Quelques notes sur l'investissement d'un portefeuille à un actif en présence d'information complémentaire au marché

Thierry Bazier-Matte

7 avril 2017

Table des matières

1 Introduction

1.1 Avant propos

[Todo: Discuter du rôle croissant que jouent l'informatique et les statistiques dans la construction de portefeuille. Contraster avec les math. stochastiques. Citer Simons et cet article de Quandl selon lequel data is the new shit.**]**

1.2 Exposition du problème et hypothèses

Ce mémoire vise à établir clairement et rigoureusement comment un investisseur averse au risque disposant *d'information complémentaire* au *marché* peut utiliser cette information pour accroître son *utilité espérée* ou, de façon équivalente, son *rendement équivalent certain*.

Modélisation du marché Nous entendrons ici par *marché* n'importe quel type d'actif financier ou spéculatif dans lequel on peut investir une partie de sa fortune dans l'espoir de la voir fructifier au cours d'une période de temps arbitraire. Ainsi, tout au long de l'exposé théorique qui suivra, il peut être pertinent d'avoir en tête les rendements quotidiens issus des grands indices boursiers (par exemple les 500 plus grandes capitalisations américaines). Cependant, le traitement qui sera développé pourrait tout aussi bien s'appliquer à une action cotée en bourse dont on considère les rendements mensuels.[Nécessaire?]Mathématiquement, l'idée de marché peut ainsi être réduite à celle d'une variable aléatoire $R(t)$ décrivant l'évolution du rendement de l'actif en question.

Relativement à l'idée de marché, nous ferons également l'hypothèse que l'univers a une influence sur ces rendements. Il serait par exemple raisonnable de croire que le prix du pétrole a une influence sur l'évolution du rendement du marché américain. De la même façon, l'annonce d'un scandale aura a son tour des répercussions sur la valeur du titre de la compagnie dont il est l'objet. En outre, il a été montré par Fama et French que le rendement d'une action pouvait s'expliquer comme une combinaison de quelques facteurs fondamentaux (la taille de l'entreprise, le risque de marché et le ratio cours/valeur). On peut alors considérer un vecteur d'information $\vec{X}(t) = (X_1(t), X_2(t), \dots)$ dont chaque composante représente une information particulière, par exemple l'absence ou la présence d'un certain type de scandale, un ratio comptable, le prix d'un certain actif financier.[Rephrase] D'un point de vue probabiliste, on dira donc qu'il existe une forme de dépendance entre $R(t)$ et $\{\vec{X}(\tau) \mid \tau < t\}$ l'ensemble des évènements antérieurs à t. Le processus joint de ces deux évènements sera désormais défini comme *la distribution totale de marché*, ou simplement le marché.

Stationarité Bien qu'un tel modèle permette de représenter de façon très générale l'évolution d'un marché, nous formulerons l'hypothèse supplémentaire selon laquelle le marché est un processus *stationnaire*. Ceci permet notamment d'évacuer la notion

temporelle afin de ne représenter qu'une distribution de causes (l'information X) et d'effet (l'observation des rendements R). Cette hypothèse est assez contraignante. Elle suppose d'une part que les réalisations passées n'ont aucun effet sur les réalisations futures (indépendance) et d'autre part que la distribution de marché est figée dans le temps, ce qui implique notamment l'absence de probabilité de faillite. Elle implique aussi que le marché ne peut être vu comme un environnement adversariel qui réagirait par exemple aux décisions d'un investisseur. Ceci vient notamment mettre en cause la théorie des marchés efficients selon laquelle une brèche dans l'absence d'arbitrage serait immédiatement colmatée par des spéculateurs (effet d'autorégulation). Nous aurons toutefois l'occasion de revenir plus en détail sur les liens à faire entre cet exposé et l'efficience des marché.

Approche mathématique et statistique Dans ce qui suit, nous noterons par M la distribution de marché. Le vecteur aléatoire d'information sera par ailleurs formé de m composantes ; pour l'instant, aucune hypothèse par rapport à la dépendance des composantes de X ne sera formulée. À ce point-ci, on a donc le modèle de marché suivant :

$$M = (R, X_1, \ldots, X_m). \tag{1}$$

On fera également l'hypothèse qu'on possède un ensemble de n éléments échantillonés à partir de M, de sorte que :

$$\{r_i, x_{i1}, \ldots, x_{im}\}_{i=1}^{n} \sim M \tag{2}$$

représente notre ensemble d'échantillonage (aussi appelé ensemble d'entraînement). Le domaine des rendements possibles de R sera noté $\boldsymbol{R} \subseteq \mathscr{R}$ et celui du vecteur d'information X sera noté $\boldsymbol{X} \subseteq \mathscr{R}^m$. Le vecteur d'observations de rendement sera noté $r \in \mathscr{R}^n$ et la matrice d'information par $X \in \mathscr{R}^{n \times m}$.

Modélisation de la préférence Indépendamment de la notion de marché, *l'aversion au risque* est modélisée par une fonction d'utilité $u : \boldsymbol{R} \to \boldsymbol{U}$, où $\boldsymbol{R} \subseteq \mathscr{R}$ est le domaine (fermé ou non) des rendements considérés et $\boldsymbol{U} \subseteq \mathscr{R}$ celui des *utilités*.

Bien qu'en pratique il soit plus facile de travailler sur des fonctions possédant des valeurs dans $\boldsymbol{U}$, en pratique cet espace est adimensionel*[Citation needed]*, de sorte que nos résultats seront présentés dans l'espace des rendements $\boldsymbol{R}$.

Fonction de décision Donnés ces éléments de base, le but de ce mémoire sera alors de déterminer une fonction de décision d'investissement $q : \boldsymbol{X} \to \boldsymbol{P} \subseteq \mathscr{R}$ maximisant l'utilité espérée de l'investissement.

Mathématiquement on a donc le problème fondamental suivant :

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{E}\, u(R \cdot q(X)), \tag{3}$$

où l'optimisation a lieu dans un espace de fonctions $\boldsymbol{Q}$ à préciser.

Cependant, comme la distribution $(X, R) = M$ est inconnue, il est impossible de déterminer la fonction $q^\star$ minimisant cet objectif. On dispose toutefois d'un échantillon de M dont on peut se servir pour approximer le problème (ainsi formulé, le problème devient un programme d'optimisation stochastique), voir [SDR09]) :

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)), \tag{4}$$

mais encore ici le problème est mal spécifié, puisqu'aucune contrainte n'a été posée sur l'espace $\boldsymbol{Q}$. Par exemple, il suffirait de prendre pour q un dictionnaire associant à x_i la valeur $\alpha \, r_i$, où $\alpha > 0$, et à toute autre valeur de x une valeur nulle pour avoir une valeur d'utilité arbitrairement grande à mesure que $\alpha \to \infty$.

Risque in-échantillon et hors échantillon une telle fonction q est qu'elle se généralise très mal. En effet pour toute observation x qui ne figurerait dans l'ensemble d'entraînement, q prescrirait alors un investissement nul. Il y a alors une énorme différence entre l'utilité observée au sein de notre échantillon et l'utilité hors échantillon.

Donnée une fonction de décision $q \in \boldsymbol{Q}$ et un échantillon de M, on définit le *risque in-échantillon* ou *risque empirique* par

$$\hat{R}(q) = n^{-1} \sum_{i=1}^{n} \ell(r_i \, q(x_i)), \tag{5}$$

où $\ell = -u$. De la même façon, on définit le *risque hors-échantillon* ou *erreur de généralisation* par

$$R(q) = \boldsymbol{E} \, \ell(R \cdot q(X)). \tag{6}$$

On peut souhaiter d'une bonne fonction de décision qu'elle performe bien hors échantillon, aussi la quantité $R(q) - \hat{R}(q)$ sera-t-elle primordiale et beaucoup d'attention lui sera consacrée dans les prochaines sections. Notons que le risque hors-échantillon étant théoriquement impossible à calculer, en pratique on segmentera l'ensemble d'échantillonage en deux parties, l'une dédiée à l'apprentissage, l'autre à évaluer la performance hors échantillon.

Régularisation Afin de contrecarrer le risque hors échantillon, la solution est en fait de pénaliser la complexité de la fonction de décision q (rasoir d'Occam). Ainsi, on étudiera en profondeur le choix d'une fonctionelle $R : Q \to \mathscr{R}$ permettant de quantifier la complexité de q. L'objectif serait alors

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - R(q). \tag{7}$$

Par exemple, comme les mesures sur x peuvent comporter de l'incertitude ou du bruit, il serait souhaitable que la décision $q(x_1)$ soit proche de $q(x_2)$, si x_1 et x_2 sont eux

même proches dans l'espace $\boldsymbol{X}$. Si R encodait une telle préférence, ne fonction discontinue comme le dictionnaire présenté plus haut sera alors hautement défavorisée, et une fonction plus lisse y serait préférée.

[Todo: Introduire la validation croisée ainsi que le paramètre λ dans l'objectif.**]**

Espaces de décision En pratique, ce mémoire ne considérera que des espaces de Hilbert pour $\boldsymbol{Q}$. Un des avantages des espaces de Hilbert, c'est qu'ils induisent naturellement une notion de norme $\|\cdot\|_H$, qu'on peut intuitivement relié au concept de complexité. Nous nous intéresserons donc aux propriétés induites par $R(q) = \|q\|_H^2 = \langle q, q\rangle$. Il y a aussi moyen, sous des conditions assez techniques (théorème de la représentation) de généraliser la norme L_2 de q à une norme L_p général. En particulier, nous verrons qu'une régularisation donnée par norme L_1 induit certaines propriétés d'éparsité dans la solution.

Décisions linéaires De façon générale, la forme de décision la plus simple est celle qui combine linéairement les p observations de $x \in \boldsymbol{X} \subseteq \mathscr{R}^p$; autrement dit lorsque qu'on contraint $\boldsymbol{Q} = \boldsymbol{X}^*$, *i.e.*, à l'espace dual de $\boldsymbol{X}$. En langage plus clair, à toute fonction $q \in \boldsymbol{Q}$ il existe un vecteur de dimension p tel que la décision dérivée de l'observation x sera donnée par $q(x) = \langle q, x\rangle = q^T x$.

La régularisation L_2 de q devient alors tout simplement $R(q) = q^T q = \|q\|^2$ et la fonction optimale de décision $q^\star$ sera alors déterminée en résolvant le problème d'optimisation suivant :

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i\, q^T x_i) - \lambda \|q\|^2. \tag{8}$$

1.3 Dimensionalité de l'information

[Todo: Discussion du phénomène big data, de l'importance de p**]**

1.4 Risque et garanties statistiques sur la décision

[Todo: Discussion sur les méthodes de risques hors échantillon, complexité de l'échantillonage, mesure Rademacher, distance par rapport à la "meilleure" décision**]**

1.5 Interprétations

Interprétation géométrique dans l'espace X

Interprétation statistique (avec matrix covariance)

Autre ?

1.6 Objectifs

2 Optimisation moderne de portefeuille

Dans ce document, nous allons tenter de classifier et de répertorier la plupart des méthodes ayant rapport, de près ou de loin, à l'intersection des méthodes statistiques avancées et de l'apprentissage machine avec la théorie du portefeuille, en présentant pour chacune d'elle leurs avantages et leurs inconvénients.

2.1 Théorie classique du portefeuille

Une revue de littérature sur la théorie du portefeuille serait fondamentalement incomplète sans l'article fondateur de Markowitz, publié en 1952 [Mar52].

Nous allons montrer que le cadre théorique développé par Markowitz peut être considéré comme un cas particulier de notre algorithme, pour autant que l'on considère un portefeuille à un seul actif.

Soit $w \in \mathscr{R}^k$ le vecteur représentant la répartition du portefeuille de Markowitz à k actifs à optimiser. Alors un investisseur *markowitzien* souhaite résoudre le problème suivant :

$$\begin{aligned} \text{minimiser} \quad & w^T \Sigma w \\ \text{tel que} \quad & \mu^T w = \mu_0, \end{aligned} \tag{9}$$

où $\Sigma \in \mathscr{R}^{k \times k}$ est la covariance du rendement des actifs et $\mu \in \mathscr{R}^k$ le vecteur d'espérance. **[Todo:** Montrer formellement.**]** Par la théorie de l'optimisation convexe, il existe une constante $\gamma \in \mathscr{R}$ telle que le problème énoncé est équivalent à

$$\text{maximiser} \quad \mu^T w + \gamma\, w^T \Sigma w. \tag{10}$$

Dans le cas où on considère un portefeuille à un seul actif, alors ce problème se réduit alors à

$$\text{maximiser} \quad \mu q - \gamma \sigma^2 q^2, \tag{11}$$

où on a posé $\mu \coloneqq \boldsymbol{E}\, R$ et $\sigma^2 \coloneqq \boldsymbol{Var}\, R$.

Supposons qu'un investisseur soit doté d'une utilité quadratique paramétrée par

$$u(r) = r - \frac{\gamma}{\sigma^2 + \mu^2} \sigma^2 r^2, \tag{12}$$

et que l'information factorielle intégrée à l'algorithme ne consiste uniquement qu'en les rendements eux mêmes ; autrement dit, le vecteur d'information X se réduirait tout simplement à un terme constant fixé à 1, *i.e.*, $X \sim 1$. **[Todo:** expliquer**]**.

Avec une utilité (12) et l'absence d'information supplémentaire, l'objectif de *[Citation needed]* devient aussitôt

$$\boldsymbol{EU}(qR) = q\, \boldsymbol{E}\, R - \frac{\gamma}{\sigma^2 + \mu^2} \sigma^2 q^2\, \boldsymbol{E}\, R^2. \tag{13}$$

Mais puisque ***Var*** $R = \boldsymbol{E}\, R^2 - (\boldsymbol{E}\, R)^2$, on déduit $\boldsymbol{E}\, R^2 = \sigma^2 + \mu^2$, ce qui entraîne alors que (à faire) s'exprime par

$$\text{maximiser} \quad \boldsymbol{EU}(qR) = \mu q - \gamma \sigma^2 q^2, \tag{14}$$

ce qui est tout à fait identique à (11).

Nous suggérons au lecteur intéressé par l'équivalence des diverses formulations d'optimisation de portefeuille dans un univers de Markowitz [BPS13] et [Mar14], tous deux publiés à l'occasion du soixantième anniversaire de [Mar52].

2.2 Portefeuille universel / Papiers d'Elad Hazan

Ce mémoire sera également consacré aux garanties statistiques de performance des estimateurs $q^\star$.

Bien que le modèle soit différent et de nature itérative, le *portefeuille universel* de [Cov91] est à notre connaissance un des premiers modèles de gestion de portefeuille à exploiter une distribution arbitraire tout en proposant des garanties statistiques de convergence.

Voir [Cov91, Haz15].

2.3 Théorie de portefeuille régularisé

[BEKL16]

2.4 Fama and French et suivants ?

[FF93]

2.5 Articles du NIPS

2.6 Papiers de Ben Van Roy

2.7 Conclusions : Notre problème par rapport à ces deux disciplines

3 Introduction aux fonctions de décisions non linéaires

Ce chapitre se veut une brève introduction aux propriétés des espaces de décision obtenus par noyaux reproduisants. En premier lieu, une discussion sur la forme duale du problème linéaire ainsi que les propriétés des espaces à noyau permettront d'obtenir une meilleure intuition (Section 3.1). Par la suite, la Section 3.2 présentera quels algorithmes permettant de trouver une politique d'investissement optimal à partir d'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 et d'une fonction d'utilité concave. Quelques exemples de noyaux courants seront présentés, suivis des dérivations des deux formes d'optimisation.

3.1 Propriétés des espaces de décision à noyau reproduisant

Formulations primales et duales Tel que discuté en introduction, le cas le plus simple pour un espace de décision $\boldsymbol{Q}$ est celui où $\boldsymbol{Q} = \boldsymbol{X}^*$, c'est-à-dire le dual de l'espace vectoriel $\boldsymbol{X}$ [1]. La *décision* prise suite à l'observation d'un vecteur d'information $x \in \boldsymbol{X}$ est simplement $q(x) = q^T x$. Le problème à résoudre est ainsi

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q^T x_i) - \lambda \|q\|^2, \tag{15}$$

duquel on tire un $\hat{q}$ optimal. Cette formulation *primale* est intuitivement claire : on cherche à maximiser l'utilité moyenne suivant une politique unique q appliquée à chaque observation x_i, tout en cherchant à éviter de favoriser excessivement une des dimensions d'information par rapport aux autres. Or, selon le théorème de la représentation qui sera présenté un peu plus loin (p. 16), la politique optimale $\hat{q}$ peut également s'exprimer comme une combinaison linéaire des observations x_i. Ainsi, en notant $\Xi \in \mathscr{R}^{n \times p}$ la matrice des n observations de x, il existe $\hat{\alpha} \in \mathscr{R}^n$ tel que

$$\hat{q} = \Xi^T \hat{\alpha}. \tag{16}$$

Cette propriété fondamentale permet donc de chercher une combinaison linéaire optimale $\hat{\alpha} \in \mathscr{R}^n$ à partir de laquelle la politique optimale peut être déduite. En substituant (16) dans (15), on obtient la *représentation duale* du problème :

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, \alpha^T \Xi x_i) - \lambda \alpha^T \Xi \Xi^T \alpha. \tag{17}$$

Si, à des fins de simplification d'interprétation, l'investisseur est neutre au risque, et en notant $K := \Xi \Xi^T \in \mathscr{R}^{n \times n}$, *i.e.*, $K_{ij} = x_i^T x_j$, alors le problème sous sa forme duale s'exprime comme

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \alpha^T K r - \lambda \alpha^T K \alpha. \tag{18}$$

1. Le *dual* $\boldsymbol{V}^*$ d'un espace vectoriel $\boldsymbol{V}$ correspond à l'ensemble des formes linéaires sur $\boldsymbol{V}$. Dans le cas fini où $\boldsymbol{V} = \mathscr{R}^m$, alors un élément $w^* \in \boldsymbol{V}^*$ est souvent représenté par un vecteur ligne w^T à m éléments, tel que $w^*(v) = w^T v$.

