve applique l'inégalité de Markov sur l'excès de risque d'un estimateur impropre intermédiaire ; cependant, cet estimateur étant impropre (en dehors de la classe $\mathcal{F}_B$), son excès de risque par rapport à la classe $\mathcal{F}_B$ peut prendre des valeurs négatives, de sorte qu'il n'est pas possible de lui appliquer l'inégalité de Markov.

La borne (1.101) a ceci de remarquable qu'elle ne dépend pas des constantes ρ, d_{eff} et n'admet donc pas de dépendance exponentielle en la norme BR. L'estimateur $\bar{f}_n$ contourne donc la borne inférieure (1.98) pour les estimateurs propres dans le pire des cas. Lorsque $BR = O(\sqrt{d})$, cette borne fournit une vitesse en $O(d \log(n)/n)$, optimale au facteur $\log n$ près.

Le principal inconvénient de cette approche est sa complexité algorithmique, soulignée par Foster et al. (2018). En effet, $\bar{f}_n$ est la moyenne des estimateurs $\widetilde{f}_t$, $1 \leqslant t \leqslant n+1$, où $\widetilde{f}_t$ est le postérieur prédictif Bayésiens calculé à partir des $t-1$ premières observations, soit

$$\widetilde{f}_t(y|x) := \int_{\mathbf{R}^d} \sigma(\langle \beta, x \rangle) \widehat{\pi}_t(\mathrm{d}\beta)\,,$$

où $\widehat{\pi}_t$ est le postérieur $\pi(\cdot|Z_1, \ldots, Z_{t-1})$, dont la densité par rapport à π est donnée par

$$\frac{\prod_{s=1}^{t-1} f_\beta(Y_s|X_s)}{\int_{\mathbf{R}^d} \prod_{s=1}^{t-1} f_\beta(Y_s|X_s)\pi(\mathrm{d}\beta)} = \frac{\prod_{s=1}^{t-1} \sigma(-\langle \beta, Z_s \rangle)}{\int_{\mathbf{R}^d} \prod_{s=1}^{t-1} \sigma(-\langle \beta, Z_s \rangle)\pi(\mathrm{d}\beta)}\,. \tag{1.102}$$

Le dénominateur de (1.102) n'admettant pas d'expression explicite, il est nécessaire de recourir à des techniques de calcul approché de postérieurs. Par exemple, Foster et al. (2018) montrent qu'il est possible de calculer approximativement $\bar{f}_n$ en un temps de $O(B^6 \max(d, BRn)^{12}/\varepsilon^{12})$, où ε désigne le niveau de précision recherché. Bien que ce temps soit polynomial en n, d, cette procédure est trop coûteuse pour être utilisable en pratique. Un problème ouvert posé par Foster et al. (2018) demande s'il est possible d'obtenir un algorithme moins coûteux satisfaisant une vitesse rapide de regret ou d'excès de risque. Comme nous le montrons dans la section suivante (ainsi que dans le Chapitre 7), le SMP apporte une réponse positive (partielle) à cette question.

1.4.7 Application du SMP à la régression logistique (Chapitre 7)

Nous appliquons maintenant le SMP à la régression logistique. Commençons pour simplifier par considérer le SMP non pénalisé. Ses prédictions sont données par

$$\widehat{g}_n^{\mathrm{SMP}}(y|x) = \frac{f_{\widehat{\beta}_n^{(x,y)}}(y|x)}{f_{\widehat{\beta}_n^{(x,1)}}(1|x) + f_{\widehat{\beta}_n^{(x,-1)}}(-1|x)} = \frac{\sigma(\langle \widehat{\beta}_n^{(x,y)}, yx \rangle)}{\sigma(\langle \widehat{\beta}_n^{(x,1)}, x \rangle) + \sigma(\langle \widehat{\beta}_n^{(x,-1)}, -x \rangle)}\,, \tag{1.103}$$

pour tout $(x, y) \in \mathbf{R}^d \times \{-1, 1\}$, où $\widehat{\beta}_n^{(x,y)}$ désigne l'EMV obtenu en rajoutant (x, y) à l'échantillon. De plus, sa borne d'excès de risque (1.93) s'écrit :

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \mathbb{E}_{Z_1^n, Z}\big[\sigma(\langle \widehat{\beta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\beta}_n^{Z}, Z \rangle)\big]\,. \tag{1.104}$$

Intuitivement, la borne d'excès de risque du SMP est plus fine qu'une borne en termes de stabilité de la perte, s'appliquant par exemple à l'EMV (voir la Section 1.1.6). En effet, en notant $v = \langle \widehat{\beta}_n^{-Z}, Z \rangle$ et $u = \langle \widehat{\beta}_n^{Z}, Z \rangle$, on a pour $u \simeq v \gg 1$, $\ell(\langle \widehat{\beta}_n^{-Z}, Z \rangle) - \ell(\langle \widehat{\beta}_n^{Z}, Z \rangle) = \ell(v) - \ell(u) \simeq v - u$, tandis que $\sigma(\langle \widehat{\beta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\beta}_n^{Z}, Z \rangle) \simeq \sigma'(u)(v-u) \simeq e^{-u}(v-u)$. Ainsi, le terme (1.104) peut être exponentiellement plus petit qu'un terme de stabilité de la perte. Ce gain permet d'éviter une dépendance exponentielle en BR dans le pire des cas.

D'un point de vue qualitatif, les prédictions du SMP sont moins "confiantes" (proches de 0 ou 1) que celles de l'EMV. En particulier, elles appartiennent toujours à $(0, 1)$, et sont toujours définies de manière unique même dans le cas séparé (si le jeu de données augmenté de (x, y) est séparé, alors $\widehat{f}_n^{(x,y)}(y|x) = 1$, quel que soit le choix de l'hyperplan séparateur). En particulier, si le point x est tel que la quantité

$$\sigma(\langle \widehat{\beta}_n^{(x,1)}, x \rangle) + \sigma(\langle \widehat{\beta}_n^{(x,-1)}, -x \rangle) = 1 + \sigma(\langle \widehat{\beta}_n^{(x,1)}, x \rangle) - \sigma(\langle \widehat{\beta}_n^{(x,-1)}, x \rangle)\,,$$

qui peut être vue comme un analogue du levier dans le cas logistique (en tant que mesure de l'influence de l'étiquette y du point x sur la prédiction associée), est élevée, alors la prédiction correspondante du SMP est "incertaine" (proche de $1/2$). Dans le cas extrême où le jeu de données est séparé, et où x est tel que le jeu de données reste séparé en y ajoutant $(x,1)$ ou $(x,-1)$, alors $\sigma(\langle\widehat{\beta}_n^{(x,1)},x\rangle)=\sigma(\langle\widehat{\beta}_n^{(x,-1)},-x\rangle)=1$, de sorte que $\widehat{g}_n^{\mathrm{SMP}}(1|x)=1/2$. À l'inverse, l'EMV $f_{\widehat{\beta}_n^{\mathrm{EMV}}}$ satisfait $f_{\widehat{\beta}_n^{\mathrm{EMV}}}(1|x)=0$ ou $f_{\widehat{\beta}_n^{\mathrm{EMV}}}(1|x)=1$, en fonction du choix de l'hyperplan séparateur, les deux cas étant possibles. Ainsi, le SMP corrige l'une des défaillances de l'EMV, qui est de produire des prédictions trop confiantes.

Afin d'obtenir des garanties non asymptotiques, nous considérons le SMP non pénalisé.

Théorème 1.14 (Théorème 7.6, Chapitre 7)**.** *Pour le modèle logistique* $\mathcal{F}=\{f_\beta:\beta\in\mathbf{R}^d\}$ (1.96)*, le SMP avec pénalité* $\beta\mapsto\lambda\|\beta\|^2/2$ *(où* $\lambda>0$*) s'écrit*

$$\widetilde{f}_{\lambda,n}(y|x)=\frac{\sigma(y\langle\widehat{\beta}_{\lambda,n}^{(x,y)},x\rangle)\,e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(x,y)}\|^2/2}}{\sigma(\langle\widehat{\beta}_{\lambda,n}^{(x,1)},x\rangle)\,e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(x,1)}\|^2/2}+\sigma(-\langle\widehat{\beta}_{\lambda,n}^{(x,-1)},x\rangle)\,e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(x,-1)}\|^2/2}} \tag{1.105}$$

où l'on note $\widehat{\beta}_\lambda^{(x,y)}=\widehat{\beta}_\lambda^{(-yx)}$*, avec pour* $z\in\mathbf{R}^d$ *:*

$$\widehat{\beta}_{\lambda,n}^{(z)}:=\operatorname*{arg\,min}_{\beta\in\mathbf{R}^d}\left\{\frac{1}{n+1}\left[\sum_{i=1}^n\ell(\langle\beta,Z_i\rangle)+\ell(\langle\beta,z\rangle)\right]+\frac{\lambda}{2}\|\beta\|^2\right\}.$$

De plus, pour toute loi jointe de (X,Y) *telle que* $\|X\|\leqslant R$ *presque sûrement, l'estimateur* (1.105) *satisfait, pour tout* $\lambda\geqslant 2R^2/(n+1)$*,*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big]\leqslant R(\beta)+3\cdot\frac{\mathrm{Tr}[(\Sigma+4\lambda I_d)^{-1}\Sigma]}{n}+\frac{\lambda}{2}\|\beta\|^2\,. \tag{1.106}$$

Cette borne est utile dans le cas non paramétrique où $d\gg n$. Il s'agit d'une vitesse rapide, similaire à celles obtenues par Bach (2010); Marteau-Ferey et al. (2019) dans le cas bien spécifié. Ces dernières admettent de plus un terme de variance exprimé en fonction de $H(\beta^*)$ (qui est toujours inférieur à $\Sigma/4$), et un terme de biais plus général que $\lambda\|\beta^*\|^2$, exprimé en fonction de différentes normes. En revanche, ces résultats nécessitent des hypothèses plus fortes, et exhibent une dépendance exponentielle en BR dans le cas général. D'un point de vue technique, notre analyse utilise également la notion d'auto-concordance de Bach (2010); Bach and Moulines (2013); Ostrovskii and Bach (2018), en conjonction avec la borne générale du SMP (Théorème 1.12).

Dans le cas paramétrique de dimension $d\ll n$, on en déduit directement le résultat suivant :

Corollaire 1.3 (Corollaire 7.2, Chapitre 7)**.** *Supposons que* $\|X\|\leqslant R$ *presque sûrement. Alors, l'estimateur* (1.105) *avec* $\lambda=2R^2/(n+1)$ *satisfait, pour tout* $B>0$*,*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big]-\inf_{\|\beta\|\leqslant B}R(\beta)\leqslant\frac{3d}{n}+\frac{B^2R^2}{n}\,. \tag{1.107}$$

La borne (1.107) est valable en supposant seulement que X est borné. Une telle garantie n'est pas possible pour un estimateur propre (tel que l'estimateur pénalisé de type Ridge (1.97)) sans hypothèse supplémentaire, par la borne inférieure (1.98) de Hazan et al. (2014). Tout comme la procédure de Foster et al. (2018) obtenue par conversion online-to-batch de stratégies

de mélange Bayésien, le SMP pénalisé admet une vitesse rapide sans dépendance exponentielle en BR.

Notons que la borne (1.101) est *logarithmique* en la norme BR, tandis que la borne (1.107) pour le SMP est quadratique en BR, à l'instar des résultats de Bach (2010); Marteau-Ferey et al. (2019) dans le cas bien spécifié. Dans le cas du modèle Gaussien, nous avions obtenu une borne avec une dépendance logarithmique en BR (Proposition 1.8, en utilisant un paramètre de régularisation λ plus faible lorsque BR est élevé ; cela n'est plus possible ici, car nous avons besoin de "localiser" le paramètre $\widehat{\beta}^{-Z}_{\lambda,n}$ dans un voisinage de diamètre constant de $\widehat{\beta}^{Z}_{\lambda,n}$, afin de pouvoir effectuer une approximation quadratique locale du risque en utilisant l'auto-concordance. Notons toutefois que dans le régime considéré précédemment où $BR = O(\sqrt{d})$, la borne (1.107) du SMP est bien d'ordre optimal $O(d/n)$, tandis que la borne (1.101) est d'ordre $O(d\log(n)/n)$.

