

Introduction à la quantification optimale finidimensionnelle

Afef SELLAMI *

Février 2006

La quantification est une méthode issue du traitement du signal et de l'information. Elle consiste à approcher un signal à valeurs dans un espace continu par un signal à valeur dans un espace discret. A l'origine, cette technique a été motivée par des raisons pratiques d'économie de transmission [6], et a vu ensuite ses applications s'élargir à différents domaines, notamment, depuis quelques années, les probabilités numériques. Cette introduction a été adaptée de l'introduction de ma thèse [18] dédiée à la résolution de problèmes de filtrage non linéaire par quantification et à son application en finance. Nous allons y adopter une présentation résolument probabiliste de la quantification en rappelant quelques résultats théoriques utiles et en mettant en évidence ses applications les plus récentes.

1 Définitions et résultats préliminaires

1.0.1 Quantification de variables aléatoires

On se place dans un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, et on se donne une variable aléatoire à valeurs dans \mathbb{R}^d , X , de loi \mathbb{P}_X supposée simulable. Un entier $N \in \mathbb{N}^*$ étant fixé, un N -quantifieur est une application borélienne h_N appliquant \mathbb{R}^d dans un ensemble fini $\Gamma = \{x^1, \dots, x^N\} \subset \mathbb{R}^d$.

Pour définir de manière unique l'application h_N , on a besoin de spécifier en plus une partition borélienne $(A^i)_{1 \leq i \leq N}$ de l'espace \mathbb{R}^d pour avoir:

$$h_N(X) = \sum_{i=1}^N x^i \mathbf{1}_{A^i}(X).$$

Le N -quantifieur est donc spécifié par la donnée de:

- $\Gamma = \{x^1, \dots, x^N\}$ la grille de quantification de taille N appelée aussi ensemble des centres, des points de quantification ou des centroïdes, ou encore N -quantifieur associé à h_N .
- Une partition borélienne $(A^i)_{1 \leq i \leq N}$ de l'espace \mathbb{R}^d . A chaque ensemble A^i sera associé un centre x^i que l'on supposera toujours appartenir à A^i .

*Laboratoire de Probabilités et Modèles Aléatoires. Université Paris VI. email: sellami@math.jussieu.fr

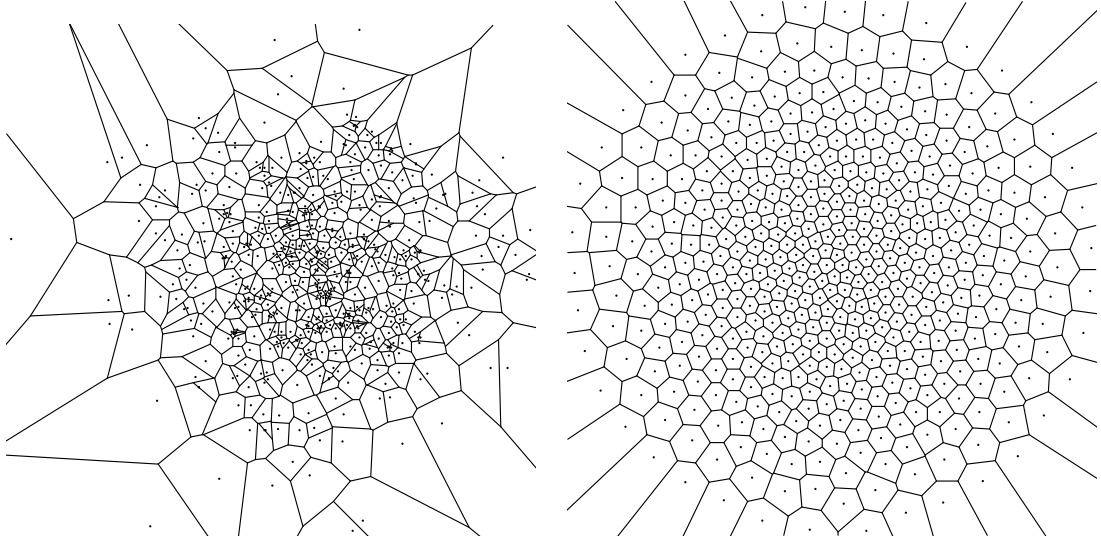


Figure 1: Partitions de l'espace associées à une grille de quantification en deux dimensions

