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

Thierry Bazier-Matte

10 février 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, on a d'autre part l'aspect d'aversion au risque qui est modélisé 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 (SAA, voir Shapiro[Citation needed]) :

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

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 $\lambda$ 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 [**?**].

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 $\sigma^2 \coloneqq \operatorname{Var} R$.

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 (**??**) devient aussitôt

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

Or, puisque $\mathrm{Var}\, R = \boldsymbol{E}R^2 - (\boldsymbol{E}R)^2$, on a que $\boldsymbol{E}R^2 = \sigma^2 + \mu^2$, ce qui entraîne donc que (12) 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 [**?**] et [**?**], tous deux publiés à l'occasion du soixantième anniversaire de [**?**].

## 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 [**?**] 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 [**?**, **?**].

## 2.3 Théorie de portefeuille régularisé

[**?**]

## 2.4 Fama and French et suivants ?

[**?**]

## 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

## 3.1 Introduction

Soit $\kappa : \boldsymbol{X} \times \boldsymbol{X} \to \mathscr{R}$ un noyau semi-défini positif, $\boldsymbol{H}$ l'espace de Hilbert à noyau reproduisant induit par $\kappa$ et $K \in \mathscr{R}^{n \times n}$ la matrice associée à l'ensemble d'échantillonage $S_n \sim M^n$. Le problème d'optimisation de portefeuille régularisé s'exprime alors par

$$\underset{q \in \boldsymbol{H}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - \lambda \|q\|_{\boldsymbol{H}}^2. \tag{15}$$

Tel que mentionné, la dimension de $\boldsymbol{H}$ 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.** *Toute solution $q^\star$ de* (15) *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 $\alpha^\star \in \mathscr{R}^n$ tel que,*

$$q^\star = \sum_{i=1}^{n} \alpha_i^\star \phi_i = (\alpha^\star)^T \phi. \tag{16}$$

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

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

$$q^\star = \sum_{i=1}^{n} \alpha_i \phi_i, \tag{17}$$

où $\alpha \in \mathscr{R}^n$ **[Todo:** Espace cotangant ? ? ?**]**, on peut donc restreindre le domaine d'optimisation à $\mathscr{R}^n$. L'objectif de (15) devient alors

$$n^{-1} \sum_{i=1}^{n} u(r_i \textstyle\sum_{j=1}^{n} \alpha_j \phi_j(x_i)) - \lambda \langle q, q \rangle_{\boldsymbol{H}}. \tag{18}$$

Le premier terme se réexprime comme

$$n^{-1} \sum_{i=1}^{n} u(r_i \, \alpha^T \phi(x_i)), \tag{19}$$

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

$$\langle q, q \rangle_{\boldsymbol{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \phi_i, \phi_j \rangle_{\boldsymbol{H}} \tag{20}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j) \tag{21}$$

$$= \alpha^T K \alpha. \tag{22}$$

De sorte que le problème général (15) peut se reformuler par

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

## 3.2 Dualité

Let us consider the following problem, optimized over $q \in \mathscr{R}^p$ :

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell(r_i\, q^T x_i) + n\lambda \|q\|^2, \tag{24}$$

where $\ell = -u$. Alternatively, this problem can be respecified using slack vector $\xi \in \mathscr{R}^n$ as

$$\begin{aligned} \text{minimiser} \quad & \sum_{i=1}^{n} \ell(\xi_i) + n\lambda \|q\|^2 \\ \text{tel que} \quad & \xi_i = r_i\, q^T x_i. \end{aligned} \tag{25}$$

Let $\alpha \in \mathscr{R}^n$. The Lagrangian of (25) can be written as

$$\mathscr{L}(q, \xi, \alpha) = \sum_{i=1}^{n} \ell(\xi_i) + n\lambda \|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i). \tag{26}$$

Because the objective (25) is convex and its constraints are affine in $q$ and $\xi$, Slater's theorem states that the duality gap of the problem is zero. In other words, solving (24) is equivalent to maximizing the Lagrange dual function $g$ over $\alpha$ :

$$\text{maximiser} \quad g(\alpha) = \inf_{q, \xi} \mathscr{L}(q, \xi, \alpha). \tag{27}$$

Now, note that

$$g(\alpha) = \inf_{q, \xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) + n\lambda \|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i) \right\} \tag{28}$$

