Bornes de généralisation

Thierry Bazier-Matte

7 avril 2017

1 Garanties statistiques

La section précédente été dédiée à l'approche algorithmique du problème : comment, donnés un ensemble d'entraînement et un espace de décision $\boldsymbol{Q}$, une fonction de décision $\hat{q} : \boldsymbol{Q} \to \mathscr{R}$ permettant de prescrire un investissment pouvait être déterminée. Cette section sera consacrée aux garanties statistiques de cette solution. Dans un premier temps, une étude de la stabilité de l'algorithme d'optimisation permettra de dériver une borne de généralisation sur la performance hors-échantillon (Section 1.1). Par la suite, le problème sera approché d'un point probabiliste (en terme de variables aléatoires) afin de comparer les performances de la décision optimale d'investissement sur M par rapport à la décision empirique (Section 1.2). Enfin, la Section 1.3 portera sur l'influence de la dimensionalité de l'espace $\boldsymbol{Q}$ sur la qualité des bornes alors obtenues, et don

Les bornes qui seront dérivées n'auront de signification qu'en terme d'*util*, c'est à dire la dimension de $u(r)$ pour un certain rendement. Comme cette notion n'a en soi aucune signification tangible, un théorème sera finalement introduit afin d'obtenir pour chacune des bornes une version sous forme de rendement équivalent.

Hypothèses et discussion Certaines hypothèses devront d'abord être formulées afin d'être en mesure d'obtenir des résultats pertinents : ce sera en fait le prix à payer pour l'absence de contraintes sur la forme de la distribution M, notamment concernant par exemple sa covariance ou la forme de ses moments d'ordre supérieurs.

Hypothèse 1. *L'amplitude de similarité d'une observation est bornée : pour tout* $x \in \boldsymbol{X}$, $\kappa(x, x) \leq \xi^2$.

Hypothèse 2. *Le rendement aléatoire est borné :* $|R| \leq \bar{r}$.

Hypothèse 3. *Un investisseur est doté d'une fonction d'utilité* u *concave, monotone et standardisée, c'est-à-dire que* $u(0) = 0$ *et* $\partial u(0) \ni 1$ [1]. *De plus,* u *est défini sur*

1. Ici, $\partial u(r)$ signifie l'ensemble des sur-gradients de u. Dans le cas dérivable, cela revient à la notion de dérivée. Dans le cas simplement continu, $\partial u(r)$ est l'ensemble des fonctions affines "touchant" à $u(r)$ et

l'ensemble de $\mathscr{R}$. Enfin, u est γ-Lipschitz, c'est-à-dire que pour tout $r_1, r_2 \in \mathscr{R}$, $|u(r_1) - u(r_2)| \leq \gamma |r_1 - r_2|$.

Avant d'aller plus loin, il convient de discuter de la plausiblité de ces contraintes. Cependant, compte tenu de l'aspect central de la première hypothèse, une discussion approfondie ne sera abordée qu'à la section 1.3.

Pour ce qui est de la seconde hypothèse, si on définit les rendements selon l'interprétation usuelle d'un changement de prix p, *i.e.*, $r = \Delta p/p$, on constatera que r est nécessairement borné par 0. De plus, selon la période de temps pendant laquelle Δp est mesuré, il y a forcément moyen de limiter l'accroissement dans le prix, pour autant que Δt soit suffisament court.

La troisième hypothèse est davantage contraignante. Elle exclut d'emblée plusieurs fonctions d'utilité courantes ; par exemple l'utilité logarithmique et racine carrée puisqu'elles ne sont définies que pour $\mathscr{R}_+$. Une utilité quadratique, comme celle de Markowitz est également inadmissible puisqu'elle est non-monotone. Les utilités de forme exponentielle inverse $u(r) = \mu(-\exp(-r/\mu) + 1)$ quant à elles violent la condition Lipschitz. On peut cependant définir une utilité exponentielle *à pente contrôlée*, c'est à dire dont la pente devient constante lorsque $r \leq r_0$. Par contre, une utilité qui serait définie par morceaux linéaires est parfaitement acceptable. Par ailleurs, on considérera souvent l'utilité *neutre au risque* $\mathbf{1} : r \mapsto r$ comme un cas limite à l'ensemble des fonctions d'utilité admissibles.

Exemple 1. FONCTIONS D'UTILITÉ LIPSCHITZ – Cette famille de fonctions est paramétrée par $\mu > 0$ et définie par $u(r) = r$ si $r \leq 0$ et $u(r) = \mu(1 - \exp(-r/\mu))$ dans le cas contraire (Fig. 1 (p. 3)). On constate les deux cas limites : lorsque $\mu \to \infty$, u se comporte comme une utilité neutre au risque, alors que $\mu \to 0$ signifie une attitude d'indifférence aux rendements supérieurs à 0. Par ailleurs, toute fonction exponentielle Lipschitz admet un coefficient Lipschitz unique $\gamma = 1$. Leur simplicité et leur expressivité feront de cette classe d'utilités l'essentiel de l'analyse numérique de ce mémoire.

Exemple 2. COPULES GAUSSIENNES – Bien que les résultats présentés dans ce travail concernent toute forme de distribution de marché respectant les hypothèses énoncées ci-haut, les simulations numériques qui seront offertes à titre d'exemples seront toutes tirées d'une distribution multivariée liée par une copule gaussienne $\mathcal{C}(\Sigma)$ où $\Sigma \in \mathscr{R}^{p+1 \times p+1}$ est la matrice de corrélation des $p+1$ *marges* de M (p variable d'information X_j et une variable de rendement R). Cette flexibilité permet notamment de contraster les résultats lorsque M est strictement bornée ou présente au contraire des queues lourdes, tout en conservant une corrélation identique.

supérieures à $u(r)$ pour tout r du domaine). Bien qu'il s'agisse d'un ensemble, la situation désigne souvent un sur-gradient optimal par rapport aux autres.

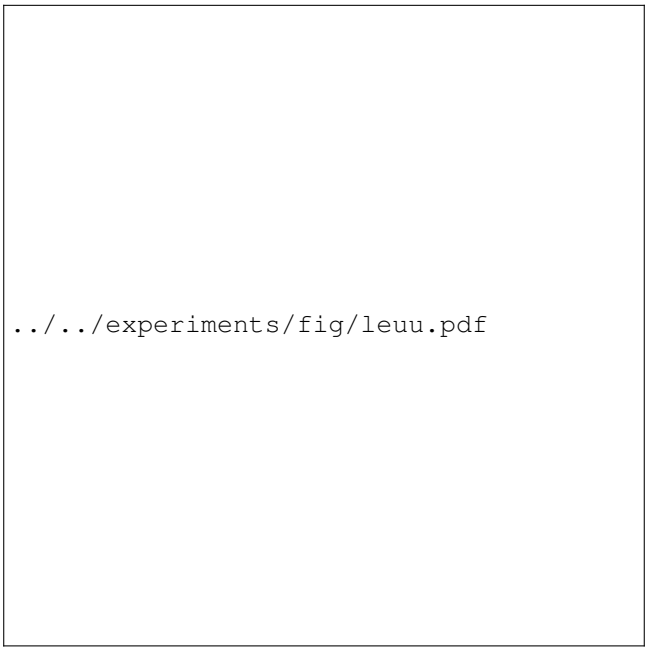

FIGURE 1 – Fonctions d'utilité Lipschitz

1.1 Bornes de généralisation

Exposition du problème Soit $\boldsymbol{Q}$ un espace de Hilbert à noyau reproduisant induit par κ et soit un ensemble d'entraînement $\mathcal{S}_n = \{(x_i, r_i)\}_{i=1}^n \sim M^n$ échantilloné à partir de la distribution de marché. Alors on peut définir *l'algorithme de décision* $\mathcal{Q} : \boldsymbol{M}^n \to \boldsymbol{Q}$ par

$$\mathcal{Q}(\mathcal{S}_n) = \arg\max_{q \in \boldsymbol{Q}} \left\{ \widehat{\boldsymbol{EU}}(\mathcal{S}_n, q) - \lambda \|q\|^2 \right\}. \tag{1}$$

Comme on l'a vu, résoudre (1) est aussi équivalent à

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i(\alpha^T \phi)(x_i)) - \lambda \alpha^T K \alpha, \tag{2}$$

où $\phi : \mathscr{R}^p \to \mathscr{R}^n$ le vecteur d'application induit par la matrice d'information Ξ. La relation $q = \alpha^T \phi$ permet de passer d'une représentation à l'autre.

