Université de Montréal

L'utilisation de règles de réseau en simulation comme technique de réduction de la variance.

par

Christiane Lemieux

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de

Philosophiæ Doctor (Ph.D.)

en informatique

mai 2000

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée:

L'utilisation de règles de réseau en simulation comme technique de réduction de la variance.

présentée par:

Christiane Lemieux

a été évaluée par un jury composé des personnes suivantes:

Felisa J. Vázquez-Abad

(président-rapporteur)

Pierre L'Ecuyer

(directeur de recherche)

Yoshua Bengio

(membre du jury)

Luc Devroye

(examinateur externe)

Christian Léger

(représentant du doyen)

Thèse acceptée le:

Sommaire

Dans cette thèse, nous étudions comment les règles de réseau peuvent être utilisées en simulation comme méthode de réduction de la variance.

Nous nous intéressons d'abord aux règles de réseau "standard" et donnons des expressions et des bornes pour la variance des estimateurs obtenus en utilisant différentes randomisations. Nous présentons ensuite un nouveau critère de sélection basé sur les projections de la règle de réseau sur les sous-espaces de l'hypercube unitaire en s dimensions sur lequel la règle est définie. Des résultats numériques obtenus sur divers problèmes pour lesquels la simulation est typiquement utilisée sont donnés, afin de comparer les règles de réseau avec la méthode Monte Carlo. Ces résultats illustrent également l'utilité du nouveau critère.

Puis, nous passons aux règles de réseau polynômiales et à leur lien avec les (t, m, s)-réseaux. Des résultats théoriques sur la variance des estimateurs associés à ces règles et un nouveau critère de sélection sont dérivés de façon analogue au cas standard. Nous donnons également des résultats numériques permettant de voir comment ces règles se comparent en pratique avec la méthode Monte Carlo et les règles standard.

Finalement, nous présentons des résultats illustrant la différence entre les règles de rang 1 et celles de type ν^r-copie. Les résultats théoriques sont appuyés par des exemples numériques comparant ces deux types de règles sur différents problèmes et par rapport à certains critères de sélection.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

$\mathbb{F}_2$ = Corps de Galois contenant deux éléments,

$\mathcal{L}^1$ = $\{f : \int_{[0,1)^s} |f(\mathbf{x})|\, d\mathbf{x} < \infty\}$,

$\mathcal{L}^2$ = $\{f : \int_{[0,1)^s} f^2(\mathbf{x}) d\mathbf{x} < \infty\}$,

$E_\alpha(c)$ = $\{f : [0,1)^s \to \mathbb{R} : |\hat{f}(\mathbf{h})| \leq c\|\mathbf{h}\|_\pi^{-\alpha}\}$, où $\|\mathbf{h}\|_\pi = \prod_{j=1}^{s} \max(1, |h_j|)$,

lg = $\log_2$,

i.i.d. = indépendants et identiquement distribués,

ppcm(a, b) = plus petit commun multiple entre a et b,

pgcd(a, b) = plus grand commun diviseur entre a et b,

sgn(h) = $|h|/h$,

$\zeta(y)$ = $\sum_{i=1}^{\infty} i^{-y}$, fonction zeta de Riemann évaluée en y,

$\mathbf{e}_j$ = j^e vecteur canonique de $\mathbb{R}^s$,

= $(0, \ldots, 0, 1, 0, \ldots, 0)$, où le 1 est en j^e position,

S = $\{1, 2, \ldots, s\}$,

I = sous-ensemble de S,

$|I|$ = cardinalité de I,

$\tilde{I}$ = ensemble obtenu à partir de $I = \{i_1, \ldots, i_t\}$ en enlevant $i_1 - 1$ à chaque i_j,

= $\{1, i_2 - i_1 + 1, \ldots, i_t - i_1 + 1\}$,

$r(I)$ = étendue de I,

$= i_t - i_1 + 1$, si $I = \{i_1, \ldots, i_t\}$,

$S(t,k)$ = ensemble des I contenant k indices dont le premier est 1 et le dernier est $\leq t$,

$= \{I = \{i_1, \ldots, i_k\} : 1 = i_1 < \ldots < i_k \leq t\}$,

$\tilde{S}(t,1)$ = ensemble des I à indices successifs, tels que $1 \leq |I| \leq t$,

$= \bigcup_{j=1}^{t} \bigcup_{w=1}^{s-j+1} \{\{w, w+1, \ldots, w+j-1\}\}$,

$H(t_1, \ldots, t_d, d)$ = ensemble des I tels que $\tilde{I}$ est dans $S(t_u, u)$, pour $2 \leq u \leq d$, ou est de la forme $\{1, \ldots, j\}$, avec $1 \leq j \leq t_1$,

$= \bigcup_{2 \leq u \leq d} \{I : |I| = u, r(I) \leq t_u\} \cup \tilde{S}(t_1, 1)$,

$\mathbf{x}_i$ = $(x_{i1}, \ldots, x_{is})$ est un point dans $[0,1)^s$,

$\mathbf{x}_I$ = $(x_{i_1}, \ldots, x_{i_t})$, si $I = \{i_1, \ldots, i_t\}$,

P_N = $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ dénote un ensemble de N points,

$P_N(I)$ = projection de P_N sur le sous-espace de $[0,1)^s$ déterminé par les indices dans I,

$= \{(x_{ni_1}, \ldots, x_{ni_t}), n = 1, \ldots, N)\}$, où $I = \{i_1, \ldots, i_t\}$,

$L^\perp$ = réseau dual à L, où L est tel que $P_N = L \cap [0,1)^s$,

$= \{\mathbf{h} \in \mathbb{Z}^s : \mathbf{h} \cdot \mathbf{x} \in \mathbb{Z}$, pour tout $\mathbf{x} \in P_N\}$,

$L_I^\perp$ = $\{\mathbf{h} \in \mathbb{Z}^{|I|} : \mathbf{h} \cdot \mathbf{x}_I \in \mathbb{Z}$, pour tout $\mathbf{x} \in P_N\}$,

$I_{\mathbf{h}}$ = ensemble des coordonnées non-nulles de $\mathbf{h}$,

$= \{j : h_j \neq 0\}$,

$\mathbb{Z}_I^*$ = $\{\mathbf{h} \in \mathbb{Z}^s : I_{\mathbf{h}} = I\}$,

$\mathbb{N}_I^*$ = $\{\mathbf{h} \in \mathbb{N}^s : I_{\mathbf{h}} = I\}$, en supposant que $\mathbb{N} = \{0, 1, 2, \ldots\}$,

$\mathbf{h} \odot \mathbf{x}$ = $\sum_{j=1}^{s} \sum_{k=1}^{\infty} h_{j,k-1} x_{j,k} \bmod 2$, où les $h_{j,k}$ et les $x_{j,k}$ sont les coefficients de l'expansion binaire de h_j et x_j, respectivement, pour $h_j \in \mathbb{N} \cup \{0\}$ et $x_j \in [0,1]$,

$\mathbf{k}_I$ = $(k_j)_{j \in I}$, où les k_j sont des entiers non négatifs,

$H(\mathbf{k}_I)$ = ensemble des vecteurs $\mathbf{h} \in \mathbb{N}_I^*$ tels que $|h_j|_p = 2^{k_j}$ pour tout $j \in I$, où $\mathbf{k}_I = (k_j)_{j \in I}$ et $k_j \geq 0$,

$\kappa = \sum_{j \in I} k_j$, pour un vecteur $\mathbf{k}_I$ donné,

ACV = *(Antithetic variates and Control Variable)*,

COND = Conditionnement,

IS = *(Importance Sampling)*,

LHS = *(Latin Hypercube Sampling)*,

LR = *(Lattice Rule)* translatée ; voir section 2.2,

LRp = *(Lattice Rule)* permutée, puis translatée ; voir section 2.3,

LRSt = *(Lattice Rule)* Stratifiée ; voir section 2.4,

MCC = Monte Carlo Conditionnelle,

VC = Variable de Contrôle.

Remerciements

Je remercie d'abord les fonds FCAR et CRSNG pour leur soutien financier tout au long de mes études de maîtrise et de doctorat.

Je tiens à exprimer ma gratitude envers mon directeur de thèse, Pierre L'Ecuyer, pour m'avoir permis de participer à plusieurs conférences, pour les nombreuses portes qu'il m'a ouvertes, les bonnes idées qu'il m'a données et pour le fait qu'il exige toujours un travail de qualité.

Ensuite, je veux remercier Raymond Couture, car il a contribué de façon importante à ce travail, notamment en ayant l'idée de la XOR-translation pour les règles de réseau polynômiales et en me redirigeant dans le bon chemin pour l'étude des fonctions monotones.

Les discussions avec Felisa Vázquez-Abad, en particulier au sujet des options asiatiques, ainsi que celles avec Art Owen et Geneviève Gauthier m'ont également aidée. Merci à Renée Touzin et François Panneton, pour les différents programmes de recherche qu'ils ont écrits et dont je me suis servi, aux membres du soutien technique du département, pour leur fiabilité et leur efficacité, et à Richard Simard, pour son aide en programmation.

De plus, je tiens à remercier les membres du jury pour leurs commentaires et suggestions qui ont permis d'améliorer la présentation de cet ouvrage.

Je ne peux passer sous silence l'apport considérable de John Watrous à la réalisation de ce travail. Il m'a non seulement aidée à résoudre certains problèmes mathématiques, mais sa sagesse et ses encouragements m'ont été d'un immense réconfort.

Finalement, je veux remercier mes frères et soeurs. Étant la plus jeune d'une famille de six enfants, je n'ai jamais eu à chercher bien loin pour trouver des modèles à tenter

d'imiter et des chemins intéressants à suivre. Merci à mes parents, Lise et Vincent Lemieux, pour la confiance qu'ils ont toujours eue en moi et en mon jugement, ainsi que pour leur soutien implicite et constant dans ma quête d'indépendance.

Chapitre 1

Introduction

Nous énonçons à la section 1.1 le problème qui est discuté dans cet ouvrage, faisons à la section 1.2 un résumé des travaux les plus importants qui lui sont reliés et énumérons les contributions qui y sont faites à la section 1.3. Les définitions formelles qui sont requises pour énoncer les résultats de cette thèse sont données au début des chapitres 2 et 4.

1.1 Problématique

Si l'intégrale

$$\mu = \int_{[0,1)^s} f(\mathbf{x}) d\mathbf{x} \tag{1.1}$$

d'une fonction f à valeur réelle définie sur l'hypercube unitaire en s dimensions n'a pas de solution analytique, on peut choisir un ensemble de points $P_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset [0,1)^s$ et estimer μ par la valeur moyenne de f sur P_N, donnée par

$$Q_N = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i).$$

Lorsque la dimension s est petite et que la fonction est suffisamment régulière, on peut construire l'ensemble P_N en appliquant sur chacune des s dimensions une règle qui fonctionne bien en une dimension : ceci correspond aux *règles de type produit*. Un cas particulier de cela consiste à prendre une *grille rectangulaire*, c.-à-d., à poser $P_N = \{(m_1/n, \ldots, m_s/n) : 0 \leq m_t \leq n-1, 1 \leq t \leq s\}$, où $N = n^s$.

Si la dimension s est modérément grande (disons, supérieure à 4 ou 5) ou que la fonction n'est pas très régulière, la *méthode de Monte Carlo* (MC) est plus appropriée. Cette méthode, introduite par Metropolis et Ulam en 1949 [91], consiste à prendre N points i.i.d (indépendants et identiquement distribués) uniformes sur $[0,1)^s$ pour construire P_N. Si on dénote par $\hat{\mu}_{\mathrm{MC}}$ la valeur de Q_N obtenue avec un tel ensemble, alors on a que

$$\mathrm{E}(\hat{\mu}_{\mathrm{MC}}) = \mu, \qquad \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) = \sigma^2/N,$$

où $\sigma^2 = \int_{[0,1)^s} f^2(\mathbf{x})d\mathbf{x} - \mu^2$ est la variance de $f(\mathbf{x})$ lorsque $\mathbf{x} \sim U([0,1)^s)$. De plus, en supposant que $\sigma^2 < \infty$, on obtient par le théorème de la limite centrale que

$$(\hat{\mu}_{\mathrm{MC}} - \mu) = O_p(N^{-1/2}), \tag{1.2}$$

où $O_p(g(n)) = \{h(n) :$ il existe $n_0 \in \mathbb{N}$, $c \in \mathbb{R}$ tels que $\Pr(h(n) \leq cg(n)) \geq p$, si $n \geq n_0\}$, pour $p \in (0,1]$. Ainsi, le comportement asymptotique (en fonction de N) de l'erreur ne dépend pas de la dimension s du problème. Par contre, la dimension s peut influencer l'erreur de façon indirecte en agissant sur σ.

Un avantage de cette méthode est que la relation (1.2) nous permet d'estimer l'erreur en se basant sur l'écart-type échantillonal $\hat{\sigma}$. De plus, la variance de l'estimateur $\hat{\mu}_{\mathrm{MC}}$ peut être estimée par $\hat{\sigma}^2/N$. Autrement dit, la méthode MC nous fournit non seulement un estimateur de μ, mais aussi une idée de la précision de cet estimateur.

Quel est le lien entre cette méthode et la *simulation stochastique*? Cette dernière est une technique utilisée pour estimer une mesure de performance donnée sur un système qui évolue de façon stochastique. Elle consiste à programmer sur ordinateur un modèle mathématique du système avec certaines entrées (aléatoires) et à observer les résultats. La mesure de performance peut habituellement s'écrire comme étant

$$\mu = \mathrm{E}\left(g(Y_1, \ldots, Y_M)\right), \tag{1.3}$$

où $\{Y_1, \ldots, Y_M\}$ est l'ensemble de variables aléatoires correspondant aux entrées du modèle. Le nombre de variables M peut également être aléatoire. Dans le programme de simulation, les variables aléatoires $Y_1, \ldots, Y_M$ sont générées en transformant des nombres pseudo-aléatoires entre 0 et 1, c.-à-d., des nombres qui imitent des variables

uniformes i.i.d. sur $[0,1)$. Pour cette raison, par simple changement de variables, la mesure de performance (1.3) peut être réécrite sous la forme (1.1), c.-à-d.,

$$\mu = \int_{[0,1)^s} f(\mathbf{u})\mathrm{d}\mathbf{u},$$

où $\mathbf{u} \in [0,1)^s$ est le vecteur de variables (pseudo-)aléatoires entre 0 et 1 nécessaires pour générer $Y_1, \ldots, Y_M$. La dimension s peut être infinie, car même si chaque simulation requiert un nombre fini γ de variables uniformes, ce nombre γ peut être aléatoire et il se peut qu'il n'y ait pas de borne s telle que la probabilité que γ soit inférieur à s soit égale à 1. En somme, la reformulation de (1.3) à (1.1) nous permet de dire que la simulation stochastique correspond en fait à utiliser la méthode MC afin d'estimer l'intégrale d'une fonction souvent compliquée et de dimension habituellement grande ou même infinie. L'exemple suivant sert à illustrer comment le passage de (1.3) à (1.1) se fait sur un problème simple. Nous verrons aux sous-sections 2.5.1 et 2.5.2 deux autres exemples de reformulation sur des problèmes un peu plus compliqués.

Exemple 1.1.1 *Supposons que l'on veut estimer* $\mathrm{E}(\bar{W}_{100})$, *l'espérance de la durée de séjour moyenne des 100 premiers clients d'un système de file d'attente de type* $GI/GI/1$ *(c.-à-d., il y a un seul serveur et les durées interarrivées entre les clients ainsi que les durées de service sont des variables aléatoires indépendantes [69]), pour lequel la politique de service est "premier arrivé premier servi". En dénotant par* S_i *la durée de service du* i^{e} *client et par* A_i *la durée interarrivée entre le* $(i-1)^{\mathrm{e}}$ *et le* i^{e} *client, on a que*

$$\mathrm{E}(\bar{W}_{100}) = \mathrm{E}(g(A_1, S_1, \ldots, A_{100}, S_{100})),$$

où

$$g(A_1, S_1, \ldots, S_{100}) = \frac{1}{100}\sum_{i=1}^{100}(W_i + S_i) \tag{1.4}$$

et W_i *est le temps d'attente dans la file du* i^{e} *client, qui satisfait l'équation de Lindley [87], c.-à-d.,*

$$W_i = \max(0, W_{i-1} + S_{i-1} - A_i), \tag{1.5}$$

en supposant que $W_0, S_0 = 0$. *On voit bien qu'en remplaçant successivement chaque* W_i *dans (1.4) par l'expression (1.5),* $\bar{W}_{100}$ *est une fonction des* A_i, S_i. *En utilisant la*

notation donnée précédemment, les variables aléatoires $A_1, S_1, \ldots, S_{100}$ *correspondent à* $Y_1, \ldots, Y_M$, *avec* $M = 200$.

Si on suppose que l'inversion est utilisée pour générer les A_i *et les* S_i *et que* $u_1, u_2, \ldots$ *représente la suite de nombres entre 0 et 1 produits par le générateur pseudo-aléatoire, alors on peut poser*

$$\begin{aligned} A_i &= F_A^{-1}(u_{2i-1}), \\ S_i &= F_S^{-1}(u_{2i}), \end{aligned}$$

où $F_A(\cdot)$ *et* $F_S(\cdot)$ *sont les fonctions de répartition de* A_i *et* S_i, *respectivement. On peut donc réécrire* $g(\cdot)$ *en fonction des* u_i :

$$g(A_1, \ldots, S_{100}) = g(F_A^{-1}(u_1), \ldots, F_S^{-1}(u_{200}))$$

et donc, si on pose

$$f(u_1, \ldots, u_{200}) = g(F_A^{-1}(u_1), \ldots, F_S^{-1}(u_{200})),$$

on obtient que

$$\mathrm{E}(\bar{W}_{100}) = \int_{[0,1)^{200}} f(\mathbf{u}) d\mathbf{u}.$$

Étant donné ce lien entre la méthode MC et la simulation, on pourrait se poser la question suivante : si d'autres méthodes que MC sont disponibles pour estimer des intégrales en grande dimension, pourrait-on les appliquer afin d'obtenir de meilleurs estimateurs pour la simulation? C'est le but de cette thèse de présenter de telles méthodes et de montrer comment elles peuvent être utiles pour contruire des estimateurs à variance réduite en simulation.

Avant d'expliquer cette idée plus en détail, nous présentons un exemple qui permet de voir quel genre de construction on doit éviter pour faire mieux que MC quand la dimension devient grande :

Exemple 1.1.2 *Supposons que l'on veut intégrer*

$$f(x_1, x_2, x_3) = 2x_1x_2 + 3x_3^2 + x_2.$$

On peut montrer qu'il est possible de décomposer f de la façon suivante :

$$\begin{aligned} f(x_1,x_2,x_3) &= f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{1,2}(x_1,x_2), \\ &= 2 + (x_1 - 1/2) + (2x_2 - 1) + (3x_3^2 - 1) + (2x_1x_2 - x_1 - x_2 + 1/2). \end{aligned}$$

Donc, même si la fonction est définie sur $[0,1)^3$, elle est en fait une somme de fonctions unidimensionnelles $(f_j(x_j),\ j = 1,2,3)$ et bidimensionnelle $(f_{1,2}(x_1,x_2))$. On peut également montrer que les composantes $f_3(x_3)$ et $f_2(x_2)$ contribuent pour 64% et 27% de la variance totale, respectivement. Ce type de décomposition est appelée "décomposition ANOVA" et s'avère très utile pour analyser les méthodes d'intégration numérique en plusieurs dimensions, comme nous le verrons à plusieurs reprises dans cet ouvrage.

Pour cet exemple, l'utilisation de la grille rectangulaire contenant $N = n^3$ points

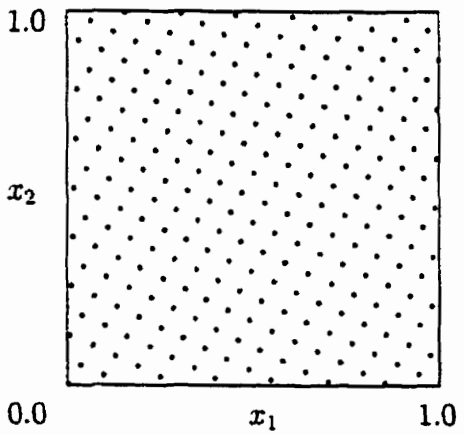
$$P_N = \left\{ \left(\frac{m_1}{n}, \frac{m_2}{n}, \frac{m_3}{n} \right), m_j = 0, \ldots, n-1, j = 1,2,3 \right\}.$$

n'est pas recommandée. En effet, on obtient alors que chaque $f_j(x_j)$ est intégrée par seulement $n = N^{1/3}$ points distincts et la fonction $f_{1,2}(x_1,x_2)$ est intégrée par seulement $n^2 = N^{2/3}$ points distincts. Or, il n'est pas difficile de croire qu'un ensemble P_N pour lequel chacune des projections $P_N(I)$ (où pour $I = \{i_1, \ldots, i_t\} \subseteq \{1, \ldots, 3\}$ non vide, $P_N(I) = \{(x_{ni_1}, \ldots, x_{ni_t}, n = 1, \ldots, N\})$ sur les $2^3 - 1$ sous-espaces de $[0,1)^3$ contient N points pourrait faire mieux que la grille et ceci deviendra de plus en plus facile à croire à mesure que la dimension s augmente puisque de façon générale, la grille rectangulaire utilise $N^{j/s}$ points différents pour intégrer les composantes f_I en $j = |I|$ dimensions.

La méthode MC a la propriété que chaque projection de P_N contient N points avec probabilité 1, car les points sont i.i.d. et donc, la probabilité que deux points $\mathbf{x}_i$ et $\mathbf{x}_j$ aient une coordonnée en commun est nulle. Mais il est également possible d'utiliser des méthodes déterministes ayant cette propriété, ainsi que celle d'avoir une distribution plus uniforme que celle d'un ensemble de points i.i.d. : un ensemble ayant cette propriété est appelé *ensemble de points à faible discrépance.* C'est ce type d'ensemble qui est utilisé dans les méthodes *quasi-Monte Carlo* (QMC) et les

règles de réseau sont un exemple de ce type de méthode. La figure 1.1 donne un exemple de règle de réseau en deux dimensions. Les 251 points illustrés sur cette figure sont définis par $x_{i1} = (i-1)/251$ et $x_{i2} = (33(i-1) \bmod 251)/251$, pour $i = 1, \ldots, 251$. Ainsi, les projections unidimensionnelles $\{x_{i1}, i = 1, \ldots, 251\}$ et $\{x_{i2}, i = 1, \ldots, 251\}$ correspondent toutes deux à l'ensemble de points également espacés donné par $\{0, 1/251, \ldots, 250/251\}$. Les différents concepts associés aux règles de réseau seront expliqués en détail aux sous-sections 2.1.1 et 2.1.2.

FIGURE 1.1: Règle de réseau en deux dimensions



La *discrépance* $D(P_N)$ mesure la non-uniformité d'un ensemble de points P_N dans $[0,1)^s$ en comparant sa distribution empirique dans $[0,1)^s$ avec celle de la loi uniforme [98, 46]. Il y a plusieurs façons de comparer ces deux distributions. Par exemple, si on considère toutes les boîtes rectangulaires dans $[0,1)^s$ qui sont alignées avec les axes et qui ont un coin à l'origine, alors on peut calculer la différence entre la fraction des points de P_N contenus dans la boîte et son volume, puis prendre $D(P_N)$ égale au supremum sur toutes les boîtes de cette différence (en valeur absolue). Ceci est la définition de la *discrépance-étoile* (*rectangular star-discrepancy*) [98]. On dit habituellement qu'une suite de points est à faible discrépance si, lorsque l'on prend les N premiers points de la suite pour former P_N, la discrépance $D(P_N)$ est significativement plus petite que celle d'un ensemble comprenant N points i.i.d. suivant la loi uniforme sur $[0,1)^s$. Dans le cas particulier où on utilise la discrépance-étoile pour définir $D(P_N)$, il est courant de dire que P_N est à faible discrépance si $D(P_N) = O(N^{-1} \log^s N)$.

La notion de discrépance permet également de construire des bornes sur l'erreur

déterministe ayant la forme suivante :

$$|Q_N - \mu| \leq D(P_N)V(f), \text{ pour toute fonction } f \in E, \tag{1.6}$$

où $V(f)$ est une mesure de la variabilité de f et E est un ensemble de fonctions, tous deux reliés à la définition précise de discrépance $D(P_N)$ utilisée [46]. L'inégalité de ce type la plus connue est sans doute celle de Koksma-Hlawka, dans laquelle $D(P_N)$ est la discrépance-étoile, $V(f)$ est la *variation au sens de Hardy et Krause* et E est l'ensemble des fonctions à variation bornée. Le lecteur peut se référer à [62, 98] pour plus de détails sur ce type de borne.

L'argument habituel qui est utilisé pour justifier la supériorité des méthodes QMC sur la méthode MC est de dire que puisque $V(f)$ ne dépend que de f, si $D(P_N)$ est dans $O(N^{-1}\log^s N)$ et que $V(f) < \infty$, alors la borne supérieure sur l'erreur donnée par (1.6) est aussi dans $O(N^{-1}\log^s N)$, ce qui, pour s fixé, est un meilleur taux de convergence que le $N^{-1/2}$ associé à la méthode MC. Pour cette raison, on s'attend à ce que les méthodes QMC approximent μ avec une plus petite erreur que la méthode MC, du moins si N est suffisamment grand. Par contre, la dimension s n'a pas besoin d'être très grande pour que l'on ait $N^{-1}\log^s N > N^{-1/2}$ pour des valeurs de N assez grandes : par exemple, quand $s = 10$, on doit avoir $N \geq 1.2 \times 10^{39}$ pour que $N^{-1}\log^s N \leq N^{-1/2}$ et donc, l'avantage de l'ordre de convergence de l'erreur de QMC sur MC prévaut seulement si on utilise plus de 1.2×10^{39} points, un ordre de grandeur extrêmement plus grand que ce que l'on veut en pratique.

Est-ce que cela signifie que les méthodes QMC ne servent pas à grand-chose en pratique ? Non, car premièrement, l'inégalité donnée en (1.6) est une borne en pire cas qui ne reflète pas nécessairement le comportement réel de Q_N pour une fonction en particulier ; deuxièmement, comme c'était le cas dans l'exemple 1.1.2, même si une fonction est définie sur $[0,1)^s$, il se peut que seulement les composantes f_I associées à des sous-ensembles I contenant un petit nombre de dimensions suffisent à expliquer la majeure partie de la variabilité de f et dans ce cas, il suffit *grosso modo* d'utiliser un ensemble de points dont les projections $P_N(I)$ associées à ces composantes sont bien distribuées. Dans ce cas, l'analyse de l'erreur peut se faire comme si la dimension était égale à $\max\{|I| : f_I \text{ est "importante"}\} \leq s$, qui définit de façon approximative la

dimension effective de f [110, 14, 46]. En résumé, les bornes du genre (1.6) ne sont pas à notre avis les meilleurs outils théoriques que l'on puisse utiliser afin d'expliquer la supériorité observée en pratique des méthodes QMC sur la méthode MC. En particulier, elles ont le désavantage de ne pas pouvoir nous fournir une bonne idée de l'erreur réelle d'intégration : nous reviendrons sur ce point sous peu.

Maintenant, comment utiliser ces méthodes pour faire de la "simulation de quasi-Monte Carlo" ? D'abord, une fois que l'ensemble de points P_N est choisi, chaque point $\mathbf{x}_i$ devrait être assigné à une "simulation". Notons que cela implique que l'on doit décider l'ordre dans lequel les composantes $x_{i1}, \ldots, x_{is}$ seront associées aux variables aléatoires constituant les entrées de la simulation. Par exemple, dans l'exemple 1.1.1, on aurait pu décider d'associer u_j à A_j et u_{j+100} à S_j plutôt que d'associer u_{2j-1} à A_j et u_{2j} à S_j. L'ordre choisi doit évidemment demeurer le même pour chaque point $\mathbf{x}_i$.

Ensuite, on doit avoir une façon d'estimer l'erreur. Pourrait-on utiliser un cas particulier de la borne (1.6) (par exemple, l'inégalité de Koksma-Hlawka) ? Non, et plusieurs raisons peuvent nous en convaincre :

(1) Ce type de borne tient pour un ensemble trop restreint de fonctions. On préfèrerait travailler dans un contexte valide pour toute fonction à variance finie, comme c'est la cas pour la méthode MC.

(2) Même en supposant que la fonction correspondant à notre problème est dans un ensemble pour lequel on a une borne de type (1.6), la vraie erreur est probablement beaucoup plus petite que la borne puisque, comme nous l'avons dit précédemment, ces bornes résultent d'une analyse en pire cas.

(3) Même en supposant que la borne pourrait nous donner une bonne idée de l'erreur, les deux quantités $V(f)$ et $D(P_N)$ sont souvent très difficiles à calculer.

Une solution beaucoup plus simple et n'ayant pas ces inconvénients est de randomiser l'ensemble P_N de façon à ce que 1) la propriété qui fait en sorte que P_N est à faible discrépance soit préservée ; 2) après randomisation, chaque point suive la loi uniforme sur $[0,1)^s$. La deuxième condition nous assure que l'estimateur formé à partir de l'ensemble randomisé $\tilde{P}_N$ est sans biais. En répétant la randomisation m fois, on obtient m copies i.i.d de $\tilde{P}_N$ et ainsi, on peut estimer l'erreur en utilisant le théorème de la

limite centrale ou calculer un estimateur sans biais de $\mathrm{Var}(\tilde{Q}_N)$, où $\tilde{Q}_N$ est l'estimateur obtenu à partir de $\tilde{P}_N$. La première condition requise pour la randomisation nous permet d'espérer que l'on obtiendra un estimateur réduisant la variance par rapport à MC. Pour donner un exemple, Cranley et Patterson ont proposé dans [20] une randomisation pour les règles de réseau satisfaisant ces deux propriétés et qui consiste à translater les règles (modulo 1) à l'aide d'un vecteur $\mathbf{u}$ généré aléatoirement et uniformément dans $[0,1)^s$. Nous reparlerons de cette technique à la sous-section 2.2. En résumé, non seulement la randomisation nous fournit une façon d'estimer l'erreur (ou la variance) de l'estimateur basé sur une règle de réseau, mais elle nous permet aussi de comparer les règles de réseau avec la méthode MC en considérant leur variance respective (théorique et empirique), ce qui est un des objectifs principaux de cet ouvrage.

Dans le contexte de la simulation, cette méthode s'applique de la façon suivante : on choisit un ensemble P_N à faible discrépance et une randomisation satisfaisant les deux conditions énoncées plus haut. Chacun des Nm points contenus dans les m copies i.i.d. $\tilde{P}_{N,1}, \ldots, \tilde{P}_{N,m}$ de la version randomisée de P_N sert à faire une simulation et nous permet de construire les m estimateurs i.i.d. et sans biais de μ donnés par

$$\tilde{Q}_{N,j} = \frac{1}{N} \sum_{\tilde{\mathbf{x}}_i \in \tilde{P}_{N,j}} f(\tilde{\mathbf{x}}_i), \qquad j = 1, \ldots, m.$$

La moyenne $\hat{\mu} = 1/m \sum_{j=1}^{m} \tilde{Q}_{N,j}$ de ces estimateurs est notre estimateur sans biais de μ,

$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_{j=1}^{m} (\tilde{Q}_{N,j} - \hat{\mu})^2 \tag{1.7}$$

nous fournit un estimateur de $\mathrm{Var}(\tilde{Q}_N)$ et pour un niveau de confiance $1-\alpha$ donné, la quantité

$$\frac{\hat{\sigma}}{\sqrt{m}} z_{1-\alpha/2}$$

nous donne une borne probabiliste sur l'erreur, c.-à-d., on a que

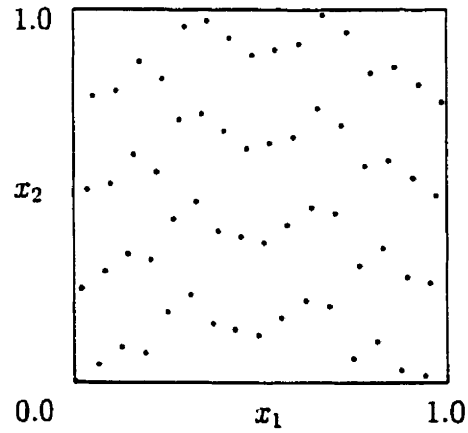
$$P\left(|\hat{\mu} - \mu| \leq \frac{\hat{\sigma}}{\sqrt{m}} z_{1-\alpha/2} \right) \approx 1 - \alpha,$$

où $z_{1-\alpha/2}$ est tel que $P(N(0,1) \leq z_{1-\alpha/2}) = 1 - \alpha/2$ [68].

Il faut maintenant préciser comment nous allons construire les ensembles de points à faible discrépance dont nous avons besoin pour appliquer la méthode QMC. Dans

cet ouvrage, nous considérons deux possibilités : les règles de réseau "standard" et les règles de réseau polynômiales. Comme leur nom l'indique, ces méthodes sont toutes deux basées sur l'utilisation d'un réseau *(lattice)* dans un espace $\mathcal{E}$ (c.-à-d., un sous-espace de $\mathcal{E}$ fermé sous l'addition et la soustraction) pour construire P_N. Dans le premier cas (standard), les réseaux sont définis sur l'espace euclidien $\mathbb{R}^s$ alors que dans le second, ils sont définis sur un espace de polynômes. Nous avons donné à la page 6 (figure 1.1) un exemple d'une règle de réseau définie sur $\mathbb{R}^2$ et la figure 1.2 montre un exemple de règle de réseau polynômiale en deux dimensions. Les 64 points illustrés sur cette figure sont basés sur le polynôme de degré six donné par $P(z) = z^6 - z - 1$ et sont définis par $(x_{01}, x_{02}) = (0, 0)$ et $x_{i1} = (z^{4(i-1)}/P(z) \bmod (P(z), 2))\big|_{z=2}$ et $x_{i2} = (z^{4i}/P(z) \bmod (P(z), 2))|_{z=2}$, pour $i = 1, \ldots, 63$, où " mod $(P(z), 2)$" signifie qu'on prend le reste de la division polynômiale par $P(z)$, en supposant que les opérations sur les coefficients sont effectuées dans $\mathbb{F}_2$, le corps de Galois contenant deux éléments. Ce type de construction sera expliqué en détail à la section 4.1.

FIGURE 1.2: Règle de réseau polynômiale avec $N = 64$



Le lien entre les deux méthodes retenues peut aussi être établi dans un contexte beaucoup plus général (qui comprend d'autres types de méthodes), mais qui peut être expliqué sans avoir recours à la définition de règle de réseau. Nous voulons donner ce point de vue, car il permet de faire des liens entre les générateurs pseudo-aléatoires et les ensembles de points à faible discrépance. D'une certaine façon, ceci établit un lien supplémentaire entre les méthodes MC et QMC, puisque les générateurs pseudo-aléatoires servent à générer l'ensemble P_N dans la méthode MC.

En gros, l'idée est la suivante : le type de méthode que nous considérons revient à choisir un générateur pseudo-aléatoire dont le nombre d'états possibles n'est pas trop grand, puis à former P_N en utilisant les s-tuplets successifs formés à partir des différentes sous-suites périodiques produites par le générateur. Historiquement, l'idée d'utiliser des suites pseudo-aléatoires comme méthode QMC a d'abord été énoncée dans [95].

Plus précisément, si on a un générateur pseudo-aléatoire défini sur l'espace d'états Ξ de cardinalité $|\Xi| = N$ avec la fonction de transition $\tau : \Xi \to \Xi$, qui nous fait passer d'un état à l'autre en posant $\xi_i = \tau(\xi_{i-1})$ et ayant une fonction de sortie $g : \Xi \to [0,1)$, qui produit les nombres entre 0 et 1 émis par le générateur, alors on construit P_N de la façon suivante :

$$P_N = \{(g(\xi_0), \ldots, g(\xi_{s-1})) : \xi_0 \in \Xi\}. \tag{1.8}$$

Voici un exemple jouet pour mieux comprendre comment se fait la construction :

Exemple 1.1.3 *Soit* $\Xi = \{0, \ldots, 10\} = \mathbb{Z}_{11}$ *et* $\tau(\xi) = 6 \cdot \xi \bmod 11$. *En prenant le germe* $\xi_0 = 1$, *on obtient la suite* $\xi_0, \xi_1, \ldots$ *de période 10 donnée par :*

$$1, 6, 3, 7, 9, 10, 5, 8, 4, 2, 1, 6, \ldots$$

et tout autre germe différent de 0 donnera lieu à la même sous-suite, mais avec un état initial différent (par exemple, $\xi_0 = 2$ *donne* $2, 1, 6, 3, 7, \ldots$*). Si* $\xi_0 = 0$, *alors on a la suite de période 1 donnée par* $0, 0, \ldots$. *Il y a donc deux sous-suites associées à ce générateur. Ainsi, si la dimension s vaut 3 et que l'on utilise la fonction de sortie* $g(\xi) = \xi/11$, *alors on a que*

$$\begin{aligned}&\{(g(\xi_0), \ldots, g(\xi_{s-1})) : \xi_0 \in \Xi, \xi_0 \neq 0\} \\ &\qquad = \left\{\left(\frac{1}{11}, \frac{6}{11}, \frac{3}{11}\right), \left(\frac{6}{11}, \frac{3}{11}, \frac{7}{11}\right), \ldots, \left(\frac{4}{11}, \frac{2}{11}, \frac{1}{11}\right), \left(\frac{2}{11}, \frac{1}{11}, \frac{6}{11}\right)\right\}\end{aligned}$$

contient 10 éléments et P_N *est obtenu en ajoutant* $(g(0), g(0), g(0)) = (0, 0, 0)$ *à cet ensemble.*

Les deux types de générateurs pseudo-aléatoires que nous allons utiliser sont les *générateurs à congruence linéaire* (GCL) et les *générateurs de Tausworthe*. Les ensembles P_N obtenus de cette façon sont des cas particuliers de règles de réseau standard

et polynômiales, respectivement. Les deux types de générateurs choisis sont rapides, faciles à implanter, mais surtout, des mesures d'uniformité peuvent être calculées en pratique afin de juger de la qualité des ensembles P_N ainsi produits. Ces mesures sont le *test spectral* [19] pour les GCL et la *résolution* [72] pour les générateurs de Tausworthe. C'est en définissant des critères de sélection basés sur ces mesures que nous pourrons choisir les paramètres des générateurs pseudo-aléatoires nous permettant de construire des ensembles de points P_N ayant les bonnes propriétés d'uniformité nécessaires pour être utilisés dans un contexte QMC.

Ces mesures sont d'ailleurs reliées à des critères de sélection typiquement utilisés pour choisir des ensembles de points à faible discrépance, dans le contexte général où ils ne proviennent pas nécessairement de générateurs pseudo-aléatoires. En effet, le test spectral est équivalent, à un choix de norme près, à *l'index de Babenko-Zaremba*, un des critères de sélection souvent utilisés pour choisir les règles de réseau. De son côté, la résolution des générateurs de Tausworthe présente des liens avec la propriété définissant les (t, m, s)-*réseaux* [98], aussi appelées *règles de réseau digitales* dans [64]. Les (t, m, s)-réseaux constituent une des familles importantes d'ensembles de points à faible discrépance. Les suites de Sobol' [121], Faure [28] et Niederreiter-Xing [101] produisent de tels ensembles. Il est intéressant de voir que le test spectral et la résolution sont reliés aux critères qui ont été développés en parallèle pour choisir des familles connues d'ensembles de points à faible discrépance.

Un autre avantage associé au type de construction défini par (1.8) est qu'il est facile de trouver des fonctions τ faisant en sorte que l'ensemble P_N soit *stationnaire dans la dimension*. En gros, cette propriété signifie que les projections $P_N(I)$ dépendent seulement de l'espacement entre les indices dans I. Par exemple, lorsque P_N a cette propriété, alors on a que $P_N(\{1,2,4\}) = P_N(\{s-3, s-2, s\})$. Nous verrons une définition plus précise de cette propriété au chapitre 3, ainsi que les avantages qui lui sont associés.

On peut maintenant formuler la question globale à laquelle nous tentons de répondre dans cette thèse : *comment choisir des règles de réseau qui, dans le contexte de la simulation, nous permettront d'obtenir des estimateurs ayant une plus petite*

variance que l'estimateur MC qui utilise le même nombre de points ?

1.2 Survol de la littérature

Le but de cette section n'est pas de faire une revue complète de la littérature traitant de notre sujet, mais plutôt de situer notre contribution parmi les travaux qui lui sont reliés de plus près. Avant de faire cela, donnons d'abord quelques références générales couvrant notre domaine de recherche. Les méthodes QMC sont exposées en détail par Niederreiter dans [98] et les règles de réseau, par Sloan et Joe dans [116]. Une revue axée davantage sur les applications pratiques des méthodes QMC est faite par Spanier et Maize dans [125]. Fishman [29] et L'Ecuyer [70] couvrent les principales techniques de réduction de la variance en simulation. Pour une introduction à la simulation en général, on peut utiliser le livre de Bratley *et al.* [12], Law et Kelton [69] ou celui de Fishman [29].

Nous allons commencer par faire, dans les deux paragraphes qui suivent, un bref historique nous permettant de voir qui a contribué à faire avancer le domaine des méthodes QMC, d'abord pour les règles de réseau et ensuite, pour les (t, m, s)-réseaux. Nous décrirons par la suite quelles ont été les principales contributions quant aux aspects de ces méthodes qui sont considérés dans cet ouvrage.

Dans le cas des règles de réseau, les premières constructions ont été proposées par Korobov [60], Bahvalov [7] et Hlawka [50]. Ces travaux et ceux qui ont suivi sont résumés par Niederreiter [94, 98], qui a lui-même apporté de nouveaux éléments, notamment en ce qui a trait au lien entre ces règles et les générateurs pseudo-aléatoires. Ensuite, Sloan et Joe et leurs collaborateurs ont fait plusieurs contributions importantes qui sont résumées en détail dans leur livre [116]. En particulier, ils ont généralisé les constructions antérieures et ont précisé la façon de représenter les règles de réseau. Plus récemment, Hickernell a généralisé plusieurs concepts, surtout au niveau des critères de sélection pouvant être utilisés pour choisir les règles et de leur lien avec l'erreur d'intégration. Ces résultats se trouvent dans [46, 47] et les références qui y sont données. La contribution d'Hellekalek [42] au développement d'une structure générale pour décrire les critères de sélection mérite également d'être mentionnée et tout comme

Hickernell, sa théorie ne s'applique pas seulement aux règles de réseau.

Pour ce qui est des (t, m, s)-réseaux, Sobol' a proposé les premières constructions dans [121], qui étaient en base 2. Elles étaient alors appelées "P_τ-réseaux" et "LP_τ-suites". Faure a généralisé à d'autres bases dans [28], définissant ainsi des suites ayant de meilleures propriétés d'équidistribution que celles de Sobol', mais qui offrent moins de flexibilité quant au choix du nombre de points N. Puis, Niederreiter a formalisé et généralisé ces constructions dans plusieurs articles, qui sont résumés dans [98]. Ses constructions les plus récentes, développées avec Xing, se trouvent dans [102, 103]. Larcher, Schmid et leurs collaborateurs [65, 64, 114] ont apporté des contributions intéressantes, notamment en ce qui a trait à l'étude de l'erreur en utilisant une décomposition en série de Walsh. Owen [104, 105, 106, 109] a aussi fait d'importantes contributions, entre autres grâce à la randomisation qu'il a proposée pour ces méthodes. Finalement, mentionnons les travaux de Halton [39], qui a proposé une suite à faible discrépance souvent utilisée en pratique.

Maintenant, puisque l'utilisation de méthodes QMC en simulation telle que nous la décrivons dans cet ouvrage repose avant tout sur une randomisation appropriée, voyons quelles ont été les principales contributions à ce niveau. Pour les règles de réseau, Cranley et Patterson ont proposé dans [20] de translater aléatoirement la règle (modulo 1). Joe a étudié d'autres randomisations dans [53] et a analysé l'erreur associée. La variance des estimateurs obtenus par translation aléatoire a été étudiée par Tuffin dans [130], mais dans le contexte assez restrictif des fonctions dont la série de Fourier est absolument convergente.

Dans le cas des (t, m, s)-réseaux, Owen a proposé une randomisation dans [104], en construisant ce qu'il appelle des (t, m, s)-*réseaux brouillés (scrambled* (t, m, s)-*nets)*. Il a également étudié la variance théorique des estimateurs ainsi obtenus et l'a comparée avec celle de l'estimateur MC dans [105, 106, 109].

Tuffin [129] ainsi que Morohosi et Fushimi [93] ont étudié la randomisation par translation aléatoire des (t, m, s)-réseaux et ont comparé la variance associée, théoriquement et empiriquement, à celle des (t, m, s)-réseaux brouillés. En gros, leurs conclusions sont que théoriquement, les (t, m, s)-réseaux brouillés ont l'avan-

tage d'avoir une variance inférieure. Par contre, empiriquement, les deux types de randomisations sont à peu près équivalentes et la translation aléatoire a un coût de calcul moins élevé. Autrement dit, leurs résultats empiriques semblent indiquer que la différence entre les variances théoriques associées à ces deux randomisations est petite, du moins pour les problèmes qu'ils ont considérés.

Plus récemment, Wang et Hickernell [133] ont proposé une nouvelle randomisation pour la suite d'Halton et l'ont comparée à d'autres randomisations pour cette suite ainsi qu'à la méthode MC, en regardant la variance empirique des estimateurs obtenus sur une fonction-test et ont trouvé que leur randomisation semblait avantageuse. La randomisation de ces suites a aussi été étudiée par Spanier dans [124].

Nous voulons mentionner les travaux qui ont été faits sur la méthode de l'échantillonnage de l'hypercube latin *(latin hypercube sampling (LHS))*, puisqu'elle peut être vue comme une méthode QMC randomisée. Cette méthode a été introduite par McKay, Beckman et Conover dans [90]. La variance est étudiée dans [90], par Stein dans [126], par Avramidis et Wilson dans [5] et d'un point de vue QMC par Owen dans [105].

Nous voulons maintenant énumérer les contributions importantes qui ont été faites quant à l'utilisation de méthodes QMC sur des problèmes pratiques (contrairement à des fonctions-tests, par exemple). Or, il semble qu'il y ait beaucoup plus de publications dans lesquelles des méthodes non randomisées sont utilisées en pratique. En plus de Spanier et Maize [125], Owen donne dans [107] plusieurs références à ce sujet, pour des problèmes en finance, en infographie et en physique. L'article de Boyle *et al.* [11] constitue une source supplémentaire de références quant à l'application de ces méthodes en finance. En ce qui concerne les méthodes QMC randomisées, des (t, m, s)-réseaux brouillés sont utilisés par Caflisch *et al.* dans [14] pour l'évaluation de produits financiers reliés à des prêts hypothécaires *(mortgage-backed securities)*. Aussi, dans [108], Owen et Tavella utilisent des (t, m, s)-réseaux brouillés pour estimer la valeur au risque *(value-at-risk)* d'un contrat financier. Plus récemment, Fushimi *et al.* [32] ont utilisé des (t, m, s)-réseaux translatés aléatoirement comme méthode d'estimation sur différents problèmes financiers. Dans le cas des règles de réseau, nous n'avons pas trouvé de publications antérieures aux nôtres dans lesquelles des versions rando-

misées (ou non) servent à construire des estimateurs pour des problèmes pratiques, par exemple en finance.

Plusieurs résultats importants dans le domaine des méthodes QMC ont été obtenus en utilisant diverses décompositions de la fonction à intégrer. On peut voir cette approche comme étant une façon de "diviser-pour-régner" afin de mieux comprendre le comportement de l'approximation de μ basée sur P_N, c.-à-d., on brise le problème en une série de petits problèmes plus faciles à traiter (car les fonctions de base utilisées dans la décomposition sont simples), puis on recombine le tout. Dans le cas des (t, m, s)-réseaux, Sobol [122], Entacher [26], Owen [105, 106, 109], Morohosi et Fushimi [93] ainsi qu'Hickernell et Yue [49] ont utilisé une décomposition en fonctions de Haar afin d'étudier les (t, m, s)-réseaux. La décomposition en série de Walsh a été utilisée pour étudier l'erreur d'intégration obtenue par les (t, m, s)-réseaux par Larcher et ses collaborateurs dans [67, 66, 65]. Pour les règles de réseau, ce sont les séries de Fourier qui sont habituellement utilisées pour étudier l'erreur et construire des critères de sélection, tel qu'expliqué par Sloan et Joe [116] et Hickernell [44, 46]. Ces décompositions peuvent également être utilisées pour définir un critère de qualité généralisant le test spectral *(weighted spectral test)*. Ce concept a été introduit par Hellekalek et est expliqué dans [42]. Un cas particulier de ce type de mesure est la *diaphonie*, qui présente des liens importants avec la discrépance.

Un des champs de recherche actuels concernant les méthodes QMC est de définir des mesures de qualité pour les ensembles de points à faible discrépance qui regardent plus en détail leurs projections. À ce niveau, Hickernell a contribué de façon importante [45, 46, 47] en généralisant les mesures classiques telles que la discrépance-étoile et le critère P_α^s [116], qui est très couramment utilisé pour mesurer la qualité des règles de réseau. Il utilise pour cela des poids qui quantifient l'importance des différentes projections. En ce qui a trait aux (t, m, s)-réseaux, Owen [109], Larcher [64], Hickernell [47] et Schmid [114] ont proposé de généraliser le paramètre t qui mesure la qualité de ces ensembles à un vecteur de paramètres t_I associés aux différentes projections sur les sous-ensembles I de S, mais n'ont pas expliqué comment obtenir des constructions minimisant ces t_I. Citons également un article de Shaw [115] antérieur à ces derniers

et dans lequel l'auteur donne un tableau de règles de Korobov choisies en considérant un certain nombre de projections en basse dimension, mais pour un nombre de points N ne dépassant pas 900.

Nous voulons également mentionner deux approches qui ont été utilisées dans l'étude des méthodes QMC. Dans les deux cas, l'idée est de faire une analyse en moyenne plutôt qu'en pire cas, comme on le fait lorsque l'on travaille avec des inégalités du genre (1.6). La première approche consiste à regarder ce qui se passe en moyenne sur une famille donnée d'ensembles de points. Disney et Sloan utilisent cela dans [22] pour donner des bornes sur l'erreur d'intégration obtenue en utilisant une règle de réseau. Disney et Sloan [23] ainsi que Joe et Disney [54] s'en servent pour analyser le comportement du P_α^s moyen pour différentes sortes de règles de réseau. La deuxième approche est en quelque sorte la contrepartie de la première, puisqu'on s'intéresse maintenant à l'erreur moyenne sur un ensemble donné de fonctions ou encore, à la variance d'un estimateur QMC lorsque la fonction est choisie aléatoirement parmi un ensemble donné. Cette approche est étudiée par Wozniakowski [136], Wasilkowski [134], Hickernell [46, 47] et Hoogland *et al.* [52].

Finalement, nous voulons mentionner certaines contributions qui ont été faites pour améliorer les résultats des méthodes QMC sur des problèmes en grande dimension. Plus précisément, nous voulons parler des techniques pour réduire la dimension effective d'une fonction. Caflisch et Moskovitz [13] ont proposé la méthode du pont brownien, qui est utile lorsque la fonction dépend de la réalisation d'un mouvement brownien. Cette technique a été utilisée dans plusieurs articles en finance (par exemple, [1, 14, 135]). Dans [1], Acworth *et al.* ont suggéré une méthode basée sur l'analyse en composantes principales, qui permet de réduire la dimension effective lorsque la fonction dépend d'un vecteur de variables aléatoires de loi normale. Fox [31] a proposé des méthodes allant dans le même sens pour la génération de processus de Poisson. Nous ne parlerons pas plus de ce type de méthode ici, bien que nous ayons étudié leur combinaison avec des règles de réseau sur des problèmes en finance dans [84, 9].

1.3 Contributions de cette thèse

Nous voulons préciser qu'une partie des résultats présentés dans cet ouvrage se trouvent dans [83, 85, 79, 80].

Le chapitre 2 donne des résultats sur la variance des estimateurs formés à partir d'une règle de réseau standard randomisée et nous considérons trois différentes randomisations. Il commence avec un des résultats importants de cet ouvrage, qui consiste à donner une expression pour la variance de l'estimateur formé à partir d'une règle de réseau translatée (modulo 1) aléatoirement. Il existe une expression similaire pour l'erreur d'intégration [116, *Theorem* 2.8], sauf qu'elle tient seulement pour les fonctions dont la série de Fourier est absolument convergente, alors que celle que nous donnons pour la variance tient pour n'importe quelle fonction à variance finie. C'est la clé qui nous permet par la suite de donner des résultats qui s'appliquent à la grande majorité des fonctions que l'on peut rencontrer en simulation. Ce résultat a été démontré de façon indépendante par Tuffin dans [130], mais sous l'hypothèse que la fonction a une représentation en série de Fourier qui est absolument convergente.