Intuitivement, la matrice K, étant semi-définie positive, représente une *covariance de similarité* entre chacune des observations x_i, où la variance de chaque observation est donnée par sa norme $\|x_i\|^2$ et la corrélation entre deux observations par le cosinus de l'angle : $\rho_{ij} = x_i^T x_j / \|x_i\| \|x_j\|$. L'expression $n^{-1} Kr \in \mathscr{R}^n$ indique quelles dimensions permettent d'obtenir le meilleur rendement en considérant l'influence pondérée de toutes les observations :

$$[Kr]_j = n^{-1} \sum_{i=1}^{n} r_i \rho_{ij} \|x_i\| \|x_j\|. \tag{19}$$

Le rôle de α est alors de choisir les dimensions les plus favorables ; enfin le terme de régularisation $\lambda \alpha^T K \alpha$ a pour effet non seulement de choisir une solution finie (puisque quadratique), mais aussi de standardiser l'effet de chaque dimension afin de limiter par exemple l'influence d'observations dotées d'une norme plus élevée que les autres.

On note finalement que la solution analytique du problème risque neutre devient

$$K\alpha = \frac{1}{2n\lambda} Kr. \tag{20}$$

entraînant sans surprise $\hat{\alpha} = (2n\lambda)^{-1} r$ si K est de plein rang. Si par contre K n'est pas de plein rang, c'est-à-dire s'il existe une observation de norme nulle ($\|x_i\| = 0$) ou colinéaire par rapport à une autre ($x_i = k x_j$ entraîne $\rho_{ij} = 1$), $\hat{\alpha}$ n'est pas défini puisqu'il existe alors une infinité de solutions. Il est à noter que le théorème de la représentation n'est pas forcément *nécessaire*, il est simplement suffisant.

Nous verrons cependant une autre forme duale au problème dont la solution $\hat{\alpha}$ est en bijection avec $\hat{q}$.

Décisions non-linéaires Si cette classe des décisions linéaires a l'avantage d'être simple, elle est en revanche fort peu adaptée à des situations pourtant peu complexes. Géométriquement, elle ne fait que séparer l'espace $\boldsymbol{X}$ en deux : un côté entraînera des décisions d'investissement positifs, l'autre des décisions négatives. **[Todo:** Problème XOR irrésoluble**]**.

La méthode des noyaux permet de circonvenir ce problème en remplaçant la notion de similarité entre deux points par une fonction de noyau semi-défini positif κ.

Définition. Un *noyau semi-défini positif*, ou simplement un *noyau* $\kappa : \boldsymbol{X} \times \boldsymbol{X} \to \mathscr{R}$ est tel que pour tout ensemble $\{x_1, \ldots, x_n\} \in \boldsymbol{X}^n$, la matrice $K_{ij} = \kappa(x_i, x_j)$ est semi-définie positive.

Proposition 1. *Tout noyau semi-défini positif κ induit un espace de décision $\boldsymbol{Q}$*[2] *doté d'un produit scalaire $\langle \cdot, \cdot \rangle : \boldsymbol{Q} \times \boldsymbol{Q} \to \mathscr{R}$ ainsi que d'une application $\phi : \boldsymbol{X} \to \boldsymbol{Q}$ donnée par $\phi(x) = \kappa(x, \cdot) = \kappa(\cdot, x)$. De plus, $\boldsymbol{Q}$ dispose de la propriété reproductive par laquelle pour tout $q \in \boldsymbol{Q}$, $q(x) = \langle q, \phi(x) \rangle$. En particulier on en conclut que*

2. Pour être tout à fait exact, $\boldsymbol{Q}$ est alors un espace de Hilbert à noyau reproduisant.

$\kappa(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$. *Finalement, l'inégalité de Cauchy-Swartz s'applique à* $\boldsymbol{Q}$ *: pour tout* $q_1, q_2 \in \boldsymbol{Q}$, $\langle q_1, q_2 \rangle^2 \leq \|q_1\| \|q_2\|$, *où la norme de* q *est définie par* $\|q\|^2 = \langle q, q \rangle$. *En particulier, on note que* $q(x)^2 \leq \|q\| \kappa(x, x)$.

Ainsi, doté d'un noyau κ, on obtient un espace de décision $\boldsymbol{Q}$ tel que le problème primal s'exprime par

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - \lambda \|q\|^2. \tag{21}$$

Il convient de noter que chaque type de noyau entraîne une classe de décision bien particulière. Ainsi, selon la géométrie de la densité de la distribution M, certains noyaux seront plus adaptés que d'autre. D'une certaine façon, il s'agit là d'une faiblesse du modèle car celui-ci est incapable de *déterminer* le bon noyau à employer et cette tâche revient alors au gestionnaire de portefeuille.

Exemples Outre le *noyau linéaire*, défini par $\kappa(x_1, x_2) = x_1^T x_2$, les *noyaux polynômiaux d'ordre* k donnés par $\kappa(x_1, x_2) = (x_1^T x_2 + c)^k$ sont également courants. Ces types de noyaux ont cependant l'inconvénient de conserver une notion d'amplitude absolue ; on peut à l'inverse définir des noyaux invariants au déplacement et à la rotation, *i.e.* tels que $\kappa(x_1, x_2) = \kappa(\|x_1 - x_2\|)$. La notion de similarité ne dépend alors plus que de la distance entre deux points. Ainsi, le noyau gaussien κ_σ sera défini par :

$$\kappa_\sigma(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2} \right), \tag{22}$$

où σ représente la sensibilité du noyau ; des valeurs élevées de σ le rendront rapidement insensible à des données pourtant rapprochées dans l'espace $\boldsymbol{X}$ alors qu'une valeur σ faible leur accordera une similarité beaucoup plus grande.

Enfin, ces noyaux peuvent se recombiner afin d'en former de nouveaux. Voir Bishop et Mohri.

3.2 Algorithmes de décision non-linéaires

Magré que le problème primal soit bien posé, l'espace $\boldsymbol{Q}$ est a priori inconnu et peut de surcroit être de dimension infinie. Il est donc nécessaire de déterminer une méthode algorithmique capable de déterminer $\hat{q}$. Si la matrice de similarité K est définie positive, alors on peut utiliser le théorème de la représentation pour résoudre le problème suivant :

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad \sum_{i=1}^{n} u(r_i \, \alpha^T \psi(x_i)) - \alpha^T K \alpha. \tag{23}$$

où $\psi : \boldsymbol{X} \to \mathscr{R}^n$ est un opérateur linéaire tel que $\psi(x_i)_j = \kappa(x_i, x_j)$; c'est en fait la contrepartie de l'application de Ξ sur x_i dans le cas linéaire. Par ailleurs la décision

optimale s'exprime comme $\hat{q} = \hat{\alpha}^T \psi$, c'est-à-dire comme une combinaison linéaire de fonctions non-linéaires. Enfin, si K n'est pas définie positive, il suffit alors de ne considérer que les points sans co-linéarité ou avec norme non nulle.

Il existe aussi une autre façon de résoudre le problème primal qui consiste à dualiser le problème primal dans le cas linéaire pour voir émerger la matrice de similarité K, ce qui permet alors de considérer n'importe quel noyau. Par ailleurs, cette méthode est valide que K soit de plein rang ou non. Ainsi, le problème primal peut se résoudre suivant le problème

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad -\sum_{i=1}^{n} \ell^*(\alpha_i / r_i) - \frac{1}{4n\lambda} \alpha^T K \alpha. \tag{24}$$

La fonction $\ell^* : \mathscr{R} \to \mathscr{R}$ est le *conjugué convexe* de la fonction de perte $\ell = -u$ (voir (35), p. 15). La décision est donnée par

$$q(x) = -\frac{1}{2n\lambda} \alpha^T \psi(x). \tag{25}$$

Par exemple, dans le cas d'une utilité risque neutre, $\ell^* = \infty$ sauf si $\alpha_i / r_i = -1$, donc nécessairement $\alpha = -r$ et alors

$$q(x) = \frac{1}{2n\lambda} r^T \psi(x). \tag{26}$$

3.3 Démonstrations

Approche duale On cherche à résoudre le problème suivant, avec $q \in \mathscr{R}^p$ comme variable d'optimisation :

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell(r_i \, q^T x_i) + n\lambda \|q\|^2, \tag{27}$$

où $\ell = -u$. De façon équivalente, en introduisant un nouveau vecteur $\xi \in \mathscr{R}^n$, on a

$$\begin{array}{rl} \text{minimiser} & \displaystyle\sum_{i=1}^{n} \ell(\xi_i) + n\lambda \|q\|^2 \\ \text{tel que} & \xi_i = r_i \, q^T x_i. \end{array} \tag{28}$$

Soit $\alpha \in \mathscr{R}^n$. Le lagrangien de (28) peut s'exprimer comme

$$\mathscr{L}(q, \xi, \alpha) = \sum_{i=1}^{n} \ell(\xi_i) + n\lambda \|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i). \tag{29}$$

Puique l'objectif de (28) est convexe et que ses contraintes sont affines en q et ξ, on peut appliquer le théorème de Slater qui spécifie que le saut de dualité du problème est

nul. En d'autres mots, résoudre (27) revient à maximiser la fonction dual de Lagrange g sur α :

$$\text{maximiser} \quad g(\alpha) = \inf_{q,\xi} \mathscr{L}(q, \xi, \alpha). \tag{30}$$

On note que

$$g(\alpha) = \inf_{q,\xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) + n\lambda \|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i) \right\} \tag{31}$$

$$= \inf_{\xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) - \alpha^T \xi \right\} + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\} \tag{32}$$

$$= -\sup_{\xi} \left\{ a^T \xi - \sum_{i=1}^{n} \ell(\xi_i) \right\} + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\} \tag{33}$$

$$= -\sum_{i=1}^{n} \ell^*(\alpha_i) + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\}. \tag{34}$$

Où ℓ^* est le conjugé convexe de la fonction de perte et est définie par

$$\ell(\alpha_i) = \sup_{\xi_i} \left\{ \alpha_i \xi_i - \ell(\xi_i) \right\}. \tag{35}$$

On note par ailleurs l'usage de l'identité

$$f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \ell(\xi_i) \Longrightarrow f^*(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \ell^*(\xi_i) \tag{36}$$

À présent, considérons le second terme de (34). Puisque l'expression est dérivable, on peut résoudre analytiquement q.

$$\nabla_q \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\} = 0 \tag{37}$$

implique que

$$q = -\frac{1}{2n\lambda} \sum_{i=1}^{n} \alpha_i r_i x_i \tag{38}$$

à l'infimum.

En utilisant (38), on peut éliminer q de (34) pour obtenir

$$g(\alpha) = -\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{2n\lambda} \sum_{i,j=1}^{n} \alpha_i \alpha_j r_i r_j x_i^T x_j + \frac{1}{4n\lambda} \sum_{i,j=1}^{n} \alpha_i \alpha_j r_i r_j x_i^T x_j \tag{39}$$

$$= -\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r). \tag{40}$$

Ainsi, sous sa forme duale, le problème (27) est équivalent à résoudre

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell^*(\alpha_i) + \frac{1}{4n\lambda}(\alpha \circ r)^T K(\alpha \circ r). \tag{41}$$

On peut finalement définir $\tilde{\alpha}_i = \alpha_i / r_i$ pour obtenir le résultat annoncé plus haut.

Approche primale Soit $\kappa : \boldsymbol{X} \times \boldsymbol{X} \to \mathscr{R}$ un noyau semi-défini positif, $\boldsymbol{Q}$ l'espace de décision induit par κ et $K \in \mathscr{R}^{n \times n}$ la matrice de similarité. Le problème d'optimisation de portefeuille régularisé s'exprime alors par

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - \lambda \|q\|^2. \tag{42}$$

Tel que mentionné, la dimension de $\boldsymbol{Q}$ est possiblement infinie, ce qui rend numériquement impossible la recherche d'une solution $q^\star$. Toutefois, le théorème de la représentation permet de rendre le problème résoluble.

Théorème 1 (Théorème de la représentation). *Toute solution $q^\star$ de* (42) *repose dans le sous-espace vectoriel engendré par l'ensemble des n fonctions $\{\phi_i\}$, où $\phi_i = \kappa(x_i, \cdot)$. Numériquement, il existe un vecteur $\alpha \in \mathscr{R}^n$ tel que,*

$$q^\star = \sum_{i=1}^{n} \alpha_i \phi_i = \alpha^T \phi. \tag{43}$$

Démonstration. Voir [MRT12], Théorème 5.4 pour une démonstration tenant compte d'un objectif régularisé général. La démonstration est dûe à [KW71]. □

Le théorème de la représentation permet donc de chercher une solution dans un espace à n dimensions, plutôt que la dimension possiblement infinie de $\boldsymbol{Q}$. En effet, puisque

$$q^\star = \sum_{i=1}^{n} \alpha_i \phi_i, \tag{44}$$

où $\alpha \in \mathscr{R}^n$, on peut donc restreindre le domaine d'optimisation à $\mathscr{R}^n$. L'objectif de (42) devient alors

$$n^{-1} \sum_{i=1}^{n} u(r_i \textstyle\sum_{j=1}^{p} \alpha_j \phi_j(x_i)) - \lambda \langle q, q \rangle_{\boldsymbol{Q}}. \tag{45}$$

Le premier terme se réexprime comme

$$n^{-1} \sum_{i=1}^{n} u(r_i \, \alpha^T \phi(x_i)), \tag{46}$$

alors qu'en employant les propriétés de linéarité du produit intérieur, on transforme le second terme par

$$\langle q, q\rangle^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \alpha_i\alpha_j\langle\phi_i, \phi_j\rangle \tag{47}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{p} \alpha_i\alpha_j\kappa(x_i, x_j) \tag{48}$$

$$= \alpha^T K\alpha. \tag{49}$$

De sorte que le problème général (42) peut se reformuler par

$$\underset{\alpha\in\mathscr{R}^n}{\text{maximiser}} \quad n^{-1}\sum_{i=1}^{n} u(r_i\alpha^T\phi(x_i)) - \lambda\,\alpha^T K\alpha \ . \tag{50}$$

4 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 4.1). 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 4.2). Enfin, la Section 4.3 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* $\partial u(0) \ni 1$ [3]. *De plus,* u *est défini 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 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 4.3.

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

3. 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 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.

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.

4.1 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

$$\mathcal{Q}(\mathcal{S}_n) = \arg\max_{q \in \boldsymbol{Q}} \left\{ \widehat{\boldsymbol{EU}}(\mathcal{S}_n, q) - \lambda \|q\|^2 \right\}. \tag{51}$$

Comme on l'a vu, résoudre (51) est aussi équivalent à

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i(\alpha^T \phi)(x_i)) - \lambda \alpha^T K \alpha, \tag{52}$$

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 $\mathcal{S}_n \sim M^n$:

$$\hat{\zeta}_u(\mathcal{S}_n) \leq \hat{\Omega}_u(n, \ldots), \tag{53}$$

où

$$\hat{\zeta}_u(\mathcal{S}_n) = \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \tag{54}$$

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

$$\boldsymbol{CE}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) \geq \widehat{\boldsymbol{CE}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \hat{\Omega}_e(n, \ldots). \tag{55}$$

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.