La question qui se pose naturellement est de savoir dans quelle mesure une fonction de décision $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ est capable d'offrir à un investisseur une utilité espérée comparable à celle qu'il aurait observée au sein de l'ensemble d'entraînement. Il serait aussi souhaitable qu'une telle garantie soit indépendante de l'ensemble d'entraînement

$\mathcal{S}_n$. Autrement dit, on cherche à déterminer une borne probabiliste $\hat{\Omega}_u$ sur l'erreur de généralisation de $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ valide pour tout $\mathcal{S}_n \sim M^n$:

$$\hat{\zeta}_u(\mathcal{S}_n) \leq \hat{\Omega}_u(n, \ldots), \tag{3}$$

où

$$\hat{\zeta}_u(\mathcal{S}_n) = \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \tag{4}$$

représente l'erreur de généralisation.

Bien que ces résultats soient intéressants d'un point de vue théorique, on veut d'un point de vue pratique pouvoir garantir au détenteur du portefeuille un intervalle de confiance sur l'équivalent certain du portefeuille. On cherchera donc une borne $\hat{\Omega}_e$ telle que

$$\boldsymbol{CE}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) \geq \widehat{\boldsymbol{CE}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \hat{\Omega}_e(n, \ldots). \tag{5}$$

Intuition et éléments de preuve En fait, la motivation derrière ces hypothèses est la suivante : combinées à l'élément de régularisation, elles parviennent d'une part à borner la perte que peut entraîner la prise de décision dans le pire cas et d'autre part à borner la différence entre deux fonctions de décision entraînées sur des ensembles à peu près identiques.

Considérons deux ensembles d'entraînement : $\mathcal{S}_n \sim M^n$ et $\mathcal{S}'_n$, où $\mathcal{S}'_n$ ne diffère de $\mathcal{S}_n$ que par un seul point (par exemple le j-ème point serait rééchantilloné de la distribution de marché M). De l'algorithme $\mathcal{Q}$ on dérivera alors deux décisions : $\hat{q}$ et $\hat{q}'$. Pour n suffisament grand, on peut alors s'attendre à ce que l'utilité dérivée de ces deux décisions soit relativement proche, et ce, pour toute observation. On aurait alors une borne $\beta(n)$ telle que pour tout $(x, r) \sim M$,

$$|u(r\,\hat{q}(x)) - u(r\,\hat{q}'(x))| \leq \beta. \tag{6}$$

C'est ce qu'on appelle dans la littérature la *stabilité algorithmique*. La plupart des algorithmes régularisés classiques disposent par ailleurs d'une telle stabilité. En particulier, le terme de régularisation $\lambda\|q\|^2$, combiné à la continuité Lipschitz de u font en sorte que $\beta = (n^{-1})$. Par le Lemme 1, p. 12 (une application directe du théorème de Bousquet), on obtient effectivement

$$\beta \leq \frac{\gamma^2 \bar{r}^2 \xi^2}{2\lambda n}. \tag{7}$$

Dotée de cette stabilité de $\mathcal{Q}$, la différence dans l'erreur de généralisation de $\mathcal{S}_n$ et $\mathcal{S}'_n$ peut alors être bornée :

$$|\hat{\zeta}(\mathcal{S}_n) - \hat{\zeta}(\mathcal{S}'_n)| = |\,\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}') + \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \tag{8}$$

$$\leq |\,\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| + |\,\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')|. \tag{9}$$

Or, par le théorème de Jensen appliqué à la fonction valeur absolue, on obtient du premier terme que

$$|\,\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| = |\,\boldsymbol{E}(u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)))| \tag{10}$$

$$\leq \boldsymbol{E}(|u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)|) \tag{11}$$
$$\leq \beta, \tag{12}$$

par définiton de la stabilité. Quant au deuxième terme de (9) on peut le borner de la même façon :

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \tag{13}$$
$$= n^{-1} \left| \sum_{i=1}^{n} \mathbb{I}_{i \neq j} u(r_i\, \hat{q}(x_i)) + u(r_j\, \hat{q}(x_j)) - \sum_{i=1}^{n} \mathbb{I}_{i \neq j} u(r_i\, \hat{q}'(x_i)) - u(r'_j\, \hat{q}'(x'_j)) \right| \tag{14}$$
$$\leq n^{-1} \left(|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j))| + \sum_{i=1}^{n} \mathbb{I}_{i \neq j} \big| u(r_i\, \hat{q}(x_i)) - u(r_i\, \hat{q}'(x_i)) \big| \right) \tag{15}$$
$$\leq n^{-1} \left(|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j))| + (n-1)\beta \right). \tag{16}$$

Considérons le premier terme. Par le Lemme 3, p. 13, on sait que $\hat{q}(x) \leq (2\lambda)^{-1} \bar{r} \xi^2$ et que $|R| \leq \bar{r}$. On peut donc borner cette différence par la différence dans l'utilité dérivée par la meilleure décision d'investissement sur le meilleur rendement et sur le pire rendement. Par hypothèse Lipschitz et de sur-gradient de 1 à $r = 0$, on sait que pour $r > 0, u(r) < r$ et que pour $r < 0$, $\gamma r \leq u(r)$. On peut donc conclure que

$$|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j)| \leq u((2\lambda)^{-1} \bar{r}^2 \xi^2) - u(-(2\lambda)^{-1} \bar{r}^2 \xi^2) \tag{17}$$
$$\leq (2\lambda)^{-1} (\gamma + 1) \bar{r}^2 \xi^2. \tag{18}$$

Ce qui entraîne donc que

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \leq \frac{\gamma + 1}{2\lambda n} \bar{r}^2 \xi^2 + \frac{n-1}{n} \beta \tag{19}$$
$$\leq \beta + \frac{\gamma + 1}{2\lambda n} \bar{r}^2 \xi^2, \tag{20}$$

d'où, après quelques simplifications algébriques, on peut enfin tirer que

$$|\hat{\zeta}(\mathcal{S}_n) - \hat{\zeta}(S'_n)| \leq \beta(2\gamma^2 + \gamma + 1). \tag{21}$$

Ainsi la différence dans l'erreur de généralisation est de convergence (n^{-1}). À ce stade, la démonstration est presque complète, puisqu'en appliquant l'inégalité de concentration de McDiarmid, on obtient que pour tout $\mathcal{S}_n$,

$$\mathrm{P}\{\hat{\zeta}(\mathcal{S}_n) \geq \epsilon + \boldsymbol{E}_{\mathcal{S}_n}\, \hat{\zeta}(\mathcal{S}_n)\} \leq \exp\left(-\frac{2\epsilon^2}{n\beta^2(2\gamma^2 + \gamma + 1)^2} \right), \tag{22}$$

ce qui revient à dire qu'avec probabilité $1 - \delta$:

$$\hat{\zeta}(\mathcal{S}_n) < \boldsymbol{E}_{\mathcal{S}_n}\, \hat{\zeta}(\mathcal{S}_n) + \frac{\sqrt{n}\beta(2\gamma^2 + \gamma + 1) \log(1/\delta)}{2}. \tag{23}$$