Nous donnons ensuite des bornes sur la variance de l'estimateur sus-mentionné, qui tiennent en imposant certaines conditions (assez fortes) sur la fonction à intégrer. Mais ce qui nous intéresse davantage, c'est de comparer cette variance avec celle de l'estimateur MC basé sur le même nombre de points. Nous y arrivons en regardant la variance moyenne obtenue sur l'ensemble des règles de rang 1, pour un ordre N premier. La borne obtenue est la somme de la variance de l'estimateur MC (multipliée par une quantité près de 1) et de la variance de l'estimateur obtenu en utilisant une grille rectangulaire contenant N^s points. En imposant des conditions sur la fonction, on montre comment cette borne peut être resserrée (en diminuant le deuxième terme de la somme).

Le reste du chapitre est consacré à l'étude de deux autres méthodes de randomisation pour les règles de réseau, qui peuvent être vues comme des extensions de la précédente. Dans la première, on permute au hasard les coordonnées des N points dans chaque dimension, avant de faire la translation modulo 1. De cette façon, la réduction de variance est garantie pour les fonctions monotones par rapport à chacun de leurs

arguments. Ce résultat était connu pour les méthodes des variables antithétiques et de l'échantillonnage de l'hypercube latin (LHS) [40, 90, 5]. Remarquons que cette randomisation détruit la propriété de réseau et nous donne un ensemble de points qui s'apparente fortement à celui obtenu par la méthode LHS, mais requiert moins de travail. En effet, on utilise ici le *même* vecteur pour translater tous les points, alors que dans LHS, on utilise un vecteur différent par point. Par contre, en comparaison avec la méthode par translation aléatoire, le temps de calcul pour générer les points est beaucoup plus long. Comme corollaire de ce résultat, on obtient qu'en une dimension, si la fonction est monotone, alors la réduction de variance est garantie en utilisant une règle de réseau translatée aléatoirement. Un résultat impliquant ce dernier était connu [30], mais en supposant des conditions plus fortes sur la fonction à intégrer.

La deuxième randomisation consiste à partitionner l'hypercube en utilisant une certaine base pour le réseau, puis à translater aléatoirement chaque point à l'intérieur du parallélépipède fondamental qui lui est ainsi associé. Puisqu'on se trouve de cette façon à utiliser la stratification, la réduction de variance est garantie pour toute fonction dont la variance est finie. Par contre, on perd ici aussi la propriété de réseau et le travail requis pour générer les points augmente par rapport à la méthode par translation aléatoire.

Nous concluons le chapitre en donnant des résultats numériques provenant de trois problèmes afin d'illustrer comment les règles de réseau translatées aléatoirement peuvent réduire la variance par rapport à la méthode MC. Nous faisons également des comparaisons avec les deux autres randomisations proposées.

Au chapitre 3, nous étudions les projections des règles de réseau standard et des fonctions à intégrer sur les sous-espaces de l'hypercube en s dimensions. Nous supposons tout au long du chapitre que la randomisation choisie est celle par translation aléatoire. À la section 3.3, nous donnons des lemmes qui relient la décomposition ANOVA d'une fonction à celle en série de Fourier. Ceci nous permet de réécrire la variance de l'estimateur obtenu par translation sous une forme qui justifie le nouveau critère de sélection que nous proposons à la section 3.5. Ce nouveau critère accorde plus d'importance que ses prédécesseurs aux projections de l'ensemble de points sur les

sous-espaces de l'hypercube en basse dimension, tout en étant calculable en pratique. Son avantage par rapport aux critères traditionnels est illustré à l'aide d'exemples numériques où l'on voit que dans certains cas, son utilisation fait la différence entre réduire la variance par rapport à MC ou non. Des tableaux énumérant les meilleurs règles par rapport à ce nouveau critère sont également donnés.

Nous donnons aussi dans ce chapitre des expressions pour la variance des composantes ANOVA lorsque la fonction est un polynôme. Ceci nous permet de donner des conditions pour qu'il y ait réduction de variance dans ce cas. Cette étude nous conduit également à la définition d'un critère de sélection spécifique au polynôme que l'on veut intégrer et dont la minimisation nous assure de choisir une règle qui minimise une partie importante de la variance. Plusieurs liens avec le critère P_α^s [116] sont faits dans cette section. Nous nous intéressons ensuite au cas des polynômes de degré 1 et 2, ce qui nous permet de donner des résultats plus précis sur la réduction de variance. De plus, pour les polynômes de degré 2, nous donnons un critère de sélection qui dépend du polynôme et qui nous permet de trouver une règle minimisant la variance. En combinant ce résultat avec un autre, qui démontre qu'en moyenne, sur l'ensemble des règles de rang 1 pour un ordre N premier, la variance de l'estimateur obtenu avec une règle de réseau est bornée supérieurement par une quantité près de 1 multipliée par la variance de l'estimateur MC, on montre que la règle choisie selon ce critère ne peut faire bien pire que l'estimateur MC.

Au chapitre 4, nous étudions les règles de réseau polynômiales. Nous donnons d'abord à l'annexe E deux lemmes que nous avons établis afin de faciliter la manipulation des séries de Walsh, car ces dernières nous servent par la suite à étudier la variance des estimateurs formés à partir de règles de réseau polynômiales. Ensuite, nous donnons des résultats expliquant le lien entre la *résolution* d'une règle [72] et un autre paramètre d'équidistribution, que nous appelons q_l et qui est relié au paramètre t définissant les (t, m, s)-réseaux. Nous présentons également un nouveau critère de sélection pour les règles de réseau polynômiales, qui est en fait l'équivalent de celui donné au chapitre 3 pour les règles de réseau standard. En analogie avec les résultats obtenus dans le cas standard, nous donnons une expression pour la variance des estima-

teurs formés à partir d'une règle de réseau polynômiale *XOR-translatée* et faisons des liens entre la décomposition ANOVA et celle en série de Walsh. Cette expression nous permet par la suite de donner des bornes sur la variance qui dépendent du nouveau critère de sélection et qui sont valides lorsque la fonction respecte de fortes conditions.

Nous comparons ensuite les propriétés de la variance de l'estimateur obtenu par XOR-translation avec celle de l'estimateur obtenu à l'aide d'un (t, m, s)-réseau brouillé, telle qu'étudiée par Owen dans [105, 106, 109]. Finalement, nous regardons comment le taux de convergence de la variance des estimateurs XOR-translatés varie en fonction des hypothèses sur le taux de décroissance des coefficients de Walsh de la fonction et de la norme utilisée pour mesurer les vecteurs $\mathbf{h}$ dans $\mathbb{N}^s$ en lesquels les coefficients de Walsh $\tilde{f}(\mathbf{h})$ sont évalués. Dans [67, 65], Larcher et ses collaborateurs ont utilisé la décomposition en série de Walsh pour analyser le comportement asymptotique de l'erreur d'intégration obtenue en utilisant des (t, m, s)-réseaux pour intégrer un certain type de fonction. Ce qui est différent dans notre cas, c'est que l'on considère la variance plutôt que l'erreur et comme dans le cas standard, cela nous permet d'affaiblir les conditions sur les fonctions afin d'obtenir les différentes expressions et bornes. Le chapitre se termine avec des exemples où l'on compare la variance empirique des estimateurs formés à partir de ces règles avec la méthode MC et aussi, avec les règles de réseau standard.

Finalement, au chapitre 5, nous étudions un certain type de règles de réseau standard, appelées *règles de type ν^r-copie*. Dans le livre de Sloan et Joe [116], on discute de certains avantages qu'ont ces règles par rapport aux règles de rang 1. Suivant cela, les règles qu'ils suggèrent d'utiliser et qui se trouvent dans leur annexe A sont de cette forme. Nous apportons un point de vue différent au chapitre 5 en donnant des résultats décrivant le mauvais comportement de ce type de règle, notamment en ce qui a trait à l'intégration des composantes f_I de la fonction à intégrer pour lesquelles I ne contient pas beaucoup d'indices, c.-à-d., les composantes à basse dimension.

Nous comparons également ce type de règle avec les *règles de Korobov*, qui sont les règles que nous utilisons habituellement en pratique (par exemple, dans tous les résultats numériques des chapitres précédents). Cette comparaison est faite de deux

façons : d'abord, en calculant certains critères de sélection pour ces deux types de règles, puis, en utilisant ces règles sur une fonction-test et sur un problème d'évaluation d'options en finance. Ce qui ressort de ces comparaisons, c'est qu'en utilisant des critères de sélection qui regardent de plus près les projections de la règle, on détecte facilement le mauvais comportement des règles de type copie. Notre deuxième niveau de comparaison fait ressortir les défauts de ces règles de façon encore plus claire, car on y voit que lorsqu'elles sont utilisées sur des problèmes où les composantes f_I en basse dimension sont importantes (par exemple, dans le cas de l'évaluation d'options en finance), on obtient des estimateurs ayant une variance parfois beaucoup plus grande que celle de l'estimateur MC. En comparaison, les règles de Korobov réduisent généralement la variance par rapport à l'estimateur MC.

Chapitre 2

Réduction de variance par des règles de réseau randomisées

Dans ce chapitre, nous examinons différentes façons de randomiser une règle de réseau standard et pour chacune de ces méthodes, nous donnons des résultats sur la variance théorique de l'estimateur ainsi formé et comparons cette variance avec celle de l'estimateur MC. Chaque section débute avec un résumé des résultats qui y sont présentés, mais voici un bref aperçu du contenu de chacune d'entre elles. Nous débutons à la section 2.1 avec une introduction aux règles de réseau, afin de donner tout le matériel nécessaire pour présenter nos résultats. Puis, à la section 2.2, nous étudions la variance théorique des estimateurs basés sur une règle de réseau translatée aléatoirement et la comparons avec celle de l'estimateur MC. À la section 2.3, nous étudions une méthode de randomisation consistant à permuter dans chaque dimension les coordonnées de la règle de réseau avant d'appliquer une translation aléatoire. Une troisième méthode de randomisation, qui est un cas particulier de la stratification [15], est considérée à la section 2.4. Finalement, nous donnons à la section 2.5 des résultats numériques provenant de diverses expériences afin d'illustrer comment les différentes méthodes se comparent en pratique.

2.1 Introduction aux règles de réseau

Nous expliquons d'abord ce qu'est une règle de réseau, puis donnons la définition d'un GCL et quel est le lien entre ces deux entités. Ensuite, nous présentons quelques critères couramment utilisés pour choisir les règles de réseau pour N et s donnés et faisons le lien avec le test spectral, que nous décrivons également.

2.1.1 Définition d'une règle de réseau

C'est dans les années 1950 et 1960 que les règles de réseau ont vu le jour, grâce aux travaux des mathématiciens E. Hlawka, N.M. Korobov, S.K. Zaremba et L.K. Hua. À cette époque, on les connaissait sous la forme de "bonnes" règles de réseau *(good lattice points)*. Depuis ce temps, ces règles ont été étudiées et généralisées par plusieurs chercheurs [117], [119], [120], [23], [55], [54] et [118].

Un *réseau* L est un sous-ensemble discret (infini) de $\mathbb{R}^s$ fermé sous l'addition et la soustraction. On dit qu'un ensemble de vecteurs linéairement indépendants $\{\mathbf{g}_1, \ldots, \mathbf{g}_s\}$, où $\mathbf{g}_j \in \mathbb{R}^s$ pour $j = 1, \ldots, s$, est une *base* pour le réseau si chaque point du réseau est une combinaison linéaire entière de $\mathbf{g}_1, \ldots, \mathbf{g}_s$, c.-à-d., $L = \{\sum_{j=1}^{s} c_j \mathbf{g}_j : c_1, \ldots, c_s \in \mathbb{Z}\}$. Ces vecteurs sont appelés les *générateurs* du réseau.

À chaque base d'un réseau donné, on peut associer le *parallélépipède fondamental*

$$\Lambda = \Lambda(\mathbf{g}_1, \ldots, \mathbf{g}_s) = \{\lambda_1 \mathbf{g}_1 + \ldots + \lambda_s \mathbf{g}_s : 0 \leq \lambda_i \leq 1, 1 \leq i \leq s\},$$

qui permet de partitionner $\mathbb{R}^s$ en cellules de même volume et de même orientation en formant les ensembles

$$\{\mathbf{x} + \Lambda\} = \{\mathbf{y} \in \mathbb{R}^s : \mathbf{y} = \mathbf{x} + \mathbf{z}, \text{ où } \mathbf{z} \in \Lambda\}, \qquad \text{pour } \mathbf{x} \in L.$$

Ces ensembles seront utilisés à la section 2.4 pour définir une randomisation des règles de réseau menant à une forme de stratification.

La valeur absolue du déterminant de la matrice A dont les colonnes sont données par $\mathbf{g}_1, \ldots, \mathbf{g}_s$ est égale au *déterminant du réseau* et correspond au volume du parallélépipède fondamental. Cette quantité est indépendante du choix de base et est donc unique. Les vecteurs colonnes de la transposée de A^{-1} forment une base pour le

réseau dual à L, qui est dénoté par $L^\perp$ et défini par

$$L^\perp = \{\mathbf{h} \in \mathbb{R}^s : \mathbf{h} \cdot \mathbf{x} \in \mathbb{Z},\ \forall \mathbf{x} \in L\}.$$

Le réseau dual peut être interprété de la façon suivante [116] : à partir de chaque point $\mathbf{h}$ dans $L^\perp$, si on considère l'ensemble d'hyperplans de dimension $s-1$ définis par $\{\mathbf{x} : \mathbf{h} \cdot \mathbf{x} = \delta\}$, pour $\delta = 0, \pm 1, \pm 2, \ldots$, alors on peut montrer que chaque point dans le réseau L se trouve sur un des hyperplans et chaque hyperplan contient au moins un point dans L. De plus, ces hyperplans sont parallèles et équidistants et on peut montrer que la distance entre eux est de $\|\mathbf{h}\|_2^{-1}$. La figure 2.1 illustre un exemple de réseau (à gauche) et montre son réseau dual (à droite). Remarquons la différence d'échelle entre les deux réseaux. Nous avons inclus sur le réseau primal la famille d'hyperplans générée par le point $\mathbf{h} = (1, 2)$ qui se trouve dans le réseau dual.

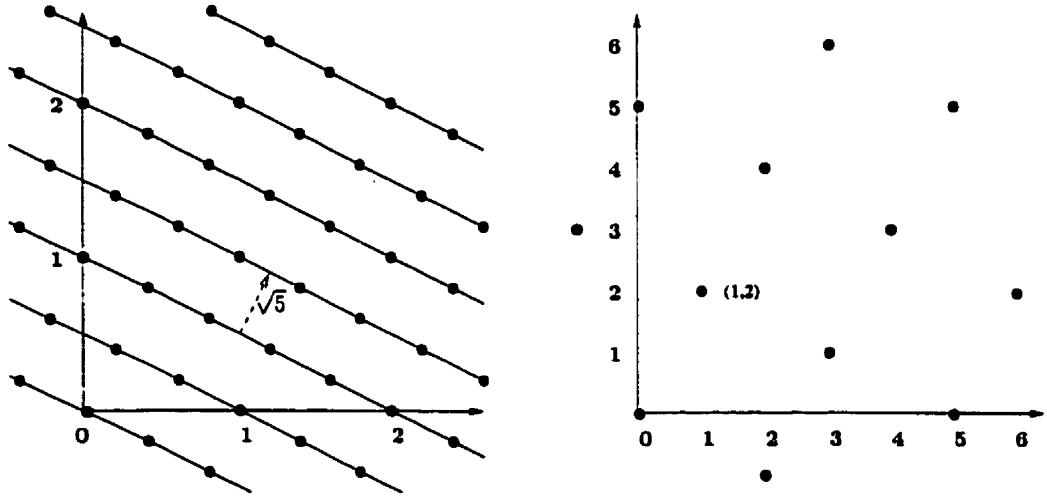


FIGURE 2.1: Un réseau avec son réseau dual

Comme nous le verrons plus loin, la notion de réseau dual est très utile afin d'étudier l'erreur et la variance des estimateurs associés aux règles de réseau. Elle permet également de définir des critères de sélection permettant de choisir ces règles.

Un *réseau d'intégration en s dimensions* est un cas particulier d'un réseau en s dimensions. La propriété supplémentaire qu'il possède est de contenir l'ensemble de vecteurs d'entiers $\mathbb{Z}^s$ comme sous-ensemble. Une *règle de réseau dans* $[0,1)^s$ est un ensemble de points P_N formé par l'intersection de $[0,1)^s$ avec un réseau d'intégration

en s dimensions. Le nombre de points N dans P_N sert à définir l'*ordre* de la règle.

Pour un réseau d'intégration, l'inverse du déterminant du réseau correspond à la *densité* du réseau d'intégration, qui est égale au nombre de points du réseau par hypercube unitaire aligné avec les axes. Par définition du réseau dual, le déterminant de $L^\perp$ est égal à l'inverse de celui du réseau L, ce qui signifie que la "densité" * du réseau dual est de $1/N$. À titre d'illustration, on peut voir à la figure 2.1 de la page 25 que le réseau à gauche donne lieu à une règle de réseau contenant cinq points et on a bien une "densité" de $1/5$ dans le réseau dual, qui est à droite (par exemple, chaque carré semi-ouvert de taille cinq par cinq contient cinq points).

Plutôt que de prendre la base engendrant L pour représenter une règle de réseau définie par $P_N = L \cap [0,1)^s$, on peut utiliser la forme suivante [118] :

$$P_N = \bigcup_{j_1=0}^{n_1-1} \cdots \bigcup_{j_r=0}^{n_r-1} \left\{ \left(j_1 \frac{\mathbf{z}_1}{n_1} + \ldots + j_r \frac{\mathbf{z}_r}{n_r} \right) \bmod 1 \right\}, \tag{2.1}$$

où le modulo 1 est effectué composante par composante. Il a été démontré dans [118] que toute règle de réseau pouvait être représentée sous cette forme non-répétitive. Les $\mathbf{z}_1, \ldots, \mathbf{z}_r \in \mathbb{Z}^s$ sont les *vecteurs générateurs de la règle*. Les entiers $n_1, \ldots, n_r > 1$ utilisés dans cette représentation sont appelés les *invariants* et leur produit est égal à N, l'ordre de la règle. Les invariants respectent la propriété que n_{i+1} divise n_i pour $i = 1, \ldots, r-1$. Le nombre $r \leq s$ est appelé le *rang* et correspond au nombre minimal de vecteurs $\mathbf{z}$ nécessaires pour générer les N points de la règle. Le rang et les invariants sont déterminés de façon unique par la règle. Cependant, les vecteurs $\mathbf{z}_1, \ldots, \mathbf{z}_r$ eux ne sont pas uniques.

En paramétrisant les règles de cette façon, cela permet de définir de façon plus compacte l'espace sur lequel les recherches pour trouver les meilleures règles par rapport à un critère de sélection donné sont définies. Nous verrons dès la prochaine sous-section que pour les règles que nous avons considérées dans nos expériences numériques, cette paramétrisation est particulièrement simple.

*Nous mettons densité entre guillemets quand on parle du réseau dual, car ce dernier n'est pas un réseau d'intégration.

2.1.2 Types de règles de réseau considérés

Dans cet ouvrage, nous porterons notre attention sur deux cas particuliers de règles de réseau : les *règles de Korobov* [61] et les *règles ν^r-copie* [23, 55]. Une règle de Korobov est une règle de rang 1, dont le vecteur générateur $\mathbf{z}$ est de la forme

$$\mathbf{z} = (1, a, \ldots, a^{s-1}) \bmod N,$$

où le modulo est appliqué à chaque composante. Ainsi, une telle règle est formée par les N points

$$\mathbf{x}_i = \left(\frac{i-1}{N}\mathbf{z}\right) \bmod 1, \qquad i = 1, \ldots, N. \tag{2.2}$$

L'ensemble de points illustré à la figure 1.1 est en fait une règle de Korobov avec $N = 251$ et $a = 33$. Les règles de Korobov ont des avantages par rapport aux autres types de règles (de rang 1 ou supérieur) qui seront discutés au début du chapitre 3.

On peut tout de suite voir que si l'on veut faire une recherche afin de trouver la meilleure règle de Korobov par rapport à un critère de sélection donné et pour un nombre de points N fixé, alors il est suffisant de chercher sur tous les entiers a entre 1 et $N-1$ afin de trouver la règle qui optimise le critère en question.

Nous définissons dès maintenant ce qu'est une règle de type copie, même si ce sujet sera mis de côté jusqu'au chapitre 5. De cette façon, le lecteur aura un exemple de ce qui existe comme alternative aux règles de Korobov. Notre définition diffère de celle donnée dans [116], car nous englobons le cas où le rang r est égal à s avec celui où $r < s$ (correspondant aux chapitres 6 et 7 de [116], respectivement). Aussi, nous supposons dès le départ que la règle copiée est de rang 1.

Définition 2.1.1 *Une* règle ν^r-copie *est une règle de rang r que l'on obtient en partionnant chacun des r premiers axes de $[0,1)^s$ en ν parties égales, obtenant ainsi ν^r rectangles de même volume, puis en "copiant" une règle de rang 1 contenant n points dans chacun de ces rectangles, en faisant une mise à l'échelle appropriée des r premières coordonnées, afin que $[0,1)^s$ soit projeté sur $[0,1/\nu)^r \times [0,1)^{s-r}$.*

Plus précisément, les $n\nu^r$ points d'une telle règle sont donnés par l'ensemble

$$\bigcup_{m_1=0}^{\nu-1} \cdots \bigcup_{m_r=0}^{\nu-1} \bigcup_{i=1}^{n} \left\{ \left[\left(\frac{m_1}{\nu}, \ldots, \frac{m_r}{\nu}, \underbrace{0, \ldots, 0}_{s-r \text{ fois}} \right) + \mathbf{x}_i \right] \bmod 1 \right\}, \tag{2.3}$$

où $\{\mathbf{x}_i,\ i = 1, \ldots, n\}$ est la règle de rang 1 que l'on a copiée, $\nu \geq 2$ est un entier et $\text{pgcd}(n, \nu) = 1$, où le symbole "pgcd" signifie "plus grand commun diviseur". Lorsque la règle de rang 1 est de type Korobov, cet ensemble est équivalent aux points utilisés à l'équation (7.5) dans [116]. La figure 2.2 donne une exemple d'une règle de type 2^2-copie, avec $n = 5$ pour la règle de rang 1 qui est copiée. Puisque le rang $r = 2$, on partitionne le carré $[0, 1)^2$ en $\nu^r = 2^2$ carrés de côté $1/\nu = 1/2$ et on copie la même règle d'ordre $n = 5$ dans chacun de ces quatre carrés, obtenant ainsi un total de 20 points.

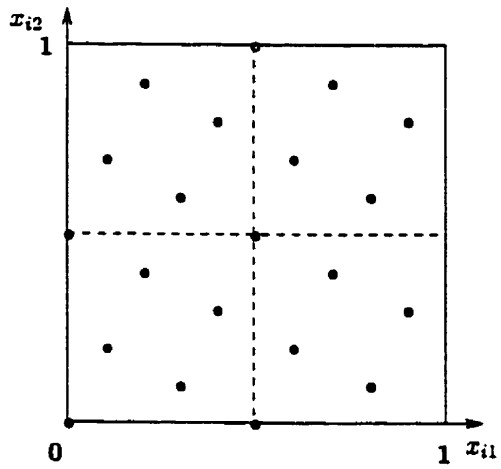


FIGURE 2.2: Règle 2^2-copie

Les règles de type ν^r-copie peuvent servir à construire des suites de règles imbriquées, en commençant avec la règle de rang 1, puis en passant successivement aux rangs supérieurs, jusqu'à $r = s$ [55]. Ainsi, chaque règle de la suite contient ν fois plus de points que la règle précédente et contient tous les points de la règle précédente. L'avantage d'avoir une telle suite de règles est que si l'on constate que l'erreur d'intégration donnée par la première règle de la suite est trop grande, la règle suivante nous permet d'ajouter des points à ceux déjà utilisés afin de tenter de diminuer cette erreur et ainsi de suite. Une telle suite peut également servir à estimer l'erreur d'intégration, tel que mentionné dans [55, 116].

2.1.3 Règle projection-régulière

On dit qu'une règle de réseau est *projection-régulière* si $P_N(\{1,\ldots,t\})$, la projection de P_N sur les t premières dimensions de $[0,1)^s$, est d'ordre $n_1 \cdots n_t$ [116], pour $t = 1,\ldots,s$. Dans [88], on dit qu'une règle de réseau est *complètement projection-régulière* si $P_N(I)$ possède $n_1,\ldots,n_{|I|}$ comme invariants, pour tout $I \subseteq S$ non vide, où $|I|$ représente la cardinalité de I et $S = \{1,\ldots,s\}$.

Dans le cas où la règle P_N est de rang 1, ces définitions peuvent être énoncées en fonction de la densité du réseau L associé [85], c.-à-d., le réseau L tel que $P_N = L \cap [0,1)^s$. En effet, une règle de rang 1 est projection-régulière si la densité de $L_{\{1,\ldots,t\}}$, la projection du réseau L sur les t premières dimensions de $\mathbb{R}^s$, est égale à celle de L pour tout $1 \leq t \leq s$ et elle est complètement projection-régulière si la densité de L_I est égale à celle de L pour tout $I \subseteq S$ non vide. Dans le second cas, cela revient à dire que chaque projection $P_N(I)$ pour $I \subseteq S$ non vide contient N points, alors que dans le premier, on n'exige seulement que cette propriété tienne pour les projections sur des sous-ensembles I contenant des dimensions successives, c.-à-d., de la forme $I = \{1,\ldots,t\}$, pour $1 \leq t \leq s$.

Nous supposons à plusieurs reprises dans cet ouvrage que les règles de réseau utilisées sont de rang 1 et complètement projection-régulières, car cela nous assure entre autres que toutes les projections unidimensionnelles sont données par $\{0, 1/N, \ldots, (N-1)/N\}$. Notons que dans le cas des règles de rang supérieur à 1, les projections $P_N(I)$ ne contiennent pas nécessairement N points distincts même si la règle P_N est complètement projection-régulière. En effet, le nombre de points dans $P_N(I)$ ne peut dépasser le produit $n_1 \cdots n_{|I|}$ des invariants et ce produit vaut N seulement si $|I| \geq r$.

2.1.4 Critères de sélection traditionnellement utilisés

La plupart des critères de sélection qui sont utilisés pour choisir des règles de réseau pour une paire (N, s) donnée et un rang r prédéterminé (c.-à-d., pour choisir les vecteurs générateurs $\mathbf{z}_1,\ldots,\mathbf{z}_r$ engendrant la règle, tels que donnés à l'équation (2.1)) sont basés sur la borne donnée en (1.6). L'idée est de choisir une définition

précise de discrépance $D(P_N)$ (rappelons que nous avons parlé de ce concept dans l'introduction en disant que $D(P_N)$ mesurait la non-uniformité de P_N, précisant qu'il y avait plusieurs façons de mesurer cette non-uniformité), de déterminer l'ensemble de fonctions E pour lequel l'inégalité (1.6) tient (ou vice-versa), puis de faire une recherche exhaustive parmi un ensemble de règles ayant une structure précise (par exemple, les règles de Korobov, si $r = 1$), afin de trouver celle qui minimise $D(P_N)$.

La règle ainsi choisie est donc celle qui minimise la *borne* sur l'erreur d'intégration pour les fonctions dans E, parmi l'ensemble de règles considérées. Autrement dit, on minimise l'erreur en pire cas. Comme alternative à ces recherches, Langtry a proposé dans [63] des algorithmes permettant de construire des règles de réseau qui, approximativement, devraient minimiser l'index de Babenko-Zaremba (que nous définirons plus loin, à l'équation (2.11)). En bout de ligne, il n'est pas clair que les règles qu'il obtient ont de meilleures propriétés que celles obtenues à l'aide d'une recherche. L'avantage de sa méthode est qu'elle est plus rapide. Nous préférons rester du côté des recherches exhaustives, mais en prenant soin d'utiliser des critères qui sont assez rapides à calculer.

Habituellement, on définit E comme un ensemble de fonctions dont les coefficients de Fourier "décroissent rapidement", car les règles de réseau ont une structure qui fait en sorte que ce type de fonction est bien intégré, c.-à-d., avec une petite erreur. En effet, en utilisant la représentation en série de Fourier de f,

$$f(\mathbf{x}) = \sum_{\mathbf{h}\in\mathbf{Z}^s} \hat{f}(\mathbf{h}) e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}}, \text{ pour tout } \mathbf{x} \in [0,1)^s,$$

où

$$\hat{f}(\mathbf{h}) = \int_{[0,1)^s} f(\mathbf{x}) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}} d\mathbf{x} \tag{2.4}$$

est le coefficient de Fourier de f évalué en $\mathbf{h}$, on a que

$$Q_N - \mu = \frac{1}{N}\sum_{i=1}^{N}\sum_{\mathbf{h}\in\mathbf{Z}^s} \hat{f}(\mathbf{h}) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i} - \hat{f}(\mathbf{0}). \tag{2.5}$$

Si on suppose que la représentation en série de Fourier de f est absolument convergente, c.-à-d., que $\sum_{\mathbf{h}\in\mathbf{Z}^s} |\hat{f}(\mathbf{h})| < \infty$, les conditions du théorème de Fubini [112] sont respectées et donc, on peut interchanger l'ordre de sommation dans (2.5) et par la

suite, utiliser le fait que pour une règle de réseau, on a [116, *Lemma 2.7*]

$$\sum_{i=1}^{N} e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i} = \begin{cases} N & \text{si } \mathbf{h} \in L^\perp \\ 0 & \text{sinon.} \end{cases} \tag{2.6}$$

On obtient ainsi que [116, *Theorem 2.8*]

$$Q_N - \mu = \sum_{\mathbf{0}\neq\mathbf{h}\in L^\perp} \hat{f}(\mathbf{h}). \tag{2.7}$$

Si on suppose que les fonctions que l'on intègre sont assez régulières et donc, que leurs coefficients de Fourier décroissent avec $\mathbf{h}$, ceci suggère d'utiliser un critère de sélection de la forme [47]

$$D_w(P_N) = \sum_{\mathbf{0}\neq\mathbf{h}\in L^\perp} w(\mathbf{h}) \qquad \text{ou} \qquad D'_w(P_N) = \sup_{\mathbf{0}\neq\mathbf{h}\in L^\perp} w(\mathbf{h}), \tag{2.8}$$

où les $w(\mathbf{h})$ sont des poids qui devraient décroître avec $\|\mathbf{h}\|$ (pour une norme quelconque $\|\cdot\|$) en fonction de nos hypothèses sur le comportement des coefficients de Fourier de la fonction à intégrer. Plus précisément, en choisissant des poids $w(\mathbf{h})$ qui imitent relativement bien le comportement des $\hat{f}(\mathbf{h})$, on peut voir la mesure $D_w(P_N)$ comme une approximation de l'erreur (mais qui est plus facile à calculer) et donc, la règle P_N minimisant $D_w(P_N)$ devrait avoir une erreur associée qui soit petite. La mesure $D'_w(P_N)$ est une approximation encore plus succinte de l'erreur que ne l'est $D_w(P_N)$ et qui peut s'avérer plus rapide à calculer dans certains cas. La plupart des critères de sélection discutés dans cet ouvrage sont de la forme (2.8).

Une définition possible pour l'ensemble E est de prendre

$$E = \left\{ f : [0,1)^s \to \mathbb{R} : |\hat{f}(\mathbf{h})| \leq \frac{c}{\|\mathbf{h}\|^\alpha},\ \forall \mathbf{h} \in \mathbb{Z}^s \right\},$$

où $\|\cdot\|$ est une certaine norme, $c > 0$ et α est un entier quantifiant le degré de régularité des fonctions dans cet ensemble.

Dans le cas où on choisit de prendre la norme produit $\|\mathbf{h}\|_\pi = \prod_{j=1}^s \max(1, |h_j|)$ et $\alpha > 1$, alors E est dénoté par $E_\alpha(c)$ dans [116] et la discrépance $D_w(P_N)$ associée est donnée par

$$P_\alpha^s = \sum_{\mathbf{0}\neq\mathbf{h}\in L^\perp} \|\mathbf{h}\|_\pi^{-\alpha}. \tag{2.9}$$

Cela revient à utiliser les poids

$$w(\mathbf{h}) = \|\mathbf{h}\|_\pi^{-\alpha}$$

dans la définition (2.8). Si la fonction a une représentation en série de Fourier qui est absolument convergente, alors P_α^s et $E_\alpha(c)$ sont respectivement les quantités $D(P_N)$ et E dans l'inégalité (1.6), c est une borne sur $V(f)$ et le produit cP_α^s nous fournit donc une borne sur l'erreur d'intégration pour toutes les fonctions dans $E_\alpha(c)$.

De plus, si α est un entier pair, on peut montrer que P_α^s est égal à l'erreur d'intégration d'une fonction bien précise qui est dans $E_\alpha(c)$ et donc, l'expression (2.9) peut être écrite comme une somme sur les points de la règle plutôt que sur le réseau dual $L^\perp$ et on obtient alors

$$P_\alpha^s = -1 + \frac{1}{N} \sum_{\mathbf{z} \in P_N} \prod_{j=1}^{s} \left[1 - \frac{(-1)^{\alpha/2} (2\pi)^\alpha}{(\alpha)!} B_\alpha(z_j) \right], \tag{2.10}$$

où $B_\alpha(\cdot)$ est le polynôme de Bernoulli de degré α [116]. Sous cette forme, P_α^s se calcule en un temps qui est dans $O(Ns)$.

Ce critère a été souvent utilisé (habituellement avec $\alpha = 2$) pour faire des recherches qui ont permis de construire des tableaux de règles minimisant P_α^s, mais dans des dimensions dépassant rarement $s = 12$ [38, 120, 55, 54, 116]. Dans [46], plusieurs généralisations de ce P_α^s sont présentées, toujours dans le contexte de l'inégalité (1.6).

Un autre critère a été assez souvent utilisé et étudié pour choisir des règles de réseau et comme pour le critère précédent, il est basé sur la borne sur l'erreur d'intégration des fonctions dans $E_\alpha(c)$, sauf qu'on utilise $D'_w(P_N)$ au lieu de $D_w(P_N)$ dans (2.8), c.-à-d., on tente plutôt de minimiser le terme dominant de la borne $\sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} \|\mathbf{h}\|_\pi^{-\alpha}$. Ceci donne lieu à l'*index de Babenko-Zaremba*

$$\rho = \min_{\mathbf{0} \neq \mathbf{h} \in L^\perp} \|\mathbf{h}\|_\pi, \tag{2.11}$$

que l'on veut maximiser. L'avantage de ce critère sur P_α^s est qu'il ne dépend pas de α, puisque le terme dominant dans $\sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} \|\mathbf{h}\|_\pi^{-\alpha}$ est le même pour toute valeur de $\alpha \geq 1$. De plus, tout comme pour P_α^s, on peut trouver un ensemble de fonctions E_ρ tel que ρ^{-1} correspond à la discrépance $D(P_N)$ dans la borne (1.6) pour les fonctions

dans E_ρ. L'inconvénient est que ρ est beaucoup plus difficile à calculer que P_α^s et donc, il devient très coûteux de faire des recherches pour trouver la règle qui maximise ρ parmi un certain ensemble de règles. On l'utilise dans [89] pour calculer des tableaux allant jusqu'en dimension 10.

Les critères P_2^s et ρ sont des cas particuliers de (2.8), mais ils peuvent aussi être englobés dans une structure plus générale qui est décrite dans [42]. On explique dans cet article que le critère P_2^s correspond à la définition de la *diaphonie classique* [137] appliquée aux règles de réseau et que ρ est une version du test spectral qui utilise la norme produit plutôt que la norme euclidienne. Nous reviendrons sur le lien entre ρ et le test spectral à la section 2.1.7.

2.1.5 Randomisation par translation aléatoire

Pour randomiser P_N, une méthode à la fois simple et naturelle est de faire une translation aléatoire modulo 1, tel que proposé par Cranley et Patterson dans [20]. Plus précisément, cette méthode consiste à générer un vecteur $\mathbf{u}$ uniformément dans $[0,1)^s$, puis à l'utiliser pour translater aléatoirement (modulo 1 dans chaque dimension) P_N, obtenant ainsi $\tilde{P}_N = \{(\mathbf{x}_i + \mathbf{u}) \bmod 1, i = 1, \ldots, N\}$. Chaque point de $\tilde{P}_N$ suit donc la loi uniforme sur $[0,1)^s$ et

$$\hat{\mu}_{\mathrm{LR}} = \frac{1}{N} \sum_{i=1}^{N} f\left((\mathbf{x}_i + \mathbf{u}) \bmod 1\right)$$

est un estimateur sans biais de μ. Si on génère m vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_m$ i.i.d. uniformément distribués dans $[0,1)^s$, on peut alors estimer la variance de $\hat{\mu}_{\mathrm{LR}}$ à l'aide de la formule (1.7), en prenant $\tilde{Q}_{N,j}$ égal à l'estimateur $\hat{\mu}_{\mathrm{LR}}$ qui utilise le j^e vecteur $\mathbf{u}_j$. L'estimateur de la variance ainsi obtenu est sans biais, tel que démontré dans [80].

2.1.6 Lien avec les générateurs à congruence linéaire

Les générateurs à congruence linéaire (GCL) (voir [71], par exemple) sont définis par une récurrence

$$\xi_n = a\,\xi_{n-1} \bmod N.$$

Les nombres entre 0 et 1 sont produits par la fonction de sortie

$$u_n = \xi_n / N.$$

Le *modulo* N est un entier positif, le *multiplicateur* a est dans $\{1, \ldots, N-1\}$ et $\xi_0 \in \{0, \ldots, N-1\}$ est le *germe*.

Il y a deux propriétés des GCL qui nous intéressent plus particulièrement. D'abord, si N est un nombre premier et que a est un élément primitif modulo N (c.-à-d., le plus petit entier k tel que $a^k = 1 \bmod N$ est $k = N-1$), alors le GCL défini par N et a atteint la période maximale de $N-1$, en autant que ξ_0 soit différent de 0. Ensuite, pour tout $p > 0$, l'ensemble formé des p-tuplets successifs de la suite $\{u_0, u_1, \ldots\}$ qui se chevauchent, et à partir de tous les germes initiaux possibles, a une structure de réseau : c.-à-d., si on pose

$$\Psi_p = \{(u_0, u_1, \ldots, u_{p-1}) : \xi_0 \in \mathbb{Z}_N\},$$

alors on a que

$$\Psi_p = L \cap [0,1)^p, \tag{2.12}$$

pour un certain réseau L dans $[0,1)^p$. Ainsi, les points contenus dans Ψ_p correspondent aux points d'une règle de réseau d'ordre N en p dimensions. Remarquons que la structure de Ψ_p correspond à un cas particulier de ce qui a été défini en (1.8), avec $s = p$, $g(\xi) = \xi/N$, $\tau(\xi) = a\xi \bmod N$ et $\Xi = \mathbb{Z}_N$. De plus, si les cycles définis par le GCL sont purement périodiques, alors Ψ_s correspond à l'ensemble de points P_N obtenu à partir d'une règle de Korobov ayant $((1, a, \ldots, a^{s-1}) \bmod N)$ comme vecteur générateur. Finalement, pour un GCL à période maximale, l'ensemble Ψ_s peut être construit très facilement, car il suffit d'ajouter le point $\mathbf{0}$ à l'ensemble des $N-1$ s-tuplets successifs que l'on obtient en énumérant les u_i produits par le GCL, comme nous l'avons fait à l'exemple 1.1.3 de la page 11.

2.1.7 Test spectral

Le fait que Ψ_p ait une structure de réseau peut sembler mauvais du point de vue de l'utilisation d'un GCL comme générateur pseudo-aléatoire, car cela signifie que les

points dans Ψ_p sont situés sur des hyperplans parallèles et équidistants. Cependant, si la distance entre eux est petite, cela signifie que les points dans Ψ_p sont assez bien distribués dans $[0,1)^p$. De plus, lorsque le GCL sert comme générateur pseudo-aléatoire, seule une petite fraction de la période est utilisée. Cela veut dire qu'on regarde seulement un échantillon de points dans Ψ_p.

En fait, la structure de réseau de Ψ_p permet de définir le *test spectral*, un des tests théoriques très couramment utilisés pour juger la qualité d'un GCL [29, 78, 75] et qui consiste à calculer la distance d_p entre ces hyperplans. Puisqu'il existe des techniques efficaces pour calculer d_p et des bornes inférieures absolues d_p^* sur d_p (obtenues en considérant toutes les formes de réseau possibles et non pas seulement les réseaux d'intégration ; voir [75] et les références qui s'y trouvent pour plus de détails), on peut définir des figures de mérite de la forme :

$$M_T = \min_{1 \le p \le T} d_p^*/d_p, \tag{2.13}$$

que l'on veut près de 1. Cette figure de mérite mesure donc l'uniformité de toutes les projections $P_N(I)$ de P_N sur des sous-ensembles I de la forme $I = \{1, \ldots, p\}$, pour $1 \le p \le T$. Le critère M_T a été utilisé dans [75, 29] pour construire des tableaux de bons GCL, ainsi que d'autres types de générateurs pseudo-aléatoires basés sur des congruences linéaires.

La raison pour laquelle mesurer d_p est appelé le test spectral est que l'on peut montrer [21, 59] que d_p^{-1}, dénoté par l_p, est égal à la longueur du plus court vecteur dans $L^\perp$, c.-à-d.,

$$l_p = \min_{\mathbf{0} \ne \mathbf{h} \in L^\perp} \|\mathbf{h}\|_2, \tag{2.14}$$

où L est le réseau utilisé dans la représentation (2.12). En fait, historiquement, le test spectral a été proposé par Coveyou et MacPherson [19] comme étant une mesure de l'éloignement entre la fonction de densité empirique associée à un GCL et celle de la loi uniforme. Le moyen utilisé pour mesurer cet éloignement était de faire une analyse en série de Fourier (analyse spectrale) de ces deux fonctions. Pour plus de détails sur le test spectral, voir [19, 58, 73].

Si on se replace dans le contexte de l'intégration numérique sur $[0,1)^s$ et que l'on regarde l'égalité donnée en (2.14), on se rend compte que l_s et le critère ρ donné en

(2.11) mesurent tous les deux la longueur du plus court vecteur dans $L^\perp$, mais en utilisant une norme différente. Ainsi, si on remplace la norme produit dans (2.9) par la norme euclidienne $\|\cdot\|_2$, la minimisation de d_s correspond à minimiser le terme dominant de cette nouvelle borne d'intégration et tout comme c'est le cas pour ρ, on peut trouver un ensemble de fonctions E_{d_s} tel que d_s correspond à la discrépance dans l'inégalité (1.6). On peut donc imaginer que l'on pourrait utiliser d_s ou d'autres critères plus généraux qui lui sont reliés, comme M_T, afin de choisir des règles de réseau et c'est ce qui est proposé dans [27]. Des inégalités reliant ρ et d_s sont également données dans cet article. Dans [82, 84, 83], nous utilisons les meilleurs GCL par rapport à M_8 comme méthode QMC sur des problèmes de finance et de théorie du risque. De plus, le critère de sélection que nous allons présenter au chapitre 3 est basé sur le test spectral.

2.2 Randomisation par translation aléatoire

Comme nous l'avons expliqué à la section 2.1.5, cette méthode consiste à translater aléatoirement et uniformément, modulo 1 dans chaque dimension, une règle de réseau. Ainsi, en répétant un certain nombre de fois, on peut estimer l'erreur d'intégration en utilisant le théorème de la limite centrale, ou estimer la variance de l'estimateur qui utilise les points d'un tel réseau translaté.

Si on définit l'*efficacité* des estimateurs comme étant l'inverse du produit du coût de calcul et de l'erreur quadratique moyenne de l'estimateur [70], cette méthode est gagnante par rapport à MC si sa variance est plus petite, car les deux sont sans biais et le coût de calcul de l'estimateur MC est supérieur. Des comparaisons du temps de calcul seront données à la section 2.5. Nous verrons aussi à cette section que pour différents problèmes, la variance empirique des estimateurs basés sur les règles de réseau translatées aléatoirement est inférieure à celle de l'estimateur MC.

Ce que nous voulons faire dans la présente section, c'est de donner des résultats théoriques comparant la variance de l'estimateur basé sur une règle de réseau translatée aléatoirement avec celle de l'estimateur MC, afin de voir si la réduction de variance par rapport à MC peut être garantie dans certains cas. Pour cela, nous donnons d'abord à la sous-section 2.2.1 une expression pour la variance de $\hat{\mu}_{\mathrm{LR}}$ en fonction des coefficients

de Fourier de f ; à la sous-section 2.2.2, nous donnons des bornes sur cette variance faisant intervenir le critère P_α^s ; à la sous-section 2.2.3, nous expliquons pourquoi il est difficile de garantir la réduction de variance par rapport à MC, mais terminons sur une note positive en donnant un résultat nous indiquant qu'"en moyenne" (sur un ensemble de règles de rang 1), on ne devrait pas faire bien pire que la méthode MC. Ensuite, nous voyons à la sous-section 2.2.4 que les résultats de réduction de variance existant pour les méthodes des variables antithétiques et de l'échantillonnage de l'hypercube latin dans le cas des fonctions monotones ne peuvent être étendus au cas des règles de réseau translatées aléatoirement en donnant un contre-exemple, puis concluons la section en discutant à la sous-section 2.2.5 que les résultats obtenus en pratique sont plus encourageants que ce que la théorie prédit.

2.2.1 Variance de l'estimateur obtenu par translation aléatoire

Dans ce qui suit, nous considérons l'estimateur $\hat{\mu}_{\mathrm{LR}}$ obtenu en utilisant cette méthode, qui est donné par

$$\hat{\mu}_{\mathrm{LR}} = \frac{1}{N}\sum_{i=1}^{N} f((\mathbf{x}_i + \mathbf{u}) \bmod 1) \tag{2.15}$$

et montrons que sa variance peut être exprimée en fonction des coefficients de Fourier de la fonction f.

Proposition 2.2.1 *Soit $f \in \mathcal{L}^2$. Alors on a que*

$$\mathrm{E}(\hat{\mu}_{\mathrm{LR}}) = \mu$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sum_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}^\perp} |\hat{f}(\mathbf{h})|^2, \tag{2.16}$$

où $\hat{f}(\mathbf{h})$ est le coefficient de Fourier de f évalué en $\mathbf{h}$, tel que donné à l'équation (2.4).

Démonstration : d'abord, on a que

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_{\mathrm{LR}}) &= \int_{[0,1)^s} \frac{1}{N}\sum_{i=1}^{N} f((\mathbf{x}_i + \mathbf{u}) \bmod 1) d\mathbf{u} \\
&= \frac{1}{N}\sum_{i=1}^{N} \int_{[0,1)^s} f((\mathbf{x}_i + \mathbf{u}) \bmod 1) d\mathbf{u},
\end{aligned}$$

où le théorème de Fubini (voir [112], par exemple) nous permet de changer l'ordre de la somme et de l'intégrale, puisque $f \in \mathcal{L}^2$ implique que $f \in \mathcal{L}^1$, c.-à-d., $\int_{[0,1)^s} f^2(\mathbf{x})d\mathbf{x} < \infty$ implique que $\int_{[0,1)^s} |f(\mathbf{x})|d\mathbf{x} < \infty$. Or, pour tout $i = 1, \ldots, N$, puisque $\mathbf{x}_i$ est fixé, on a que $\mathbf{y} = (\mathbf{x}_i + \mathbf{u}) \bmod 1$ est uniformément distribué sur $[0,1)^s$ et donc,

$$\mathrm{E}(\hat{\mu}_{\mathrm{LR}}) = \int_{[0,1)^s} f((\mathbf{x}_i + \mathbf{u}) \bmod 1)d\mathbf{u} = \int_{[0,1)^s} f(\mathbf{y})d\mathbf{y} = \mu. \tag{2.17}$$

Ensuite, on définit la fonction $g(\cdot) : [0,1)^s \to \mathbb{R}$ telle que $g(\mathbf{u}) = \sum_{i=1}^N f((\mathbf{x}_i + \mathbf{u}) \bmod 1)/N$. Ainsi, g est dans $\mathcal{L}^2$, $\mathrm{Var}(g(\mathbf{u})) = \mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ et par l'égalité de Parseval [112], on a

$$\mathrm{Var}(g(\mathbf{u})) = \sum_{\mathbf{0} \neq \mathbf{h} \in \mathbf{Z}^s} |\hat{g}(\mathbf{h})|^2. \tag{2.18}$$

Il faut maintenant calculer $\hat{g}(\mathbf{h})$:

$$\begin{aligned}
\hat{g}(\mathbf{h}) &= \int_{[0,1)^s} g(\mathbf{u}) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{u}} d\mathbf{u} \\
&= \int_{[0,1)^s} \left(\frac{1}{N} \sum_{i=1}^N f((\mathbf{x}_i + \mathbf{u}) \bmod 1) \right) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{u}} d\mathbf{u} \\
&= \frac{1}{N} \sum_{i=1}^N \int_{[0,1)^s} f((\mathbf{x}_i + \mathbf{u}) \bmod 1) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{u}} d\mathbf{u} \\
&= \frac{1}{N} \sum_{i=1}^N \int_{[0,1)^s} f(\mathbf{v}_i) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot(\mathbf{v}_i - \mathbf{x}_i)} d\mathbf{v}_i \\
&= \frac{1}{N} \sum_{i=1}^N e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i} \int_{[0,1)^s} f(\mathbf{v}_i) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{v}_i} d\mathbf{v}_i \\
&= \frac{1}{N} \sum_{i=1}^N e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i} \hat{f}(\mathbf{h}) \\
&= \begin{cases} \hat{f}(\mathbf{h}) & \text{si } \mathbf{h} \in L^\perp, \\ 0 & \text{sinon.} \end{cases}
\end{aligned}$$

Dans la série d'équations précédentes, la troisième égalité est obtenue en interchangeant la somme et l'intégrale et ce changement d'ordre est légal par le théorème de Fubini, puisque f est dans $\mathcal{L}^2$; la quatrième égalité est obtenue en appliquant le changement de variable $\mathbf{v}_i = (\mathbf{x}_i + \mathbf{u}) \bmod 1$; la dernière égalité est obtenue en appliquant (2.6).

En remplaçant dans (2.18), on obtient bien que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} |\hat{f}(\mathbf{h})|^2$. ∎

Remarque 2.2.1 *Une démonstration que $\hat{\mu}_{\mathrm{LR}}$ est un estimateur sans biais de μ se trouve dans [116, Theorem 4.11], mais elle requiert que f ait une représentation en série de Fourier qui soit absolument convergente alors que la nôtre ne requiert en fait que f soit dans $\mathcal{L}^1$ (pour que l'espérance existe). De même, il a été démontré (indépendamment de nos résultats) dans [130] que* $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sum_{\mathbf{0}\neq\mathbf{h}\in L^\perp} |\hat{f}(\mathbf{h})|^2$, *mais là aussi, on suppose que f a une représentation en série de Fourier qui est absolument convergente.*

Ainsi, la variance de $\hat{\mu}_{\mathrm{LR}}$ est donnée par une expression qui ressemble beaucoup à celle donnant l'erreur d'intégration (2.7) lorsque la règle de réseau P_N (non-translatée) est utilisée pour approximer μ. La différence est que pour la variance, on somme les normes au carré des coefficients $\hat{f}(\mathbf{h})$, alors que pour l'erreur, on les somme directement. Dans un cas comme dans l'autre, cela signifie que pour une fonction f donnée, la règle P_N permet d'approximer μ avec une petite erreur si les coefficients de Fourier sont petits lorsqu'évalués aux points $\mathbf{h}$ faisant partie du réseau dual $L^\perp$. En particulier, si f a des coefficients de Fourier qui décroissent quand $\mathbf{h}$ s'éloigne de l'origine, on voudra que $L^\perp$ contienne le moins de points possible près de l'origine. Remarquons que c'est justement ce que les critères de sélection $\rho = \min_{\mathbf{0}\neq\mathbf{h}\in L^\perp} \|\mathbf{h}\|_\pi$ et $d_s^{-1} = \min_{\mathbf{0}\neq\mathbf{h}\in L^\perp} \|\mathbf{h}\|_2$ regardent, c.-à-d., ils cherchent à maximiser la longueur du plus court vecteur dans $L^\perp$, en utilisant chacun leur propre définition de norme pour mesurer les $\mathbf{h}$.

On doit cependant signaler que contrairement à la condition requise pour que l'erreur d'intégration soit donnée par (2.7), soit que f ait une représentation en série de Fourier qui soit absolument convergente, la condition pour que la variance de $\hat{\mu}_{\mathrm{LR}}$ soit donnée par (2.16) est simplement que f soit dans $\mathcal{L}^2$. Est-ce que cela veut dire que le fait de translater aléatoirement une règle de réseau permet d'intégrer une plus grande variété de fonctions que si on ne la translate pas? Non, cela veut simplement dire que moins de conditions sont nécessaires pour analyser la variance que ce qui est requis pour analyser l'erreur déterministe. Cette observation est importante, car cela signifie que l'expression pour la variance tient pour la plupart des fonctions que l'on peut rencontrer en simulation, alors que celle pour l'erreur ne tient pour pratiquement aucune de ces fonctions. En effet, f a une représentation en série de Fourier absolument

convergente seulement si la continuation périodique de f, définie par $\bar{f} : \mathbf{R}^s \to \mathbf{R}$, avec $\bar{f}(\mathbf{y}) = f(\mathbf{y} \bmod 1)$, est continue, ce qui requiert que $\lim_{x_j \to 1, j \in I} f(\mathbf{x}) = f(\mathbf{0})$ pour n'importe quel sous-ensemble I de S, une condition qui n'est pratiquement jamais respectée en pratique (à moins d'effectuer une transformation sur f de façon à ce qu'elle respecte cela, comme on l'explique dans [116, section 2.12]).