Théorème 2 (Borne sur l'erreur de généralisation (util)). *L'erreur de généralisation sur $\hat{q}$ est bornée par*

$$\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \leq \hat{\Omega}_u, \tag{56}$$

où

$$\hat{\Omega}_u = \frac{\bar{r}^2 \xi^2}{2\lambda} \left(\frac{\gamma^2}{n} + (2\gamma^2 + \gamma + 1) \sqrt{\frac{\log(1/\delta)}{2n}} \right). \tag{57}$$

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 $(x, r) \sim M$,

$$|u(r\, \hat{q}(x)) - u(r\, \hat{q}'(x))| \leq \beta. \tag{58}$$

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. 28 (une application directe du théorème de Bousquet), on obtient effectivement

$$\beta \leq \frac{\gamma^2 \bar{r}^2 \xi^2}{2\lambda n}. \tag{59}$$

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 :

$$|\hat{\zeta}(\mathcal{S}_n) - \hat{\zeta}(\mathcal{S}'_n)| = |\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}') + \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \tag{60}$$

$$\leq |\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| + |\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')|. \tag{61}$$

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

$$|\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| = |\boldsymbol{E}(u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)))| \tag{62}$$

$$\leq \boldsymbol{E}(|u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)|) \tag{63}$$
$$\leq \beta, \tag{64}$$

par définiton de la stabilité. Quant au deuxième terme de (61) on peut le borner de la même façon :

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \tag{65}$$
$$= n^{-1} \left| \sum_{i=1}^{n} \mathbb{I}_{i \neq j} u(r_i\, \hat{q}(x_i)) + u(r_j\, \hat{q}(x_j)) - \sum_{i=1}^{n} \mathbb{I}_{i \neq j} u(r_i\, \hat{q}'(x_i)) - u(r'_j\, \hat{q}'(x'_j)) \right| \tag{66}$$
$$\leq n^{-1} \left(|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j))| + \sum_{i=1}^{n} \mathbb{I}_{i \neq j} \big| u(r_i\, \hat{q}(x_i)) - u(r_i\, \hat{q}'(x_i)) \big| \right) \tag{67}$$
$$\leq n^{-1} \left(|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j))| + (n-1)\beta \right). \tag{68}$$

Considérons le premier terme. Par le Lemme 3, p. 28, 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

$$|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j)| \leq u((2\lambda)^{-1} \bar{r}^2 \xi^2) - u(-(2\lambda)^{-1} \bar{r}^2 \xi^2) \tag{69}$$
$$\leq (2\lambda)^{-1} (\gamma + 1) \bar{r}^2 \xi^2. \tag{70}$$

Ce qui entraîne donc que

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \leq \frac{\gamma + 1}{2\lambda n} \bar{r}^2 \xi^2 + \frac{n-1}{n} \beta \tag{71}$$
$$\leq \beta + \frac{\gamma + 1}{2\lambda n} \bar{r}^2 \xi^2, \tag{72}$$

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

$$|\hat{\zeta}(\mathcal{S}_n) - \hat{\zeta}(S'_n)| \leq \beta(2\gamma^2 + \gamma + 1). \tag{73}$$

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 $\mathcal{S}_n$,

$$\mathrm{P}\{\hat{\zeta}(\mathcal{S}_n) \geq \epsilon + \boldsymbol{E}_{\mathcal{S}_n}\, \hat{\zeta}(\mathcal{S}_n)\} \leq \exp\left(-\frac{2\epsilon^2}{n\beta^2(2\gamma^2 + \gamma + 1)^2} \right), \tag{74}$$

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

$$\hat{\zeta}(\mathcal{S}_n) < \boldsymbol{E}_{\mathcal{S}_n}\, \hat{\zeta}(\mathcal{S}_n) + \frac{\sqrt{n}\beta(2\gamma^2 + \gamma + 1) \log(1/\delta)}{2}. \tag{75}$$

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.

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. 29, un investisseur aura un rendement équivalent hors échantillon *CE* tel que

$$CE \geq \widehat{CE} - (1/(\lambda\sqrt{n})). \tag{76}$$

De façon explicite :

$$CE \geq \widehat{CE} - \partial u^{-1}(\widehat{CE}) \cdot \frac{\bar{r}^2\xi^2}{2\lambda}\left(\frac{\gamma^2}{n} + (2\gamma^2+\gamma+1)\sqrt{\frac{\log(1/\delta)}{2n}}\right). \tag{77}$$

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 4.3).

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...**]**

4.2 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

$$\underset{q\in\boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{E}\ u(R\cdot q(X)). \tag{78}$$

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 ***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 :

$$q = \alpha\, \boldsymbol{E}(R\, \kappa(X, \cdot)) \tag{79}$$

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

$$\begin{aligned} \boldsymbol{E1}(q) &= \boldsymbol{E}(R\, q(X)) && (80)\\ &= \boldsymbol{E}(R\langle q, \kappa(X, \cdot)\rangle) && (81)\\ &= \boldsymbol{E}\langle q, R\, \kappa(X, \cdot)\rangle && (82)\\ &= \langle q, \boldsymbol{E}(R\, \kappa(X, \cdot))\rangle && (83)\\ &= \alpha\|q\|^2 \geq 0. && (84) \end{aligned}$$

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.* $u(r) = o(r)$ [4].

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* $\mathcal{X} \subseteq \boldsymbol{X}$*,*

$$\mathrm{P}\{R \geq 0 \mid X \in \mathcal{X}\} < 1, \tag{85}$$

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.

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

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

$$\mathrm{P}\{R \cdot \theta(X) \leq -\delta\} > \varrho \tag{86}$$

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 $r \in \boldsymbol{R}$,

$$\mathrm{P}\{B \geq r\} \geq \mathrm{P}\{R \cdot \theta(X) \geq r\} \tag{87}$$

[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

$$\begin{aligned}
\lim_{s\to\infty} \boldsymbol{E}\, u(R \cdot s\theta(X)) &\leq \lim_{s\to\infty} u(sB) && (88)\\
&= \lim_{s\to\infty} (\varrho u(-s\delta) + (1-\varrho)u(s\bar{r}\xi)) && (89)\\
&\leq \lim_{s\to\infty} -\varrho s\delta + (1-\varrho)o(s) = -\infty, && (90)
\end{aligned}$$

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 $\hat{q} \sim \mathcal{Q}(M^n)$:

$$\boldsymbol{EU}(\hat{q}) \geq \boldsymbol{EU}(q^\star) - \Omega_u. \tag{91}$$

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

$$\omega = \frac{4\gamma^2\xi^2(32 + \log(1/\delta))}{\lambda n}, \tag{92}$$

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

$$\lambda\|\hat{q} - q_\lambda^\star\|^2 \leq \boldsymbol{EU}_\lambda(q_\lambda^\star) - \boldsymbol{EU}_\lambda(\hat{q}) \leq \omega. \tag{93}$$

De la deuxième inégalité, on obtient alors que

$$\begin{aligned}
\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(q_\lambda^\star) &\geq -\omega + \lambda\|\hat{q}\|^2 - \lambda\|q_\lambda^\star\|^2 && (94)\\
&\geq -\omega - 2\lambda\|\hat{q}\|\|q_\lambda^\star - \hat{q}\| - \lambda\|q_\lambda^\star - \hat{q}\|^2. && (95)
\end{aligned}$$

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. 28, $\|\hat{q}\| \leq \bar{r}\xi/(2\lambda)$, d'où on obtient

$$\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(q_\lambda^\star) \geq -2\omega - \bar{r}\xi\sqrt{\frac{\omega}{\lambda}}. \tag{96}$$

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

$$\boldsymbol{EU}(q_\lambda^\star) - \boldsymbol{EU}(q^\star) \geq \lambda\|q_\lambda^\star\|^2 - \lambda\|q^\star\|^2 \geq -\lambda\|q^\star\|^2, \tag{97}$$

ce qui donne finalement

$$\begin{aligned}\boldsymbol{EU}(\hat{q}) &= \boldsymbol{EU}(q^\star) + \boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(q_\lambda^\star) + \boldsymbol{EU}(q_\lambda^\star) - \boldsymbol{EU}(q^\star) &(98)\\ &\geq \boldsymbol{EU}(q^\star) - 2\omega - \bar{r}\xi\sqrt{\omega/\lambda} - \lambda\|q^\star\|^2. &(99)\end{aligned}$$

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** 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

$$CE \geq CE^\star - (1/(\lambda\sqrt{n})). \tag{100}$$

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

$$CE \geq CE^\star - \partial u^{-1}(CE)\cdot\left(\lambda\|q^\star\|^2 + \frac{8\gamma^2\xi^2(32+\log(1/\delta))}{n\lambda} + \frac{2\gamma\bar{r}\xi^2}{\lambda}\sqrt{\frac{32+\log(1/\delta)}{n}}\right) \tag{101}$$

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 $\partial u^{-1}(\widehat{CE})$.

4.3 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 [5]. Sous sa forme généralisée, l'inégalité de Bernstein implique qu'avec haute probabilité,

$$\mathrm{P}\{|\|X\|^2 - \boldsymbol{E}\,\|X\|^2| \geq t\} \leq \exp\left(\frac{t^2}{(p)}\right). \tag{102}$$

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

$$\Omega = (p/\lambda\sqrt{n}). \tag{103}$$

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.

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

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

$$\Omega = (p^k/\lambda\sqrt{n}). \tag{104}$$

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

$$\underset{q\in\mathscr{R}^p}{\text{maximiser}} \quad n^{-1}\sum_{i=1}^{n} r_i\, q^T x_i - \lambda\|q\|^2. \tag{105}$$

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

$$\underset{q\in\mathscr{R}^p}{\text{maximiser}} \quad \boldsymbol{E}(R\,X^T q) - \lambda\|q\|^2, \tag{106}$$

d'où on tire

$$q_\lambda^\star = \frac{1}{2\lambda}\,\boldsymbol{Cov}(R, X), \tag{107}$$

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 $q_\lambda^\star$:

$$P = \frac{1}{2\lambda} R\,X^T\,\boldsymbol{Cov}(R, X). \tag{108}$$

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

Puisque toutes nos variables sont centrées et réduites,

$$\boldsymbol{Cov}(R, X) = \sum_{j=1}^{p} \boldsymbol{E}\,R\,X_j. \tag{109}$$

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

4.4 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.

4.5 Lemmes

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

Lemme 1 (Stabilité). On montre ici que

$$\beta \leq \frac{(\gamma \bar{r} \xi)^2}{2\lambda n}. \tag{110}$$

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q) \tag{111}$$

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EI}}_\lambda(q), \tag{112}$$

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

$$\widehat{\boldsymbol{EI}}(\hat{q}_1) - \lambda \|\hat{q}_1\|^2 \geq \widehat{\boldsymbol{EI}}(\hat{q}_u) - \lambda \|\hat{q}_u\|^2 \geq 0, \tag{113}$$

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q). \tag{114}$$

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

$$q(x) = \langle q, \kappa(x, \cdot) \rangle_{\boldsymbol{Q}} \leq \|q\| \sqrt{\kappa(x, x)}, \tag{115}$$

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,

$$\widehat{\boldsymbol{EI}}_\lambda(q) = n^{-1}\sum_{i=1}^{n} r_i\, q(x_i) - \lambda\|q\|^2 \tag{116}$$

$$\leq n^{-1}\sum_{i=1}^{n} r_i\sqrt{\kappa(x_i,x_i)}\|q\| - \lambda\|q\|^2 \tag{117}$$

$$\leq \bar{r}\xi\|q\| - \lambda\|q\|^2. \tag{118}$$

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

$$\|q\| = \frac{\bar{r}\xi}{2\lambda}, \tag{119}$$

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

$$\hat{q}(x) \leq \frac{\bar{r}\xi^2}{2\lambda}. \tag{120}$$

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

$$\boldsymbol{EU}_\lambda(tq_1+(1-\alpha)q_2) \geq \alpha\,\boldsymbol{EU}_\lambda(q_1)+(1-\alpha)\,\boldsymbol{EU}_\lambda(q_2)+\lambda\alpha(1-\alpha)\|q_1-q_2\|^2, \tag{121}$$

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

$$\boldsymbol{EU}_\lambda(\alpha q_1 + (1-\alpha)q_2) \tag{122}$$

$$= \boldsymbol{E}\, u(R\cdot(\alpha q_1 + (1-\alpha)q_2)(X)) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{123}$$

$$= \boldsymbol{E}\, u(\alpha(R\cdot q_1(X)) + (1-\alpha)(R\cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{124}$$

$$\geq \boldsymbol{E}(\alpha u(R\cdot q_1(X)) + (1-\alpha)u(R\cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{125}$$

$$= \alpha\,\boldsymbol{EU}(q_1) + (1-\alpha)\,\boldsymbol{EU}(q_2) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{126}$$

$$= \alpha\,\boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\,\boldsymbol{EU}_\lambda(q_2) - \lambda(\|\alpha q_1 + (1-\alpha)q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2). \tag{127}$$

Mais d'autre part,

$$-\lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 + \lambda\alpha\|q_1\|^2 + \lambda(1-\alpha)\|q\|^2 \tag{128}$$

$$= \lambda\alpha(1-\alpha)(\|q_1\|^2 + \|q_2\|^2 - 2\langle q_1, q_2\rangle) \tag{129}$$

$$= \lambda\alpha(1-\alpha)\|q_1 - q_2\|^2, \tag{130}$$

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

$$EU_1 \geq EU_2 - \Omega_u. \tag{131}$$

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 :

$$-\Omega_u \leq EU_1 - EU_2 = u(CE_1) - u(CE_2) \leq \partial u(CE_2)(CE_1 - CE_2). \tag{132}$$

On trouve ainsi :

$$CE_1 \geq CE_2 - \Omega_u \cdot \partial u^{-1}(CE_2). \tag{133}$$

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 $t \in \boldsymbol{Q}$,

$$\boldsymbol{E}\, e^{\langle t, Q \rangle} \leq \exp\left(\frac{\beta^2 \|t\|^2}{2}\right). \tag{134}$$

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

$$\exp(sa + (1-s)b) \leq s \exp a + (1-s) \exp b. \tag{135}$$

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

$$g(q) = \frac{1}{2}\left(\frac{\langle t, q \rangle}{\beta \|t\|} + 1\right) \tag{136}$$

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

$$ag(q) = \frac{1}{2}(\langle t, q \rangle + \beta\|t\|), \tag{137}$$

$$b(1 - g(q)) = -\frac{1}{2}(\beta\|t\| - \langle t, q \rangle), \tag{138}$$

et donc

$$\exp(ag(q) + (1 - g(q))b) = e^{\langle t, q \rangle}. \tag{139}$$

La branche droite de l'inégalité devient quant à elle

$$\left(\frac{\langle t, q \rangle}{\beta \|t\|} + 1\right) e^{\beta\|t\|} + \left(1 - \frac{\langle t, q \rangle}{\beta \|t\|}\right) e^{-\beta\|t\|} \tag{140}$$

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

$$\boldsymbol{E}\, e^{\langle t, Q \rangle} \leq \boldsymbol{E}\left(\left(\frac{\langle t, Q \rangle}{\beta \|t\|} + 1\right) e^{\beta\|t\|} + \left(1 - \frac{\langle t, Q \rangle}{\beta \|t\|}\right) e^{-\beta\|t\|}\right) \tag{141}$$

$$= e^{\beta\|t\|} + e^{-\beta\|t\|} \tag{142}$$

$$= e^{\phi(\beta\|t\|)} \tag{143}$$

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,

$$\mathrm{P}\{\|Q\| \geq \epsilon\} = \mathrm{P}\{s\langle t, Q\rangle \geq s\epsilon\} = \mathrm{P}\{e^{s\langle t,X\rangle} \geq e^{s\epsilon}\} \tag{144}$$

$$\leq e^{-s\epsilon}\, \boldsymbol{E}\, e^{\langle t,Q\rangle}. \tag{145}$$

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

$$\|f(\mathcal{S}_n) - f(\mathcal{S}'_n)\| \leq c, \tag{146}$$

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

$$\mathrm{P}\{\|f(\mathcal{S}_n) - \boldsymbol{E}\, f(\mathcal{S}_n)\| \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{nc^2}\right). \tag{147}$$

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,

$$\|\mathcal{Q}(\mathcal{S}_n) - \mathcal{Q}(\mathcal{S}'_n)\| \leq \frac{\bar{r}\xi}{\lambda n}. \tag{148}$$

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

$$\|\hat{q} - \boldsymbol{E}\, \mathcal{Q}(\mathcal{S}_n)\| \leq \frac{\bar{r}\xi}{\lambda}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{149}$$

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

$$\boldsymbol{E}\, \mathcal{Q}(\mathcal{S}_n) = \boldsymbol{E}_{M^n}\left(\frac{1}{2n\lambda}\sum_{i=1}^{n} r_i\, \kappa(\cdot, x_i)\right) \tag{150}$$

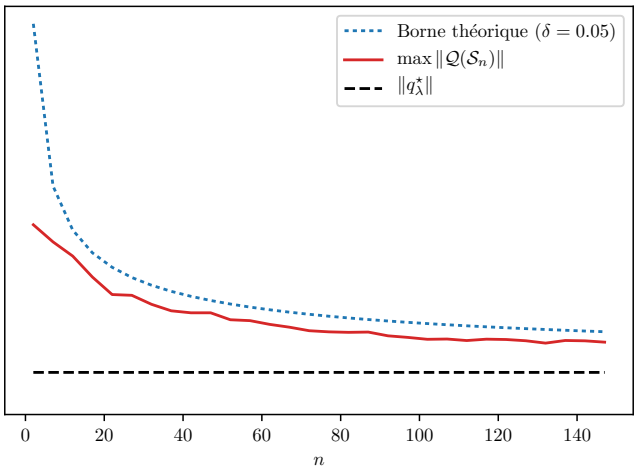

FIGURE 1 – Illustration du Lemme 9.