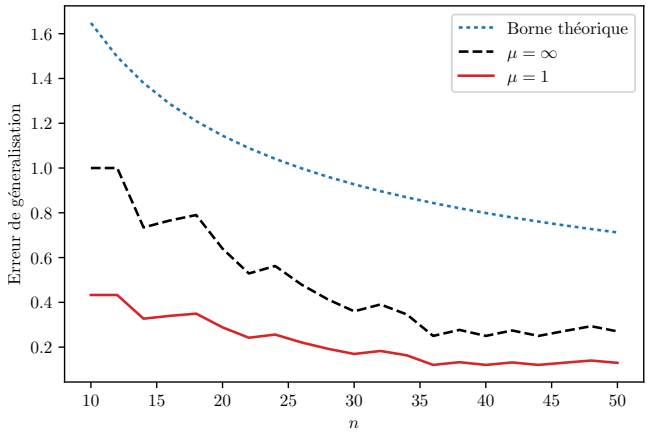

FIGURE 2 – Erreur de généralisation pour distribution Rademacher à deux dimensions

Or, $\boldsymbol{E}_{\mathcal{S}_n} \hat{\zeta}(\mathcal{S}_n) \leq \beta$ (voir [MRT12] pour une preuve technique mais complète), d'où on a finalement la borne recherchée :

$$\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \leq \hat{\Omega}_u, \tag{24}$$

où

$$\hat{\Omega}_u = \frac{\bar{r}^2 \xi^2}{2\lambda} \left(\frac{\gamma^2}{n} + (2\gamma^2 + \gamma + 1) \sqrt{\frac{\log(1/\delta)}{2n}} \right). \tag{25}$$

Exemple 3. DISTRIBUTION RADEMACHER – Pour concrétiser le comportement asymptotique des équations (24) et (25), on peut considérer la situation extrêmement simple où M n'est formée que de deux marges : une seule variable d'information X et une variable de rendement, l'une indépendante de l'autre et toutes deux Rademacher [2]. Dans le cas d'un noyau linéaire, on a trivialement $\bar{r} = 1$ et $\xi^2 = 1$; avec $\lambda = 1/2$, on a alors $\hat{\Omega}_u = 1/n + 2\sqrt{2\log(1/\delta)/n}$. La Fig. 2 (p. 6) illustre cette situation pour une confiance de 95% avec une utilité Lipschitz et le cas limite neutre au risque.

Équivalent certain À ce point-ci, il ne reste plus qu'à inverser le domaine de cette garantie pour l'exprimer en unités de rendements. En effet, si à partir d'un échantillon d'entraînement on a pu calculer un rendement équivalent $\widehat{CE} = u^{-1}(\widehat{EU})$, en utilisant le résultat du Lemme 5, p. 14, un investisseur aura un rendement équivalent hors échantillon *CE* tel que

$$CE \geq \widehat{CE} - (1/(\lambda\sqrt{n})). \tag{26}$$

2. Une variable aléatoire *Rademacher* $\mathcal{R}$ retourne ± 1 avec probabilité $1/2$. Elle possède notamment la propriété d'être standard, *i.e.* $\boldsymbol{E}\ \mathcal{R} = 0$ et $\boldsymbol{Var}\ \mathcal{R} = 1$.

De façon explicite :

$$CE \geq \widehat{CE} - \partial u^{-1}(\widehat{CE}) \cdot \frac{\bar{r}^2 \xi^2}{2\lambda} \left(\frac{\gamma^2}{n} + (2\gamma^2 + \gamma + 1) \sqrt{\frac{\log(1/\delta)}{2n}} \right). \tag{27}$$

Cette borne permet ainsi d'appréhender dans quelle mesure un large échantillonage est nécessaire pour obtenir un degré de confiance élevé. On notera l'influence de plusieurs facteurs sur la qualité de la borne (la discussion sur l'influence du terme $\bar{r}^2 \xi^2$ est repoussé à la Section 1.3).

Ainsi, la constante γ et le terme du sur-gradient inverse $\partial u^{-1}(\widehat{CE})$ sont tous deux susceptibles de dégrader considérablement la borne, particulièrement lorsque l'investisseur est doté d'une utilité très averse au risque ; dans des cas extrêmes, par exemple une utilité exponentielle inverse, ces deux valeurs divergeront très rapidement. Il convient cependant de prendre note que la constante Lipschitz est globalement plus importante puisqu'on considère son carré. Il devient alors essentiel de contrôler l'agressivité de l'algorithme en choisissant des valeurs élevées pour la régularisation λ de manière à chercher une utilité espérée relativement proche de $u(0)$.

On constate par ailleurs le rôle de premier plan que joue le terme de régularisation. Avec une régularisation élevée, on obtiendra sans surprise une borne très serrée, mais aux dépens de la politique d'investissement qui varie selon $(1/\lambda)$. Il est donc primordial de faire une validation croisée sur λ pour déterminer le meilleur compromis entre la variance des résultats et l'objectif à atteindre. La constante de confiance δ est quant à elle très performante ; une confiance de 99.9% n'accroît la borne que par un facteur de 2.63. Enfin, compte tenu du théorème limite centrale, l'ordre de convergence de $(1/\sqrt{n})$ n'a finalement rien de surprenant.**[Todo:** Plus de détails...**]**

1.2 Bornes de sous optimalité

Introduction et hypothèses supplémentaires Jusqu'ici, les efforts théoriques ont été déployés pour déterminer comment se comportait la fonction de décision $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ dans un univers probabiliste par rapport à l'univers statistique dans lequel elle avait été construite. Notre attention va maintenant se tourner vers la performance de $\hat{q}$ dans l'univers probabiliste par rapport à la meilleure décision disponible, c'est à dire la solution $q^\star$ de

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{E}\ u(R \cdot q(X)). \tag{28}$$

Il convient cependant de réaliser que l'existence d'une borne sur $q^\star$ n'est pas assurée. En effet, supposons d'une part que l'on dispose d'une utilité neutre au risque $\boldsymbol{1}$, telle que $\boldsymbol{1}(r) = r$, et d'autre part que $\boldsymbol{E}\, R = 0$. Soit $\alpha > 0$. On pourrait alors définir la fonction suivante :

$$q = \alpha\, \boldsymbol{E}(R\, \kappa(X, \cdot)) \tag{29}$$

On aurait alors, du fait de la linéarité du produit scalaire,

$$\boldsymbol{E1}(q) = \boldsymbol{E}(R\, q(X)) \tag{30}$$

$$= \boldsymbol{E}(R\langle q, \kappa(X, \cdot)\rangle) \tag{31}$$
$$= \boldsymbol{E}\langle q, R\,\kappa(X, \cdot)\rangle \tag{32}$$
$$= \langle q, \boldsymbol{E}(R\,\kappa(X, \cdot))\rangle \tag{33}$$
$$= \alpha\|q\|^2 \geq 0. \tag{34}$$

On peut alors obtenir une utilité espérée non bornée à mesure que $\alpha \to \infty$. Par ailleurs, ainsi défini, q représente effectivement la covariance entre R et la projection de X dans l'espace dual de $\boldsymbol{Q}$; par exemple dans le cas d'un noyau linéaire on aurait $q = \boldsymbol{E}(RX^T) = \boldsymbol{Cov}(R, X)$. On sait qu'en espérance l'application de q à X variera de la même façon que celle de R et donc qu'on aura une utilité infinie, puisque l'utilité est neutre.

Pour empêcher une telle situation d'exister on introduit l'hypothèse suivante. Elle exclut toute forme d'utilité à pente constante pour $r \geq r_0$, notamment l'utilité risque neutre.

