Bornes de généralisation

Thierry Bazier-Matte

15 février 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 $\hat{q} : \boldsymbol{Q} \rightarrow \mathscr{R}$ 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* $u(r)|_{r=0} = 0$ *et* $\partial u(r)|_{r=0} \ni 1$ [1]. *De plus,* u *est défini*

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

sur 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 toutefois 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.

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

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

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

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.

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 $\beta = \mathcal{O}(n^{-1})$. Par le Lemme 1, p. 8 (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. 8, 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 $\mathcal{O}(n^{-1})$. À ce stade, la démonstration est presque complète, puisqu'en appliquant l'inégalité de concentration de McDiarmid, on obtient qu'avec probabilité $1 - \delta$:

$$\hat{\zeta}(\mathcal{S}_n) < \boldsymbol{E}_{\mathcal{S}_n} \hat{\zeta}(\mathcal{S}_n) + \sqrt{\beta(2\gamma^2 + \gamma + 1) \log(1/\delta)}. \tag{22}$$

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 :

$$\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \leq \hat{\Omega}_u, \tag{23}$$

où

$$\hat{\Omega}_u = \beta + \sqrt{\beta(2\gamma^2 + \gamma + 1)\log(1/\delta)} = \mathcal{O}(1/(\sqrt{n}\lambda)). \tag{24}$$

Équivalent certain Puis inverser pour obtenir l'équivalent certain.

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 Mohri[Citation needed]. La démonstration de la borne de la décision bornée est un résultat inédit, dûe à Delage dans le cas linéaire.

1.2 Bornes de sous optimalité

Exposition du problème 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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{EU}(R \cdot q(X)). \tag{25}$$

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, telle que $u(r) = r$, et d'autre part que $\boldsymbol{E}R = 0$. Soit $\alpha > 0$. On pourrait alors définir la fonction suivante :

$$q = \alpha \boldsymbol{E}(R\,\kappa(X, \cdot)) \tag{26}$$

On aurait alors

$$\boldsymbol{EU}(q) = \boldsymbol{E}(Rq(X)) = \boldsymbol{E}(R\boldsymbol{E}(R\,\kappa(X, X))) \tag{27}$$

$$= \boldsymbol{E}(R^2\,\kappa(X, X)) \geq 0, \tag{28}$$

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 $\boldsymbol{Q}$. Puisque l'utilité est neutre, 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. On verra plus loin au cours d'une démonstration la motivation derrière cette hypothèse supplémentaire :

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

Une autre hypothèse est maintenant nécessaire pour s'assurer que $q^\star$ soit borné : l'efficience des marchés. Dans notre cadre théorique, ceci se traduit par l'absence de l'existence d'une fonction $q \in \boldsymbol{Q}$ telle que

$$\boldsymbol{P}\{R \cdot q(X) > 0\} = 1. \tag{29}$$

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**]**.

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

$$\boldsymbol{P}\{R \geq 0 \mid X \in \mathcal{R}\} < 1, \tag{30}$$

et de la même façon avec l'évènement $\boldsymbol{P}\{R \leq 0\}$.

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

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 $\boldsymbol{Q}$. Si à première vue l'erreur de généralisation et de sous-optimalité du problème de portefeuille se comporte comme $\mathcal{O}(1/(\lambda\sqrt{n}))$, dans un contexte ou p est comparable à n, on souhaite comprendre comment l'ajout d'information dans $\boldsymbol{Q}$ peut venir ralentir ces bornes.

Discussion sur la première hypothèse La première hypothèse dépend pour sa part de la forme de κ. 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. 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 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 en ajoutant une 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 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 $\xi^2 = \mathcal{O}(p)$. Pour les noyaux polynomiaux d'ordre k, ce serait plutôt $\xi^2 = \mathcal{O}(p^k)$.

Cette 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 [2]. Dans sa forme généralisée, l'inégalité de Bernstein implique alors que

$$\boldsymbol{P}\{|\|X\|^2 - \boldsymbol{E}\|X\|^2| \geq t\} \leq \exp\left(\frac{t^2}{\mathcal{O}(p)}\right). \tag{31}$$

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 $\boldsymbol{E}\|X\|^2 = \sum_{j=1}^{p} \boldsymbol{Var}X_j = \mathcal{O}(p)$, et on aura donc une borne $\xi^2 = \mathcal{O}(p)$, mais nettement plus forte que celle considérée au dernier paragraphe, puisque la composante $\|\nu\|^2$ devient inutile.