2.2.2 Bornes sur la variance

Par la même procédure que celle utilisée pour borner l'erreur d'intégration (voir les équations (2.7) et (2.9)), on peut imposer des conditions sur la vitesse de décroissance des coefficients de Fourier de f afin de borner la variance de $\hat{\mu}_{\mathrm{LR}}$ donnée à la proposition 2.2.1. Nous supposons dans ce qui suit que l'ensemble $E_\alpha(c) = \{f : [0,1)^s \to \mathbf{R} : |\hat{f}(\mathbf{h})| \leq c\|\mathbf{h}\|_\pi^{-\alpha}$, pour tout $\mathbf{h} \in \mathbf{Z}^s\}$ introduit dans [116] peut être défini avec $\alpha \geq 1$.

Corollaire 2.2.1 *Si $f \in \mathcal{L}^2$ est telle que $f \in E_\alpha(c)$ pour un certain $\alpha \geq 1$ entier et une constante $c > 0$, alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq c^2 P_{2\alpha}^s = c^2 \left(-1 + \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{s} \left[1 - \frac{(-1)^\alpha (2\pi)^{2\alpha}}{(2\alpha)!} B_{2\alpha}(x_{ij}) \right] \right).$$

où P_α^s est défini en (2.10) et $B_{2\alpha}(\cdot)$ est le polynôme de Bernoulli de degré 2α.

Démonstration : on a que

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) &= \sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} |\hat{f}(\mathbf{h})|^2 \\
&\leq c^2 \sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} \|\mathbf{h}\|_\pi^{-2\alpha} \\
&= c^2 P_{2\alpha}^s, \\
&= c^2 \left(-1 + \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{s} \left[1 - \frac{(-1)^\alpha (2\pi)^{2\alpha}}{(2\alpha)!} B_{2\alpha}(x_{ij}) \right] \right),
\end{aligned}$$

où la première inégalité tient par hypothèse ; l'égalité suivante suit par définition de P_α^s (voir (2.9)) et la dernière égalité est obtenue en utilisant la formule (2.10), puisque 2α est nécessairement pair. ∎

Ainsi, le critère $P_{2\alpha}^s$ a une interprétation également au niveau de la variance des estimateurs provenant des règles de réseau translatées aléatoirement, lorsque l'on suppose certaines conditions sur f. En effet, une des interprétations possibles pour ce

corollaire est de dire qu'en choisissant une règle qui minimise $P_{2\alpha}^s$, on minimise une certaine borne sur la variance de $\hat{\mu}_{\mathrm{LR}}$ qui est valide pour les fonctions dans $E_\alpha(c)$.

Remarque 2.2.2 *Notons que pour analyser l'erreur d'intégration des fonctions dans $E_\alpha(c)$ et utiliser la formule pour P_α^s qui se calcule en $O(Ns)$ opérations, on doit prendre $\alpha \geq 2$ pair, alors que pour la variance, on n'a qu'à prendre $\alpha \geq 1$ entier.*

Remarque 2.2.3 *Une condition suffisante pour qu'il existe $c > 0$ tel que f soit dans $E_\alpha(c)$ pour $\alpha > 1$ un entier, est que f soit périodique avec une période de 1 (dans chaque dimension) et que ses dérivées partielles*

$$\frac{\partial^{q_1+\ldots+q_s} f}{\partial x_1^{q_1} \ldots \partial x_s^{q_s}},\ 0 \leq q_k \leq \alpha,\ 1 \leq k \leq s$$

existent et soient continues sur $[0,1)^s$ [116, page 71]. Donc, cela nous donne une façon de vérifier que le corollaire 2.2.1 s'applique sans connaître explicitement les coefficients de Fourier de f. Par contre, la périodicité semble une condition difficile à respecter en pratique. De plus, puisque cette condition tient pour $\alpha > 1$, l'avantage de considérer la variance plutôt que l'erreur, discuté à la remarque précédente, tombe.

Dans [116, *Theorem 5.2*], on donne un résultat (basé sur une borne due à Bahvalov [7]) disant que pour tout rang r, pour tout ensemble d'invariants $\{n_2, \ldots, n_r\}$ tels que $n_r \geq 2$ et n_{k+1} divise n_k, $k = 2, \ldots, r-1$ et pour tout nombre premier n, on peut trouver une règle de réseau de rang r ayant les invariants $n_1 = n, n_2, \ldots, n_r$ telle que

$$P_\alpha^s \leq e(s,\alpha)\frac{(\log N)^{\alpha(s-1)}}{N^\alpha},$$

où $N = n_1 n_2 \ldots, n_r$ et $e(s,\alpha)$ est indépendant de N. En combinant ceci avec le corollaire 2.2.1, cela signifie que pour toute forme de règle telle que spécifiée ci-dessus, on peut en trouver une dont la variance associée sera bornée par une certaine constante multipliée par $(\log N)^{2\alpha(s-1)}/N^{2\alpha}$, pour des fonctions dans $E_\alpha(c)$. Cependant, puisque ce résultat n'est pas constructif, on doit quand même faire une recherche sur toutes les règles ayant la structure sus-mentionnée afin de trouver celle qui minimise le critère $P_{2\alpha}^s$.

Tuffin a donné un résultat similaire dans [130], mais en utilisant une borne sur le P_α^s valide pour les règles de rang 1 et qui est donnée dans [99]. L'ordre de convergence

obtenu est légèrement moins bon que $O((\log N)^{2\alpha(s-1)}/N^{2\alpha})$, puisque la borne tirée de [99] tient pour tout N et non pas seulement pour N premier. Par contre, il est plus facile de faire une recherche pour trouver la meilleure règle par rapport à $P_{2\alpha}^s$ en se restreignant de cette façon. Mais ce n'est pas l'approche que nous voulons adopter : plutôt que d'essayer de minimiser une borne sur la variance, nous préférons essayer de minimiser directement celle-ci ou encore, démontrer que cette variance est inférieure à celle de l'estimateur MC. Malheureusement, il ne semble pas possible d'obtenir un tel résultat, à moins de se resteindre soit du côté des fonctions, soit du côté des règles utilisées. Par contre, nous verrons qu'en moyenne (sur un certain ensemble de règles), on ne peut généralement pas faire pire que l'estimateur MC. La prochaine sous-section traite de ces différentes idées.

2.2.3 Réduction de variance par rapport à Monte Carlo

Nous donnons d'abord une expression pour la variance de l'estimateur obtenu par la méthode MC, qui peut être comparée avec celle donnée en (2.16) pour les règles de réseau translatées aléatoirement.

Proposition 2.2.2 *Soit $f \in \mathcal{L}^2$ et $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC utilisant N points. Alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) = \frac{1}{N} \sum_{\mathbf{0} \neq \mathbf{h} \in \mathbb{Z}^s} |\hat{f}(\mathbf{h})|^2. \tag{2.19}$$

Démonstration : le théorème de Parseval s'applique car $f \in \mathcal{L}^2$ et donc, on a que

$$\int_{[0,1)^s} f^2(\mathbf{x}) d\mathbf{x} = \sum_{\mathbf{h} \in \mathbb{Z}^s} |\hat{f}(\mathbf{h})|^2.$$

Le résultat suit puisque $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) = \frac{1}{N}\left(\int_{[0,1)^s} f^2(\mathbf{x}) d\mathbf{x} - \mu^2\right)$ et que $\mu = \hat{f}(\mathbf{0})$. ∎

Si on regarde le rapport $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})/\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$, on peut voir le facteur $1/N$ dans (2.19) comme étant une normalisation qui tient compte du fait que la densité des points dans $L^\perp$ est de $1/N$. Ainsi, $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ sera inférieure à $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ si et seulement si, "en moyenne", les $|\hat{f}(\mathbf{h})|^2$ sont plus petits sur $L^\perp$ que sur $\mathbb{Z}^s$.

Évidemment, pour n'importe quelle règle P_N, on peut toujours trouver une fonction f dans $\mathcal{L}^2$ pour laquelle $\hat{\mu}_{\mathrm{LR}}$ sera un très mauvais estimateur de μ : en comparant (2.16)

et (2.19), on voit bien que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ pourrait être N fois plus grande que $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ dans le pire cas. Quel est ce pire cas? Si f est telle que les $|\hat{f}(\mathbf{h})|$ sont nuls lorsque $\mathbf{h} \notin L^{\perp}\backslash\{\mathbf{0}\}$. Cela se produit lorsque peu importe $\mathbf{u} \in [0,1)^s$, $f((\mathbf{x}_i + \mathbf{u}) \bmod 1) = f((\mathbf{x}_j + \mathbf{u}) \bmod 1)$, pour tout $1 \leq i,j \leq N$. Autrement dit, le pire cas se produit lorsque la *fréquence* de f est dans $L^{\perp}$, c.-à-d., s'il existe $\mathbf{h} \in L^{\perp}$ tel que $f(\mathbf{x}) = f(\mathbf{0})$ à chaque fois que $\mathbf{h} \cdot \mathbf{x}$ est entier, mais que f n'est pas constante.

En ce qui concerne l'erreur, ce pire cas pour la variance de $\hat{\mu}_{\mathrm{LR}}$ correspond à avoir $|Q_N - \mu| = |f(\mathbf{x}_1) - \mu|$, qui peut valoir 0 si $f(\mathbf{x}_1) = \mu$. Donc, les règles de réseau qui sont mauvaises du point de vue de la variance ne sont pas nécessairement mauvaises quand on considère l'erreur. Par contre, une mauvaise règle par rapport à la variance peut être détectée en comparant la variance empirique obtenue avec celle de l'estimateur MC, alors qu'avec l'erreur, on n'a pas vraiment de moyen de savoir si la règle est bonne ou non.

Pour démontrer que la variance de l'estimateur $\hat{\mu}_{\mathrm{LR}}$ est inférieure à celle de l'estimateur MC en utilisant les expressions (2.16) et (2.19), deux approches semblent possibles à première vue. La première serait d'imposer des conditions sur la fonction f par l'intermédiaire de ses coefficients de Fourier. Or, ces conditions doivent être très spécifiques à la règle employée pour que l'on puisse effectivement obtenir une réduction de variance. En effet, si on ne fait qu'imposer des conditions sur la "vitesse" de décroissance des $|\hat{f}(\mathbf{h})|$ en fonction de $\|\mathbf{h}\|$, cela ne nous garantit pas que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$, puisque cette condition profite autant aux deux estimateurs.

La deuxième approche possible serait d'essayer de démontrer que la variance moyenne prise sur un certain ensemble $R(L)$ de règles est inférieure à celle de la méthode MC. On voudrait idéalement que le résultat soit valide pour n'importe quelle fonction dans $\mathcal{L}^2$ et on serait tenté de croire que cela est possible, puisque l'on regarde la variance moyenne sur l'ensemble $R(L)$ et ainsi, les pires cas associés à chaque règle sont amoindris par le fait qu'ils ne sont pas les pires cas pour toutes les règles considérées. Une approche "en moyenne" comme celle-ci a été utilisée dans [22] pour démontrer des résultats sur l'erreur d'intégration des règles de réseau et dans [23, 54], pour comparer le P_α^s moyen de règles de différents rangs. C'est ce qui nous a suggéré

l'idée de considérer la variance moyenne.

Si, en plus de ce résultat, on était en possession d'une figure de mérite $\gamma(L)$ telle que pour deux règles $L_1, L_2 \in R(L)$, on avait que

$$\gamma(L_1) \leq \gamma(L_2) \Rightarrow \sum_{0 \neq \mathbf{h} \in L_1^\perp} |\hat{f}(\mathbf{h})|^2 \leq \sum_{0 \neq \mathbf{h} \in L_2^\perp} |\hat{f}(\mathbf{h})|^2, \tag{2.20}$$

alors on pourrait montrer que la règle faisant partie de $R(L)$ qui minimise $\gamma(L)$ donne un estimateur ayant une variance inférieure ou égale à celle de l'estimateur MC. Le problème ici, c'est que la condition (2.20) est non seulement spécifique à la fonction f, mais il semble également très difficile de trouver un critère $\gamma(\cdot)$ la satisfaisant, à moins de se restreindre à un ensemble bien précis de fonctions. Par exemple, nous verrons au chapitre 3 que dans le cas où la fonction f est un polynôme de degré deux, il est possible de trouver un critère $\gamma(\cdot)$ dépendant de f qui satisfait (2.20).

En résumé, il semble que pour démontrer que la variance est réduite avec les règles de réseau translatées aléatoirement, il faille soit se concentrer sur une règle en particulier et regarder l'ensemble de fonctions pour lesquelles on peut obtenir le résultat désiré, soit se concentrer sur une fonction et trouver un critère de sélection taillé sur mesure pour elle, qui nous permettra de choisir la règle appropriée réduisant la variance. En pratique, l'éventail de fonctions que l'on peut vouloir intégrer est large et donc, la restriction à un ensemble bien particulier de fonctions n'est pas souhaitable. De plus, cela nécessite que l'on soit capable de vérifier les conditions pour que la fonction appartienne à l'ensemble en question, ce qui n'est pas nécessairement facile à faire. Aussi, on préfère utiliser un critère qui ne soit pas spécifique à une fonction en particulier. De cette façon, les meilleures règles par rapport à ce critère peuvent être prédéterminées et ensuite utilisées sur un large éventail de fonctions. Donc, il vaut peut-être mieux penser dès maintenant à d'autres approches que les deux dont nous venons de discuter pour essayer de comparer la variance des règles de réseau et celle de la méthode MC.

Comme solutions alternatives, nous suggérons deux possibilités, toutes deux dérivées de la deuxième approche mentionnée précédemment. Nous allons voir que ces deux méthodes s'appliquent chacune dans un contexte différent.

- **Méthode 1 :** Définir un critère de sélection qui, pour la plupart des fonctions rencontrées en pratique, *devrait* satisfaire la relation (2.20). Cette approche est développée dans le chapitre 3, où l'idée clé servant à définir de tels critères est de regarder plus en détail les projections de la fonction et de l'ensemble de points P_N sur les sous-espaces de $[0,1)^s$.
- **Méthode 2 :** Choisir au hasard la règle de réseau parmi l'ensemble des règles de rang 1. Cette approche est traitée par la proposition 2.2.3, qui est énoncée formellement ci-dessous et qui sera démontrée à la suite de quelques explications.

Pour énoncer la proposition 2.2.3, nous avons besoin de la notation suivante :

$$\begin{aligned}
\hat{\mu}_{\mathrm{LR}}(\mathbf{A}) &= \text{estimateur venant de la règle d'ordre } N \text{ générée par le vecteur } \mathbf{A}, \\
&= \frac{1}{N}\sum_{i=0}^{N-1} f\left(\left(\frac{i}{N}\mathbf{A}\right) \bmod 1\right), \\
\hat{\mu}_{\mathrm{grid}} &= \text{estimateur obtenu en utilisant les } N^s \text{ points d'une grille rectangulaire} \\
&\quad \text{translatée aléatoirement et comptant } N \text{ points dans chaque dimension,} \\
&= \frac{1}{N^s}\sum_{m_1=0}^{N-1}\cdots\sum_{m_s=0}^{N-1} f\left(\left(\left(\frac{m_1}{N},\ldots,\frac{m_s}{N}\right)+\mathbf{u}_{\mathbf{g}}\right) \bmod 1\right), \\
&\quad \text{où } \mathbf{u}_{\mathbf{g}} \text{ est uniformément distribué dans } [0,1)^s.
\end{aligned}$$

Proposition 2.2.3 *Soit $f \in \mathcal{L}^2$ et $\mathbf{A}$ un vecteur aléatoire et uniforme dans $[1,\ldots,N-1]^s$. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC construit à partir de N points. Alors*

$$\mathrm{E}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) = \mu$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) \leq \frac{N}{N-1}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) + \frac{N-2}{N-1}\mathrm{Var}(\hat{\mu}_{\mathrm{grid}}).. \tag{2.21}$$

La proposition 2.2.3 peut être interprétée de deux façons différentes : la première interprétation motive la méthode 2 et la deuxième, la méthode 1. D'abord, dans le cas où nous n'avons pas suffisamment d'information sur la fonction f afin de déterminer si elle entre dans le contexte pour lequel les critères du chapitre 3 sont définis, la proposition nous indique que si on choisit une règle de façon aléatoire et uniforme parmi les $(N-1)^s$ règles de rang 1 d'ordre N et qu'on la translate aléatoirement, alors la variance de l'estimateur obtenu est bornée par la somme d'une quantité légèrement

supérieure à 1 multipliée par la variance de l'estimateur MC et du produit d'une quantité légèrement inférieure à 1 et de la variance de l'estimateur $\hat{\mu}_{\text{grid}}$ obtenu en translatant aléatoirement une grille rectangulaire comptant N^s points. Il est important de remarquer que c'est bien la grille avec N points par dimension et non N points au total, même si l'estimateur que l'on utilise ne contient que N points au total. Cela signifie que le deuxième terme de la borne (2.21) sera habituellement petit.

La deuxième interprétation de cette proposition s'applique davantage à la simulation, car nous croyons que dans ce contexte, on a habituellement une bonne idée de la structure de la fonction. Elle est basée sur le fait que la variance de l'estimateur basé sur une règle de rang 1 choisie au hasard est égale à la variance moyenne sur toutes les règles de rang 1. Ainsi, le résultat de la proposition signifie que cette variance moyenne est bornée approximativement par la somme de $\text{Var}(\hat{\mu}_{\text{MC}})$ et de $\text{Var}(\hat{\mu}_{\text{grid}})$. Donc, cela nous permet de dire que si $\text{Var}(\hat{\mu}_{\text{grid}}) \ll \text{Var}(\hat{\mu}_{\text{MC}})$ (ce qui devrait être le cas pour plusieurs fonctions), alors il doit y avoir des règles de rang 1 donnant lieu à des estimateurs réduisant la variance par rapport à $\hat{\mu}_{\text{MC}}$. En effet, dans la moyenne qui est faite sur les règles de rang 1, il se trouve des règles de piètre qualité qui doivent être compensées par des "bonnes" règles afin que la borne (2.21) sur la variance moyenne tienne. Ceci suggère qu'avec un critère de sélection motivé par une connaissance suffisante du type de fonction à intégrer, on pourra trouver ces "bonnes" règles. Ainsi, la proposition 2.2.3 peut être vue également comme étant une justification de la méthode 1.

Nous tenons à inclure dans le texte la démonstration de cette proposition, car elle nous permettra de trouver une borne plus serrée que celle donnée dans l'énoncé de la proposition. Ceci sera utilisé par la suite pour tenter d'améliorer le résultat. Le lemme qui suit sert à simplifier la présentation de la démonstration et requiert la notation suivante :

$$
\begin{aligned}
L^{\perp}(\mathbf{a}) &= \text{réseau dual à la règle de rang 1 dont le vecteur générateur est } \mathbf{a}, \text{ pour} \\
&\quad \mathbf{a} \in [1, \ldots, N-1]^s, \\
M_N^s(\mathbf{h}) &= \sum_{\mathbf{a} \in [1,\ldots,N-1]^s} \mathbf{1}_{\{\mathbf{h} \in L^{\perp}(\mathbf{a})\}},
\end{aligned}
$$

$$
\begin{aligned}
&= \text{nombre de vecteurs } \mathbf{a} \text{ dans } [1,\ldots,N-1]^s \text{ tels que } \mathbf{h} \text{ est dans } L^\perp(\mathbf{a}),\\
\eta(\mathbf{h}) &= \sum_{j=1}^{s} \mathbf{1}_{\{h_j \neq 0 \bmod N\}},\\
&= \text{nombre de coordonnées de } \mathbf{h} \text{ qui sont différentes de 0, modulo } N.
\end{aligned}
$$

Le lemme 2.2.1 nous indique, pour un vecteur $\mathbf{h}$ donné, combien de vecteurs $\mathbf{a}$ sont tels que $\mathbf{h}$ est dans le réseau dual au réseau généré par $\mathbf{a}$. Cela va nous servir à savoir combien de fois on doit inclure $|\hat{f}(\mathbf{h})|^2$ dans le calcul de la variance moyenne.

Lemme 2.2.1 *Soit N premier et $\mathbf{h} = (h_1,\ldots,h_s) \in \mathbb{Z}^s$. Alors*

$$M_N^s(\mathbf{h}) = \frac{(N-1)^s}{N}\left(1 + (-1)^{\eta(\mathbf{h})}(N-1)^{-\eta(\mathbf{h})+1}\right).$$

Démonstration : voir l'annexe B, page xxiii.

Démonstration de la proposition 2.2.3 : d'abord, on a que

$$\mathrm{E}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) = \mathrm{E}\left(\mathrm{E}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) \mid \mathbf{A} = \mathbf{a}\right) = \frac{1}{(N-1)^s}\sum_{\mathbf{a}} \mathrm{E}(\hat{\mu}_{\mathrm{LR}}(\mathbf{a})) = \mu,$$

par la proposition 2.2.1. Pour la variance, on procède comme suit :

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) &= \mathrm{Var}\left[\mathrm{E}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A}) \mid \mathbf{A} = \mathbf{a})\right] + \mathrm{E}\left[\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A}) \mid \mathbf{A} = \mathbf{a})\right]\\
&= \mathrm{Var}(\mu) + \frac{1}{(N-1)^s}\sum_{\mathbf{a}\in[1,\ldots,N-1]^s} \mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{a}))\\
&= \frac{1}{(N-1)^s}\sum_{\mathbf{a}\in[1,\ldots,N-1]^s}\ \sum_{0\neq\mathbf{h}\in L^\perp(\mathbf{a})} |\hat{f}(\mathbf{h})|^2\\
&= \frac{1}{(N-1)^s}\sum_{0\neq\mathbf{h}\in\mathbb{Z}^s} M_N^s(\mathbf{h})|\hat{f}(\mathbf{h})|^2\\
&= \frac{1}{(N-1)^s}\sum_{j=0}^{s}\ \sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=j}} M_N^s(\mathbf{h})|\hat{f}(\mathbf{h})|^2\\
&= \frac{1}{N}\left[\sum_{j=0}^{s}\ \sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=j}} \left(1 + (-1)^j(N-1)^{-j+1}\right)|\hat{f}(\mathbf{h})|^2\right]\\
&= \frac{1}{N-1}\left[\sum_{j=0}^{s}\ \sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=j}} \left(1 - \frac{1}{N} + \frac{(-1)^j}{N}(N-1)^{-j+2}\right)|\hat{f}(\mathbf{h})|^2\right] \qquad (2.22)\\
&= \frac{1}{N-1}\sum_{0\neq\mathbf{h}\in\mathbb{Z}^s} |\hat{f}(\mathbf{h})|^2 + \frac{N-2}{N-1}\sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=0}} |\hat{f}(\mathbf{h})|^2 - \frac{1}{N-1}\sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=1}} |\hat{f}(\mathbf{h})|^2
\end{aligned}
$$

$$+\frac{1}{N-1}\sum_{j=2}^{s}\sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=j}}\left(-\frac{1}{N}+\frac{(-1)^j}{N}(N-1)^{-j+2}\right)|\hat{f}(\mathbf{h})|^2 \tag{2.23}$$

$$\leq \frac{1}{N-1}\sum_{0\neq\mathbf{h}\in\mathbb{Z}^s}|\hat{f}(\mathbf{h})|^2+\frac{N-2}{N-1}\sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=0}}|\hat{f}(\mathbf{h})|^2$$

$$-\frac{1}{N-1}\sum_{\substack{0\neq\mathbf{h}\in\mathbb{Z}^s\\ \eta(\mathbf{h})=1}}|\hat{f}(\mathbf{h})|^2 \tag{2.24}$$

$$\leq \frac{N}{N-1}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})+\frac{N-2}{N-1}\mathrm{Var}(\hat{\mu}_{\mathrm{grid}}). \tag{2.25}$$

Dans ce qui précède, la troisième égalité est obtenue en appliquant la proposition 2.2.1 ; la sixième est obtenue en appliquant le lemme 2.2.1 ; la première inégalité tient car le terme entre parenthèses à la ligne 2.23 est inférieur ou égal à 0 ; la deuxième vient de l'expression (2.19) et du fait que le réseau dual correspondant à la grille $G = \{(m_1/N, \ldots, m_s/N), 0 \leq m_j < N, 1 \leq j \leq s\}$ est $\{\mathbf{h} \in \mathbb{Z}^s : \eta(\mathbf{h}) = 0\}$. ∎

Remarque 2.2.4 *On doit passer de* $1/N$ *à* $1/(N-1)$ *à la ligne 2.22, afin que le terme entre parenthèses à la ligne 2.23 soit inférieur ou égal à 0.*

Même si on s'attend à ce que la variance de $\hat{\mu}_{\mathrm{grid}}$ soit petite pour la plupart des fonctions, si on ne fait aucune hypothèse sur la fonction f, la borne (2.21) donnée à la proposition 2.2.3 peut être très grande. Par exemple, si f a une période de $1/N$ dans chaque dimension, alors $\mathrm{Var}(\hat{\mu}_{\mathrm{grid}}) = N^s\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ et donc, on aurait pu utiliser directement les expressions (2.16) et (2.19) pour $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ et $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ et obtenir ainsi une meilleure borne sur la variance de $\hat{\mu}_{\mathrm{LR}}(\mathbf{A})$, soit que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) \leq N\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. Par contre, le terme que nous avons laissé tomber de (2.24) à (2.25) peut être assez grand en valeur absolue, car il contient entre autres les coefficients de Fourier évalués en des points près de l'origine (de la forme $(0, \ldots, 0, k, 0, \ldots, 0)$, avec $1 \leq k \leq N-1$). Ainsi, en n'éliminant pas complètement ce terme et/ou en imposant des conditions sur f, on peut trouver des bornes plus serrées que celle donnée à la proposition 2.2.3 : c'est ce que nous faisons dans les deux corollaires qui suivent.

Corollaire 2.2.2 *Si* f *est dans* $E_\alpha(c)$ *pour une constante* $c > 0$ *et un entier* $\alpha \geq 1$, *alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) \leq \frac{N}{N-1}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})+c^2\left[-1+\left(1+\frac{\pi^2}{3N^{2\alpha}}\right)^s\right].$$

Démonstration : on a que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{grid}}) = \sum_{\substack{\mathbf{0}\neq\mathbf{h}\in\mathbb{Z}^s \\ \eta(\mathbf{h})=0}} |\hat{f}(\mathbf{h})|^2 \leq c^2 \sum_{\substack{\mathbf{0}\neq\mathbf{h}\in\mathbb{Z}^s \\ \eta(\mathbf{h})=0}} \|\mathbf{h}\|_\pi^{-2\alpha} = c^2\, P^s_{2\alpha,\mathrm{grid}},$$

où $P^s_{2\alpha,\mathrm{grid}}$ est le $P^s_{2\alpha}$ de la grille rectangulaire $G = \{(m_1/N, \ldots, m_s/N) : 0 \leq m_j \leq N-1, 1 \leq j \leq s\}$. Or, pour calculer cette quantité, on peut utiliser le théorème 6.7 dans [46] et ainsi, on obtient que

$$\begin{aligned}
P^s_{2\alpha,\mathrm{grid}} &= \left[-1 + \prod_{j=1}^{s}\left(1 - \frac{(-1)^\alpha(2\pi)^{2\alpha}}{N^{2\alpha}(2\alpha)!}B_{2\alpha}(0)\right)\right] \\
&= \left[-1 + \left(1 - \frac{(-1)^\alpha(2\pi)^{2\alpha}}{N^{2\alpha}(2\alpha)!}B_{2\alpha}(0)\right)^s\right] \\
&= \left[-1 + \left(1 + \frac{2\zeta(2\alpha)}{N^{2\alpha}}\right)^s\right] \\
&\leq \left[-1 + \left(1 + \frac{\pi^2}{3N^{2\alpha}}\right)^s\right],
\end{aligned}$$

puisque $\zeta(2j) \leq \zeta(2) = \pi^2/6$, pour tout entier $j \geq 1$, où $\zeta(y)$ est la fonction zeta de Riemann évaluée en y. ∎

Remarque 2.2.5 *Pour s fixé, la borne sur* $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A}))$ *donnée au corollaire 2.2.2 converge vers* $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ *quand* $N \to \infty$.

Une autre façon d'améliorer le résultat obtenu à la proposition 2.2.3 est de ne pas laisser tomber le terme négatif à la ligne (2.24), mais cela requiert que les coefficients de Fourier de f satisfassent une condition particulière qui est donnée au corollaire 2.2.3. Avant d'énoncer formellement ce résultat, nous voulons expliquer "en mots" quelle est cette condition. La valeur du terme que l'on a laissé tomber à la ligne (2.24) est

$$\frac{1}{N-1} \sum_{\substack{\mathbf{0}\neq\mathbf{h}\in\mathbb{Z}^s \\ \eta(\mathbf{h})=1}} |\hat{f}(\mathbf{h})|^2, \tag{2.26}$$

où $\eta(\mathbf{h}) = \sum_{j=1}^{s} \mathbf{1}_{\{h_j \neq 0 \bmod N\}}$ représente le nombre de coordonnées de $\mathbf{h}$ qui sont différentes de 0, modulo N. La condition donnée au corollaire 2.2.3 revient à demander que ce terme soit plus grand que le deuxième terme de la ligne (2.24), donné par

$$\frac{N-2}{N-1} \sum_{\substack{\mathbf{0}\neq\mathbf{h}\in\mathbb{Z}^s \\ \eta(\mathbf{h})=0}} |\hat{f}(\mathbf{h})|^2, \tag{2.27}$$

de façon à annuler ce dernier et faire en sorte qu'il ne reste que le premier, donné par $N\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})/(N-1)$. Or, plutôt que de vérifier cette condition de façon globale, nous allons décomposer les expressions (2.26) et (2.27) et vérifier la condition morceau par morceau. Plus précisément, à chaque terme dans (2.27), on associe un sous-ensemble de termes de (2.26) de la façon suivante : si $|\hat{f}(\mathbf{h}_0)|^2$ est le terme considéré dans (2.27), alors on prend chaque composante non nulle de $\mathbf{h}_0$ (on sait qu'il y en a au moins une) et si cette composante $h_{0,j}$ est positive, on regarde les $N-1$ vecteurs $\mathbf{h}_0^{l,j} = (h_{0,1}^{l,j}, \ldots, h_{0,s}^{l,j})$ définis par

$$h_{0,k}^{l,j} = \begin{cases} h_{0,k} - l & \text{si } k = j \\ h_{0,k} & \text{si } k \neq j, \end{cases} \tag{2.28}$$

pour $l = 1, \ldots, N-1$. Donc, tous ces $\mathbf{h}_0^{l,j}$ sont tels que $\eta(\mathbf{h}) = 1$, ne sont différents de $\mathbf{h}_0$ que par une coordonnée et ont la propriété que $\|\mathbf{h}_0^{l,j}\|_2 < \|\mathbf{h}_0\|_2$. Si $h_{0,j} < 0$, on fait la même chose, sauf que l'on remplace $h_{0,k} - l$ par $h_{0,k} + l$ dans (2.28). La figure 2.3 illustre comment on définit les $\mathbf{h}_0^{l,j}$ à partir de $\mathbf{h}_0$: on y donne deux points $\mathbf{h}_0$, $\tilde{\mathbf{h}}_0$ et on indique tous les points $\mathbf{h}_0^{l,j}$, $\tilde{\mathbf{h}}_0^{l,j}$ qui sont définis à partir de ces deux points respectifs.

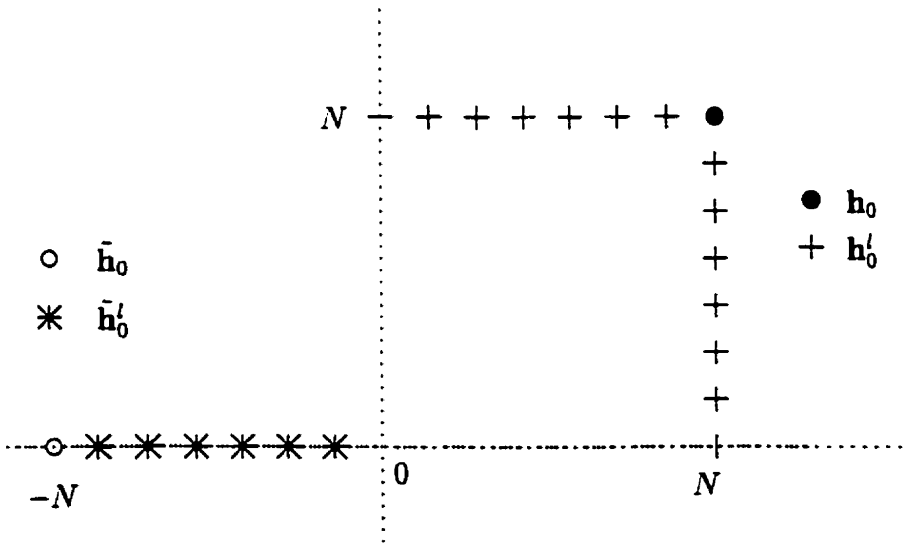


FIGURE 2.3: Illustration du regroupement : cas où $N = 7$

Dans la démonstration du corollaire 2.2.3, on montre que les sous-ensembles de vecteurs $\{\mathbf{h}_0^{l,j}, l = 1, \ldots, N-1; j : h_j \neq 0\}$ obtenus de cette façon partitionnent l'ensemble sur lequel on somme dans (2.26) (c.-à-d., les $\mathbf{h}$ tels que $\eta(\mathbf{h}) = 1$). Il suffit

alors de vérifier que chaque $\mathbf{h}_0$ satisfait

$$(N-2)|\hat{f}(\mathbf{h}_0)|^2 \leq \sum_{\substack{j=1 \\ h_{0,j}\neq 0}}^{s} \sum_{l=1}^{N-1} |\hat{f}(\mathbf{h}_0^{l,j})|^2,$$

afin que (2.26) annule (2.27). Le facteur $(N-2)$ qui multiplie $|\hat{f}(\mathbf{h}_0)|^2$ est contrebalancé par le fait que l'on somme $(N-1)|I_{\mathbf{h}_0}| \geq N-1$ termes à droite, où $I_{\mathbf{h}_0} = \{j : h_{0,j} \neq 0\}$. De plus, puisque $\|\mathbf{h}_0^{l,j}\|_2 < \|\mathbf{h}_0\|_2$, on peut penser que l'on devrait souvent avoir $|\hat{f}(\mathbf{h}_0)|^2 \leq |\hat{f}(\mathbf{h}_0^{l,j})|^2$ en pratique, si les coefficients de Fourier de f décroissent avec la longueur $\|\mathbf{h}\|_2$ de $\mathbf{h}$. Voici l'énoncé formel du résultat :

Corollaire 2.2.3 *Si pour tout* $\mathbf{h} \in \mathbb{Z}^s \setminus \{\mathbf{0}\}$ *tel que* $\eta(\mathbf{h}) = 0$, *on a que*

$$|\hat{f}(\mathbf{h})|^2 \leq \frac{1}{N-2} \left[\sum_{\substack{j=1 \\ h_j \neq 0}}^{s} \sum_{l=1}^{N-1} |\hat{f}(\mathbf{h} - l \cdot \mathrm{sgn}(h_j)\mathbf{e}_j)|^2 \right], \tag{2.29}$$

où $\mathrm{sgn}(h_j) = |h_j|/h_j$ *et* $\mathbf{e}_j \in \mathbb{Z}^s$ *est un vecteur de 0 avec un 1 en* j^e *position, alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) \leq \frac{N}{N-1}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}),$$

où $\mathbf{A}$ *est choisi uniformément et aléatoirement dans* $[1, \ldots, N-1]^s$.

Démonstration : voir l'annexe B, page xxv.

Le résultat est intéressant, car il nous montre que sous la condition (2.29), on obtient une réduction de variance garantie par rapport à MC en utilisant l'estimateur $\hat{\mu}_{\mathrm{LR}}(\mathbf{A})$ avec $\mathbf{A}$ choisi aléatoirement, à un facteur $N/(N-1)$ près. De façon équivalente, ce résultat signifie que si f satisfait (2.29), alors la variance moyenne sur les règles de rang 1 est bornée par la variance de $\hat{\mu}_{\mathrm{MC}}$, à un facteur près. Afin de rendre ce résultat plus attrayant en pratique, un objectif à atteindre dans le futur serait de traduire la condition (2.29) en une plus facile à vérifier, qui, par exemple, ne ferait pas intervenir les coefficients de Fourier de f.

2.2.4 Autres approches que celles utilisant les séries de Fourier

Le corollaire 2.2.3 impose des conditions sur les coefficients de Fourier de la fonction f afin de démontrer un résultat relié à la réduction de variance. De façon plus

générale, l'approche "Fourier" est celle qui a été favorisée jusqu'à présent dans cette section. Par contre, un autre type de condition qui pourrait être considéré serait de demander que f soit monotone par rapport à chacun de ses arguments. En effet, pour certaines méthodes de réduction de variance connexes comme les variables antithétiques et l'échantillonnage de l'hypercube latin, la monotonicité de f par rapport à chacun des arguments garantit la réduction de variance [40, 90, 5]. De plus, nous verrons à la sous-section 2.3.3 que ce résultat est vrai en dimension $s = 1$ pour une règle de réseau translatée qui est complètement projection-régulière. Par contre, nous verrons également à cette section que la généralisation à $s > 1$ ne peut se faire, du moins, pas en utilisant le concept de *dépendance négative par quadrant* (d.n.q.) introduit par Lehmann dans [81] et utilisé dans [5] pour démontrer la réduction de variance amenée par les variables antithétiques et l'échantillonnage de l'hypercube latin pour les fonctions monotones, en dimension $s \geq 1$.

On pourrait se demander si une approche différente de celle utilisée dans [5] pourrait permettre de démontrer la réduction de variance pour $s > 1$. La réponse est non, du moins, pas si on veut que le résultat soit vrai pour n'importe quelle règle de réseau translatée. Même si on ajoute comme condition que la règle soit complètement projection-régulière et de rang 1 (ce qui semble raisonnable, si on veut pouvoir appliquer le résultat en dimension 1), le résultat n'est pas vrai, c.-à-d., on peut trouver un contre-exemple :

Exemple 2.2.1 *Le contre-exemple est donné en dimension $s = 2$ et la règle de réseau est telle que $\mathbf{x}_i = ((i-1)/N, (i-1)/N)$, pour tout $i = 1, \ldots, N$ (donc, les points se trouvent sur la diagonale $x = y$). Notons que cette règle est complètement projection-régulière et de rang 1. La fonction f est définie par*

$$f(x,y) = \begin{cases} c_1 & si\ y < x - b_\epsilon, \\ c_2 & si\ x - b_\epsilon \leq y < x + b_\epsilon, \\ c_3 & si\ y \geq x + b_\epsilon, \end{cases}$$

où $\epsilon < (1 + 2(N-1))/2N^2$, $b_\epsilon = (1 - \sqrt{1-2\epsilon})$ et c_1, c_2, c_3 sont tels que $c_1 < c_2 < c_3$ et $c_2 = (c_1 + c_3)/2$. Donc, f est monotone décroissante par rapport à x et monotone croissante par rapport à y.

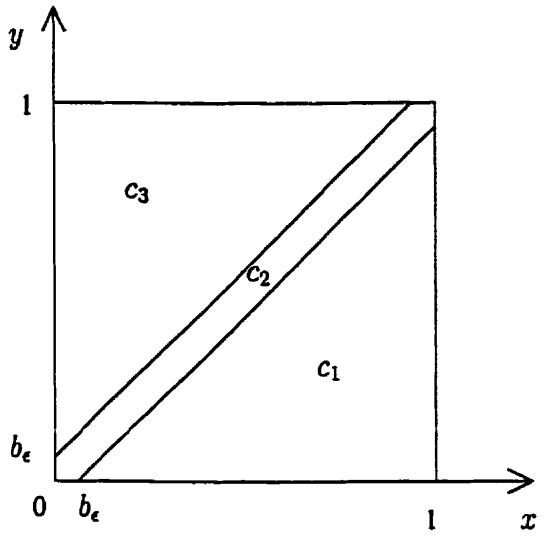


FIGURE 2.4: $f(x, y)$

Or, on peut montrer que si $N > 8$, *alors* $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) > \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. *La démonstration se trouve à l'annexe A.*

Ce contre-exemple détruit donc tout espoir de démontrer le résultat " *Si* $f \in \mathcal{L}^2$ *est monotone par rapport à chacun de ses arguments et qu'on utilise une règle de réseau complètement projection-régulière de rang 1, alors* $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$".

2.2.5 La pratique à la rescousse de la théorie

Même si nous n'avons pas donné dans cette section des résultats qui garantissent que les règles de réseau translatées aléatoirement réduisent la variance, il semble qu'en pratique, lorsque la règle est choisie adéquatement, par exemple à l'aide d'une figure de mérite basée sur le test spectral, les règles de réseau translatées ont une variance empirique généralement inférieure (souvent significativement) à celle de l'estimateur MC et ce, pour différents types de problèmes et donc, de fonctions. Nous donnons à la fin de ce chapitre trois exemples pour lesquels on réussit toujours à réduire la variance, parfois par des facteurs aussi grands que 460. Dans le premier cas, on s'intéresse au calcul de la probabilité qu'un projet soit terminé à temps pour un réseau stochastique d'activités, dans le deuxième, on cherche à estimer le prix d'options asiatiques et dans le troisième, on veut calculer la probabilité de ruine d'une compagnie d'assurances.

Pour ces problèmes, les fonctions regardées sont parfois définies en très grande dimension (jusqu'à 40000 dans le cas de la probabilité de ruine) et même avec un nombre de points N relativement petit (de l'ordre de 256), on réussit à réduire la variance par rapport à MC, empiriquement. Ceci semble aller à l'encontre de ce que le taux de convergence asymptotique des méthodes QMC nous dit par rapport à celui de la méthode MC, mais ces taux de convergence sont pour les *bornes* sur l'erreur et non pas pour l'erreur en tant que telle. Cela confirme ce dont nous avons déjà discuté dans l'introduction, à l'effet que les bornes de type (1.6) ne sont pas très utiles afin d'expliquer le succès des méthodes QMC, particulièrement lorsque l'on travaille en grande dimension.

À ce point-ci, on peut utiliser des raisonnements heuristiques basés sur les propositions 2.2.1 et 2.2.3 (avec l'interprétation de la variance moyenne) afin d'expliquer ces résultats en disant : "Le fait d'utiliser un critère basé sur le test spectral afin de sélectionner la règle de réseau nous permet de construire l'estimateur $\hat{\mu}_{\mathrm{LR}}$ de façon à ce que plusieurs des coefficients de Fourier qui contribuent de façon significative à la variance de $\hat{\mu}_{\mathrm{MC}}$ soient éliminés de la variance de $\hat{\mu}_{\mathrm{LR}}$ et ainsi, on obtient que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) < \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. Le résultat donné à la proposition 2.2.3 suggère que de telles règles existent." Au chapitre 3, le fait de regarder plus en détail les caractéristiques de la fonction et de la règle sur les espaces en basse dimension nous permettra de mieux voir comment la variance peut être réduite si on choisit les règles de façon appropriée.

2.3 Permutations des coordonnées

Dans cette section, nous démontrons que si l'on permute aléatoirement les points contenus dans chacune des s projections unidimensionnelles d'une règle de réseau et que l'on applique ensuite une translation aléatoire et uniforme dans $[0,1)^s$ (modulo 1) au réseau, alors la variance de l'estimateur ainsi obtenue est inférieure ou égale à celle de l'estimateur MC, à condition que la fonction f soit monotone par rapport à chacun de ses arguments. Nous expliquons à la sous-section 2.3.2 les liens entre cette méthode et celle de l'échantillonnage de l'hypercube latin, puis voyons à la sous-section 2.3.3 comment notre résultat s'applique aux règles de réseau simplement translatées

aléatoirement en une dimension.

2.3.1 Variance de l'estimateur obtenu par permutation

Nous considérons l'estimateur

$$\hat{\mu}_{\mathrm{LRp}} = \frac{1}{N}\sum_{i=1}^{N} f\left((\mathbf{w}_i + \mathbf{u}) \mod 1\right), \tag{2.30}$$

où $\mathbf{u}$ est un vecteur aléatoire, uniformément distribué sur $[0,1)^s$ et les $\mathbf{w}_i$ sont obtenus à partir de s permutations aléatoires uniformes et indépendantes de $[1,\ldots,N]$ données par $\pi^j = (\pi_1^j,\ldots,\pi_N^j)$, $j = 1,\ldots,s$, en posant

$$\mathbf{w}_i = (x_{\pi_i^1 1},\ldots,x_{\pi_i^s s}),\ i = 1,\ldots,N,$$

où $P_N = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ est une règle de réseau. Donc, π^j permute l'ordre des points de $P_N(\{j\})$, la projection de P_N en dimension j. Notons que pour avoir $\{w_{1j},\ldots,w_{Nj}\} = \{0,1/N,\ldots,(N-1)/N\}$ pour $j = 1,\ldots,s$, on doit demander que la règle soit de rang 1 et complètement projection-régulière.

Remarquons que l'on applique la *même* translation $\mathbf{u}$ à tous les points. Ceci diffère de la méthode de l'échantillonnage de l'hypercube latin, dans laquelle chaque point est translaté indépendamment des autres. Nous reviendrons sur cette différence à la sous-section 2.3.2.

La proposition qui suit montre que l'estimateur $\hat{\mu}_{\mathrm{LRp}}$ est sans biais et que sa variance est bornée par $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$, en autant que la fonction soit monotone par rapport à chaque variable. Ce résultat et sa démonstration sont largement inspirés du résultat similaire qui existe pour la méthode de l'échantillonnage de l'hypercube latin, que l'on trouve dans [90, 5].

Proposition 2.3.1 *Soit $f \in \mathcal{L}^2$, monotone par rapport à chacun de ses arguments. Soit P_N une règle de réseau de rang 1 complètement projection-régulière et $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC basé sur N points. Alors*

$$\mathrm{E}(\hat{\mu}_{\mathrm{LRp}}) = \mu$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LRp}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : dans ce qui suit, lorsque nous utilisons la notation $\mathrm{E}_{\mathbf{x}}(\cdot)$, c'est pour indiquer que l'espérance est définie par rapport à la variable aléatoire $\mathbf{x}$ et $\mathrm{Cov}_{\mathbf{x},\mathbf{y}}(\cdot,\cdot)$ signifie que la covariance est définie sur les variables aléatoires $\mathbf{x}, \mathbf{y}$.

D'abord, puisque les permutations $\pi^1, \ldots, \pi^s$ sont aléatoires, uniformes et indépendantes, on a que chaque $\mathbf{w}_i$ suit une loi uniforme discrète sur $[0, 1/N, \ldots, (N-1)/N]^s$. Donc,

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_{\mathrm{LRp}}) &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{\mathbf{w}_i}(\mathrm{E}_{\mathbf{u}}(f((\mathbf{w}_i + \mathbf{u}) \mod 1) \mid \mathbf{w}_i)) \\
&= \frac{1}{(N-1)^s} \sum_{\mathbf{x}\in[0,\ldots,(N-1)/N]^s} \int_{[0,1)^s} f((\mathbf{x}+\mathbf{u}) \mod 1) d\mathbf{u} \\
&= \mu,
\end{aligned}$$

par le même argument qu'en (2.17).

Posons $X_i = (\mathbf{w}_i + \mathbf{u}) \mod 1$, $i = 1, \ldots, N$. Par (2.30), on a que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LRp}}) = \frac{1}{N}\mathrm{Var}(f(X_i)) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} \mathrm{Cov}(f(X_i), f(X_j)),$$

et $\mathrm{Var}(f(X_i)) = \sigma^2$, $i = 1, \ldots, N$, où $\sigma^2 = \int_{[0,1)^s} f(\mathbf{x})d\mathbf{x} - \mu^2$, car $X_i = (X_{i1}, \ldots, X_{is}) \sim U[0,1)^s$. Puisque $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) = \sigma^2/N < \infty$, on doit donc montrer que

$$\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} \mathrm{Cov}(f(X_i), f(X_j)) \leq 0. \tag{2.31}$$

Étant donné que les paires $(X_i, X_j), i = 1, \ldots, N, j = 1, \ldots, N, j \neq i$ ont toutes la même loi de probabilité conjointe, il est suffisant de montrer que $\mathrm{Cov}(f(X_1), f(X_2)) \leq 0$ pour démontrer (2.31). Pour faire cela, nous allons procéder en trois étapes : d'abord, nous allons décomposer autrement les points $X_i, i = 1, \ldots, N$, puis, nous allons conditionner la covariance sur une des variables aléatoires nouvellement définies, pour ensuite appliquer un résultat tiré de [90] qui nous permettra de démontrer que $\mathrm{Cov}(f(X_1), f(X_2)) \leq 0$ à l'aide de définitions et du théorème 1 tirés de [81].

Étant donné que $\mathbf{u} \sim U([0,1)^s)$, chaque u_k, $k = 1, \ldots, s$, peut être réécrit comme étant $u_k = n_k + v_k$, où n_k suit une loi uniforme discrète sur $[0, 1/N, \ldots, (N-1)/N]$ et $v_k \sim U([0, 1/N))$, pour $k = 1, \ldots, s$, et ces $2s$ variables aléatoires sont mutuellement indépendantes. Ainsi, on a que

$$X_{ik} = (w_{ik} + u_k) \mod 1 = (w_{ik} + n_k + v_k) \mod 1 = \tilde{w}_{ik} + v_k,$$

où $\tilde{w}_{ik} = (w_{ik} + n_k) \bmod 1$.

Or, dans chaque dimension $k = 1, \ldots, s$, la paire $(\tilde{w}_{1k}, \tilde{w}_{2k})$ a la même loi de probabilité conjointe que la paire (w_{1k}, w_{2k}) : en effet, pour a, b dans $[0, 1/N, \ldots (N-1)/N]$, on a

$$\begin{aligned}
P[(\tilde{w}_{1k}, \tilde{w}_{2k}) = (a,b)] &= \int_0^1 P[(\tilde{w}_{1k}, \tilde{w}_{2k}) = (a,b) \mid u_k] du_k \\
&= \frac{1}{N} \sum_{p=0}^{N-1} P[(\tilde{w}_{1k}, \tilde{w}_{2k}) = (a,b) \mid n_k = p/N] \\
&= \frac{1}{N} \sum_{p=0}^{N-1} P[(w_{1k}, w_{2k}) = ((a - p/N) \bmod 1, (b - p/N) \bmod 1)] \\
&= \begin{cases} 1/(N(N-1)) & \text{si } a \neq b \\ 0 & \text{sinon} \end{cases} \\
&= P[(w_{1k}, w_{2k}) = (a,b)].
\end{aligned}$$

De plus, la loi de chaque $\tilde{w}_{ik}$ est la même que celle de w_{ik} et les $\tilde{w}_{i1}, \ldots, \tilde{w}_{is}$ sont indépendants entre eux, puisque les w_{ik} et les u_k le sont. Donc, on peut réécrire X_i comme étant

$$X_i = \tilde{\mathbf{w}}_i + \mathbf{v},$$

où $\mathbf{v} \sim U([0, 1/N)^s)$ et les $\tilde{\mathbf{w}}_i, i = 1, \ldots, N$ ont les mêmes propriétés que les $\mathbf{w}_i, i = 1, \ldots, N$. Autrement dit, l'ensemble de points $\{X_1, \ldots, X_N\}$ est équivalent à celui qu'on obtiendrait en additionnant à chaque $\mathbf{w}_i$ un vecteur uniforme sur $[0, 1/N)^s$.

Ensuite, on utilise le fait que

$$\begin{aligned}
\mathrm{Cov}(f(X_1), f(X_2)) &= \mathrm{E}_{\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2}(\mathrm{Cov}(f(X_1), f(X_2) \mid \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)) \\
&\quad + \mathrm{Cov}_{\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2}(\mathrm{E}(f(X_1) \mid \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2), \mathrm{E}(f(X_2) \mid \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)) \quad (2.32)
\end{aligned}$$

et nous allons montrer que cette quantité est inférieure ou égale à 0 en montrant que chacune des deux paires $(f(X_1 \mid \tilde{\mathbf{w}}_1), f(X_2 \mid \tilde{\mathbf{w}}_2))$ et $(\mathrm{E}(f(X_1) \mid \tilde{\mathbf{w}}_1), \mathrm{E}(f(X_2) \mid \tilde{\mathbf{w}}_2))$ est d.n.q. (à dépendance négative par quadrant), ce qui nous permettra de conclure. Nous expliquons d'abord la définition de *fonctions concordantes*, de variables aléatoires *à dépendance négative par quadrant (d.n.q.) (negatively quadrant dependant)* et rappelons un théorème dû à Lehmann [81], qui nous servira à démontrer que les deux paires en question sont d.n.q. .