D'un point de vue computationnel, la prédiction du SMP s'obtient en calculant les deux solutions $\widehat{\beta}^{(x,1)}_{\lambda,n}$ et $\widehat{\beta}^{(x,-1)}_{\lambda,n}$ de régressions logistiques "perturbées", obtenues en ajoutant un échantillon. Ainsi, comparé à une approche fondée sur le mélange Bayésien (Kakade and Ng, 2005; Foster et al., 2018), le SMP remplace un problème d'échantillonnage selon le postérieur par un problème d'optimisation, qui est typiquement moins coûteux. En ce sens, le SMP apporte une réponse au problème ouvert posé par Foster et al. (2018). Le SMP reste cependant plus coûteux qu'une régression logistique simple (lorsque l'on cherche à calculer la prédiction en plusieurs valeurs de x), car il requiert de calculer les solutions modifiées $\widehat{\beta}^{(x,1)}_{\lambda,n}$ et $\widehat{\beta}^{(x,-1)}_{\lambda,n}$ pour chaque point de test x. Cette différence de complexité s'observe également dans le cas du modèle Gaussien (Section 1.4.5) : une fois calculés $\widehat{\beta}^{\mathrm{LS}}_n$ et $\widehat{\Sigma}^{-1}_n$, avec un coût de $O(nd^2+d^3) = O(nd^2)$ (pour $n \geqslant d$), la prédiction de l'EMV en $x \in \mathbf{R}^d$, soit $\mathcal{N}(\langle\widehat{\beta}^{\mathrm{LS}}_n, x\rangle, 1)$, se calcule en temps $O(d)$, tandis que celle du SMP $\widehat{g}^{\mathrm{SMP}}_n(\cdot|x) = \mathcal{N}(\langle\widehat{\beta}^{\mathrm{LS}}_n, x\rangle, (1+\langle\widehat{\Sigma}^{-1}_n x, x\rangle)^2)$ s'obtient en temps $O(d^2)$. Il y a donc dans ce cas aussi un compromis entre garanties statistiques et coût computationnel.

1.5 Forêts aléatoires

Cette section porte sur les méthodes de forêts aléatoires, étudiées dans la Partie I. Dans la Section 1.5.1, nous décrivons le principe des méthodes de forêts aléatoires, avant de passer en revue les garanties théoriques existantes sur leur performance prédictive en Section 1.5.2. La Section 1.5.3 porte sur notre principale contribution, à savoir l'analyse précise d'une variante de forêts, les *forêts de Mondrian* (introduites par Lakshminarayanan et al., 2014), menée dans le Chapitre 2, ainsi que celle d'une variante séquentielle de l'algorithme (Chapitre 3).

1.5.1 Forêts aléatoires : principes généraux

Méthodes ensemblistes. Dans cette thèse, nous étudions une famille particulière de procédures utilisés tant en classification qu'en régression, à savoir les *forêts aléatoires* (*Random Forests* en anglais). Proposées par Breiman (2001a), et inspirées par les travaux de Amit and Geman (1997); Ho (1998); Dietterich (2000); Breiman (2000); Cutler and Zhao (2001), les forêts aléatoires appartiennent à la famille des méthodes dites d'*ensemble* ou *ensemblistes* (*ensemble methods* en anglais), qui consistent à combiner plusieurs prédicteurs individuels afin d'obtenir un prédicteur "agrégé" de bonne qualité. Les approches ensemblistes les plus courantes sont :

- le *boosting* (Freund and Schapire, 1997; Friedman, 2001; Schapire and Freund, 2012), qui combine des prédicteurs simples de manière séquentielle afin d'obtenir un prédicteur combiné ayant une erreur faible, chaque prédicteur élémentaire étant choisi de manière à corriger les erreurs des précédents. Cette procédure part de prédicteurs simples (ayant une forte erreur d'approximation) et opère une réduction du biais.

- le *bagging* (contraction de *bootstrap aggregating*), introduit par Breiman (1996). À l'inverse du boosting, cette approche part d'un prédicteur complexe caractérisé par une forte attache aux données, ce qui est généralement associé à un faible biais mais à une variance élevée. Le bagging consiste alors à construire plusieurs réalisations aléatoires indépendantes de ce prédicteur, obtenues en l'évaluant sur des sous-échantillons tirés aléatoirement, et à effectuer la moyenne de ces réalisations aléatoires. L'idée sous-jacente est de réduire la variance du prédicteur, en effectuant la moyenne de plusieurs réalisations faiblement ou modérément corrélées.

La philosophie des forêts aléatoires se rapproche davantage de celle du bagging — qui est d'ailleurs l'un des ingrédients de la procédure proposée par Breiman (2001a) — que de celle du boosting : elles partent de prédicteurs individuels complexes (les arbres de décision, décrits plus bas) construits de manière randomisée, dont elles forment la moyenne non pondérée. De plus, à l'instar du bagging et contrairement au boosting, les prédicteurs individuels sont construits en parallèle, indépendamment les uns des autres.

Arbres de décision. Les prédicteurs individuels que combinent les forêts aléatoires sont des *arbres de décision* (Breiman et al., 1984; Devroye et al., 1996). Ces derniers sont construits en partitionnant l'espace de manière récursive, en appliquant successivement des *coupures* qui séparent chaque cellule en deux. Afin de simplifier la discussion, nous supposons que $\mathcal{X} = [0,1]^d$; notons toutefois que les arbres de décision ont la propriété agréable qu'elles permettent de traiter simultanément des variables continues et discrètes (Breiman et al., 1984).

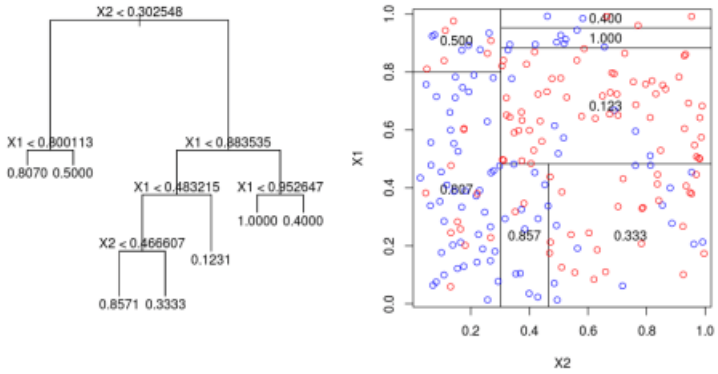

Figure 1.2: Un arbre de décision en dimension $d = 2$. Le schéma de gauche indique la structure d'arbre binaire (ainsi que les coupures associées à chaque nœud), tandis que la partie droite illustre la partition de $[0,1]^2$ correspondante. À chaque nœud de l'arbre, correspondant à une cellule de la partition, est associée une prédiction. Source : https://www.r-bloggers.com/regression-tree-using-ginis-index/.

Un arbre de décision (à valeurs dans $\widehat{\mathcal{Y}}$) est un prédicteur $g : [0,1]^d \to \widehat{\mathcal{Y}}$, caractérisé les éléments suivants (voir la Figure 1.2).

- Une partition finie $\mathscr{C}$ de $[0,1]^d$ dite *arborescente* (en anglais *kd-tree* ou *tree partition*), c'est-à-dire associée à
  - un arbre binaire $\mathcal{T}$ fini, enraciné et ordonné. Ainsi, chaque nœud $\mathbf{v}$ distinct de la racine admet un unique parent, et admet soit deux nœuds fils (un fils gauche noté $\mathbf{v}0$, et un fils droit noté $\mathbf{v}1$), auquel cas $\mathbf{v}$ est dit *intérieur*, soit aucun nœud fils, auquel cas $\mathbf{v}$ est une *feuille* (ou *nœud terminal*) ;
  - ainsi qu'une famille de *coupures* $\Sigma = (\sigma_{\mathbf{v}})_{\mathbf{v}}$ indexée par les nœuds intérieurs de $\mathcal{T}$. Une coupure σ est caractérisée par une coordonnée de coupure $j \in \{1, \dots, d\}$ ainsi que d'un seuil $s \in \mathbf{R}$.

  À partir de la donnée de l'arbre $\mathcal{T}$ et des coupures Σ, il est possible d'associer récursivement à tout nœud $\mathbf{v}$ de $\mathcal{T}$ une *cellule* $C_{\mathbf{v}} \subset [0,1]^d$, de la façon suivante :
  - la cellule associée à la racine ε de $\mathcal{T}$ est $C_\varepsilon = [0,1]^d$;
  - pour tout nœud intérieur $\mathbf{v}$, étant donné sa cellule $C_{\mathbf{v}}$ et la coupure $\sigma_{\mathbf{v}} = (j_{\mathbf{v}}, s_{\mathbf{v}})$, on définit $C_{\mathbf{v}0} = \{x \in C_{\mathbf{v}} : x_{j_{\mathbf{v}}} \leqslant s_{\mathbf{v}}\}$ (où $x = (x_j)_{1 \leqslant j \leqslant d}$), et $C_{\mathbf{v}1} = C_{\mathbf{v}} \setminus C_{\mathbf{v}0}$.

  Ainsi, les cellules $C_{\mathbf{v}}$ des feuilles $\mathbf{v}$ de $\mathcal{T}$ forment une partition $\mathscr{C}$ de $[0,1]^d$ associée à la structure d'arbre et aux coupures.
- Une prédiction $\widehat{y}_{\mathbf{v}} = \widehat{y}_C \in \widehat{\mathcal{Y}}$ associée à toute feuille $\mathbf{v}$ de $\mathcal{T}$, c'est-à-dire à toute cellule $C = C_{\mathbf{v}}$ de $\mathscr{C}$. Ces prédictions définissent une fonction $g : [0,1]^d \to \widehat{\mathcal{Y}}$ par $g(x) = \widehat{y}_{C(x)}$, où $C(x) \in \mathscr{C}$ est la cellule de x, telle que $x \in C(x)$.

Une fois la structure de la partition déterminée, les arbres de décision prédisent généralement en utilisant la minimisation du risque empirique dans chaque cellule ; dans le cas de la classification (Exemple 1.3), cela revient à choisir la classe la plus fréquente dans la cellule,

tandis que dans le cas de la régression (Exemple 1.4), cela correspond à la moyenne des Y_i tels que X_i appartienne à la cellule.

Ainsi, les algorithmes d'arbres de décision sont caractérisés par le choix de la partition de $[0,1]^d$. L'algorithme de référence en classification et régression à partir d'arbres de décision est l'algorithme CART (*Classification And Regression Trees*, Breiman et al., 1984). Il s'agit d'un algorithme *glouton*, qui optimise le choix de la coupure dans chaque nœud en fonction de la réduction immédiate de l'erreur apportée par cette coupure, plutôt que par une optimisation globale prenant en compte l'effet des coupures suivantes. Cela est dû au fait que le problème de la minimisation du risque empirique parmi tous les arbres de décision est NP-complet (Hyafil and Rivest, 1976).

Il reste alors à choisir la complexité des arbres, c'est-à-dire le critère d'arrêt qui détermine quand un nœud n'est plus coupé. L'algorithme CART (Breiman et al., 1984; Friedman et al., 2001) procède de la façon suivante : dans un premier temps, un arbre de décision complètement développé (ne contenant qu'un point par cellule) est formé de manière gloutonne ; dans un second temps, cet arbre est *élagué* (en enlevant certaines coupures), en minimisant l'erreur empirique pénalisée par le nombre de feuilles.

Les arbres de décision, et en particulier l'algorithme CART, sont simples, rapides à évaluer (une fois l'arbre de décision formé, calculer les prédictions ne demande que de comparer les coordonnées de l'entrée x aux seuils des coupures, le long du chemin de x dans l'arbre) et interprétables (l'arbre renseigne sur les variables qui ont déterminé la prédiction). Cependant, du point de vue de la performance prédictive, les arbres individuels souffrent de certaines limitations : en particulier les prédictions de l'algorithme CART s'avèrent instables, car fortement sensibles au choix des coupures (Breiman, 1996).

Forêts aléatoires. Comme nous l'avons annoncé, les forêts aléatoires sont obtenues en combinant des arbres individuels. La procédure proposée par Breiman (2001a) repose sur les ingrédients suivants :

- des arbres de décision construits en parallèle de manière randomisée, et non élagués : chaque arbre est "profond" et contient un faible nombre de points par cellule. Les arbres individuels sont combinés en faisant la moyenne de leurs prédictions (dans le cas de la régression), ou par un vote majoritaire (en classification) ;
- chaque arbre est construit sur un sous-échantillon de $(X_1, Y_1), \ldots, (X_n, Y_n)$, obtenu par exemple en tirant n points avec répétition parmi l'échantillon (on parle alors d'échantillon de type *bootstrap*, utilisé dans le cas du bagging) ;
- le choix de la partition (et en particulier des coupures) est partiellement randomisé, notamment en tirant aux hasard les coordonnées considérées pour chaque coupure potentielle.