Quand $X \in L^p$, on définit un N -quantifieur L^p -optimal de X par l'application h_N^* solution du problème d'optimisation paramétré par la taille de la grille de quantification N :

$$\inf\{\|X - h(X)\|_p^p, h : \mathbb{R}^d \rightarrow \mathbb{R}^d, \text{ application borelienne t.q. } |h(\mathbb{R}^d)| \leq N\}. \quad (1.1)$$

D'après les résultats établis par exemple dans [7], ce problème admet une solution h_N^* définie par un N -quantifieur $\Gamma_N^* = \{x^1, \dots, x^N\}$, vérifiant:

$$\mathbb{E}|X - h_N^*(X)|^p = \mathbb{E} \min_{x \in \Gamma_N^*(\mathbb{R}^d)} |X - x|^p,$$

et par la partition $(\mathbf{C}_i(\Gamma^*))_{1 \leq i \leq N}$, dite de Voronoï, associée à cet ensemble, définissant h_N^* comme une projection au plus proche voisin sur l'ensemble des centres $(x^i)_{1 \leq i \leq N}$ (voir Figure 1). Soit:

$$\mathbf{C}_i(\Gamma^*) \subset \{\xi \in \mathbb{R}^d \text{ t.q. } |\xi - x^i| = \min_{1 \leq k \leq N} |\xi - x^k|\}.$$

La distorsion¹ s'écrit alors:

$$\underline{\mathcal{D}}_N^{X,p} := \|X - h^*(X)\|_p^p = \left\| \min_{1 \leq i \leq N} |X - x^i| \right\|_p^p.$$

Elle converge vers zero quand la taille du quantifieur N tend vers $+\infty$, et sa vitesse de convergence est régie par le théorème de Zador énoncé comme suit:

Théorème 1.1 (Cf. [7, 2]) *On suppose que $\int_{\mathbb{R}^d} |\xi|^{p+r} \mathbb{P}_X(d\xi) < +\infty$ pour $r > 0$. Alors,*

$$\lim_N (N^{\frac{p}{d}} \underline{\mathcal{D}}_N^{X,p}) = J_{p,d} \|\varphi\|_{\frac{d}{d+p}},$$

où $\mathbb{P}_X(d\xi) = \varphi(\xi) \lambda_d(d\xi) + \bar{\mu}(d\xi)$, $\bar{\mu} \perp \lambda_d$ (λ_d mesure de Lebesgue sur \mathbb{R}^d) et pour tout $q \in \mathbb{R}_+^*$, $\|g\|_q := (\int |g|^q(u) du)^{\frac{1}{q}}$.

¹On notera par ailleurs l'erreur de quantification $\|X - h_N^*(X)\|_p$.

Remarque 1.1 $J_{p,d}$ correspond à la limite pour la loi uniforme sur $[0, 1]$. On sait que $J_{p,1} = \frac{1}{2^p(p+1)}$ et que $J_{2,2} = \frac{5}{18\sqrt{3}}$. D'une manière générale, on ne connaît pas la valeur de cette constante pour $d > 2$, mais on a $J_{2,d} \sim \frac{d}{2\pi e}$. (Cf. [7]).