Définition 2.3.1 ([81]) *Deux fonctions $g(x_1,\ldots,x_s)$ et $h(y_1,\ldots,y_s)$ sont dites* concordantes *si soit elles augmentent, soit elles diminuent ensemble en fonction de $x_k = y_k$, lorsque tous les $x_j, j \neq k$ et $y_j, j \neq k$ sont fixés, pour tout k.*

Définition 2.3.2 ([81]) *On dit qu'une paire de variables aléatoires (X,Y) est* d.n.q. *si*

$$P(X \leq x, Y \leq y) \leq P(X \leq x)P(Y \leq y).$$

Dans [81, *Theorem 1*], on dit que si (1) les paires $(X_1,Y_1),(X_2,Y_2),\ldots,(X_s,Y_s)$ sont indépendantes, (2) (X_k,Y_k) est d.n.q. pour tout k, (3) $X = g(X_1,\ldots,X_s)$ et $Y = h(Y_1,\ldots,Y_s)$ sont *concordantes* en chaque argument, alors (X,Y) est d.n.q. .

Dans [90], on montre que les paires (w_{1k},w_{2k}), $k = 1,\ldots,s$ sont d.n.q., ce qui implique que les paires $(\tilde{w}_{1k},\tilde{w}_{2k})$, $k = 1,\ldots,s$ le sont aussi, puisqu'elles ont la même loi de probabilité conjointe et on sait aussi que ces paires sont mutuellement indépendantes, puisque les permutations $\pi^1,\ldots,\pi^s$ et les translations $u_1,\ldots,u_s$ sont indépendantes d'une dimension à l'autre. De plus, si on pose

$$\begin{aligned} g(\tilde{w}_{11},\ldots,\tilde{w}_{1s}) &= f(X_1 \mid \tilde{w}_{11},\ldots,\tilde{w}_{1s}) = f(\tilde{\mathbf{w}}_1 + \mathbf{v}) \\ \text{et } h(\tilde{w}_{21},\ldots,\tilde{w}_{2s}) &= f(X_2 \mid \tilde{w}_{21},\ldots,\tilde{w}_{2s}) = f(\tilde{\mathbf{w}}_2 + \mathbf{v}), \end{aligned}$$

alors g et h sont concordantes, puisque f est monotone par rapport à chaque variable. De même, si on pose

$$\begin{aligned} g(\tilde{w}_{11},\ldots,\tilde{w}_{1s}) &= \mathrm{E}(f(X_1 \mid \tilde{w}_{11},\ldots,\tilde{w}_{1s})) \\ h(\tilde{w}_{21},\ldots,\tilde{w}_{2s}) &= \mathrm{E}(f(X_2 \mid \tilde{w}_{21},\ldots,\tilde{w}_{2s})), \end{aligned}$$

et que l'on utilise la notation $\int_{c_i} f(\tilde{\mathbf{w}}_i + \mathbf{v})d\mathbf{v}$ pour signifier que l'on intègre $f(\cdot)$ sur la cellule déterminée par $\tilde{\mathbf{w}}_i$, alors on a que

$$\begin{aligned} g(\tilde{w}_{11},\ldots,\tilde{w}_{1s}) &= N\int_{c_1} f(\tilde{\mathbf{w}}_1 + \mathbf{v})d\mathbf{v} \\ \text{et } h(\tilde{w}_{21},\ldots,\tilde{w}_{2s}) &= N\int_{c_2} f(\tilde{\mathbf{w}}_2 + \mathbf{v})d\mathbf{v}, \end{aligned}$$

et donc on a que g et h sont concordantes, puisque f est monotone par rapport à chaque variable.

Puisque ces deux paires sont d.n.q. et que l'équation d'Hoeffding (voir [81], par exemple) nous indique que

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[P(X \le x, Y \le y) - P(X \le x)P(Y \le y)]dxdy,$$

cela signifie que $\mathrm{Cov}(f(X_1), f(X_2) \mid \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2) \le 0$ avec probabilité 1 et que

$$\mathrm{Cov}_{\tilde{\mathbf{w}}_1,\tilde{\mathbf{w}}_2}[\mathrm{E}(f(X_1 \mid \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)), \mathrm{E}(f(X_2 \mid \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2))] \le 0,$$

ce qui, en utilisant l'expression (2.32) pour $\mathrm{Cov}(f(X_1), f(X_2))$, nous démontre que $\mathrm{Cov}(f(X_1), f(X_2)) \le 0$. ∎

Donc, en randomisant davantage une règle de réseau que ce qui est fait dans la méthode par simple translation aléatoire, on a un résultat de réduction de variance garantie, pour les fonctions monotones. Notons cependant que l'application des permutations π^j, $1 \le j \le s$ sur le réseau $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ fait en sorte que l'on perd complètement la structure de réseau, c.-à-d., les points $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ ainsi obtenus ne forment pas un réseau.

2.3.2 Comparaison avec la méthode de l'échantillonage de l'hypercube latin

En fait, l'ensemble de points $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ correspond à celui qui est utilisé dans la méthode de l'échantillonage de l'hypercube latin (LHS) [90], avant qu'ils ne soient translatés, c.-à-d., pour la méthode LHS, on utilise les N points $\{\mathbf{w}_1 + \mathbf{u}_1, \ldots, \mathbf{w}_N + \mathbf{u}_N\}$, où les $\mathbf{u}_i$, $i = 1, \ldots, N$ sont i.i.d. uniformément distribués sur $[0, 1/N)^s$. Donc, la différence entre l'estimateur $\hat{\mu}_{\mathrm{LRp}}$ et celui obtenu par la méthode LHS, que l'on dénote par $\hat{\mu}_{\mathrm{LHS}}$, est que dans le premier cas, on utilise le même vecteur $\mathbf{u}$ pour translater tous les points modulo 1, alors que dans le second, chaque point est translaté indépendamment des autres, dans l'hypercube de côté $1/N$ défini par $\prod_{j=1}^{s}[w_{ij}, w_{ij} + 1/N)$. Cette ressemblance entre les deux méthodes nous a permis de démontrer la proposition 2.3.1 assez facilement, car nous n'avons eu qu'à modifier légèrement ce qui avait été fait pour $\hat{\mu}_{\mathrm{LHS}}$ dans [90, 5]. On a donc le même résultat de réduction de variance dans les deux cas, mais le coût de calcul de $\hat{\mu}_{\mathrm{LHS}}$ est plus élevé que celui de $\hat{\mu}_{\mathrm{LRp}}$, puisque l'on doit générer N fois plus de vecteurs aléatoires en s dimensions.

Nous verrons dans les résultats présentés à la fin de ce chapitre que les deux estimateurs ont une variance empirique à peu près semblable et donc, en termes d'efficacité, $\hat{\mu}_{\mathrm{LRp}}$ est avantageux par rapport à $\hat{\mu}_{\mathrm{LHS}}$ sur ces exemples. De façon ironique, ces résultats nous démontreront également que la variance empirique de l'estimateur $\hat{\mu}_{\mathrm{LR}}$ obtenu par simple translation aléatoire est habituellement inférieure à celle de ces deux estimateurs théoriquement plus sûrs, en plus d'avoir un coût de calcul inférieur. Donc, $\hat{\mu}_{\mathrm{LR}}$ est doublement gagnant du point de vue de l'efficacité sur ces exemples. La variance empirique de $\hat{\mu}_{\mathrm{LR}}$ démontre également un taux de convergence meilleur que le $O(N^{-1})$ de $\hat{\mu}_{\mathrm{MC}}$, alors que cela ne semble pas être le cas pour $\hat{\mu}_{\mathrm{LRp}}$ et $\hat{\mu}_{\mathrm{LHS}}$. Évidemment, la règle de réseau doit être choisie de façon appropriée pour que $\hat{\mu}_{\mathrm{LR}}$ ait ces propriétés.

2.3.3 Application aux règles de réseau translatées aléatoirement

Le résultat suivant sur la variance de $\hat{\mu}_{\mathrm{LR}}$ en une dimension découle directement de la proposition 2.3.1 :

Corollaire 2.3.1 *Si $s = 1$, que la fonction est monotone et que P_N est (complètement) projection-régulière, alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : la règle est de rang 1 puisque $s = 1$ et on applique directement la proposition 2.3.1, puisque la permutation des coordonnées n'affecte pas P_N étant donné que $s = 1$. ∎

Notons qu'en une dimension, on a automatiquement une règle complètement projection-régulière, en autant que la règle contienne N points distincts. Dans [30, *Theorem 2*], on démontre un résultat plus fort (soit que la variance est minimisée par $\hat{\mu}_{\mathrm{LR}}$ en dimension 1), mais en plus de la monotonicité, on demande que la covariance associée à la fonction f soit convexe et symétrique par rapport à l'axe $x = 1/2$.

La raison pour laquelle on ne peut généraliser ce résultat à $s > 1$ dans le cas des règles de réseau translatées est que pour ces dernières, dans une dimension k donnée, les paires $((x_{ik} + u_k) \bmod 1, (x_{jk} + u_k) \bmod 1), 1 \leq i, j \leq N, i \neq j$ ne sont pas toutes

d.n.q. et donc, on ne peut appliquer directement le théorème 1 de [81]. Par exemple, si les deux points x_{ik} et x_{jk} sont espacés de $1/N$, avec disons $x_{ik} = (i-1)/N$ et $x_{jk} = i/N$, alors on a $P((((i-1)/N+u) \bmod 1) \leq x, ((i/N+u) \bmod 1) \leq y) = \min(x,y) - 1/N$, qui n'est pas nécessairement inférieur ou égal à xy. Remarquons également que le résultat pour $s = 1$ nous indique que si f est monotone, alors pour tout $N \geq 1$ entier,

$$\sum_{0 \neq h \in \mathbb{Z}} |\hat{f}(hN)|^2 \leq \frac{1}{N} \sum_{0 \neq h \in \mathbb{Z}} |\hat{f}(h)|^2.$$

2.4 Stratification

Dans cette section, nous proposons une autre façon de randomiser une règle de réseau, qui revient en fait à utiliser la méthode de la stratification [15], une des techniques de réduction de la variance souvent utilisée en simulation. L'idée est de choisir une base pour le réseau et d'utiliser le parallélépipède fondamental associé à ce choix de base pour partitionner l'hypercube $[0,1)^s$ en N cellules de même volume. Ensuite, on génère un point aléatoirement et uniformément dans chacune de ces cellules et les points ainsi obtenus sont ceux que l'on utilise afin de construire une approximation pour μ. La figure 2.5 illustre comment se fait ce type de randomisation en deux dimensions.

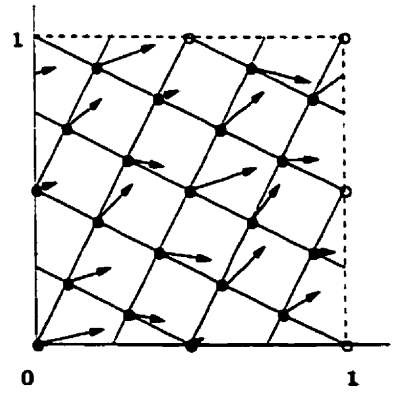


FIGURE 2.5: Illustration de la stratification

Les résultats sur la stratification nous garantissent que la variance de l'estimateur ainsi construit sera inférieure à celle de l'estimateur MC, pour n'importe quelle fonction

f dans $\mathcal{L}^2$. Ainsi, comme à la section précédente, en utilisant une randomisation "plus aléatoire" que la translation modulo 1, on réussit à démontrer un résultat garantissant la réduction de variance.

Proposition 2.4.1 *Soit $f \in \mathcal{L}^2$. Soit $\mathbf{v}_1, \ldots, \mathbf{v}_s$ une base engendrant un réseau L (c.-à-d., , $L = \{\sum_{j=1}^s c_j \mathbf{v}_j : c_1, \ldots, c_s \in \mathbb{Z}\}$) et Λ le parallélépipède associé. Posons*

$$\mathbf{C}_i = (\mathbf{x}_i + \Lambda) \mod 1, \qquad i = 1, \ldots, N,$$

où $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} = L \cap [0,1)^s$ est la règle de réseau associée à L. Soit $\mathbf{w}_i$ un point aléatoirement et uniformément distribué dans $\mathbf{C}_i$, pour $i = 1, \ldots, N$, et supposons que les points $\mathbf{w}_1, \ldots, \mathbf{w}_N$ sont indépendants. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC et posons

$$\hat{\mu}_{\mathrm{LRSt}} = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{w}_i).$$

Alors

$$\mathrm{E}(\hat{\mu}_{\mathrm{LRSt}}) = \mu$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LRSt}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : d'abord, posons

$$\mu_i = \mathrm{E}(f(\mathbf{w}_i)), \qquad \sigma_i^2 = \mathrm{Var}(f(\mathbf{w}_i)).$$

Ainsi,

$$\mathrm{E}(\hat{\mu}_{\mathrm{LRSt}}) = \frac{1}{N} \sum_{i=1}^{N} \mu_i = \frac{1}{N} \sum_{i=1}^{N} N \int_{\mathbf{x} \in \mathbf{C}_i} f(\mathbf{x}) d\mathbf{x} = \mu$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LRSt}}) = \frac{1}{N^2} \sum_{i=1}^{N} \sigma_i^2,$$

par indépendance des $\mathbf{w}_i$, où

$$\sigma_i^2 = N \int_{\mathbf{x} \in \mathbf{C}_i} f^2(\mathbf{x}) d\mathbf{x} - \mu_i^2, \; i = 1, \ldots, N,$$

alors que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) \;=\; \frac{1}{N} [\mathrm{E}_i(\mathrm{Var}(f(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathbf{C}_i)) + \mathrm{Var}_i(\mathrm{E}(f(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathbf{C}_i))]$$

$$
\begin{aligned}
&= \frac{1}{N}\left[\mathrm{E}_i\mathrm{Var}(f(\mathbf{w}_i)) + \mathrm{Var}_i\mathrm{E}(f(\mathbf{w}_i))\right] \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\sigma_i^2 + \sum_{i=1}^{N}(\mu_i - \mu)^2\right) \\
&\geq \mathrm{Var}(\hat{\mu}_{\mathrm{LRSt}}). \quad \blacksquare
\end{aligned}
$$

On peut voir que $\mathrm{Var}(\hat{\mu}_{\mathrm{LRSt}})$ est strictement inférieure à $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ si, pour au moins un $i, 1 \leq i \leq N$, on a $\mu_i \neq \mu$. Le cas extrême apparaît lorsque la partition de $[0,1)^s$ est telle que f est constante sur chaque $\mathbf{C}_i$ et ainsi, $\mathrm{Var}(\hat{\mu}_{\mathrm{LRSt}}) = 0$.

Il est intéressant de voir comment, pour une fonction donnée, on peut essayer de minimiser la variance de $\hat{\mu}_{\mathrm{LRSt}}$. Il faut trouver la base qui minimise

$$\frac{1}{N}\sum_{i=1}^{N}\left(\int_{\mathbf{x}\in C_i} f^2(\mathbf{x})d\mathbf{x} - \mu_i^2\right).$$

Voici quelques arguments heuristiques expliquant comment on peut essayer de réaliser cela : si on suppose que la fonction est telle que $|f(\mathbf{x}) - f(\mathbf{y})|$ croît avec $\|\mathbf{x} - \mathbf{y}\|_2$ (e.g., si f est monotone par rapport à chacun de ses arguments), alors on voudra partitionner $[0,1)^s$ en cellules qui minimisent la distance maximale $\|\mathbf{x} - \mathbf{y}\|_2$ que l'on peut obtenir entre deux points $\mathbf{x}, \mathbf{y}$ d'une cellule. Pour faire cela en pratique, on peut utiliser la base réduite de Minkowski. En effet, cette base est construite en choisissant $\mathbf{v}_1$ égal au plus court vecteur du réseau, puis en ajoutant successivement les vecteurs $\mathbf{v}$ les plus courts qui permettent à l'ensemble $\{\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}\}$ d'être complété en une base pour le réseau, pour $i = 2, \ldots, s$ (voir [77] pour plus de détails sur ce type de base). En gros, cette construction produit le parallélépipède fondamental qui est le plus "orthogonal" possible et ainsi, on satisfait (approximativement) notre critère de minimiser la distance maximale pouvant être obtenue à l'intérieur d'une cellule. D'un autre côté, si on sait que la fonction f est particulièrement variable par rapport à certaines dimensions, on préférera choisir une base qui fait en sorte que l'intervalle couvert par une cellule soit minimisé dans ces dimensions.

Bien sûr, en comparaison avec $\hat{\mu}_{\mathrm{LR}}$, cet estimateur nécessite un temps de calcul supérieur, puisque l'on doit générer N vecteurs aléatoires plutôt qu'un seul. En effet, chaque $\mathbf{w}_i$ est généré à partir d'un $\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,s}) \in [0,1)^s$ en posant

$$w_{ij} = \left(x_{ij} + \sum_{k=1}^{s} u_{i,k}v_{kj}\right) \bmod 1, \qquad i = 1, \ldots, N, j = 1, \ldots, s.$$

Donc, en excluant le travail requis pour déterminer la base $\{\mathbf{v}_1, \ldots, \mathbf{v}_s\}$, $\hat{\mu}_{\mathrm{LRSt}}$ nécessite un temps de calcul égal à celui requis pour évaluer $\hat{\mu}_{\mathrm{LR}}$ plus le temps requis pour générer $N-1$ vecteurs (pseudo-)aléatoires dans $[0,1)^s$. Remarquons cependant que le nombre de vecteurs aléatoires à générer est le même que pour l'estimateur $\hat{\mu}_{\mathrm{MC}}$, mais on a des additions modulo 1 à effectuer en plus.

2.5 Résultats numériques

Dans cette section, nous illustrons à l'aide de trois exemples comment les règles de réseau randomisées peuvent réduire empiriquement la variance par rapport à la méthode MC. Dans tous les exemples, les variables aléatoires sont générées par inversion et donc, la dimension du problème est égale au nombre de variables aléatoires qui doivent être générées dans une simulation. Chacun des exemples correspond à un problème de dimension petite, moyenne et grande, respectivement. Dans le premier exemple, nous comparons entre autres les estimateurs $\hat{\mu}_{\mathrm{LR}}$ et $\hat{\mu}_{\mathrm{LRp}}$, c.-à-d., la randomisation par translation aléatoire avec celle par permutation discutée à la section 2.3. Dans la deuxième partie du second exemple, la randomisation par translation aléatoire est comparée avec la stratification, telle que discutée à la section 2.4.

De façon générale dans cet ouvrage, notre façon de comparer les règles de réseau avec la méthode MC est de donner les facteurs de réduction de variance estimés par rapport à la méthode MC. Ces facteurs de réduction de variance empiriques peuvent être utilisés pour calculer un intervalle de confiance approximatif sur le rapport des variances théoriques, en utilisant la statistique de Fischer [68]. Par exemple, au niveau de confiance 98%, si la variance de la règle de réseau translatée est estimée à l'aide de 100 translations, alors $(F^{-1}_{0.99;100N-1,100}\hat{R}, F_{0.99;100,100N-1}\hat{R})$ est un intervalle de confiance approximatif pour le rapport $R = \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})/\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$. Les valeurs de la statistique de Fischer dans ce cas sont $F^{-1}_{0.99;100N-1,100} \approx F^{-1}_{0.99;\infty,100} = 1/1.43$ et $F_{0.99;100,100N-1} \approx F_{0.99;100,\infty} = 1.36$. Nous allons utiliser cela dans ce qui suit pour donner une idée de la précision des résultats présentés dans cette section. Notons que le fait que $F_{0.99;\infty,100}$ soit égal à 1.43 signifie que lorsqu'un rapport est supérieur à 1.43, on peut dire qu'il y a une différence significative entre les deux variances au niveau 98%.

2.5.1 Réseau stochastique d'activités

Cet exemple est tiré de [5]. On considère un *réseau stochastique d'activités* (RSA), qui est représenté par un graphe orienté acyclique $(\mathcal{N}, \mathcal{A})$, où $\mathcal{N}$ est un ensemble de *noeuds* qui contient une source et une destination et $\mathcal{A}$ est l'ensemble d'arcs correspondant aux *activités*. La figure 2.6 illustre un exemple de RSA. Chaque activité k dans $\mathcal{A}$ possède une durée aléatoire V_k ayant la fonction de répartition $F_k(\cdot)$, $k = 1, \ldots, |\mathcal{A}|$. Certaines *activités-bidon* représentent des relations de précédence et ont une durée nulle. On dénote par $N(A)$ le nombre d'activités ayant une durée non nulle, $N(P)$ représente le nombre de chemins orientés de la source à la destination et $C_j \subseteq \mathcal{A}$ est l'ensemble d'activités formant le chemin j, pour $1 \leq j \leq N(P)$. Le *temps de complétion du réseau*, dénoté par T, est la longueur du plus long chemin allant de la source à la destination.

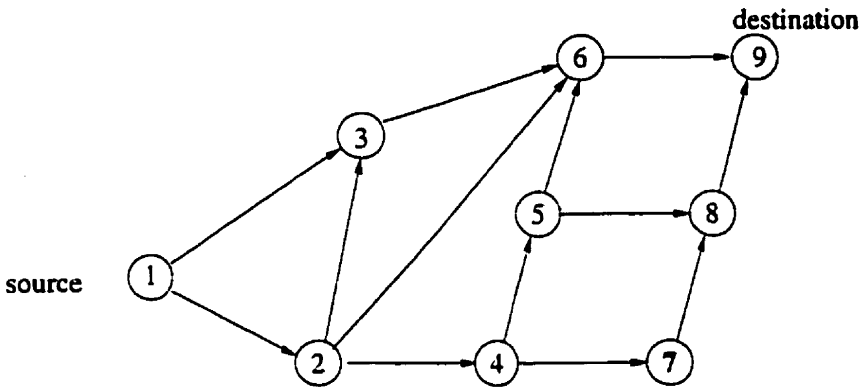


FIGURE 2.6: Exemple d'un RSA, tiré de [5]

On veut estimer $\mu = F_T(x) = \Pr[T \leq x]$ pour un seuil donné x. Si on utilise la méthode standard, que ce soit avec MC ou QMC, la dimension du problème est $s = N(A)$, puisqu'une variable uniforme est requise pour générer la durée de chaque activité, à l'aide de la relation $V_k = F_k^{-1}(u_k)$. On peut donc écrire μ comme étant l'intégrale

$$\mu = F_T(x) = \int_{[0,1)^{N(A)}} \prod_{j=1}^{N(P)} \mathbf{1}\left[\sum_{k \in C_j} F_k^{-1}(u_k) \leq x\right] du_1 \ldots du_{N(A)}.$$

La dimension du problème et la variance de l'estimateur MC peuvent être toutes deux réduites en appliquant la *méthode Monte Carlo conditionnelle* (MCC), comme suit [5] :

on doit choisir un sous-ensemble d'activités $\mathcal{L} \subseteq \mathcal{A}$ tel que chaque chemin orienté j de la source à la destination contient exactement une activité l_j faisant partie de $\mathcal{L}$. Un tel ensemble est appelé *ensemble de coupe uniforme (uniformly directed cutset)*. L'idée de MCC est de générer seulement la durée des activités dans $\mathcal{B} = \mathcal{A} \setminus \mathcal{L}$ et d'estimer μ par la probabilité conditionnelle que T soit inférieur ou égal à x étant donné ces durées. La dimension du problème est ainsi réduite à $s = N(B)$, où $N(B)$ est le nombre d'activités non bidon dans $\mathcal{B}$. L'estimateur MCC est donné par

$$Y = \Pr[T \leq x \mid \{V_j,\, j \in \mathcal{B}\}] = \prod_{l \in \mathcal{L}} F_l \left[\min_{\{j=1,\ldots,N(P):\, l_j = l\}} \left(x - \sum_{k \in C_j \setminus \{l_j\}} V_k \right) \right].$$

Cet estimateur est sans biais, car $\mathrm{E}[\Pr(T \leq x \mid \{V_j,\, j \in \mathcal{B}\})] = \Pr(T \leq x)$ (ceci vient du fait que $\mathrm{E}(\mathrm{E}(X \mid Y)) = \mathrm{E}(X)$ [68]). Remarquons que cette méthodologie pourrait être appliquée pour estimer la longueur du plus long chemin dans un réseau en général, qui n'est pas nécessairement un RSA.

Dans les résultats numériques qui suivent, les s nombres uniformes requis pour chaque simulation sont générés soit par la méthode MC, soit par une règle de réseau randomisée. Dans [5], on propose de les générer à l'aide de la méthode LHS.

Nous avons fait des expériences sur le réseau donné à la figure 2.6 (réseau 1 dans [5, 6]), afin de comparer MC, LHS, une règle de réseau translatée aléatoirement (LR) et une règle de réseau permutée (LRp), dont l'estimateur est défini à l'équation (2.30) de la page 55, avec ou sans MCC. Les règles de réseau ont été choisies à l'aide du critère $M_{32,24,12,8}$ (que l'on définit à la section 3.5) et se trouvent au tableau 3.2. Pour la durée des activités, nous avons utilisé les mêmes lois que dans [6] : les activités (1, 2), (1, 3), (2, 4), (6, 9) et (7, 8) sont de loi normale avec $\mu = 13.0$, 5.5, 5.2, 3.2 et 3.2, respectivement et $\sigma = \mu/4$ dans chaque cas. La loi normale est ajustée de façon à ce que la densité associée aux valeurs négatives soit reportée en une masse de probabilité en 0. Les activités (2, 3), (2, 6), (3, 6), (4, 5), (4, 7), (5, 6), (5, 8) et (8, 9) sont de loi exponentielle de moyenne 7.0, 16.5, 14.7, 6.0, 10.3, 20.0, 4.0 et 16.5, respectivement.

L'ensemble $\mathcal{L}$ contient les cinq arcs $\{(3,6),(2,6),(5,6),(5,8),(4,7)\}$ [5]. La dimension du problème est donc $s = 13$ sans MCC et $s = 8$ avec MCC. Nous avons choisi un seuil x égal à 90, ce qui implique que $F_T(x) \approx 0.89$, tel que calculé par notre es-

timateur ayant la plus petite variance empirique. Pour LR, LRp et LHS, nous avons utilisé différentes valeurs pour le nombre de points N et $m = 100$ copies de la règle translatée. Pour la méthode MC, nous avons fait mN répétitions i.i.d., afin que la comparaison soit juste. Le tableau 2.1 donne les facteurs de réduction de variance estimés par rapport à l'estimateur MC naïf, c.-à-d., sans MCC.

TABLEAU 2.1: Facteurs de réduction de variance estimés pour l'exemple du RSA

		N		
Méthode	s	4093	16381	65521
LHS	13	3.2	4.3	3.4
LRp	13	3.1	3.0	3.4
LR	13	6.2	4.2	25
MCC	8	4.1	4.1	4.1
LHS+MCC	8	58	56	63
LRp+MCC	8	54	61	57
LR+MCC	8	268	839	3086

La combinaison de LR avec MCC (dernière ligne) est décidément gagnante pour ce problème. De plus, les facteurs de réduction de variance augmentent quasiment de façon linéaire avec N : nous avons estimé le facteur de réduction de variance en continuant à quadrupler (environ) le nombre de points et avons obtenu 6423 (avec $N = 262139$ points), ce qui confirme cette tendance. Autrement dit, la variance de l'estimateur LR+MCC semble avoir un taux de convergence dans l'ordre de N^{-2}. Le temps de calcul de cet estimateur est inférieur à celui de MC pour une taille d'échantillon équivalente, puisque LR et MCC réduisent tous deux l'effort de calcul, en plus de réduire la variance. La combinaison de LHS avec MCC réduit la variance par des facteurs non négligeables, tout comme celle combinant LRp et MCC. En fait, LHS et LRp donnent environ les mêmes facteurs. Ainsi, du point de vue de l'efficacité des méthodes, LRp est mieux que LHS et LR est doublement gagnante, puisqu'elle est plus rapide que ces deux méthodes, en plus d'avoir une variance inférieure. Le tableau 2.2 donne les rapports de temps CPU de chacune de ces trois méthodes par rapport à l'estimateur MC naïf (sans MCC). Nous avons effectué les mêmes expériences pour différentes valeurs de x ainsi que pour l'autre réseau donné dans [5] et les conclusions étaient semblables.

TABLEAU 2.2: Rapports CPU des différentes méthodes

	N		
Méthode	4093	16381	65521
LHS	2.2	2.3	2.3
LRp	1.8	1.9	2.0
LR	0.8	0.8	0.8

Simplement pour donner une idée de la précision des facteurs donnés au tableau 2.1, les intervalles de confiance au niveau 98% pour les rapports de la première ligne (LHS) sont (2.2, 4.4), (3.0, 5.8) et (2.4, 4.6), respectivement. Ceux pour les rapports de la dernière ligne sont (187, 364), (586, 1141) et (2158, 4197), respectivement. Ainsi, on peut dire que même en tenant compte du bruit sur ces facteurs de réduction de variance empiriques, la méthode LR+MCC demeure significativement supérieure aux autres qui sont données au tableau 2.1.

2.5.2 Options asiatiques

Combinaison de LR translatée avec d'autres méthodes de réduction de la variance

Considérons le problème où l'on veut évaluer le prix d'une option asiatique sur la moyenne arithmétique, pour un actif dont la valeur au temps u est dénoté par $S(u)$. On suppose que l'évolution de $S(u)$ obéit au modèle de Black-Scholes [10], avec un taux d'intérêt sans risque r et une volatilité de σ. Sous la mesure *neutre au risque*, cela implique que $S(\cdot)$ satisfait l'équation différentielle stochastique

$$dS(\zeta) = rS(\zeta)d\zeta + \sigma S(\zeta)dB(\zeta), \tag{2.33}$$

où $B(\cdot)$ est un mouvement brownien standard. Pour plus de détails sur ce modèle, voir [24] ou [56]. La solution de l'équation (2.33) est

$$S(\zeta) = S(0)\exp\left[(r - \sigma^2/2)\zeta + \sigma B(\zeta)\right].$$

Le prix d'exercice de l'option est dénoté par K et T représente sa date d'expiration. La valeur finale de l'option est donnée par $\max(0, (1/s)\sum_{i=1}^{s} S(t_i) - K)$, où $t_i =$

$T_1 + i(T - T_1)/s$, T_1 est la période de temps qui s'écoule avant que l'on observe les prix qui entrent dans la moyenne et s est le nombre de prix entrant dans la moyenne. On doit donc générer une trajectoire de $B(\cdot)$ en déterminant sa valeur aux points $t_1, \ldots, t_s$ et ceci peut être accompli en générant s variables aléatoires i.i.d. suivant la loi normale standard, puisque $B(t_i) - B(t_{i-1}) \sim N(0, \Delta)$, où $\Delta = (T - T_1)/s$. L'espérance de la valeur finale, qui correspond au prix que l'on veut estimer, peut être écrite à l'aide de l'intégrale en s dimensions suivante :

$$\mu = \int_{[0,1)^s} e^{-rT} \max\left(0, \frac{1}{s}\sum_{i=1}^{s} S(0) \exp\left[\left(r - \frac{\sigma^2}{2}\right) t_i + \sigma\sqrt{\Delta}\sum_{j=1}^{i} \Phi^{-1}(u_j)\right] - K\right) d\mathbf{u}, \tag{2.34}$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale standard.

Afin de réduire la variance, on peut utiliser le prix de l'option sur la moyenne *géométrique* comme variable de contrôle (VC) [57], ainsi que des variables antithétiques. Des résultats numériques obtenus en combinant ces méthodes avec des règles de réseau translatées aléatoirement sont donnés dans [82, 84]. Dans [35], l'*importance sampling* (IS) et la stratification (STR) sont utilisés afin de réduire la variance et dans [132], la combinaison IS et VC est étudiée.

En simulation, la méthode IS permet de traiter les "événements rares". Ainsi, dans notre cas, elle est surtout utile lorsque l'option est "out-of-the-money", c.-à-d., lorsque $K > S(0)$. Elle consiste à générer les variables aléatoires de loi normale $Z_j = B(t_j) - B(t_{j-1})$ sous une loi de probabilité modifiée (en changeant les espérances $\mathrm{E}(Z_j)$) de façon à ce que la valeur finale de l'option prenne plus souvent une valeur non-nulle. Bien sûr, l'estimateur doit être multiplié par un *rapport de vraisemblance* [35, 70] approprié afin de demeurer sans biais. Notons que dans l'équation (2.34), les variables aléatoires normales Z_j sont écrites en fonction des uniformes u_j qui sont utilisées pour les générer par inversion, c.-à-d., on a $Z_j = \Phi^{-1}(u_j)$. De son côté, la stratification est appliquée sur la variable aléatoire $Y = \mathbf{a} \cdot (Z_1, \ldots, Z_s)$, où $\mathbf{a}$ est un vecteur "optimal" donné. Les Z_j sont ensuite générés en conditionnant sur Y.

Dans nos expériences, nous utilisons le changement de mesure proposé dans [35] quand nous appliquons IS. Une méthode alternative est décrite dans [132], qui est basée sur un estimé du gradient de la variance de l'estimateur et qui détermine le

changement de loi de probabilité au fur et à mesure que la simulation progresse.

Nous avons fait des expériences afin de comparer différentes combinaisons des méthodes ci-dessus et de leur association avec une règle de réseau translatée aléatoirement. On dénote par COND la méthode qui génère les Z_i en conditionnant sur Y, en prenant $\mathbf{a}$ égal au vecteur de tendance optimal pour la méthode IS, tel que suggéré dans [35] : ce vecteur comprend les espérances $\mathrm{E}(Z_j)$ sous la loi modifiée que l'on utilise pour IS et dépend des paramètres du problème. Lorsque nous combinons COND avec LR, la première coordonnée de chaque point translaté sert à générer $Y = \mathbf{a} \cdot (Z_1, \ldots, Z_s)$: c'est ce qui remplace la stratification. Les $s - 1$ coordonnées suivantes du point translaté sont suffisantes pour générer le vecteur $(Z_1, \ldots, Z_s)$, tel qu'expliqué dans [35]. Les méthodes IS et STR sont appliquées exactement de la même façon que dans [35]. Les règles de réseau ont été choisies à l'aide du critère $M_{32,24,12,8}$ et sont données au tableau 3.2.

Maintenant, voici comment les différentes méthodes de réduction de la variance sont combinées avec LR dans nos expériences. D'abord, on peut voir VC et IS comme étant deux façons de réécrire la fonction f que l'on intègre dans (2.34), qui se trouve à être la valeur finale de l'option actualisée au taux sans risque r. Par exemple, dans le cas de VC, $f(\mathbf{u})$ est remplacée par $f(\mathbf{u}) + \hat{\beta}(\mu_g - g(\mathbf{u}))$, où $g(\mathbf{u})$ est la valeur finale de l'option sur la moyenne géométrique actualisée au taux sans risque r (en utilisant le vecteur $\mathbf{u}$ pour générer les s normales requises dans le calcul de cette valeur finale), $\mu_g = \int_{[0,1)^s} g(\mathbf{u})d\mathbf{u}$ représente le prix exact de l'option sur la moyenne géométrique (que l'on peut calculer, voir [57]) et $\hat{\beta}$ est le coefficient habituel associé à la méthode de la variable de contrôle, que l'on peut estimer en utilisant les mêmes simulations [69, 70]. Lorsque l'on combine VC et IS, nous appliquons d'abord IS, car le changement de mesure proposé dans [35] a été défini pour fonctionner avec l'estimateur naïf et non avec l'estimateur contrôlé. Ce faisant, on obtient une nouvelle fonction à intégrer, sur laquelle on peut, ou non, appliquer une règle de réseau. Pour plus de détails sur la façon de combiner IS et CV dans le cas de l'estimateur MC, voir [132]. Finalement, on peut voir COND + LR simplement comme une alternative à LR pour générer les s normales requises pour simuler le mouvement brownien définissant la trajectoire de

l'actif sous-jacent, une fois VC et/ou IS appliqué(s) à la fonction.

TABLEAU 2.3: Facteurs de réduction de variance estimés pour l'exemple des options asiatiques

	N		
Méthode	4093	16381	65521
MC+IS+COND+STR	1502	1596	1598
LR	6.4	6.7	12
IS	12	12	12
VC	376	378	381
IS+LR	23	30	33
VC+LR	703	620	597
IS+VC	899	899	903
IS+VC+LR	1772	1895	1605
VC+IS+COND+LR	6092	6167	6858

Le tableau 2.3 contient les facteurs de réduction de variance estimés, pour certaines combinaisons de méthodes. Les paramètres de l'option sont $S(0) = 50, \sigma = 0.3$, $r = 0.05$, $K = 55$, $T = 1$ année, $T_1 = 0$ et $s = 64$. Parmi les combinaisons données dans le tableau (et toutes les autres que nous avons essayées), celle qui gagne est VC+IS+COND+LR. Cet estimateur est meilleur que la combinaison MC+IS+COND+STR de [35] par un facteur d'environ 4. LR seul est mieux que MC par un facteur d'au moins 6. Remarquons que VC+LR, qui est très simple, performe déjà plutôt bien et fait mieux que IS+LR, même si nous avons choisi un prix d'exercice K supérieur à $S(0)$, ce qui devrait en principe favoriser IS par rapport à VC. En effet, d'autres expériences (telles que rapportées dans [82, 84]) nous indiquent que l'efficacité de VC décroît généralement avec K et s, alors que celle d'IS augmente avec K, tel qu'expliqué dans [132, 35]. De son côté, LR seule a la même tendance que VC, c.-à-d., son efficacité décroît avec K. Cela est dû au fait que lorsque K est grand, la fonction f que l'on intègre est nulle sur la plupart du domaine $[0,1)^s$ et donc, la bonne distribution des points de la règle n'est pas très utile. Ces deux tendances complémentaires (celle de VC ou LR et celle de IS) semblent créer un effet positif lorsque IS et VC sont combinées et de même pour la combinaison IS + LR. De plus, en utilisant les trois méthodes (IS+VC+LR), on fait mieux que n'importe quelle combinaison de deux.

Encore une fois, on peut calculer des intervalles de confiance au niveau 98% pour les facteurs théoriques de réduction de variance en utilisant les rapports empiriques du tableau 2.3. Par exemple, on obtient pour la méthode MC+IS+COND+STR les intervalles (1050, 2043), (1116, 2171) et (1117, 2173). Pour la méthode LR, on obtient (4.5, 8.7), (4.7, 9.1) et (8.4, 16) et pour la combinaison gagnante VC+IS+COND+LR, on obtient (4260, 8285), (4313, 8387) et (4796, 9327).

Comparaison de LR stratifiée avec LR translatée

Dans le tableau suivant, nous donnons les facteurs de réduction de variance obtenus en utilisant l'estimateur $\hat{\mu}_{\text{LRSt}}$: cette méthode est dénotée par "LRSt" dans le tableau. Pour construire la base, nous avons utilisé le logiciel `LatMRG` [76] : nous avons pris les vecteurs formant la base réduite de Minkowski en dimension s (nous avons introduit ce concept à la page 63 ; voir [77] pour plus de détails). Nous donnons aussi les facteurs de réduction obtenus avec la règle de réseau qui a servi à construire $\hat{\mu}_{\text{LRSt}}$, mais qui est simplement translatée aléatoirement : cette méthode est dénotée par "LR" dans le tableau 2.4. Les règles de réseau ont été choisies à l'aide du critère M_8 (dont la définition se trouve à l'équation (2.13), page 35) et se trouvent dans [75]. La colonne "CPU" indique le rapport entre le temps moyen requis pour calculer l'estimateur $\hat{\mu}_{\text{LRSt}}$ (ou $\hat{\mu}_{\text{LR}}$, pour la méthode LR) et celui requis pour $\hat{\mu}_{\text{MC}}$. Lorsque des variables antithétiques et la variable de contrôle mentionnée précédemment sont utilisées, nous dénotons cette méthode par "ACV" et les facteurs sont donnés par rapport à l'estimateur MC qui utilise ces deux techniques. Sinon, on a l'estimateur "naïf". Les paramètres du modèle pour l'option sont $S(0) = 100$, $r = \ln 1.09$, $\sigma = 0.2$, $T = 120$ jours et $T_1 = T - s$ jours.

Ces résultats nous indiquent que la stratification n'est pas très payante pour cet exemple : la variance empirique n'est jamais significativement supérieure à celle de l'estimateur MC, mais on ne gagne jamais non plus par des facteurs importants (au-dessus de 1.8). En tenant compte du temps supplémentaire requis pour calculer $\hat{\mu}_{\text{LRSt}}$ par rapport à la méthode MC, on peut même dire que l'on perd pour ce qui est de l'efficacité. Nous avons essayé d'autres règles de réseau, mais sans plus de succès. En comparaison, l'estimateur $\hat{\mu}_{\text{LR}}$ obtient de très bons résultats et est plus rapide que MC.

TABLEAU 2.4: Facteurs de réduction de variance amenés par $\hat{\mu}_{\text{Str}}$ par rapport à $\hat{\mu}_{\text{LR}}$

				naïf				ACV			
s	N	Méth.	$K =$	90	100	110	CPU	90	100	110	CPU
10	1021	LRSt		1.6	1.4	0.98	2.5	1.3	0.96	1.2	2.4
	1021	LR		251	131	75	0.8	7.0	5.2	3.1	0.9
	4093	LRSt		1.6	1.8	1.5	2.5	0.93	0.92	0.96	2.4
	4093	LR		462	297	86	0.8	3.7	3.7	2.8	0.9
30	1021	LRSt		0.94	1.0	1.2	4.4	1.2	0.85	1.0	4.3
	1021	LR		169	56	23	0.8	3.1	2.5	2.4	0.9
	4093	LRSt		0.94	1.1	0.99	4.3	1.1	0.79	1.3	3.8
	4093	LR		224	94	44	0.8	3.2	3.1	2.5	0.8

Ainsi, même si l'estimateur $\hat{\mu}_{\text{LRSt}}$ est plus sûr théoriquement que $\hat{\mu}_{\text{LR}}$ (étant donné que la réduction de variance est garantie pour le premier), il semble qu'en pratique, il est possible d'obtenir des estimateurs ayant une bien plus petite variance en utilisant une règle de réseau translatée aléatoirement.

2.5.3 Probabilité de ruine

Nous décrivons d'abord le modèle, puis les trois estimateurs qui sont testés dans la deuxième partie afin de comparer l'utilisation d'une règle de réseau translatée aléatoirement avec la méthode MC.

Modèle et méthodes d'estimation

Une compagnie d'assurances reçoit des réclamations à des temps aléatoires et de taille aléatoire. Plus précisément, on dénote par $S(t) = \sum_{k=1}^{N(t)} Y_k$ le processus de réclamations accumulées, où $N(t)$ représente le nombre de réclamations reçues pendant l'intervalle de temps $(0, t]$ et Y_k est le montant de la k^e réclamation. En retour du paiement des réclamations, la compagnie reçoit des primes à un certain taux $c(\cdot)$, qui peut dépendre du processus de surplus $U(\cdot)$. Ainsi, on a $dU(t) = c(U(t))dt - dS(t)$ ou, de façon équivalente, $U(t) = \int_0^t c(U(v))dv - S(t)$. On cherche à évaluer la probabilité de ruine étant donné un surplus initial u,

$$\psi(u) = P(U(t) < 0 \text{ pour un certain } t \geq 0 \mid U(0) = u).$$

Nous allons faire l'hypothèse que $N(\cdot)$ est un processus de Poisson de taux λ et que les tailles des réclamations Y_k, $k = 1, 2, \ldots$, sont des variables aléatoires i.i.d. de moyenne β. De plus, on suppose que la fonction de prime est constante et que le processus de surplus gagne de l'intérêt à un taux δ. Ces deux hypothèses reviennent en fait à supposer que $c(x) = c + \delta x$, où c et δ sont deux constantes non négatives [34].

Habituellement, $\psi(u)$ est très petite : la ruine est un *événement rare.* Ainsi, l'approche naïve consistant à simuler le processus d'arrivée des réclamations jusqu'à ce que la ruine se produise risque d'être très inefficace. Comme approche alternative, on peut utiliser l'*importance sampling* pour faire les simulations et c'est une des trois méthodes que nous considérons dans les résultats numériques. On explique dans [2] comment les lois de probabilité doivent être changées dans le cas où $\delta = 0$. Dans le cas particulier où les tailles des réclamations sont de loi exponentielle, le changement de mesure optimal est connu [2]. Lorsque $\delta > 0$, le problème est plus compliqué, car on doit alors modifier de façon locale les paramètres de la fonction de répartition au cours de la simulation : voir [4] pour plus de détails et [131], qui présente des algorithmes permettant d'appliquer l'*importance sampling* dans ce contexte.

Afin d'estimer $\psi(u)$, on peut également utiliser la dualité entre $U(\cdot)$ et le *processus de stockage* $X(\cdot)$ défini par

$$dX(t) = \begin{cases} -c(X(t))dt + dS(t) & \text{si } X(t) > 0, \\ dS(t) & \text{si } X(t) = 0 \end{cases}$$

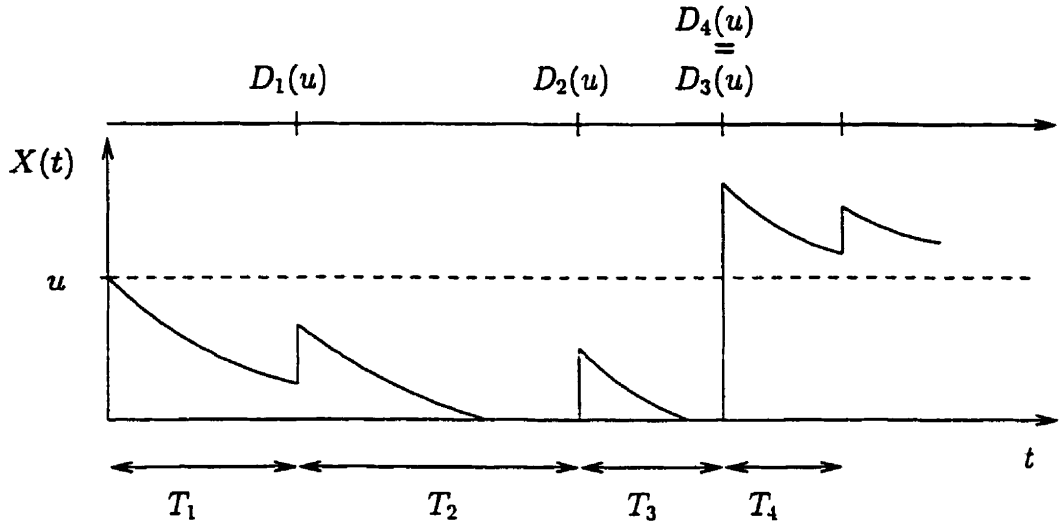
et une valeur initiale $X(0)$ posée égale à u. En supposant que $c > \lambda\beta$, ce processus est récurrent positif, ce qui signifie que l'espérance du temps qui s'écoule entre deux visites au point u est finie. Ainsi, on peut écrire [3] :

$$\psi(u) = 1 - \lim_{n\to\infty} \frac{D_n(u)}{\sum_{i=1}^{n} T_i},$$

où $D_n(u)$ est le temps total passé par $X(\cdot)$ sous le niveau u avant son n^e saut et T_i est le temps entre le $(i-1)^e$ saut et le i^e saut de $X(\cdot)$. La figure 2.7 illustre un exemple de trajectoire pour le processus $X(\cdot)$ ainsi que la valeur de $D_n(u)$ et T_i.

Donc, une deuxième méthode pour estimer $\psi(u)$ est de simuler l réclamations, puis

FIGURE 2.7: Exemple d'une trajectoire de $X(t)$



de calculer l'estimateur

$$\hat{\psi}(u) = 1 - \frac{D_l(u)}{\sum_{i=1}^{l} T_i}. \tag{2.35}$$

Cette approche est utilisée dans [92]. Le nombre l de réclamations doit être choisi suffisamment grand pour que la plupart du biais, qui décroît lentement, disparaisse. Toutefois, les propriétés de $X(\cdot)$ font en sorte que l'on peut utiliser la *méthode regénérative* pour estimer $\psi(u)$, comme cela est fait dans [131], par exemple. Les temps de regénération sont les instants où $X(t)$ atteint u. Avec cette troisième méthode, les simulations i.i.d. correspondent aux cycles de $X(\cdot)$, plutôt qu'à une suite de l réclamations. Cela signifie que dans l'estimateur (2.35), le nombre l devient aléatoire et correspond au nombre de réclamations dans le cycle regénératif simulé.

Pour chacune des trois approches, MC ou QMC peut être utilisé. En supposant que toutes les variables sont générées par inversion, on a besoin de deux $U(0,1)$ par réclamation : une pour la taille et une pour la durée interarrivée. Ainsi, la dimension du problème est égale à deux fois le nombre de réclamations par simulation. Avec l'*importance sampling* et l'approche regénérative, le nombre de réclamations est aléatoire et augmente (stochastiquement) avec u, pour une simulation donnée. Avec l'estimateur (2.35), ce nombre est égal à $2l$, mais la dimension effective (que nous avons brièvement définie à la page 8 de l'introduction) est probablement plus petite que cela, car les interactions qui contribuent le plus à la variance sont celles qui

ont lieu entre des variables associées à un même cycle regénératif. Remarquons que pour ce problème, il est très important d'utiliser des ensembles de points à faible discrépance pour lesquels les projections $P_N(I)$ sont bonnes de façon uniforme. En effet, puisque l'on peut associer chaque cycle de la simulation à un sous-ensemble I de coordonnées successives dans $\{1,\ldots,2l\}$, cela nous garantit que mêmes si les simulations sont longues (car l est grand), chaque cycle est simulé en utilisant un ensemble $P_N(I)$ bien distribué. Les ensembles de points *stationnaires dans la dimension* (voir page 12 de l'introduction) choisis à l'aide d'un critère qui regarde plusieurs projections ont cette propriété. Nous verrons à la section 3.2 que les règles de Korobov (voir section 2.1.2) avec N premier sont stationnaires dans la dimension. De plus, la structure de ces règles permet de traiter sans problème le cas où la dimension est aléatoire. Ceci est dû au fait que ce type de règle est déterminé par un seul paramètre, soit l'entier a définissant le vecteur générateur de la règle, une fois le nombre de points N choisi. Ainsi, un point $\mathbf{x}_i$ d'une règle de Korobov en t dimensions donné par $\mathbf{x}_i = (i-1)((1,a,\ldots,a^{t-1}) \bmod N)/N$ peut être "allongé" autant que l'on veut en utilisant la relation $x_{ij} = (i-1)(a^{j-1} \bmod N)/N$ afin de générer les coordonnées des dimensions $j = t+1, t+2, \ldots$. Nous utilisons donc ce type de règle dans nos expériences numériques.

Résultats numériques

Nous présentons maintenant les résultats obtenus à la suite d'expériences conduites afin de voir si les règles de réseau translatées aléatoirement pouvaient faire mieux que MC pour ce problème, dépendamment de l'approche choisie pour faire les simulations et des paramètres du modèle.

Nous avons utilisé une loi de probabilité exponentielle pour la taille des réclamations dans tous les exemples. Cette hypothèse n'est pas très réaliste, mais dans ce cas, $\psi(u)$ a une solution analytique [34], ce qui permet de comparer les résultats de la simulation avec les résultats théoriques et donc, on peut calculer le biais de chaque estimateur. On dénote par DUAL la méthode où l'on simule l réclamations et où l'on utilise l'estimateur (2.35), par REG, la méthode où la simulation regénérative est

utilisée et par IS celle où on utilise l'*importance sampling*.

Les tableaux donnent les facteurs de réduction de variance estimés par rapport à la méthode MC. Les règles de réseau ont été choisies à l'aide du critère M_8 (que l'on a défini à l'équation (2.13), page 35) : on peut les trouver dans [75]. Nous faisons toujours $m = 100$ translations aléatoires modulo 1 de la règle de réseau que l'on utilise et l'estimateur MC est basé sur mN répétitions i.i.d., afin que la comparaison soit juste. Dans le tableau 2.5, nous donnons aussi le rapport moyen du temps CPU requis pour calculer l'estimateur LR sur celui requis pour l'estimateur MC. Dans chaque tableau, nous ne nous occupons pas de comparer le biais (empirique) de l'estimateur LR avec celui de MC, car le biais est le même pour les deux méthodes. Par contre, une fois les tableaux présentés, nous ferons quelques remarques sur le biais associé à chacune des trois méthodes de simulation.