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\lambda}\, \boldsymbol{E}_M(R\,\kappa(\cdot, X)) \tag{151}$$

$$= \frac{1}{n}\sum_{i=1}^{n} q^\star_\lambda \tag{152}$$

$$= q^\star_\lambda. \tag{153}$$

On obtient ainsi

$$\|\hat{q} - q^\star_\lambda\| \leq \frac{\bar{r}\xi}{\lambda}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{154}$$

Exemple 1. 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 Rademacher. Ainsi, dans la Fig. 1 (p. 32), 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

$$\underset{q\in\boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{EI}_\lambda(q) = \hat{\boldsymbol{E}}\,\langle q|t\rangle - \tfrac{\lambda}{2}\|q\|^2. \tag{155}$$

est donnée par

$$\langle \hat{q}_1| = \lambda^{-1}\,\hat{\boldsymbol{E}}\,\langle t| \tag{156}$$

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

$$\hat{q}_1^T = \lambda^{-1}\hat{\boldsymbol{E}}(r^T x) \tag{157}$$

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

$$\boldsymbol{EI} = \lambda \langle \hat{q}_1 | \cdot \rangle . \tag{158}$$

et donc que

$$\boldsymbol{EI}(\hat{q}_1) = \lambda \|\hat{q}_1\|^2. \tag{159}$$

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

$$-\lambda/2\|\hat{q}_1 + \Delta q\|^2 = -\lambda/2\|\hat{q}_1\|^2 - \lambda \langle \hat{q}_1 | \Delta q \rangle - \lambda/2\|\Delta q\|^2. \tag{160}$$

On a donc

$$\begin{aligned} \boldsymbol{EI}_\lambda(\hat{q}_1) - \boldsymbol{EI}_\lambda(\hat{q}_1 + \Delta q) &= -\boldsymbol{EI}(\Delta q) + \lambda \langle \hat{q}_1 | \Delta q \rangle + \lambda/2\|\Delta q\|^2 && (161)\\ &= -\lambda \langle \hat{q}_1 | \Delta q \rangle + \lambda \langle \hat{q}_1 | \Delta q \rangle + \lambda/2\|\Delta q\|^2 && (162)\\ &= \lambda/2\|\Delta q\|^2 \geq 0, && (163) \end{aligned}$$

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,

$$\|\hat{q}_1\| \geq \|\hat{q}_u\|. \tag{164}$$

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

$$\begin{aligned} \widehat{\boldsymbol{EU}}(\hat{q}_u) \leq \widehat{\boldsymbol{EI}}(\hat{q}_u) = \lambda \langle \hat{q}_1, \hat{q}_u \rangle \leq \lambda \|\hat{q}_1\| \|\hat{q}_u\| \leq \lambda \|\hat{q}_1\|^2 && (165)\\ = \widehat{\boldsymbol{EI}}(\hat{q}_1) && (166) \end{aligned}$$

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

$$\lambda \langle \hat{q}_1 | \hat{q}_u \rangle \geq \lambda/2\|\hat{q}_u\|^2 \tag{167}$$

Mais par Cauchy-Schwartz, on a aussi

$$\|\hat{q}_1\| \|\hat{q}_u\| \geq \langle \hat{q}_1, \hat{q}_u \rangle \geq \|\hat{q}_u\|^2/2 \tag{168}$$

Et donc

$$\|\hat{q}_1\| \geq \|\hat{q}_u\|/2. \tag{169}$$

□

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

$$\widehat{\boldsymbol{EU}}(\hat{q}_u) - \boldsymbol{EU}(\hat{q}_u) \leq \gamma(\widehat{\boldsymbol{EI}}(\hat{q}_1) - \boldsymbol{EI}(\hat{q}_1)). \tag{170}$$

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

$$u(R_1) - u(R_2) \leq \partial u(R_2)(R_1 - R_2) \leq \gamma(R_1 - R_2), \tag{171}$$

par définition du coefficient Lipschitz. On tire donc

$$\boldsymbol{E}\, u(R_1) - \boldsymbol{E}\, u(R_2) \leq \gamma(\boldsymbol{E}\, R_1 - \boldsymbol{E}\, R_2). \tag{172}$$

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

$$\begin{aligned} \widehat{\boldsymbol{EU}}(\hat{q}_u) - \boldsymbol{EU}(\hat{q}_u) &\leq \gamma(\widehat{\boldsymbol{EI}}(\hat{q}_u) - \boldsymbol{EI}(\hat{q}_u)) && (173)\\ &= \gamma\lambda(\langle \hat{q}_1 | \hat{q}_u \rangle - \langle q^\star_\lambda | \hat{q}_u \rangle). && (174) \end{aligned}$$

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

5 Expériences empiriques

Cette section sera l'occasion de valider numériquement les garanties présentées à la *[Citation needed]*quant aux erreurs de généralisation et de sous optimalité inhérentes à l'algorithme d'investissement présenté dans ce mémoire.

Il va sans dire que le cadre théorique général qui a été développé jusqu'à maintenant présente plusieurs paramètres (dimensionalité du problème, loi de marché, fonction d'utilité, noyau employé, etc.) ; tous les décrire représenterait une tâche titanesque, aussi certains choix devront être faits pour restreindre la quantité de paramètres étudiés ; la Section 5.1 énumérera le choix fait pour chacun de ces paramètres.

Par la suite, les Sections 5.2, 5.3 et 5.4 étudieront la qualité des garanties de généralisation et de sous optimalité dans un contexte où, respectivement, la taille de l'échantillonage augmente, la taille de l'échantillonage est fixe mais la dimensionalité du problème augmente et enfin, la taille de l'échantillonage et de la dimensionalité augmentent toutes les deux, mais à des rythmes différents.

5.1 Méthodologie

Noyau Le noyau employé dans nos expériences sera linéaire. En particulier, c'est avec un tel noyau que la dépendance entre la dimensionalité du problème est les erreurs de sous optimalité et de généralisation est la plus facilement caractérisable.

Fonctions d'utilité Chaque expérience sera conditionnée par une fonction d'utilité exponentielle Lipschitz LEU_μ (voir Fig. 2 (p. 38) pour une description de cette famille).

Ces utilités sont idéales pour deux raisons : d'abord elles ont toutes un coefficient Lipschitz $\gamma = 1$; ensuite, leur paramètre $\mu \geq 0$ permet de quantifier facilement l'aversion au risque qu'elles convoient, $\mu \to \infty$ correspondant à une attitude neutre au risque et $\mu = 0$ correspondant à l'attitude extrêmement averse où aucune utilité n'est accordée aux rendements supérieurs à zéro. Mathématiquement, les fonctions exponentielles Lipschitz sont définies par

$$\mathit{LEU}_\mu(r) = \begin{cases} r & r < 0 \\ \mu(1 - e^{-r/\mu}) & r \geq 0 \end{cases}. \tag{175}$$

La fonction d'utilité inverse $\mathit{LEU}_\mu^{-1} : \boldsymbol{U} \to \boldsymbol{R}$, nécessaire pour exprimer en terme de rendement équivalent les erreurs exprimées en util, est illustrée à la Fig. 3 (p. 38). On peut vérifier algébriquement que

$$\mathit{LEU}_\mu^{-1}(r) = \begin{cases} r & r < 0 \\ -\mu \log(1 - r/\mu) & r \geq 0 \end{cases}. \tag{176}$$

Finalement, les bornes d'erreur de généralisation et de sous optimalité, lorsqu'elles sont exprimées en équivalent certain, font intervenir l'inverse du sous-gradient de u^{-1}. Dans le cas de l'utilité LEU, celui-ci correspond tout simplement à l'inverse de la dérivée de $\textit{LEU}_\mu$ et est donc donné par

$$\left(\frac{d}{dr}LEU_\mu^{-1}(r)\right)^{-1} = \begin{cases} 1 & r < 0 \\ e^{r/\mu} & r \geq 0 \end{cases}. \tag{177}$$

Régularisation Sauf exception, le facteur de régularsation $\lambda = 1/2$ sera employé au cours de toutes les expériences.

Loi de marché La loi de marché M sera construite en deux temps. D'abord, une loi de marché théorique $\tilde{M} \in \mathscr{R}^{p+1\times p+1}$ sera définie. Toutes ses marges seront décrites par des variables aléatoires Rademacher (retournant ± 1 avec probabilité $1/2$). La dépendance entre les marges sera modélisée à l'aide d'une copule gaussienne dont la matrice de corrélation sera définie en début de chaque section.

Puis, à partir de cette loi de marché théorique $\tilde{M}$, un échantillon fini $M \sim \tilde{M}^{5000}$ de 5000 points en sera tiré afin de former une loi de marché discrète M à partir de laquelle toutes les expériences seront réalisées. En quelque sorte, M fournira alors une approximation à $\tilde{M}$, mais permettra de déterminer exactement des variables comme l'utilité hors échantillon $\boldsymbol{EU}(q)$ d'une politique q ou l'utilité espérée optimale $\textit{EU}^\star$, qu'il serait autrement impossible à déterminer théoriquement (sauf dans le cas de l'utilité neutre au risque).

Précision de la borne et quantiles d'erreur Les bornes sur les erreurs présentées aux Théorèmes *[Citation needed]*s'appliquent à tout échantillon $\mathcal{S}_n$ avec une probabilité $1-\delta$. Elles s'appliquent donc, de façon équivalente, au $1-\delta$-ième quantile avec probabilité 1.

Ainsi, pour confirmer ces bornes, celles-ci seront évaluées à $\delta = 5\%$ et le 95$^{\text{e}}$ percentile d'erreur sera mesuré.

Échantillonnage Les échantillons d'entraînement $\mathcal{S}_n$ seront traités de façon équivalente pour toutes les sections.

Par exemple, à la Section 5.2, où c'est la taille de l'échantillon qui augmente linéairement, on tirera d'abord $m \times \bar{n}$ réalisations de M afin d'obtenir m échantillons d'entraînement $\mathcal{S}_{\bar{n}}$. Puis, on exposera progressivement à l'algorithme n des $\bar{n}$ points afin d'obtenir peu à peu une meilleure représentation de M. Les m points serviront à déterminer le 95$^{\text{e}}$ percentile d'erreur.

À la Section 5.3, où c'est la dimensionalité du problème qui varie, l'idée demeure la même, cette fois avec n fixe et p variable. On tire donc tout d'abord $m \times \bar{n}$ réalisations

de M. Comme chacune de ces réalisations est consituée de $\bar{p}$ variables d'information, où $\bar{p}$ note le nombre de marges d'information de M, on a alors qu'à présenter à l'algorithme des réalisations "incomplètes", dont seules les p premières dimensions sont connues.

Enfin, à la section 5.4, la situation est un mélange des deux précédentes, où de plus en plus de points provenant d'un même échantillon sont présentés à l'algorithme, leur dimension dévoilée progressant en fonction de n.

Toutes nos expériences disposeront de $m = 100$ échantillons d'entraînement.

Environnement de calcul L'identification numériques des politiques optimales $\hat{q}$ se fera à partir de l'implémentation CVXPY[DB16] et du solveur ECOS[DCB13]. Les calculs numériques se feront à partir de la librairie BLAS et de l'interface NUMPY.

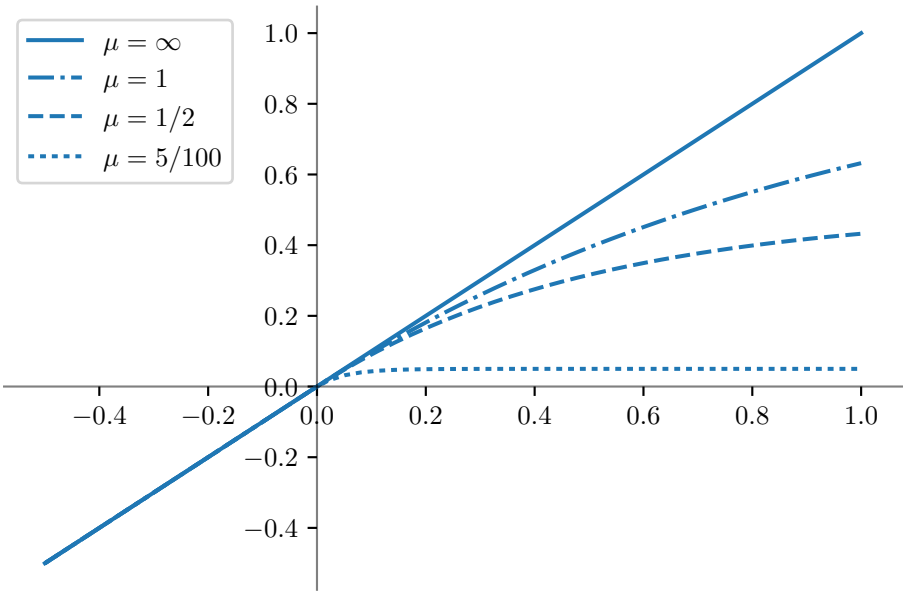

FIGURE 2 – Fonctions d'utilité exponentielles Lipschitz (LEU)

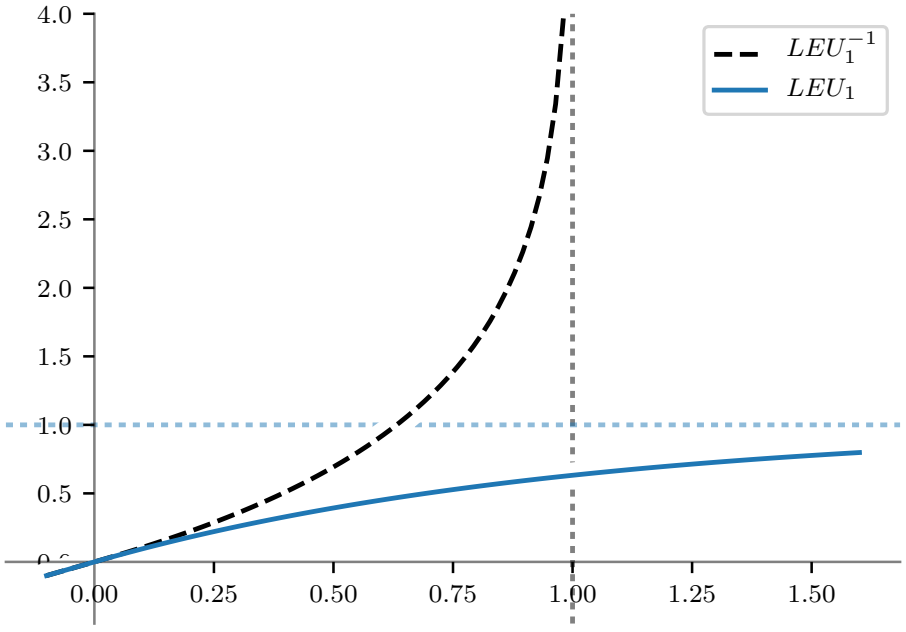

FIGURE 3 – Utilité et utilité inverse

5.2 *n* variable, *p* constant

L'objet de cette section est l'étude du cas canonique où la taille n de l'échantillon $\mathcal{S}_n$ augmente linéairement.

Loi de marché Tel qu'expliqué à la Section 5.1, une loi de marché discrète M sera dérivée d'une loi théorique M. Cette loi théorique disposera ici de trois marges (deux variables d'information et une variable de rendement, toutes trois Rademacher). La loi théorique de marché M sera modélisé à partir de la matrice corrélation Σ donnée par

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ R \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & R \end{array} \\ \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \end{array}, \tag{178}$$

où $\rho = 1/\sqrt{2}$, ce qui correspond à la plus grande valeur de corrélation permettant à Σ d'être semi-définie positive. Ainsi, X_1 et X_2 seront mutuellement indépendants, mais auront toutes deux une influence égale sur la réalisation de R (la corrélation entre X_j et R correspond au tau de Kendall : $\boldsymbol{Corr}(X_j, R) = \frac{2}{\pi}\arcsin(\rho) = 1/2$. Voir [Rém13] pour des précisions).

On en déduit évidemment que $\xi = \sqrt{2}$ et $\bar{r} = 1$.

5.2.1 Erreur de généralisation

On rappelle que l'erreur de généralisation d'une politique d'investissement q consiste à mesurer la différence entre l'utilité (resp. équivalent certain) espérée observée en échantillon avec l'utilité (resp. équivalent certain) espérée hors échantillon, ou, mathématiquement, de déterminer $\widehat{\boldsymbol{EU}}(q) - \boldsymbol{EU}(q)$ (resp. $\widehat{\boldsymbol{CE}}(q) - \boldsymbol{CE}(q)$).

Quantiles d'erreur – Figure 4 Tout d'abord, la Fig. 4 (p. 41) indique les quantiles d'erreur de généralisation des m échantillons d'entraînement, incluant la valeur maximale et minimale pour chaque n. Nos bornes ne donnent que des garanties partielles sur les maximums ni sur les minimums, cependant il est intéressant d'observer le comportement convergenant vers zéro de chacun des quantiles d'erreur.

Erreur de généralisation et aversion au risque – Figure 5 Par ailleurs, si intuitivement on peut s'attendre à observer une relation entre l'erreur de généralisation et l'aversion au risque, la Fig. 5 (p. 42) montre qu'effectivement, une plus forte aversion au risque (caractérisée par μ) entraîne une erreur de généralisation plus faible, alors qu'au contraire, une faible aversion au risque entraîne une erreur de généralisation plus importante. Cette relation est importante puisqu'elle permet de généraliser les observations empiriques faites à partir d'une seule utilité à d'autres utilités. Dans les expériences suivantes, l'utilité étalon sera celle caractérisée par un coefficient $\mu = 1$.

Borne sur l'erreur – Figures 6 et 7 La Fig. 6 (p. 43) permet de constater la validité des garanties théoriques offertes par l'algorithme d'investissement. On constate ici que la borne n'est pas exactement serrée, les courbes théoriques et empiriques différant d'un ordre de grandeur. Néanmoins, il faut conserver à l'idée que ces bornes sont valides pour toute distribution de marché M de dimension $\xi \leq \sqrt{2}$ et $\bar{r} \leq 1$ et toute courbe d'utilité u de coefficient Lipschitz 1. C'est toutefois avec cette forme particulière de M (marges Rademacher) qu'on a pu observer les bornes plus serrées.