[Todo: Montrer que la borne est proportionelle à $\max R \cdot q(X)$.**]**

Introduction au cas linéaire 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

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} r_i \, q^T x_i - \lambda \|q\|^2. \tag{32}$$

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

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad \boldsymbol{E}(R\, X^T q) - \lambda \|q\|^2, \tag{33}$$

d'où on tire

$$q_\lambda^\star = \frac{1}{2\lambda} \boldsymbol{Cov}(R, X), \tag{34}$$

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 $q_\lambda^\star$:

$$P = \frac{1}{2\lambda} R\, X^T \boldsymbol{Cov}(R, X). \tag{35}$$

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

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

Puisque toutes nos variables sont centrées et réduites,

$$\boldsymbol{Cov}(R, X) = \sum_{j=1}^{p} \boldsymbol{E} R\, X_j. \tag{36}$$

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

1.4 Lemmes

Lemme 1 (Stabilité). On montre ici que

$$\beta \leq \frac{(\gamma \bar{r} \xi)^2}{2\lambda n}. \tag{37}$$

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q) \tag{38}$$

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EI}}_\lambda(q), \tag{39}$$

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

$$\widehat{\boldsymbol{EI}}(\hat{q}_1) - \lambda \|\hat{q}_1\|^2 \geq \widehat{\boldsymbol{EI}}(\hat{q}_u) - \lambda \|\hat{q}_u\|^2 \geq 0, \tag{40}$$

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q). \tag{41}$$

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

$$q(x) = \langle q, \kappa(x, \cdot) \rangle_{\boldsymbol{Q}} \leq \|q\| \sqrt{\kappa(x, x)}, \tag{42}$$

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 2, il suffit en fait de borner la solution de $\widehat{\boldsymbol{EI}}_\lambda(q)$. Mais,

$$\widehat{\boldsymbol{EI}}_\lambda(q) = n^{-1} \sum_{i=1}^{n} r_i\, q(x_i) - \lambda \|q\|^2 \tag{43}$$

$$\leq n^{-1} \sum_{i=1}^{n} r_i \sqrt{\kappa(x_i, x_i)} \|q\| - \lambda \|q\|^2 \tag{44}$$

$$\leq \bar{r}\xi \|q\| - \lambda \|q\|^2. \tag{45}$$

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

$$\|q\| = \frac{\bar{r}\xi}{2\lambda}, \tag{46}$$

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

$$\hat{q}(x) \leq \frac{\bar{r}\xi^2}{2\lambda}. \tag{47}$$

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

$$\boldsymbol{EU}_\lambda(tq_1 + (1-\alpha)q_2) \geq \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) + \lambda\alpha(1-\alpha)\|q_1 - q_2\|^2, \tag{48}$$

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

$$\boldsymbol{EU}_\lambda(\alpha q_1 + (1-\alpha)q_2) \tag{49}$$

$$= \boldsymbol{E}u(R \cdot (\alpha q_1 + (1-\alpha)q_2)(X)) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{50}$$

$$= \boldsymbol{E}u(\alpha(R \cdot q_1(X)) + (1-\alpha)(R \cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{51}$$

$$\geq \boldsymbol{E}(\alpha u(R \cdot q_1(X)) + (1-\alpha)u(R \cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{52}$$

$$= \alpha \boldsymbol{EU}(q_1) + (1-\alpha)\boldsymbol{EU}(q_2) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{53}$$

$$= \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) - \lambda(\|\alpha q_1 + (1-\alpha)q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2) \tag{54}$$

$$\geq \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) - \lambda(\alpha\|q_1\|^2 + (1-\alpha)\|q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2) \tag{55}$$

$$= \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2). \tag{56}$$

La preuve est la même lorsqu'on considère $\widehat{\boldsymbol{EU}}_\lambda$.

Lemme 5 (Borne sur la décision optimale). 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

$$\boldsymbol{P}\{R \cdot \theta(X) \leq -\delta\} > \varrho \tag{57}$$

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 $r \in \boldsymbol{R}$,

$$\boldsymbol{P}\{B \geq r\} \geq \boldsymbol{P}\{R \cdot \theta(X) \geq r\} \tag{58}$$

[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

$$\begin{aligned} \lim_{s\to\infty} \boldsymbol{E}u(R \cdot s\theta(X)) &\leq \lim_{s\to\infty} u(sB) && (59)\\ &= \lim_{s\to\infty} (\varrho u(-s\delta) + (1-\varrho)u(s\bar{r}\xi)) && (60)\\ &\leq \lim_{s\to\infty} -\varrho s\delta + (1-\varrho)o(s) = -\infty, && (61) \end{aligned}$$

ce qui démontre bien que s est borné.

Lemme 6 (Borne sur l'équivalent certain).

Références

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