Nous pouvons ainsi écrire que $\|X - h_N^*(X)\|_p = O(N^{-\frac{1}{d}})$ au voisinage de $N \rightarrow +\infty$. Par ailleurs, il sera essentiel de noter que les quantifieurs L^2 -optimaux vérifient une propriété dite de stationnarité, à savoir que:

$$\mathbb{E}[X|h_N^*(X)] = h_N^*(X). \quad (1.2)$$

Cette propriété permet d'utiliser des termes correcteurs de premier ordre dans les différentes applications de la quantification, comme nous allons le voir dans le paragraphe suivant pour l'intégration numérique puis, plus loin, dans les exemples d'évaluation d'options américaines [3] ou de filtrage [18]).

D'un point de vue pratique, définir la fonction h_N^* pour une taille de quantifieur fixée N s'avère être un problème d'optimisation assez délicat à résoudre notamment dans le cas multidimensionnel. En effet, pour plusieurs lois en dimensions 1, il existe des solutions analytiques fermées ou semi-fermées qui peuvent aisément être calculées. On citera par exemple la quantification de la loi normale par la méthode de Newton [14], ou de la loi exponentielle, entre autres, dans [5]. Par ailleurs, dans le cas particulier d'une loi à densité log-concave, la solution au problème (1.1) est unique [9], ce qui permet l'utilisation efficace d'algorithmes numériques d'optimisation tels que l'algorithme du gradient ou celui du point fixe appelé aussi méthode de Lloyd I [14, 1].

Ces méthodes déterministes deviennent rapidement difficiles à mettre en œuvre en dimensions supérieures à 1. D'une part parce qu'elles impliquent des calculs d'intégrales trop complexes au delà de la dimension 1, d'autres part, parce que la solution de (1.1) n'est pas unique ce qui diminue encore l'efficacité des résolutions numériques. Dans ce cas alors, on préfère utiliser des algorithmes stochastiques d'apprentissage fondés sur la simulation. On citera dans ce contexte l'algorithme *Competitive Learning Vector Quantization* (CLVQ algorithm) aussi appelé algorithme de Kohonen à zéro voisins, les algorithmes génétiques [8] ou l'algorithme de Lloyd I multidimensionnel [1]. Une étude détaillée des méthodes de quantification quadratique optimale est proposée dans [14]. Il est à noter que c'est le cas quadratique ($p = 2$) qui est le plus souvent étudié sur le plan numérique, même si les résultats théoriques sont généralement énoncés pour p quelconque.

Nous ne détaillerons pas plus avant cet aspect de la quantification car il n'est pas au centre de notre travail. Nous nous intéresserons plutôt à l'exploitation de ces quantifieurs optimaux pour différentes applications pratiques. A cet effet, il est important d'introduire tout d'abord une définition de processus quantifié (ou quantification de processus différente de la notion de quantification fonctionnelle de processus. Cf. [10, 15]). Ce sera l'objet du prochain paragraphe.

1.0.2 Quantification de processus

Dans les différentes applications de la quantification, il est souvent requis de considérer la quantification d'un processus markovien à temps discret $(X_k)_{k \geq 0}$ dont on connaît la dynamique d'évolution de manière à pouvoir en simuler la trajectoire. Une approche possible dans ce cas est de quantifier chaque variable X_k en tenant compte de sa loi marginale, on parle donc de *quantification marginale*. Pour cela, on doit se fixer une taille de grille N_k à chaque pas de temps et un N_k -quantifieur L^p -optimal de $X_k \in L^p$ que l'on notera $\Gamma_k = \{x_k^1, \dots, x_k^{N_k}\}$.