Ceci dit, si les bornes ne sont en tant que telles pas particulièrement fortes, l'ordre $\mathcal{O}(n^{-1/2})$ qu'elles indiquent est lui très bien respecté empiriquement et il pourrait donc être possible d'anticiper de combien l'erreur empirique peut décroître selon la taille de l'échantillonnage en interpolant tout simplement les erreurs déjà observées avec un polynôme $\mathcal{O}(n^{-1/2})$.

Erreur en util et en équivalent certain Quant à l'erreur théorique et sa borne, on constate qu'il y a en fait peu de distorsion entre le domaine de l'util et celui du rendement. Pour expliquer ce phénomène, la Fig. 7 (p. 44) décompose l'erreur de généralisation : d'une part sa partie erreur en échantillon et hors échantillon. On y observe que les valeurs obtenues en util, toutes inférieures à 0.6, entraînent une faible distorsion si on se fie à la Fig. 3 (p. 38).

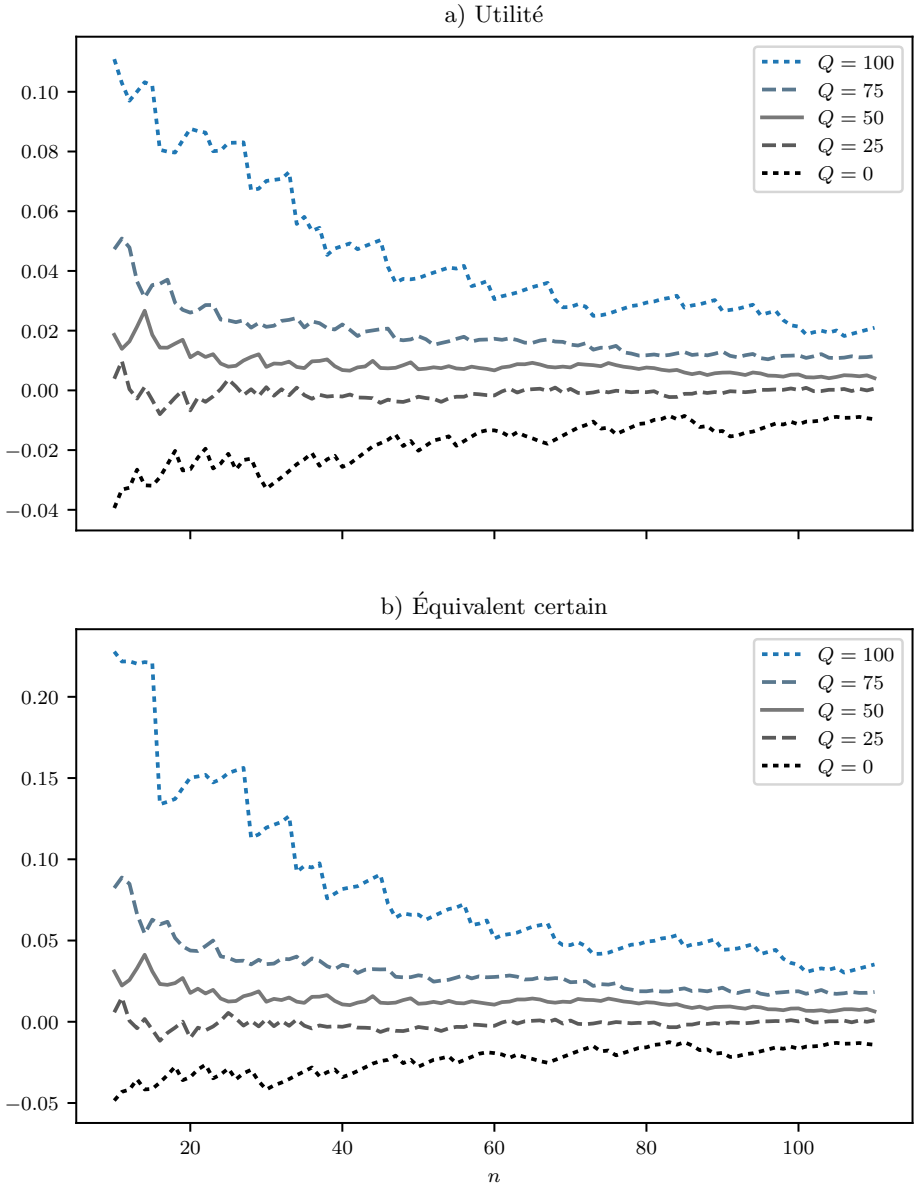

FIGURE 4 – Quartiles et valeurs maximales de l'erreur de généralisation

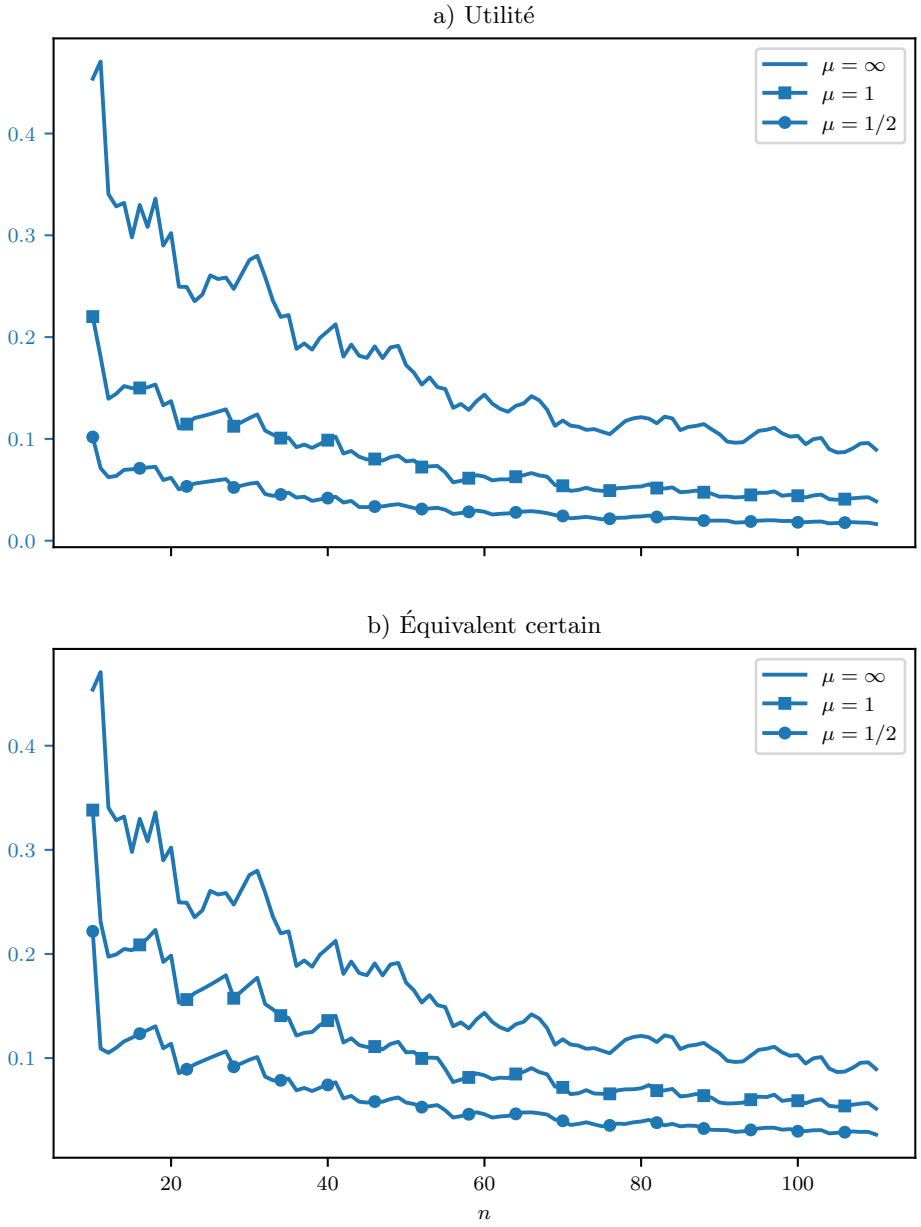

FIGURE 5 – Aversion au risque et erreur de généralisation (95^{e} percentile)

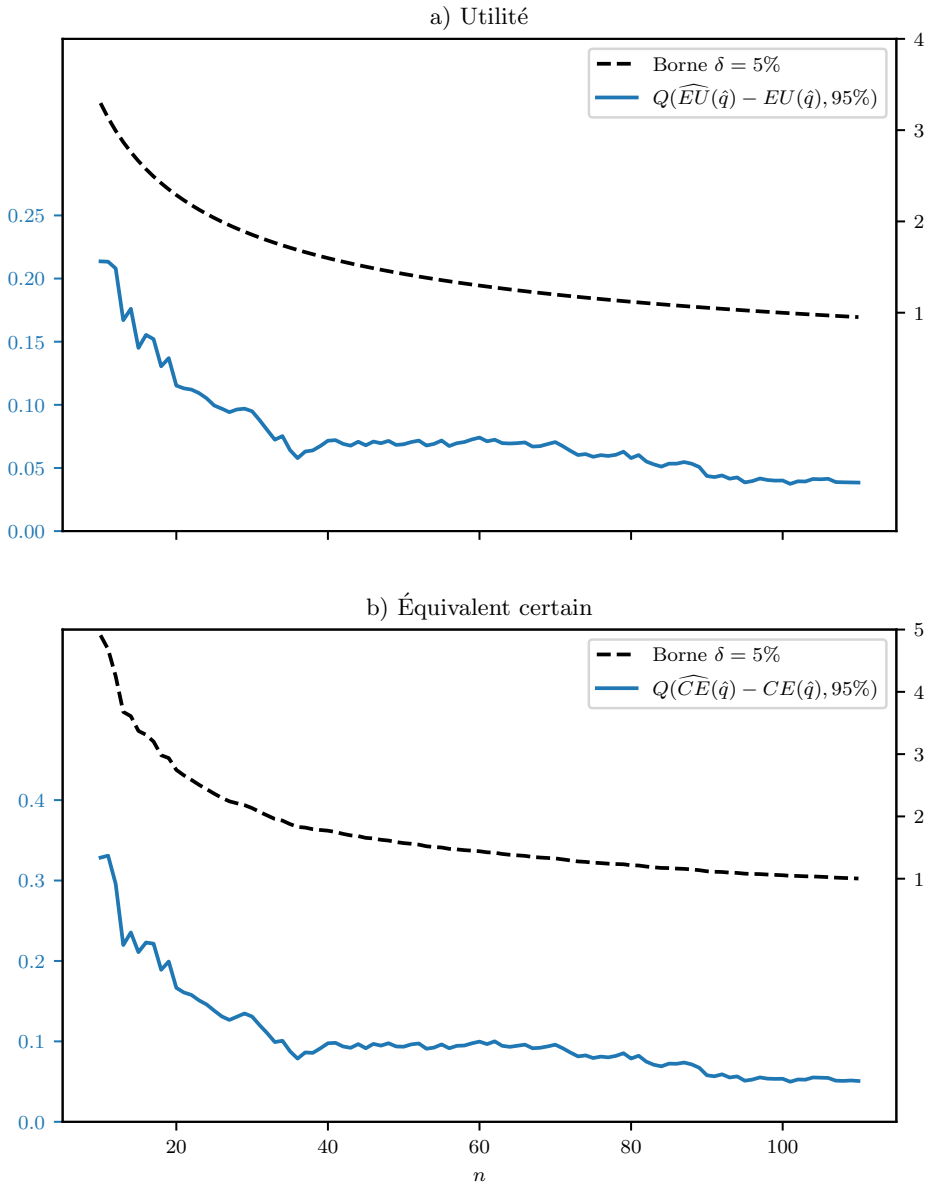

FIGURE 6 – Borne sur le 95e percentile de l'erreur de généralisation. L'axe de gauche indique la valeur de l'erreur empirique, tandis que l'axe de droite indique celle de la borne théorique.

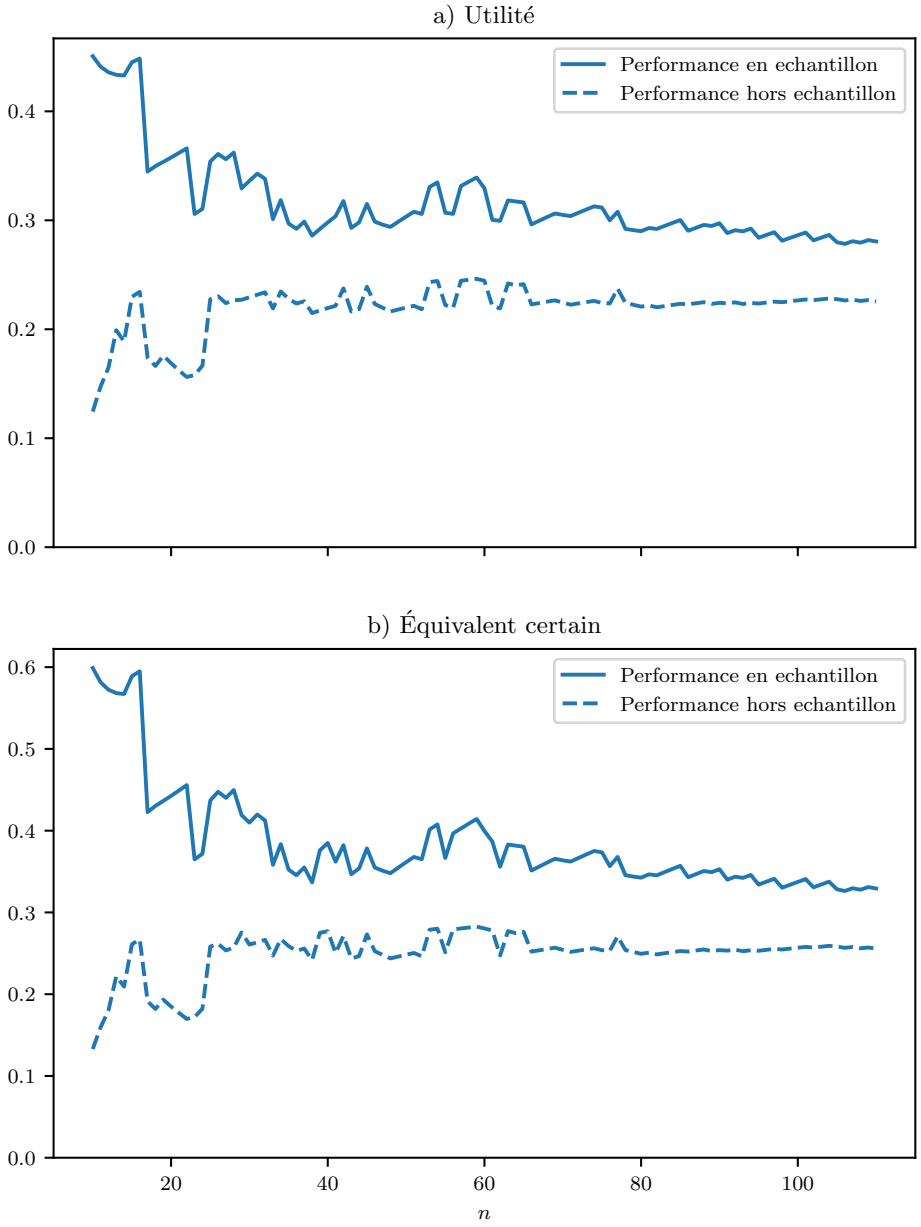

FIGURE 7 – Composantes en échantillon et hors échantillon de l'erreur maximale.

5.2.2 Erreur de sous optimalité

λ constant – Figure 8 Contrairement à l'erreur de généralisation, l'erreur (en util) de sous optimalité $\boldsymbol{EU}(q^\star) - \boldsymbol{EU}(\hat{q})$ (resp. $\boldsymbol{CE}(q^\star) - \boldsymbol{CE}(\hat{q})$ dans le domaine des rendements) ne bénéficie pas d'une convergence vers zéro du fait de la présence du terme de régularisation dans l'algorithme $\mathcal{Q}(\mathcal{S}_n)$. En fait, tel que vu au théorème *[Citation needed]*, pour $n \to \infty$, la meilleure borne qu'on puisse avoir est proportionelle à $\lambda\|q^\star\|^2$ (domaine des utils).

Par exemple la Fig. 8 (p. 46) illustre précisément comment les erreurs empiriques et théoriques plafonnent toutes les deux à des constantes non nulles. De plus, contrairement à la borne de généralisation, la borne de sous-optimalité se trouve à deux ordres de grandeur de l'erreur empirique. Ceci a un effet particulièrement néfaste lorsqu'on considère la borne dans le domaine des rendements où l'effet de l'utilité inverse (voir Fig. 3 (p. 38)) se fait violemment sentir : puisque l'utilité de la politique optimale est proche de la limite asymptotique $\lim_{r\to\infty} \mathrm{LEU}_1(r) = 1$, l'inversion $u^{-1}(EU^\star)$ retourne une valeur très élevée.

Néanmoins, tout comme c'était le cas pour l'erreur de généralisation, si la borne de sous optimalité ne donne pas nécessairement de fortes garanties, en revanche elle suggère un ordre de convergence qui lui semble être en adéquation avec l'erreur de sous optimalité empirique maximale (ou plutôt, avec son 95$^{\text{e}}$ percentile d'erreur).