Hypothèse 4. *L'utilité croît sous-linéairement, ie.* $u(r) = o(r)$ [3].

Une autre hypothèse est maintenant nécessaire pour s'assurer que $q^\star$ soit borné : l'absence d'arbitrage. D'un point de vue strictement financier, cela fait certainement du sens en vertu de l'efficience des marchés, version semi-forte*[Citation needed]*. D'un point de vue théorique, ceci exige en fait qu'il n'y ait pas de région dans $\boldsymbol{X}$ telle que tous les rendements s'y produisant soient nécessairement positifs ou négatifs.**[Todo:** Inserer image**]**. Ainsi, même en ayant une conaissance parfaite du monde, il subsistera toujours un terme de bruit rendant incertains la réalisation des rendements.

Hypothèse 5. *Pour toute région* $\mathcal{X} \subseteq \boldsymbol{X}$*,*

$$\mathrm{P}\{R \geq 0 \mid X \in \mathcal{X}\} < 1, \tag{35}$$

et de la même façon avec l'évènement $\mathrm{P}\{R \leq 0 \mid X \in \mathcal{X}\}$.

Décision optimale finie On veut montrer que $\|q^\star\|$ est borné. Pour ce faire, on va tout d'abord décomposer $q = s\theta$, où on pose $\|\theta\| = 1$ et $s > 0$; ainsi on peut poser notre problème d'optimisation comme la recherche d'une 'direction' θ et d'une magnitude s dans $\boldsymbol{Q}$. De plus, puisque $\|q\| = s$, il suffit de montrer que $s^\star$ est borné.

Notons d'abord que l'hypothèse 5 entraîne en particulier qu'il existe $\delta > 0$ et $\varrho \geq 0$ tels que

$$\mathrm{P}\{R \cdot \theta(X) \leq -\delta\} > \varrho \tag{36}$$

pour tout $\theta \in \boldsymbol{Q}$ tel que $\|\theta\| = 1$. Définissons maintenant une variable aléatoire à deux états : $B = -\delta$ avec probabilité ϱ et $B = \bar{r}\xi$ avec probabilité $1 - \varrho$. Puisque $R \cdot \theta(X) \leq \bar{r}\xi$, on a alors que, pour tout $r \in \boldsymbol{R}$,

$$\mathrm{P}\{B \geq r\} \geq \mathrm{P}\{R \cdot \theta(X) \geq r\} \tag{37}$$

3. Mathématiquement, on exige donc que $u(r)/r \to 0$.

[Todo: voir figure a produire.**]**

Puisque par hypothèse u est concave et puisque que B domine stochastiquement $R \cdot \theta(X)$, on a nécessairement que $\boldsymbol{E}\, u(sB) \geq \boldsymbol{E}\, u(R \cdot s\theta(X))$, pour tout $s > 0$. Or, par hypothèse de sous-linéarité on obtient que

$$\lim_{s\to\infty} \boldsymbol{E}\, u(R \cdot s\theta(X)) \leq \lim_{s\to\infty} u(sB) \tag{38}$$

$$= \lim_{s\to\infty} (\varrho u(-s\delta) + (1-\varrho)u(s\bar{r}\xi)) \tag{39}$$

$$\leq \lim_{s\to\infty} -\varrho s\delta + (1-\varrho)o(s) = -\infty, \tag{40}$$

ce qui démontre bien que s est borné.

Dérivation de la borne On cherchera donc à établir une borne Ω_u sur *l'erreur de sous-optimalité* de $\hat{q} \sim \mathcal{Q}(M^n)$:

$$\boldsymbol{EU}(\hat{q}) \geq \boldsymbol{EU}(q^\star) - \Omega_u. \tag{41}$$

Pour ce faire, on utilisera le résultat suivant, montré par *[Citation needed]*Shalev. En posant

$$\omega = \frac{4\gamma^2\xi^2(32 + \log(1/\delta))}{\lambda n}, \tag{42}$$

on obtient qu'avec probabilité $1-\delta$,

$$\lambda\|\hat{q} - q_\lambda^\star\|^2 \leq \boldsymbol{EU}_\lambda(q_\lambda^\star) - \boldsymbol{EU}_\lambda(\hat{q}) \leq \omega. \tag{43}$$

De la deuxième inégalité, on obtient alors que

$$\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(q_\lambda^\star) \geq -\omega + \lambda\|\hat{q}\|^2 - \lambda\|q_\lambda^\star\|^2 \tag{44}$$

$$\geq -\omega - 2\lambda\|\hat{q}\|\|q_\lambda^\star - \hat{q}\| - \lambda\|q_\lambda^\star - \hat{q}\|^2. \tag{45}$$

Or, pour un même δ, le résultat de Shalev*[Citation needed]*implique que $\|q_\lambda^\star - \hat{q}\| \leq \sqrt{\omega/\lambda}$. De plus, par le lemme 3, p. 13, $\|\hat{q}\| \leq \bar{r}\xi/(2\lambda)$, d'où on obtient

$$\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(q_\lambda^\star) \geq -2\omega - \bar{r}\xi\sqrt{\frac{\omega}{\lambda}}. \tag{46}$$

Enfin, puisque par définition de $q_\lambda^\star$, $\boldsymbol{EU}(q_\lambda^\star) - \lambda\|q_\lambda^\star\|^2 \geq \boldsymbol{EU}(q^\star) - \lambda\|q^\star\|^2$, on trouve alors que

$$\boldsymbol{EU}(q_\lambda^\star) - \boldsymbol{EU}(q^\star) \geq \lambda\|q_\lambda^\star\|^2 - \lambda\|q^\star\|^2 \geq -\lambda\|q^\star\|^2, \tag{47}$$

ce qui donne finalement

$$\boldsymbol{EU}(\hat{q}) = \boldsymbol{EU}(q^\star) + \boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(q_\lambda^\star) + \boldsymbol{EU}(q_\lambda^\star) - \boldsymbol{EU}(q^\star) \tag{48}$$

$$\geq \boldsymbol{EU}(q^\star) - 2\omega - \bar{r}\xi\sqrt{\omega/\lambda} - \lambda\|q^\star\|^2. \tag{49}$$

Équivalent certain et analyse À partir du résultat obtenu au dernier paragraphe, on peut à nouveau inverser le domaine de garantie afin de l'exprimer en rendement équivalent. En définissant *CE* l'équivalent certain hors échantillon suivant la politique $\hat{q}$ et $CE^\star$ l'équivalent certain optimal compte tenu de l'utilité donnée, l'application directe du Lemme 5, permet de garantir une performance de l'ordre de

$$CE \geq CE^\star - (1/(\lambda\sqrt{n})). \tag{50}$$

Plus précisément, avec probabilité $1-\delta$,

$$CE \geq CE^\star - \partial u^{-1}(CE) \cdot \left(\lambda \|q^\star\|^2 + \frac{8\gamma^2\xi^2(32+\log(1/\delta))}{n\lambda} + \frac{2\gamma\bar{r}\xi^2}{\lambda}\sqrt{\frac{32+\log(1/\delta)}{n}} \right) \tag{51}$$

Les bornes de sous-optimalité convergent ainsi environ à la même vitesse que celle de sous-optimalité, c'est-à-dire dans un régime de $(1/\sqrt{n})$. Bien sûr, une différence majeure est la présence de $\|q^\star\|$ qui est a priori impossible à déterminer, dans la mesure où aucune hypothèse n'est faite sur la distribution de M. On constate d'ailleurs sans surprise qu'une faible valeur de régularisation permet au résultat algorithmique de se raprocher du résultat optimal, bien que les autres termes de la borne aient un effet inverse. Par ailleurs, le sur-gradient inverse de u à *CE* ne peut lui non plus être déterminé précisément, aussi pour estimer la borne on lui substituera $\partial u^{-1}(\widehat{CE})$.