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

$$= - \sup_{\xi} \left\{ a^T \xi - \sum_{i=1}^{n} \ell(\xi_i) \right\} + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\} \tag{30}$$

$$= -\sum_{i=1}^{n} \ell^*(\alpha_i) + \inf_q \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\}. \tag{31}$$

Where $\ell^*$ is the convex conjugate of the loss function and is defined by

$$\ell(\alpha_i) = \sup_{\xi_i} \{\alpha_i \xi_i - \ell(\xi_i)\}. \tag{32}$$

Note that the identity

$$f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \ell(\xi_i) \Longrightarrow f^*(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \ell^*(\xi_i) \tag{33}$$

was used. Consider now the second part of (31). Since the expression is differentiable, we can analytically solve for $q$ :

$$\nabla_q \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda \|q\|^2 \right\} = 0 \tag{34}$$

implies that

$$q = -\frac{1}{2n\lambda} \sum_{i=1}^{n} \alpha_i r_i x_i \tag{35}$$

at the infimum.

Using (35), we can eliminate $q$ from (31), so that

$$g(\alpha) = -\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{2n\lambda} \sum_{i,j=1}^{n} \alpha_i \alpha_j r_i r_j x_i^T x_j + \frac{1}{4n\lambda} \sum_{i,j=1}^{n} \alpha_i \alpha_j r_i r_j x_i^T x_j \tag{36}$$

$$= -\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r). \tag{37}$$

Therefore, in its dual form, the problem (24) is equivalent to solving

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell^*(\alpha_i) + \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r). \tag{38}$$

**Prescribed investment** In its original form, given a feature vector $\tilde{x}$, the algorithm (24) suggests an investment size of $p_0 = q^T \tilde{x}$, where $q$ is the trained value obtained by optimizing (24). In the dual formulation (38), with optimal value $\alpha$, we have from (35) :

$$p_0 = q^T x_0 \tag{39}$$

$$= -\frac{1}{2n\lambda} \sum_{i=1}^{n} \alpha_i r_i x_i^T x_0. \tag{40}$$

**[Todo:** Insert kernel formulation with vector $\phi$.**]**

### 3.3 Alternate problem

We now consider a new problem, slightly different from (24) where a regularization based on the sum of the square of the investment sizes $q^T x_i$ is applied :

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell(r_i\, q^T x_i) + \gamma \sum_{i=1}^{n} (q^T x_i)^2 + n\lambda \|q\|^2. \tag{41}$$

Again, this problem can be respecified using slack vector $\xi \in \mathscr{R}^n$ as

$$\begin{array}{rl} \text{minimiser} & \displaystyle\sum_{i=1}^{n} \ell(\xi_i) + \gamma \sum_{i=1}^{n} (\xi_i / r_i)^2 + n\lambda \|q\|^2 \\ \text{tel que} & \xi_i = r_i\, q^T x_i. \end{array} \tag{42}$$

The constraints in (42) are again affine, so that Slater's theorem apply.

The lagrangian of (42) is

$$\mathscr{L}(q, \xi, \alpha) = \sum_{i=1}^{n} \ell(\xi_i) + \gamma \sum_{i=1}^{n} (\xi_i / r_i)^2 + n\lambda \|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i), \tag{43}$$

and we seek its infimum over $(q, \xi)$.

$$\inf_{q,\xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) + \gamma \sum_{i=1}^{n} (\xi_i / r_i)^2 + n\lambda \|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i) \right\} \tag{44}$$

$$= \inf_{\xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) + \gamma \sum_{i=1}^{n} (\xi_i / r_i)^2 - \alpha^T \xi \right\} + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i - n\lambda \|q\|^2 \right\} \tag{45}$$

$$= -\sup_{\xi} \left\{ a^T \xi - \left( \sum_{i=1}^{n} \ell(\xi_i) + \gamma \sum_{i=1}^{n} (\xi_i / r_i)^2 \right) \right\} - \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r). \tag{46}$$

Let $f_i(\xi_i) := h_1(\xi_i) + h_2(\xi_i) = \ell(\xi_i) + \gamma(\xi_i / r_i)^2$. Then, using (33), the first expression of (46) can be restated as

$$-\sup_{\xi} \left\{ \alpha^T \xi - \sum_{i=1}^{n} f_i(\xi_i) \right\} = -\sum_{i=1}^{n} f_i^*(\alpha_i). \tag{47}$$