Signalons également l'existence d'une variante de forêts aléatoires couramment utilisée, à savoir l'algorithme *Extra-Trees* (Geurts et al., 2006). Cette procédure combine également les ingrédients ci-dessus, à l'exception du sous-échantillonnage (bagging).

Les forêts aléatoires comptent parmi les algorithmes de classification et de régression les plus couramment utilisés en pratique (Fernández-Delgado et al., 2014). Elles combinent en effet une très bonne performance prédictive, un coût d'entraînement et d'évaluation modeste, et le peu (voire l'absence) de paramètres libres à calibrer. Ces succès empiriques contrastent

cependant avec une compréhension et des garanties théoriques limitées pour cette famille de procédures (Arlot and Genuer, 2014; Wager and Walther, 2015; Biau and Scornet, 2016).

1.5.2 Revue des résultats théoriques sur les forêts aléatoires

Dans cette section, nous passons en revue les garanties théoriques disponibles sur la performance prédictive des forêts aléatoires. Nous nous restreignons au cas de la régression avec perte quadratique (pour lequel $\mathcal{Y} = \widehat{\mathcal{Y}} = \mathbf{R}$), ce qui permet d'avoir recours à une décomposition biais-variance précise. Signalons également que des résultats généraux permettent de convertir toute garantie de risque en régression en une garantie en classification (Devroye et al., 1996). Au delà de l'erreur de prédiction, de nombreux autres questions relatives aux forêts, tant théoriques que méthodologiques, ont été étudiées ; nous renvoyons à Criminisi et al. (2012); Boulesteix et al. (2012); Biau and Scornet (2016) pour des revues plus complètes sur ce sujet.

Biais, variance et randomisation. Afin de clarifier la discussion ultérieure, nous commençons par des préliminaires généraux sur le risque, le biais et la variance de prédicteurs randomisés et de leurs ensembles. Ces résultats s'appliquent à tous les estimateurs de type ensembliste (qui effectuent la moyenne simple d'estimateurs randomisés individuels) et donc en particulier aux différentes variantes de forêts aléatoires.

Soit $\widehat{g}_n(\cdot;\Theta) : [0,1]^d \to \mathbf{R}$ un prédicteur randomisé, construit à partir du jeu de données i.i.d. $\mathscr{D}_n = \big((X_1,Y_1),\dots,(X_n,Y_n)\big)$ et de l'aléa Θ introduit par l'algorithme (pour les forêts, à travers le choix aléatoire des coupures, le sous-échantillonnage, etc.). Définissons pour tout $x \in [0,1]^d$:

$$\bar{g}_n(x;\Theta) = \mathbb{E}\big[\widehat{g}_n(x;\Theta)|\Theta\big]\,,$$

où l'espérance conditionnelle par rapport à Θ revient à intégrer sur l'échantillon aléatoire $\mathscr{D}_n$. Ainsi, en notant $g^*(X) = \mathbb{E}[Y|X]$ la fonction de régression (Exemple 1.4), on a pour tout $x \in \mathbf{R}^d$

$$\mathbb{E}\big[(\widehat{g}_n(x;\Theta) - g^*(x))^2|\Theta\big] = \mathbb{E}\big[(\bar{g}_n(x;\Theta) - g^*(x))^2|\Theta\big] + \mathrm{Var}\big(\widehat{g}_n(x;\Theta)|\Theta\big)\,. \tag{1.108}$$

En prenant $x = X$ dans (1.108), en considérant l'espérance sur (X,Θ) et en notant que $R(g) - R(g^*) = \mathbb{E}[(g(X) - g^*(X))^2]$ pour toute fonction $g : [0,1]^d \to \mathbf{R}^d$ (Exemple 1.4), on obtient la *décomposition biais-variance* suivante[17] :

$$\mathbb{E}\big[R(\widehat{g}_n(\cdot;\Theta))\big] - R(g^*) = \mathbb{E}\big[(\bar{g}_n(X;\Theta) - g^*(X))^2\big] + \mathbb{E}\big[\mathrm{Var}(\widehat{g}_n(X;\Theta)|X,\Theta)\big]\,. \tag{1.109}$$

Le premier terme du membre de droite de (1.109) constitue le *biais* de $\widehat{g}_n$; il s'agit de l'espérance sur (X,Θ) d'un terme indépendant de l'échantillon. Le second terme correspond à la *variance* de $\widehat{g}_n$. Notons que le terme de "variance" se réfère ici à la variabilité induite par l'échantillon $\mathscr{D}_n$; l'aléa Θ introduit par l'algorithme contribue quant à lui à la fois au biais et à la variance. Un estimateur de variance élevée est instable, en général car il s'attache trop aux spécificités du jeu de données. À l'inverse, une procédure présentant un biais élevé est insuffisamment flexible pour approcher la fonction de régression g^*.

[17] Cette décomposition diffère légèrement de celle que nous utilisons au Chapitre 2, qui introduit un terme intermédiaire distinct (mais proche) de $\bar{g}_n$.

Risque d'estimateurs ensemblistes. Considérons maintenant $M \geqslant 1$ réalisations i.i.d. $\Theta^{(1)}, \ldots, \Theta^{(M)}$ de Θ, indépendantes de l'échantillon $\mathscr{D}_n$, et définissons l'estimateur d'ensemble (moyenné)

$$\widehat{g}_{n,M}(x;\Theta_M) = \frac{1}{M}\sum_{m=1}^{M} \widehat{g}_n(x;\Theta^{(m)})$$

avec $\Theta_M = (\Theta^{(1)}, \ldots, \Theta^{(M)})$. À $\mathscr{D}_n$ et x fixés, la loi des grands nombres implique que, lorsque $M \to \infty$, $\widehat{g}_{n,M}(x;\Theta_M)$ converge presque sûrement vers $\widehat{g}_{n,\infty}(x) := \mathbb{E}[\widehat{g}_{n,M}(x;\Theta^{(1)})|\mathscr{D}_n]$ dès lors que cette espérance est bien définie. $\widehat{g}_{n,\infty}$ correspond à un estimateur *non-randomisé* de g^*. En écrivant une décomposition "biais-variance" par rapport à Θ_M, conditionnellement à $\mathscr{D}_n$, on obtient pour tout $x \in \mathbf{R}^d$:

$$\begin{aligned}\mathbb{E}\big[(\widehat{g}_{n,M}(x;\Theta_M) - g^*(x))^2|\mathscr{D}_n\big] &= \big(\mathbb{E}[\widehat{g}_{n,M}(x;\Theta_M)|\mathscr{D}_n] - g^*(x)\big)^2 + \mathrm{Var}\big(\widehat{g}_{n,M}(x;\Theta_M)|\mathscr{D}_n\big)\\ &= \big(\widehat{g}_{n,\infty}(x) - g^*(x)\big)^2 + \frac{1}{M}\mathrm{Var}\big(\widehat{g}_n(x;\Theta^{(1)})|\mathscr{D}_n\big)\,;\end{aligned}$$

en considérant l'espérance sur $x = X$ et $\mathscr{D}_n$, on obtient :

$$\mathbb{E}\big[R(\widehat{g}_{n,M})\big] = \mathbb{E}\big[R(\widehat{g}_{n,\infty})\big] + \frac{1}{M}\mathbb{E}\big[\mathrm{Var}\big(\widehat{g}_n(X;\Theta^{(1)})|X,\mathscr{D}_n\big)\big]\,. \tag{1.110}$$

L'expression (1.110) montre que le risque moyen de $\widehat{g}_{n,M}$ décroît avec le nombre de répétitions M, et converge lorsque $M \to \infty$ vers celui de l'estimateur idéalisé $\widehat{g}_{n,\infty}$, correspondant à un ensemble infini. Bien souvent, $\widehat{g}_{n,\infty}$ est complexe et n'est pas calculable explicitement ; $\widehat{g}_{n,M}$ en constitue un approximation de type Monte-Carlo, d'autant plus précise que M est élevé. En pratique, M est contraint par des limites computationnelles, et l'expression (1.110) quantifie le risque additionnel encouru pour une valeur donnée de M.

Remarque 1.8 (Nombre de répétitions). En considérant le cas $M = 1$ dans (1.110), il vient

$$\mathbb{E}\big[\mathrm{Var}\big(\widehat{g}_n(X;\Theta^{(1)})|X,\mathscr{D}_n\big)\big] = \mathbb{E}\big[R(\widehat{g}_{n,M})\big] - \mathbb{E}\big[R(\widehat{g}_{n,\infty})\big] \leqslant \mathbb{E}\big[R(\widehat{g}_{n,M})\big] - R(g^*)\,,$$

de sorte que (1.110) implique, en notant $\mathcal{R}(\widehat{h}_n) = \mathbb{E}[R(\widehat{h}_n)] - R(g^*)$ l'excès de risque en espérance de l'estimateur $\widehat{h}_n$:

$$\mathcal{R}(\widehat{g}_{n,M}) \leqslant \mathcal{R}(\widehat{g}_{n,\infty}) + \frac{1}{M}\cdot\mathcal{R}(\widehat{g}_{n,1})\,.$$

Ainsi, on a $\mathcal{R}(\widehat{g}_{n,M}) \lesssim \mathcal{R}(\widehat{g}_{n,\infty})$ dès lors que $M \gtrsim \mathcal{R}(\widehat{g}_{n,1})/\mathcal{R}(\widehat{g}_{n,\infty})$. Le nombre de prédicteurs individuels M nécessaires à une performance optimale dépend donc de la qualité relative des prédicteurs individuels et ensemblistes.

Considérons à présent l'effet de la moyenne $\widehat{g}_{n,M}$ sur le biais et la variance (au sens de la décomposition (1.109), c'est-à-dire par rapport à l'échantillon aléatoire $\mathscr{D}_n$). Définissons comme précédemment les intermédiaires

$$\bar{g}_{n,M}(x;\Theta_M) = \mathbb{E}[\widehat{g}_{n,M}(x;\Theta_M)|\Theta_M] = \frac{1}{M}\sum_{m=1}^{M}\bar{g}_n(x;\Theta^{(m)})$$

(qui ne dépend que de Θ_M et pas de $\mathscr{D}_n$) et $\bar{g}_{n,\infty}(x) = \mathbb{E}[\widehat{g}_{n,\infty}(x)] = \mathbb{E}[\widehat{g}_n(X;\Theta)]$ (qui est purement déterministe). Le terme de biais dans la décomposition (1.109) vaut alors :

$$\mathbb{E}\big[(\bar{g}_{n,M}(X;\Theta_M) - g^*(X))^2\big] = \mathbb{E}\big[(\bar{g}_{n,\infty}(X) - g^*(X))^2\big] + \frac{1}{M}\mathbb{E}\big[\mathrm{Var}\big(\bar{g}_n(X;\Theta)|X\big)\big]\,; \tag{1.111}$$

le premier terme de la décomposition (1.111) correspond au biais de la procédure idéalisée $\widehat{g}_{n,\infty}$, tandis que le second constitue le biais additionnel de son approximation $\widehat{g}_{n,M}$. De même, le terme de variance dans (1.109) s'écrit

$$\begin{aligned}
&\mathbb{E}\big[\mathrm{Var}(\widehat{g}_{n,M}(X;\Theta_M)|X,\Theta_M)\big] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\big[\mathrm{Var}(\widehat{g}_n(X;\Theta^{(m)})|X,\Theta^{(m)})\big] + \\
&\quad + \frac{1}{M^2}\sum_{m\neq m'}\mathbb{E}\big[\mathrm{Cov}(\widehat{g}_n(X;\Theta^{(m)}),\widehat{g}_n(X;\Theta^{(m')})|X,\Theta^{(m)},\Theta^{(m')})\big] \\
&= \frac{1}{M}\mathbb{E}\big[\mathrm{Var}(\widehat{g}_n(X;\Theta^{(1)})|X,\Theta^{(1)})\big] + \frac{M-1}{M}\mathbb{E}\big[\mathrm{Cov}\big(\widehat{g}_n(X;\Theta^{(1)}),\widehat{g}_n(X;\Theta^{(2)})|X,\Theta^{(1)},\Theta^{(2)}\big)\big].
\end{aligned}$$

En particulier, pour M grand, la variance de l'estimateur moyenné $\widehat{g}_{n,M}$ est proche de celle de l'ensemble infini, qui vaut

$$\mathbb{E}\big[\mathrm{Var}(\widehat{g}_{n,\infty}(X)|X)\big] = \mathbb{E}\big[\mathrm{Cov}\big(\widehat{g}_n(X;\Theta^{(1)}),\widehat{g}_n(X;\Theta^{(2)})|X,\Theta^{(1)},\Theta^{(2)}\big)\big]\,. \tag{1.112}$$

Ainsi, la variance de l'estimateur ensembliste infini dépend de la corrélation (par rapport à l'échantillon $\mathscr{D}_n$) entre les prédictions de deux réalisations indépendantes de l'estimateur randomisé individuel.