1.3 Garanties et dimensionalité du problème

Toutes les bornes considérées jusqu'à présent ont été dérivées sans faire apparaître explicitement la relation qui les lient avec avec la dimension p de l'espace $\boldsymbol{Q}$; autrement dit, on a implicitement considéré que $p = o(n)$. Or, si à première vue l'erreur de généralisation et de sous-optimalité du problème de portefeuille se comportent comme $(1/(\lambda\sqrt{n}))$, dans un contexte ou p est comparable à n, on souhaite comprendre comment l'ajout d'information dans $\boldsymbol{Q}$ peut venir affecter ces bornes.

Discussion sur la première hypothèse Revenons dans un premier temps sur la première hypothèse qu'on a employé allègrement dans nos résultats ; celle-ci stipule que $\kappa(x,x) \leq \xi^2$. Pour les espaces de décision affines, par exemple ceux engendrés par les noyaux de la forme $\kappa(x_1,x_2) = f(\|x_1 - x_2\|)$, cette propriété est naturellement observée puisqu'alors $\kappa(x,x) = f(0)$, peu importe la taille de $\boldsymbol{X}$. Pour d'autres types de noyaux, par exemple les décisions linéaires $\kappa(x_1,x_2) = x_1^T x_2$, il devient alors nécessaire de borner le support de X pour respecter la condition. Deux approches peuvent alors être employées : soit chaque variable d'information est bornée individuellement, soit on borne simplement $\kappa(X,X)$ par une borne probabiliste.

Le premier cas se prête bien à la situation où on dispose d'une bonne compréhension des variables d'information et de leur distribution. Par exemple, X_j peut naturellement et/ou raisonnablement reposer sur un support fini ; pour d'autres types de

distributions, par exemple les variables normales et sous-normales (dominées stochastiquement par une variable normale), on peut borner avec un haut degré de confiance la déviation de leur espérance. Les cas problématiques seront plutôt présentés par des variables X_j présentants des moments supérieurs élevés. En pratique, on pourra alors soit *saturer* l'information par une borne arbitraire, *i.e.* en posant $\tilde{X}_j = X_j(\nu_j/|X_j|)$, puis en ajoutant une nouvelle dimension d'information vrai/faux indiquant si la borne a été atteinte, ou simplement décider de l'incorporer telle qu'elle, mais en n'ayant alors aucune garantie sur les performances hors échantillon. Pour un noyau linéaire, si chaque variable $|X_j| \leq \nu_j$, alors par le théorème de Pythagore on a simplement que $\|X\|^2 \leq \|\nu\|^2 = \xi^2$. On remarquera alors que $\xi^2 = (p)$. Pour les noyaux polynomiaux d'ordre k, ce serait plutôt $\xi^2 = (p^k)$.

Penchons-nous un moment sur le cas linéaire. La situation où X dispose d'une borne explicite sur son support peut en fait être relaxée, moyennant que chacune des composantes soient indépendantes l'une à l'autre et que leur carré soient de forme sous-exponentielle [4]. Sous sa forme généralisée, l'inégalité de Bernstein implique qu'avec haute probabilité,

$$\mathrm{P}\{|\|X\|^2 - \boldsymbol{E}\,\|X\|^2| \geq t\} \leq \exp\left(\frac{t^2}{(p)}\right). \tag{52}$$

Autrement dit, à mesure que p est grand, la norme $\|X\|^2$ sera concentrée autour de son espérance. Si $\boldsymbol{E}\,X_j = 0$, alors $\|X\|^2 \approx \boldsymbol{E}\,\|X\|^2 = \sum_{j=1}^{p} \boldsymbol{Var}\,X_j = (p)$, et on aura donc une borne $\xi^2 = (p)$, mais nettement plus forte que celle considérée au dernier paragraphe, puisque les bornes deviennent alors inutiles. De plus, l'ajout d'une seule dimension d'information vient automatiquement rendre inexacte la borne statique ξ^2.

Dans un contexte où p est de l'ordre de n, les bornes dérivées aux deux dernières sous-sections peuvent donc se révéler trompeuses, puisqu'elles suggèrent à un potentiel investisseur des garanties ne dépendant que de n. En particulier, puisque toutes nos bornes sont en fait de la forme $\Omega = (\xi^2/\lambda\sqrt{n})$, il serait plus exact de postuler l'existence d'une variable ξ^2 telle que les bornes se comportent en fait selon la dynamique

$$\Omega = (p/\lambda\sqrt{n}). \tag{53}$$

En particulier, dans des régimes où $\sqrt{n} = (p)$, il devient impossible d'avoir des bornes convergeant vers 0, celles-ci restant en fait stationnaires. En outre, si $\sqrt{n} = o(p)$, par exemple si $p = (n)$, alors une divergence devient assurée.

Cependant, cette discussion n'est valide que dans le cas particulier des noyaux linéaires. Les noyaux gaussiens conservent quant à eux une indépendance par rapport à la dimensionalité, alors que les noyaux polynomiaux l'exacerbent ; pour un noyau de degré k il devient plus juste d'indiquer

$$\Omega = (p^k/\lambda\sqrt{n}). \tag{54}$$

4. Voir Boucheron et/ou Wainwright et/ou définir brièvement

Introduction au cas linéaire **[Todo:** Ne pas lire cette section ! !**]** Pour le moment, nous allons considérer le cas plus simple où $\boldsymbol{Q} = \boldsymbol{X}^*$, c'est à dire que le problème revient simplement à

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} r_i \, q^T x_i - \lambda \|q\|^2. \tag{55}$$

Pour simplifier la présentation, une utilité neutre au risque sera considérée comme cas limite au problème plus général (voir lemme de borne*[Citation needed]*).

D'un point de vue probabiliste, on peut définir $q_\lambda^\star$ comme la solution de

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad \boldsymbol{E}(R \, X^T q) - \lambda \|q\|^2, \tag{56}$$

d'où on tire

$$q_\lambda^\star = \frac{1}{2\lambda} \boldsymbol{Cov}(R, X), \tag{57}$$

puisque les deux variables sont centrées. On retrouve alors l'inégalité montrée en lemme *[Citation needed]*(nécessaire ? ?) Considérons maintenant P le rendement aléatoire obtenu en utilisant la décision $q_\lambda^\star$:

$$P = \frac{1}{2\lambda} R \, X^T \boldsymbol{Cov}(R, X). \tag{58}$$

On a alors $\boldsymbol{E} \, P = 1/2\lambda \; \boldsymbol{Cov}^2(R, X)$.

Puisque toutes nos variables sont centrées et réduites,

$$\boldsymbol{Cov}(R, X) = \sum_{j=1}^{p} \boldsymbol{E} \, R \, X_j. \tag{59}$$

En supposant que notre problème est pleinement déterminé en supposant l'existance d'une matrice A telle que $R = AX$

1.4 Note bibliographique

La théorie de la stabilité algorithmique remonte en fait aux années 70 avec les travaux de Luc Devroye appliqués à l'algorithme des k plus proches voisins*[Citation needed]*. Jusqu'alors, les bornes de généralisation étaient présentées pour toute décision $q \in \boldsymbol{Q}$ (ie Vapnik). Bousquet*[Citation needed]*a été le premier a présenter des résultats dans des espaces de Hilbert à noyau reproduisant. La démonstration est fortement inspirée de l'excellente référence [MRT12]. La démonstration de la borne sur la décision bornée est un résultat inédit, dûe à Delage dans le cas linéaire. On doit également à Rudin l'idée de la dimensionalité sur la qualité des garanties, et plus généralement l'idée d'employer une fonction de perte pour parvenir à autre chose qu'une question de régression/classification comme c'est souvent le cas.