TABLEAU 2.5: Méthode DUAL

N	$u = 0$	$u = 10$
$\psi(u) \rightarrow$	0.841108	3.9123e-2
251	14	2.2
509	14	2.2
1021	18	2.3
2039	27	3.5
4093	28	3.6
8191	30	4.0
16381	52	4.8
CPU	0.74	0.73

Dans le tableau 2.5, LR réduit la variance de façon significative : par un facteur d'au moins 13 quand $u = 0$ et d'au moins 2 quand $u = 10$. En termes d'efficacité, LR a l'avantage supplémentaire de nécessiter moins de temps CPU que MC. Par exemple, quand $u = 0$ et $N = 16381$, LR améliore l'efficacité par un rapport de 70. Les facteurs augmentent avec N, mais pas aussi rapidement que dans l'exemple du réseau stochastique d'activités. Les facteurs de réduction de variance décroissent avec u, sans doute parce que la dimension effective est plus grande dans ce cas. Remarquons que la dimension du problème est égale à 8000 ici. Ainsi, il serait probablement impossible d'utiliser un (t, m, s)-réseau dans ce cas (ce type de méthode QMC a été mentionné

à la page 12 de l'introduction). En effet, les (t, m, s)-réseaux (avec $t > 0$) ne sont généralement pas stationnaires dans la dimension et de plus, ils sont habituellement construits d'une façon telle que les propriétés d'équidistribution des projections $P_N(I)$ se détériorent à mesure que I contient des indices qui deviennent de plus en plus près de s et donc, les dernières réclamations dans la simulation seraient intégrées par des points mal distribués.

Pour donner une idée de la précision des facteurs donnés au tableau 2.5, les intervalles de confiance au niveau 98% pour le rapport des variances théoriques associées à la première ligne ($N = 251$) sont (9.7, 19) et (1.6, 3.0) et ceux de la dernière ligne ($N = 16381$) sont (36, 71) et (3.3, 6.5).

Nous avons également effectué des expériences afin de voir comment les facteurs de réduction de variance étaient affectés par la valeur de l. Le tableau 2.6 résume ces expériences : on y indique, pour la règle de Korobov d'ordre N =251 qui est la meilleure par rapport à M_8, les facteurs de réduction de variance empiriques par rapport à la méthode MC, pour trois valeurs de l (5000, 10000 et 20000).

TABLEAU 2.6: Méthode DUAL, cas où $N < 2l$

N	l	$u = 0$	$u = 10$
	$\psi(u) \rightarrow$	0.841108	3.9123e-2
251	5000	11.4	2.7
251	10000	12.4	2.1
251	20000	12.2	2.1

On voit donc ici que pour un N fixé, si l augmente, les facteurs demeurent à peu près constants. Ceci confirme en quelque sorte notre hypothèse que la dimension effective du problème dépend de la longueur des cycles et est indépendante de l. Remarquons que dans le tableau 2.6, le nombre de points $N = 251$ est inférieur à la dimension $2l$ (10000, 20000 et 40000). Cela n'affecte pas la qualité de la méthode, car même si à la base, chaque simulation i utilise alors un point $(x_{i,1}, \ldots, x_{i,2l})$ dont les composantes se répètent, la translation aléatoire fait en sorte que ces $2l$ points deviennent des $U(0,1)$ i.i.d. et ainsi, la forte corrélation (qui serait destructive) qu'on avait à l'intérieur du point disparaît. Par contre, la corrélation entre les simulations (que l'on espère être

constructive) demeure malgré la randomisation, puisque la structure de réseau de P_N est préservée sous la translation.

TABLEAU 2.7: Méthode IS

N	Méthode	$u = 3.54$	$u = 29.6$
	$\psi(u) \to$	0.5	0.01
8191	MC	44	4278
	LR	59	4471
16381	MC	44	4278
	LR	79	5291
	r	26.7	174.5

Le tableau 2.7 donne les résultats obtenus avec l'*importance sampling* (IS). Le taux d'intérêt est fixé à 0, afin que le changement de mesure puisse facilement être fait. Les rapports sont donnés par rapport à la variance de l'estimateur MC sans IS. La dernière ligne donne le nombre moyen de réclamations par simulation. Ce nombre, multiplié par deux, nous donne une idée de la dimension du problème pour ce type de simulation. Pour cet exemple simplifié, l'IS réduit la variance par des facteurs importants (première ligne pour chaque valeur de N) : cela rend la tâche plus ardue pour que LR amène une amélioration additionnelle. Malgré cela, LR réussit quand même à aller chercher une réduction de variance supplémentaire, allant jusqu'à un facteur de 1.8 quand $\psi(u) = 0.5$ et $N = 16381$. Comme prévu, LR réussit mieux quand u est plus petit. Ceci est intéressant, car IS a la propriété inverse : il améliore davantage par rapport à MC lorsque $\psi(u)$ diminue, ou, de façon équivalente, lorsque u augmente. Ainsi, LR et IS sont deux méthodes complémentaires qui semblent se combiner avantageusement, comme c'était le cas pour les options asiatiques (voir au tableau 2.3).

TABLEAU 2.8: Méthode REG

N	Méthode	$u = 0$	$u = 10$
	$\psi(u) \to$	0.841108	3.9123e-2
8191	LR	3.9	1.9
16381	LR	2.6	2.4
	r	7.3	42.7

Finalement, nous donnons les résultats obtenus en utilisant la méthode regénéra-

tive au tableau 2.8. Dans ce cas-ci, LR réduit la variance par des facteurs qui varient entre 2 et 4. Lorsque $\delta = 0$, il est connu que IS peut donner des estimateurs plus précis que la méthode regénérative [2]. Toutefois, lorsque $\delta > 0$, IS est plus difficile à appliquer, alors que la méthode regénérative fonctionne bien pour n'importe quelle valeur de δ.

On peut donc se demander laquelle parmi les méthodes DUAL et REG est la plus appropriée lorsque IS est compliqué à utiliser. Si on compare ces deux méthodes du point de vue du biais, celui dans REG n'est pas aussi persistant qu'avec DUAL : par exemple, quand $u = 10$, on sait que $\psi(10) = 3.9123\text{e-}2$ et l'estimateur obtenu avec REG vaut $\hat{\psi}(10) = 3.92\text{e-}2$, alors qu'avec la méthode DUAL, on obtient $\hat{\psi}(10) = 3.98\text{e-}2$. C'est d'autant plus décevant que ce dernier est basé sur beaucoup plus d'observations ($4000/43 \approx 93$ fois plus).

En ce qui concerne la variance, il faut faire attention lorsque l'on compare les estimateurs, puisque les méthodes n'utilisent pas le même nombre de réclamations par simulation, même une fois les paramètres du modèle fixés. Pour faire des comparaisons justes, une possibilité est de multiplier la variance par le nombre de réclamations par simulation. En faisant cela pour comparer DUAL et REG, on se rend compte que même si $\hat{\sigma}^2$ est beaucoup plus petit avec DUAL (1.90e-12, au lieu de 2.46e-8 pour REG), après normalisation, les deux variances sont quasiment égales (2.82e-7 et 2.86e-7, respectivement).

Ainsi, lorsque IS ne peut être appliqué aisément, la méthode regénérative semble être plus appropriée. Par contre, si le modèle ne donne pas lieu à un processus regénératif, nous n'avons pas vraiment d'autre choix que d'utiliser DUAL.

Chapitre 3

Critères de sélection et projections sur les sous-espaces de l'hypercube

Dans ce chapitre, nous regardons comment les propriétés des projections de la règle de réseau choisie et de la fonction à intégrer peuvent nous informer sur le comportement de la variance de l'estimateur $\hat{\mu}_{\mathrm{LR}}$ défini en (2.15), page 37, obtenu par translation aléatoire. Cette étude nous conduit à la définition d'un nouveau critère de sélection basé sur le test spectral pour choisir les règles de réseau. Pour parvenir à ce but, nous faisons d'abord des rappels sur la décomposition ANOVA d'une fonction, car nous utilisons cette décomposition afin de justifier la forme de notre critère de sélection. Nous revenons aussi sur le concept d'ensemble de points stationnaire dans la dimension et de règles de réseau (complètement) projection-régulières, car plusieurs résultats présentés dans ce chapitre supposent que les règles ont ces propriétés. Ensuite, nous expliquons le lien entre la décomposition ANOVA d'une fonction et celle en série de Fourier, ce qui nous permet de relier notre critère de sélection avec l'expression (2.16) pour la variance de $\hat{\mu}_{\mathrm{LR}}$, puis nous donnons des résultats sur la variance de l'estimateur $\hat{\mu}_{\mathrm{LR}}$ lorsque la fonction est polynômiale.

3.1 Décomposition ANOVA

La décomposition d'une fonction en composantes ANOVA [51, 25] a été souvent utilisée [44, 45, 46, 105, 106, 107] récemment pour tenter d'expliquer le comportement des méthodes QMC, notamment en ce qui a trait à l'intégration de fonctions en très grande dimension. Motivé par des exemples numériques en 360 dimensions reliés à l'évaluation de produits financiers, Paskov et Traub [110] ont d'abord introduit le concept de *dimension effective* d'une fonction f, qui correspond en gros au nombre de variables qui suffisent pour expliquer la majeure partie de la variabilité de f. Nous donnons une définition plus précise tirée de [14, 46] aux équations (3.4) et (3.5). Parallèlement, Owen a utilisé dans [105, 106] la décomposition ANOVA d'une fonction pour démontrer des résultats sur la variance des (t, m, s)-réseaux brouillés et dans [107] pour montrer comment optimiser sa méthode de l'*échantillonnage du supercube latin* (LSS). Hickernell [44, 45] a aussi utilisé ce concept pour proposer de nouvelles mesures de discrépance et donc, de nouvelles bornes sur l'erreur d'intégration.

Dans cet ouvrage, la décomposition ANOVA d'une fonction sera utilisée pour justifier la forme du critère de sélection que nous proposons à la section 3.5. L'idée est la suivante : cette décomposition réécrit la fonction $f : [0,1)^s \to \mathbb{R}$ comme une somme de 2^s fonctions f_I associées à chacun des sous-ensembles $I \subseteq S = \{1, \ldots, s\}$. Chacune des composantes f_I a une variance associée σ_I^2 et les rapports σ_I^2/σ^2 peuvent être utilisés afin de quantifier l'importance de chaque composante f_I, puisqu'on peut montrer que $\mathrm{Var}(f) = \sigma^2 = \sum_I \sigma_I^2$. Or, on sait qu'à mesure que la dimension augmente, il devient de plus en plus difficile de trouver un ensemble de points P_N tel que chaque projection $P_N(I)$ est bonne et donc, dans ce cas, on peut utiliser les rapports σ_I^2/σ^2 afin de nous indiquer quelles sont les projections $P_N(I)$ pour lesquelles on devrait s'assurer d'avoir une bonne distribution, c.-à-d., qui devraient être considérées dans notre critère de sélection.

Les deux prochaines sous-sections donnent de façon formelle les définitions associées à la décomposition ANOVA et les références utilisées pour cette partie sont [46, 107].

3.1.1 Composantes ANOVA de f

Définition 3.1.1 *([46, 107]) Soit $f \in \mathcal{L}^2$. La* décomposition ANOVA de f *est donnée par*

$$f(\mathbf{x}) = \sum_{I \subseteq S} f_I(\mathbf{x}),$$

où $f_I(\mathbf{x})$ est définie par

$$f_I(\mathbf{x}) = \int_{[0,1)^{I^c}} f(\mathbf{x}) d\mathbf{x}_{I^c} - \sum_{J \subset I} f_J(\mathbf{x}),$$

et I^c est le complément de I dans S.

Donc, la composante ANOVA f_I correspond à la partie de f qui dépend seulement de $\mathbf{x}_I$, où $\mathbf{x}_I = (x_j)_{j \in I}$. De plus, cette décomposition est telle que

$$\int_0^1 f_I(\mathbf{x}_I) dx_j = 0 \text{ pour tout } j \in I \tag{3.1}$$

ce qui signifie que $\int_{[0,1)^s} f_I(\mathbf{x}) d\mathbf{x} = 0$ si I est non vide et $\int_{[0,1)^s} f_\emptyset(\mathbf{x}) d\mathbf{x} = \mu$. De plus, si $I \neq J$, alors $\int_{[0,1)^s} f_I(\mathbf{x}) f_J(\mathbf{x}) d\mathbf{x} = 0$. Ainsi, si on pose

$$\sigma_I^2 = \mathrm{Var}(f_I(\mathbf{x})) = \int_{[0,1)^s} f_I^2(\mathbf{x}) d\mathbf{x}, \tag{3.2}$$

alors on a que

$$\sigma^2 = \sum_{\emptyset \neq I \subseteq S} \sigma_I^2. \tag{3.3}$$

L'importance relative des σ_I^2 nous indique quelles variables ou quels sous-ensembles de variables sont les plus importantes [46]. Dans cet ouvrage, c'est davantage la décomposition de la variance σ^2 telle que définie par (3.3) que nous utilisons, plutôt que celle qui réécrit f comme on le fait à la définition 3.1.1.

3.1.2 Dimension effective

On dit qu'une fonction est *de dimension tronquée s_t en proportion p_t* [14, 46] si

$$\sum_{I \subseteq \{1,2,\ldots,s_t\}} \sigma_I^2 = p_t \sigma^2 \tag{3.4}$$

et elle est *de dimension superposée s_s en proportion p_s* [14, 46] si

$$\sum_{I : |I| \leq s_s} \sigma_I^2 = p_s \sigma^2. \tag{3.5}$$

Dans le premier cas, on a que les s_t premières variables contribuent à $100p_t\%$ de la variance de f, alors que dans le second, ce sont les sous-ensembles comptant au plus s_s variables qui contribuent à $100p_s\%$ de la variance de f. Ces définitions laissent entendre que s_t (ou s_s) sont fixées à l'avance et qu'on trouve le p_t (p_s) correspondant. De façon équivalente, on pourrait d'abord fixer p_t (p_s), remplacer les égalités par des $\geq$ dans les définitions et trouver le s_t (s_s) minimal satisfaisant l'inégalité.

Dans [85], nous avons défini un autre concept de dimension effective, basé sur l'idée que certaines fonctions ont des composantes f_I qui sont plus importantes lorsque le sous-ensemble I est composé d'indices successifs. Plus précisément, on dit qu'une fonction est *de dimension successive superposée s_u en proportion p_u* si

$$\sum_{I \subseteq \tilde{S}(s_u,1)} \sigma_I^2 = p_u \sigma^2, \tag{3.6}$$

où $\tilde{S}(s_u, 1)$ est l'ensemble des I à indices successifs, tels que $1 \leq |I| \leq s_u$.

Ici, ce sont les sous-ensembles comprenant des indices successifs et n'ayant pas plus que s_u éléments qui contribuent à $100p_u\%$ de la variance totale de f. Contrairement aux deux définitions précédentes, en prenant p_u égal à 1, il n'est pas certain que l'on puisse trouver un entier s_u entre 1 et s qui satisfasse la définition, puisque même lorsque s_u est égal à s, on ne considère pas tous les sous-ensembles $I \subseteq S$ quand on fait la somme des σ_I^2.

Afin de motiver la définition de dimension successive superposée, revenons à l'exemple de la probabilité de ruine, vu à la section 2.5.3. Dans le cas où on utilise l'estimateur DUAL, dans lequel on a l réclamations par simulation, on peut supposer qu'en posant s_u égal au double de l'espérance du nombre de réclamations par cycle regénératif, la relation (3.6) devrait être satisfaite pour une valeur p_u assez près de 1, puisque les interactions les plus importantes entre les $2l$ variables aléatoires de la simulation se font à l'intérieur des cycles regénératifs.

3.2 Propriétés des projections de l'ensemble de points P_N

Dans cette section, nous étudions deux propriétés importantes qui décrivent le comportement des projections $P_N(I)$ d'un ensemble de points P_N. Donnons d'abord

la définition de stationnarité dans la dimension. Dans ce qui suit, $\tilde{I} = \{1, i_2 - i_1 + 1, \ldots, i_t - i_1 + 1\}$ lorsque $I = \{i_1, \ldots, i_t\}$.

Définition 3.2.1 *Un ensemble de points P_N est* stationnaire dans la dimension *si, pour tout $I, J \subseteq S$ tels que $\tilde{I} = \tilde{J}$, on a que*

$$P_N(I) = P_N(J).$$

La stationnarité dans la dimension d'un ensemble de points P_N nous garantit l'uniformité de la qualité de P_N à mesure que l'on "avance" de la dimension 1 à la dimension s. Dans le contexte de la simulation, cette propriété est particulièrement importante si on doit faire de longues simulations pour lesquelles la mesure de performance à estimer dépend de façon assez uniforme de tous les événements de la simulation, comme nous l'avons vu pour l'exemple de la probabilité de ruine à la section 2.5.3, dans le cas où on a l réclamations par simulation (estimateur DUAL).

Dans [85], nous démontrons que les ensembles de points P_N obtenus à partir d'un générateur pseudo-aléatoire pour lequel la récurrence définissant la fonction de transition est inversible ont cette propriété. Plus précisément :

Proposition 3.2.1 *[85, Proposition 2] Étant donné une récurrence de la forme $\xi_i = \tau(\xi_{i-1})$, où $\tau : \Xi \to \Xi$ et Ξ est un ensemble fini, si $\tau(\cdot)$ est inversible et $g : \Xi \to [0, 1)$ est une fonction de sortie donnée, alors $P_N = \{(g(\xi_0), \ldots, g(\xi_{s-1})) : \xi_0 \in \Xi\}$, l'ensemble de tous les vecteurs (se chevauchant) sur tous les cycles de la récurrence, est un ensemble de points stationnaire dans la dimension.*

Une condition équivalente à celle que $\tau(\cdot)$ soit inversible est de demander que chaque cycle déterminé par la récurrence soit purement périodique. Cela signifie que pour tout germe $\xi_0 \in \Xi$, il existe un entier ρ égal à la période de la suite $\xi_0, \xi_1, \ldots$ et tel que $\xi_i = \xi_{i+\rho}$ pour tout $i \geq 0$.

Comment savoir si $\tau(\cdot)$ est inversible ? Dans [86, *Theorem 6.11*], on dit que si Ξ est un corps fini, que $\tau(\cdot)$ est de la forme $\xi_{n+k} = b_{k-1}\xi_{n+k-1} + b_{k-2}\xi_{n+k-2} + \ldots + b_0\xi_n + b$, avec $b, b_0, \ldots, b_{k-1} \in \Xi$ et que b_0 est non nul, alors la suite $\xi_0, \xi_1, \ldots$ est purement périodique. Donc, si on prend un GCL avec un modulo N premier, alors pour n'importe quel multiplicateur $a \in [1, \ldots, N-1]$, tous les cycles engendrés par la récurrence

sont purement périodiques puisque dans ce cas, $\Xi = \mathbb{Z}_N$ est un corps fini, $k = 1$ et $b_0 = a \neq 0$. Si le GCL a un modulo N qui est une puissance de deux, on peut montrer qu'il suffit de prendre un multiplicateur a impair pour s'assurer que la fonction de transition soit inversible.

En ce qui a trait aux règles de réseau (complètement) projection-régulières, le résultat suivant donne des conditions nécessaires et suffisantes pour avoir ces propriétés dans le cas d'une règle de rang 1. Rappelons qu'une règle de rang 1 qui est complètement projection-régulière a la propriété que pour tout sous-ensemble $I \subseteq S$ non vide, la projection $P_N(I)$ contient N points distincts.

Proposition 3.2.2 *[80, Proposition 1] Une règle de rang 1 ayant le vecteur générateur* $\mathbf{z} = (z_1, \ldots, z_s)$ *est projection-régulière si et seulement si* $pgcd(N, z_1) = 1$. *Elle est complètement projection-régulière si et seulement si* $pgcd(N, z_d) = 1$ *pour* $1 \leq d \leq s$.

Donc, une règle de rang 1 d'ordre N premier est toujours complètement projection-régulière, puisque $\text{pgcd}(N, z) = 1$ pour n'importe quel $z \in [1, \ldots, N-1]$. Si N est une puissance de deux, il suffit de choisir des composantes z_d impaires pour le vecteur générateur.

Comme conséquence de ces deux propositions, on peut voir que toute règle de Korobov (voir la sous-section 2.1.2 pour la définition de ce type de règle) basée sur un vecteur générateur $((1, a, \ldots, a^{s-1}) \bmod N)$ est stationnaire dans la dimension et complètement projection-régulière si N est premier ou si N est une puissance de deux et que a est impair. Cela justifie le fait que l'on utilise ce type de règle dans nos exemples numériques, avec N premier ou une puissance de deux. De plus, le fait de se restreindre à ces règles simplifie les algorithmes de recherche visant à trouver la meilleure règle par rapport à un critère de sélection donné, puisque pour N fixé, chaque règle est complètement déterminée par le paramètre a. Rappelons également que les règles de rang supérieur à 1 ne peuvent avoir la propriété que *toutes* leurs projections $P_N(I)$ pour I non vide contiennent N points (voir page 29, avant la sous-section 2.1.4).

3.3 Décomposition ANOVA et série de Fourier

Dans cette section, nous relions la décomposition ANOVA d'une fonction à celle en série de Fourier, en regardant comment se décompose la variance dans chacun des cas. Ceci est exprimé par le résultat donné à la proposition 3.3.1, qui nous montre entre autres que l'on peut calculer les variances σ_I^2 définies en (3.2) à l'aide des coefficients de Fourier de f et donc, sans avoir à définir explicitement les composantes ANOVA f_I introduites à la définition 3.1.1. De plus, ce résultat va nous permettre de motiver la forme du critère de sélection qui sera présenté à la section 3.5 et qui consiste à utiliser le test spectral, que l'on a défini à la section 2.1.7, sur les projections $P_N(I)$ associées aux composantes f_I jugées importantes dans la décomposition ANOVA.

Pour démontrer la proposition 3.3.1, le résultat intermédiaire suivant est utilisé :

Lemme 3.3.1 *Soit $f \in \mathcal{L}^2$. Pour un vecteur $\mathbf{h} \in \mathbb{Z}^s$, si on dénote par $I_{\mathbf{h}}$ l'ensemble $\{j \in S : h_j \neq 0\}$, alors pour tout $I \subseteq S$ non vide, les coefficients de Fourier de la composante ANOVA f_I de f (voir définition 3.1.1) sont :*

$$\hat{f}_I(\mathbf{h}) = \begin{cases} \hat{f}(\mathbf{h}) & \text{si } I = I_{\mathbf{h}}, \\ 0 & \text{sinon .} \end{cases}$$

Démonstration : on a que

$$\begin{aligned} \hat{f}_I(\mathbf{h}) &= \int_{[0,1)^s} f_I(x_1,\ldots,x_s) e^{-2\pi\sqrt{-1}(h_1x_1+\ldots+h_sx_s)} dx_1 \ldots dx_s \\ &= \int_{[0,1)^I} f_I(\mathbf{x}_I) e^{-2\pi\sqrt{-1}\sum_{j\in I} h_j x_j} d\mathbf{x}_I \left(\int_{[0,1)^{I^c}} e^{-2\pi\sqrt{-1}\sum_{j\notin I} h_j x_j} d\mathbf{x}_{I^c} \right) \qquad (3.7) \\ &= \begin{cases} 0 & \text{si } h_j \neq 0 \text{ pour au moins un } j \in I^c, \\ \int_{[0,1)^I} f_I(\mathbf{x}_I) e^{-2\pi\sqrt{-1}\sum_{j\in I} h_j x_j} d\mathbf{x}_I & \text{sinon,} \end{cases} \end{aligned}$$

où I^c est le complément de I dans S. En effet, si un des h_j avec $j \in I^c$ est différent de 0, alors l'intégrale entre parenthèses en (3.7) vaut 0, puisque $\int_0^1 e^{-2\pi\sqrt{-1}hx}dx = \left(\sin(2\pi h) - \sin(0) \ + \sqrt{-1}\left(\cos(2\pi h) - \cos(0)\right)\right)/(2\pi h) = 0$. Maintenant, supposons que $h_j = 0$ pour tout $j \in I^c$ et que $h_j = 0$ pour au moins un $j \in I$. Posons $I_0 = \{j \in I : h_j = 0\}$ et dénotons par I_0^c le complément de I_0 dans I. Alors

$$\hat{f}_I(\mathbf{h}) = \int_{[0,1)^{I_0^c}} e^{-2\pi\sqrt{-1}\sum_{j\in I_0^c} h_j x_j} \left(\int_{[0,1)^{I_0}} f_I(\mathbf{x}_I) d\mathbf{x}_{I_0} \right) d\mathbf{x}_{I_0^c} = 0,$$

puisque $I_0 \neq \emptyset$ et donc, $\int_{[0,1)^{I_0}} f_I(\mathbf{x}_I) d\mathbf{x}_{I_0} = 0$ (voir la propriété donnée en (3.1)).

Étant donné que $I \neq I_{\mathbf{h}}$ est équivalent à

$$h_j = 0 \text{ pour au moins un } j \in I \text{ ou } h_j \neq 0 \text{ pour au moins un } j \in I^c$$

et que l'on a montré que $\hat{f}_I(\mathbf{h}) = 0$ si une de ces deux conditions est satisfaite, il reste à montrer que si $I = I_{\mathbf{h}}$, alors

$$\hat{f}_I(\mathbf{h}) = \hat{f}(\mathbf{h}).$$

Or,

$$\begin{aligned} \hat{f}(\mathbf{h}) &= \int_{[0,1)^s} \left(\sum_{J \subseteq S} f_J(\mathbf{x}) \right) e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}} d\mathbf{x}, \\ &= \sum_{J \subseteq S} \hat{f}_J(\mathbf{h}) = \hat{f}_{I_{\mathbf{h}}}(\mathbf{h}), \end{aligned}$$

où la deuxième égalité suit en interchangeant la somme avec l'intégrale et la troisième égalité vient du fait que pour tout $J \neq I_{\mathbf{h}}$, $\hat{f}_J(\mathbf{h}) = 0$. ∎

Dans [46, page 127], une version légèrement moins forte du résultat précédent est donnée, car on dit que $\hat{f}_I(\mathbf{h}) = \hat{f}(\mathbf{h})$ si $I \subseteq I_{\mathbf{h}}$, et $\hat{f}_I(\mathbf{h}) = 0$ sinon. Le résultat qui suit nous montre que la décomposition ANOVA de la variance σ^2 de f et de $\hat{\mu}_{\text{LR}}$ peut être vue comme une façon de partitionner les vecteurs $\mathbf{h} \in \mathbb{Z}^s$ selon l'espace minimal auxquels ils appartiennent, déterminé par les composantes de $\mathbf{h}$ qui sont non nulles.

Proposition 3.3.1 *Soit $f \in \mathcal{L}^2$. Alors pour tout $I \subseteq S$ non vide,*

$$\sigma_I^2 = \sum_{\mathbf{h} \in \mathbb{Z}_I^*} |\hat{f}(\mathbf{h})|^2$$

où $\mathbb{Z}_I^* = \{\mathbf{h} \in \mathbb{Z}^s : I_{\mathbf{h}} = I\}$. De plus, si on écrit l'estimateur $\hat{\mu}_{\text{LR}}$ comme étant

$$\hat{\mu}_{\text{LR}} = \sum_{I \subseteq S} \left(\frac{1}{N} \sum_{i=1}^{N} f_I((\mathbf{x}_i + \mathbf{u}) \bmod 1) \right)$$

et que pour I non vide, on pose

$$\sigma_{I,\text{LR}}^2 = \text{Var}\left(\frac{1}{N} \sum_{i=1}^{N} f_I((\mathbf{x}_i + \mathbf{u}) \bmod 1) \right),$$

alors

$$\left.\begin{aligned} \text{Var}(\hat{\mu}_{\text{LR}}) &= \textstyle\sum_{\emptyset \neq I \subseteq S} \sigma_{I,\text{LR}}^2, \\ \text{où } \sigma_{I,\text{LR}}^2 &= \textstyle\sum_{\mathbf{h} \in L^\perp \cap \mathbb{Z}_I^*} |\hat{f}(\mathbf{h})|^2. \end{aligned}\right\} \tag{3.8}$$

Démonstration : d'abord, on a que

$$\sigma_I^2 = \int_{[0,1)^s} f_I^2(\mathbf{x})d\mathbf{x} = \sum_{\mathbf{h}\in\mathbb{Z}^s} |\hat{f}_I(\mathbf{h})|^2,$$

la première égalité venant du fait que $\int_{[0,1)^s} f_I(\mathbf{x})d\mathbf{x} = 0$ pour tout $\emptyset \neq I \subseteq S$, et la deuxième, de l'égalité de Parseval appliquée à f_I. Or, par le résultat du lemme 3.3.1, on a que

$$\sum_{\mathbf{h}\in\mathbb{Z}^s} |\hat{f}_I(\mathbf{h})|^2 = \sum_{\mathbf{h}\in\mathbb{Z}_I^*} |\hat{f}(\mathbf{h})|^2,$$

puisque $I = I_{\mathbf{h}}$ est équivalent à avoir $\mathbf{h} \in \mathbb{Z}_I^*$. Par définition de $\sigma_{I,\mathrm{LR}}^2$ et par orthogonalité des f_I, on a que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sum_{\emptyset\neq I\subseteq S} \sigma_{I,\mathrm{LR}}^2$. Finalement, on peut calculer

$$\sigma_{I,\mathrm{LR}}^2 = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} f_I(\mathbf{x}_i + \mathbf{u})\right) = \sum_{\mathbf{0}\neq\mathbf{h}\in L^\perp} |\hat{f}_I(\mathbf{h})|^2 = \sum_{\mathbf{h}\in L^\perp\cap\mathbb{Z}_I^*} |\hat{f}(\mathbf{h})|^2,$$

où la deuxième égalité est obtenue par application de la proposition 2.2.1 à la fonction f_I et la troisième, par le lemme 3.3.1. ∎

La proposition 3.3.1 fournit aussi une définition alternative aux notions de dimensions tronquée et (successive) superposée, en utilisant les coefficients de Fourier de f pour décomposer les variances σ_I^2 utilisées dans les définitions données en (3.4), (3.5) et (3.6).

Intuitivement, cette décomposition nous fait voir que pour une fonction donnée, la variance sera d'autant plus réduite si le réseau choisi pour l'intégrer a la propriété que pour les ensembles d'indices I tels que σ_I^2 est grand, les vecteurs $\mathbf{h} \in \mathbb{Z}_I^*$ dont le coefficient $|\hat{f}(\mathbf{h})|^2$ contribuent le plus à σ_I^2 ne sont pas dans le réseau dual $L^\perp$. Si les coefficients $|\hat{f}(\mathbf{h})|$ les plus importants sont associés à des "petits" vecteurs $\mathbf{h}$ (c.-à-d., ayant une petite norme $\|\mathbf{h}\|$, pour un choix de norme à déterminer), alors la propriété mentionnée à la phrase précédente devrait être satisfaite de façon approximative si les petits vecteurs $\mathbf{h} \in \mathbb{Z}_I^*$ ont été évités par le réseau dual $L^\perp$. Cela correspond en gros à ce que font les critères comme le test spectral pour choisir les règles de réseau (dans ce cas, la norme utilisée est la norme euclidienne), si on les applique au sous-ensemble $P_N(I)$. En effet, si on s'assure qu'il n'y a pas de petit vecteur dans $L_I^\perp = \{\mathbf{h} : \sum_{j\in I} h_j x_{ij} \in \mathbb{Z}, i = 1,\ldots,N\}$, alors il n'y en aura pas non plus dans

$L^\perp \cap \mathbb{Z}_I^*$, puisque $L^\perp \cap \mathbb{Z}_I^* \subseteq L_I^\perp$. C'est l'idée que nous exploitons à la section 3.5, lorsque nous présentons notre critère de sélection pour choisir les règles de réseau.

3.4 Cas où f est un polynôme

Nous allons maintenant nous intéresser au cas où l'on veut intégrer une fonction f qui est un polynôme en s variables de degré d, c.-à-d., f est de la forme

$$f(x_1, \ldots, x_s) = \sum_{\mathbf{d} \in D(d)} c(\mathbf{d}) \prod_{j=1}^{s} x_j^{d_j}, \tag{3.9}$$

où $D(d) = \{(d_1, \ldots, d_s) \in [0 \ldots d]^s : \sum_{j=1}^s d_j \leq d\}$ et $c(\mathbf{d}) \in \mathbb{R}$.

Bien sûr, pour ce type de fonction, il n'est pas nécessaire d'utiliser l'intégration numérique puisque μ peut être calculé de façon analytique. Nous croyons qu'il est tout de même intéressant de regarder quelle est la variance de l'estimateur obtenu en utilisant une règle de réseau translatée aléatoirement pour ce type de fonction et de la comparer avec celle de l'estimateur MC et voici pourquoi.

Si on peut approximer une fonction non-polynômiale f par une somme de polynômes, alors l'étude que nous nous proposons de faire est une façon de briser un problème compliqué (la fonction f) en petits problèmes (les polynômes) plus faciles à traiter. En particulier, nous allons donner dans le cas des polynômes de degré un et deux des conditions suffisantes pour réduire la variance par rapport à MC, qui sont assez simples à vérifier. En fait, dans le cas des polynômes de degré 1, nous montrerons que la variance est réduite par un facteur N, si on utilise une règle de rang 1 et d'ordre N qui est complètement projection-régulière. Ainsi, ces résultats pourraient être très utiles en pratique afin d'étudier la réduction de variance (théorique) lorsque le problème pour lequel on utilise la simulation est représenté par une fonction f non-polynômiale qui peut être approximée par une somme de polynômes de degré un ou deux. Nous ne nous penchons pas dans cet ouvrage sur l'approximation de fonctions par des polynômes, mais donnons des outils qui pourront être utilisés si on étudie ce problème dans des travaux futurs.

En prime, l'étude faite dans cette section nous amènera à mieux comprendre la justification derrière certains critères de sélection utilisés pour choisir les règles de

réseau. De plus, elle illustre comment la décomposition de $\mathrm{Var}(f) = \sigma^2$ en une somme $\sigma^2 = \sum_I \sigma_I^2$ peut nous aider à comparer la variance des estimateurs $\hat{\mu}_{\mathrm{LR}}$ et $\hat{\mu}_{\mathrm{MC}}$ et cette décomposition soutient la structure du critère de sélection que nous présentons à la section 3.5. Ces deux raisons nous ont incités à présenter les résultats sur les polynômes avant de définir notre critère de sélection.

3.4.1 Calcul des coefficients de Fourier

Nous donnons d'abord l'expression pour les coefficients de Fourier lorsque f est un polynôme, car c'est l'outil qui nous sert à étudier la variance de l'estimateur $\hat{\mu}_{\mathrm{LR}}$ défini en (2.15), page 37, sur ce type de fonction. Dans ce qui suit, l'ensemble $D(I,d)$ représente les combinaisons de degrés à considérer lorsque l'on étudie la projection f_I :

$$D(I,d) = \{(d_1,\ldots,d_s) \in D(d) : d_j \geq 1 \text{ si } j \in I\}.$$

Remarquons que l'on peut avoir $d_j \geq 1$ même si $j \notin I$ dans la définition de $D(I,d)$: autrement dit, $(d_1,\ldots,d_s) \in D(I,d)$ si et seulement si $I \subseteq \{j : d_j \geq 1\}$.

Lemme 3.4.1 *Soit f un polynôme tel que donné en (3.9) et soit $\mathbf{h} \in \mathbb{Z}^s$. Alors*

$$\hat{f}(\mathbf{h}) = \begin{cases} \mu & si\ |I_{\mathbf{h}}| = 0, \\ \sum_{\mathbf{d} \in D(I_{\mathbf{h}},d)} c(\mathbf{d}) \left(\prod_{j \notin I_{\mathbf{h}}} \frac{1}{(d_j+1)}\right) \left(\prod_{j \in I_{\mathbf{h}}} F(h_j,d_j)\right) & si\ 0 < |I_{\mathbf{h}}| \leq d, \\ 0 & si\ |I_{\mathbf{h}}| > d, \end{cases}$$

où

$$\begin{aligned} F(h_j,d_j) &= \int_0^1 x_j^{d_j} e^{-2\pi\sqrt{-1}h_j x_j} dx_j \\ &= \sum_{w=1}^{d_j} \left(\frac{\sqrt{-1}}{2\pi h_j}\right)^w (-1)^{w+1} \prod_{l=0}^{w-2} (d_j - l). \end{aligned}$$

Démonstration : voir l'annexe B, page xxvi.

Cela signifie entre autres que si f est un polynôme de degré d, alors sa dimension superposée (voir équation (3.5)) est d en proportion 1.

L'exemple suivant illustre comment le résultat de cette sous-section peut être utilisé en combinaison avec ceux de la section précédente afin de déterminer l'importance de

chaque composante ANOVA d'un polynôme quelconque, sans calculer explicitement les f_I.

Exemple 3.4.1 *Soit la fonction tridimensionnelle* $f(x_1, x_2, x_3) = 2x_1x_2 + 3x_3^2 + x_2$. *En utilisant le lemme 3.4.1, on peut calculer les coefficients de Fourier de* f *et on obtient :*

$$\begin{aligned}
\hat{f}(h_1, 0, 0) &= \sqrt{-1}/(2\pi h_1), \text{ si } h_1 \neq 0, \\
\hat{f}(0, h_2, 0) &= \sqrt{-1}/(\pi h_2), \text{ si } h_2 \neq 0. \\
\hat{f}(0, 0, h_3) &= 3\left[\sqrt{-1}/(2\pi h_3) + 1/(2\pi^2 h_3^2)\right], \text{ si } h_3 \neq 0, \\
\hat{f}(h_1, h_2, 0) &= -1/(2\pi^2 h_1 h_2), \text{ si } h_1 h_2 \neq 0,
\end{aligned}$$

et $\hat{f}(\mathbf{h}) = 0$ *pour tous les autres cas. La variance totale est* $\sigma^2 = 56/45$ *: en utilisant la proposition 3.3.1 et les coefficients ci-dessus, elle peut être décomposée comme étant la somme de* $\sigma^2_{\{1\}} = 1/12$, $\sigma^2_{\{2\}} = 1/3$, $\sigma^2_{\{3\}} = 4/5$, *et* $\sigma^2_{\{1,2\}} = 1/36$ *(les autres* σ^2_I *étant nuls). Dans ce cas-ci, les composantes unidimensionnelles* $f_{\{3\}}(x_3)$ *et* $f_{\{2\}}(x_2)$, *contribuent à environ 64% et 27% de la variance totale, respectivement.*

3.4.2 Définition de la mesure de qualité $P_{\alpha_I}(I)$

Maintenant que nous connaissons les coefficients de Fourier de f, nous pouvons calculer explicitement $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ et $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ à l'aide des propositions respectives 2.2.1 (page 37) et 2.2.2 (page 42), les comparer et essayer de trouver des conditions suffisantes pour avoir $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. Pour arriver à cela, il est plus facile de regarder d'abord quelle est la variance au niveau des projections, c.-à-d., de comparer $\sigma^2_{I,\mathrm{LR}}$ (qui est définie à la proposition 3.3.1) et σ^2_I. Nous allons donner à la prochaine sous-section une expression pour $\sigma^2_{I,\mathrm{LR}}$ qui isole les caractéristiques de la règle employée de celles de la fonction, un peu comme le font les bornes sur l'erreur déterministe de type $D(P_N)V(f)$ introduites à l'équation (1.6). Nous utilisons la notation $P_{\alpha_I}(I)$ pour représenter les composantes de $\sigma^2_{I,\mathrm{LR}}$ dépendant de la règle et pour chaque $\sigma^2_{I,\mathrm{LR}}$, on a un ensemble de termes $P_{\alpha_I}(I)$ qui lui sont rattachés. Nous définissons $P_{\alpha_I}(I)$ ci-dessous, passons le restant de la sous-section à donner des formules pour calculer cette quantité et l'utilisons pour définir une généralisation du critère P^s_α.

Définition 3.4.1 *Soit L un réseau, $I \subseteq S$ non vide et $\alpha_I = (\alpha_j)_{j\in I}$ un vecteur d'entiers tel que $\alpha_j \geq 2$ pour tout $j \in I$ et que $\sum_{j\in I} \alpha_j = 0 \bmod 2$. On définit*

$$P_{\alpha_I}(I) = \sum_{\mathbf{h}\in L^\perp \cap \mathbb{Z}_I^*} \prod_{j\in I} h_j^{-\alpha_j}. \tag{3.10}$$

La condition $\sum_{j\in I} \alpha_j = 0 \bmod 2$ est là simplement pour s'assurer que $P_{\alpha_I}(I)$ ne vaut pas 0.

Cette mesure $P_{\alpha_I}(I)$ est reliée au critère P_α^s défini à l'équation (2.9). La différence est que l'on ne somme que sur les vecteurs $\mathbf{h} \in L^\perp \cap \mathbb{Z}_I^*$ et que les différentes composantes h_j de $\mathbf{h}$ ne sont pas nécessairement élevées à la même puissance. Plus précisément, étant donné que $\{\mathbb{Z}_I^*\}_{\emptyset\neq I\subseteq S}$ forme une partition de $\mathbb{Z}^s\backslash\{\mathbf{0}\}$, la relation entre les deux définitions est que

$$P_\alpha^s = \sum_{\emptyset\neq I\subseteq S} P_{\underbrace{(\alpha,\ldots,\alpha)}_{|I| \text{ fois}}}(I).$$

Si on utilise des puissances α_j paires, alors on peut obtenir une formule pour $P_{\alpha_I}(I)$ qui se calcule en $O(N|I|)$:

Lemme 3.4.2 *Soit $P_N = L \cap [0,1)^s$ une règle de réseau d'ordre N, $I \subseteq S$ et $\alpha_I = (\alpha_j)_{j\in I}$ un vecteur d'entiers pairs avec $\alpha_j \geq 2$ pour tout $j \in I$. Alors*

$$P_{\alpha_I}(I) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j\in I}\left[-\frac{(-1)^{\alpha_j/2}(2\pi)^{\alpha_j}}{(\alpha_j)!}B_{\alpha_j}(x_{ij})\right].$$

Démonstration : on doit utiliser le fait que la décomposition en série de Fourier d'un polynôme de Bernoulli dont le degré α est pair est donnée par [46]

$$B_\alpha(x) = \frac{-(-1)^{\alpha/2}\alpha!}{(2\pi)^\alpha}\sum_{h\neq 0}\frac{e^{2\pi\sqrt{-1}hx}}{|h|^\alpha},\ 0 \leq x \leq 1.$$

Ainsi, on obtient que

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\prod_{j\in I}\left[-\frac{(-1)^{\alpha_j/2}(2\pi)^{\alpha_j}}{(\alpha_j)!}B_{\alpha_j}(x_{ij})\right] &= \frac{1}{N}\sum_{i=1}^{N}\prod_{j\in I}\left[\sum_{h_j\neq 0}\frac{e^{2\pi\sqrt{-1}h_j x_{ij}}}{|h_j|^{\alpha_j}}\right] \\
&= \frac{1}{N}\sum_{\mathbf{h}\in\mathbb{Z}_I^*}\frac{\sum_{i=1}^{N} e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i}}{\prod_{j\in I} h_j^{-\alpha_j}} \\
&= \sum_{\mathbf{h}\in L^\perp\cap\mathbb{Z}_I^*}\prod_{j\in I} h_j^{-\alpha_j},
\end{aligned}$$

où la deuxième égalité est obtenue en changeant l'ordre des sommes sur $\mathbf{h}$ et sur i et la troisième égalité vient de [116, *Lemma 2.7*] (voir l'équation (2.6)) et du fait que α_j pair implique que $|h_j|^{-\alpha_j} = h_j^{-\alpha_j}$. ∎

Cette formule ressemble à celle pour le P_α^s, sauf qu'ici, le terme qui est multiplié dans le produit sur $j \in I$ est $\left[-(-1)^{\alpha_j/2}(2\pi)^{\alpha_j} B_{\alpha_j}(x_{ij})/(\alpha_j)!\right]$ au lieu de $\left[1-(-1)^{\alpha_j/2}(2\pi)^{\alpha_j} B_{\alpha_j}(x_{ij})/(\alpha_j)!\right]$ et il n'y a pas de (-1) en avant du produit de $1/N$ avec la somme sur les $i = 1, \ldots N$.

À l'aide des $P_{\alpha_I}(I)$, on peut généraliser la définition de P_α^s à une mesure $P_{\alpha_1,\ldots,\alpha_s}$ et obtenir une formule se calculant dans $O(Ns)$ de la façon suivante :

Lemme 3.4.3 *Soient $\alpha_1, \ldots, \alpha_s$ des entiers. Posons*

$$\begin{aligned} P_{\alpha_1,\ldots,\alpha_s} &= \sum_{\emptyset \neq I \subseteq S} P_{\alpha_I}(I), \\ &= \sum_{0 \neq \mathbf{h} \in L^\perp} \prod_{j \in I_{\mathbf{h}}} h_j^{-\alpha_j}. \end{aligned}$$

Si chaque α_j est pair, alors

$$P_{\alpha_1,\ldots,\alpha_s} = -1 + \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{s} \left[1 - \frac{(-1)^{\alpha_j/2}(2\pi)^{\alpha_j/2}}{\alpha_j!} B_{\alpha_j}(x_{ij})\right].$$

Démonstration : on a que

$$\begin{aligned} P_{\alpha_1,\ldots,\alpha_s} &= \sum_{\emptyset \neq I \subseteq S} P_{\alpha_I}(I), \\ &= \sum_{\emptyset \neq I \subseteq S} \left\{ \frac{1}{N} \sum_{i=1}^{N} \prod_{j \in I} \left[-\frac{(-1)^{\alpha_j/2}(2\pi)^{\alpha_j}}{(\alpha_j)!} B_{\alpha_j}(x_{ij}) \right] \right\} \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{\emptyset \neq I \subseteq S} \prod_{j \in I} \left[-\frac{(-1)^{\alpha_j/2}(2\pi)^{\alpha_j}}{(\alpha_j)!} B_{\alpha_j}(x_{ij}) \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \prod_{j=1}^{s} \left[1 - \frac{(-1)^{\alpha_j/2}(2\pi)^{\alpha_j}}{(\alpha_j)!} B_{\alpha_j}(x_{ij}) \right] - 1 \right\}, \end{aligned}$$

la deuxième égalité vient du lemme 3.4.2 et la troisième est obtenue en changeant l'ordre de sommation. ∎

Quel est l'avantage de $P_{\alpha_1,\ldots,\alpha_s}$ sur P_α^s? Avec $P_{\alpha_1,\ldots,\alpha_s}$, on peut accorder à chaque dimension $j \in S$ plus ou moins de poids en ajustant les α_j : plus α_j est petit, plus la j^e coordonnée a de l'importance.

Dans le cas où $|I| = 1$ et que la règle est de rang 1, on a la formule suivante pour $P_\alpha(\{j\})$, qui sera utilisée à plusieurs reprises dans les sous-sections suivantes, ainsi que dans le chapitre 5 :

Lemme 3.4.4 *Si on utilise une règle de rang 1 et d'ordre N qui est complètement projection-régulière, alors pour $\alpha \geq 2$ pair, on a que*

$$P_\alpha(\{j\}) = \frac{2\zeta(\alpha)}{N^\alpha}, \qquad j = 1, \ldots, s,$$

où $\zeta(\alpha)$ est la fonction zeta de Riemann évaluée en α.

Démonstration : puisque la règle de rang 1 est complètement projection-régulière, par la proposition 3.2.2, cela signifie que $\text{pgcd}(z_j, N) = 1$, où $(z_1, \ldots, z_s)$ est le vecteur générateur de la règle. Ainsi, les seules solutions à $hz_j = 0 \bmod N$ sont de la forme $h = lN, l \in \mathbb{Z}$ pour $j = 1, \ldots, s$. Donc, $\mathbf{h} \in L^\perp \cap Z^*_{\{j\}}$ si et seulement si ($h_j = 0 \bmod N$, $h_j \neq 0$ et $h_k = 0$ si $k \neq j$), ce qui implique que

$$P_\alpha(\{j\}) = \sum_{0 \neq l \in \mathbb{Z}} (Nl)^{-\alpha} = \frac{2\zeta(\alpha)}{N^\alpha},$$

qui est défini, puisque $\alpha > 1$. La deuxième égalité vient du fait que α est pair. ∎

Puisque les projections unidimensionnelles d'une règle de rang 1 qui est complètement projection-régulière sont toujours données par $\{0, 1/N, \ldots, (N-1)/N\}$ et donc, sont indépendantes de la règle, il est normal qu'on ait une formule pour $P_\alpha(\{j\})$ se calculant dans $O(1)$ plutôt que dans $O(N)$. Le lemme suivant est lui aussi très facile à démontrer, mais nous le donnons quand même, car il sera utilisé dans la démonstration de résultats ultérieurs.

Lemme 3.4.5 *Soit $I \subseteq S$ et $\alpha_I = (\alpha_j)_{j \in I}$ un vecteur d'entiers avec $\alpha_j \geq 2$ pour tout $j \in I$. Alors*

$$|P_{\alpha_I}(I)| \leq P_{(2,\ldots,2)}(I).$$

Démonstration : puisque $\alpha_j \geq 2$ pour tout $j \in I$, alors $\left|h_j^{-\alpha_j}\right| \leq h_j^{-2}$ pour tout $h_j \in \mathbb{Z}$ et donc

$$\prod_{j \in I} \left|h_j^{-\alpha_j}\right| \leq \prod_{j \in I} h_j^{-2} \Rightarrow \left|\sum_{\mathbf{h} \in L^\perp \cap \mathbb{Z}^*_I} \prod_{j \in I} h_j^{-\alpha_j}\right| \leq \sum_{\mathbf{h} \in L^\perp \cap \mathbb{Z}^*_I} \left|\prod_{j \in I} h_j^{-\alpha_j}\right| \leq \sum_{\mathbf{h} \in L^\perp \cap \mathbb{Z}^*_I} \prod_{j \in I} h_j^{-2}.$$

∎

3.4.3 Expression et borne pour la variance

Nous pouvons maintenant donner le résultat qui décompose la variance $\sigma^2_{I,\mathrm{LR}}$ en une somme pondérée de ces $P_{\alpha_I}(I)$. Afin de faciliter la présentation, nous définissons les ensembles $A_{I,d}$ et $A^+_{I,d}$ qui sont utilisés afin de décrire la somme sus-mentionée :

$$\begin{aligned} A_{I,d} &= \{\alpha_I : 2 \le \alpha_j \le 2(d-|I|+1) \text{ pour tout } j \in I, \sum_{j\in I} \alpha_j = 0 \bmod 2\}, \\ A^+_{I,d} &= \{\alpha_I : 2 \le \alpha_j \le 2(d-|I|+1), \alpha_j = 0 \bmod 2 \text{ pour tout } j \in I\}. \end{aligned}$$

Donc, $A^+_{I,d}$ est un sous-ensemble de $A_{I,d}$.

La proposition qui suit nous donne une expression pour σ^2_I et pour $\sigma^2_{I,\mathrm{LR}}$ qui sont toutes deux exprimées sous la même forme, de façon à ce qu'il soit plus facile ensuite de comparer ces deux quantités. Dans les deux cas, on a une somme sur un vecteur α_I et les termes de cette somme sont formés du produit de $\gamma_{\alpha_I}(I)$ (qui dépend seulement de la fonction) avec $2^t \prod_{j\in I} \zeta(\alpha_j)$ dans le cas de σ^2_I et $P_{\alpha_I}(I)$ dans le cas de $\sigma^2_{I,\mathrm{LR}}$. La somme qui nous donne σ^2_I contient moins de termes que celle définissant $\sigma^2_{I,\mathrm{LR}}$ car dans le premier cas, on somme sur les $\alpha_I \in A^+_{I,d}$ et dans le deuxième, sur les $\alpha_I \in A_{I,d}$. *

Proposition 3.4.1 *Soit f un polynôme de degré d et $I \subseteq S$, tel que $1 \le |I| = t \le d$. Alors*

$$\sigma^2_I = \sum_{\alpha_I \in A^+_{I,d}} \gamma_{\alpha_I}(I)\, 2^t \prod_{j\in I} \zeta(\alpha_j) \tag{3.11}$$

$$\text{et} \qquad \sigma^2_{I,\mathrm{LR}} = \sum_{\alpha_I \in A_{I,d}} \gamma_{\alpha_I}(I)\, P_{\alpha_I}(I), \tag{3.12}$$

où

$$\gamma_{\alpha_I}(I) = \sum_{\substack{\mathbf{d},\mathbf{d}' \in D(I,d) \\ d_j + d'_j \ge \alpha_j,\ j\in I}} c(\mathbf{d})c(\mathbf{d}') \left(\prod_{k\notin I} \frac{1}{(d_k+1)(d'_k+1)} \right) g(\alpha_I, \mathbf{d}, \mathbf{d}') \tag{3.13}$$

*En fait, on aurait pu dire que σ^2_I est également donné par l'expression $\sum_{\alpha_I \in A_{I,d}} \gamma_{\alpha_I}(I) P_{\alpha_I,0}(I)$, où $P_{\alpha_I,0}(I)$ est le $P_{\alpha_I}(I)$ de la règle contenant le point $\{\mathbf{0}\}$, puisque $\sigma^2_I = \mathrm{Var}(f_I(\mathbf{0}+\mathbf{u})) = \sigma^2_{I,\mathrm{LR},0}$, où $\sigma^2_{I,\mathrm{LR},0}$ est le $\sigma^2_{I,\mathrm{LR}}$ de la règle formée du point $\{\mathbf{0}\}$. Nous avons décidé d'être plus explicite et d'utiliser le fait que

$$P_{\alpha_I,0}(I) = \begin{cases} 2^t \prod_{j\in I} \zeta(\alpha_j) & \text{si } \alpha_j = 0 \bmod 2 \text{ pour tout } j \in I, \\ 0 & \text{sinon.} \end{cases}$$

et

$$g(\alpha_I, \mathbf{d}, \mathbf{d}') = \prod_{j \in I} \left[\left(\frac{\sqrt{-1}}{2\pi} \right)^{\alpha_j} \sum_{v_j = \max(1, \alpha_j - d_j')}^{\min(\alpha_j - 1, d_j)} (-1)^{v_j} \prod_{l=0}^{v_j - 2} (d_j - l) \prod_{l=0}^{\alpha_j - v_j - 2} (d_j' - l) \right]. \tag{3.14}$$

Démonstration : voir l'annexe B, page xxvii.