Garanties théoriques sur les forêts. L'article original de Breiman (2001a) établit des bornes sur l'erreur des forêts (une pour la classification et une pour la régression), en fonction de la qualité des arbres individuels et de la corrélation entre les différents arbres ; la borne obtenue en régression est similaire à celle issue de la décomposition biais-variance indiquée plus haut.

Un pan important de la littérature étudie les propriétés de *consistance* de forêts aléatoires. Biau et al. (2008) établit la consistance de certains types de forêts en classification en la déduisant de celle des arbres individuels[18], et montre également que certains classifieurs ensemblistes sont consistants même lorsque les classifieurs individuels ne le sont pas.

Pour montrer la consistance d'arbres individuels, il est possible d'utiliser des résultats généraux de consistance d'histogrammes (Devroye et al., 1996). Par exemple, un classifieur en histogramme dont (1) la partition est construite indépendamment des réponses Y_i ; (2) le diamètre de la cellule du point de test X tend en loi vers 0 lorsque $n \to \infty$; et (3) le nombre de points de l'échantillon dans la cellule du point X tend vers l'infini lorsque $n \to \infty$, est consistant indépendamment de la loi de (X, Y) en classification (Devroye et al., 1996). Ces résultats découlent eux-mêmes de résultats généraux de consistance d'estimateurs à moyenne locale dûs à Stone (1977), dont la consistance des k-plus proche voisins (énoncée dans le Théorème 1.2) est un cas particulier. Garantir les conditions nécessaires à la consistance pour des forêts (ou pour des arbres) requiert en général de considérer le mécanisme de construction des arbres, tant du point de vue de la structure combinatoire de l'arbre (en particulier le nombre de cellules) que de la loi des coupures (qui affecte notamment le diamètre des cellules).

À la suite de Biau et al. (2008) (qui étudie des modèles simples de forêts), un axe de recherche récent consiste à établir la consistance de variantes de forêts plus sophistiquées (Denil

[18]Dans le cas de la classification (contrairement à celui de la régression), le risque d'un ensemble de prédicteurs randomisés combinés par vote majoritaire n'est pas nécessairement inférieur à celui des classifieurs individuels, à cause de la non-convexité de l'erreur de classification.

et al., 2014; Scornet et al., 2015; Wager and Walther, 2015; Mentch and Hooker, 2016; Cui et al., 2017; Wager and Athey, 2018; Athey et al., 2019), se rapprochant davantage des forêts utilisées en pratique. Bien que plus complexes (en raison de la complexité des algorithmes considérés), les preuves reposent sur les mêmes principes généraux. Wager and Athey (2018) et Mentch and Hooker (2016) ont indépendamment établi la normalité asymptotique de l'estimateur des forêts aléatoires, lorsque la taille n de l'échantillon tend vers l'infini ; en particulier, ils montrent que la variance asymptotique de l'estimateur des forêts peut être estimée à l'aide du Jackknife. Ces résultats sont utiles dans le cadre de tâches inférentielles plus générales que la prédiction, comme par exemple pour former des intervalles de confiance sur les valeurs de la fonction de régression. En revanche, ils ne fournissent pas de vitesse théorique sur la variance asymptotique de l'estimateur, et donc de garantie sur la qualité de celui-ci.

Difficultés liées au choix glouton des coupures. Dans le cas des forêts de Breiman, l'obtention de garanties théoriques précises est délicate. Biau et al. (2008) montrent notamment que les forêts de Breiman $\widehat{g}_n$ sont en fait *inconsistantes*, au sens où il existe une loi jointe de (X, Y) telle que $R(g^*) = 0$ mais $\mathbb{E}[R(\widehat{g}_n)] \geqslant c > 0$ pour tout n. Cette limitation tient au fait que ces forêts sont construites de manière *gloutonne*, en optimisant le choix de coupures en fonction de la réduction immédiate de l'erreur. Il est en effet possible de construire des lois de (X, Y) pour lesquelles certaines coupures nécessaires (car permettant des coupures ultérieures réduisant fortement l'erreur) mais n'entraînant pas de réduction immédiate de l'erreur ne seront jamais réalisées, l'algorithme lui préférant indéfiniment des coupures induisant un (faible) gain immédiat.

Scornet et al. (2015) démontrent la consistance d'un algorithme de forêts proche de celui de Breiman, mais avec des arbres suffisamment élagués[19]. Cette garantie de consistance est obtenue en régression, en supposant que la loi de Y sachant X suit un modèle additif (Stone, 1985) :

$$Y = \sum_{j=1}^{d} f_j(X^j) + \varepsilon\,,$$

où X^j désigne la $j$ième coordonnée de X, et où $\mathbb{E}[\varepsilon|X] = 0$. En d'autres termes, la fonction de régression $g^*(X) = \mathbb{E}[Y|X]$ se décompose comme une somme de fonctions des coordonnées, et ne présente pas de terme d'interaction. L'absence d'interactions permet d'éviter les configurations mettant en défaut le choix glouton des coupures, où le bénéfice d'une coupure n'apparaît qu'après des coupures ultérieures. Notons cependant que les arbres et les forêts, qui combinent des coupures selon plusieurs variables, permettent en principe d'approcher des fonctions de régression avec des effets d'interaction complexes.

Dans une autre approche, Wager and Walther (2015) établissent un résultat garantissant la validité de l'estimation des moyennes dans les feuilles, uniformément sur une famille de partitions "médianes" où chaque coupure partage les données en deux fractions équilibrées. Ce résultat s'applique donc aux partitions "adaptatives", qui utilisent les données (et en particulier les sorties Y_i) pour choisir les coupures, à condition que les Y_i utilisés pour choisir les coupures soient disjoints de ceux utilisés pour estimer les moyennes. Afin d'obtenir la consistance, Wager and Walther (2015) considèrent également une hypothèse sur la fonction de régression (d'effet minoré de coupures individuelles) qui assure la qualité du choix glouton de coupures.

[19] Dans le cas des forêts de Breiman classiques qui ne sont pas élaguées, Scornet et al. (2015) réduit la consistance à une conjecture relativement délicate à établir.

Vitesses de convergence pour des forêts stylisées. La seule consistance ne donne que peu d'information sur l'effet de différents choix de paramètres de l'algorithme, ainsi que sur les facteurs qui influencent la qualité des prédictions. Il est en effet possible d'établir la consistance même au moyen d'une analyse lâche, qui ne capture que très partiellement le comportement effectif de l'algorithme. Pour cette raison, plusieurs travaux cherchent à quantifier l'erreur de prédiction des forêts, ainsi que sa dépendance en les paramètres de l'algorithme, sous certaines hypothèses sur la fonction de régression.

En raison des difficultés liées au choix glouton des coupures, il est commode de considérer des forêts simplifiées, où le choix des coupures ne se fait plus par minimisation de l'erreur mais par randomisation. Cela correspond dans une certaine mesure à la motivation initiale des forêts, qui cherchent à pallier la trop forte instabilité des arbres CART (due notamment à la sensibilité au choix des coupures, Breiman, 1996, 2001a). Cette approche conduit à considérer des forêts stylisées dites *purement aléatoires* (en anglais *purely random forests*, abrégé PRF), pour lesquelles la partition est générée indépendamment du jeu de données. Des forêts de ce type ont été considérées par Cutler and Zhao (2001); Breiman (2000), et analysées par Breiman (2004); Biau et al. (2008); Biau (2012); Genuer (2012); Arlot and Genuer (2014); Klusowski (2018) ; nous étudions également un tel algorithme dans le Chapitre 2. Signalons également que des idées similaires de partitions aléatoires ont été étudiées par Rahimi and Recht (2008, 2009) ; en particulier, l'algorithme de *random binning* de Rahimi and Recht (2008) repose sur des partitions aléatoires du cube $[0,1]^d$. À l'inverse, chaque arbre aléatoire (en tant qu'histogramme) effectue une régression linéaire avec pour variables (aléatoires) les indicatrices des différentes cellules de la partition, à l'instar de la procédure de *Random Kitchen Sinks* (RKS) de Rahimi and Recht (2009). La façon dont les différentes partitions sont utilisées diffère cependant entre les PRF et les RKS ; tandis que les premières effectuent la moyenne simple des histogrammes associés aux arbres individuels, les seconds les combinent avec des poids eux-mêmes optimisés.

Breiman (2004); Biau (2012) considèrent un algorithme de type PRF, les *forêts centrées* (Biau and Scornet, 2016), obtenues de la façon suivante :

- les arbres sont complets de profondeur p (à 2^p feuilles), où $p = p_n$ est fixé ;
- les coupures sont obtenues dans chaque nœud en coupant au milieu d'une coordonnée j sélectionnée au hasard, uniformément sur $\{1, \dots, d\}$.

Biau (2012) établit que de telles forêts (avec un choix convenable de p_n) atteignent un risque de $O(n^{-1/((4/3\cdot\log 2)d+1)})$ lorsque la fonction de régression $g^* : [0,1]^d \to \mathbf{R}$ est Lipschitz. De plus, si la fonction de régression ne dépend que de $s \leqslant d$ variables, et si l'algorithme choisit en fait chacune de ces variables avec probabilité proche de $1/s$ (et les variables restantes avec une probabilité faible), alors la vitesse de convergence devient $O(n^{-1/((4/3\cdot\log 2)s+1)})$, ce qui apporte un gain significatif lorsque la dimension d est élevée mais le nombre s de variables informatives est faible. Biau (2012) propose également un mécanisme en partie heuristique de sélection de variables, permettant de sélectionner les s variables aléatoires en ayant recours à un échantillon indépendant. Duroux and Scornet (2018); Wager and Walther (2015) considèrent des forêts proches (dites *médianes*), et obtiennent la même vitesse de $O(n^{-1/((4/3\cdot\log 2)d+1)})$, ainsi que $O(n^{-1/((4/3\cdot\log 2)s+1)})$ dans le cas de forêts coupant selon les s variables informatives. Klusowski (2018) complète ces résultats en établissant des bornes inférieures pour les forêts centrées.

Les vitesses de convergences précédentes s'avèrent sous-optimales par rapport à la vitesse minimax sur la classe des fonctions Lipschitz, qui est de $O(n^{-2/(d+2)})$ (Stone, 1980, 1982; Györfi et al., 2002). Intuitivement, cela tient au fait que le choix aléatoire (uniforme) des coordonnées de coupure conduit à des cellules déséquilibrées, où certaines coordonnées ont été choisies moins souvent que d'autres ; le diamètre de ces cellules est donc élevé à cause de ces coordonnées. De manière générale, la nature récursive des arbres aléatoires, qui introduisent de l'aléa à chaque nouvelle coupure, peut conduire à des partitions déséquilibrées ou délicates à contrôler de manière théorique. Arlot and Genuer (2014) obtiennent la vitesse optimale $O(n^{-2/3})$ dans le cas $d = 1$ pour une autre variante de PRF, appelée *forêts purement aléatoires uniformes* (*purely uniformly random forest*, PURF, dont la partition associée de $[0, 1]$ s'obtient en tirant un nombre fixe $k - 1 \geqslant 0$ d'extrémités uniformément dans $[0, 1]$), mais une vitesse sous-optimale en dimension $d \geqslant 2$ pour une autre variante de PRF. Enfin, dans le Chapitre 2, nous étudions une variante de PRF, les *forêts de Mondrian* (Lakshminarayanan et al., 2014), dont nous montrons qu'elles atteignent la vitesse minimax $O(n^{-2/(d+2)})$ (voir également la Section 1.5.3).

Avantage des forêts par rapport aux arbres. Comme nous venons de le voir, les vitesses de convergence évoquées précédemment permettent de distinguer différentes façons de construire les arbres de manière randomisée, en montrant que certaines constructions conduisent à des vitesses moins bonnes que d'autres, en raison du caractère plus au moins bien équilibré des partitions. Cependant, tous les résultats précédents s'appliquent tant aux forêts qu'aux arbres individuels, et ne mettent donc pas en évidence d'avantage des forêts par rapport aux arbres.