1.5 Lemmes

[Todo: Ordonner les lemmes selon l'ordre dans lequel ils sont invoqués.]

Lemme 1 (Stabilité). On montre ici que

$$\beta \leq \frac{(\gamma \bar{r} \xi)^2}{2\lambda n}. \tag{60}$$

Lemme 2 (Décision neutre au risque comme cas limite). Soient $\hat{q}_u$ la solution de

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q) \tag{61}$$

et $\hat{q}_1$ la solution de

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EI}}_\lambda(q), \tag{62}$$

où $\widehat{\boldsymbol{EI}}(q) := n^{-1} \sum_{i=1}^n r_i \, q(x_i)$. On note tout d'abord avec l'inégalité de Jensen que $u(\widehat{\boldsymbol{EI}}(\hat{q}_u)) \geq \widehat{\boldsymbol{EU}}(\hat{q}_u) \geq \lambda \|\hat{q}_u\|^2 \geq 0$. Mais puisque u a un sur-gradient de 1 à 0, on déduit que $u(x) \geq 0$ entraîne $x \geq u(x)$. On a ainsi $\widehat{\boldsymbol{EI}}(\hat{q}_u) - \lambda \|\hat{q}_u\|^2 \geq 0$. Mais comme $\hat{q}_1$ maximise $\widehat{\boldsymbol{EI}}_\lambda$, on obtient

$$\widehat{\boldsymbol{EI}}(\hat{q}_1) - \lambda \|\hat{q}_1\|^2 \geq \widehat{\boldsymbol{EI}}(\hat{q}_u) - \lambda \|\hat{q}_u\|^2 \geq 0, \tag{63}$$

d'où on tire finalement $\|\hat{q}_u\| \leq \|\hat{q}_1\|$.

Lemme 3 (Borne sur la décision algorithmique). On va ici démontrer que la décision $\hat{q}(x)$ est bornée, et ce, pour tout $x \in \boldsymbol{X}$ et pour toute solution $\hat{q}$ de

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q). \tag{64}$$

Pour ce faire, on va mettre à profit la propriété reproductive de $\boldsymbol{Q}$ induite par κ qui stipule que

$$q(x) = \langle q, \kappa(x, \cdot) \rangle_{\boldsymbol{Q}} \leq \|q\| \sqrt{\kappa(x, x)}, \tag{65}$$

où l'inégalité découle de l'inégalité Cauchy-Schwartz appliquée au produit interne de $\boldsymbol{Q}$. On rappelle que, par hypothèse, $\forall x \in \boldsymbol{X}, \kappa(x, x) \leq \xi^2$; il suffit donc de borner $\|q\|$. De plus, par le Lemme 11, il suffit en fait de borner la solution de $\widehat{\boldsymbol{EI}}_\lambda(q)$. Mais,

$$\widehat{\boldsymbol{EI}}_\lambda(q) = n^{-1} \sum_{i=1}^n r_i \, q(x_i) - \lambda \|q\|^2 \tag{66}$$

$$\leq n^{-1} \sum_{i=1}^n r_i \sqrt{\kappa(x_i, x_i)} \|q\| - \lambda \|q\|^2 \tag{67}$$

$$\leq \bar{r} \xi \|q\| - \lambda \|q\|^2. \tag{68}$$

Puisque l'expression $\bar{r} \xi \|q\| - \lambda \|q\|^2$ est quadratique, elle atteint son maximum à

$$\|q\| = \frac{\bar{r} \xi}{2\lambda}, \tag{69}$$

on en conclut que $\|\hat{q}\| \leq (2\lambda)^{-1}\bar{r}\xi$ et donc que

$$\hat{q}(x) \leq \frac{\bar{r}\xi^2}{2\lambda}. \tag{70}$$

Lemme 4 (Forte concavité). L'objectif est fortement concave, que ce soit sous sa version statistique $\widehat{\boldsymbol{EU}}_\lambda$ ou probabiliste $\boldsymbol{EU}_\lambda$. Autrement dit, pour tout $\alpha \in [0,1]$, on a

$$\boldsymbol{EU}_\lambda(tq_1+(1-\alpha)q_2) \geq \alpha\,\boldsymbol{EU}_\lambda(q_1)+(1-\alpha)\,\boldsymbol{EU}_\lambda(q_2)+\lambda\alpha(1-\alpha)\|q_1-q_2\|^2, \tag{71}$$

et de même pour $\widehat{\boldsymbol{EU}}_\lambda$. Effectivement, puisque u est concave et $\|\cdot\|^2$ est convexe, on a successivement :

$$\begin{aligned}
&\boldsymbol{EU}_\lambda(\alpha q_1 + (1-\alpha)q_2) && (72)\\
&\quad = \boldsymbol{E}\, u(R\cdot(\alpha q_1 + (1-\alpha)q_2)(X)) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (73)\\
&\quad = \boldsymbol{E}\, u(\alpha(R\cdot q_1(X)) + (1-\alpha)(R\cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (74)\\
&\quad \geq \boldsymbol{E}(\alpha u(R\cdot q_1(X)) + (1-\alpha)u(R\cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (75)\\
&\quad = \alpha\,\boldsymbol{EU}(q_1) + (1-\alpha)\,\boldsymbol{EU}(q_2) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (76)\\
&\quad = \alpha\,\boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\,\boldsymbol{EU}_\lambda(q_2) - \lambda(\|\alpha q_1 + (1-\alpha)q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2). && (77)
\end{aligned}$$

Mais d'autre part,

$$\begin{aligned}
&-\lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 + \lambda\alpha\|q_1\|^2 + \lambda(1-\alpha)\|q\|^2 && (78)\\
&\quad = \lambda\alpha(1-\alpha)(\|q_1\|^2 + \|q_2\|^2 - 2\langle q_1, q_2\rangle) && (79)\\
&\quad = \lambda\alpha(1-\alpha)\|q_1 - q_2\|^2, && (80)
\end{aligned}$$

Ce qui complète la démonstration. La dérivation demeure exactement la même lorsqu'on considère $\widehat{\boldsymbol{EU}}_\lambda$.

Lemme 5 (Borne sur l'équivalent certain). Soient $CE_1 = u^{-1}(EU_1)$ et $CE_2 = u^{-1}(EU_2)$ et soit une borne Ω_u telle que

$$EU_1 \geq EU_2 - \Omega_u. \tag{81}$$

Par définition du sur-gradient, pour tout $r \in \mathscr{R}$, $u(r+\Delta) \leq u(r) + \Delta \cdot \partial u(r)$. Donc en posant $\Delta = CE_1 - CE_2$ et $r = CE_2$, on obtient ces deux inégalités :

$$-\Omega_u \leq EU_1 - EU_2 = u(CE_1) - u(CE_2) \leq \partial u(CE_2)(CE_1 - CE_2). \tag{82}$$

On trouve ainsi :

$$CE_1 \geq CE_2 - \Omega_u \cdot \partial u^{-1}(CE_2). \tag{83}$$

Typiquement, CE_1 et EU_1 seront des quantités inobservables, alors que CE_2 et EU_2 seront des quantités calculables. De plus, si $\partial u^{-1}(CE_2)$ comporte plusieurs éléments (*e.g.* si la dérivée de u est discontinue à CE_2), on choisira l'élément le plus favorable ; la plupart du temps ce sera équivalent à $\lim_{r \to CE_2^-} 1/u'(r)$ dans la région où $1/u'(r)$ est défini. Enfin, on note que cette limite existe puisque u est strictement monotone, et donc sa pente ne s'annule nulle part.