L'expression pour $\sigma^2_{I,\mathrm{LR}}$ donnée à la proposition 3.4.1 nous a conduit naturellement à la définition de $P_{\alpha_I}(I)$, qui utilise une norme de type produit pour mesurer les vecteurs $\mathbf{h}$. Intuitivement, cela justifie l'utilisation de ce type de norme. De plus, nous avons vu à la section 2.1.4 que le fait d'utiliser des poids de la forme $w(\mathbf{h}) = \prod_{j \in I} h_j^{-\alpha_j} \mathbf{1}_{\{h_j \neq 0\}}$ avec α pair dans la mesure de qualité générale $D_w(P_N) = \sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} w(\mathbf{h})$ donnée en (2.8) permet de passer d'une somme infinie sur les $\mathbf{h}$ dans $L^\perp$ à une somme finie sur les points de P_N, puisque l'on peut alors utiliser le lien entre cette définition de $D_w(P_N)$ et la représentation en série de Fourier des polynômes de Bernoulli. Ceci est un autre point en faveur de l'utilisation de la norme produit pour mesurer les vecteurs $\mathbf{h}$.

Nous donnons maintenant un résultat qui borne la variance de $\hat{\mu}_{\mathrm{LR}}$ par une constante, qui dépend de la fonction, multipliée par $P_2^{\mathrm{sup}}(d)$, qui est défini par

$$P_2^{\mathrm{sup}}(d) = \sum_{\emptyset \neq I \subseteq S, |I| \leq d} P_{\{2,\ldots,2\}}(I).$$

L'abréviation "sup" signifie "superposée". On peut voir $P_2^{\mathrm{sup}}(d)$ comme une version de P_2^s qui ne considère que les projections dont la dimension superposée est inférieure ou égale à d.

Corollaire 3.4.1 *Soit f un polynôme de degré d. Alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq P_2^{\mathrm{sup}}(d) \max_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \right).$$

Démonstration : on a que

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) &= \sum_{\emptyset \neq I \subseteq S, |I| \leq d} \sigma^2_{I,\mathrm{LR}} \\ &= \sum_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} \gamma_{\alpha_I}(I) P_{\alpha_I}(I) \right) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \, |P_{\alpha_I}(I)| \right) \\
&\leq \sum_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \, P_{(2,\ldots,2)}(I) \right) \\
&\leq \left(\sum_{\emptyset \neq I \subseteq S, |I| \leq d} P_{(2,\ldots,2)}(I) \right) \max_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \right) \\
&= P_2^{\sup}(d) \max_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \right).
\end{aligned}
$$

Dans la série d'égalités précédentes, la première tient car $\sigma^2_{I,\mathrm{LR}} = 0$ si $|I| > d$, en combinant les résultats de la proposition 3.3.1 et du lemme 3.4.1 ; la deuxième inégalité est une conséquence du lemme 3.4.5. ∎

On peut utiliser le résultat précédent pour borner le taux auquel la variance converge lorsque l'on intègre un polynôme de degré d. En effet, puisque le terme multipliant $P_2^{\sup}(d)$ dans la borne donnée au corollaire 3.4.1 ne dépend pas de N, ce taux de convergence peut être borné par le taux de convergence de $P_2^{\sup}(d)$. Or, $P_2^{\sup}(d) \leq P_2^s$ et comme nous l'avons déjà expliqué au chapitre 2 (page 41), puisqu'il est possible de trouver une règle de réseau telle que

$$
P_\alpha^s \leq e(s,\alpha) \frac{(\log N)^{\alpha(s-1)}}{N^\alpha}, \tag{3.15}
$$

cela signifie que l'on peut trouver une règle telle que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = O(N^{-2}(\log N)^{2(s-1)})$ pour intégrer un polynôme de degré d [†]. Par contre, rappelons que la méthode utilisée pour démontrer (3.15) n'est pas constructive.

3.4.4 Conditions pour que la variance soit réduite

Nous donnons maintenant des conditions suffisantes afin d'avoir $\sigma^2_{I,\mathrm{LR}} \leq k\sigma_I^2$, pour une constante k réelle positive quelconque. En posant $k = 1/N$ et en vérifiant que cette condition tient pour tout I non vide, cela nous assure alors que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. Nous sommes conscients que l'énoncé de ces conditions n'offre peut-être pas beaucoup d'intuition pour l'instant, mais il nous permet d'avoir un résultat général que l'on

[†] On aurait pu également arriver à ce résultat en utilisant le lemme 3.4.1 pour montrer que les polynômes satisfont la condition du corollaire 2.2.1 avec $\alpha = 1$.

pourra appliquer au cas où $d = 1, 2$ à la sous-section suivante, obtenant ainsi des conditions et expressions plus concrètes.

En gros, les conditions ont la forme suivante : on compare terme par terme les expressions pour σ_I^2 et $\sigma_{I,\mathrm{LR}}^2$ données en (3.11) et (3.12), respectivement. Ainsi, pour tous les α_I qui sont dans l'ensemble $A_{I,d}$, on détermine d'abord si α_I fait partie de $A_{I,d}^+$ ou non, c.-à-d., si α_I est dans $A_{I,d}^+$ ou $A_{I,d}^-$, où

$$A_{I,d}^- = \{\alpha_I : 2 \le \alpha_j \le 2(d - |I| + 1)\ \forall j \in I, \sum_{j \in I} \alpha_j = 0 \bmod 2, \sum_{j \in I} \mathbf{1}_{\{\alpha_j = 1 \bmod 2\}} > 0\}.$$

Rappelons que les α_I dans $A_{I,d}^+$ sont ceux qui sont également sommés pour calculer σ_I^2. Une fois l'appartenance déterminée, on doit vérifier si $\gamma_{\alpha_I}(I)$ est inférieur ou supérieur à 0 et selon ce résultat, dans le cas où $\alpha_I \in A_{I,d}^+$, il faut vérifier si $P_{\alpha_I}(I)$ respecte une certaine inégalité, qui revient en fait à comparer $P_{\alpha_I}(I)$ avec la quantité équivalente dans l'expression pour σ_I^2.

Proposition 3.4.2 *Soit f un polynôme de degré d, $k > 0$ une constante réelle, $I \subseteq S$, tel que $1 \le |I| = t \le d$ et $\gamma_{\alpha_I}(I)$, la fonction définie en (3.13).*

Si, pour tout $\alpha_I \in A_{I,d}^+$, on a que

$$P_{\alpha_I}(I) \begin{cases} \le & k2^t \prod_{j \in I} \zeta(\alpha_j) \quad si\ \gamma_{\alpha_I}(I) \ge 0 \\ \ge & k2^t \prod_{j \in I} \zeta(\alpha_j) \quad si\ \gamma_{\alpha_I}(I) < 0 \end{cases} \tag{3.16}$$

et que

$$\sum_{\alpha_I \in A_{I,d}^-} \gamma_{\alpha_I}(I) P_{\alpha_I}(I) \le 0, \tag{3.17}$$

alors

$$\sigma_{I,\mathrm{LR}}^2 \le k\sigma_I^2.$$

Démonstration : on a que

$$\begin{aligned} \sigma_I^2 &= \sum_{\alpha_I \in A_{I,d}^+} \gamma_{\alpha_I}(I) 2^t \prod_{j \in I} \zeta(\alpha_j) \\ &\ge \frac{1}{k} \sum_{\alpha_I \in A_{I,d}^+} \gamma_{\alpha_I}(I) P_{\alpha_I}(I) \\ &= \frac{1}{k} \left(\sigma_{I,\mathrm{LR}}^2 - \sum_{\alpha_I \in A_{I,d}^-} \gamma_{\alpha_I}(I) P_{\alpha_I}(I) \right) \\ &\ge \sigma_{I,\mathrm{LR}}^2 / k, \end{aligned}$$

où la première inégalité tient puisque la condition (3.16) est vérifiée et la deuxième, puisque la condition (3.17) est vérifiée. L'égalité à la troisième ligne vient du fait que $A^+_{I,d}$ et $A^-_{I,d}$ partitionnent l'ensemble $A_{I,d}$ de α_I sur lequel on somme dans la définition de $\sigma^2_{I,\mathrm{LR}}$. ∎

Remarque 3.4.1 *Les conditions énoncées dans le proposition 3.4.2 sont suffisantes, mais pas nécessaires. En effet, une partie des inégalités données en (3.16) et (3.17) peuvent être violées si les autres arrivent à compenser, surtout si les vecteurs α_I pour lesquels les inégalités sont violées dans (3.16) ont un poids $|\gamma_{\alpha_I}(I)|$ associé qui est près de 0.*

Remarque 3.4.2 *Dans le cas où $\alpha_j = \alpha$ pair pour tout j dans I, c'est la condition (3.16) qui doit être vérifiée, puisque α_I est alors dans $A^+_{I,d}$. Dans ce cas, les indices dans I ne servent qu'à déterminer si $P_{\alpha_I}(I)$ doit être inférieur ou supérieur à la borne $(2\zeta(\alpha))^{|I|}/N$ qui elle, ne dépend que de la cardinalité de I.*

Évidemment, on peut aller plus loin et utiliser le résultat précédent pour donner des conditions suffisantes pour avoir une réduction de variance globale, lorsque l'on compare les deux estimateurs $\hat{\mu}_{\mathrm{MC}}$ et $\hat{\mu}_{\mathrm{LR}}$.

Corollaire 3.4.2 *Soit f un polynôme de degré d. Si, pour tout $I \subseteq S$ tel que $1 \leq |I| \leq d$, la condition (3.16) est respectée avec $k = 1/N$ pour tout $\alpha_I \in A^+_{I,d}$ et que la condition (3.17) est respectée, alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : par hypothèse, la proposition 3.4.2 s'applique avec $k = 1/N$ et donc,

$$\sigma^2_{I,\mathrm{LR}} \leq \frac{1}{N}\sigma^2_I,$$

pour tout $I \subseteq S$ tel que $1 \leq |I| \leq d$. On obtient ainsi que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sum_{\substack{I \subseteq S \\ 1 \leq |I| \leq d}} \sigma^2_{I,\mathrm{LR}} \leq \frac{1}{N} \sum_{\substack{I \subseteq S \\ 1 \leq |I| \leq d}} \sigma^2_I = \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

∎

Pour utiliser ce résultat afin de trouver une règle qui nous garantit que la variance sera réduite pour un polynôme donné, il faudrait calculer le signe des différents $\gamma_{\alpha_I}(I)$ impliqués dans le corollaire 3.4.2 et faire une recherche sur les règles afin d'en trouver qui respectent les inégalités sur les différents $P_{\alpha_I}(I)$, en prenant $k = 1/N$. Nous croyons que cette approche est trop lourde à utiliser. En fait, la difficulté vient du fait que l'on ne connaît pas *a priori* le signe de $\gamma_{\alpha_I}(I)$ pour les différents α_I impliqués dans la vérification des conditions. Par exemple, cela nous simplifierait beaucoup la tâche si on savait que pour tout polynôme, $\alpha_I \in A^+_{I,d}$ implique que $\gamma_{\alpha_I}(I) > 0$ et $\alpha_I \in A^-_{I,d}$ implique que $\gamma_{\alpha_I}(I)P_{\alpha_I}(I) \leq 0$ pour toute règle, car alors on n'aurait qu'à vérifier que les $P_{\alpha_I}(I)$ tels que $\alpha_I \in A^+_{I,d}$ sont suffisamment petits afin de montrer que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$, ce qui, sans être facile, est un peu plus raisonnable. Malheureusement, ces conditions sur le signe de $\gamma_{\alpha_I}(I)$ ne tiennent pas en général.

Comme alternative, nous établissons à la définition 3.4.2 un critère pour choisir les règles qui *dépend de la fonction* et qui représente une partie de la variance de $\hat{\mu}_{\mathrm{LR}}$. Cette définition utilise le fait que lorsque $\alpha_j = 2$ pour tout $j \in I$, alors $\gamma_{\alpha_I}(I) \geq 0$, comme l'indique le lemme suivant :

Lemme 3.4.6 *Soit la fonction $\gamma_{\alpha_I}(I)$ telle que définie en (3.13) et $I \subseteq S$, telle que $|I| = t$, $1 \leq t \leq d$ et que $\alpha_j = 2$ pour tout $j \in I$. Alors*

$$\gamma_{\alpha_I}(I) = (2\pi)^{-2t}\left(\sum_{\mathbf{d}\in D(I,d)} c(\mathbf{d}) \prod_{k\notin I} \frac{1}{(d_k+1)}\right)^2 \geq 0.$$

Démonstration : il suffit de regarder ce que vaut $g(2,\ldots,2,\mathbf{d},\mathbf{d}')$. Or, on a que

$$g(2,\ldots,2,\mathbf{d},\mathbf{d}') = (2\pi)^{-2t}(-1)^t \sum_{v_1=1}^{1} \cdots \sum_{v_t=1}^{1} (-1)^t = (2\pi)^{-2t},$$

pour tous $\mathbf{d},\mathbf{d}' \in D(I,d)$. Donc,

$$\begin{aligned}
\gamma_{(2,\ldots,2)}(I) &= \sum_{\mathbf{d},\mathbf{d}'\in D(I,d)} c(\mathbf{d})c(\mathbf{d}') \prod_{k\notin I} \frac{1}{(d_k+1)(d_k'+1)} (2\pi)^{-2t} \\
&= (2\pi)^{-2t}\left(\sum_{\mathbf{d}\in D(I,d)} c(\mathbf{d}) \prod_{k\notin I} \frac{1}{(d_k+1)}\right)^2 \\
&\geq 0.
\end{aligned}$$

∎

Puisque le coefficient $\gamma_{(2,\ldots,2)}(I)$ est toujours non négatif, cela signifie que pour les α_I de la forme $(2,\ldots,2)$, on doit vérifier que $P_{(2,\ldots,2)}(I) \leq k2^{|I|}\zeta(2)^{|I|}$ dans la proposition 3.4.2. Ainsi, afin d'obtenir un estimateur réduisant la variance par rapport à MC, on pourrait chercher une règle qui minimise les différents critères $P_{(2,\ldots,2)}(I)$, pour $I \subseteq S$ tel que $1 \leq |I| \leq d$, où d est le degré de la fonction à intégrer. Le lemme 3.4.2 nous fournit une formule pour évaluer ces quantités. Ensuite, il faut pondérer l'importance que l'on doit accorder à chacun de ces $P_{(2,\ldots,2)}(I)$. En effet, il semble peu probable qu'une seule règle minimise chacun des $P_{(2,\ldots,2)}(I)$. Si c'était le cas, cela signifierait que cette règle serait bonne pour intégrer n'importe quel polynôme. L'idée la plus intuitive pour résoudre cette difficulté est d'utiliser les poids $\gamma_{(2,\ldots,2)}(I)$ que l'on retrouve dans l'expression pour $\sigma^2_{I,\mathrm{LR}}$. C'est ce que l'on utilise afin de définir le critère suivant :

Définition 3.4.2 *Pour f un polynôme de degré d, on définit le critère $\hat{P}_2^d(f)$ comme étant :*

$$\hat{P}_2^d(f) = \sum_{I:1\leq|I|=t\leq d} (2\pi)^{-2t} \left(\sum_{\mathbf{d}\in D(I,d)} c(\mathbf{d}) \prod_{k\notin I} \frac{1}{d_k+1} \right)^2 \underbrace{P_{(2,\ldots,2)}}_{|I| \text{ fois}}(I). \tag{3.18}$$

La quantité $\hat{P}_2^d(f)$ représente une partie de la variance de $\hat{\mu}_{\mathrm{LR}}$, donc la minimiser signifie que l'on minimise une partie de la variance. De plus, puisque $P_{(2,\ldots,2)}(I) \geq |P_{\alpha_I}(I)|$ pour $\alpha_j \geq 2$ et que $\gamma_{(2,\ldots,2)}(I) \geq 0$ (voir les lemmes 3.4.5 et 3.4.6), cela signifie que l'on minimise probablement une partie importante de la variance en choisissant la règle selon le critère $\hat{P}_2^d(f)$. En fait, nous verrons à la prochaine sous-section que dans le cas où $d = 2$, la minimisation de $\hat{P}_2^d(f)$ nous assure de minimiser *toute* la variance.

Remarque 3.4.3 *Évidemment, puisque le critère $\hat{P}_2^d(f)$ est spécifique aux polynômes et que nous n'avons pas besoin d'utiliser une approximation (basée sur une règle de réseau ou autre) afin de calculer l'intégrale μ dans ce cas, ce critère n'a pas d'utilité directe en pratique. Cependant, la forme de $\hat{P}_2^d(f)$ peut nous suggérer un critère à utiliser dans un contexte général (pas seulement pour les polynômes) ayant la structure $\sum_I c(\sigma_I^2)P_{(2,\ldots,2)}(I)$, où les $c(\sigma_I^2)$ devraient approximer le comportement des σ_I^2, puisque les $\gamma_{(2,\ldots,2)}(I)$ utilisés dans la définition de $\hat{P}_2^d(f)$ sont des composantes de σ_I^2 et ces variances σ_I^2 sont définies pour toute fonction dans $\mathcal{L}^2$ et non pas seulement pour les polynômes.*

3.4.5 Cas particuliers où $d = 1$ ou $d = 2$

Pour obtenir des résultats plus forts quant à la réduction de variance que peut amener l'estimateur $\hat{\mu}_{\text{LR}}$ par rapport à l'estimateur MC, on regarde d'abord le cas particulier où le degré du polynôme est 1, c.-à-d., lorsque la fonction est linéaire.

Lemme 3.4.7 *Soit f un polynôme de degré 1. Soit $\hat{\mu}_{\text{LR}}$ l'estimateur obtenu en utilisant une règle de rang 1 d'ordre N qui est complètement projection-régulière. Alors pour tout $j = 1, \ldots, s$,*

$$\sigma^2_{\{j\},\text{LR}} = \frac{1}{N^2}\sigma^2_{\{j\}}.$$

Démonstration : en utilisant la proposition 3.4.1, on a que

$$\sigma^2_{\{j\},\text{LR}} = \gamma_2(\{j\})P_2(\{j\}) = c^2(\mathbf{e}_j)(2\pi)^{-2}P_2(\{j\}) = \frac{1}{N^2}c^2(\mathbf{e}_j)(2\pi)^{-2}2\zeta(2) = \frac{1}{N^2}\sigma^2_{\{j\}},$$

où $\mathbf{e}_j \in \mathbb{Z}^s$ est un vecteur de 0 avec un 1 en j^e position. Dans ce qui précède, la deuxième égalité est obtenue en appliquant le lemme 3.4.6 et la troisième vient du lemme 3.4.4. ∎

On peut utiliser ce résultat pour montrer que dans le cas d'un polynôme de degré 1, l'estimateur provenant d'une règle de réseau de rang 1 qui est complètement projection-régulière a une variance N fois plus petite que celle de l'estimateur MC correspondant.

Corollaire 3.4.3 *Soit f un polynôme de degré 1. Soit $\hat{\mu}_{\text{MC}}$ l'estimateur MC obtenu avec N points et $\hat{\mu}_{\text{LR}}$ l'estimateur obtenu en utilisant une règle de rang 1 d'ordre N qui est complètement projection-régulière. Alors*

$$\text{Var}(\hat{\mu}_{\text{LR}}) = \frac{1}{N}\text{Var}(\hat{\mu}_{\text{MC}}).$$

Démonstration : on a que

$$\text{Var}(\hat{\mu}_{\text{MC}}) = \frac{1}{N}\sum_{j=1}^{s}\sigma^2_{\{j\}}$$

et

$$\text{Var}(\hat{\mu}_{\text{LR}}) = \sum_{j=1}^{s}\sigma^2_{\{j\},\text{LR}} = \frac{1}{N^2}\sum_{j=1}^{s}\sigma^2_{\{j\}} = \frac{1}{N}\text{Var}(\hat{\mu}_{\text{MC}}),$$

la deuxième égalité suivant par application du lemme 3.4.7. ∎

Remarque 3.4.4 *Étant donné qu'un polynôme de degré 1 est une somme de s fonctions unidimensionnelles monotones, on savait déjà par le corollaire 2.3.1 que la variance était réduite par rapport à MC en utilisant une règle de réseau translatée de rang 1 complètement projection-régulière. Le corollaire 3.4.3 nous donne l'information supplémentaire que le facteur de réduction est de* $1/N$.

Dans le cas particulier où f est un polynôme de degré 2, la réduction de variance dépend des caractéristiques de la règle employée. En effet, on doit maintenant considérer les projections de la règle sur des sous-ensembles I contenant une *ou deux* dimensions et ce n'est que dans le cas où $|I| = 1$ que les règles de réseau de rang 1 complètement projection-régulières sont toutes les mêmes. Pour vérifier qu'il y a réduction de variance, on utilise le corollaire 3.4.2.

Corollaire 3.4.4 *Soit f un polynôme de degré 2. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC obtenu avec N points et $\hat{\mu}_{\mathrm{LR}}$ l'estimateur obtenu en utilisant une règle de rang 1 d'ordre N qui est complètement projection-régulière. Si on a*

$$P_{(2,2)}(\{i,j\}) \leq \frac{\pi^4}{9N} \tag{3.19}$$

pour tout $1 \leq i < j \leq s$, alors

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : on n'a qu'à utiliser le corollaire 3.4.2, mais dans le cas où $d = 2$. La condition (3.17) doit donc être vérifiée pour tous les I tels que $|I| \leq 2$, mais $A^-_{I,2} = \emptyset$ si $|I| \leq 2$: la condition est donc vide dans ce cas. Ensuite, on doit vérifier la condition (3.16) pour les I tels que $|I| \leq 2$ et tous les α_I dans $A^+_{I,2}$: or,

$$A^+_{I,2} = \begin{cases} \{(2),(4)\} & \text{si } |I| = 1, \\ \{(2,2)\} & \text{si } |I| = 2. \end{cases}$$

Par le lemme 3.4.6, on sait que $\gamma_{(2,2)}(\{i,j\}) \geq 0$, $\gamma_2(\{j\}) \geq 0$. De plus,

$$\gamma_4(\{j\}) = \sum_{\substack{\mathbf{d},\mathbf{d}' \in D(\{j\},2) \\ d_j + d'_j \geq 4}} c(\mathbf{d}) c(\mathbf{d}') \left(\prod_{k \neq j} \frac{1}{(d_k+1)(d'_k+1)} \right) g(4,\mathbf{d},\mathbf{d}')$$

et le seul couple $(\mathbf{d}, \mathbf{d}')$ tel que $\mathbf{d}, \mathbf{d}' \in D(\{j\}, 2)$ et $d_j + d'_j \geq 4$ est $(2\mathbf{e}_j, 2\mathbf{e}_j)$. Puisque

$$g(4, 2\mathbf{e}_j, 2\mathbf{e}_j) = (2\pi)^{-4}(-1)^2 \sum_{v_1=2}^{2} (-1)^{v_1} \prod_{l=0}^{v_1-2} (2-l) \prod_{l=0}^{4-v_1-2} (2-l) = 4(2\pi)^{-4} > 0,$$

on a que $\gamma_4(\{j\}) \geq 0$. Étant donné que tous les coefficients $\gamma_{\alpha_I}(I)$ associés aux α_I dans $A^+_{I,d}$ sont non négatifs, la vérification de la condition (3.16) avec $k = 1/N$ se résume donc à :

$$P_{\alpha_I}(I) \leq \begin{cases} \frac{2}{N}\zeta(\alpha_{i_1}) & \text{si } |I| = 1, \text{ pour } \alpha_{i_1} = 2, 4, \\ \frac{2^2}{N}\zeta(\alpha_{i_1})\zeta(\alpha_{i_2}) & \text{si } |I| = 2, \text{ pour } \alpha_{i_1} = \alpha_{i_2} = 2. \end{cases}$$

Par le lemme 3.4.4, qui s'applique puisque P_N est complètement projection-régulière et de rang 1, pour $\alpha \in \{2, 4\}$, on a que

$$P_\alpha(\{j\}) = \frac{2\zeta(\alpha)}{N^\alpha} \leq \frac{2\zeta(\alpha)}{N},$$

puisque $N \geq 1$ et par hypothèse, pour tout $I \subseteq S$ tel que $|I| = 2$, on a que

$$P_{(2,2)}(I) \leq \frac{\pi^4}{9N} = \frac{1}{N} 2^2 \zeta^2(2),$$

puisque $\zeta(2) = \pi^2/6$. Donc, la condition (3.16) est vérifiée avec $k = 1/N$. ∎

Remarque 3.4.5 *Dans le cas où la règle utilisée est une règle de Korobov, la condition (3.19) doit seulement être vérifiée pour les $s-1$ paires (i, j) telles que $i = 1, j = 2, \ldots, s$.*

La condition énoncée dans ce corollaire est forte, puisque l'on demande que le $P_{(2,2)}(I)$ de *chacun* des sous-ensembles I de deux variables soit inférieur à $\pi^4/9N$. À cause de cela, on ne peut réécrire cette condition en une plus globale, qui soit donnée pour P_2^s, par exemple. Par contre, si on est prêt à tenir compte des caractéristiques de la fonction, cette condition peut être affaiblie de la façon suivante :

Corollaire 3.4.5 *Soit f un polynôme de degré 2. Soit $\hat{\mu}_{\mathrm{LR}}$ l'estimateur obtenu avec une règle de réseau de rang 1 et d'ordre N qui est complètement projection-régulière. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC correspondant. Si*

$$\frac{\sum_{1 \leq i < j \leq s} c^2(\mathbf{e}_i + \mathbf{e}_j) P_{(2,2)}(\{i, j\})}{\sum_{1 \leq i < j \leq s} c^2(\mathbf{e}_i + \mathbf{e}_j)} \leq \frac{\pi^4}{9N},$$

alors

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : on a que

$$
\begin{aligned}
&\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \\
&= \sum_{j=1}^{s} \sigma^2_{\{j\},\mathrm{LR}} + \sum_{1\le i<j\le s} \sigma^2_{\{i,j\},\mathrm{LR}} \\
&= \sum_{j=1}^{s} (\gamma_2(\{j\})P_2(\{j\}) + \gamma_4(\{j\})P_4(\{j\})) + \sum_{1\le i<j\le s} \gamma_{(2,2)}(\{i,j\})P_{(2,2)}(\{i,j\}) \\
&= \sum_{j=1}^{s} \left(\frac{\gamma_2(\{j\})}{N^2} 2\zeta(2) + \frac{\gamma_4(\{j\})}{N^4} 2\zeta(4) \right) + \sum_{1\le i<j\le s} (2\pi)^{-4} c^2(\mathbf{e}_i + \mathbf{e}_j) P_{(2,2)}(\{i,j\}) \\
&\le \sum_{j=1}^{s} \left(\frac{\gamma_2(\{j\})}{N^2} 2\zeta(2) + \frac{\gamma_4(\{j\})}{N^4} 2\zeta(4) \right) + \sum_{1\le i<j\le s} c^2(\mathbf{e}_i + \mathbf{e}_j) \frac{1}{144N} \\
&= \frac{1}{N} \left(\sum_{j=1}^{s} \left(\frac{\gamma_2(\{j\})}{N} 2\zeta(2) + \frac{\gamma_4(\{j\})}{N^3} 2\zeta(4) \right) + \sum_{1\le i<j\le s} (2\pi)^{-4} c^2(\mathbf{e}_i + \mathbf{e}_j) 2^2 \zeta^2(2) \right) \\
&\le \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}),
\end{aligned}
$$

la deuxième égalité venant de la proposition 3.4.1, la troisième des lemmes 3.4.4 et 3.4.6 et l'inégalité à la quatrième ligne tient par hypothèse. ∎

Ainsi, en pondérant les $P_{(2,2)}(\{i,j\})$ par les coefficients $c^2(\mathbf{e}_i+\mathbf{e}_j)$, cela nous permet de demander une condition moyenne sur l'ensemble des couples (i,j), qui est plus facile à respecter. En effet, une mauvaise projection $P_N(\{i_1,j_1\})$ de l'ensemble de points est moins néfaste si le coefficient correspondant $c(\mathbf{e}_{i_1}+\mathbf{e}_{j_1})$ est petit en valeur absolue et peut être compensée par de bons résultats sur les autres projections.

Voici un exemple simplifié qui donne une idée de ce que cette condition requiert au niveau des coefficients $c(\mathbf{e}_i+\mathbf{e}_j)$ lorsque les $P_{(2,2)}(\{i,j\})$ sont connus et ne respectent pas tous la condition (3.19) du corollaire 3.4.4.

Exemple 3.4.2 *Supposons que $s=4$ et que la règle est stationnaire dans la dimension, avec*

$$
\begin{aligned}
P_{2,2}(\{1,2\}) &= \frac{\pi^4}{9N} - 6\text{e-}6 \\
P_{2,2}(\{1,3\}) &= \frac{\pi^4}{9N} + 5\text{e-}6 \\
P_{2,2}(\{1,4\}) &= \frac{\pi^4}{9N} + 3\text{e-}6.
\end{aligned}
$$

Donc, $P_{2,2}(\{1,3\})$ et $P_{2,2}(\{1,4\})$ ne satisfont pas la condition (3.19). Pour simplifier

la notation, posons $c_{i,j} = c(\mathbf{e}_i + \mathbf{e}_j)$. *On doit donc vérifier que*

$$\left(c_{1,2}^2 + c_{2,3}^2 + c_{3,4}^2\right) P_{2,2}(\{1,2\}) + \left(c_{1,3}^2 + c_{2,4}^2\right) P_{2,2}(\{1,3\}) + c_{1,4}^2 P_{2,2}(\{1,4\}) \leq \left(\sum_{i=1}^{3}\sum_{j=i+1}^{4} c_{i,j}^2\right) \frac{\pi^4}{9N},$$

ce qui revient à demander que

$$-6\text{e-}6\left(c_{1,2}^2 + c_{2,3}^2 + c_{3,4}^2\right) + 5\text{e-}6\left(c_{1,3}^2 + c_{2,4}^2\right) + 3\text{e-}6\ c_{1,4}^2 \leq 0.$$

Ainsi, on peut dire que cette règle permet de définir un estimateur $\hat{\mu}_{\mathrm{LR}}$ *réduisant la variance par rapport à l'estimateur MC pour tout polynôme de degré deux satisfaisant*

$$\frac{5c_{1,3}^2 + 5c_{2,4}^2 + 3c_{1,4}^2}{6(c_{1,2}^2 + c_{2,3}^2 + c_{3,4}^2)} \leq 1.$$

Cette inégalité est respectée si, par exemple, les coefficients $c_{1,2}, c_{2,3}$ *et* $c_{3,4}$ *sont les plus grands en valeur absolue. Elle ne l'est pas si, par exemple, on a* $|c_{1,3}| > 1.5|c_{1,2}| > 0$, $|c_{2,4}| > 1.5|c_{2,3}|$ *et* $|c_{1,4}| > \sqrt{2}|c_{3,4}|$.

À partir d'ici jusqu'à la fin de la présente sous-section, nous donnons des résultats qui nous permettront de montrer, au corollaire 3.4.8, que pour tout polynôme de degré deux, il existe une règle dont la variance associée est bornée par la variance de l'estimateur MC multipliée par un terme dans $O(1 + 1/N)$. Nous débutons avec le corollaire 3.4.6, qui montre que la minimisation de la somme pondérée de $P_{(2,2)}(\{i,j\})$ utilisée dans l'énoncé du corollaire 3.4.5 équivaut à minimiser la variance de $\hat{\mu}_{\mathrm{LR}}$.

Corollaire 3.4.6 *Soit f un polynôme de degré 2. Parmi les règles de rang 1 et d'ordre N qui sont complètement projection-régulières, celles qui minimisent la quantité*

$$\sum_{1 \leq i < j \leq s} c^2(\mathbf{e}_i + \mathbf{e}_j) P_{(2,2)}(\{i,j\})$$

minimisent également la variance de $\hat{\mu}_{\mathrm{LR}}$.

Démonstration : dans la démonstration du corollaire 3.4.5, nous avons vu que pour une règle complètement projection-régulière de rang 1 et d'ordre N,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) &= \frac{2}{N^2}\sum_{j=1}^{s}\left(\gamma_2(\{j\})\zeta(2) + \frac{\gamma_4(\{j\})}{N^2}\zeta(4)\right) + \sum_{1 \leq i < j \leq s} \frac{c^2(\mathbf{e}_i + \mathbf{e}_j)}{(2\pi)^4} P_{(2,2)}(\{i,j\}) \\ &= c_1(N) + c_2 \sum_{1 \leq i < j \leq s} c^2(\mathbf{e}_i + \mathbf{e}_j) P_{(2,2)}(\{i,j\})\end{aligned}$$

où $c_2 = (2\pi)^{-4}$ et $c_1(N)$ ne dépendent pas de la règle. Donc, si P_N est tel que

$$\sum_{1\le i<j\le s} c^2(\mathbf{e}_i+\mathbf{e}_j)P_{(2,2)}(\{i,j\})$$

est minimal parmi toutes les règles complètement projection-régulières de rang 1, alors $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ est aussi minimale. ∎

Remarque 3.4.6 *Dans le cas d'un polynôme de degré 2, si la règle est complètement projection-régulière et de rang 1, alors on a que*

$$\begin{aligned}\hat{P}_2^d(f) &= (2\pi)^{-2}\sum_{j=1}^{s}\left(\sum_{\mathbf{d}\in D(\{j\},2)} c(\mathbf{d})\prod_{k\neq j}\frac{1}{d_k+1}\right)^2\frac{2\pi^2}{6N^2}\\ &\quad+(2\pi)^{-4}\sum_{1\le i<j\le s} c^2(\mathbf{e}_i+\mathbf{e}_j)P_{(2,2)}(\{i,j\}) \qquad (3.20)\end{aligned}$$

et donc, minimiser $\sum_{1\le i<j\le s} c^2(\mathbf{e}_i+\mathbf{e}_j)P_{(2,2)}(\{i,j\})$ *est équivalent à minimiser* $\hat{P}_2^d(f)$ *dans ce cas, puisque le premier terme dans (3.20) ne dépend pas de* P_N. *Autrement dit, dans le cas où* $d=2$, *la minimisation du critère* $\hat{P}_2^d(f)$ *nous assure de minimiser* toute *la variance.*

On doit maintenant regarder ce qui se passe en moyenne, sur tout l'ensemble des règles de rang 1 telles que N est premier, d'une façon similaire à ce qui a été fait au chapitre 2 (proposition 2.2.3). Ceci va nous permettre de démontrer l'existence de la règle dont la variance associée est bornée par la variance de l'estimateur MC multipliée par un terme dans $O(1+1/N)$, pour un polynôme de degré deux donné.

Pour faire cela, nous avons besoin du lemme suivant, qui donne une borne supérieure sur le $P_{\alpha_I}(I)$ moyen des règles de rang 1, pour un N premier. Dans ce qui suit, $P_{\alpha_I}(I,\mathbf{a})$ dénote la mesure $P_{\alpha_I}(I)$ correspondant à la règle de rang 1 dont le vecteur générateur est $\mathbf{a}$.

Lemme 3.4.8 *Soit* $I\subseteq S$ *tel que* $1\le|I|=t\le s$ *et* $\alpha_I=(\alpha_j)_{j\in I}$ *un vecteur d'entiers pairs avec* $\alpha_j\ge 2$ *pour tout* $j\in I$. *Si* N *est premier, alors*

$$\frac{1}{(N-1)^s}\sum_{\mathbf{a}\in[1\ldots N-1]^s} P_{\alpha_I}(I,\mathbf{a}) \le \frac{1}{N-1}\left(1+\left(\frac{N-2}{N^{2t}}\right)\right)2^t\prod_{j\in I}\zeta(\alpha_j).$$

Démonstration : voir l'annexe B, page xxix.

Cette borne supérieure sur le $P_{\alpha_I}(I)$ moyen est donc formée du produit de $2^t \prod_{j\in I} \zeta(\alpha_j)$ et d'une quantité légèrement supérieure à $1/N$. Cela signifie que le $P_{\alpha_I}(I)$ moyen respecte la condition (3.16) requise pour avoir $\sigma^2_{I,\mathrm{LR}} \leq k\sigma^2_I$ dans le cas où $\gamma_{\alpha_I}(I) \geq 0$, avec $k = (1 + (N-2)/N^{2t})/(N-1)$. Malheureusement, cela ne suffit pas pour nous assurer qu'il existe une règle telle que $\sigma^2_{I,\mathrm{LR}} \leq \sigma^2_I/k$, puisque la *même* règle doit satisfaire la condition (3.16) pour tous les α_I tels que $\gamma_{\alpha_I}(I) \geq 0$, en plus de satisfaire cette condition dans le cas où $\gamma_{\alpha_I}(I) < 0$. Ainsi, même si le résultat précédent tient pour tout d, ce n'est que dans le cas où $d = 2$ que l'on peut montrer qu'il existe une règle telle que $\sigma^2_{I,\mathrm{LR}} \leq k\sigma^2_I$ pour tout $|I| \leq d$ et donc, telle que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq k\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ avec une constante k légèrement supérieure à 1. Pour montrer cela, on doit d'abord borner la variance moyenne :

Corollaire 3.4.7 *Soit f un polynôme de degré 2 et N un nombre premier. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC contenant N points et $\hat{\mu}_{\mathrm{LR}}(\mathbf{a})$ l'estimateur obtenu en utilisant une règle de rang 1 et d'ordre N basée sur le vecteur générateur* $\mathbf{a}$. *Alors*

$$\frac{1}{(N-1)^s} \sum_{\mathbf{a}\in[1,\ldots,N-1]^s} \mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{a})) \leq \left(1 + \frac{1}{N} + \frac{1}{N^2} + \frac{2}{N^3}\right) \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : on a que

$$\begin{aligned}
&\frac{1}{(N-1)^s} \sum_{\mathbf{a}\in[1,\ldots,N-1]^s} \mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{a})) \\
&= \frac{1}{(N-1)^s} \sum_{\mathbf{a}\in[1,\ldots,N-1]^s} \sum_{\substack{I\subseteq S \\ 1\leq|I|\leq 2}} \sigma^2_{I,\mathrm{LR},\mathbf{a}} \\
&= \frac{1}{(N-1)^s} \sum_{\mathbf{a}\in[1,\ldots,N-1]^s} \left[\sum_{j=1}^{s} (\gamma_2(\{j\})P_2(\{j\},\mathbf{a}) + \gamma_4(\{j\})P_4(\{j\},\mathbf{a})) \right. \\
&\quad \left. + \sum_{I:|I|=2} \gamma_{(2,2)}(I)P_{(2,2)}(I,\mathbf{a})\right] \\
&\leq \sum_{j=1}^{s} \left(\frac{\gamma_2(\{j\})}{N^2} 2\zeta(2) + \frac{\gamma_4(\{j\})}{N^4} 2\zeta(4)\right) + \frac{1}{N-1}\left(1 + \frac{N-2}{N^4}\right) \sum_{I:|I|=2} \gamma_{(2,2)}(I) 2^2\zeta^2(2) \\
&\leq \frac{N}{N-1}\left(1 + \frac{N-2}{N^4}\right) \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) \\
&= \left(1 + \frac{1}{N} + \frac{1}{N^2} + \frac{2}{N^3}\right) \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).
\end{aligned}$$

Dans ce qui précède, la deuxième égalité découle de la proposition 3.4.1 et la première inégalité vient du lemme 3.4.4, qui s'applique pour tout $\mathbf{a}$ puisque N est premier (et

ainsi $\text{pgcd}(a_j, N) = 1$, ce qui garantit la projection-régularité), du lemme 3.4.8 et du fait que $\gamma_2(\{j\})$, $\gamma_4(\{j\})$ et $\gamma_{(2,2)}(I)$ sont non négatifs (voir la démonstration du corollaire 3.4.4 pour $\gamma_4(\{j\}) \geq 0$). ∎

En combinant le résultat précédent avec le corollaire 3.4.6, on arrive à donner un critère dont la minimisation nous assure de trouver une règle dont la variance ne peut excéder par beaucoup la variance de l'estimateur MC basé sur le même nombre de points :

Corollaire 3.4.8 *Pour N premier, si $\mathbf{a}_1 \in [1 \dots N-1]^s$ est tel que*

$$\sum_{1 \leq i < j \leq s} c^2(\mathbf{e}_i + \mathbf{e}_j) P_{(2,2)}(\{i,j\}, \mathbf{a}_1) \leq \sum_{1 \leq i < j \leq s} c^2(\mathbf{e}_i + \mathbf{e}_j) P_{(2,2)}(\{i,j\}, \mathbf{a}_2) \qquad (3.21)$$

pour tout $\mathbf{a}_2 \in [1 \dots N-1]^s$, alors

$$\text{Var}(\hat{\mu}_{\text{LR}}(\mathbf{a}_1)) \leq \left(1 + \frac{1}{N} + \frac{1}{N^2} + \frac{2}{N^3}\right) \text{Var}(\hat{\mu}_{\text{MC}}).$$

Démonstration : tout vecteur $\mathbf{a} \in [1 \dots N-1]^s$ donne lieu à une règle de rang 1 complètement projection-régulière puisque N est premier et donc, par le corollaire 3.4.6, la condition (3.21) implique que

$$\text{Var}(\hat{\mu}_{\text{LR}}(\mathbf{a}_1)) \leq \text{Var}(\hat{\mu}_{\text{LR}}(\mathbf{a}_2))$$

et donc

$$\text{Var}(\hat{\mu}_{\text{LR}}(\mathbf{a}_1)) \leq \frac{1}{(N-1)^s} \sum_{\mathbf{a} \in [1 \dots N-1]^s} \text{Var}(\hat{\mu}_{\text{LR}}(\mathbf{a})) \leq \left(1 + \frac{1}{N} + \frac{1}{N^2} + \frac{2}{N^3}\right) \text{Var}(\hat{\mu}_{\text{MC}}),$$

la deuxième inégalité suivant par le corollaire 3.4.7. ∎

3.5 Nouveau critère de sélection pour les règles de réseau

Dans cette section, nous définissons un critère de sélection plus approprié que P_α^s ou M_T, qui ont été définis à la sous-section 2.1.4, et qui devrait, pour la plupart des fonctions rencontrées en simulation, nous fournir des règles dont les estimateurs associés ont une variance inférieure à celle de l'estimateur MC. Cela sera vérifié dans le cas du problème des options asiatiques à la section 3.6.3.

Avant de définir ce nouveau critère, nous expliquons d'abord de quelle façon les critères P_α^s et M_T devraient être améliorés et parlons de critères plus généraux qui existent déjà dans la littérature, soit le $\tilde{P}_\alpha^s$ [45, 46] et le *test spectral pondéré* [42].

3.5.1 Motivation

Le résultat de la proposition 3.3.1 (page 88) et la discussion faite à la remarque 3.4.3 (page 102) nous suggèrent qu'il est préférable de choisir la règle en accordant plus de poids aux projections $P_N(I)$ pour lesquelles σ_I^2/σ^2 est grand. Cela nous indique que les critères comme P_α^s devraient être remplacés par des critères plus généraux, qui regardent de plus près la qualité des différentes projections en leur donnant plus ou moins de poids. Le critère $P_{\alpha_1,\ldots,\alpha_s}$ introduit au lemme 3.4.3 permet de donner des poids différents à chaque dimension mais n'a pas la flexibilité désirée car si on veut, par exemple, que toutes les projections unidimensionnelles aient le même poids (ce qui n'est pas inhabituel comme requête), alors on doit prendre $\alpha_1 = \ldots = \alpha_s = \alpha$ et on retombe alors sur le critère P_α^s.

Le critère $\hat{P}_2^d(f)$ que nous avons donné en (3.18) a la propriété de considérer les différentes projections, mais dépend explicitement de la fonction à intégrer et est valide seulement pour les polynômes. On préférerait un critère plus général, qui devrait nous permettre de choisir des règles qui soient bonnes pour un large éventail de fonctions. Nous sommes conscients que cette approche est une heuristique qui ne nous garantit évidemment pas que les critères de sélection définis en suivant cette idée vont choisir la meilleure règle pour un problème donné. Pour faire cela, il faudrait plutôt utiliser une méthode *adaptative* qui, pour une fonction donnée, nous permettrait de construire au fur et à mesure que l'on recueille de l'information sur f un estimateur ayant une très petite erreur. Ceci pourrait être l'objet de travaux futurs mais pour l'instant, on préfère continuer à travailler sur la définition de meilleurs critères de sélection.

3.5.2 Critère $\tilde{P}_\alpha^s$ d'Hickernell

Hickernell a proposé dans [45, 46] des généralisations pour le P_α^s qui tiennent compte des projections. Son idée est d'utiliser des poids β_I quantifiant l'importance relative

de chaque sous-ensemble $I \subseteq S$, en fonction de leur apport à la variabilité σ_I^2 de la composante ANOVA f_I associée. Autrement dit, les poids devraient être choisis de façon à ce que si σ_I^2 est grand par rapport aux autres σ_J^2, alors β_I devrait aussi être grand par rapport aux autres β_J et vice-versa. De façon plus précise, en fonction des $P_{\alpha_I}(I)$, cette mesure est donnée par

$$\tilde{P}_\alpha^s = \sum_{\emptyset \neq I \subseteq S} \beta_I^2 P_{(\alpha,\ldots,\alpha)}(I), \tag{3.22}$$

qui correspond à $D^2_{\mathcal{F},\alpha/2,2}(P)$ en utilisant la notation dans [46, équation (4.8a), page 128]. Par rapport au critère général $D_w(P_N)$ défini en (2.8), cela correspond à utiliser $D_w(P_N) = \sum_{\mathbf{0} \neq \mathbf{h} \in L^\perp} w(\mathbf{h})$ avec les poids $w(\mathbf{h}) = \beta_{I_\mathbf{h}}^2 \|\mathbf{h}\|_\pi^{-\alpha}$. La discussion faite à la remarque 3.4.3 présente une façon alternative d'arriver à un critère de cette forme.

Dans le cas particulier où les β_I sont de type produit, c.-à-d., de la forme

$$\beta_I = \beta_0 \prod_{j \in I} \beta_j^{\alpha/2}, \tag{3.23}$$

pour $\alpha \geq 2$ un entier pair et $\beta_0, \ldots, \beta_s > 0$, le $\tilde{P}_\alpha^s$ pour une règle de réseau P_N est donné par [46]

$$\tilde{P}_\alpha^s = \beta_0^2 \left\{ -1 + \frac{1}{N} \sum_{\mathbf{z} \in P_N} \prod_{j=1}^{s} \left[1 - \frac{(-4\pi^2 \beta_j^2)^{\alpha/2}}{\alpha!} B_\alpha(z_j) \right] \right\}. \tag{3.24}$$

Si l'on veut être aussi sévère envers toutes les projections I ayant la même cardinalité $|I|$ et être plus exigeant avec celles pour lesquelles $|I|$ est petit, alors on doit poser $\beta_1 = \ldots = \beta_s = \beta < 1$. Le critère P_α^s est obtenu en posant $\beta_0 = \ldots = \beta_s = 1$, ce qui signifie qu'il accorde autant d'importance à chaque projection, puisqu'alors $\beta_I = 1$ pour tout $I \subseteq S$ non vide.

Remarque 3.5.1 *Nous avons vu à la remarque 3.4.2 à la page 100 que dans le cas des polynômes, la condition (3.16) que doit satisfaire $P_{(2,\ldots,2)}(I)$ dans le but de réduire la variance par rapport à la méthode MC ne dépend pas des indices spécifiques dans I, mais de la cardinalité $|I|$. Aussi, cette condition est de moins en moins sévère à mesure que $|I|$ augmente, car la borne supérieure donnée par $(2\zeta(2))^{|I|}/N$ croît avec $|I|$. En supposant que cette propriété des polynômes peut être étendu au cas général, on peut voir ces deux faits comme une justification pour choisir $\beta_1 = \ldots = \beta_s = \beta < 1$.*

Tout comme dans le cas de la mesure $P_{\alpha_1,\ldots,\alpha_s}$, on ne peut fixer qu'un paramètre par dimension avec $\tilde{P}_\alpha^s$. Ce qui est différent avec $\tilde{P}_\alpha^s$, c'est que l'on peut fixer les β_j de façon à ce que chaque projection unidimensionnelle ait la même importance sans retomber sur le critère P_α^s, c.-à-d., tout en ayant des poids différents pour les différentes projections.

Finalement, on peut voir que notre critère $\hat{P}_2^d(f)$ donné en (3.18) est un cas particulier de $\tilde{P}_\alpha^s$, avec $\alpha = 2$ et les poids

$$\beta_I = \begin{cases} (2\pi)^{-t} \left| \sum_{\mathbf{d} \in D(I,d)} c(\mathbf{d}) \prod_{k \notin I} \frac{1}{d_k+1} \right| & \text{si } |I| \leq d, \\ 0 & \text{sinon.} \end{cases}$$

Ces poids ne peuvent être décomposés de façon à satisfaire la condition (3.23) et donc, on ne peut utiliser la formule (3.24) pour calculer $\hat{P}_2^d(f)$. Avec ces poids, on ne peut faire autrement que de calculer chaque $P_{(2,\ldots,2)}(I)$ pour $1 \leq |I| \leq d$, en utilisant le lemme 3.4.2.

3.5.3 Liens entre $\tilde{P}_\alpha^s$ et le test spectral pondéré (*weighted spectral test* [42])

Nous voulons brièvement expliquer dans cette sous-section comment le critère $\tilde{P}_\alpha^s$ peut être relié au test spectral pondéré, tel que défini dans [42]. Ceci n'est qu'une extension du lien déjà connu comme quoi P_2^s est un cas particulier de la *diaphonie classique (classical diaphony)* [137], qui elle-même constitue une des réalisations du concept plus général qu'est le test spectral pondéré, dont nous donnons maintenant la définition :

Définition 3.5.1 *[42, Definition 6.1] Soit un système de fonctions $\mathcal{F} = \{\chi_\mathbf{h}\}$ (avec soit $\chi_\mathbf{h} = e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}}$, qui correspond aux fonctions de base de la série de Fourier, ou soit les $\{\chi_\mathbf{h}\}$ correspondent aux fonctions de base de la série de Walsh en base q [42, Definition 3.1]), une norme $\|\cdot\|$ sur $\mathbb{R}^s$ et une fonction de poids $r : \mathbf{h} \to r(\mathbf{h})$ satisfaisant*

(1) $r(\mathbf{h}) > 0 \qquad \forall \mathbf{h},$

(2) $r(\mathbf{0}) = 1,$

(3) $\sum_{\mathbf{h}} 1/r(\mathbf{h})^2 < \infty$.

Pour un ensemble de points $\omega = (\mathbf{x}_i)_{i \geq 1}$ *dans* $[0,1)^s$, *on définit le* test spectral pondéré *par la quantité*

$$F_N(\omega) = \left(\sum_{\mathbf{h} \neq 0} \frac{1}{r(\mathbf{h})^2} |S_N(\chi, \omega)|^2 \right)^{1/2},$$

où

$$S_N(\chi, \omega) = \frac{1}{N} \sum_{i=1}^{N} \chi_{\mathbf{h}}(\mathbf{x}_i).$$

Donc, si on choisit $\chi_{\mathbf{h}} = e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}}$, $r(\mathbf{h}) = \beta_{I_{\mathbf{h}}}^{-1} \|\mathbf{h}\|_\pi^{\alpha/2}$ et que les N premiers points de ω forment une règle de réseau, alors $(F_N(\omega))^2 = \tilde{P}_\alpha^s$. Nous tenions à expliquer ce lien, car d'après les résultats numériques de la sous-section 3.6.4, il semble que le critère $\tilde{P}_\alpha^s$ permette de choisir des règles qui obtiennent de bons résultats en pratique, ce qui suggère qu'il y aurait peut-être lieu de s'intéresser également à d'autres versions du test spectral pondéré afin de définir des critères de sélection.