Par conséquent, on définit la quantification (Voronoi) de X_k par:

$$\hat{X}_k = h_{N_k}^*(X_k) = \sum_{i=1}^{N_k} x_k^i \mathbf{1}_{\mathbf{C}_i(\Gamma_k)}(X_k). \quad (1.3)$$

Le processus ainsi quantifié $(\hat{X}_k)_{k \geq 0}$ ne vérifie plus la propriété de Markov. Cependant, une approximation de la probabilité de transition entre différents états à deux dates successives reste possible à travers les paramètres compagnons p_k^{ij} , pour $i \in \{1, \dots, N_k\}$ et $j \in \{1, \dots, N_{k+1}\}$ définis par:

$$\begin{aligned} p_k^{ij} &= \mathbb{P}[X_{k+1} \in \mathbf{C}_j(\Gamma_{k+1}) | X_k \in \mathbf{C}_i(\Gamma_k)] \\ &= \mathbb{P}[\hat{X}_{k+1} = x_{k+1}^j | \hat{X}_k = x_k^i]. \end{aligned}$$

D'une manière générale, pour $0 \leq k < n$, $i \in \{1, \dots, N_k\}$ et f borélienne dans \mathbb{R}^d , on notera:

$$\hat{\mathbf{P}}_k f(x_k^i) = \mathbb{E}[f(\hat{X}_{k+1}) | \hat{X}_k = x_k^i] = \sum_{j=1}^{N_{k+1}} f(x_{k+1}^j) p_k^{ij}.$$

Pour des horizons n raisonnables (≤ 100), il est donc possible de calculer et de stocker dans des tables facilement accessibles les grilles de quantifications et les paramètres compagnons i.e. les (x_k^i) et les (p_k^{ij}) pour $0 \leq k \leq n$, $1 \leq i \leq N_k$ et $1 \leq j \leq N_{k+1}$. Ce pré-traitement des données, dit *off-line*, permet de minimiser les calculs d'éventuels estimateurs utilisant la quantification. Un exemple de ces grilles pour la loi normale centrée réduite est disponible et téléchargeable à partir de <http://www.proba.jussieu.fr/pageperso/pages/quant3.html>.

2 Application à l'intégration numérique

Une application immédiate de la quantification est le calcul d'approximations numériques d'intégrales par rapport à une mesure donnée. On se pose le problème de l'évaluation de l'intégrale $\mathbb{E}[f(X)]$, pour X de loi \mathbb{P}_X sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Si \hat{X} désigne une N -quantification L^2 -optimale de X nous pouvons nous donner comme estimateur : $\mathbb{E}[f(\hat{X})]$. Comme \hat{X} est une variable aléatoire discrète, le calcul de cet estimateur se résumera à une somme

pondérée finie, dont les termes sont lus à partir de tables pré-calculées. En reprenant les notations du paragraphe précédent, on pose:

$$\mathbb{E}[f(X)] \approx \mathbb{E}[f(\hat{X})] = \sum_{i=1}^N x^i \int \mathbf{1}_{\mathbf{C}_i(\Gamma)}(x) \mathbb{P}_X(dx) = \sum_{i=1}^N x^i \hat{\mathbf{p}}^i,$$

où $\hat{\mathbf{p}}^i = \mathbb{P}_X(\mathbf{C}_i(\Gamma)) = \mathbb{P}(X \in \mathbf{C}_i(\Gamma))$.

Les pondérations $\hat{\mathbf{p}}^i$ sont aussi des paramètres compagnons qui peuvent être calculés en même temps que la grille de quantification Γ et stockés dans des tables accessibles pendant l'estimation. L'erreur d'estimation est contrôlée par l'erreur de quantification $\Delta = X - \hat{X}$. En effet, quand f est continue, dérivable à dérivée bornée, il existe $\xi \in (X, \hat{X})$ tel que:

$$f(X) - f(\hat{X}) = \langle Df(\xi), \Delta \rangle,$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire canonique sur \mathbb{R}^d .