Let us introduce another identity :

$$(h_1 + h_2)^*(\alpha_i) = \inf_{\alpha_i' + \alpha_i'' = \alpha_i} \{h_1^*(\alpha_i') + h_2^*(\alpha_i'')\}. \tag{48}$$

Using (48), (47) can be written as

$$-\sum_{i=1}^{n} f_i^*(\alpha_i) = -\sum_{i=1}^{n} (h_1 + h_2)^*(\xi_i) \tag{49}$$

$$= -\sum_{i=1}^{n} \inf_{\alpha_i' + \alpha_i'' = \alpha_i} \{h_1^*(\alpha_i') + h_2^*(\alpha_i'')\}. \tag{50}$$

The first conjugate function $h_1^*$ is simply $\ell^*$. The second conjugate function can be derived analytically :

$$h_2^*(\alpha_i'') = \sup_{\xi_i} \{\alpha_i'' \xi_i - h_2(\xi_i)\} \tag{51}$$

$$= \sup_{\xi_i} \{\alpha_i'' \xi_i - \gamma(\xi_i / r_i)^2\}. \tag{52}$$

The supremum occurs when

$$\xi_i = \frac{r_i^2}{2\gamma} \alpha_i''. \tag{53}$$

Therefore, (52) simplifies to

$$h_2^*(\alpha_i'') = \frac{r_i^2}{4\gamma} (\alpha_i'')^2. \tag{54}$$

Putting it all back together, the dual of (41) is

$$-\sum_{i=1}^{n} \inf_{\alpha_i' + \alpha_i'' = \alpha_i} \left\{ \ell^*(\alpha_i') + \frac{r_i^2}{4\gamma} (\alpha_i'')^2 \right\} - \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r), \tag{55}$$

which is equivalent to

$$-\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{4\gamma} \sum_{i=1}^{n} (r_i \, \beta_i)^2 - \frac{1}{4n\lambda} (r \circ (\alpha + \beta))^T K (r \circ (\alpha + \beta)), \tag{56}$$

with new optimization variables $\alpha = \alpha', \beta = \alpha'' \in \mathscr{R}^n$. The dual optimization problem is therefore

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell^*(\alpha_i) + \frac{1}{4\gamma} \|r \circ \beta\|^2 + \frac{1}{4n\lambda} (r \circ (\alpha + \beta))^T K (r \circ (\alpha + \beta)). \tag{57}$$

**Prescribed investment [Todo: ]**

# 4 Garanties statistiques

## 4.1 Bornes de généralisation

**Exposition du problème** Soit $\boldsymbol{Q}$ un espace de Hilbert à noyau reproduisant induit par $\kappa$ et soit un ensemble d'entraînement $\mathcal{S}_n = \{(x_i, y_i)\}_{i=1}^n \sim M^n$ 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

$$\mathcal{Q}(\mathcal{S}_n) = \arg\max_{q \in \boldsymbol{Q}} \left\{ \widehat{\boldsymbol{EU}}(\mathcal{S}_n, q) - \lambda \|q\|^2 \right\}. \tag{58}$$

Comme on l'a vu, résoudre (58) est aussi équivalent à

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^n u(r_i\, \alpha^T \phi(x_i)) - \lambda \|\alpha\|_K^2, \tag{59}$$

où $\phi : \mathscr{R}^p \to \mathscr{R}^n$ le vecteur d'application induit par la matrice d'information $\Xi$. 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 $\Omega$ sur l'erreur de généralisation de $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ valide pour tout $\mathcal{S}_n \sim M^n$ :

$$\hat{\zeta}(\mathcal{S}_n) \leq \Omega(n, \ldots), \tag{60}$$

où

$$\hat{\zeta}(\mathcal{S}_n) = \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \tag{61}$$

représente l'erreur de généralisation. On peut en fait montrer que $\Omega \to 0$ à mesure que $n \to \infty$. En fait, on peut démontrer que $\Omega = O(n^{-1/2})$, ce qui permet de quantifier la "vitesse" à laquelle la convergence à lieu.

**Démonstration** 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 $(x, r) \sim M$,

$$|u(r\, \hat{q}(x)) - u(r\, \hat{q}'(x))| \leq \beta. \tag{62}$$

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 $\beta = O(n^{-1})$.