λ décroissant – Figure 9 Comme il fut discuté à la section *[Citation needed]*, en utilisant un facteur de régularisation $\lambda = \mathcal{O}(n^{-1/2})$, on peut garantir une convergence de l'erreur de sous optimalité vers zéro (voir Fig. 9 (p. 47)). Cependant, si dans ce cas-ci l'erreur empirique de sous-optimalité, qu'elle soit exprimée en util ou en rendement, semble bien converger à un rythme $\mathcal{O}(n^{-1/2})$, la borne théorique elle ne progresse qu'à un rythme de $\mathcal{O}(n^{-1/4})$, ce qui est particulièrement lent. Par contre, même une progression aussi lente permet quand même d'obtenir des garanties d'équivalent certain un peu plus raisonnables puisqu'on force alors la limite $\lambda\|q^\star\|^2$ de la borne en util à s'éloigner de la région où u^{-1} retourne des valeurs très grandes.

Par contre, si cette décroissance de λ entraîne une convergence de l'erreur de sous optimalité, c'est au prix de la garantie sur l'erreur de généralisation, qui est elle proportionelle à $\mathcal{O}(\lambda^{-1})$.

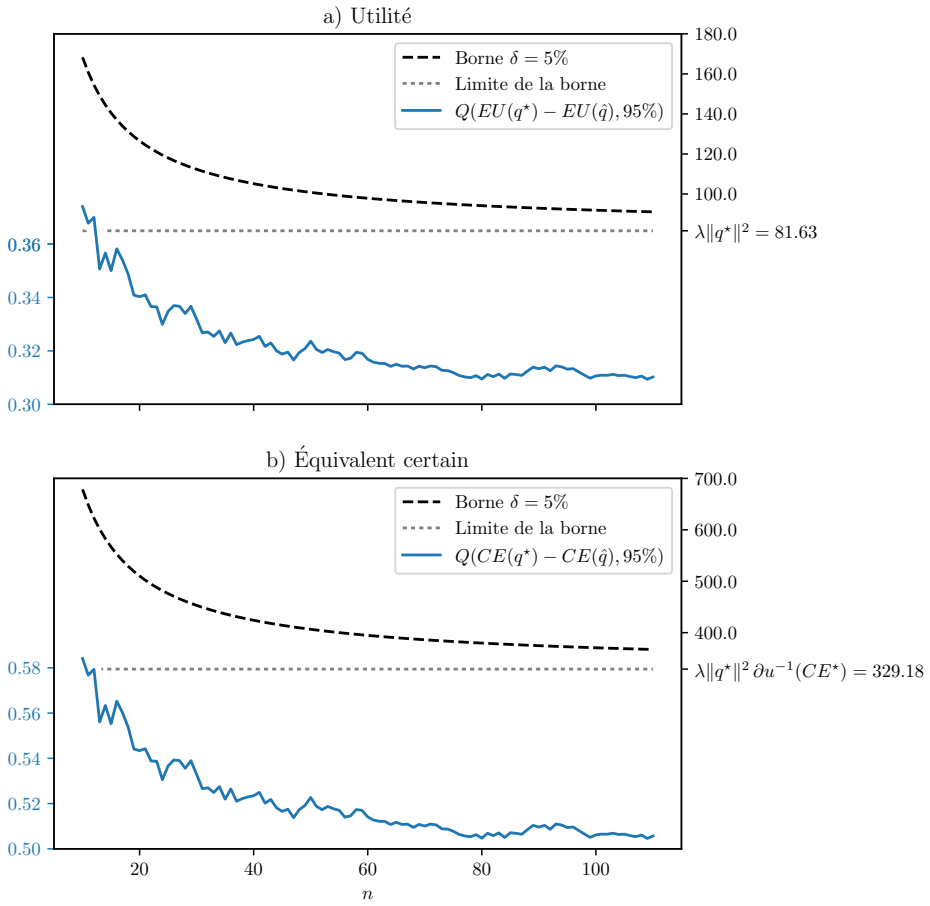

FIGURE 8 – Borne de sous optimalité, λ constant

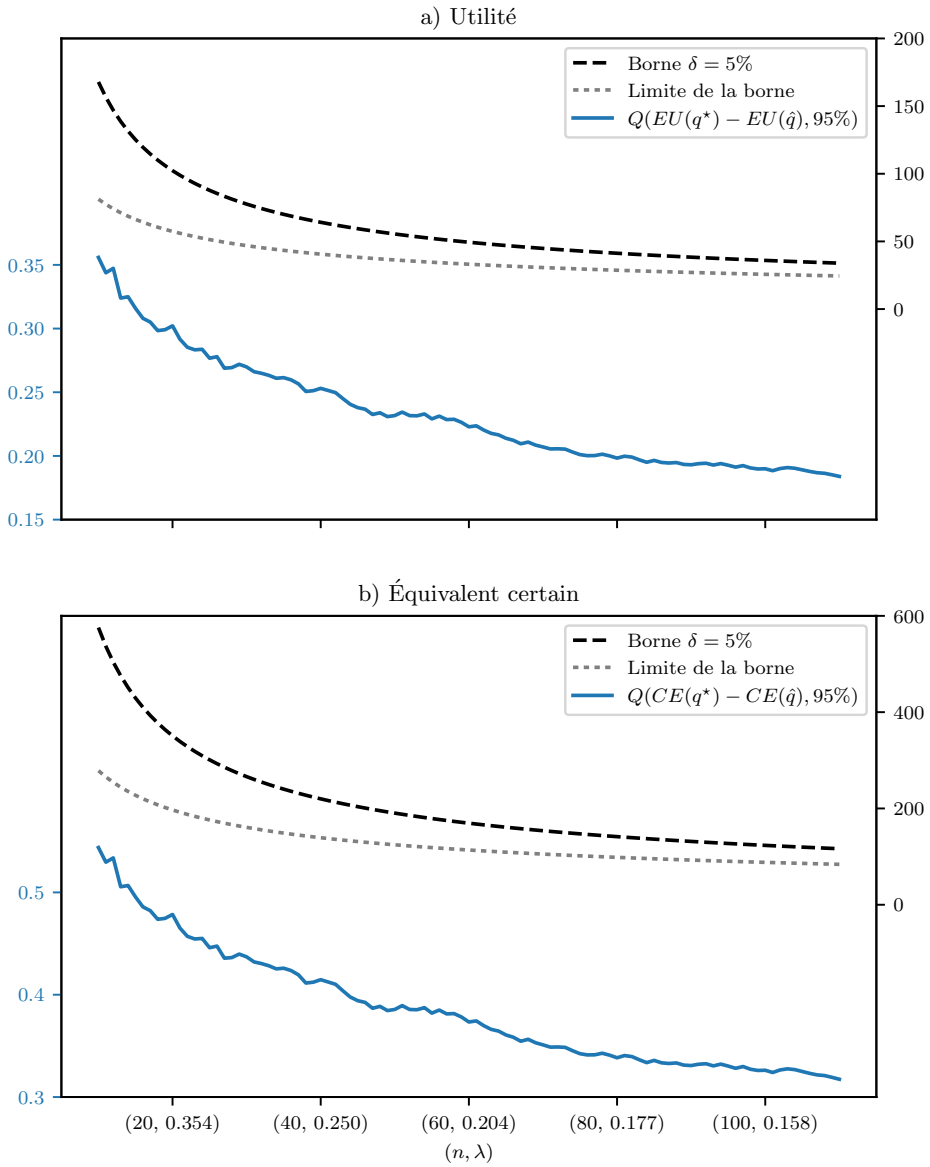

FIGURE 9 – Borne de sous optimalité, λ décroissant

5.3 *n* constant, *p* variable

On peut aussi considérer le rapport qu'entretiennent les bornes de généralisation et de sous-optimalité de l'algorithme de maximisation d'utilité régularisé lorsqu'on ajoute de nouvelles informations indépendantes des précédantes, tout en conservant la taille d'échantillonnage constante.

Protocole d'expérience Afin de bien comprendre l'effet que peut avoir un régime en haute dimension sur les deux types d'erreur étudiées, une copule gaussienne avec marges Rademacher sera encore employée pour modéliser $\tilde{M}$. Cette fois cependant, cette copule disposera de $\bar{p}$ marges d'information indépendantes dont chaque marge sera dévoilée progressivement. De plus, trois situations seront étudiées : celle où toute l'information est concentrée à la première marge, les autres étant indépendantes de R, celle où chaque marge dispose d'une corrélation de $1/\sqrt{\bar{p}}$ avec R (information dispersée) et finalement celle où aucune information n'est présente pour déterminer R, c'est-à-dire que toutes les marges X_j sont indépendantes de R.

Mathématiquement, $\tilde{M}$ est donc décrit par une copule gaussienne à $\bar{p}+1$ marges Rademacher dont la matrice de corrélation est paramétrée par un vecteur de corrélation $\rho \in \mathscr{R}^{\bar{p}}$:

$$\Sigma = \begin{array}{c} \\ X_1 \\ \vdots \\ X_{\bar{p}} \\ R \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & \cdots & X_{\bar{p}} & R \end{array} \\ \begin{pmatrix} \ddots & & & | \\ & I_{\bar{p}\times\bar{p}} & & \rho \\ & & \ddots & | \\ — & \rho & — & 1 \end{pmatrix} \end{array}. \tag{179}$$

Le cas de l'information concentrée se traduira par un vecteur de corrélation donné par

$$\rho = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}, \tag{180}$$

celui de l'information dispersée par le vecteur de corrélation

$$\rho = \begin{pmatrix} 1/\sqrt{\bar{p}} & \cdots & 1/\sqrt{\bar{p}} \end{pmatrix}, \tag{181}$$

et celui sans aucune information par le vecteur de corrélation

$$\rho = \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix}. \tag{182}$$

Enfin, les expériences qui suivent fixent le nombre total de variables d'information à $\bar{p} = 50$.

Erreur de généralisation – Figure 10 On a déjà remarqué que la borne de généralisation affiche une croissance $\mathcal{O}(\xi^2)$ ce qui, dans le cas d'un noyau linéaire, devrait se traduire par une progression linéaire $\mathcal{O}(p)$. Ainsi, la Fig. 10 (p. 50) montre effectivement une telle progression pour les trois situations énumérées à la section précédante.

En fait, la plupart des observations qui ont été faites dans le cas où p est constant et n est variable peuvent être réutilisées. Par exemple, on remarque que la même différence d'un ordre de grandeur entre l'erreur empirique et théorique persiste à mesure qu'on dévoile de nouvelles variables d'informations X_j. Et ici encore, on perçoit une faible dilatation de valeur entre la borne exprimée en util et en rendement.

Pour ce qui concerne les trois situations d'information, bien que chacune d'entre elles affichent à peu près la même progression linéaire, la situation où toute l'information est concentrée dès $p = 1$ entraîne d'abord une erreur nulle, qui augmente à mesure que de nouvelles variables "de bruit" sont ajoutées. On remarquera par ailleurs la forte similarité entre les courbes *Information dispersée* et *Aucune information*. En effet, comme $\bar{p}$ est assez important, lorsque $p = 1$ et que le signal est dispersé, le signal perçu à partir d'une seule caractéristique est très faible. Néanmoins, la courbe d'information diluée fléchit par rapport à celle de l'absence complète d'information à mesure que p augmente vers $\bar{p}$, conformément à l'intuition qu'on pourrait en avoir.

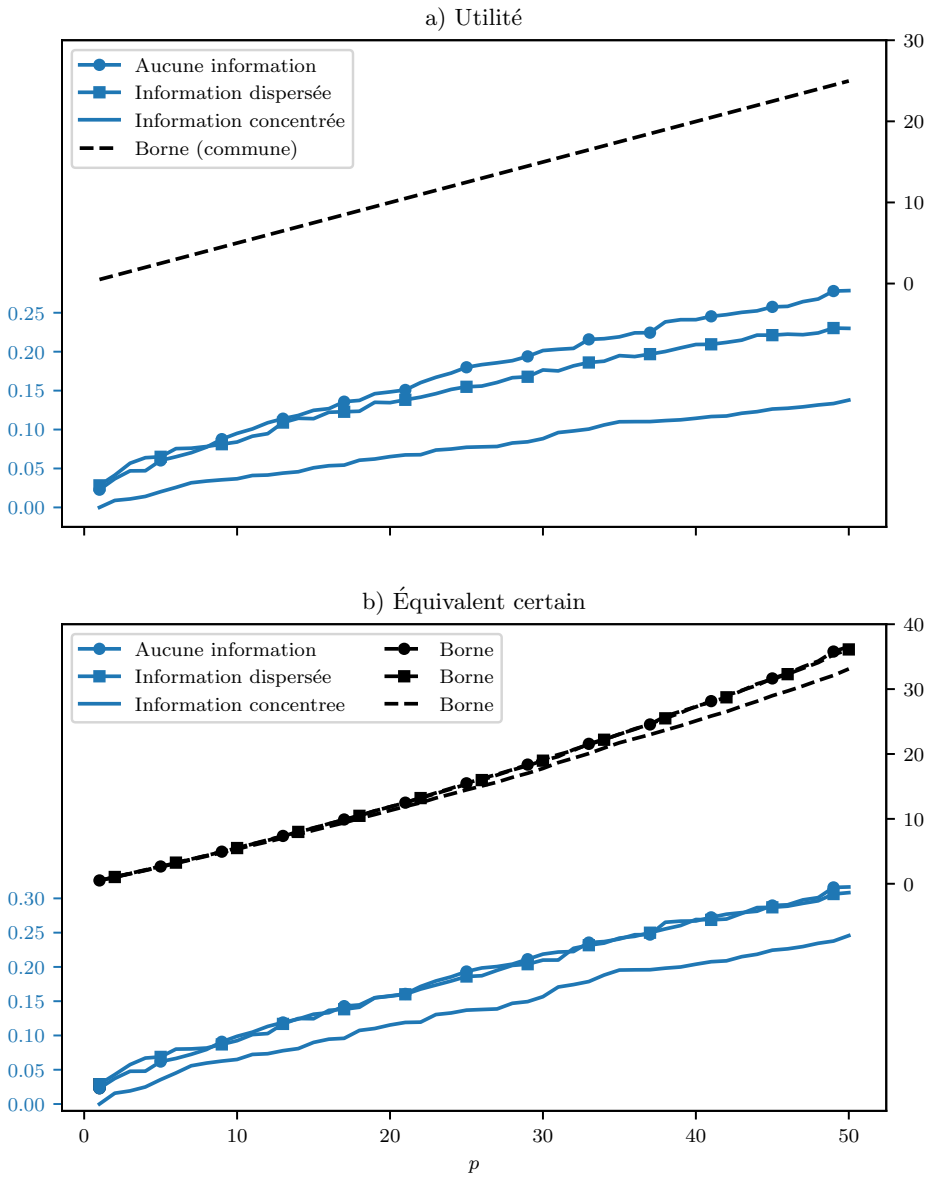

FIGURE 10 – Erreur de généralisation avec ajout d'information

5.3.1 Sous optimalité

Utilité espérée optimale Dans le cas où on ajoute de l'information, la sous optimalité, contrairement à l'erreur de généralisation, peut référer à deux types d'erreur. Soit on compare la performance hors échantillon de $\hat{q}$ à celle de la politique optimale qui ne dispose que de $p < \bar{p}$ variables d'information, soit à la politique optimale qui dispose de $\bar{p}$ variables d'information nécessaires pour décrire M. Cependant, le développement théorique qui a été mené au cours de la dernière section ne s'est implicitement préoccupé que de la première situation.

Par exemple, la Fig. 11 (p. 52) indique la progression de l'utilité espérée optimale à mesure que de nouvelles variables d'information sont dévoilées. On y observe sans surprise que le cas où toute l'information est disponible dès $p = 1$ affiche une utilité espérée optimale constante, alors qu'on a une progression à peu près linéaire lorsqu'on dévoile progressivement des variables d'information qui sont chacunes faiblement corrélées à R, mais indépendantes l'une à l'autre. Enfin, aucune information se traduit inévitablement par une utilité espérée optimale nulle.

Erreurs de sous optimalité La Fig. 12 (p. 53) elle, indique la sous optimalité relative à mesure que de nouvelles variables d'information sont dévoilées. On y note d'abord une progression de l'erreur pour les trois situations étudiées ; de plus, la différence entre le panneau a) de la Fig. 12 et la Fig. 11 est une manifestation de la présence du facteur de régularisation constant à mesure que p augmente. Finalement, si les courbes *information dispersée* et *aucune information* subissent peu de distortion entre le domaine des utils et celui des rendements, la courbe *information concentrée* affiche une énorme sous optimalité lorsqu'elle est exprimée en rendement. Cela s'explique par le fait que $EU^\star$ est très proche de 1 (numériquement $1 - EU^\star = 1.89 \times 10^{-10}$ lorsque $p = 50$), ce qui entraîne un équivalent certain de l'ordre de 10, puisque $CE^\star = -\log_e(1.89 \times 10^{-10})$.