Ceci donne la majoration d'erreur qu'on appellera *d'ordre zero*:

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(\hat{X})]| \leq C \|\Delta\|_1 \leq C \|\Delta\|_2. \quad (2.4)$$

Quand f est continue, 2 fois dérivable, à dérivée seconde bornée, on peut développer f à un ordre supérieur afin d'établir une majoration d'erreur *d'ordre 1*. En effet, il existe $\xi \in (X, \hat{X})$ tel que:

$$f(X) - f(\hat{X}) = \langle Df(\hat{X}), \Delta \rangle + \frac{1}{2} \Delta' D^2 f(\xi) \Delta.$$

Ainsi, comme \hat{X} vérifie la propriété de stationnarité (1.2), il est possible d'établir:

$$\begin{aligned} |\mathbb{E}[f(X)] - \mathbb{E}[f(\hat{X})]| &\leq \mathbb{E}|\mathbb{E}[f(X) - f(\hat{X})|\hat{X}]| \\ &\leq C \mathbb{E}|\langle \Delta, \Delta \rangle| \leq C \|\Delta\|_2^2. \end{aligned} \quad (2.5)$$

En utilisant le théorème de Zador 1.1, on obtient un taux de convergence en $O(N^{-\frac{1}{d}})$ dans le cas de l'inégalité (2.4), et moyennant l'hypothèse plus restrictive sur la fonction f , on a un taux de convergence deux fois plus rapide $O(N^{-\frac{2}{d}})$ à partir de (2.5).

L'intégration numérique par quantification est une méthode qui s'approche dans son principe des méthodes de Monte Carlo [17]: elle s'appuie sur une représentation de la loi de X par un ensemble discret fini pondéré. L'estimateur de Monte Carlo s'écrit en effet comme la somme équipondérée sur un échantillon iid de taille M :

$$\mathbb{E}[f(X)] \approx \frac{1}{M} \sum_{i=1}^M f(X^i) \quad \text{où} \quad (X^1, \dots, X^M) \quad iid \quad X^1 \sim \mathbb{P}_X.$$

Mais si le principe reste le même, de grandes différences séparent les deux méthodes:

- Les grilles de quantification ainsi que les pondérations peuvent être calculées offline, et stockées dans des tables accessibles par plusieurs applications à la fois. La complexité du calcul exclut donc la procédure d'optimisation des quantifieurs et compte

seulement les opérations élémentaires de somme et de pondération. Au contraire, les méthodes de Monte Carlo utilisent une partie de la capacité de calcul dans la simulation online des échantillons X^i .

- L'estimateur Monte Carlo est un estimateur aléatoire, dont il faudra gérer la variance lors des applications par des procédures de contrôle et de minimisation de variance. A son opposé, l'estimateur par quantification est un estimateur déterministe.
- La vitesse de convergence (en loi) des estimateurs de Monte Carlo est $O(N^{-\frac{1}{2}})$ (TCL); il est indépendant de la dimension. La loi du logarithme itéré règle la vitesse de convergence p.s. en $\sqrt{\frac{\log \log N}{N}}$. La convergence des estimateurs par quantification, bien que dépendant de la dimension, est asymptotiquement plus rapide jusqu'à la dimension 2 pour celles du type *ordre 0* et jusqu'à la dimension 4 pour celles du type *ordre 1*. Au delà, la méthode de quantification ne reste pas compétitive lorsque $N \rightarrow +\infty$; cependant elle se révèle encore très efficace en pratique notamment en dimension moyenne ($d \leq 10$) lorsque N n'est pas très grand.

Il est délicat de comparer une méthode déterministe comme la quantification et une méthode fournissant un résultat aléatoire comme la méthode de Monte Carlo. Néanmoins, si l'on se tient à la pratique des utilisations, il paraît naturel de comparer l'erreur de quantification avec la taille d'un intervalle de confiance. Typiquement, avec un choix adéquat d'intervalle de confiance, ceci revient à comparer $\|f(X) - f(\hat{X})\|_2$ et $\frac{2\sigma_f(X)}{\sqrt{N}}$ pour \hat{X} une N -quantification L^2 -optimale et f Lipschitzienne; ou bien, plus universellement à comparer $\|X - \hat{X}\|_2$ et $\frac{2\sigma_X}{\sqrt{N}}$ pour les schémas de quantification d'ordre 0 et $\|X - \hat{X}\|_2^2$ et $\frac{2\sigma_X}{\sqrt{N}}$ pour les schémas d'ordre 1. Il est intéressant de voir alors qu'il existe des seuils critiques N_c^0 pour les schémas d'ordre 0 et N_c^1 pour les schémas d'ordre 1, pour lesquels $\forall N \leq N_c^0 \leq N_c^1$ l'erreur par quantification est inférieure à la longueur de l'intervalle de confiance donné par la méthode de Monte Carlo. Ce résultat est détaillé dans [13].