Un tel effet a été établi pour la première fois par Arlot and Genuer (2014) : pour l'algorithme PURF décrit plus haut (qui combine des histogrammes aléatoires en dimension 1), lorsque la fonction de régression est deux fois continûment différentiable, la forêt infinie atteint une vitesse améliorée de[20] $O(n^{-4/5})$, qui est optimale pour cette classe de fonctions, à l'inverse des arbres simples qui admettent toujours une vitesse de $O(n^{-2/3})$ dans ce cas. Arlot and Genuer (2014) établissent également des vitesses améliorées pour les forêts dans le cas de PRF en dimension $d \geqslant 2$, bien que les vitesses soient dans ce cas sous-optimales. Dans le Chapitre 2, nous étendons les résultats optimaux de Arlot and Genuer (2014) pour les PURF dans le cas $d = 1$, au cas général $d \geqslant 1$ pour les forêts de Mondrian, en montrant que les forêts infinies (ou avec suffisamment d'arbres) atteignent la vitesse optimale $O(n^{-4/(d+4)})$ sur la classe des fonctions deux fois différentiables, tandis que les arbres atteignent seulement la vitesse en $O(n^{-2/(d+2)})$ du cas Lipschitz.

Dans ces résultats, il est important de noter que les vitesses améliorées pour les forêts proviennent d'une réduction du *biais*, et non de la *variance* (qui n'est réduite que d'un facteur constant). L'effet sous-jacent est un phénomène de lissage : tandis qu'un arbre exhibe des discontinuités aux bords des cellules, une forêt présente plusieurs petites discontinuités provenant de chacun des arbres, ayant lieu à des seuils différents. La fonction de régression associée à une forêt est par conséquent plus lisse que celle d'un arbre, et approche donc mieux les fonctions régulières.

Cet effet de réduction du biais (et non de la variance) va à rebours de la motivation initiale ayant conduit aux forêts, et plus généralement aux méthodes d'agrégation ensembliste de type bagging. En effet, l'objectif de ces méthodes est de partir de prédicteurs individuels complexes

[20] À un effet de bord près, commun à toutes les procédures par moyennes locales (Wasserman, 2006).

(de faible biais mais de variance élevée), et d'en réduire la variance. En particulier, les forêts de Breiman combinent des arbres plus profonds (non élagués) que ceux de l'algorithme CART ; à l'inverse, les résultats obtenus pour les PRF (réduction du biais) suggèrent une profondeur optimale des forêts inférieure à celle des arbres simples. Il y a deux interprétations complémentaires à ces résultats : d'une part, le lissage opéré par les forêts entraîne une réduction du biais, davantage que de la variance (le phénomène de lissage a également été relevé par Bühlmann and Yu, 2002, qui l'associent à une réduction d'un facteur constant de la variance) ; d'autre part, les modèles de forêts *purement aléatoires* ne permettent pas de mettre en évidence un effet de réduction de la variance dans les forêts, qui justifierait d'utiliser des arbres profonds.

Ainsi, à notre connaissance, aucun résultat existant ne permet de justifier l'usage des forêts aléatoires telles que proposées par Breiman (c'est-à-dire constituées d'arbres individuels profonds et non élagués) ou de montrer leur avantage par rapport à des arbres simples, convenablement élagués. Les propriétés des forêts avancées pour justifier leur performance, telles que leur capacité à sélectionner des variables informatives (Breiman, 2004; Biau, 2012; Scornet et al., 2015; Wager and Walther, 2015) ou à s'adapter à la "dimension intrinsèque" (notamment aux corrélations) des variables (Dasgupta and Freund, 2008; Verma et al., 2009), bien que pertinentes, sont également applicables aux arbres individuels. Comme souligné précédemment, l'effet de réduction du biais (par lissage des prédiction) conduit à sélectionner des forêts *moins profondes* que les arbres individuels. Il est bien possible de montrer que le sous-échantillonnage permet de réduire la variance, notamment dans le cas de l'estimateur du plus proche voisin (Biau and Devroye, 2010; Biau et al., 2010; Samworth, 2012) ; toutefois, cette réduction de variance suppose que la taille k_n des sous-échantillons satisfait $k_n/n \to 0$; dans le cas où $k_n \asymp n$, qui correspond au *bagging* utilisé en pratique, les résultats existants ne montrent qu'une réduction d'un facteur constant de la variance.

Une piste possible est d'étudier des partitions aléatoires partiellement adaptatives (contrairement à celles des estimateurs de forêts purement aléatoires, pour lequel la moyennisation ne réduit que le biais), pour lesquelles il est possible d'escompter une réduction de la variance. Une approche plus abordable consisterait à étudier l'effet de la moyenne de prédicteurs individuels "complexes" (qui interpolent le jeu de données), construits par un sous-échantillonnage des observations ou des variables, pour des procédures plus simples que les méthodes de forêts, telles que des méthodes linéaires, pour lesquelles une analyse explicite semble envisageable.

1.5.3 Analyse des forêts de Mondrian (Chapitres 2 et 3)

Notre principale contribution à l'étude des forêts est l'analyse d'une variante de forêts purement aléatoires (PRF), à savoir les *forêts de Mondrian* (Lakshminarayanan et al., 2014). Ces forêts reposent sur une loi particulière sur les partitions arborescentes du cube $[0,1]^d$, appelée *partitions de Mondrian*, introduites par Roy and Teh (2009). Dans la discussion qui suit, nous allons donner une définition récursive informelle de cette loi ; pour une définition rigoureuse, nous renvoyons à Roy (2011) ou à la Section 2.8.1 du Chapitre 2.

Définition des forêts de Mondrian. Le *processus de Mondrian* $\mathsf{MP}([0,1]^d)$ (Roy and Teh, 2009) est en fait (la loi d') un processus $(\Pi_\lambda)_{\lambda \in \mathbf{R}^+}$ de partitions Π_λ de $[0,1]^d$, telles que la partition $\Pi_{\lambda'}$ est un raffinement de la partition Π_λ pour tous $\lambda' \geqslant \lambda \geqslant 0$, obtenue en effectuant d'éventuelles coupures supplémentaires des cellules de Π_λ. Le paramètre λ, appelé *durée de vie* de la partition Π_λ, gouverne la complexité de cet arbre. La partition Π_λ, dont la loi est

notée $\mathsf{MP}(\lambda, [0,1]^d)$ s'obtient de la façon suivante (nous définissons en fait la loi $\mathsf{MP}(\lambda, C)$ pour tout hyperrectangle $C = \prod_{j=1}^d [a_j, b_j] \subset \mathbf{R}^d$) :

En notant ε la racine de l'arbre associé, on pose $C_\varepsilon = C$. En notant $C = \prod_{j=1}^d [a_j, b_j]$, soient E_j, $j = 1, \dots, d$, des variables exponentielles indépendantes d'intensités respectives $b_j - a_j$, et soit $E = \min_{1 \leqslant j \leqslant d} E_j$.

- Si $E > \lambda$, la cellule C_ε n'est pas coupée, la partition Π_λ est simplement $\{C\}$.
- Sinon, soit $J = \arg\min_{1 \leqslant j \leqslant d} E_j$, et S une variable uniforme sur $[a_J, b_J]$. La cellule C_ε est alors coupée selon la coordonnée J au seuil S, en deux cellules $C_0 := \{x \in C_\varepsilon : x_J \leqslant S\}$ et $C_1 := C_\varepsilon \backslash C_0$. Les cellules C_0 et C_1 sont alors elles-mêmes partitionnées indépendamment en répétant ce procédé, mais avec un budget de $\lambda - E$: $\Pi^0 \sim \mathsf{MP}(\lambda - E, C_0)$ et $\Pi^1 \sim \mathsf{MP}(\lambda - E, C_1)$, et la partition Π_λ est $\Pi^0 \cup \Pi^1$.

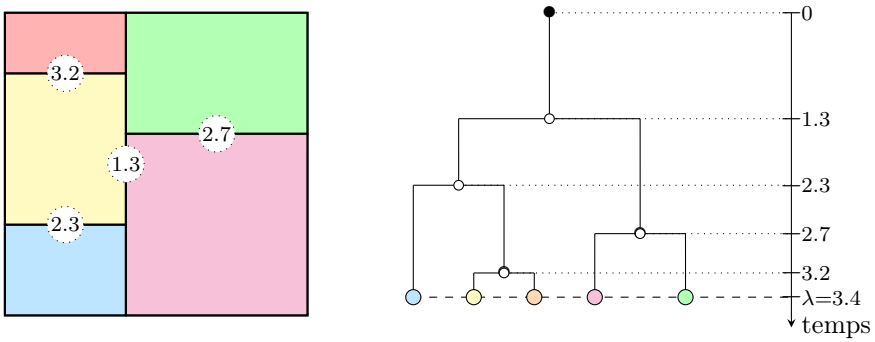

Figure 1.3: Une partition de Mondrian (à gauche), avec la structure d'arbre correspondante (à droite). Les temps des coupures sont indiqués sur l'axe vertical, tandis que les coupures sont signalés par des cercles (∘).

On obtient ainsi une partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$ de C, avec ici $C = [0,1]^d$. En faisant varier le budget λ, c'est-à-dire l'instant à partir duquel on cesse de couper les cellules, on obtient le processus $(\Pi_\lambda)_{\lambda \in \mathbf{R}^+} \sim \mathsf{MP}([0,1]^d)$. Par définition, $\Pi_{\lambda'}$ raffine Π_λ pour tout $\lambda' \geqslant \lambda$. De plus, on vérifie (en utilisant l'absence de mémoire de la loi exponentielle) que $(\Pi_\lambda)_\lambda$ est un processus de Markov, au sens où pour $\lambda \leqslant \lambda'$, $\Pi_{\lambda'}$ est indépendant de $(\Pi_t)_{t<\lambda}$ conditionnellement à Π_λ.

En dimension $d = 1$, il est possible de montrer que $\mathsf{MP}(\lambda, [0,1])$ est la loi de la partition de $[0,1]$ dont les lieux de coupures forment un processus ponctuel de Poisson d'intensité λ (Roy and Teh, 2009; Roy, 2011). En outre, une propriété fondamentale du processus de Mondrian est celle de restriction ; cette propriété découle des propriétés des variables exponentielles.

Proposition 1.9 (Roy, 2011)**.** *Soient $C_0 \subset C_1$ deux hyperrectangles de $\mathbf{R}^d$. Si $\Pi_\lambda \sim \mathsf{MP}(\lambda, C_1)$, alors la partition $\Pi_\lambda|_{C_0} = \{A \cap C_0 : A \in \Pi_\lambda\}$ de C_0 induite par restriction de la partition Π_λ suit la loi $\mathsf{MP}(\lambda, C_0)$.*

Les *forêts de Mondrian* ont été proposées par Lakshminarayanan et al. (2014). Cet algorithme effectue la moyenne d'arbres de décision tirés indépendamment selon la loi $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$; cet estimateur utilise également une forme de régularisation au sein de chaque arbre (contrairement aux estimateurs par histogrammes), au moyen d'une procédure bayésienne hiérarchique sur l'arbre calculée de manière approximative (Lakshminarayanan et al.,

2014). Nous omettons ce second aspect de la procédure, et appelons *forêt de Mondrian* l'estimateur de type PRF obtenu en utilisant des partitions de loi $\mathsf{MP}(\lambda, [0,1]^d)$; dans ce cas, la régularisation est obtenue par le choix du paramètre de complexité λ des partitions. Nous reviendrons sur cet aspect à la fin de cette section et dans le Chapitre 3.

Analyse des forêts de Mondrian (Chapitre 2). Les forêts de Mondrian ont été introduites pour des raisons computationnelles, en tant que procédure calculable de manière séquentielle (en ligne) à partir des propriétés de Markov et de restriction des partitions de Mondrian (Lakshminarayanan et al., 2014).

Dans le Chapitre 2, nous montrons que cette variante de PRF se prête à une analyse théorique précise. Celle-ci repose sur le fait qu'il est possible d'obtenir directement une description exacte des propriétés locales et globales utiles à l'analyse statistique. Le théorème suivant (qui correspond à la Proposition 2.1 du Chapitre 2) fournit en particulier la loi *exacte* de la cellule $C_\lambda(x)$ contenant un point $x \in [0,1]^d$ dans une partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$.