Borne sur l'erreur de sous optimalité – Figures 13, 14 et 15 Ces trois figures traduisent comment la borne théorique sur l'erreur de sous optimalité se comporte pour chacune des trois situations explorées ici. Tout d'abord, on remarque que pour chacune d'elle la courbe théorique permet bel et bien de borner la courbe empirique. Cependant, comme c'était le cas pour l'erreur de généralisation, on observe à nouveau un décalage de deux ordres de grandeur entre les courbes. Cependant, le caractère *linéaire* qu'annonce la borne théorique semble se matérialiser empiriquement. Pour ce qui concerne la borne de la situation avec information concentrée dès la première variable d'information, on constate qu'exprimer la sous optimalité en terme d'équivalent certain peut donner lieu à des situations aberrantes, puisqu'on obtient en effet une borne théorique de l'ordre de 10^{13}.

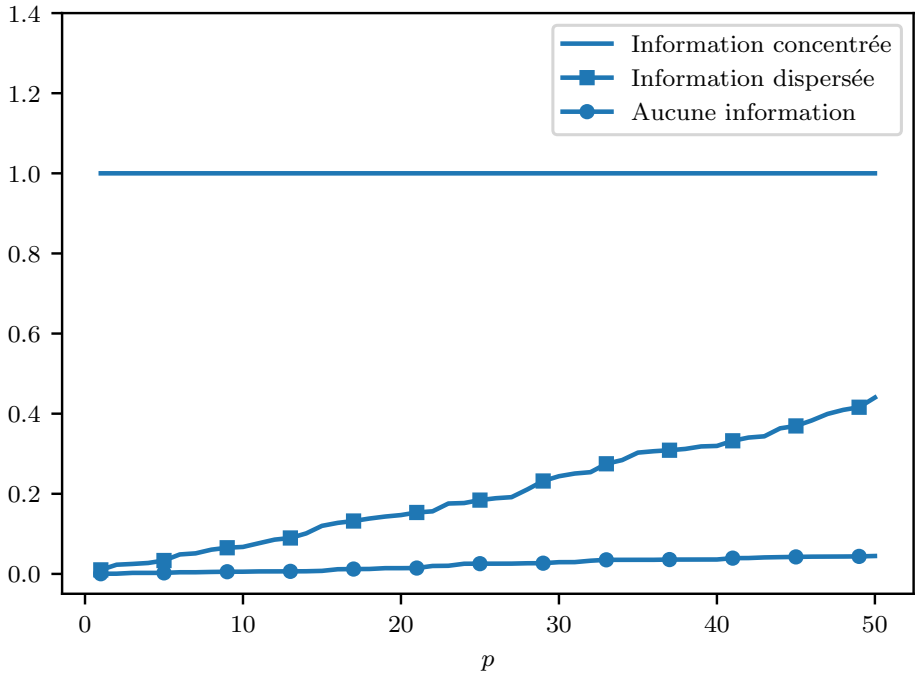

FIGURE 11 – Progression de $EU^{\star}$ relatif

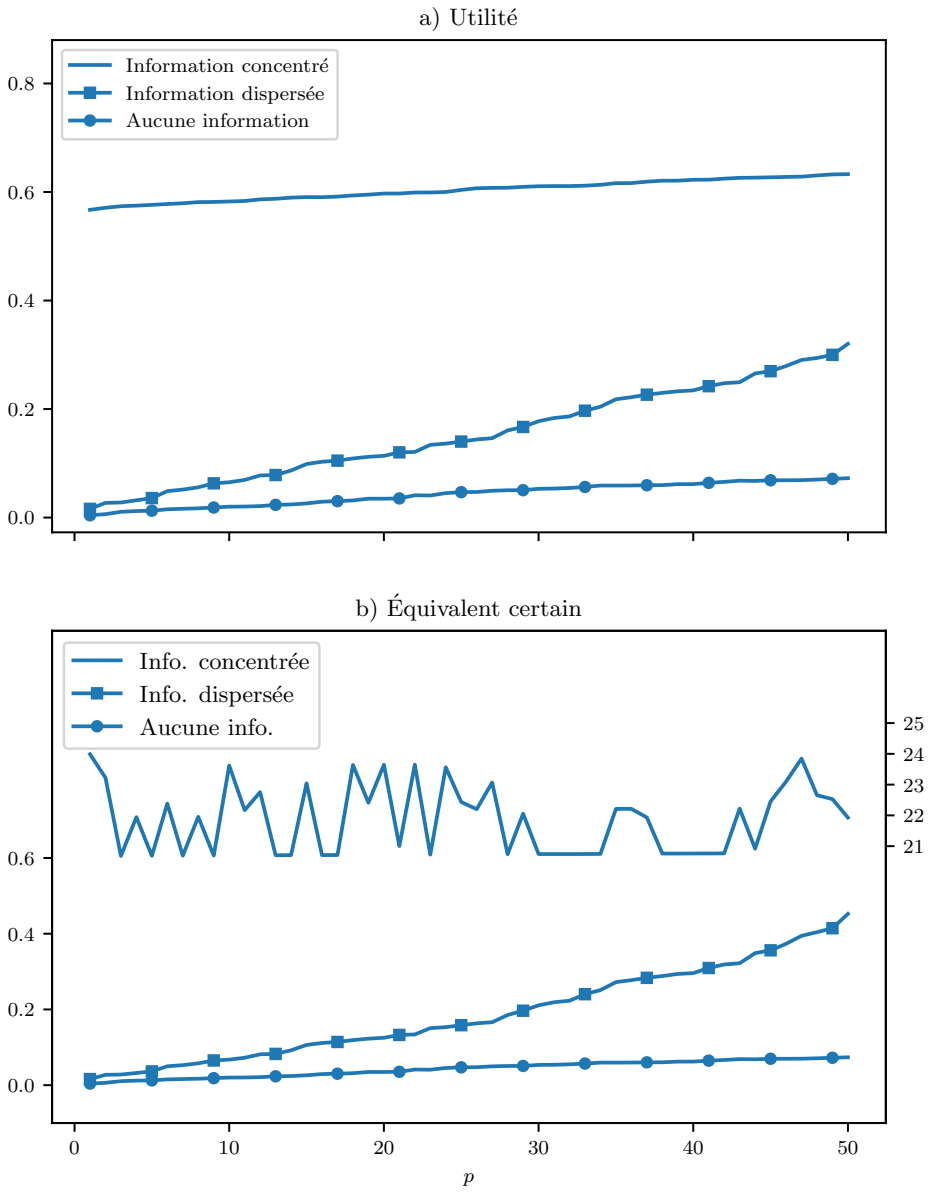

FIGURE 12 – Progression de la sous optimalité relative

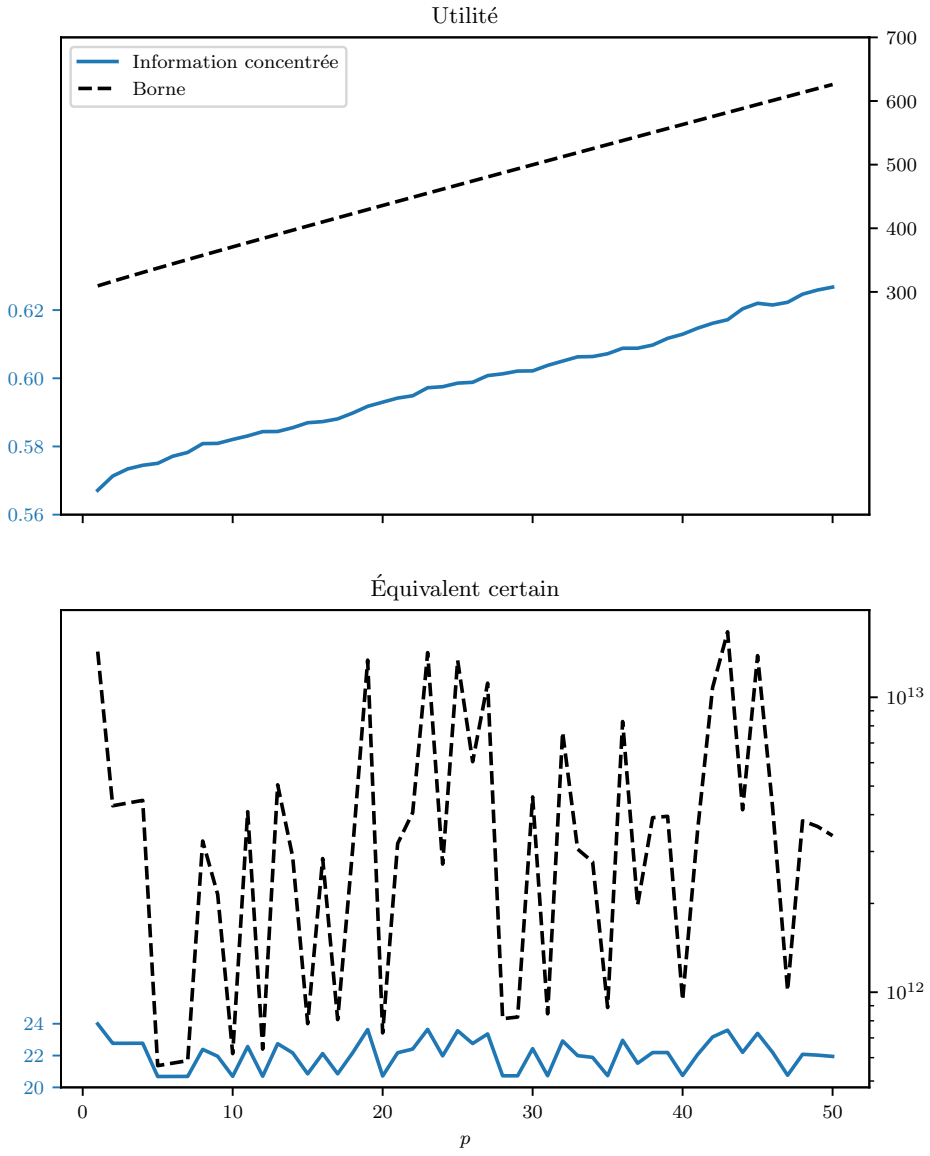

FIGURE 13 – Borne sur l'erreur de sous optimalité, information concentrée

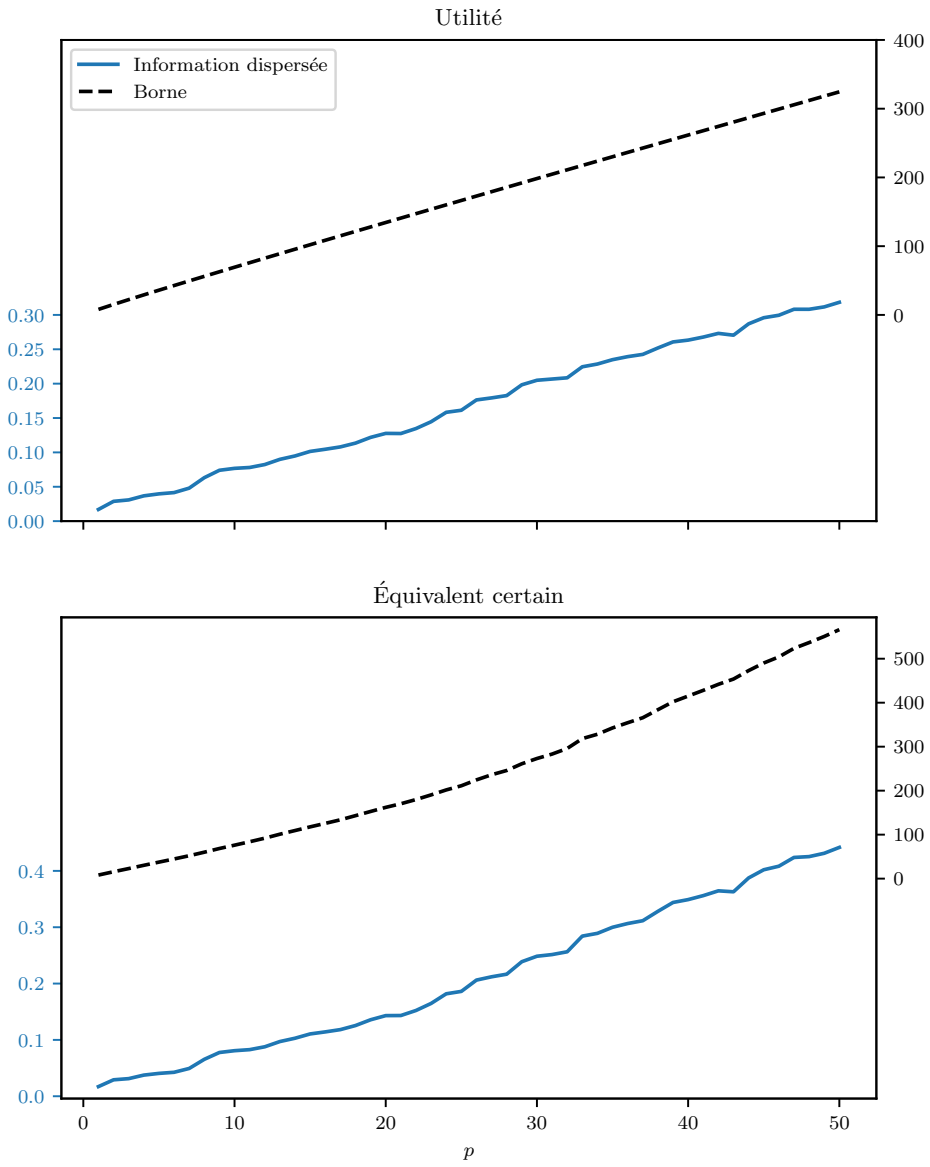

FIGURE 14 – Borne sur l'erreur de sous optimalité, information dispersée

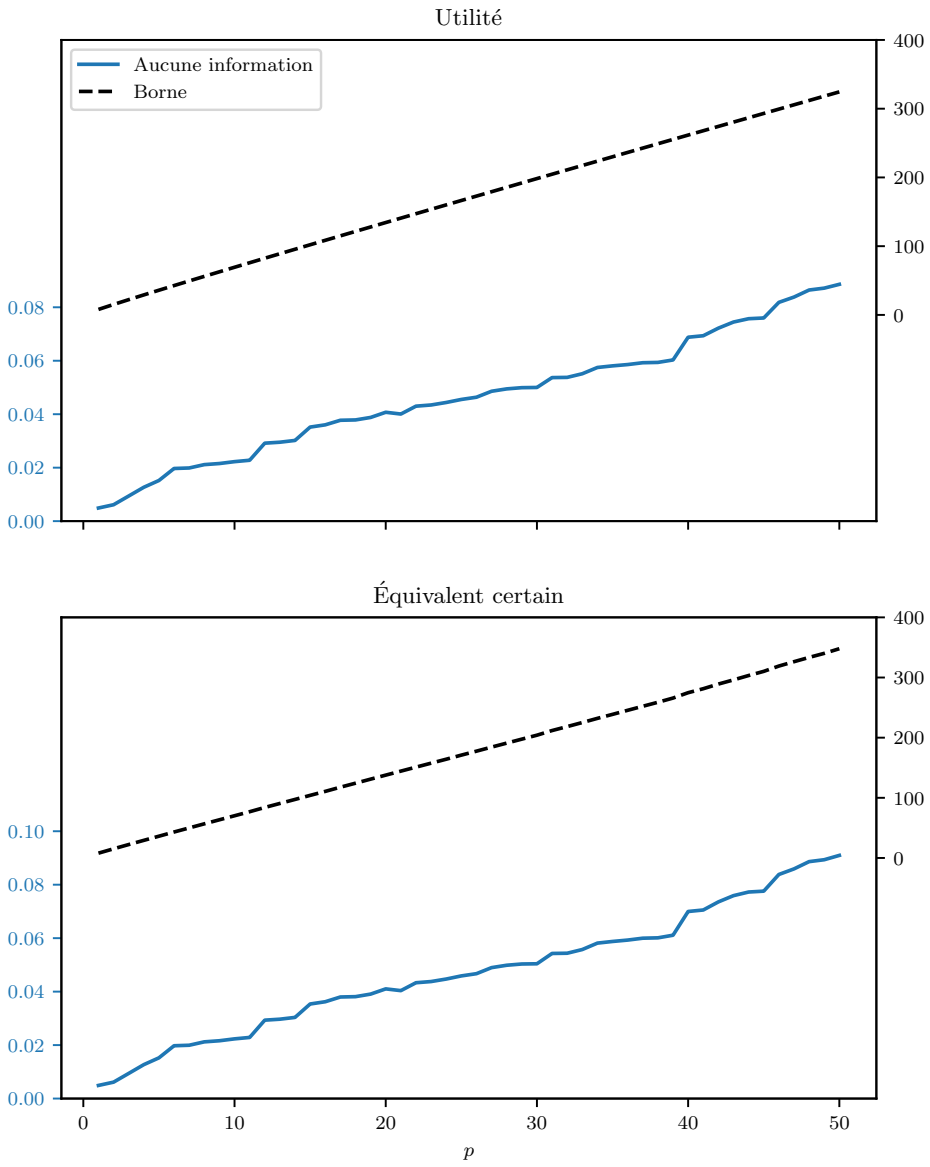

FIGURE 15 – Borne sur l'erreur de sous optimalité, aucune information

5.4 Ajout d'information et d'échantillons

Cette section cherche à illustrer les comportements des deux bornes lorsqu'on est en présence de régimes dynamiques en n et en p, *i.e.* lorsque $p = (n^k)$.