Il est par ailleurs possible d'appliquer un raisonnement d'échantillonnage préférentiel [17] dans l'intégration numérique par quantification.

Définition 2.1 *Echantillonnage Préférentiel:* *C'est la procédure par laquelle on approche de manière empirique une mesure de probabilité difficile à simuler ν en utilisant un échantillon iid (ξ_1, \dots, ξ_N) d'une autre mesure de probabilité μ dite loi d'importance, plus facile à simuler. S'il existe une fonction m et une constante \bar{m} vérifiant:*

$$\nu(dx) = m(x)\mu(dx) \quad \text{et} \quad m(x) \leq \bar{m},$$

l'approximation est alors donnée par $\nu \approx \frac{1}{N} \sum_{i=1}^N m(\xi_i)\delta_{\xi_i}$.

Si on désigne par \mathbf{p} la densité de X et par \mathbf{q} une densité d'importance qui vérifiera $f \frac{\mathbf{p}}{\mathbf{q}} \in \mathcal{C}_b$, on pourra utiliser la quantification d'une variable Y de densité q pour estimer

$\mathbb{E}[f(X)]$.

$$\begin{aligned}\mathbb{E}[f(X)] &= \mathbb{E}\left[f(Y) \frac{\mathbf{p}(Y)}{\mathbf{q}(Y)}\right] \approx \mathbb{E}\left[f(\hat{Y}) \frac{\mathbf{p}(\hat{Y})}{\mathbf{q}(\hat{Y})}\right], \\ |\mathbb{E}[f(X)] - \mathbb{E}\left[f(\hat{Y}) \frac{\mathbf{p}(\hat{Y})}{\mathbf{q}(\hat{Y})}\right]| &\leq C\|Y - \hat{Y}\|_2.\end{aligned}\tag{2.6}$$

3 Autres applications

La quantification a historiquement connu plusieurs applications dans le domaine de la théorie de l'information, du traitement du signal et de sa compression. Cependant, la solution qu'elle offre au problème d'intégration numérique permet de l'appliquer dans de nouveaux domaines pour résoudre des problèmes impliquant un calcul numérique d'intégrales, d'espérances ou d'espérances conditionnelles. De tels problèmes se posent en finance dans le cadre de modèles d'évaluation de produits dérivés, où un calcul d'espérance conditionnelle est requis et souvent, ce type de problèmes est reformulé de manière rétrograde en utilisant un principe de programmation dynamique. La quantification, de par son principe de pré-traitement et de calcul off-line des grilles s'adapte bien à ce type d'approche. Parmi les applications qui s'inscrivent dans ce cadre, on cite [4], où est proposé un algorithme d'évaluation d'options américaines et d'estimation de temps d'arrêt optimal, la date d'exercice de l'option, lorsque l'actif sous jacent suit une diffusion brownienne. Comme pour le problème d'intégration numérique, l'utilisation de quantifieurs stationnaires permet dans cette application d'améliorer l'estimation numérique par le passage à un ordre supérieur de convergence (Cf. [3]). Par ailleurs, Pagès et Pham [12] définissent une quantification *markovienne* de processus qui préserve la propriété de Markov du processus original et permet de proposer une solution numérique à un problème de contrôle stochastique apparaissant dans des problématiques financières de gestion de portefeuille. Dans ces applications, la résolution numérique rétrograde est rendue possible grâce aux grilles de quantification précalculées.