3.5.4 Définition du nouveau critère

Nous proposons maintenant un nouveau critère de sélection pour choisir les règles de réseau, qui est basé sur le test spectral. C'est un cas particulier de la mesure $D'_w(P_N) = \sup_{\mathbf{0} \neq \mathbf{h} \in L^\perp} w(\mathbf{h})$ donnée en (2.8). Le choix des $w(\mathbf{h})$ est motivé par l'expression pour la variance donnée en (3.8). En gros, l'idée est de maximiser la longueur du plus court vecteur dans chacun des réseaux duaux $L_I^\perp$ pour lesquels on juge que σ_I^2 est important. Ainsi, on s'assure que la projection $P_N(I)$ associée est bien distribuée. L'utilisation du supremum plutôt que de la somme fait en sorte que le calcul est plus rapide. En effet, étant donné qu'on n'utilise pas la norme produit dans la définition de $w(\mathbf{h})$ mais plutôt la norme euclidienne (comme le fait le test spectral), la simplification à l'aide des polynômes de Bernoulli n'est pas possible lorsque l'on utilise la mesure $D_w(P_N)$ et donc, il serait pratiquement impossible de faire le calcul si on utilisait la somme plutôt que le supremum. De toute façon, le critère $\tilde{P}_\alpha^s$, qui utilise la somme et la norme produit, devient assez difficile à utiliser comme critère de sélection quand N et/ou s est grand, même si on peut utiliser la formule (3.24) qui se calcule en $O(Ns)$. Des comparaisons du temps CPU requis pour calculer P_2^s et d_s sont faites dans [27].

En utilisant l'algorithme de "branch-and-bound" donné dans [77] pour calculer d_s, ces auteurs ont calculé en une seconde la valeur de d_s pour $s = 2, \ldots, 32$, pour une règle de Korobov d'ordre $N \approx 2^{30}$, alors que cela leur a pris environ 31 minutes pour calculer P_2^s, $s = 2, \ldots, 32$, pour la même règle et sur le même ordinateur.

De plus, avec le test spectral, il existe des bornes absolues sur les critères associés aux différentes projections $P_N(I)$, ce qui nous permet de combiner de diverses façons les mesures associées à chacune de ces projections. Pour les $P_{\alpha_I}(I)$, on peut essayer de "normaliser" en utilisant la valeur correspondante de $P_{\alpha_I}(I)$ (donnée par $(2^{|I|}/N) \prod_{j \in I} \zeta(\alpha_j)$) pour un ensemble de N points i.i.d. uniformes, mais cette approche n'est pas aussi satisfaisante, entre autres parce qu'on ne peut borner ce rapport. Donc, nous croyons que le test spectral possède une rapidité et une flexibilité qui le rendent attrayant pour construire des critères de sélection.

Dans ce qui suit, nous supposons que P_N est stationnaire dans la dimension. Cela nous permet de réduire le nombre de projections $P_N(I)$ à considérer, puisque plusieurs sont alors équivalentes. Nous supposons également que la règle est de rang 1 et complètement projection-régulière. Ainsi, il n'est pas nécessaire de regarder les projections unidimensionnelles $P_N(\{j\})$, pour $j = 1, \ldots, s$, puisqu'elles sont toujours données par l'ensemble $\{0, 1/N, \ldots, (N-1)/N\}$, qui est optimal pour une réseau de N points en une dimension.

Le nouveau critère que nous définissons est une généralisation de M_T qui tient compte de façon explicite de la projection de P_N non seulement sur les I de la forme $I = \{1, \ldots, t\}$ pour $1 \leq t \leq t_1$, mais aussi sur les I dans

$$\bigcup_{u=2}^{d} S(t_u, u),$$

où $S(t_u, u) = \{I = \{i_1, \ldots, i_u\} : 1 = i_1 < \ldots < i_u \leq t_u\}$. Autrement dit, cette nouvelle mesure calcule d_t pour tous les $P_N(I)$ tels que I est à indices successifs et contient au plus t_1 indices, mais aussi pour les I contenant u indices, dont le premier est 1 et le dernier est inférieur ou égal à t_u, pour $u = 2, \ldots, d$. Avec le critère M_T (qui correspond à choisir $u = 1$ et $t_1 = T$), les projections $P_N(I)$ telles que I n'est pas à indices successifs ne sont pas regardées explicitement : elles ne sont considérées qu'à travers le plus petit ensemble J de la forme $J = \{1, \ldots, t\}$ tel que $I \subset J$. La raison

pour laquelle on pose $i_1 = 1$ dans la définition de $S(t_u, u)$ est que l'on a supposé que P_N était stationnaire dans la dimension.

Définition 3.5.2 *Soient $t_1, \ldots, t_d$ des entiers positifs. On définit le critère*

$$M_{t_1,\ldots,t_d} = \min\left[\min_{2\le j\le t_1} l_j/l_j^*(N), \min_{2\le u\le d}\left(\min_{I\in S(t_u,u)} l_I/l_{|I|}^*(N)\right)\right].$$

Rappelons que $l_t^{-1} = d_t$ est la quantité calculée dans le test spectral, c.-à-d., l_t correspond à la longueur du plus court vecteur dans $L_{\{1,\ldots,t\}}^{\perp}$, le réseau dual associé à $P_N(\{1,\ldots,t\})$, et l_I correspond à la longueur du plus court vecteur dans $L_I^{\perp}$. La quantité $l_t^*(N) = c_t N^{1/t}$ est une borne supérieure absolue pour l_t, qui correspond à la meilleure valeur possible de l_t pour un réseau contenant N points dans $[0,1)^t$. La valeur de c_t peut être trouvée dans [58] pour $t \le 8$ (constante d'Hermite) et des bornes sur c_t sont données dans [16, 74, 75] pour $t > 8$: dans nos calculs, nous avons pris pour $t > 8$ la borne (inférieure) de Rogers [16].

Donc, $M_{t_1,\ldots,t_d}$ est toujours entre 0 et 1 et plus il est près de 1, meilleure est la qualité de P_N. La proposition suivante spécifie comment $M_{t_1,\ldots,t_d}$ est équivalent à la mesure de discrépance générale $D_w'(P_N)$ définie en (2.8). Ceci va nous permettre de voir que notre critère n'est pas aussi sévère envers toutes les projections et de proposer une heuristique pour le choix des paramètres $d, t_1, \ldots, t_d$. Rappelons la signification des symboles $r(I)$ et $H(t_1, \ldots, t_d, d)$, puisqu'ils sont utilisés dans l'énoncé : $r(I) = i_t - i_1 + 1$ est l'étendue de I, pour $I = \{i_1, \ldots, i_t\}$, et $H(t_1, \ldots, t_d, d)$ est l'ensemble des I tels que $\tilde{I}$ est soit dans $S(t_u, u)$, pour $2 \le u \le d$, ou soit de la forme $\{1, \ldots, j\}$, avec $1 \le j \le t_1$.

Proposition 3.5.1 *Supposons que $t_1 \ge t_u$ pour $2 \le u \le d$. Avec les poids*

$$w(\mathbf{h}) = \begin{cases} l_{|I_\mathbf{h}|}^*(N)/\|\mathbf{h}\|_2 & \text{si } \mathbf{h} \in L^{\perp} \text{ et } I_\mathbf{h} \in H(t_1, \ldots, t_d, d), \\ l_{r(I_\mathbf{h})}^*(N)/\|\mathbf{h}\|_2 & \text{si } \mathbf{h} \in L^{\perp},\ I_\mathbf{h} \notin H(t_1, \ldots, t_d, d) \text{ mais que } r(I_\mathbf{h}) \le t_1, \\ 0 & \text{sinon,} \end{cases}$$

la mesure de discrépance $D_w'(P_n)$ définie par (2.8) est égale à $M_{t_1,\ldots,t_d}^{-1}$.

Démonstration : voir l'annexe B, page xxx.

Remarque 3.5.2 *La définition des poids $w(\mathbf{h})$ pour obtenir l'équivalence dans la proposition 3.5.1 n'est pas unique. Nous avons choisi celle qui donnait au plus grand nombre possible de vecteurs $\mathbf{h}$ un poids non nul.*

On peut voir que les $\mathbf{h}$ qui sont dans $H(t_1, \ldots, t_d, d)$ ont plus de poids que les autres. En effet, la borne $l^*_{|I_\mathbf{h}|}(N)$ sur la norme des vecteurs qui sont dans $H(t_1, \ldots, t_d, d)$ est sévère (grande) : pour un $\mathbf{h}$ donné, $\mathbf{h}_{I_\mathbf{h}}$ ne peut appartenir à $L_I^\perp$ si $|I| < |I_\mathbf{h}|$ et $l^*_t(N)$ décroît avec t, donc la borne supérieure la plus grande (sévère) que l'on puisse utiliser sur $\|\mathbf{h}\|_2 = \|\mathbf{h}_{I_\mathbf{h}}\|_2$ est $l^*_{|I_\mathbf{h}|}(N)$. Les vecteurs qui ne sont pas dans $H(t_1, \ldots, t_d, d)$ mais pour lesquels $r(I_\mathbf{h}) \leq t_1$ ont une borne $l^*_{r(I_\mathbf{h})}(N)$ moins sévère car $r(I_\mathbf{h}) > |I_\mathbf{h}|$ dans ce cas : on parlera alors de borne "lâche". Par exemple, si $s = 30$, $d = 3$, $t_1 = 16$, $t_2 = 12$ et $t_3 = 8$, alors pour un vecteur $\mathbf{h}_1$ tel que $I_{\mathbf{h}_1} = \{3, 8, 10\}$, on a $w(\mathbf{h}_1) = l^*_3(N)/\|\mathbf{h}_1\|_2$. Par contre, si $\mathbf{h}_2$ est tel que $I_{\mathbf{h}_2} = \{1, 5, 16\}$, alors $\mathbf{h}_2 \notin H(16, 12, 8, 3)$, mais $r(I_{\mathbf{h}_2}) = 16 \leq t_1$ et donc, $w(\mathbf{h}_2) = l^*_{16}(N)/\|\mathbf{h}_2\|_2$. Ainsi, on accorde beaucoup moins d'importance à $\mathbf{h}_2$ qu'à $\mathbf{h}_1$, puisque $(l^*_{16}(N) = 0.05887N^{1/16}) < (l^*_3(N) = \sqrt{2}N^{1/3})$, en utilisant les constantes de Rogers et d'Hermite, respectivement.

3.5.5 Choix des paramètres d, $t_1, \ldots, t_d$

Ceci nous amène à la question suivante : comment choisir d et la suite $t_1, \ldots, t_d$? Si on prend $d = 1$, on retombe sur le critère M_{t_1}. Si on prend $d = s$, $t_1 = \ldots = t_s = s$, la fonction $w(\mathbf{h})$ correspondante est alors définie par $w(\mathbf{h}) = l^*_{|I_\mathbf{h}|}/\|\mathbf{h}\|_2$ pour tout $\mathbf{h}$ et est donc plus régulière, mais ce choix de paramètres requiert de calculer l_I pour tous les $I \subseteq S$ non vides tels que $i_1 = 1$. On doit donc faire un compromis entre la rapidité de calcul et la régularité de la fonction $w(\mathbf{h})$ déterminée par les paramètres $d, t_1, \ldots, t_d$.

La raison pour laquelle on préfère avoir une fonction $w(\mathbf{h})$ qui soit régulière est que, comme nous l'avons vu en introduisant les critères $D_w(P_N)$ et $D'_w(P_N)$ en (2.8) à la page 31, les $w(\mathbf{h})$ sont supposés approximer le comportement des $\hat{f}(\mathbf{h})$. Ainsi, si les $w(\mathbf{h})$ changent de façon abrupte sur $\mathbb{Z}^s$, cela revient à dire que l'on suppose que les $\hat{f}(\mathbf{h})$ ont aussi ce comportement, ce qui ne réflète pas nécessairement les hypothèses réelles que l'on fait sur la fonction.

En ce qui concerne la rapidité de calcul, on doit regarder le nombre de projections

à considérer (pour lesquels l_I est calculé) et ce nombre est donné par

$$N(t_1,\ldots,t_d) = \sum_{u=2}^{d} \binom{t_u - 1}{u-1} + (t_1 - d).$$

Regardons comment se comporte $N(t_1,\ldots,t_d)$ quand d augmente :

$$\begin{aligned} N(t_1,\ldots,t_d) &= \sum_{u=2}^{d} \frac{(t_u-1)\ldots(t_u-u+1)}{(u-1)\ldots 1} + (t_1 - d) \\ &\approx \sum_{u=1}^{d} t_u^u. \end{aligned} \tag{3.25}$$

Donc, si $t_1 = \ldots = t_d = t$, alors $N(t_1,\ldots,t_d) \in O(dt^d)$. Cela signifie que d pourra difficilement dépasser 4 ou 5 en pratique. En fait, l'approximation (3.25) nous indique que les t_u devraient être choisis de façon à ce que t_u décroisse avec u, puisque chaque t_u est élevé à la puissance u dans cette approximation de $N(t_1,\ldots,t_d)$. Par exemple, si on prend $t_1 = \ldots = t_4 = 32$, $N(t_1,\ldots,t_4) = 5022$ alors que $N(32,24,12,8) = 144$. De plus, le temps de calcul de l_I augmente avec $|I|$.

Pour ce qui est de la fonction $w(\mathbf{h})$, on peut voir qu'il y a deux frontières dans $L^\perp$ où elle change de façon abrupte :

(1) Lorsque les $\mathbf{h}$ passent de $H(t_1,\ldots,t_d,d)$ à $\{\mathbf{h} \in L^\perp \backslash H(t_1,\ldots,t_d,d) : r(I_\mathbf{h}) \leq t_1\}$: à ce moment, on passe de la borne serrée $l^*_{|I_\mathbf{h}|}(N)$ à la borne lâche $l^*_{r(I_\mathbf{h})}(N)$ dans $w(\mathbf{h})$.

(2) Lorsque $\mathbf{h}$ est tel que $r(I_\mathbf{h})$ devient supérieur à t_1 : on passe alors à un poids nul.

Or, plus ces frontières se situent en un endroit où les $\mathbf{h}$ sont grands, moins le changement est abrupt. Comment choisir les t_u pour que cela arrive ? Deux facteurs doivent être pris en compte, qui correspondent respectivement aux deux types de frontières décrits plus haut : 1) plus u est petit, plus on veut que t_u soit grand pour que les $\mathbf{h}$ soient assez grands ; 2) on veut que t_1 soit assez grand. Suivant cela, le compromis semble être de choisir $t_1 \geq t_2 \geq \ldots \geq t_d$ avec t_1 assez grand et fixer d et $t_2,\ldots,t_d$ pour que le critère soit calculable en pratique.

3.5.6 Lien avec d'autres critères

Nous voulons maintenant comparer le critère $M_{t_1,\ldots,t_d}$ avec le $\tilde{P}^s_\alpha$ introduit par Hickernell. Premièrement, $M_{t_1,\ldots,t_d}$ utilise le supremum dans (2.8) plutôt que la somme.

Ensuite, les poids β_I dans $\tilde{P}_\alpha^s$ sont de type produit alors que ceux dans $M_{t_1,\ldots,t_d}$ sont de la forme générale β_I (et correspondent aux quantités $(l_I^*(N))^{1/2}, (l_{r(I_h)}^*(N))^{1/2}$ et 0 dans la définition de $w(\mathbf{h})$ donnée à la proposition 3.5.1). Une autre différence est que les poids β_I dans $M_{t_1,\ldots,t_d}$ dépendent du nombre de points N. Cela pourrait être aussi le cas pour $\tilde{P}_\alpha^s$, mais dans [46, 47], on dit que les poids devraient plutôt dépendre du type de fonction à intégrer et donc, être indépendants de la règle utilisée.

Une variante de $M_{t_1,\ldots,t_d}$ serait de regrouper les projections ayant le même nombre de dimensions et de faire une moyenne parmi elles, puis de prendre le minimum parmi 1) ces moyennes et 2) les $l_t/l_t^*(N)$ pour les projections à dimensions successives. Plus précisément, on peut définir :

$$\bar{M}_{t_1,\ldots,t_d} = \min\left\{ \min_{2\le j\le t_1} l_j/l_j^*(N), \min_{2\le u\le d}\left[\frac{1}{n_u}\sum_{I\in S(t_u,u)} l_I/(l_I^*(N))\right]\right\}, \tag{3.26}$$

où $n_u = \binom{t_u-1}{u-1}$ est le nombre de sous-ensembles I entrant dans la moyenne. Cette définition constitue une alternative intéressante si l'on juge que les projections $P_N(I)$ sur les sous-ensembles I à indices successifs sont plus importantes que celles sur des indices non successifs. Par contre, la mesure $M_{t_1,\ldots,t_d}$ détecte plus facilement les ensembles de points P_N ayant de mauvaises projections. En effet, si un de ces ensembles d'indices I est tel que la qualité de $P_N(I)$ telle que mesurée par l_I est beaucoup moins bonne que celle du meilleur réseau en $|I|$ dimensions, donnée par $l_{|I|}^*(N)$, alors ce mauvais I ne pourra être caché dans la moyenne comme c'est le cas pour $\bar{M}_{t_1,\ldots,t_d}$. Nous avons observé ce fait empiriquement dans des expériences numériques qui se trouvent à l'annexe C.

Une autre façon de généraliser $M_{t_1,\ldots,t_d}$ serait de pondérer les différents $l_I/l_I^*(N)$ par des poids β_I entre 0 et 1, pour $I \in \bigcup_{u=2}^d S(t_u,u)$ ou $I \subseteq \{1,\ldots,t_1\}$:

$$\tilde{M}_{t_1,\ldots,t_d} = \min\left(\min_{2\le j\le t_1} \beta_j l_j/l_j^*(N), \min_{2\le u\le d}\min_{I\in S(t_u,u)} \beta_I l_I/l_{|I|}^*(N)\right),$$

où $\beta_j = \beta_{\{1,\ldots,j\}}$. Le désavantage de ce critère est que tout comme pour le critère $\tilde{P}_\alpha^s$, on doit choisir (de façon arbitraire) plusieurs poids β_I, en plus des paramètres $d, t_1, \ldots, t_d$.

3.6 Résultats numériques

À la sous-section 3.6.1, nous comparons pour plusieurs valeurs de N les règles obtenues à la suite de recherches utilisant différentes valeurs de d et $t_1, \ldots, t_d$ dans la définition du critère de sélection $M_{t_1,\ldots,t_d}$. Le logiciel `LatMRG` [76, 77] que nous avons utilisé pour faire ces recherches exhaustives permet de calculer l_I efficacement pour n'importe quel I, disons, tel que $|I| \leq 40$. Ensuite, aux sous-sections 3.6.2 et 3.6.3, nous comparons certaines règles obtenues à la sous-section 3.6.1 sur une fonction-test et sur le problème des options asiatiques, respectivement. Finalement, à la sous-section 3.6.4, nous comparons les règles choisies à l'aide du critère $M_{t_1,\ldots,t_d}$ avec celles choisies à l'aide de $\tilde{P}_\alpha^s$ sur le problème des options asiatiques.

3.6.1 Tableaux de règles choisies avec le nouveau critère

Dans le tableau 3.1, nous donnons les meilleurs a obtenus en utilisant les critères M_8, $M_{8,8}$, $M_{8,8,8}$ et $M_{8,8,8,8}$. Pour chaque valeur de N, la recherche est faite sur tous les éléments primitifs modulo N, c.-à-d., sur tous les a définissant des GCL à période maximale. Le choix de prendre $t_1 = \ldots = t_d$ correspond en fait à la première version que nous avons proposée pour généraliser le critère M_T [85]. Les étoiles $*$ dénotent le meilleur a par rapport au critère indiqué dans la colonne associée. Nous donnons également pour chaque a la valeur obtenue par rapport aux autres critères. La dernière ligne du tableau indique le nombre de projections pour lesquelles la longueur du plus court vecteur a été évaluée afin de calculer le critère correspondant.

Pour $N = 131071$, le meilleur a par rapport à M_8 performe très mal par rapport aux autres critères. Les ensembles de points basés sur ce GCL ont de mauvaises projections $P_N(I)$ pour certaines paires $I = \{i, j\}$, avec $j - i \leq 7$. D'un autre côté, on peut voir que les meilleurs a par rapport à $M_{8,8,8}$ n'obtiennent jamais de petites valeurs par rapport aux autres critères. De plus, il semble qu'à mesure que d augmente, il arrive de moins en moins souvent que le meilleur a par rapport au critère allant jusqu'à $d-1$ ne soit pas très bon par rapport à celui qui va jusqu'à d, en comparaison avec le meilleur a par rapport à ce critère, c.-à-d., il arrive que le meilleur a par rapport à M_8 soit très mauvais par rapport à $M_{8,8}$, mais cela se produit beaucoup moins souvent quand

TABLEAU 3.1: Meilleurs a par rapport à $M_{t_1,\ldots,t_d}$, où $t_1 = \ldots = t_d$, $1 \leq d \leq 4$, pour différentes valeurs de N.

N	a	M_8	$M_{8,8}$	$M_{8,8,8}$	$M_{8,8,8,8}$
32749	219	0.71802*	0.54037	0.22793	0.22793
	3042	0.65164	0.65164*	0.27985	0.27955
	8801	0.49395	0.46549	0.46549*	0.32481
	307	0.41208	0.41208	0.39577	0.38533*
65521	17364	0.70713*	0.39281	0.33148	0.28788
	6060	0.61743	0.61743*	0.24005	0.22910
	15409	0.48026	0.48026	0.46670*	0.22299
	332	0.56644	0.53533	0.41988	0.38979*
131071	43165	0.70941*	0.02425	0.02425	0.02425
	52344	0.66838	0.66695*	0.17004	0.17004
	38429	0.48814	0.47986	0.47986*	0.34233
	9290	0.58448	0.47631	0.47516	0.41693*
Nb. de projections		7	13	33	67

on regarde la valeur de $M_{8,8,8,8}$ du meilleur a par rapport à $M_{8,8,8}$. Donc, parmi les quatre critères utilisés dans le tableau 3.1, on peut penser que de choisir le meilleur a par rapport à $M_{8,8,8}$ pour construire un estimateur semble un compromis raisonnable. Dans les deux prochaines sous-sections, nous allons comparer le meilleur a par rapport à M_8 avec le meilleur par rapport à $M_{8,8,8}$ sur différents problèmes.

La structure du tableau 3.2 est similaire à celle du tableau 3.1, mais nous comparons maintenant les meilleurs a par rapport aux critères M_{32}, $M_{32,24,12,8}$ et $M_{32,24,16,12}$.

Dans le tableau 3.2, on voit que parmi les meilleures règles par rapport à M_{32}, certaines sont mauvaises par rapport aux critères qui considèrent les projections sur des dimensions non successives (par exemple, lorsque $N = 8191$, 65521 et 131071). Les meilleurs a par rapport à $M_{32,24,12,8}$ ont une valeur relativement bonne pour M_{32} et sont habituellement bons également par rapport à $M_{32,24,16,12}$. Évidemment, puisque $M_{32,24,16,12}$ est le critère qui regarde le plus de projections parmi les trois, les meilleurs GCL par rapport à ce critère ne sont jamais mauvais par rapport aux deux autres critères.

En résumé, nous suggérons d'utiliser les règles se trouvant dans les tableaux 3.1 et 3.2 et qui maximisent les critères $M_{8,8,8}$, $M_{8,8,8,8}$, $M_{32,24,12,8}$ ou $M_{32,24,16,12}$. Le choix

TABLEAU 3.2: Meilleurs a par rapport à $M_{t_1,\ldots,t_d}$ pour certaines valeurs de $(d, t_1, \ldots, t_d)$ et N.

N	a	M_{32}	$M_{32,24,12,8}$	$M_{32,24,16,12}$
1021	331	0.61872*	0.09210	0.09210
	76	0.53757	0.29344*	0.21672
	306	0.30406	0.26542	0.26542*
2039	393	0.65283*	0.15695	0.15695
	1487	0.49679	0.32196*	0.17209
	280	0.29807	0.25156	0.25156*
4093	219	0.66150*	0.13642	0.13642
	1516	0.39382	0.28399*	0.20839
	1397	0.40722	0.27815	0.27815*
8191	1716	0.64854*	0.05243	0.05243
	5130	0.50777	0.30676*	0.10826
	7151	0.47395	0.28809	0.28299*
16381	665	0.65508*	0.15291	0.14463
	4026	0.50348	0.29139*	0.23532
	5693	0.52539	0.26800	0.25748*
32749	9515	0.67356*	0.29319	0.13061
	14251	0.50086	0.32234*	0.12502
	8363	0.41099	0.29205	0.28645*
65521	2469	0.63900*	0.17455	0.06630
	8950	0.55678	0.34307*	0.20965
	944	0.39593	0.28813	0.26280*
131071	29803	0.66230*	0.03137	0.03137
	28823	0.44439	0.33946*	0.15934
	26771	0.54482	0.29403	0.29403*
Nb. de projections		31	141	321

entre ces quatre critères devrait être basé sur l'information que l'on a quant à la dimension effective de la fonction et en cas de doute, nous suggérons de prendre $M_{32,24,16,12}$. Nous pensons que les règles basées sur ce critère permettent de construire un estimateur $\hat{\mu}_{\mathrm{LR}}$ qui devrait avoir une variance inférieure à celle de l'estimateur MC pour une bonne majorité des problèmes que l'on rencontre en pratique. Ceci sera illustré par l'exemple des options asiatiques à la sous-section 3.6.3.

3.6.2 Résultats sur une fonction-test

Nous considérons ici le même exemple que dans [27]. Le polynôme utilisé dans cet article, défini par

$$f(x_1, \ldots, x_s) = (c_1 x_1 + \ldots + c_s x_s)^\alpha ,$$

a une dimension superposée égale à $\alpha = 2$ ou $\alpha = 3$, indépendamment de la dimension s. Pour cette raison, on s'attend à ce qu'un critère de la forme M_{t_1,t_2,t_3} soit plus approprié qu'un de la forme M_{t_1}. Nous avons modifié l'expérience par rapport à ce qui est fait dans [27]. En effet, dans cet article, on mesure la qualité des approximations à l'aide de leur erreur relative moyenne, qui est obtenue en générant aléatoirement et de façon indépendante 20 fois les paramètres c_j alors qu'ici, nous utilisons la variance empirique des estimateurs basés sur 50 copies i.i.d. de la règle de réseau translatée aléatoirement. Le tableau 3.3 contient la moyenne des facteurs de réduction de variance (par rapport à l'estimateur MC qui utilise $50N$ répétitions i.i.d.) qui ont été obtenus sur 10 exemplaires i.i.d. de la fonction f ; à chaque fois, les paramètres c_j sont générés aléatoirement sur $[0, 1)$ de façon indépendante. Les deux règles de réseau qui sont utilisées ont été choisies à l'aide des critères M_8 et $M_{8,8,8}$. Pour chaque combinaison (N, s), le plus grand facteur moyen est indiqué par une étoile (*).

TABLEAU 3.3: Facteurs de réduction de variance moyens, $\alpha = 3$

	N	$s = 5$	$s = 10$	$s = 15$	$s = 20$	$s = 25$	$s = 30$
$M_{8,8,8}$	32749	3000*	2340*	1960*	2320*	2080*	1630*
M_8	32749	170	468	746	1170	1540	1180
$M_{8,8,8}$	65521	4380*	4310*	5820*	5530	3410	3230*
M_8	65521	2980	3840	3520	5870*	4090*	2700
$M_{8,8,8}$	131071	4140*	2640*	3070*	4210*	4520*	2920*
M_8	131071	638	11.1	16.9	20.8	31.3	42.1

Comme on peut le voir, les meilleurs a par rapport à $M_{8,8,8}$ donnent généralement de meilleurs résultats que les meilleurs par rapport à M_8. En fait, ces derniers ont parfois une variance plus de 50 fois supérieure à celle de l'estimateur basé sur $M_{8,8,8}$ lorsque leur valeur de $M_{8,8,8}$ est petite (par exemple, pour $N = 131071$, quand $s \geq 10$). Quand $M_{8,8,8}$ est utilisé pour choisir la règle, la variance est réduite par un facteur

d'au moins 1600 en comparaison avec MC. Quand le meilleur a par rapport à M_8 a une valeur raisonnable de $M_{8,8,8}$, les deux estimateurs LR ont des variances à peu près comparables.

Dans [27], les règles basées sur M_8 sont comparées à des (t, m, s)-réseaux. Parmi les valeurs de N considérées (voir leur tableau 3), le cas où $N = 131071$ est le seul où la règle de réseau a une erreur significativement supérieure à celle du (t, m, s)-réseau. Comme c'était le cas pour la comparaison avec MC, la petite valeur de $M_{8,8,8}$ pour cette règle explique cette mauvaise performance. En utilisant la meilleure règle par rapport à $M_{8,8,8}$ pour cette valeur de N, l'erreur se rapproche de celle du (t, m, s)-réseau, même qu'elle est inférieure en dimension $s \geq 20$. Nous reproduisons au tableau 3.4 les résultats de [27] pour $N = 131071$ et donnons l'erreur obtenue en utilisant la règle choisie à l'aide du critère $M_{8,8,8}$.

TABLEAU 3.4: Erreurs relatives moyennes, $\alpha = 3$ [27]

	N	$s = 5$	$s = 10$	$s = 15$	$s = 20$	$s = 25$	$s = 30$
tms	131072	1.65e-5*	3.16e-5*	5.45e-5*	1.31e-4	1.65e-4	1.85e-4
M_8	131071	1.34e-3	2.58e-3	2.25e-3	1.85e-3	1.64e-3	1.43e-3
$M_{8,8,8}$	131071	1.79e-4	2.00e-4	1.55e-4	1.03e-4*	6.92e-5*	8.96e-5*

Puisque la dimension s va jusqu'à 30 dans cet exemple, nous avons regardé si le fait d'utiliser un critère de la forme M_{t_1,t_2,t_3} avec $t_2, t_3 > 8$ permettait d'obtenir de meilleurs résultats. Nos expériences numériques sembent indiquer que cela n'améliore pas de façon importante la performance des règles de réseau. Ces résultats se trouvent à l'annexe D.

Précisons que le but de cette expérience n'était pas de déterminer quel est le meilleur critère pour ce type de fonction-test, mais plutôt d'illustrer l'importance d'utiliser des règles de réseau pour lesquelles les projections $P_N(I)$ sont de bonne qualité pour plusieurs sous-ensembles I, surtout ceux pour lesquels $|I|$ est petit. L'exemple étudié à la sous-section suivante va dans le même sens et puisqu'il s'agit d'un problème "naturel", cela montre que l'utilisation de critères de la forme $M_{t_1,\ldots,t_d}$ avec $d > 1$ n'est pas seulement importante dans le contexte (plutôt artificiel) considéré dans la présente sous-section.

3.6.3 Résultats sur le problème des options asiatiques

Nous donnons maintenant des résultats comparant différentes règles données dans les tableaux de la section 3.6.1 sur le problème des options asiatiques, que nous avons décrit à la section 2.5.2. Dans le tableau 3.5, les paramètres du modèle sont $T = 120$ jours, $T_1 = T - s$ jours, $\sigma = 0.2$, $r = \ln 1.09$ et $S(0) = 100$. Lorsque des variables antithétiques et la variable de contrôle décrite à la section 2.5.2 sont utilisées, nous dénotons cela par l'estimateur "ACV" : sans ces techniques, on a l'estimateur "naïf". Nous utilisons 100 translations aléatoires indépendantes pour estimer la variance des estimateurs basés sur une règle de réseau. Chaque valeur dans le tableau correspond au facteur de réduction de variance amené par la règle de réseau considérée, par rapport à l'estimateur MC correspondant qui utilise $100N$ répétitions indépendantes.

TABLEAU 3.5: Facteurs de réduction de variance estimés

s	N	a	$K = 90$	$K = 100$	$K = 110$	$K = 90$	$K = 100$	$K = 110$
			estimateur naïf			estimateur ACV		
10	32749	$M_{8,8,8}$	3600	2500	910	36	19	8.3
	32749	M_8	3100	1400	550	7.3	6.6	5.3
	65521	$M_{8,8,8}$	7900	4500	1500	45	25	9.8
	65521	M_8	3800	2900	1000	41	14	7.3
	131071	$M_{8,8,8}$	7200	2300	1100	43	19	14
	131071	M_8	47	46	14	0.40	0.49	0.81
60	32749	$M_{8,8,8}$	270	81	16	13	6.4	2.5
	32749	M_8	230	58	22	9.4	5.7	2.9
	65521	$M_{8,8,8}$	320	88	23	20	7.2	2.9
	65521	M_8	373	130	23	12	7.3	2.8
	131071	$M_{8,8,8}$	350	60	14	16	6.8	1.8
	131071	M_8	20	7.0	2.4	0.29	0.36	0.52

Comme pour l'exemple de la sous-section précédente, le fait d'utiliser le critère M_8 pour choisir a peut donner de mauvais résultats lorsque l'une des projections $P_N(I)$ a une mauvaise distribution pour certains triplets ou paires I. Le critère $M_{8,8,8}$ évite ces règles et fournit des estimateurs qui réduisent significativement la variance en comparaison avec MC pour ce problème, même en 60 dimensions et lorsque les techniques de réduction de la variance sont utilisées. Les règles basées sur M_8 font pire

que MC pour les estimateurs ACV lorsque $N = 131071$. La variance est réduite par des facteurs allant jusqu'à 45 en utilisant $M_{8,8,8}$ plutôt que M_8. Quand le meilleur a par rapport à M_8 a une valeur raisonnable de $M_{8,8,8}$, les deux estimateurs LR offrent une performance comparable, avec une léger avantage envers ceux basés sur $M_{8,8,8}$.

Pour donner une idée de la précision des facteurs donnés au tableau 3.5, les intervalles de confiance au niveau 98% pour le rapport théorique des variances associés aux deux dernières lignes (cas où $(N, s) = (131071, 60)$) sont : (245, 476), (42, 82), (9.8, 19), (11, 22), (4.8, 9.2) et (1.3, 2.4) pour $M_{8,8,8}$ et (14, 27), (4.9, 9.5), (1.7, 3.3), (0.20, 0.39), (0.25, 0.49) et (0.36, 0.71) pour M_8. Donc, même en tenant compte du bruit sur ces estimations, on peut dire que la variance associée au critère M_8 pour cette valeur de (N, s) et pour la méthode ACV est significativement supérieure à celle de l'estimateur MC.

Dans le tableau suivant, on refait la même expérience, mais en comparant le meilleur a par rapport à M_{32} avec le meilleur par rapport à $M_{32,24,16,12}$. Le problème est en 60 dimensions et les paramètres du modèle sont $T = 1$ an, $T_1 = 0$, $\sigma = 0.2$, $r = 0.05$ et $S(0) = 100$.

TABLEAU 3.6: Facteurs de réduction de variance estimés, $s = 60$

N	a	$K =$	90	100	110	90	100	110
			estimateur naïf			estimateur ACV		
1021	$M_{32,24,16,12}$		25	11	3.8	2.5	2.5	1.7
1021	M_{32}		0.4	0.4	0.4	0.3	0.3	0.3
8191	$M_{32,24,16,12}$		53	20	4.8	5.1	3.8	2.9
8191	M_{32}		3.2	1.2	0.7	0.2	0.2	0.3
32749	$M_{32,24,16,12}$		126	44	9.5	13	7.5	3.8
32749	M_{32}		108	32	9.2	7.7	5.7	2.4
131071	$M_{32,24,16,12}$		91	25	12	9.0	6.4	3.5
131071	M_{32}		13	4.4	3.4	1.2	1.3	1.2

Ici encore, avec des paramètres légèrement différents, on observe le même phénomène : lorsque le meilleur a par rapport au critère qui ne considère que les projections successives a de mauvaises projections sur des indices non successifs, comme c'est le cas pour tous les N sauf $N = 32749$, l'estimateur associé performe significativement moins bien que celui associé au a choisi à l'aide du critère plus général, qui ici est

$M_{32,24,16,12}$, puisque ce dernier détecte le mauvais comportement de ce type de règle et choisit une règle n'ayant pas ces problèmes.

3.6.4 Comparaison entre $M_{t_1,\ldots,t_d}$ et $\tilde{P}_\alpha^s$

Dans cette sous-section, nous comparons les critères $M_{t_1,\ldots,t_d}$ et $\tilde{P}_\alpha^s$ sur le problème des options asiatiques, afin de voir si l'un de ces critères considérant les projections $P_N(I)$ permet d'obtenir des estimateurs à plus petite variance que l'autre sur ce problème pratique. Ces deux critères seront également comparés sur des fonctions-test et sur le même problème d'options (avec des valeurs de N et s différentes) au chapitre 5, lorsque nous comparerons différentes règles de Korobov à des règles de type ν^r-copie. Dans ce qui suit, nous nous sommes surtout intéressés au cas où N est assez petit, car il n'est pas nécessaire que N soit grand pour obtenir de bons estimateurs sur ce type de problème et notre but est de voir comment ces deux critères se comparent dans un contexte qui soit le plus réaliste possible.

Dans le tableau 3.8, nous donnons les facteurs de réduction de variance empiriques par rapport à l'estimateur $\hat{\mu}_{\mathrm{MC}}$ obtenus à l'aide des règles choisies soit avec $M_{8,8,8}$, soit avec le critère $\tilde{P}_2^8$ qui utilise les poids $\beta_1 = \ldots = \beta_8 = \sqrt{3/(8\pi^2)}$. La justification de ce choix de poids est donnée dans [46, page 144] : cette valeur de β fait en sorte que $\tilde{P}_2^s$ devient ainsi égal à la "discrépance $\mathcal{L}^2$-étoile", un cas particulier de la mesure générale $D(P_N)$ donnée à l'inégalité (1.6) et qui est souvent utilisée dans l'étude des (t, m, s)-réseaux. Dans les deux cas, la recherche est faite sur tous les éléments primitifs modulo N. Le tableau 3.7 donne la valeur de a pour les règles de Korobov ainsi trouvées ainsi que leur valeur de $M_{8,8,8}$ et $\tilde{P}_2^8$. Nous utilisons 100 translations aléatoires pour estimer la variance de chacun de ces deux estimateurs et l'estimateur $\hat{\mu}_{\mathrm{MC}}$ est basé sur $100N$ répétitions i.i.d., afin que la comparaison soit juste. Nous étudions les estimateurs "naïf" et "ACV", comme à la sous-section précédente et considérons le problème lorsque $s = 10$ et $s = 60$. Les paramètres de l'option sont définis de la même façon qu'au tableau 3.5 de la sous-section précédente.

D'après ces résultats, on ne peut pas dire qu'un critère fait généralement mieux que l'autre sur ce problème. Dans plusieurs cas, il n'y a pas de différence significative

TABLEAU 3.7: Meilleurs a par rapport à $M_{8,8,8}$ et $\tilde{P}_2^8$.

N	Crit.	a	$M_{8,8,8}$	$\tilde{P}_2^8$
1021	$M_{8,8,8}$	325	0.40544*	1.1544e-4
	$\tilde{P}_2^8$	103	0.23481	1.0245e-4*
2039	$M_{8,8,8}$	1067	0.40966*	4.6785e-5
	$\tilde{P}_2^8$	110	0.37835	2.7617e-5*
4093	$M_{8,8,8}$	444	0.42049*	1.8267e-5
	$\tilde{P}_2^8$	2665	0.29993	1.0541e-5*
8191	$M_{8,8,8}$	1932	0.42850*	6.1724e-6
	$\tilde{P}_2^8$	6664	0.29819	3.7431e-6*

TABLEAU 3.8: Facteurs de réduction de variance estimés

s	N	a	$K=90$	$K=100$	$K=110$	$K=90$	$K=100$	$K=110$
			estimateur naïf			estimateur ACV		
10	1021	$M_{8,8,8}$	215	224	76	5.9	5.9	2.5
	1021	$\tilde{P}_2^8$	239	149	92	4.1	5.0	3.5
	2039	$M_{8,8,8}$	444	239	117	13	8.3	2.5
	2039	$\tilde{P}_2^8$	454	416	144	6.7	7.5	1.6
	4093	$M_{8,8,8}$	537	298	140	8.4	5.2	5.3
	4093	$\tilde{P}_2^8$	818	427	228	18	12	6.0
	8191	$M_{8,8,8}$	1198	464	277	21	12	5.9
	8191	$\tilde{P}_2^8$	1308	1047	375	18	13	6.6
60	1021	$M_{8,8,8}$	78	25	8.0	4.0	3.3	1.6
	1021	$\tilde{P}_2^8$	104	26	10	3.6	2.5	2.3
	2039	$M_{8,8,8}$	107	25	8.6	3.1	2.6	1.9
	2039	$\tilde{P}_2^8$	77	19	5.9	1.9	2.1	1.5
	4093	$M_{8,8,8}$	47	18	4.4	1.4	1.5	1.2
	4093	$\tilde{P}_2^8$	140	39	12	4.1	3.2	2.2
	8191	$M_{8,8,8}$	215	79	19	9.4	6.6	2.7
	8191	$\tilde{P}_2^8$	108	35	12	3.3	3.8	1.9

entre les deux variances au niveau de confiance 98% (lorsque le rapport est entre 0.70 et 1.36, selon la statistique de Fischer). Par contre, il semble que lorsque $s = 60$ et que N vaut 4093, la règle choisie avec $\tilde{P}_2^8$ obtient une plus petite variance que celle choisie avec $M_{8,8,8}$ et c'est le contraire lorsque N vaut 8191.

Pour donner une idée de la précision de ces rapports, au niveau de confiance 98%, les intervalles de confiance pour les rapports théoriques des variances lorsque $N = 8191$ et $s = 10$ sont (838, 1629), (324, 631), (194, 377), (15, 29), (8.4, 16), (4.1, 8.0) pour

la règle choisie avec $M_{8,8,8}$ et (915, 1779), (732, 1424), (262, 510), (13, 24), (9.1, 18) et (4.6, 9.0) pour la règle choisie avec $\tilde{P}_2^8$.

Que conclure de cela? Les résultats de cette section nous indiquent qu'il est important d'utiliser un critère qui considère plus attentivement les projections que ce que fait P_α^s, ρ ou M_T. Les critères $M_{t_1,\ldots,t_d}$ et $\tilde{P}_\alpha^s$ font tous les deux cela et semblent permettre de construire des estimateurs d'à peu près semblable qualité. Nous avons tendance à favoriser le critère $M_{t_1,\ldots,t_d}$, étant donné qu'il permet de faire des recherches plus rapides que $\tilde{P}_\alpha^s$, surtout lorsque N et/ou s sont grands. Un autre avantage est qu'étant donné que $M_{t_1,\ldots,t_d}$ est toujours entre 0 et 1, il est plus facile d'avoir une idée de la qualité d'une règle en regardant sa valeur de $M_{t_1,\ldots,t_d}$ qu'en regardant sa valeur de $\tilde{P}_\alpha^s$, puisque cette dernière peut prendre n'importe quelle valeur positive. Des normalisations ont été proposées pour le critère $\tilde{P}_\alpha^s$ [47, 48], mais elles sont basées sur des taux de convergence asymptotiques optimaux plutôt que sur une borne provenant du réseau optimal et ne sont donc pas aussi utiles; entre autres, elles ne donnent pas nécessairement un nombre entre 0 et 1.

Une heuristique qui pourrait être étudiée dans le futur serait de faire d'abord une recherche exhaustive avec $M_{t_1,\ldots,t_d}$, retenir un certain nombre de règles parmi les meilleures qu'on a trouvées, puis utiliser $\tilde{P}_\alpha^s$ pour choisir une règle parmi ces meilleures. Autrement dit, on utiliserait $M_{t_1,\ldots,t_d}$ pour éliminer les mauvaises règles, mais on ferait le choix final à l'aide de $\tilde{P}_\alpha^s$. Ceci nous éviterait d'avoir à faire de trop longs calculs, car la recherche sur toutes les règles serait faite en utilisant $M_{t_1,\ldots,t_d}$.

Chapitre 4

Règles de réseau polynômiales

Nous étudions dans ce chapitre les règles de réseau polynômiales. Nous verrons dès la sous-section 4.1.6 que ces règles sont reliées de très près à la famille des (t, m, s)-réseaux, qui rappelons-le, est une famille de méthodes QMC très importante, souvent utilisée en pratique et étudiée par plusieurs (voir [97] et les références qui s'y trouvent, ainsi que [43, 100]). Nous parlons ici de certaines propriétés qu'ont les règles de réseau polynômiales en ce qui a trait à leurs liens avec les (t, m, s)-réseaux et à la variance des estimateurs qui leur sont associés. Nous proposons également un nouveau critère de sélection pour les choisir.

Afin de motiver la pertinence de ce critère de sélection, mentionnons que dans la littérature récente sur les méthodes QMC, plusieurs auteurs [109, 64, 47, 114] ont exprimé l'idée qu'il serait préférable de construire des (t, m, s)-réseaux qui minimisent (d'une certaine manière) un vecteur $(t_I)_{I \subseteq S}$ de paramètres tels que $P_N(I)$ est un $(t_I, m, |I|)$-réseau, plutôt que de ne regarder que la valeur de t, qui correspond en fait à $\max_{\emptyset \neq I \subseteq S} t_I$. Comme alternative à la construction de (t, m, s)-réseaux pour lesquels les t_I ont été minimisés, nous présentons dans ce chapitre une façon de construire des règles de réseau polynômiales pour lesquelles les *résolutions* ℓ_I associées aux différentes projections sont maximisées. Un avantage de cette approche est qu'il semble que le calcul de ℓ_I soit plus rapide à effectuer que celui de t_I.

Afin d'établir les différences entre ces deux constructions, nous donnons à la section 4.3 des résultats comparant ℓ_I, t_I ainsi qu'un paramètre q_I relié à t_I et définissons à

la section 4.4 un critère de sélection basé sur les ℓ_t pour choisir des règles de réseau polynômiales. Ensuite, à la section 4.5, nous donnons des résultats sur la variance des estimateurs formés à partir de règles de réseau polynômiales *XOR-translatées*. En posant de fortes conditions sur les fonctions à intégrer, nous trouvons des bornes sur la variance qui font intervenir le nouveau critère. Nous établissons aussi des liens avec les résultats qui existent sur la variance des (t, m, s)-réseaux brouillés et étudions l'ordre de convergence de la variance pour certains ensembles de fonctions. Puis, nous concluons le chapitre en présentant des tableaux de règles qui minimisent le critère présenté à la section 4.4 et des résultats numériques où ces règles sont utilisées afin de construire des estimateurs dont la variance empirique est inférieure à celle de l'estimateur MC. Ces résultats numériques nous permettent également de comparer empiriquement les règles de réseau standard avec la version polynômiale.

Mais tout d'abord, nous commençons par donner les notions de base sur les règles de réseau polynômiales et nous expliquons à la section 4.2 ce qu'est la décomposition en série de Walsh d'une fonction. Cette représentation joue le rôle qu'ont les séries de Fourier pour les règles de réseau standard. Notons qu'un certain nombre de résultats présentés dans ce chapitre sont carrément la "version polynômiale" de résultats discutés aux chapitres 2 et 3.

4.1 Introduction aux règles de réseau polynômiales

Nous définissons d'abord ce qu'est un générateur de Tausworthe, puisque c'est à partir de ces générateurs que nous construisons les règles de réseau polynômiales. Nous définissons aussi ce qu'est la résolution, avant d'introduire le concept de règle de réseau polynômiale et de XOR-translation. Nous nous référons à [72, 79, 80] pour cette partie. Ensuite, nous donnons la propriété définissant les (t, m, s)-réseaux et expliquons le lien entre cette propriété et la résolution. Nous expliquons également la randomisation proposée par Owen pour les (t, m, s)-réseaux, car nous la comparerons à la XOR-translation à la section 4.5.3.

4.1.1 Définition d'un générateur de Tausworthe

Un générateur de Tausworthe produit une suite de bits à l'aide d'une récurrence linéaire modulo 2 et les nombres u_n entre 0 et 1 sont formés en utilisant des blocs de bits successifs. Plus précisément, ce type de générateur est basé sur un polynôme $P(z) = z^m - a_1 z^{m-1} - \cdots - a_m$ de degré m à coefficients dans $\mathbb{F}_2$, le corps de Galois contenant deux éléments, et tel que $a_m \neq 0$. Il utilise la récurrence

$$\xi_n = a_1 \xi_{n-1} + \cdots + a_m \xi_{n-m} \bmod 2 \tag{4.1}$$

pour former la sortie

$$u_n = \sum_{i=1}^{L} \xi_{nv+i-1} 2^{-i}, \tag{4.2}$$

où v et L sont des entiers positifs. Le germe de la récurrence est dénoté par $\mathbf{s}_0 = (\xi_0, \ldots, \xi_{m-1}) \in \mathbb{F}_2^m$. Si $P(\cdot)$ est un polynôme primitif (voir [86] pour la définition de cette propriété), que $\mathbf{s}_0 \neq \mathbf{0}$ et que $\rho = 2^m - 1$ est relativement premier avec v, alors la suite des u_n (et celle des ξ_n) est purement périodique avec une période de ρ, qui est la période maximale pour une récurrence d'ordre m. La prochaine sous-section explique comment sont construits les générateurs de Tausworthe utilisés dans les expériences numériques de la section 4.6.

4.1.2 Implantation et générateurs combinés

Il existe des moyens efficaces pour calculer u_n à partir de u_{n-1} si on impose une certaine condition sur $P(\cdot)$ [72]. Cette condition est que $P(\cdot)$ doit être un trinôme primitif de la forme $P(z) = z^m - z^q - 1$, avec $0 < 2q < m$, $0 < v \leq m - q < m \leq L$, $\text{pgcd}(v, 2^m - 1) = 1$ et L doit être égal à la taille des mots de l'ordinateur. L'inconvénient est que les générateurs basés sur des polynômes qui respectent cette condition ont des défauts statistiques importants. Pour corriger ce problème, on peut combiner J générateurs de Tausworthe à période maximale basés respectivement sur les polynômes P_j de degré m_j et avec $v_j = v$, $j = 1, \ldots, J$. De plus, on peut montrer que ce générateur combiné est équivalent à un autre générateur de Tausworthe, mais qui a un polynôme caractéristique réductible correspondant au produit $P_1(z) \cdot \ldots \cdot P_J(z)$, dont le degré est $m = \sum_{j=1}^{J} m_j$. La période du générateur combiné est de

$\rho = \text{ppcm}(2^{m_1} - 1, \ldots, 2^{m_J} - 1)$, où "ppcm" signifie "plus petit commun multiple". Lorsque $\rho = \prod_{j=1}^{J}(2^{m_j} - 1)$, cet autre générateur de Tausworthe comprend 2^J sous-cycles correspondant aux différentes combinaisons (à période maximale) que l'on peut obtenir à partir des J composantes. Le sous-cycle $0, 0, \ldots$ correspond à ne combiner aucune composante. Dans les expériences numériques de la section 4.6, nous utilisons des générateurs à deux ou trois composantes.

4.1.3 Résolution

Nous expliquons dans cette sous-section ce qu'est la résolution associée à un générateur de Tausworthe. Ce concept est utilisé pour définir notre critère de sélection, ainsi que pour expliquer les connections avec les (t, m, s)-réseaux. Tout ce qui suit se trouve dans [72].

Pour mesurer l'uniformité de la suite des u_n produite par un générateur de Tausworthe, on peut choisir un entier positif p et regarder la distribution de l'ensemble des p-tuplets successifs pouvant être générés à partir de tous les états possibles, comme on le faisait pour les GCL, c.-à-d., on regarde

$$\Psi_p = \{(u_0, u_1, \ldots, u_{p-1}) : \mathbf{s}_0 \in \{0,1\}^m\}. \tag{4.3}$$

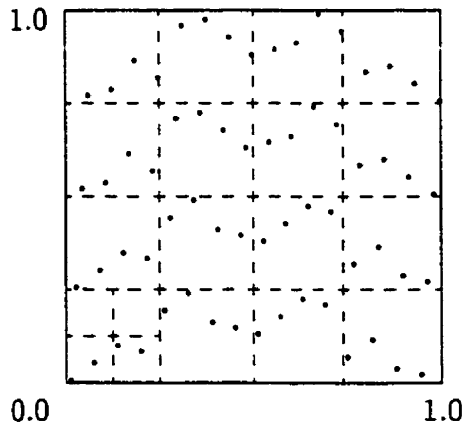
Une façon de mesurer l'uniformité de Ψ_p pour un générateur de Tausworthe dont le polynôme associé (réductible ou non) est d'ordre m est de former une partition de $[0,1)^p$ en 2^{pl} cellules cubiques de même volume, où l'on doit imposer $l \leq L$. On dit que Ψ_p est (p,l)-*équidistribué* si chacune des 2^{pl} cellules contient le même nombre de points, c.-à-d., 2^{m-pl} points. Évidemment, cela est possible seulement si $pl \leq m$. On donne dans [33, 18, 17] des conditions permettant de vérifier cette propriété, que nous expliquerons brièvement à la prochaine sous-section. La plus grande valeur de l pour laquelle Ψ_p est (p,l)-équidistribué est appelée la *résolution* et est dénotée par ℓ_p. La valeur maximale que peut prendre ℓ_p est donnée par

$$\ell_p^* = \min\left(L, \lfloor m/p \rfloor\right).$$

L'*écart de résolution* est défini par $\Gamma_p = \ell_p^* - \ell_p$ et un générateur est dit *à résolution maximale en dimension* p si $\Gamma_p = 0$. De plus, si $\Gamma_p = 0$ pour $p = 1, \ldots, m$, alors on dit

que le générateur est *à équidistribution maximale* (ME) [128]. Quand $p > m$, on obtient nécessairement que $\ell_p = 0$, puisqu'on a un plus grand nombre de cubes que de points. La figure 4.1 nous donne un exemple où le générateur est $(2,2)$-équidistribué, mais non pas $(2,3)$-équidistribué. Le polynôme utilisé pour générer cette règle est donné par $P(z) = z^6 - z - 1$ et les paramètres v et L sont $v = 4$ et $L = 6$.