Méthodologie Cette fois, la loi de marché théorique $\tilde{M}$ disposera de $\bar{p} = 100$ variables d'information, encore liées à R à partir d'une copule gaussienne Σ donnée par

$$\Sigma = \begin{array}{c} \\ X_1 \\ \vdots \\ X_{\bar{p}} \\ R \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & \cdots & X_{\bar{p}} & R \end{array} \\ \begin{pmatrix} \ddots & & & | \\ & I_{\bar{p}\times\bar{p}} & & \rho \\ & & \ddots & | \\ — & \rho & — & 1 \end{pmatrix} \end{array}. \tag{183}$$

où cette fois, $\rho = \begin{pmatrix} 1/\sqrt{100} & \cdots & 1/\sqrt{100} \end{pmatrix}$.

Trois régimes seront étudiés : celui où $p = \mathcal{O}(n^{1/2})$, $p = \mathcal{O}(n^{3/4})$ et $p = \mathcal{O}(n)$. À des fins de simplicité, chacun de ces régimes sera décrit par la puissance k de n, de manière à avoir $k = 1/2, 3/4$ et 1. La figure 16 illustre précisément les trois régimes.

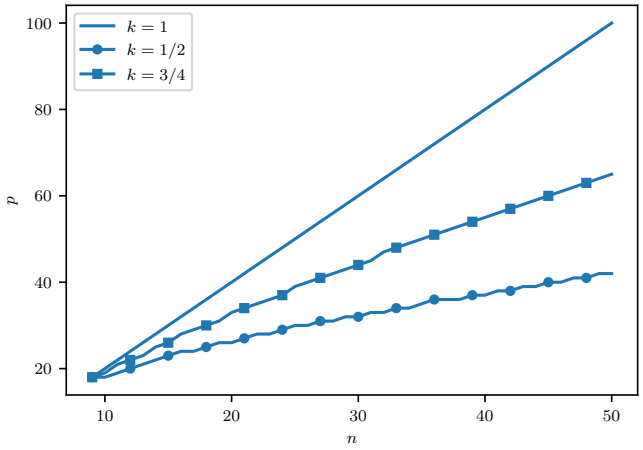

FIGURE 16 – Progression de p par rapport à n

5.4.1 Erreur de généralisation

Borne de généralisation – Figure 17 La Fig. 17 (p. 59) illustre enfin le problème avec la théorie présentée jusqu'à présent. Jusqu'à présent, même si les bornes théoriques dérivées n'étaient pas nécessairement serrées, elles servaient tout de même de guide pour connaître le régime dans lequel l'erreur empirique progressait. Par exemple,

lorsque p était fixe, on obtenait systématiquement une erreur de généralisation décroissant avec un rythme $\mathcal{O}(n^{-1/2})$. De la même façon, avec n constant, ces mêmes erreurs évoluaient selon un régime $\mathcal{O}(p)$. On se serait donc attendu à ce qu'on obtienne un régime global d'erreur $\mathcal{O}(p/\sqrt{n})$. Cependant, la Fig. 17 (p. 59) démontre clairement que tel n'est pas le cas, puisque peu importe le régime, là où on se serait attendu soit au moins à une erreur constante, on observe plutôt une diminution de l'erreur de généralisation !

Régime d'ordre plus élevé – Figure 18 Il faut effectivement attendre de voir apparaître des régimes très élevés (par exemple $k = 3/2$) avant d'enfin observer une augmentation de l'erreur. Cependant, l'ordre exact k ne semble pas clair et certaines expériences laissent croire que k pourrait plutôt dépendre des paramètres du problème.

5.4.2 Erreur de sous optimalité

La Figure 19 présente quant à elle l'erreur de sous optimalité pour les trois régimes étudiés et une fois encore, on constate une vive différence entre l'ordre donné par la borne théorique et ce qu'on observe en pratique. La courbe empirique $k = 1/2$ à peu près stable laisse croire qu'on est en présence d'un régime $\mathcal{O}(p/\sqrt{n})$, mais la courbe $k = 1$ vient en fait démentir cette interprétation puisque la progression paraît alors linéaire, un peu comme ce qui était observé lorsque n était constant.

De plus, la forme même des bornes théoriques permet de constater les deux régimes théoriques superposés, c'est-à-dire de la forme $\mathcal{O}(p/n + p/\sqrt{n})$. Par ailleurs, la présence du terme de régularisation vient une fois de plus brouiller les interprétations, puisqu'il faut conserver à l'esprit qu'un terme $\lambda\|q^\star\|^2$ est aussi compris dans la borne, et dont la progression est encore mal comprise (voir Fig. 20 (p. 62)).

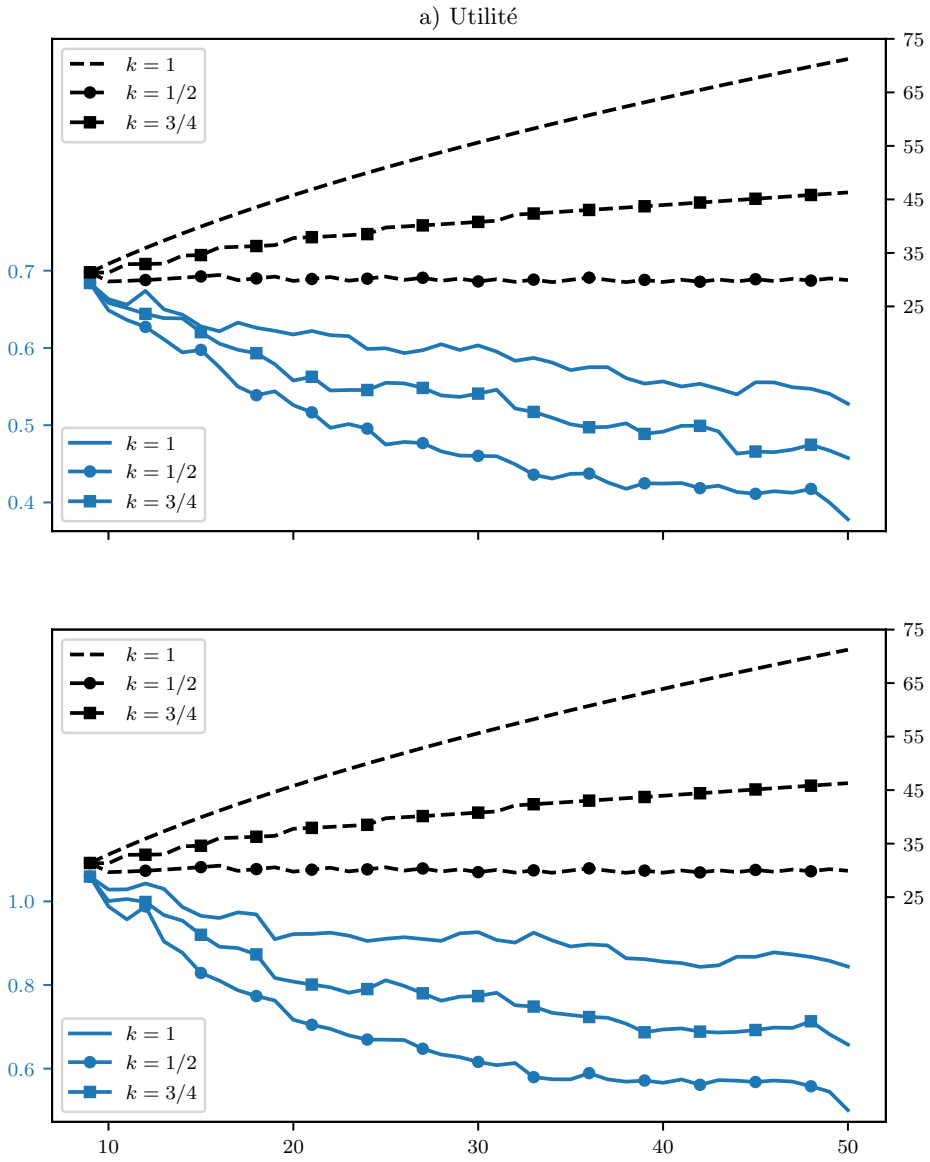

FIGURE 17 – Borne sur l'erreur de généralisation, n et p variables

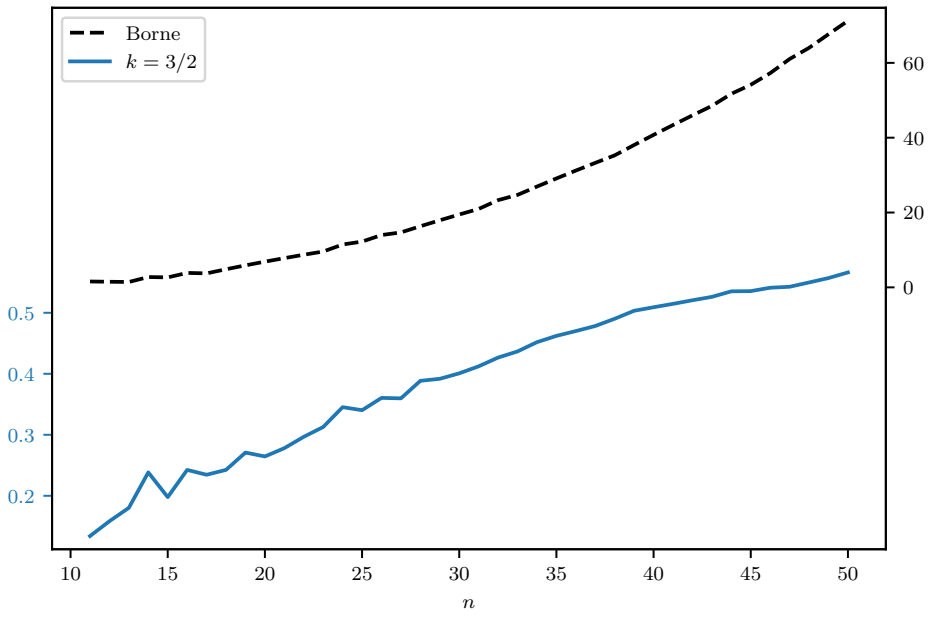

FIGURE 18 – Erreur de généralisation à très haut régime

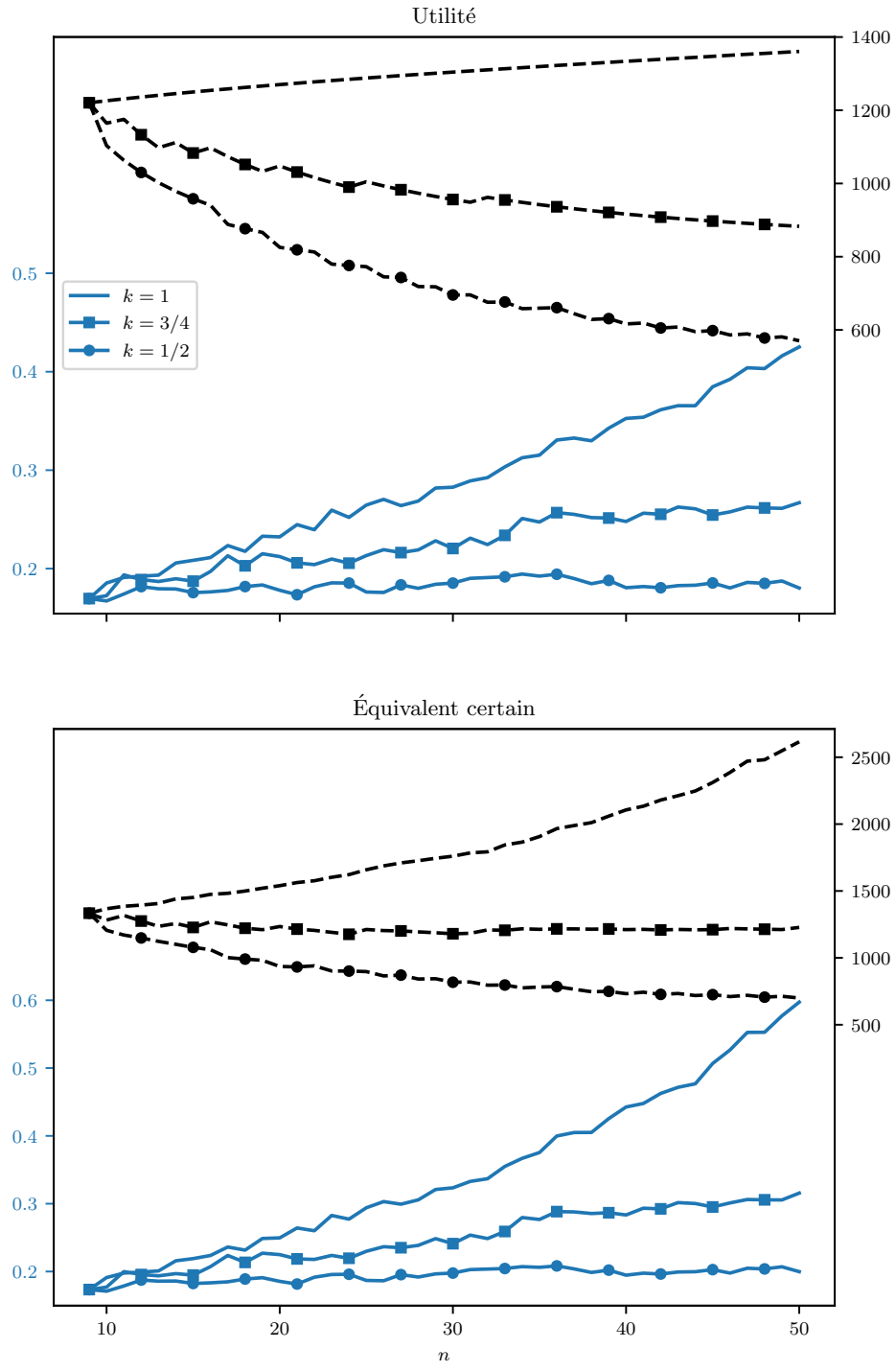

FIGURE 19 – Borne sur l'erreur de sous-optimalité, n et p variables

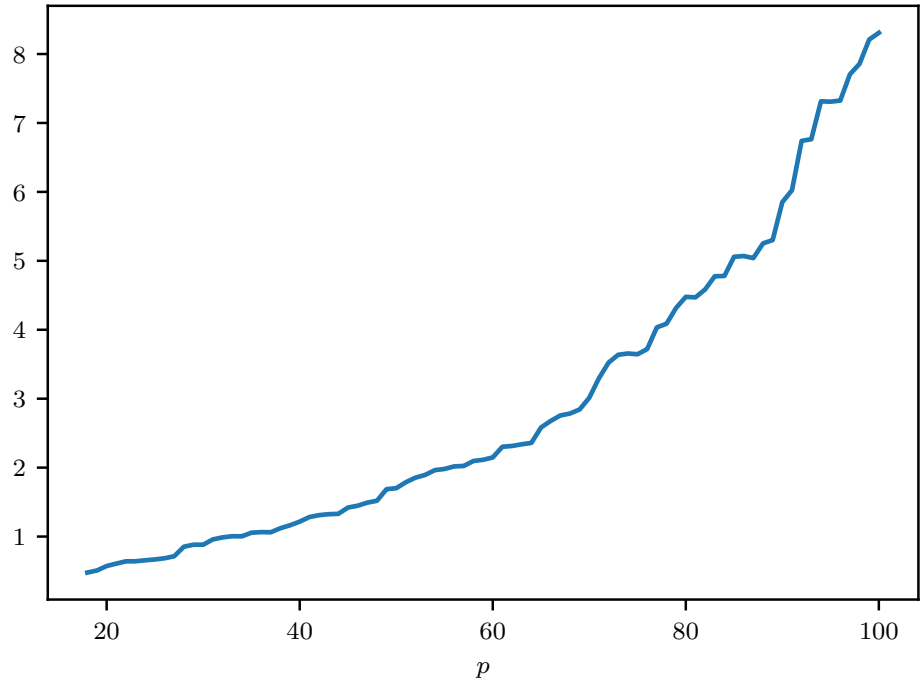

FIGURE 20 – Progression de $\|q^\star\|^2$

6 Conclusion

SVM multiclasse

Time series et learning

Références

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

[BEKL16] Gah-Yi Ban, Noureddine El Karoui, and Andrew EB Lim. Machine learning and portfolio optimization. *Management Science*, 2016.

[BPS13] Taras Bodnar, Nestor Parolya, and Wolfgang Schmid. On the equivalence of quadratic optimization problems commonly used in portfolio theory. *European Journal of Operational Research*, 229(3) :637–644, 2013.

[Cov91] Thomas M Cover. Universal portfolios. *Mathematical finance*, 1(1) :1–29, 1991.

[DB16] Steven Diamond and Stephen Boyd. CVXPY : A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83) :1–5, 2016.

[DCB13] A. Domahidi, E. Chu, and S. Boyd. ECOS : An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.

[FF93] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1) :3–56, 1993.

[Haz15] Elad Hazan. Introduction to online convex optimization. *Foundations and trends in optimization*, 2(3-4) :157–325, 2015.

[KW71] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1) :82–95, 1971.

[Mar52] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1) :77–91, 1952.

[Mar14] Harry Markowitz. Mean–variance approximations to expected utility. *European Journal of Operational Research*, 234(2) :346–355, 2014.

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

[Rém13] Bruno Rémillard. *Statistical Methods for Financial Engineering*. CRC Press, 2013.

[SDR09] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming : modeling and theory*. SIAM, 2009.