Théorème 1.15 (Proposition 2.1, Chapitre 2)**.** *Soit $x \in [0,1]^d$, et soit $C_\lambda(x)$ la cellule d'une partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ contenant x. Alors, $C_\lambda(x)$ a la même loi que*

$$\prod_{j=1}^{d} \left[(x_j - \lambda^{-1} E_{j,L}) \vee 0, (x_j + \lambda^{-1} E_{j,R}) \wedge 1 \right], \tag{1.113}$$

où $x = (x_j)_{1 \leqslant j \leqslant d}$, et où $E_{1,L}, E_{1,R}, \ldots, E_{d,L}, E_{d,R}$ sont des variables i.i.d. de loi $\mathsf{Exp}(1)$.

Pour les variantes de PRF considérées dans la littérature (Breiman, 2000; Biau et al., 2008; Arlot and Genuer, 2014; Scornet, 2016), la loi des cellules est en général complexe et n'admet pas de description aussi précise. En effet, le contrôle de la partition requiert de considérer l'effet des coupures successives, ce qui conduit à une analyse délicate. Dans le cas des partitions de Mondrian, la caractérisation de la loi des cellules s'obtient directement, en exploitant la propriété de restriction des partitions de Mondrian, sans avoir à raisonner conditionnellement à la structure combinatoire de l'arbre ou au nombre de coupures.

Une seconde quantité importante pour l'analyse est le nombre de cellules des partitions. Pour cette variante de forêts, il est également de possible de calculer l'espérance de cette quantité :

Proposition 1.10 (Proposition 2.2, Chapitre 2)**.** *Si $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$, alors le nombre $|\Pi_\lambda|$ de cellules dans la partition Π_λ satisfait :*

$$\mathbb{E}[|\Pi_\lambda|] = (1+\lambda)^d . \tag{1.114}$$

Une esquisse de la preuve de la Proposition 1.10 figure en Section 2.7, tandis que la preuve complète de ce résultat se trouve dans la Section 2.8.2. L'idée de la preuve consiste à construire une version modifiée du processus de Mondrian, pour laquelle l'espérance du nombre de coupures est inchangée, et dont les partitions sont des "produits" de partitions de Mondrian unidimensionnelles ; on conclut alors en utilisant le lien entre ces dernières et les processus de Poisson sur $[0,1]$.

Comme première application du Théorème 1.15 et de l'équation (1.114), nous obtenons une vitesse de convergence minimax sur la classe des fonctions Lipschitz :

Théorème 1.16 (Théorème 1.16, Chapitre 2). *Supposons la fonction de régression* $g^*(x) := \mathbb{E}[Y|X=x]$ *Lipschitz, et la variance conditionnelle* $\mathrm{Var}(Y|X)$ *bornée. Alors, l'estimateur des forêts de Mondrian* $\widehat{g}_{\lambda,M,n}$ *avec* $M \geqslant 1$ *arbres et de paramètre* $\lambda \asymp n^{1/(d+2)}$ *satisfait*

$$\mathbb{E}\big[(\widehat{g}_{\lambda,M,n}(X) - g^*(X))^2\big] = O(n^{-2/(d+2)}),$$

qui est la vitesse optimale au sens minimax sur cette classe de fonctions (Stone, 1982; Györfi et al., 2002).

En effet, le Théorème 1.15 implique que le diamètre de la cellule $C_\lambda(X)$ du point de test $X \sim P_X$ est d'ordre $O(1/\lambda)$, ce qui permet de contrôler le biais de l'estimateur $\widehat{g}_{\lambda,M,n}$. De plus, la formule (1.114) sur le nombre de cellules permet de contrôler la variance de cet estimateur. Le Théorème 1.16 en découle. Comme signalé précédemment, les variantes de forêts purement aléatoires considérées dans la littérature (telles les forêts centrées) n'atteignent pas la vitesse minimax $O(n^{-2/(d+2)})$ (Breiman, 2004; Biau, 2012; Arlot and Genuer, 2014; Klusowski, 2018), hormis en dimension 1 (Arlot and Genuer, 2014). Cela tient au choix uniforme de la coordonnée de coupure à chaque étape, qui conduit à des cellules déséquilibrées.

Le Théorème 1.16 est valable quel que soit le nombre d'arbres $M \geqslant 1$, et en particulier pour des arbres de Mondrian ($M = 1$). L'avantage des forêts par rapports aux arbres individuels se manifeste dans le cas d'une fonction de régression plus régulière que simplement Lipschitz :

Théorème 1.17 (Théorème 2.3, Chapitre 2). *Supposons que la fonction de régression* g^* *est de classe* $\mathscr{C}^2$, *que* $\mathrm{Var}(Y|X)$ *bornée, et que* X *admet une densité Lipschitz et positive. Alors, l'estimateur des forêts de Mondrian* $\widehat{g}_{\lambda,M,n}$ *avec* $\lambda \asymp n^{1/(d+4)}$ *et* $M \gtrsim n^{2/(d+4)}$ *satisfait*

$$\mathbb{E}\big[(\widehat{g}_{\lambda,M,n}(X) - g^*(X))^2 | X \in B_\varepsilon\big] = O(n^{-4/(d+4)}),$$

où $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$, *pour tout* $\varepsilon > 0$.

Par ailleurs, on vérifie directement pour $d = 1$ qu'un arbre de Mondrian calibré de manière optimale admet un risque d'au moins $\Theta(n^{-2/3})$, correspondant à la vitesse Lipschitz, même dans le cas $\mathscr{C}^2$ (Proposition 2.3). La vitesse en $O(n^{-4/(d+4)})$ correspond à la vitesse minimax d'estimation des fonctions de classe $\mathscr{C}^2$ (Györfi et al., 2002). La restriction à B_ε permet d'éviter un effet de bord, commun aux estimateurs par moyennes locales (Györfi et al., 2002; Wasserman, 2006) ; sans cette restriction, la vitesse obtenue est de $O(n^{-3/(d+3)})$, qui est plus lente mais meilleure que celle du cas Lipschitz.

Le Théorème 1.17 repose sur un contrôle plus précis du biais de forêts avec $M \gg 1$ (c'est-à-dire des forêts infinies) sous l'hypothèse $g^* \in \mathscr{C}^2([0,1]^d)$. Pour obtenir ce contrôle, le seul diamètre de $C_\lambda(x)$ (pour $x \in [0,1]^d$) n'est pas assez précis, nous utilisons donc la loi exacte de $C_\lambda(x)$ décrite par le Théorème 1.15. Par la propriété d'absence de mémoire des lois exponentielles, on déduit notamment de celle-ci la loi de $C_\lambda(x)$ conditionnellement à $z \in C_\lambda(x)$, pour tous $x, z \in [0,1]^d$. Le biais des forêts infinies s'étudie alors en considérant le "noyau" :

$$F_{p,\lambda}(x,z) = e^{-\lambda\|x-z\|_1}\,\mathbb{E}\left[\left\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\right\}^{-1}\right],$$

pour $x, z \in [0,1]^d$, où p désigne la densité de X, et où

$$C_\lambda(x,z) := \prod_{j=1}^{d} \left[(x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1\right]$$

avec $E_{1,L}, E_{1,R}, \ldots, E_{d,L}, E_{d,R}$ des variables i.i.d. de loi $\mathsf{Exp}(1)$.

Forêts de Mondrian en ligne par agrégation (Chapitre 3). Ce chapitre complète le précédent d'un point de vue méthodologique. Nous revenons à la motivation initiale des forêts de Mondrian (Lakshminarayanan et al., 2014), qui est de fournir un algorithme de forêt calculable en ligne, c'est-à-dire pouvant être mis à jour de manière efficace à l'arrivée d'un nouvel échantillon (X_t, Y_t). Comme indiqué précédemment, l'algorithme proposé par Lakshminarayanan et al. (2014) utilise les propriétés des processus de Mondrian afin d'obtenir un tel algorithme en ligne ; cet algorithme admet cependant deux limitations principales. D'une part, il ne s'agit pas d'un algorithme exact, mais d'une procédure qui calcule un estimateur bayésien sur chaque arbre de manière approchée. D'autre part, cette procédure n'admet pas de garantie théorique.

Nous introduisons un algorithme de forêts de Mondrian, constitué par un ensemble d'arbres individuels. Chaque arbre s'obtient en considérant une réalisation d'un processus de Mondrian $\Pi_\infty = (\Pi_\lambda)_{\lambda \in \mathbf{R}^+}$, dont la donnée équivaut à celle d'un arbre binaire complet infini, où le temps, la coordonnée et le seuil de coupure sont indiqués à chaque nœud et aléatoires. L'estimateur associé à un tel arbre infini correspond à l'agrégation à poids exponentiels de tous les prédicteurs constitués par des sous-arbres de décision finis T de Π_∞, avec une prédiction $\widehat{y}_{\mathbf{v}}$ associée à chaque feuille $\mathbf{v}$ de T. D'un point de vue computationnel, un tel estimateur pose a priori deux difficultés :

- il nécessite a priori de tirer une réalisation du processus de Mondrian infini, ce qui n'est pas possible avec des moyens finis ;
- même pour un arbre fini à n feuilles (par exemple avec un point par feuille), le nombre de sous-arbres est exponentiel en n ; le coût d'une agrégation à poids exponentiels naïve avec un poids par sous-arbre est donc prohibitif.

La première difficulté est levée en ne tirant que les coupures nécessaires à séparer les points du jeu de données ; une telle partition se met à jour en utilisant les propriétés de restriction des partitions de Mondrian (Lakshminarayanan et al., 2014). Afin de contourner la seconde difficulté, nous utilisons une loi a priori d'une certaine forme sur les sous-arbres (un processus de branchement), telle que le postérieur soit également de cette forme. Le calcul de l'agrégation à poids exponentiels se "factorise" alors, de sorte qu'il suffit de maintenir un poids par nœud plutôt qu'un poids par sous-arbre, ce qui permet une réduction exponentielle de la complexité. La procédure d'agrégation correspond alors à celle des *arbres experts* introduite par Helmbold and Schapire (1997), intimement liée à l'algorithme du *Context Tree Weighting* (Willems et al., 1995; Willems, 1998) de compression de données, qui repose sur la même factorisation.

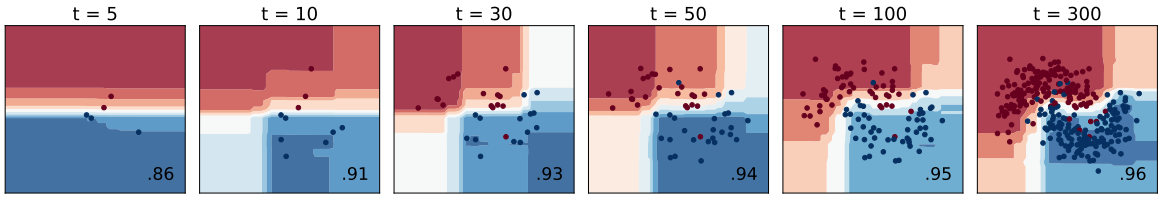

Figure 1.4: Évolution de la fonction de décision en classification, pour l'algorithme AMF introduit, en fonction du nombre de points (Chapitre 3).

L'estimateur $\widehat{g}_n$ ainsi obtenu satisfait la garantie de risque suivante, par exemple en régression bornée avec $\mathcal{Y} = [-B, B]$ (Exemple 1.4) : si Π est une réalisation aléatoire du processus

de Mondrian (c'est-à-dire un arbre infini, avec une coupure associée à chaque nœud), alors

$$\mathbb{E}[R(\widehat{g}_n)] \leqslant \mathbb{E}_\Pi\Big[\inf_{\mathcal{T},g_\mathcal{T}}\Big\{R(g_\mathcal{T}) + \frac{4B^2(|\mathcal{T}|+1)\log n}{n}\Big\}\Big] \tag{1.115}$$

où l'infimum porte sur les sous-arbres finis $\mathcal{T}$ (à $|\mathcal{T}|$ feuilles) de Π et les fonctions $g_\mathcal{T} : [0,1]^d \to \mathbf{R}$ constantes sur les cellules de $\mathcal{T}$, et l'espérance du membre de droite porte sur le tirage de Π. Ce résultat découle du Corollaire 3.2 et de la conversion online-to-batch. En particulier, en considérant les sous-arbres $\mathcal{T}$ de la forme Π_λ ($\lambda > 0$), et en utilisant les résultats du Chapitre 2, il est possible d'obtenir des vitesses de convergence non paramétriques sur les classes de régularité Hölder (Théorème 3.2). Nous renvoyons au Chapitre 3 pour des expériences numériques sur cette procédure ; bien que non compétitive avec les forêts de Breiman pour des tâches de classification pures, elle offre une bonne calibration des probabilités des différentes classes (mesurée par la perte logarithmique ou l'AUC), tout en étant implémentable de manière séquentielle et efficace, et moins sensible au nombre d'arbres.