Doté de cette stabilité de $\mathcal{Q}$, on peut alors borner la différence dans l'erreur de généralisation de $\mathcal{S}_n$ et $\mathcal{S}'_n$ :

$$|\hat{\zeta}(\mathcal{S}_n) - \hat{\zeta}(\mathcal{S}'_n)| = |\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}') + \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \tag{63}$$

$$\leq |\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| + |\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')|. \tag{64}$$

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

$$|\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| = |\boldsymbol{E}(u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)))| \tag{65}$$

$$\leq \boldsymbol{E}(|u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)|) \tag{66}$$

$$\leq \beta, \tag{67}$$

par définiton de la stabilité. Quant au deuxième terme de (64) on peut le borner de la même façon :

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \tag{68}$$

$$= n^{-1} \left| \sum_{i=1}^{n} \mathbb{I}_{i \neq j} u(r_i\, \hat{q}(x_i)) + u(r_j\, \hat{q}(x_j)) - \sum_{i=1}^{n} \mathbb{I}_{i \neq j} u(r_i\, \hat{q}'(x_i)) - u(r'_j\, \hat{q}'(x'_j)) \right| \tag{69}$$

$$\leq n^{-1} \left( |u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j))| + \sum_{i=1}^{n} \mathbb{I}_{i \neq j} \big| u(r_i\, \hat{q}(x_i)) - u(r_i\, \hat{q}'(x_i)) \big| \right) \tag{70}$$

$$\leq n^{-1} \left( |u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j))| + (n-1)\beta \right). \tag{71}$$

Considérons le premier terme. Par le lemme **[Todo: ]**, 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 sous-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

$$|u(r_j\, \hat{q}(x_j) - u(r'_j\, \hat{q}'(x'_j)| \leq u((2\lambda)^{-1}\bar{r}^2\xi^2) - u(-(2\lambda)^{-1}\bar{r}^2\xi^2) \tag{72}$$

$$\leq (2\lambda)^{-1}(\gamma + 1)\bar{r}^2\xi^2. \tag{73}$$

Ce qui entraîne donc que

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| \leq \frac{\gamma + 1}{2\lambda n}\bar{r}^2\xi^2 + \frac{n-1}{n}\beta \tag{74}$$

$$\leq \beta + \frac{\gamma + 1}{2\lambda n}\bar{r}^2\xi^2. \tag{75}$$

Next McDiarmid et autre. Easy stuff.

**Équivalent certain** Puis inverser pour obtenir l'équivalent certain.

**Note bibliographique** La théorie de la stabilité algorithmique remonte en fait aux années 70 avec les travaux de Luc Devroye appliqués à l'algorithme des $k$ plus proches voisins[Citation needed]. Jusqu'alors, les bornes de généralisation étaient présentées pour toute décision $q \in \boldsymbol{Q}$ (ie Vapnik). Bousquet[Citation needed]a été le premier a présenter des résultats dans des espaces de Hilbert à noyau reproduisant. La démonstration est fortement inspirée de l'excellente référence Mohri[Citation needed]. La démonstration de la borne de la décision bornée est un résultat inédit, dûe à Delage dans le cas linéaire.

## 4.2 Bornes de sous optimalité

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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{EU}(R \cdot q(X)). \tag{76}$$

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

$$q = \alpha \boldsymbol{E}(R\,\kappa(X, \cdot)) \tag{77}$$

On aurait alors

$$\boldsymbol{EU}(q) = \boldsymbol{E}(Rq(X)) = \boldsymbol{E}(R\boldsymbol{E}(R\,\kappa(X, X))) \tag{78}$$

$$= \boldsymbol{E}(R^2\,\kappa(X, X)) \geq 0, \tag{79}$$

On peut alors obtenir une utilité espérée non bornée à mesure que $\alpha \to \infty$. Par ailleurs, ainsi défini, $q$ représente effectivement la covariance entre $R$ et la projection de $X$ dans l'espace dual de $\boldsymbol{Q}$. Puisque l'utilité est neutre, on sait qu'en espérance l'application de $q$ à $X$ variera de la même façon que celle de $R$ et donc qu'on aura une utilité infinie. On verra plus loin au cours d'une démonstration la motivation derrière cette hypothèse supplémentaire :

**Hypothèse 1.** *L'utilité croît sous-linéairement, ie.* $u(r) = o(r)$.

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

$$\boldsymbol{P}\{R \cdot q(X) > 0\} = 1. \tag{80}$$

D'un point de vue strictement financier, cela fait certainement du sens en vertu de l'efficience des marchés, version semi-forte[Citation needed]. D'un point de vue théorique, ceci exige en fait qu'il n'y ait pas de région dans $\boldsymbol{X}$ telle que tous les rendements s'y produisant soient nécessairement positifs ou négatifs.**[Todo:** Inserer image**]**.