Une autre application issue du domaine du traitement du signal est celle de la résolution de problèmes de filtrage non linéaire. Elle a été introduite dans [11], où la construction d'une solution séquentielle rétrograde a permis de voir l'efficacité théorique de la quantification, tout en laissant possible une implémentation numérique classique dite forward. Le passage aux schémas d'ordre un a été suggéré dans [13] et étudié en détail dans [18], où une comparaison avec les méthodes Monte Carlo (filtres à particules) est effectuée sur les plans théoriques et numériques. En finance, la problématique du filtrage est omniprésente dans les modèles d'actifs à volatilité stochastique : la volatilité est vue comme un signal dont le cours de l'actif est l'observation bruitée. L'introduction d'approximations numériques du filtre permet alors de résoudre des problèmes en observation partielle tels que l'évaluation d'options américaines [16] et de gestion de portefeuille.

References

- [1] J. G. Attali. *1. Méthodes de stabilité pour des chaînes de Markov non féllériennes 2. Sur quelques autres problèmes issus des réseaux de neurones*. PhD thesis, Université de Paris I Panthéon Sorbonne, 1999.
- [2] V. Bally and G. Pagès. A quantization algorithm for solving discrete time multi-dimensional optimal stopping problems. *Bernoulli*, 9:1003–1049, 2003.
- [3] V. Bally, G. Pagès, and J. Printems. First order schemes in the numerical quantization method. *Mathematical finance*, 13(1):1–16, 2003.
- [4] V. Bally, G. Pagès, and J. Printems. A quantization tree method for pricing and hedging multidimensional american options. *Mathematical Finance*, 15(1):119–168, 2005.
- [5] J. C. Fort and G. Pagès. Asymptotics of optimal quantizers for some scalar distributions. *Journal of Computational and Applied Mathematics*, 146:253–275, 2002.
- [6] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. The Kluwer International Series in Engineering and Computer Science. Springer, 2000.
- [7] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Lecture Notes in Mathematics. Springer, 2000.
- [8] S. Ben Hamida and M. Mrad. Optimal Quantization using Evolutionary Algorithms. CMAP Preprint 564, Ecole Polytechnique-France, 2005.
- [9] J. Kieffer. Exponential rate of convergence for the Lloyd’s method I. *IEEE on Information Theory, Special issue on Quantization*, 28(2):205–210, 1982.
- [10] H. Luschgy and G. Pagès. Functional quantization of Gaussian processes. *Journal of Functional Analysis*, 196(2):486–531, 2002.
- [11] G. Pagès and H. Pham. Optimal quantization methods for nonlinear filtering with discrete-time observations. *Bernoulli*, 11(5):893–932, 2005.
- [12] G. Pagès, H. Pham, and J. Printems. An optimal Markovian quantization algorithm for multidimensional stochastic problems. *Stochastics and Dynamics*, 4:501–545, 2004.
- [13] G. Pagès, H. Pham, and J. Printems. Optimal quantization methods and applications to numerical problems in finance. In *Handbook of Computational and Numerical Methods in Finance*. S.T. Rachev, Birkhauser, Boston, 2004.
- [14] G. Pagès and J. Printems. Optimal quadratic quantization for numerics : the Gaussian case. *Monte Carlo methods and applications*, 9(2):135–168, 2003.
- [15] G. Pagès and J. Printems. Functional quantization for pricing derivatives. Preprint of Laboratoire de Probabilités et Modèles Aléatoires PMA-930, 2004.
- [16] H. Pham, W. Runggaldier, and A. Sellami. Approximation by quantization of the filter process and applications to optimal stopping problems under partial observation. *Monte Carlo methods and applications*, 11:57–81, 2005.
- [17] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer texts in statistics. Springer, 2nd edition, 1999.
- [18] A. Sellami. Optimal quantization methods for filtering and applications to finance. PhD thesis, Université Paris Dauphine. 2005.