Par rapport à l'estimateur des forêts de type PRF considéré au Chapitre 2, la régularisation n'est plus assurée par le paramètre λ de complexité des arbres (les arbres considérés ne sont pas élagués), mais par l'agrégation à poids exponentiels sur tous les sous-arbres. L'estimateur ainsi obtenu est donc compétitif avec le meilleur élagage $\mathcal{T}$ de l'arbre infini ; en pratique, cela permet de sélectionner des partitions plus adaptatives, qui approchent mieux la fonction de régression lorsque celle-ci varie davantage dans certaines régions que d'autres. Cependant, contrairement au cas des forêts de Mondrian de type PRF, l'estimateur n'admet a priori pas de vitesses minimax dans le cas $\mathscr{C}^2$: cela tient au fait que la complexité des arbres est optimisée individuellement pour chaque arbre. Cet estimateur peut être vu comme un algorithme de plus proches voisins, mais où le nombre de voisins dépend du point considéré et est choisi de manière adaptative.

1.6 Annexe technique

Dans cette section figurent des définitions et résultats techniques mentionnés ou utilisés dans cette introduction.

1.6.1 Variables sous-Gaussiennes et inégalités de concentration

Dans cette section, nous collectons quelques définitions et résultats élémentaires de concentration de variables i.i.d. ; nous renvoyons à Boucheron et al. (2013) pour davantage de détails sur ce sujet.

Nous commençons par rappeler l'inégalité de Hoeffding (Hoeffding, 1963), qui est utilisée dans la preuve de la Proposition 1.5.

Lemme 1.2 (Boucheron et al., 2013, Lemme 2.2)**.** *Soit X une variable aléatoire à valeurs dans $[0,1]$. Alors, pour tout $\lambda \in \mathbf{R}$, $\log \mathbb{E}[e^{\lambda X}] \leqslant \lambda \mathbb{E}[X] + \lambda^2/8$.*

Proof. Soit $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$ pour tout $\lambda \in \mathbf{R}$. Définissons pour $\lambda \in \mathbf{R}$ la mesure de probabilité $\mathbb{P}_\lambda := \{e^{\lambda X}/\mathbb{E}[e^{\lambda X}]\} \cdot \mathbb{P}$, et notons $\mathbb{E}_\lambda[Z]$, $\mathrm{Var}_\lambda(Z)$ l'espérance et la variance d'une variable aléatoire Z selon $\mathbb{P}_\lambda$. Par dérivation sous le signe intégral (en utilisant le fait que $0 \leqslant X \leqslant 1$), la fonction ψ est deux fois dérivable, et vérifie $\psi'(\lambda) = \mathbb{E}[Xe^{\lambda X}]/\mathbb{E}[e^{\lambda X}] = \mathbb{E}_\lambda[X]$ et $\psi''(\lambda) = \mathrm{Var}_\lambda(X)$.

Or, la variable X appartient à $[0,1]$ $\mathbb{P}$-presque sûrement, donc $\mathbb{P}_\lambda$-presque sûrement, de sorte que $\psi''(\lambda) = \mathrm{Var}_\lambda(X) \leqslant 1/4$. L'inégalité de Taylor montre alors que, pour tout $\lambda \in \mathbf{R}$, $\psi(\lambda) \leqslant \psi(0) + \psi'(0) \cdot \lambda + (1/4) \cdot \lambda^2/2$, ce qui établit le Lemme 1.2 puisque $\psi(0) = 0$ et $\psi'(0) = \mathbb{E}_0[X] = \mathbb{E}[X]$. □

L'inégalité de Hoeffding équivaut à dire que $X - \mathbb{E}[X]$ est $1/4$-sous-Gaussienne, au sens de la définition suivante :

Définition 1.6 (Variable sous-Gaussienne)**.** Soit $\sigma^2 > 0$. Une variable aléatoire réelle centrée X est dite *σ^2-sous-Gaussienne* si, pour tout $\lambda \in \mathbf{R}$, $\mathbb{E}[e^{\lambda X}] \leqslant e^{\sigma^2\lambda^2/2}$.

Le terme *sous-Gaussien* vient du fait que, si $X \sim \mathcal{N}(0, \sigma^2)$, alors $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$ pour tout $\lambda \in \mathbf{R}$. Si X est σ^2-sous-Gaussienne, alors pour tout $t \geqslant 0$,

$$\mathbb{P}(X \geqslant t) \vee \mathbb{P}(X \leqslant -t) \leqslant e^{-t^2/(2\sigma^2)}\,, \tag{1.116}$$

en appliquant l'inégalité de Markov aux variables $e^{\lambda X}$ et $e^{-\lambda X}$ et en optimisant le choix de λ. En outre, il existe une constante universelle C telle que pour tout $p \geqslant 1$:

$$\|X\|_{L^p} = \mathbb{E}[|X|^p]^{1/p} \leqslant C\sigma\sqrt{p}\,. \tag{1.117}$$

Réciproquement, pour une variable centrée X, les inégalités (1.116) et (1.117) impliquent que X est $c\sigma^2$-sous-Gaussiennes pour une certaine constante absolue c (Vershynin, 2012; Boucheron et al., 2013). La condition (1.117) permet d'étendre la définition de variables sous-Gaussiennes aux variables non centrées, et revient à dire que $X - \mathbb{E}[X]$ est $C'\sigma^2$-sous-Gaussienne et que $|\mathbb{E}[X]| \leqslant C''\sigma$.

Plus généralement, on dit qu'un vecteur aléatoire X à valeurs dans $\mathbf{R}^d$ est σ^2-sous-Gaussien ($\sigma > 0$) si $\langle \theta, X \rangle$ est σ^2-sous-Gaussien pour tout $\theta \in S^{d-1}$.

Si $X_1, \dots, X_M$ sont des variables centrées σ^2-sous-Gaussiennes, alors pour tout $\lambda > 0$:

$$\mathbb{E}[e^{\lambda \max(X_1,\dots,X_M)}] = \mathbb{E}[\max(e^{\lambda X_1}, \dots, e^{\lambda X_M})] \leqslant \sum_{i=1}^{M} \mathbb{E}[e^{\lambda X_i}] \leqslant M e^{\sigma^2\lambda^2/2}. \tag{1.118}$$

Par convexité de la fonction exponentielle, on en déduit que $\exp(\lambda \mathbb{E}[\max(X_1, \dots, X_M)]) \leqslant M e^{\sigma^2\lambda^2/2}$, c'est-à-dire (en optimisant λ) que $\mathbb{E}[\max(X_1, \dots, X_M)] \leqslant \sigma\sqrt{2\log M}$. De même, l'inégalité de Markov implique que, pour tout $t > 0$,

$$\mathbb{P}\big(\max(X_1, \dots, X_M) \geqslant \sigma\sqrt{2\log M} + \sigma t\big) \leqslant e^{-t^2/2}. \tag{1.119}$$

Enfin, soient $X_1, \dots, X_n$ des variables indépendantes centrées et σ^2-sous-Gaussiennes. Il découle de la Définition 1.6 que la moyenne $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ est σ^2/n-sous-Gaussienne. En particulier, si $(X_{i,j})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant M}$ sont des variables centrées à valeurs dans $[-1, 1]$, telles que les lignes $X_{i,\cdot} = (X_{i,j})_{1\leqslant j\leqslant M}$ sont indépendantes, alors en notant $\bar{X}_i := n^{-1}\sum_{i=1}^n X_{i,j}$, on a

$$\mathbb{E}\big[\max_{1\leqslant j\leqslant M} \bar{X}_j\big] \leqslant \sqrt{\frac{2\log M}{n}}, \quad \text{et} \quad \mathbb{P}\Big(\max_{1\leqslant j\leqslant M} \bar{X}_j \geqslant \sqrt{\frac{2\log M}{n}} + \sqrt{\frac{2t}{n}}\Big) \leqslant e^{-t} \tag{1.120}$$

pour tout $t > 0$. Ce résultat s'applique notamment à $X_{i,j} = \ell(f_j, Z_i) - R(f_j)$, où $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$ est une fonction de perte, $\mathcal{F} = \{f_1, \dots, f_M\}$ est une classe finie et $Z_1, \dots, Z_n$ sont des observations i.i.d., et permet dans ce cas de contrôler l'erreur de généralisation et l'excès de risque (Section 1.1.4).

1.6.2 Entropie relative et dualité

Nous décrivons à présent l'entropie relative ainsi qu'un résultat fondamental de dualité, qui est utilisé à plusieurs reprises dans ce texte.

Définition 1.7. Soit Θ un espace mesurable, et ρ, π deux mesures de probabilité sur Θ. L'*entropie relative*, ou *divergence de Kullback-Leibler* entre ρ et π, notée $\mathrm{KL}(\rho, \pi)$, est définie par

$$\mathrm{KL}(\rho, \pi) := \int_\Theta \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho \tag{1.121}$$

lorsque ρ est absolument continue par rapport à π, et $+\infty$ dans le cas contraire.

L'entropie relative $\mathrm{KL}(\rho, \pi)$ mesure la différence entre les mesures π et ρ, ou la qualité de π en tant qu'approximation de ρ ; cette quantité n'est pas symétrique en ρ, π. Plus précisément, l'entropie relative constitue l'excès de risque de π par rapport à ρ en estimation de densité avec perte logarithmique, lorsque la vraie loi est ρ (Exemple 1.2). Cette quantité est toujours positive, avec égalité si et seulement si $\pi = \rho$; cela se vérifie en appliquant l'inégalité de Jensen à la fonction $-\log$ et le fait que $\rho(\mathrm{d}\rho/\mathrm{d}\pi = 0) = 0$:

$$\begin{aligned}\mathrm{KL}(\rho, \pi) &= \int_\Theta -\log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\rho}\Big)\mathbf{1}\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} > 0\Big)\mathrm{d}\rho \\ &\geqslant -\log\Big(\int_\Theta \frac{\mathrm{d}\pi}{\mathrm{d}\rho}\mathbf{1}\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} > 0\Big)\mathrm{d}\rho\Big) \geqslant -\log\Big(\int_\Theta \mathrm{d}\pi\Big) = 0\end{aligned}$$

avec égalité si et seulement si $\mathrm{d}\rho/\mathrm{d}\pi \equiv 1$, c'est-à-dire $\rho = \pi$.

Notons également, pour toute mesure finie π sur Θ et toute fonction $f : \Theta \to \mathbf{R}$ mesurable, $\langle \pi, f \rangle := \int_\Theta f \mathrm{d}\pi$ dès lors que l'intégrale est bien définie dans $\mathbf{R} \cup \{+\infty, -\infty\}$. Pour toute fonction positive $h : \Theta \to \mathbf{R}^+$ telle que $\langle \pi, h \rangle \in (0, +\infty)$, notons $\pi_h := \frac{h}{\langle \pi, h \rangle} \cdot \pi$ la mesure de probabilité sur Θ de densité $h/\langle \pi, h \rangle$ par rapport à π.

Théorème 1.18 (Donsker-Varadhan). *Soit π, ρ deux mesures de probabilité sur Θ ; pour toute fonction bornée $f : \Theta \to \mathbf{R}$,*

$$\log \left(\int_\Theta \exp(f) \, \mathrm{d}\pi \right) + \mathrm{KL}(\rho, \pi) - \int_\Theta f \mathrm{d}\rho = \mathrm{KL}(\rho, \pi_{\exp(f)}) \, . \tag{1.122}$$

En particulier, pour toute fonction mesurable $f : \Theta \to \mathbf{R}$,

$$\log \langle \pi, \exp(f) \rangle = \sup_\rho \left\{ \langle \rho, f \rangle - \mathrm{KL}(\rho, \pi) \right\} , \tag{1.123}$$

le supremum étant atteint pour $\rho = \pi_{\exp(f)}$ lorsque le terme de gauche est fini.