Lemme 6 (Généralisation du lemme de Hoeffding). Ce lemme généralise le lemme de Hoeffding à un espace vectoriel de dimension arbitraire $\boldsymbol{Q}$. Soit un vecteur aléatoire $Q \in \boldsymbol{Q}$ tel que $\|Q\| \leq \beta$ et $\boldsymbol{E}\, Q = 0$. Alors pour tout $t \in \boldsymbol{Q}$,

$$\boldsymbol{E}\, e^{\langle t, Q\rangle} \leq \exp\left(\frac{\beta^2 \|t\|^2}{2}\right). \tag{84}$$

En effet, on sait que par définition de la convexité de la fonction exponentielle, pour tout $s \in [0, 1]$,

$$\exp(sa + (1-s)b) \leq s \exp a + (1-s) \exp b. \tag{85}$$

Donc en définissant $g : \{q \in \boldsymbol{Q} : \|q\| \leq \beta\} \to [0, 1]$ par

$$g(q) = \frac{1}{2}\left(\frac{\langle t, q\rangle}{\beta\|t\|} + 1\right) \tag{86}$$

et en posant $a = \beta\|t\|$ et $b = -\beta\|t\|$, alors pour tout $q \in \boldsymbol{Q}$,

$$ag(q) = \frac{1}{2}(\langle t, q\rangle + \beta\|t\|), \tag{87}$$

$$b(1 - g(q)) = -\frac{1}{2}(\beta\|t\| - \langle t, q\rangle), \tag{88}$$

et donc

$$\exp(ag(q) + (1 - g(q))b) = e^{\langle t, q\rangle}. \tag{89}$$

La branche droite de l'inégalité devient quant à elle

$$\left(\frac{\langle t, q\rangle}{\beta\|t\|} + 1\right) e^{\beta\|t\|} + \left(1 - \frac{\langle t, q\rangle}{\beta\|t\|}\right) e^{-\beta\|t\|} \tag{90}$$

et donc, puisque $\boldsymbol{E}\langle t, Q\rangle = \langle t, \boldsymbol{E}\, Q\rangle = 0$,

$$\begin{aligned} \boldsymbol{E}\, e^{\langle t, Q\rangle} &\leq \boldsymbol{E}\left(\left(\frac{\langle t, Q\rangle}{\beta\|t\|} + 1\right) e^{\beta\|t\|} + \left(1 - \frac{\langle t, Q\rangle}{\beta\|t\|}\right) e^{-\beta\|t\|}\right) && (91)\\ &= e^{\beta\|t\|} + e^{-\beta\|t\|} && (92)\\ &= e^{\phi(\beta\|t\|)} && (93) \end{aligned}$$

où $\phi(x) = \log(e^x + e^{-x})$. Or, avec le résultat de [MRT12], p. 370, on a $\phi(x) \leq x^2/2$, d'où on tire le résultat annoncé.

Lemme 7 (Généralisation de la borne de Chernoff). Ce lemme généralise la borne de Chernoff à un espace vectoriel de dimension arbitraire $\boldsymbol{Q}$. Soit un vecteur aléatoire $Q \in \boldsymbol{Q}$. Alors l'évènement $\|Q\| \geq \epsilon$ aura lieu si et seulement s'il existe $t \in \boldsymbol{Q}$, $\|t\| = 1$ tel que $\langle t, Q\rangle \geq \epsilon$. Ainsi, pour tout $s > 0$, en employant l'inégalité de Markov,

$$\begin{aligned} \mathrm{P}\{\|Q\| \geq \epsilon\} = \mathrm{P}\{s\langle t, Q\rangle \geq s\epsilon\} &= \mathrm{P}\{e^{s\langle t, X\rangle} \geq e^{s\epsilon}\} && (94)\\ &\leq e^{-s\epsilon}\, \boldsymbol{E}\, e^{\langle t, Q\rangle}. && (95) \end{aligned}$$

Lemme 8 (Généralisation de l'inégalité de McDiarmid). L'inégalité de McDiarmid peut également se généraliser à des fonctions prenant leurs valeurs dans des espaces vectoriels À élaborer !

Soit une distribution $\mathscr{F}$ à valeur dans un espace quelconque $\boldsymbol{F}$, un espace vectoriel $\boldsymbol{Q}$ et une fonction $f : \boldsymbol{F}^n \to \boldsymbol{Q}$. S'il existe une constante $c \in \mathscr{R}$ telle que pour deux ensembles d'échantillons i.i.d. $\mathcal{S}_n \sim \mathscr{F}^n$ et $\mathcal{S}'_n$, où $\mathcal{S}_n$ et $\mathcal{S}'_n$ ne diffèrent que d'un seul point rééchantilloné de $\mathscr{F}$, on a

$$\|f(\mathcal{S}_n) - f(\mathcal{S}'_n)\| \leq c, \tag{96}$$

alors pour tout échantillon aléatoire $\mathcal{S}_n \sim \mathscr{F}^n$,

$$\mathrm{P}\{\|f(\mathcal{S}_n) - \boldsymbol{E}\, f(\mathcal{S}_n)\| \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{nc^2}\right). \tag{97}$$

Lemme 9 (Borne sur la décision). Considérons le cas d'une utilité neutre au risque puisqu'on sait que toute solution à $\max_q \boldsymbol{EU}_\lambda(q)$ sera bornée par celle de $\max_q \boldsymbol{EI}_\lambda(q)$. La stabilité de l'algorithme $\mathcal{Q}$ fournie par [BE02] établit que pour deux échantillons $\mathcal{S}_n$ et $\mathcal{S}'_n$ tirés de M^n et ne différant que d'un seul point,

$$\|\mathcal{Q}(\mathcal{S}_n) - \mathcal{Q}(\mathcal{S}'_n)\| \leq \frac{\bar{r}\xi}{\lambda n}. \tag{98}$$

En posant $\hat{q} \sim \mathcal{Q}(M^n)$, on peut donc appliquer directement le résultat de l'inégalité de McDiarmid (Lemme 8) pour obtenir avec probabilité $1 - \delta$ que

$$\|\hat{q} - \boldsymbol{E}\, \mathcal{Q}(\mathcal{S}_n)\| \leq \frac{\bar{r}\xi}{\lambda}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{99}$$

Or, $\mathcal{Q}$ est un estimateur non-biaisé de $q^\star_\lambda$. En effet, pour une utilité neutre au risque,

$$\boldsymbol{E}\, \mathcal{Q}(\mathcal{S}_n) = \boldsymbol{E}_{M^n}\left(\frac{1}{2n\lambda}\sum_{i=1}^{n} r_i\, \kappa(\cdot, x_i)\right) \tag{100}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\lambda}\boldsymbol{E}_M(R\,\kappa(\cdot, X)) \tag{101}$$

$$= \frac{1}{n}\sum_{i=1}^{n} q^\star_\lambda \tag{102}$$

$$= q^\star_\lambda. \tag{103}$$

On obtient ainsi

$$\|\hat{q} - q^\star_\lambda\| \leq \frac{\bar{r}\xi}{\lambda}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{104}$$

Exemple 4. CONVERGENCE DE $\hat{q}$ VERS $q^\star_\lambda$ – La propriété du Lemme 9 s'illustre particulièrement bien dans le cas où M n'est formé à ses marges que de distributions

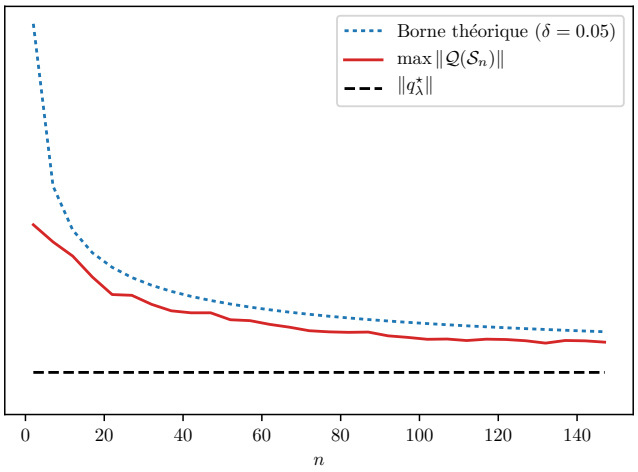

FIGURE 3 – Illustration du Lemme 9.