**Hypothèse 2.** *Pour toute région $\mathcal{R} \subseteq \boldsymbol{X}$,*

$$\boldsymbol{P}\{R \geq 0 \mid X \in \mathcal{R}\} < 1, \tag{81}$$

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

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

## 4.3 Lemmes

**Stabilité** On montre ici que

$$\beta \leq \frac{(\gamma \bar{r} \xi)^2}{2\lambda n}. \tag{82}$$

**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

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - \lambda \|q\|^2. \tag{83}$$

Pour ce faire, on va mettre à profit la propriété reproductive de $\boldsymbol{Q}$ induite par $\kappa$. En effet, celle-ci stipule que

$$q(x) = \langle q, \kappa(x, \cdot) \rangle_{\boldsymbol{Q}} \leq \|q\| \sqrt{\kappa(x, x)}, \tag{84}$$

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 $\|q\|_{\boldsymbol{Q}}$. Or, puisque $u(r) \leq r$, on remarque que

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) &\leq n^{-1} \sum_{i=1}^{n} r_i \, q(x_i) && (85) \\ &\leq n^{-1} \sum_{i=1}^{n} r_i \sqrt{\kappa(x_i, x_i)} \|q\| && (86) \\ &\leq \bar{r} \xi \|q\|. && (87) \end{aligned}$$

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

$$\|q\| = \frac{\bar{r}\xi}{2\lambda}, \tag{88}$$

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

$$\hat{q}(x) \leq \frac{\bar{r}\xi^2}{2\lambda}. \tag{89}$$

**[Todo:** Montrer qu'on peut effectivement montrer que la borne de l'expression est plus grande que celle du problème initial... Pour ce faire, utiliser a) le sous gradient de $u$ et b) la dominance de $q'$ sur $q$.**]**

**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

$$\boldsymbol{EU}_\lambda(tq_1+(1-\alpha)q_2) \geq \alpha\boldsymbol{EU}_\lambda(q_1)+(1-\alpha)\boldsymbol{EU}_\lambda(q_2)+\lambda\alpha(1-\alpha)\|q_1-q_2\|^2, \tag{90}$$

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

$$\begin{aligned}
&\boldsymbol{EU}_\lambda(\alpha q_1 + (1-\alpha)q_2) && (91)\\
&\quad = \boldsymbol{E}u(R\cdot(\alpha q_1 + (1-\alpha)q_2)(X)) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (92)\\
&\quad = \boldsymbol{E}u(\alpha(R\cdot q_1(X)) + (1-\alpha)(R\cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (93)\\
&\quad \geq \boldsymbol{E}(\alpha u(R\cdot q_1(X)) + (1-\alpha)u(R\cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (94)\\
&\quad = \alpha\boldsymbol{EU}(q_1) + (1-\alpha)\boldsymbol{EU}(q_2) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 && (95)\\
&\quad = \alpha\boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) - \lambda(\|\alpha q_1 + (1-\alpha)q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2) && (96)\\
&\quad \geq \alpha\boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) - \lambda(\alpha\|q_1\|^2 + (1-\alpha)\|q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2) && (97)\\
&\quad = \alpha\boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2). && (98)
\end{aligned}$$

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

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

Notons d'abord que l'hypothèse 2 entraîne en particulier qu'il existe $\delta > 0$ tel que

$$\boldsymbol{P}\{R\cdot\theta(X) \leq -\delta\} > \varrho \geq 0 \tag{99}$$

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

$$\boldsymbol{P}\{B \geq r\} \geq \boldsymbol{P}\{R\cdot\theta(X) \geq r\} \tag{100}$$

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

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

$$\lim_{s\to\infty} \boldsymbol{E}u(R \cdot s\theta(X)) \leq \lim_{s\to\infty} u(sB) \tag{101}$$

$$= \lim_{s\to\infty} (\varrho u(-s\delta) + (1-\varrho)u(s\bar{r}\xi)) \tag{102}$$

$$\leq \lim_{s\to\infty} -\varrho s\delta + (1-\varrho)o(s) = -\infty, \tag{103}$$

ce qui démontre bien que $s$ est borné.

# 5 Distributions synthétiques

# 6 Conclusion

SVM multiclasse

Time series et learning