Proof. Commençons par supposer f bornée, de sorte que les intégrales sont bien définies. Si ρ n'est pas absolument continue par rapport à π, il ne l'est pas non plus par rapport à $\pi_{\exp(f)} = [\exp(f)/\langle \pi, \exp(f) \rangle]\pi$, et donc les membres de gauche et de droite de (1.122) sont infinis. Si ρ est absolument continue par rapport à π, alors $\rho = [\mathrm{d}\rho/\mathrm{d}\pi]\pi = [\mathrm{d}\rho/\mathrm{d}\pi] \times [\exp(f)/\langle \pi, \exp(f) \rangle]^{-1} \pi_{\exp(f)}$, de sorte que

$$\begin{aligned} \mathrm{KL}(\rho, \pi_{\exp(f)}) &= \int_\Theta \log \Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \cdot \frac{\langle \pi, \exp(f) \rangle}{\exp(f)} \Big) \mathrm{d}\rho \\ &= \int_\Theta \log \Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \Big) \mathrm{d}\rho + \log \big(\langle \pi, \exp(f) \rangle \big) - \int_\Theta f \mathrm{d}\rho \, , \end{aligned}$$

qui coïncide précisément avec (1.122). Il en découle que, pour tous ρ, π, f (avec f bornée), par positivité de l'entropie relative,

$$\langle \rho, f \rangle - \mathrm{KL}(\rho, \pi) = \log \langle \pi, \exp(f) \rangle - \mathrm{KL}(\rho, \pi_{\exp(f)}) \leqslant \log \langle \pi, \exp(f) \rangle \, ,$$

avec égalité si et seulement si $\rho = \pi_{\exp(f)}$. Le résultat dans le cas général où f n'est pas bornée s'en déduit en considérant $f^B := \min(f, B)$ avec $B \to +\infty$. □

Le Théorème 1.18 affirme que, pour toute mesure de probabilité π, l'entropie relative $\rho \mapsto \mathrm{KL}(\rho, \pi)$ (définie sur l'espace des mesures de probabilité sur Θ, et à valeurs dans $[0, +\infty]$) est la transformée de Fenchel-Legendre (Boyd and Vandenberghe, 2004) de la transformée de Laplace logarithmique $f \mapsto \log \langle \pi, \exp(f) \rangle \in \mathbf{R} \cup \{+\infty\}$ (définie sur les fonctions $\Theta \to \mathbf{R}$). Ce résultat (ou une variante "inversée" de (1.123), qui découle de la même manière de (1.122)) est parfois appelé *formule variationnelle de Donsker-Varadhan.*

1.6.3 Convexité et forte convexité

Nous indiquons dans cette section la définition de la (forte) convexité (Boyd and Vandenberghe, 2004), ainsi que le lien entre la forte convexité et la stabilité des minima.

Définition 1.8 (Convexité, forte convexité)**.** Soit E un espace vectoriel normé (de norme notée $\|\cdot\|$), et $f : E \to \mathbf{R} \cup \{+\infty\}$. On dit que f est *convexe* si $\Omega := \{x \in E : f(x) \in \mathbf{R}\}$ est non vide et convexe, et si pour tous $x, x' \in \Omega$ et $t \in [0,1]$,

$$f\big(tx + (1-t)x'\big) \leqslant t \cdot f(x) + (1-t) \cdot f(x')\,.$$

Supposons également f différentiable sur Ω, et notons $\nabla f(x) \in E^*$ (où E^* est le dual de E) le gradient de f en x. Pour tout $\lambda > 0$, on dit que f est *λ-fortement convexe* si, pour tous $x, x' \in \Omega$,

$$f(x') - f(x) - \langle \nabla f(x), x' - x\rangle \geqslant \frac{\lambda}{2}\|x - x'\|^2\,. \tag{1.124}$$

La convexité est équivalente à la condition de λ-forte convexité avec $\lambda = 0$. On vérifie directement que la fonction $x \mapsto \|x\|^2/2$ est 1-fortement convexe sur $\mathbf{R}^d$ (la condition (1.124) est alors une égalité), et que la somme d'une fonction λ-fortement convexe et d'une fonction convexe est λ-fortement convexe (Boyd and Vandenberghe, 2004). Notons que, si $f : \Omega \to \mathbf{R}$ (où $\Omega \subset E$ est convexe) est différentiable et $x^* \in \arg\min f$, alors $\nabla f(x^*) = 0$. Réciproquement, si f est convexe et $\nabla f(x^*) = 0$, alors x^* est un minimiseur de f. Si de plus f est λ-fortement convexe, alors pour tout $x \in \Omega$ (puisque $\nabla f(x^*) = 0$) :

$$f(x) - f(x^*) \geqslant \frac{\lambda}{2}\|x - x^*\|^2\,. \tag{1.125}$$

De plus, en inversant x et x' dans l'inégalité (1.124) et en sommant l'inégalité obtenue avec (1.125), on obtient, pour tous $x, x' \in \Omega$,

$$\langle \nabla f(x') - \nabla f(x), x' - x\rangle \geqslant \lambda\|x - x'\|^2\,. \tag{1.126}$$

De ce qui précède, nous déduisons le résultat de stabilité suivant (rappelons qu'une fonction $f : \mathbf{R}^d \to \mathbf{R}$ est dite *L-Lipschitz* si $|f(x') - f(x)| \leqslant L\|x' - x\|$ pour tous $x, x' \in \mathbf{R}^d$) :

Lemme 1.3. *Soit $F : \mathbf{R}^d \to \mathbf{R}$ une fonction λ-fortement convexe, et $f : \mathbf{R}^d \to \mathbf{R}$ une fonction L-Lipschitz. Soient*[21] *$x^* = \arg\min F$, et $\widetilde{x} \in \arg\min(F + f)$. Alors*

$$\|x^* - \widetilde{x}\| \leqslant \frac{L}{\lambda}\,, \qquad f(x^*) - f(\widetilde{x}) \leqslant \frac{L^2}{\lambda}\,.$$

Proof. L'inégalité $(F + f)(\widetilde{x}) \leqslant (F + f)(x^*)$ et le caractère L-Lipschitz de f impliquent que

$$F(\widetilde{x}) - F(x^*) \leqslant f(x^*) - f(\widetilde{x}) \leqslant L\|\widetilde{x} - x^*\|\,.$$

De plus, par λ-forte convexité de F et par l'inégalité (1.125), on a $F(\widetilde{x}) - F(x^*) \geqslant \lambda\|\widetilde{x} - x^*\|^2/2$. Il en découle que $\|\widetilde{x} - x^*\| \leqslant 2L/\lambda$ et donc $f(x^*) - f(\widetilde{x}) \leqslant 2L^2/\lambda$. Cela établit la borne du Lemme 1.3, au facteur 2 près.

Pour éliminer le facteur 2, procédons comme suit. Pour toute fonction $g : \mathbf{R}^d \to \mathbf{R}$, soit $\phi_g : [0,1] \to \mathbf{R}$ la fonction définie par $\phi_g(t) = g(\widetilde{x} + t(x^* - \widetilde{x}))$. Par définition de $\widetilde{x}$, $\phi_{F+f} = \phi_F + \phi_f$ atteint son minimum en 0, de sorte que, pour tout $t \in (0,1]$,

$$0 \leqslant \frac{\phi_{F+f}(t) - \phi_{F+f}(0)}{t} = \Delta_F(t) + \Delta_f(t)\,, \tag{1.127}$$

[21] L'existence de $x^*, \widetilde{x}$ provient du fait que $F, F + f$ tendent vers $+\infty$ en $+\infty$ (par forte convexité de F et en utilisant le fait que f est Lipschitz), et que ces deux fonctions sont continues (une fonction convexe sur $\mathbf{R}^d$ étant continue). L'unicité de x^* provient de la forte convexité de F et de (1.125) ; $\widetilde{x}$ n'est pas nécessairement unique (il l'est cependant si f est convexe), mais le résultat s'applique à tout choix possible de $\widetilde{x}$.

où l'on note $\Delta_g(t) = (\phi_g(t) - \phi_g(0))/t$. Or, la fonction f est L-Lipschitz, de sorte que $\phi_f(t) - \phi_f(0) \leqslant L\|t \cdot (\widetilde{x} - x^*)\|$, c'est-à-dire $\Delta_f(t) \leqslant L\|\widetilde{x} - x^*\|$ pour tout $t \in (0,1)$. En outre, la fonction F est différentiable, donc ϕ_F également ; ainsi, lorsque $t \to 0^+$, $\Delta_F(t) \to \phi'_F(0) = \langle \nabla F(\widetilde{x}), x^* - \widetilde{x} \rangle$. Or, la λ-forte convexité de F implique, par l'inégalité (1.126), que

$$\langle \nabla F(\widetilde{x}), x^* - \widetilde{x} \rangle = -\langle \nabla F(\widetilde{x}) - \nabla F(x^*), \widetilde{x} - x^* \rangle \leqslant -\lambda \|\widetilde{x} - x^*\|^2 ,$$

où l'on a utilisé que $\nabla F(x^*) = 0$. Ainsi, en prenant $t \to 0^+$, l'inégalité (1.127) implique que $0 \leqslant -\lambda\|\widetilde{x} - x^*\|^2 + L\|\widetilde{x} - x^*\|$, c'est-à-dire $\|\widetilde{x} - x^*\| \leqslant L/\lambda$ et donc $f(x^*) - f(\widetilde{x}) \leqslant L^2/\lambda$ comme annoncé. □

1.6.4 Regret de la descente de gradient en ligne

Dans cette annexe, nous établissons une borne de regret pour l'algorithme de *descente de gradient en ligne* (en anglais *Online gradient descent*, OGD, Zinkevich, 2003). Cet algorithme coïncide avec l'algorithme de descente de gradient stochastique décrit dans la Proposition 1.2, mais considéré comme algorithme d'optimisation en ligne. Le cadre du problème est ici celui de l'optimisation convexe en ligne (voir la Section 1.2.1)

Proposition 1.11 (Zinkevich, 2003)**.** *Soit Θ une partie convexe fermée de $\mathbf{R}^d$. Supposons que, pour tout $t = 1, \dots, n$, la fonction $\ell_t : \Theta \to \mathbf{R}$ est convexe, différentiable et L-Lipschitz. Considérons l'algorithme de* descente de gradient en ligne projetée :

- $\widehat{\theta}_1 := \theta_1 \in \Theta$ *fixe ;*
- *pour $t = 1, \dots, n$, $\widehat{\theta}_{t+1} := \mathrm{proj}_\Theta(\widehat{\theta}_t - \eta \nabla \ell_t(\widehat{\theta}_t))$.*

Soit $B > 0$; posons $\eta = B/(L\sqrt{n})$, et notons $\Theta_B := \{\theta \in \Theta : \|\theta - \theta_1\| \leqslant B\}$. Alors, on a la borne de regret suivante :

$$\sum_{t=1}^n \ell_t(\widehat{\theta}_t) - \inf_{\theta \in \Theta_B} \sum_{t=1}^n \ell_t(\theta) \leqslant BL\sqrt{n} . \tag{1.128}$$

Proof. Posons $h_t := \nabla \ell_t(\widehat{\theta}_t)$ pour tout $t = 1, \dots, n$; puisque ℓ_t est L-Lipschitz, on a $\|h_t\| \leqslant L$. De plus, l'inégalité (1.27) implique qu'il suffit de contrôler le regret sur la suite de pertes linéaires $\langle h_t, \cdot \rangle$. Comme $\widehat{\theta}_{t+1} = \mathrm{proj}_\Theta(\widehat{\theta}_t - \eta h_t)$, on a pour tout $\theta \in \Theta$:

$$\|\widehat{\theta}_{t+1} - \theta\|^2 \leqslant \|\widehat{\theta}_t - \eta h_t\|^2 = \|\widehat{\theta}_t - \theta\|^2 - 2\eta \langle h_t, \widehat{\theta}_t - \theta \rangle + \eta^2 \|h_t\|^2$$

de sorte que

$$\langle h_t, \widehat{\theta}_t - \theta \rangle \leqslant \frac{1}{2\eta}\big(\|\widehat{\theta}_t - \theta\|^2 - \|\widehat{\theta}_{t+1} - \theta\|^2\big) + \frac{\eta}{2}\|h_t\|^2 .$$

En sommant l'inégalité précédente sur $t = 1, \dots, n$, on obtient pour tout $\theta \in \Theta_B$:

$$\sum_{t=1}^n \ell_t(\widehat{\theta}_t) - \sum_{t=1}^n \ell_t(\theta) \leqslant \sum_{t=1}^n \langle h_t, \widehat{\theta}_t - \theta \rangle \leqslant \frac{\|\theta_1 - \theta\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^n \|h_t\|^2 \leqslant \frac{B^2}{2\eta} + \frac{\eta L^2 n}{2} ,$$

qui vaut $BL\sqrt{n}$ pour $\eta = B/(L\sqrt{n})$. □