Rademacher. Ainsi, dans la Fig. 3 (p. 17), une distribution de marché à deux variables d'information indépendantes et toutes deux de corrélation $\rho = 0.5$ avec R sous copule gaussienne a été simulée 10 000 fois, pour constituer une "vraie" distribution pour laquelle $q^\star_\lambda$ peut être calculé ; 2000 échantillons de $\mathcal{S}_n$ ont été simulés.

Lemme 10. La solution $\hat{q}_1$ de

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{EI}_\lambda(q) = \hat{\boldsymbol{E}} \langle q|t\rangle - \tfrac{\lambda}{2}\|q\|^2. \tag{105}$$

est donnée par

$$\langle \hat{q}_1| = \lambda^{-1} \hat{\boldsymbol{E}} \langle t| \tag{106}$$

où $\langle x_i| = \kappa(x_i, \cdot)$ est l'élément dual de x sous $\boldsymbol{Q}$. Sous un noyau linéaire cela revient donc à

$$\hat{q}_1^T = \lambda^{-1} \hat{\boldsymbol{E}}(r^T x) \tag{107}$$

c'est à dire la covariance décentrée entre r et x. On observera aussi que

$$\boldsymbol{EI} = \lambda \langle \hat{q}_1|\cdot\rangle . \tag{108}$$

et donc que

$$\boldsymbol{EI}(\hat{q}_1) = \lambda \|\hat{q}_1\|^2. \tag{109}$$

Démonstration. Si on considère un déplacement de décision $\hat{q}_1 + \Delta q$, alors par linéarité le premier terme de l'objectif devient $\boldsymbol{EI}(\hat{q}_1 + \Delta q) = \boldsymbol{EI}(\hat{q}_1) + \boldsymbol{EI}(\Delta q)$ et le terme de régularisation devient

$$-\lambda/2\|\hat{q}_1 + \Delta q\|^2 = -\lambda/2\|\hat{q}_1\|^2 - \lambda \langle \hat{q}_1|\Delta q\rangle - \lambda/2\|\Delta q\|^2. \tag{110}$$

On a donc

$$\boldsymbol{EI}_\lambda(\hat{q}_1) - \boldsymbol{EI}_\lambda(\hat{q}_1 + \Delta q) = -\boldsymbol{EI}(\Delta q) + \lambda \langle \hat{q}_1 | \Delta q \rangle + \lambda/2 \|\Delta q\|^2 \tag{111}$$

$$= -\lambda \langle \hat{q}_1 | \Delta q \rangle + \lambda \langle \hat{q}_1 | \Delta q \rangle + \lambda/2 \|\Delta q\|^2 \tag{112}$$

$$= \lambda/2 \|\Delta q\|^2 \geq 0, \tag{113}$$

Ce qui entraîne $\boldsymbol{EI}_\lambda(\hat{q}_1) \geq \boldsymbol{EI}_\lambda(\hat{q}_1 + \Delta q)$. □

Lemme 11 (Borne sur la décision utilitaire). Pour toute fonction d'utilité u respectant les hypothèses,

$$\|\hat{q}_1\| \geq \|\hat{q}_u\|. \tag{114}$$

Ce lemme entraîne notamment que l'utilité en échantillon $\widehat{\boldsymbol{EU}}(\hat{q}_u) \leq \widehat{\boldsymbol{EI}}(\hat{q}_1)$: puisque $u(x) \leq x$,

$$\widehat{\boldsymbol{EU}}(\hat{q}_u) \leq \widehat{\boldsymbol{EI}}(\hat{q}_u) = \lambda \langle \hat{q}_1, \hat{q}_u \rangle \leq \lambda \|\hat{q}_1\| \|\hat{q}_u\| \leq \lambda \|\hat{q}_1\|^2 \tag{115}$$

$$= \widehat{\boldsymbol{EI}}(\hat{q}_1) \tag{116}$$

Démonstration. On note tout d'abord avec l'inégalité de Jensen que $u(\widehat{\boldsymbol{EI}}(\hat{q}_u)) \geq \widehat{\boldsymbol{EU}}(\hat{q}_u) \geq \lambda/2 \|\hat{q}_u\|^2 \geq 0$ puisque la valeur de l'objectif $\widehat{\boldsymbol{EI}}_\lambda(q)$ est d'au moins 0 à $q = 0$. Mais puisque u a un sur-gradient de 1 à 0, on déduit que $u(x) \geq 0$ entraîne $x \geq u(x)$. On a ainsi $\widehat{\boldsymbol{EI}}(\hat{q}_u) - \lambda/2 \|\hat{q}_u\|^2 \geq 0$. Ce qui entraîne alors que

$$\lambda \langle \hat{q}_1 | \hat{q}_u \rangle \geq \lambda/2 \|\hat{q}_u\|^2 \tag{117}$$

Mais par Cauchy-Schwartz, on a aussi

$$\|\hat{q}_1\| \|\hat{q}_u\| \geq \langle \hat{q}_1, \hat{q}_u \rangle \geq \|\hat{q}_u\|^2/2 \tag{118}$$

Et donc

$$\|\hat{q}_1\| \geq \|\hat{q}_u\|/2. \tag{119}$$

□

Lemme 12. L'erreur de généralisation du problème averse au risque est bornée par celle du problème neutre au risque :

$$\widehat{\boldsymbol{EU}}(\hat{q}_u) - \boldsymbol{EU}(\hat{q}_u) \leq \gamma(\widehat{\boldsymbol{EI}}(\hat{q}_1) - \boldsymbol{EI}(\hat{q}_1)). \tag{120}$$

Démonstration. Puisque u est monotone, on peut tout d'abord noter que pour tout $r + \Delta \in \boldsymbol{R}$, on a l'inégalité $u(r + \Delta) \leq u(r) + \Delta \partial u(r)$. Ainsi, pour deux variables aléatoires $R_1, R_2 \in \boldsymbol{R}$, en posant $\Delta = R_1 - R_2$, on a nécessairement

$$u(R_1) - u(R_2) \leq \partial u(R_2)(R_1 - R_2) \leq \gamma(R_1 - R_2), \tag{121}$$

par définition du coefficient Lipschitz. On tire donc

$$\boldsymbol{E}\, u(R_1) - \boldsymbol{E}\, u(R_2) \leq \gamma(\boldsymbol{E}\, R_1 - \boldsymbol{E}\, R_2). \tag{122}$$

En appliquant cette inégalité aux opérateurs $\widehat{\boldsymbol{EU}}$ et $\boldsymbol{EU}$ on obtient alors

$$\widehat{\boldsymbol{EU}}(\hat{q}_u) - \boldsymbol{EU}(\hat{q}_u) \leq \gamma(\widehat{\boldsymbol{E1}}(\hat{q}_u) - \boldsymbol{E1}(\hat{q}_u)) \tag{123}$$

$$= \gamma\lambda(\langle \hat{q}_1 | \hat{q}_u \rangle - \langle q_\lambda^\star | \hat{q}_u \rangle). \tag{124}$$

Mais par le Lemme 11, $\langle \hat{q}_1 | \hat{q}_u \rangle \geq 0$ et $\|\hat{q}_u\| \leq 2\|\hat{q}_1\|$. □

Références

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar) :499–526, 2002.

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.