FIGURE 4.1: Exemple d'ensemble Ψ_2 pour un générateur de Tausworthe



4.1.4 Règles de réseau polynômiales

Afin d'expliquer le concept de règle de réseau polynômiale, il faut d'abord dire qu'un générateur de Tausworthe peut être vu comme étant un GCL, mais qui évolue dans un espace de polynômes plutôt que dans $\mathbb{R}$ [18, 127]. L'équivalence se fait de la façon suivante [71] : on définit une transformation bijective entre l'espace $\mathbb{F}_2^m$ de la récurrence (4.1) et l'espace $\mathbb{F}_2[z]/(P)$ des polynômes de degré inférieur à m ayant des coefficients dans $\mathbb{F}_2$. À l'état $\mathbf{s}_n = (\xi_n, \ldots, \xi_{n+m-1})$, on associe le polynôme

$$p_n(z) = \sum_{j=1}^{m} c_{n,j} z^{m-j},$$

où

$$\begin{pmatrix} c_{n,1} \\ c_{n,2} \\ \vdots \\ c_{n,m} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ a_1 & 1 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ a_{m-1} & \ldots & a_1 & 1 \end{pmatrix} \begin{pmatrix} \xi_n \\ \xi_{n+1} \\ \vdots \\ \xi_{n+m-1} \end{pmatrix}. \tag{4.4}$$

On a alors

$$p_n(z) = zp_{n-1}(z) \bmod (P(z), 2), \tag{4.5}$$

où " mod $(P(z), 2)$" signifie qu'on prend le reste de la division polynômiale par $P(z)$, en supposant que les opérations sur les coefficients sont effectuées dans $\mathbb{F}_2$. Autrement dit, on a un GCL dans $\mathbb{F}_2[z]/(P)$, avec $P(z)$ comme modulo et z comme multiplicateur. Pour obtenir la sortie u_n définie en (4.2) (en supposant que $L = \infty$), on n'a qu'à évaluer $p_{nv}(z)/P(z)$ en $z = 2$.

Puisque la structure de réseau des $p_n(z)$ est préservée sous la transformation linéaire qui permet d'obtenir les u_n, l'ensemble de points Ψ_p défini en (4.3) a également une structure de réseau. C'est pour cette raison que nous parlons de "règle de réseau polynômiale" lorsque nous prenons P_N égal à Ψ_s, où $N = 2^m$. En dimension p, le réseau dual est l'espace $\mathcal{L}_p^*$ des polynômes multivariés $\mathbf{h}(z) = (h_1(z), \ldots, h_p(z))$, où $h_t(z) = \sum_{j=0}^{l-1} h_{t,j} z^j$, $h_{t,j} \in \mathbb{F}_2, l \in \mathbb{N}$, et tels que $\sum_{t=1}^{p} h_t(z) z^{(t-1)v} \bmod (P(z), 2) = 0$ [71, 127]. Par abus de notation, on peut identifier chaque polynôme $\mathbf{h}(z)$ avec le vecteur d'entiers $\mathbf{h} = (h_1, \ldots, h_p)$, où $h_t = \sum_{j=0}^{l-1} h_{t,j} 2^j \in \mathbb{N}$, et ainsi $\mathcal{L}_p^*$ peut également être vu comme un espace de vecteurs d'entiers $\mathbf{h}$. En utilisant la norme *sup* pour mesurer $\|\mathbf{h}\|$, on peut montrer [17] que la longueur du plus court vecteur dans $\mathcal{L}_p^*$ est 2^{ℓ_p}, c.-à-d.,

$$2^{\ell_p} = \min_{0 \neq \mathbf{h} \in \mathcal{L}_p^*} \|\mathbf{h}\| = \min_{0 \neq \mathbf{h} \in \mathcal{L}_p^*} \max_{1 \leq t \leq p} |h_t|_{\mathrm{p}}, \tag{4.6}$$

où $|h|_{\mathrm{p}} = 2^k$ si $2^k \leq h < 2^{k+1}$ et $|h|_{\mathrm{p}} = 0$ si $h = 0$. Plus précisément, l'égalité (4.6) est obtenue en combinant [17, *Theorem 1*] et [18, *Theorem 2*], ce qui nous permet d'obtenir la relation suivante :

$$\ell_p = \ell \text{ si et seulement si } \sum_{j=1}^{p} \max(0, \ell - \lg \sigma_j') = 0 \text{ et } \sum_{j=1}^{p} \max(0, \ell + 1 - \lg \sigma_j') > 0, \tag{4.7}$$

où $\lg = \log_2$ et σ_j' est la longueur (en utilisant la norme *sup*) du j^e vecteur d'une base réduite de Minkowski pour le réseau $\mathcal{L}_p^*$ (voir [17, *Theorem 1*]). Les propriétés de cette base font en sorte que l'on a $\sigma_1' \leq \ldots \leq \sigma_p'$ et $\sigma_1' = \min_{0 \neq \mathbf{h} \in \mathcal{L}_p^*} \|\mathbf{h}\|$. Ainsi, on obtient que $\ell_p = \ell$ si et seulement si $\lg \sigma_1' = \ell$. La relation (4.6) nous fournit donc une autre interprétation pour la résolution ℓ_p.

En pratique, lorsque l'on veut vérifier si $\ell_p = \ell_p^*$, on peut regarder si chacune des $2^{p\ell_p^*}$ combinaisons de points dans $[0, 1)^p$ contenant ℓ_p^* bits dans chaque dimension

apparaît autant de fois, lorsque l'on considère les bits correspondant qui représentent les 2^m points de Ψ_p. Pour faire cela, on doit vérifier que le système d'équations linéaires exprimant ces $2^{p\ell_p^*}$ combinaisons comme fonction de $(c_{0,1}, \ldots, c_{0,m})$ est indépendant, ce qui revient à vérifier que la matrice correspondante est de plein rang $p\ell_p^*$ [33, 72]. Dans [17], on montre qu'il est possible d'utiliser la relation (4.6) afin de calculer ℓ_p de façon plus efficace.

4.1.5 Randomisation par XOR-translation

Tout comme c'est le cas pour les règles de réseau standard, on peut randomiser les règles de réseau polynômiales en additionnant à chaque point de P_N un vecteur $\boldsymbol{U}$ uniformément distribué sur $[0,1)^s$. Cependant, comme R. Couture nous l'a fait remarquer, la contrepartie polynômiale est de faire un ou-exclusif bit par bit entre $\boldsymbol{U} = (U_1, \ldots, U_s)$ et chaque point de la règle, plutôt qu'une addition modulo 1. En pratique, le ou-exclusif est effectué sur les bits que l'on utilise pour représenter les nombres réels U_j et x_{ij}, pour $j = 1, \ldots, s$, où x_{ij} représente la j^e coordonnée du i^e point de la règle. Tout comme dans le cas standard, il suffit de répéter cette procédure un certain nombre de fois, disons avec M vecteurs i.i.d. uniformes $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_M$, afin d'obtenir un estimateur sans biais de la variance de l'estimateur basé sur les points $\{\mathbf{x}_i \oplus \boldsymbol{U}, i = 1, \ldots, N\}$. Cette *XOR-translation* préserve la propriété de réseau de l'ensemble de points et ceci sera exploité à la section 4.5, afin de donner des expressions pour la variance des estimateurs ainsi formés.

4.1.6 Définition des (t, m, s)-réseaux

Comme nous l'avons indiqué au début de ce chapitre, une des motivations derrière notre étude des règles de réseau polynômiales est que ces règles sont reliées de très près aux (t, m, s)-réseaux, qui constituent une des familles importantes de méthodes QMC. Or, la propriété définissant un (t, m, s)-réseau est basée sur la notion d'*intervalle élémentaire de* $[0,1)^s$ *en base* b ou *b-boîte*, qui est défini comme étant un ensemble de la forme

$$\prod_{j=1}^{s} \left[\frac{v_j - 1}{b^{k_j}}, \frac{v_j}{b^{k_j}} \right),$$

où k_j et $v_j \leq b^{k_j}$ sont des entiers non négatifs. Autrement dit, pour $j = 1, \ldots, s$, on coupe le j^e axe $[0,1)$ en b parties égales k_j fois, on prend le v_j^e sous-intervalle et on garde l'intersection formée par ces s sous-intervalles. Un (t, m, s)-réseau en base b est un ensemble de b^m points dans $[0,1)^s$ tel que chaque b-boîte de volume b^{t-m} contient exactement b^t points de l'ensemble. Évidemment, pour que cette propriété soit non triviale, on doit avoir $0 \leq t \leq m$. De plus, on suppose que le t donné est le plus petit entier pour lequel la propriété est vraie.

Ainsi, l'uniformité de cet ensemble dépend du paramètre t : plus t est petit, plus les points sont uniformément répartis. Par exemple, les suites de Niederreiter-Xing [102] ont la propriété de fournir des (t, m, s)-réseaux qui, pour b et m fixés, ont une valeur de t qui croît linéairement avec s, ce qui est le meilleur taux possible pour t en fonction de s. Associées aux (t, m, s)-réseaux, on a les (τ, s)-*suites*, qui sont des suites infinies de points $\mathbf{x}_1, \mathbf{x}_2, \ldots$ ayant la propriété que pour tout $m \geq 0$ et pour tout $k \geq 0$, l'ensemble $\{\mathbf{x}_{kb^m+1}, \ldots, \mathbf{x}_{(k+1)b^m}\}$ est un (t, m, s)-réseau, avec $t \leq \tau$. Pour plus de détails sur les propriétés et la construction des (t, m, s)-réseaux, voir [103, 64] et les références qui s'y trouvent.

La propriété qui définit les (t, m, s)-réseaux en base 2 est différente de celle qui est vérifiée pour mesurer la résolution d'une règle de réseau polynômiale. En effet, pour vérifier la propriété des (t, m, s)-réseaux, on ne regarde pas seulement les cubes, puisque l'on n'impose pas que $k_1 = \cdots = k_s$ dans la définition de b-boîte. Notons également que la définition de (t, m, s)-réseau est assez générale pour inclure n'importe quelle règle de réseau polynômiale, c.-à-d., on peut toujours trouver une valeur de t qui fait en sorte que la règle est un (t, m, s)-réseau en base 2.

D'ailleurs, la construction utilisée pour définir les (t, m, s)-*réseaux digitaux* [96, 98, 64] inclut la définition de règle de réseau polynômiale. Aussi, la grande majorité des constructions pertinentes aux applications de la méthode QMC entrent dans cette catégorie de (t, m, s)-réseaux [114]. Larcher et ses collaborateurs donnent dans [67, 65] des expressions pour l'erreur obtenue en utilisant un (t, m, s)-réseau digital sur un certain type de fonction et obtiennent des bornes sur cette erreur qui sont similaires à ce qui existe pour les règles de réseau standard, mais en utilisant le développement

en série de Walsh d'une fonction plutôt que celui en série de Fourier. Pour cette raison, ils appellent ces (t, m, s)-réseaux digitaux des "règles de réseau digitales". Nous utilisons aussi le développement en série de Walsh d'une fonction, mais pour donner des résultats sur la *variance* des estimateurs basés sur une règle de réseau polynômiale XOR-translatée.

Dans [104], Owen a proposé une façon de randomiser les (t, m, s)-réseaux, obtenant ainsi ce qu'il appelle des (t, m, s)-*réseaux brouillés*. Cette randomisation va comme suit pour un réseau en base b : dans chaque dimension, on divise l'intervalle $[0, 1)$ en b parties égales que l'on permute aléatoirement et uniformément ; chacun de ces b sous-intervalles est divisé en b parties égales qui sont à leur tour permutées aléatoirement et uniformément et ainsi de suite. En pratique, on arrête lorsque les sous-intervalles deviennent de longueur inférieure à la précision de l'ordinateur, mais les résultats sur la variance donnés dans [105, 106, 109] sont obtenus en supposant que ces subdivisions continuent indéfiniment. La variance des (t, m, s)-réseaux brouillés a été étudiée par Owen dans [105, 106, 109]. En comparaison avec la randomisation que nous utilisons, la méthode d'Owen requiert plus de temps de calcul, car on doit générer les différentes permutations servant à brouiller le réseau. Par contre, sa méthode détruit plus de corrélation que la XOR-translation, ce qui lui permet d'obtenir des résultats théoriques plus forts sur la réduction de variance, comme nous le verrons à la sous-section 4.5.3.

4.2 Décomposition en série de Walsh

Par léger abus de notation, nous supposons dans ce qui suit que $\mathbb{N} = \{0, 1, 2, \ldots\}$. La *décomposition de f en série de Walsh* en base 2 [8, 36] est donnée par

$$f(\mathbf{x}) = \sum_{\mathbf{h} \in \mathbf{N}^s} \tilde{f}(\mathbf{h})(-1)^{\mathbf{h} \odot \mathbf{x}},$$

où

$$\tilde{f}(\mathbf{h}) = \int_{[0,1)^s} f(\mathbf{x})(-1)^{\mathbf{h} \odot \mathbf{x}} d\mathbf{x}$$

et pour $\mathbf{h} = (h_1, \ldots, h_s) \in \mathbf{N}^s$ et $\mathbf{x} = (x_1, \ldots, x_s) \in [0, 1)^s$,

$$\mathbf{h} \odot \mathbf{x} = \sum_{j=1}^{s} \sum_{k=1}^{\infty} h_{j,k-1} x_{j,k} \bmod 2,$$

où les $h_{j,k}$ et les $x_{j,k}$ sont les coefficients de l'expansion binaire de h_j et x_j, respectivement, c.-à-d.,

$$\begin{aligned} h_j &= \sum_{k=0}^{\infty} h_{j,k} 2^k, \\ x_j &= \sum_{k=1}^{\infty} x_{j,k} 2^{-k}. \end{aligned}$$

Les fonctions de base $(-1)^{\mathbf{h} \odot \mathbf{x}}$ sont donc constantes par morceaux sur $[0,1)^s$ et valent 1 ou -1 ; les "morceaux" sont déterminés par le vecteur $\mathbf{h} \in \mathbb{N}^s$.

Tout au long de ce chapitre, la norme de h_j est définie par

$$|h_j|_{\mathrm{p}} = \begin{cases} 2^{k_j} & \text{si } 2^{k_j} \leq h_j < 2^{k_j+1}, \\ 0 & \text{si } h_j = 0. \end{cases} \tag{4.8}$$

Nous donnons maintenant un résultat analogue à celui donné dans [116, *Lemma 2.7*] pour les règles de réseau standard, qui est utilisé afin de démontrer l'expression pour l'erreur d'intégration et la variance de ces règles. Le lemme suivant aura la même utilité pour les règles de réseau polynômiales, ce qui lui confère donc de l'importance. Il est donné sans démonstration dans [79] et un résultat équivalent pour les (t, m, s)-réseaux digitaux est utilisé dans la démonstration du théorème 3 dans [67].

Lemme 4.2.1 *([79, Proposition 1]) Soit P_N une règle de réseau polynômiale où $N = 2^m$ et soit $\mathbf{h} \in \mathbb{N}^s$. Si $L = \infty$, alors*

$$\sum_{i=1}^{N} (-1)^{\mathbf{h} \odot \mathbf{x}_i} = \begin{cases} N & \textit{si } \mathbf{h} \in \mathcal{L}_s^*, \\ 0 & \textit{sinon.} \end{cases}$$

Démonstration : voir l'annexe B, page xxxi.

À ce point-ci, le lecteur intéressé à suivre de près toutes les démonstrations de ce chapitre devrait lire l'annexe E, dans laquelle nous expliquons comment représenter les fonctions de base de la série de Walsh à l'aide d'une matrice d'Hadamard. Deux lemmes qui sont utilisés dans des démonstrations ultérieures y sont énoncés.

4.3 Liens avec les (t, m, s)-réseaux

Cette section contient les résultats reliant le paramètre de qualité t des (t, m, s)-réseaux et la résolution, lorsque l'on considère les projections $P_N(I)$ d'un ensemble

de points P_N. Dans ce qui suit, nous supposons que le paramètre L dans (4.2) est supérieur ou égal au degré m de la récurrence (4.1), ce qui semble raisonnable puisque L est habituellement choisi égal à 32 ou 64 (le nombre de bits utilisés pour représenter les nombres sur un ordinateur) et $N = 2^m \leq 2^{32}$ est amplement suffisant comme nombre de points pour les méthodes QMC. Cela nous permet de poser $\ell_p^* = \lfloor m/p \rfloor$.

Les définitions qui suivent s'appliquent à n'importe quel ensemble de points et requièrent la notion-clé suivante :

Définition 4.3.1 *Soit P_N un ensemble contenant 2^m points. Le vecteur $\mathbf{k}_I = (k_j)_{j \in I}$, $k_j \geq 0$,* induit une équidistribution sur $P_N(I)$ *si, pour tout vecteur $(v_j)_{j \in I}$ tel que $0 < v_j \leq 2^{k_j}$, l'ensemble*

$$\prod_{j \in I} \left[\frac{v_j - 1}{2^{k_j}}, \frac{v_j}{2^{k_j}} \right)$$

contient $2^{m-\kappa}$ points de P_N.

Nous donnons maintenant une définition précise de t_I, le paramètre mesurant l'équidistribution de $P_N(I)$, en se basant ce qui est défini dans [109, 64, 47, 49]. Nous proposons également une définition alternative, dénotée par q_I, servant aussi à quantifier la qualité de $P_N(I)$.

Définition 4.3.2 *Soit un ensemble P_N contenant $N = 2^m$ points. Pour $I \subseteq S$ non vide, le paramètre de qualité t_I est défini par*

$$t_I = \text{argmin}_{t \geq 0} \Big\{ \sum_{j \in I} k_j \leq m - t \text{ implique que } \mathbf{k}_I \text{ induit une équidistribution sur } P_N(I) \Big\}.$$

Autrement dit, t_I est la plus petite valeur entière telle que $P_N(I)$ est un $(t_I, m, |I|)$-réseau en base 2. Plus t_I est petit, meilleure est l'équidistribution de $P_N(I)$ et on a toujours que $t_I \leq m$. Aussi, plus $|I|$ est petit, plus on s'attend à ce que t_I soit petit, puisque la somme $\sum_{j \in I} k_j$ contient moins de termes.

La définition suivante peut être utile si on veut un paramètre de qualité pour $P_N(I)$ qui soit "spécialisé" à cette projection. En effet, dans la définition de t_I, on mesure non seulement l'équidistribution de $P_N(I)$, mais aussi celle de toutes les projections $P_N(J)$ telles que $J \subset I$, puisque les k_j peuvent valoir 0 lorsque l'on vérifie si $\sum_{j \in I} k_j \leq m - t$.

Dans la définition qui suit, on restreint les k_j à être supérieurs à 0, ce qui signifie que l'on vérifie une condition plus faible que dans la définition de t_I. Par contre, on obtient de cette façon un critère de qualité, dénoté par q_I, qui présente davantage de similarité avec la résolution, comme nous le verrons à la proposition 4.3.2.

Définition 4.3.3 *Soit un ensemble P_N contenant $N = 2^m$ points. Pour $I \subseteq S$ non vide, le paramètre de qualité q_I est défini par*

$$q_I = \text{argmin}_{0 \le q \le m-|I|} \Big\{ \sum_{j \in I} k_j \le m - q \text{ et } k_j \ge 1 \text{ pour tout } j \in I \text{ implique que } \mathbf{k}_I \text{ induit une équidistribution sur } P_N(I) \Big\}.$$

Si aucun q ne satisfait cette condition, on pose $q_I = m - |I| + 1$.

La proposition suivante fait le lien entre q_I, t_I et le paramètre t définissant les (t, m, s)-réseaux.

Proposition 4.3.1 *On a les relations suivantes :*

$$\begin{aligned} t_J &\le t_I \text{ si } J \subseteq I, \\ t_I &\ge q_I, \\ t_I &= \max_{\emptyset \ne J \subseteq I} t_J = \max_{\emptyset \ne J \subseteq I} q_J \\ \text{et} \quad t &= t_{\{1,\ldots,s\}} = \max_{\emptyset \ne I \subseteq S} t_I = \max_{\emptyset \ne I \subseteq S} q_I. \end{aligned}$$

Démonstration : d'abord, le fait que $t_J \le t_I$ si $J \subseteq I$ découle de la définition de t_I. En effet, si $\mathbf{k}_J$ est tel que $\sum_{j \in J} k_j \le m - t_I$, il induit une équidistribution puisqu'on sait, par définition de t_I, que $\tilde{\mathbf{k}}_I$ induit une équidistribution si on pose $\tilde{k}_j = k_j$ si $j \in J$ et 0 sinon. Cela nous donne aussi que $t_I = \max_{\emptyset \ne J \subseteq I} t_J$ et que $t = \max_{\emptyset \ne I \subseteq S} t_I$, puisque $t = t_{\{1,\ldots,s\}}$ par définition.

Ensuite, si $q_I = m - |I| + 1$, cela signifie que le vecteur $\mathbf{k}_I$ tel que $k_j = 1$ pour tout $j \in I$ n'induit pas une équidistribution et donc, on doit avoir $t_I \ge m - |I| + 1 = q_I$. Si $q_I \le m - |I|$ et que $t_I < q_I$, cela signifie que tout vecteur $\tilde{\mathbf{k}}_I$ tel que $\sum_{j \in I} \tilde{k}_j \le m - t_I$ induit une équidistribution. On peut choisir $\tilde{k}_j \ge 1$ pour tout $j \in I$, puisque $m - t_I > m - q_I \ge |I|$, contredisant ainsi le fait que q_I est la valeur minimum que l'on peut prendre dans la définition 4.3.3. Donc, on doit avoir $t_I \ge q_I$.

Pour démontrer que $t_I = \max_{\emptyset \neq J \subseteq I} q_J$, remarquons d'abord que $t_I \geq \max_{\emptyset \neq J \subseteq I} q_J$, puisque $t_I \geq t_J \geq q_J$ pour tout $J \subseteq I$.

On doit donc montrer que $t_I \leq \max_{\emptyset \neq J \subseteq I} q_J$. Il faut donc vérifier que tout $\mathbf{k}_I$ tel que $\sum_{j \in I} k_j \leq m - \max_{\emptyset \neq J \subseteq I} q_J = \min_{\emptyset \neq J \subseteq I}(m - q_J)$ induit une équidistribution. Posons $G = \{j : k_j \geq 1\}$ et puisque, si G est vide, l'équidistribution est trivialement obtenue, supposons que $G \neq \emptyset$. Si $q_G \leq m - |G|$, alors on sait que $\sum_{j \in G} k_j \leq \min_{\emptyset \neq J \subseteq I}(m - q_J) \leq m - q_G$ est suffisant pour induire une équidistribution. Si $q_G = m - |G| + 1$, alors $\min_{\emptyset \neq J \subseteq I}(m - q_J) \leq |G| - 1$. Puisque $\sum_{j \in I} k_j = \sum_{j \in G} k_j \geq |G|$, cela signifie que $\mathbf{k}_I$ ne respecte pas $\sum_{j \in I} k_j \leq m - \max_{\emptyset \neq J \subseteq I} q_J$ et donc, on n'a pas à vérifier l'équidistribution dans ce cas. ∎

Passons maintenant à la définition de la résolution de $P_N(I)$, qui généralise celle donnée dans l'introduction, où l'on se restreignait au cas où $I = \{1, \ldots, j\}$, $1 \leq j \leq s$.

Définition 4.3.4 *Soit un ensemble P_N contenant $N = 2^m$ points. Pour $I \subseteq S$ non vide, la résolution de $P_N(I)$ est définie par*

$$\ell_I = \operatorname{argmax}_{\ell \geq 0} \Big\{ \mathbf{k}_I \textit{ avec } k_j = \ell \textit{ pour tout } j \in I \textit{ induit une équidistribution sur } P_N(I) \Big\},$$

et donc, $\ell^*_{|I|} = \lfloor m/|I| \rfloor$ est une borne supérieure sur ℓ_I.

La résolution contient donc une information différente de celle que nous donne t_I ou q_I. On dit dans [98] que la condition que l'on vérifie pour calculer la résolution est plus faible que celle qui est vérifiée lorsque l'on calcule t. Nous voulons toutefois préciser qu'il n'est pas vrai que la connaissance de t nous permet de savoir quelle est la résolution. En fait, les deux définitions seraient équivalentes si on restreignait les k_j à être égaux entre eux ($k_j = k$ pour tout $j \in I$) dans la définition de t_I. Pour illustrer en quoi cela implique plus de conditions à vérifier, rappelons qu'afin de vérifier si $\ell_I = \ell$ pour une règle de réseau polynômiale, il suffit de calculer le rang d'une matrice $|I|\ell \times |I|\ell$ alors qu'en principe, pour déterminer si $t_I = \tau$, il faudrait calculer le rang de n matrices $(m - \tau) \times (m - \tau)$, où n est égal au nombre de vecteurs $\mathbf{k}_I$ qui satisfont $\left(\sum_{j \in I} k_j = m - \tau\right)$ et donc,

$$n = \binom{m - \tau + |I| - 1}{|I| - 1}.$$

Pour vérifier si $q_I = \tau$, on doit considérer les mêmes matrices que pour t_I, sauf qu'on se restreint aux vecteurs $\mathbf{k}_I$ tels que $k_j \geq 1$ pour tout $j \in I$. Le temps de calcul est ainsi réduit, puisque le nombre de matrices est maintenant $\binom{m-\tau-1}{|I|-1}$. Nous voulons mentionner cependant que lorsque les gens construisent des (t, m, s)-réseaux digitaux, ils utilisent des techniques plus efficaces que celles que l'on vient de mentionner pour vérifier la valeur de t (voir [113], par exemple). De même, rappelons qu'il existe des méthodes plus rapides [17] que celle décrite ci-dessus pour calculer la résolution.

Une des propriétés que partagent ℓ_I et q_I est que dans le cas où le vecteur $\mathbf{k}_I$ tel que $k_j = 1$ pour tout $j \in I$ n'induit pas une équidistribution, on ne cherche pas à quantifier l'équidistribution de $P_N(I)$ en allant regarder ce qui se passe sur les sous-ensembles $J \subset I$ comme le fait t_I, mais on fixe plutôt le paramètre à 0 et $m - |I| + 1$, respectivement. Une autre façon de voir le lien entre q_I et ℓ_I est en notant que 2^{ℓ_I} est la longueur du plus court vecteur dans $\mathcal{L}_I^*$, défini par

$$\mathcal{L}_I^* = \{\mathbf{h} \in \mathbb{N}^s : \sum_{t \in I} h_t(z) z^{(t-1)v} \bmod (P(z), 2) = 0\}, \tag{4.9}$$

lorsque l'on utilise la norme *sup* définie par $\|\mathbf{h}\| = \max_{j \in I} |h_j|_\mathrm{p}$; le cas où $I = \{1, \ldots, p\}$ a été traité en (4.7) à la sous-section 4.1.4. Le résultat suivant démontre que $2^{m-q_I-|I|+1}$ est aussi la longueur du plus court vecteur dans $\mathcal{L}_I^*$, mais en utilisant la norme produit plutôt que la norme *sup*. Ce résultat nous sera utile dans la section 4.5, lorsque nous étudierons la variance de l'estimateur XOR-translaté, car dans ce contexte, il semble que la norme produit soit plus "naturelle" que la norme *sup* pour mesurer les vecteurs $\mathbf{h}$.

Proposition 4.3.2 *Soit P_N une règle de réseau polynômiale contenant $N = 2^m$ points et définie par $P(z)$ et v. Pour $I \subseteq S$ non vide, si $L = \infty$, alors on a que*

$$2^{m-q_I-|I|+1} = \min_{0 \neq \mathbf{h} \in \mathcal{L}_I^*} \|\mathbf{h}\|_\pi,$$

où $\|\mathbf{h}\|_\pi = \prod_{j \in I} \max(1, |h_j|_\mathrm{p}) = 2^{\sum_{j \in I_\mathbf{h}} k_j}$ et $\mathcal{L}_I^*$ est défini en (4.9).

Pour démontrer ce résultat, nous avons besoin du lemme suivant, qui va nous aider à appliquer le lemme E.2 de l'annexe E, pour un cas particulier de vecteurs $\mathbf{k}_I$:

Lemme 4.3.1 *Soit* $\mathbf{k}_I = (k_j)_{j\in I}$ *un vecteur n'induisant pas une équidistribution sur* P_N, *mais pour lequel* $\tilde{\mathbf{k}}_u = (\tilde{k}_j)_{j\in I}$ *défini par*

$$\tilde{k}_j = \begin{cases} k_j - 1 & si\ j = u \\ k_j & sinon, \end{cases}$$

induit une équidistribution, pour tout $u \in I$ *tel que* $k_u \geq 1$. *Posons* $J = \{j : k_j \geq 1\}$ *et* $\mathbf{g}_J = (k_j - 1)_{j\in J}$. *Alors pour toute chaîne* b *contenant* $\gamma = \sum_{j:k_j>0}(k_j - 1)$ *bits, on a que*

$$\left|\sum_{i=1}^{N} \mathbf{1}_{\{\mathbf{x}_i(\mathbf{g}_J)=b\}}(-1)^{\sum_{j\in I} x_{i,j,k_j}}\right| = \frac{N}{2^\gamma},$$

où $\mathbf{x}_i(\mathbf{g}_J)$ *représente la troncation de* $\mathbf{x}_i$ *à ses* $k_j - 1$ *premiers bits dans chaque dimension* $j \in J$.

Démonstration : voir l'annexe B, page xxxii.

Démonstration de la proposition 4.3.2 : pour démontrer l'appartenance d'un vecteur $\mathbf{h}$ à $\mathcal{L}_I^*$, nous allons utiliser le lemme 4.2.1. Ce lemme nous indique que si le vecteur $(k_j + 1)_{j\in I_\mathbf{h}}$ induit une équidistribution, où k_j est tel que $|h_j|_\mathrm{p} = 2^{k_j}$, alors $\mathbf{h}$ n'est pas dans $\mathcal{L}_I^*$, puisque l'équidistribution nous assure que $\sum_{i=1}^{2^m}(-1)^{\mathbf{h}\odot\mathbf{x}_i} = 0$. Ainsi, pour montrer que

$$\min_{0\neq\mathbf{h}\in\mathcal{L}_I^*} \|\mathbf{h}\|_\pi \geq 2^{m-q_I-|I|+1},$$

il suffit de montrer que si $\mathbf{h}$ est tel que $\|\mathbf{h}\|_\pi < 2^{m-q_I-|I|+1}$, alors $(k_j + 1)_{j\in I_\mathbf{h}}$ induit une équidistribution.

Si $q_I = m - |I| + 1$, l'inégalité est respectée trivialement puisque $2^{m-q_I-|I|+1} = 1$ dans ce cas et donc, aucun vecteur différent de 0 ne peut être tel que $\|\mathbf{h}\|_\pi < 1$.

Supposons donc que $q_I \leq m - |I|$. Soit $\mathbf{h} \neq \mathbf{0}$ tel que $\|\mathbf{h}\|_\pi < 2^{m-q_I-|I|+1}$. Alors on a que $\sum_{j\in I_\mathbf{h}} k_j < m - q_I - |I| + 1$ et on veut vérifier que $(k_j + 1)_{j\in I_\mathbf{h}}$ induit une équidistribution. Or, $\sum_{j\in I_\mathbf{h}}(k_j + 1) + (|I| - |I_\mathbf{h}|) \leq m - q_I$ est suffisant pour que $(k_j+1)_{j\in I_\mathbf{h}}$ induise une équidistribution et cette condition est équivalente à demander que

$$\sum_{j\in I_\mathbf{h}} k_j \leq m - q_I - |I|,$$

qui est respectée puisque l'on n'a considéré que les $\mathbf{h}$ tels que $\sum_{j\in I_\mathbf{h}} k_j < m-q_I-|I|+1$.

On doit maintenant montrer qu'il existe un $\mathbf{h}$ dans $\mathcal{L}_I^*$ tel que $\|\mathbf{h}\|_\pi = 2^{m-q_I-|I|+1}$. Par définition de q_I, on sait qu'il existe un vecteur $\mathbf{g} = (g_j)_{j\in I}$ et un sous-ensemble $J \subseteq I$ non vide tels que $\sum_{j\in I} g_j = m - q_I + 1$ avec $g_j \geq 1$ pour tout $j \in J$ qui n'induit pas une équidistribution. Si $q_I \leq m - |I|$, alors $J = I$ et pour tout $u \in I$, $\tilde{\mathbf{g}}_u$ tel que défini au lemme 4.3.1 induit une équidistribution. Si $q_I = m - |I| + 1$, alors il existe un sous-ensemble $J \subseteq I$ non vide tel que $\mathbf{g}$ avec $g_j = 1$ si et seulement si $j \in J$ n'induit pas une équidistribution, mais $\tilde{\mathbf{g}}_u$ tel que défini au lemme 4.3.1 en induit une pour tout $u \in J$. Notons que peu importe la valeur de q_I, on a que $\sum_{j\in J}(g_j - 1) = m - q_I - |I| + 1$ et donc, si

$$|h_j|_{\mathrm{p}} = \begin{cases} 2^{g_j-1} & \text{si } j \in J \\ 0 & \text{sinon,} \end{cases}$$

alors $\|\mathbf{h}\|_\pi = 2^{m-q_I-|I|+1}$.

Il suffit donc de trouver un vecteur $\mathbf{h}$ tel que $|h_j|_{\mathrm{p}} = 2^{g_j-1}$ pour tout $j \in J$, $h_j = 0$ sinon, qui soit dans $\mathcal{L}_I^*$. Posons $\gamma = \sum_{j\in J}(g_j - 1)$ et $\hat{\mathbf{g}}_J = (g_j - 1)_{j\in J}$. Nous allons utiliser le lemme E.2 de l'annexe E afin de montrer que $|H(\hat{\mathbf{g}}_J) \cap \mathcal{L}_s^*| = 1$. On doit donc montrer que

$$2^\gamma \sum_{n=1}^{2^\gamma} y_n^2 = N^2,$$

où

$$|y_n| = \left| \sum_{i=1}^{N} \mathbf{1}_{\mathbf{x}_i(\hat{\mathbf{g}}_J)=b_n} (-1)^{\sum_{j\in J} x_{i,j} g_j} \right| = \frac{N}{2^\gamma}.$$

par le lemme 4.3.1. Donc, on a bien que

$$2^\gamma \sum_{n=1}^{2^\gamma} y_n^2 = 2^\gamma 2^\gamma N^2 / 2^{2\gamma} = N^2.$$

∎

Remarque 4.3.1 *Dans [48], on explique quelque chose de similaire pour les règles de réseau standard : la longueur du plus court vecteur dans le réseau dual obtenue en utilisant la norme $\mathcal{L}^1$ nous donne une borne sur la somme $\sum |h_j|$ telle que toute fonction $e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}}$ est intégrée parfaitement par la règle de réseau.*

Remarque 4.3.2 *Dans la proposition précédente, si on avait remplacé q_I par t_I, la première partie de la démonstration se serait appliquée, c.-à-d., on peut montrer que*

$\min_{0\neq \mathbf{h}\in\mathcal{L}_I^*} \|\mathbf{h}\|_\pi \geq 2^{m-t_I-|I|+1}$. *Le problème survient dans la deuxième partie de la démonstration, car la définition de* t_I *ne nous permet pas de supposer que* $J = I$ *même si* $t_I \leq m-|I|$ *et donc, on n'a pas nécessairement que* $\sum_{j\in J}(g_j-1) = m-t_I-|I|+1$. *Par contre, on démontre dans [127, Corollary 4.3] que pour les générateurs de type GFSR* (generalized feedback shift-register) *et en utilisant une norme légèrement différente de la norme produit,* $(m-t-s+1)$ *est la longueur du plus court vecteur dans* $\mathcal{L}_s^*$. *Cette norme est définie par* $\sum_{j=1}^{s} \deg(h_j)$, *où* $\deg(h_j) = k_j$ *si* $|h_j|_\mathrm{p} = 2^{k_j} \geq 1$ *et* $\deg(h_j) = -1$ *si* $h_j = 0$. *Notons que lorsque les paramètres des générateurs GFSR sont choisis de façon appropriée, ils sont équivalents aux générateurs de Tausworthe [71].*

La proposition suivante nous donne des bornes en fonction de ℓ_I sur la valeur de t_I et de q_I. Ceci nous donne donc de l'information sur la valeur que t_I et q_I peuvent prendre sans avoir à les calculer explicitement. Nous croyons que dans un futur très rapproché, les gens travaillant à la construction de (t, m, s)-réseaux vont essayer d'en construire pour lesquels au moins un sous-ensemble de paramètres t_I (ou q_I) sont minimisés [114]. Si on veut comparer les ensembles de points ainsi obtenus avec des règles de réseau polynômiales pour lesquelles seules les résolutions ℓ_I sont connues, les bornes suivantes seront utiles, du moins comme première comparaison.

Proposition 4.3.3 *Soit un ensemble* P_N *contenant* $N = 2^m$ *points et soit* $I \subseteq S$ *non vide. Alors*

$$\max(0, m - |I|(\ell_I + 1) + 1) \leq q_I \leq m - |I| + 1 - \ell_I.$$

et

$$\max(0, m - |I|(\ell_I + 1) + 1) \leq t_I \leq m - \ell_I$$

Démonstration : pour les bornes sur q_I, on utilise la proposition 4.3.2. Par définition de ℓ_I, on sait que

(1) il existe un vecteur $\mathbf{h} \in \mathcal{L}_I^*$ avec $|h_j|_\mathrm{p} \leq 2^{\ell_I}$ pour tout $j \in I$,

(2) si $\mathbf{h}$ est tel que $|h_j|_\mathrm{p} < 2^{\ell_I}$ pour tout $j \in I$, alors $\mathbf{h} \notin \mathcal{L}_I^*$.

Puisque $\max_{\mathbf{h}:|h_j|_\mathrm{p}\leq 2^{\ell_I}} \|\mathbf{h}\|_\pi = 2^{|I|\ell_I}$, par (1), on a que

$$2^{m-q_I-|I|+1} \leq 2^{|I|\ell_I}$$

et donc, on doit avoir $q_I \geq m - |I|(\ell_I + 1) + 1$. Ensuite, puisque si $\|\mathbf{h}\|_\pi \leq 2^{\ell_I - 1}$, alors $|h_j|_p < 2^{\ell_I}$ pour tout $j \in I$, par (2), on a que

$$2^{m - q_I - |I| + 1} > 2^{\ell_I - 1}$$

et donc, $q_I \leq m - |I| - \ell_I + 1$.

Pour les bornes sur t_I, nous allons d'abord montrer que la borne supérieure tient. Pour cela, il faut s'assurer que si $\sum_{j \in I} k_j \leq m - (m - \ell_I) = \ell_I$, alors $\mathbf{k}_I$ induit une équidistribution. Or, $\sum_{j \in I} k_j \leq \ell_I$ implique que chaque $k_j \leq \ell_I$ et donc, on a équidistribution par définition de ℓ_I.

Puisque $t_I \geq 0$ par définition, pour démontrer la validité de la borne inférieure, il suffit de montrer que $t_I > m - |I|(\ell_I + 1)$: supposons que l'on peut trouver $t^* \leq m - |I|(\ell_I + 1)$ satisfaisant la condition énoncée à la définition 4.3.2. Cela signifie que si $\mathbf{k}_I$ est tel que $\sum_{j \in I} k_j \leq m - t^*$, alors $\mathbf{k}_I$ induit une équidistribution. On peut donc prendre $k_j = \ell_I + 1$ pour tout $j \in I$, puisqu'on obtient ainsi $\sum_{j \in I} k_j = |I|(\ell_I + 1) \leq m - t^*$, mais ceci contredit le fait que ℓ_I est la résolution de $P_N(I)$. ∎

De façon concrète, qu'est-ce que ces bornes nous indiquent sur la différence entre une règle de réseau polynômiale de résolution ℓ_I et un (t, m, s)-réseau en base 2 ayant une certaine valeur de $q_I = q$? Si $q > m - |I| + 1 - \ell_I$, alors la propriété d'équidistribution de la règle telle que mesurée à l'aide de q_I est meilleure que celle du (t, m, s)-réseau pour la projection $P_N(I)$. De la même façon, si $q \leq m - |I|(\ell_I + 1)$, alors la propriété d'équidistribution du (t, m, s)-réseau telle que mesurée à l'aide de q_I est meilleure que celle de la règle pour la projection $P_N(I)$. Nous trouvons qu'il est intéressant de voir qu'en ne connaissant que la résolution ℓ_I pour une règle donnée, on puisse dire que sa projection $P_N(I)$ est mieux équidistribuée (si q_I est notre façon de mesurer l'équidistribution) que celle d'un (t, m, s)-réseau pour lequel le paramètre q_I est suffisamment grand.

Remarquons que pour les projections unidimensionnelles, la proposition 4.3.3 nous indique que $\ell_I = m - q_I = m - t_I$, puisque la borne supérieure est égale à la borne inférieure dans ce cas.

Le corollaire suivant peut sembler évident, car il nous montre que dans le cas où la projection $P_N(I)$ a la meilleure équidistribution possible par rapport à q_I (et t_I), c.-

à-d., lorsque $t_I = q_I = 0$, on a que la résolution est maximale, mais l'inverse n'est pas vrai. Nous le donnons quand même, simplement pour comparer ces deux cas limites.

Corollaire 4.3.1 *Soit un ensemble P_N contenant $N = 2^m$ points et $I \subseteq S$ non vide.*

(1) Si $q_I = 0$, alors $\ell_I = \ell^*_{|I|} = \lfloor m/|I| \rfloor$.

(2) Si $\ell_I = \ell^*_{|I|} = \lfloor m/|I| \rfloor$, alors $0 \leq t_I \leq m - \lfloor m/|I| \rfloor$ et $0 \leq q_I \leq m - |I| + 1 - \lfloor m/|I| \rfloor$.

Démonstration : si $q_I = 0$, la borne inférieure donnée à la proposition 4.3.3 nous indique que l'on doit avoir $m - |I|(\ell_I + 1) + 1 \leq 0$ et ceci tient si et seulement si

$$\ell_I \geq \frac{m+1}{|I|} - 1. \tag{4.10}$$

Supposons que $\ell_I < \lfloor m/|I| \rfloor$: alors on doit avoir $\ell_I \leq m/|I| - 1$, puisque ℓ_I doit être entier, mais cela contredit (4.10). On doit donc avoir $\ell_I = \lfloor m/|I| \rfloor$.

Si $\ell_I = \ell^*_I$, alors

$$\begin{aligned} m - |I|(\ell^*_I + 1) + 1 &= m - |I|(\lfloor m/|I| \rfloor + 1) + 1 \\ &= m - |I|[m/|I| - (m/|I| \bmod 1) + 1 + 1 \\ &= |I|(m/|I| \bmod 1 - 1) + 1 \\ &\leq |I| - 1 - |I| + 1 = 0 \end{aligned}$$

et donc, par la proposition 4.3.3, on a que la valeur de t_I et de q_I pour la règle de réseau polynômiale est supérieure ou égale à 0. La borne supérieure provient directement de la proposition 4.3.3, en remplaçant ℓ_I par $\ell^*_{|I|}$. ∎

4.4 Nouveau critère de sélection

Nous présentons maintenant un nouveau critère de sélection afin de choisir les règles de réseau polynômiales. Ce critère est défini de façon similaire à celui que nous avons présenté à la section 3.5 pour les règles de réseau standard, qui était dénoté par $M_{t_1,\ldots,t_d}$. Dans ce qui suit, on utilise à plusieurs reprises le fait que, par définition, les règles de réseau polynômiales que nous utilisons sont stationnaires dans la dimension. Ceci vient du fait que la récurrence (4.1) est toujours inversible puisqu'elle est définie

sur le corps fini $\mathbf{F}_2$ et que l'on a supposé que $a_m \neq 0$ dans (4.1). Ainsi, on peut appliquer la proposition 3.2.1.

Définition 4.4.1 *Pour choisir des règles de réseau polynômiales, on définit le critère*

$$\Delta_{w_1,\ldots,w_d} = \max\left(\max_{1\leq u\leq w_1} \ell^*_u - \ell_u, \max_{2\leq u\leq d} \max_{I\in S(w_u,u)} \ell^*_{|I|} - \ell_I\right).$$

Ce critère calcule donc la résolution de $P_N(I)$ pour tous les I à indices successifs tels que $|I| \leq w_1$, $i_1 = 1$, et pour tous les I contenant u indices espacés d'au plus w_u et dont le premier est 1, pour $2 \leq u \leq d$. Ensuite, il compare chacune de ces résolutions ℓ_I avec la résolution maximale $\ell^*_{|I|}$, puis prend l'écart maximal ainsi obtenu.

Cette définition généralise celle d'être ME, qui correspond en fait à avoir $\Delta_m = 0$. On dira qu'un ensemble de points est $\mathrm{ME}(w_1,\ldots,w_d)$ si $\Delta_{w_1,\ldots,w_d} = 0$ [79]. Par exemple, on sait qu'un $(0,m,s)$-réseau est $\mathrm{ME}(\underbrace{s,\ldots,s}_{s \text{ fois}})$, puisque $t = \max_{\emptyset\neq I\subseteq S} q_I$ et on a vu au corollaire 4.3.1 que $q_I = 0$ implique que $\ell_I = \ell^*_{|I|}$. En fait, on peut voir le paramètre t comme étant la valeur de $\Delta_{w_1,\ldots,w_s}$ en utilisant la norme produit pour calculer la longueur du plus court vecteur dans $\mathcal{L}^*_I$ et en prenant $w_1 = \ldots = w_s = s$. En effet, si P_N est stationnaire dans la dimension et que l'on dénote par $\Delta^\pi_{w_1,\ldots,w_s}$ la valeur du critère utilisant la norme produit, on obtient alors

$$\Delta^\pi_{\underbrace{s,\ldots,s}_{s \text{ fois}}} = \max_{I:i_1=1}[(m-|I|+1)-(m-q_I-|I|+1))] = \max_{I:i_1=1} q_I = \max_{\emptyset\neq I\subseteq S} q_I = t,$$

en supposant que $t = 0$ peut être atteint, car alors on sait qu'il est possible d'avoir $q_I = 0$ pour tout I. Notons que si on est en base 2, il est impossible de construire un $(0,m,s)$-réseau, à moins que $s \leq b+1 = 3$ [65]. Autrement dit, la borne supérieure $m-|I|+1$ est atteignable, mais pour l'ensemble des (t,m,s)-réseaux définis sur n'importe quelle base. Pour être plus juste, on pourrait définir un critère $\tilde{t}$ en utilisant la borne inférieure [102, *Theorem 7*]

$$t^*_{|I|}(b) = \frac{|I|-1}{b} - \log_b\left(\frac{(b-1)|I|+b}{2}\right),$$

qui assure l'existence d'un $(t_I,m,|I|)$-réseau en base b pour tout $t_I \geq t^*_{|I|}(b), m \geq t_I$. On poserait alors

$$\tilde{t} = \max_{\emptyset\neq I\subseteq S}(t_I - \lceil t^*_{|I|}(2)\rceil)$$

pour choisir des (t, m, s)-réseaux en base 2.

Tout comme pour le critère $M_{t_1,\ldots,t_d}$ utilisé pour choisir les règles de réseau standard, on peut démontrer que $\Delta_{w_1,\ldots,w_d}$ correspond à la mesure de discrépance générale $D'_w(P_N)$ donnée en (2.8), pour un certain choix de $w(\mathbf{h})$ et en remplaçant $L^\perp$ par $\mathcal{L}_s^*$. La structure des poids $w(\mathbf{h})$ est similaire à celle obtenue dans le cas standard et donc, on peut invoquer les mêmes raisons qu'à la section 3.5 pour dire que les paramètres $w_1, \ldots, w_d$ devraient être choisis tels que $w_1 \geq \ldots \geq w_d$, avec w_1 assez grand ; d et les autres w_u devraient être sélectionnés de façon à ce que le critère puisse être calculé assez rapidement en pratique. Nous verrons à la section 4.5.2 d'autres considérations pouvant nous guider dans le choix de ces paramètres.

Proposition 4.4.1 *Supposons que $w_1 \geq w_u$ pour $2 \leq u \leq d$. Si*

$$w(\mathbf{h}) = \begin{cases} \ell^*_{|I_\mathbf{h}|} - \lg \|\mathbf{h}\| & \text{si } \mathbf{h} \in \mathcal{L}_s^* \text{ et } I_\mathbf{h} \in H(w_1, \ldots, w_d, d), \\ \ell^*_{r(I_\mathbf{h})} - \lg \|\mathbf{h}\| & \text{si } \mathbf{h} \in \mathcal{L}_s^*,\ I_\mathbf{h} \notin H(w_1, \ldots, w_d, d) \text{ mais que } r(I_\mathbf{h}) \leq w_1, \\ 0 & \text{sinon,} \end{cases} \tag{4.11}$$

alors la mesure de discrépance $D'_w(P_n)$ définie par (2.8) est égale à $\Delta_{w_1,\ldots,w_d}$.

Démonstration : voir l'annexe B, page xxxiii.

Remarque 4.4.1 *Comme dans le cas standard, la définition des poids $w(\mathbf{h})$ pour obtenir l'équivalence n'est pas unique. Nous avons choisi celle qui donnait au plus grand nombre possible de vecteurs $\mathbf{h}$ un poids non nul.*

Le résultat suivant nous donne une borne supérieure sur la valeur maximale que peut prendre q_I pour un ensemble de points pour lequel le critère $\Delta_{w_1,\ldots,w_d}$ est connu. L'idée est la suivante : si on a une règle de réseau polynômiale pour laquelle on connaît $\Delta_{w_1,\ldots,w_d}$ et qu'on veut la comparer avec un (t, m, s)-réseau pour lequel on ne connaît que t et non pas les q_I individuellement, alors le résultat suivant nous donne une valeur avec laquelle on peut comparer ce t. Autrement dit, on peut voir $t = \max_{\emptyset \neq I \subseteq S} q_I$ comme étant une borne sur $\max_{\emptyset \neq I \in H(w_1,\ldots,w_d,d)} q_I$ et le résultat suivant nous donne aussi une borne sur $\max_{\emptyset \neq I \in H(w_1,\ldots,w_d,d)} q_I$, mais qui est basée sur les résolutions ℓ_I.

Proposition 4.4.2 *Soit P_N un ensemble contenant 2^m points. Si $\Delta_{w_1,\ldots,w_d} = \Delta$, où les w_u sont tels que $w_u \leq w_1$ pour $2 \leq u \leq d$ et que $w_1 \geq \lfloor\sqrt{m}\rfloor + 1$, alors*

$$\max_{I \in H(w_1,\ldots,w_d,d)} q_I \leq m + 1 + \Delta - \lfloor\sqrt{m}\rfloor - \min\left(\left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor}\right\rfloor, 1 + \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor + 1}\right\rfloor\right).$$

Démonstration : voir l'annexe B, page xxxiv.

On peut voir que pour Δ fixé, cette borne croît avec m, alors que si on prend des (t, m, s)-réseaux construits à partir d'une (τ, s)-suite donnée, le paramètre t est borné uniformément par τ [64]. Par contre, cette borne croît avec Δ, qui dépend de m et par expérience, il semble que plus m est grand, plus il est facile de trouver des règles de réseau polynômiales avec Δ petit.

TABLEAU 4.1: Borne donnée à la proposition 4.4.2

		τ de NX2					
m	$m+1+\Delta+\max g(p)$	$s=8$	$s=9$	$s=10$	$s=11$	$s=12$	$s=13$
10	$5+\Delta$	5	6	8	9	10	11
11	$6+\Delta$	⋮	6	8	9	10	11
12	$6+\Delta$		⋮	8	9	10	11
13	$7+\Delta$			8	9	10	11
14	$8+\Delta$			8	9	10	11
15	$9+\Delta$			⋮	9	10	11
16	$9+\Delta$				⋮	10	11
17	$10+\Delta$					10	11
18	$11+\Delta$					⋮	11
19	$12+\Delta$						⋮
20	$12+\Delta$						

Le tableau 4.1 donne la valeur de $(m+1+\Delta+\max_{1\leq p\leq w_1} g(p))$ pour m allant de 10 à 20. Nous donnons également la valeur de τ pour la (τ, s)-suite NX2 de Niederreiter-Xing, telle qu'elle apparaît dans [102, *Table* 1]. Le paramètre τ pour cette construction a le meilleur taux de convergence possible en fonction de s, soit $O(s)$. La valeur de τ est constante en fonction de m et nous l'inscrivons jusqu'à la ligne correspondant à la valeur de m pour laquelle $m + 1 + \max g(p)$ vaut τ. Cela veut dire que pour un couple (m, s) donné, si la règle de réseau est telle que $\Delta = 0$ et qu'une valeur de τ est inscrite à la ligne m dans la colonne s, alors la qualité de ses projections $P_N(I)$ pour

lesquelles $I \in H(w_1, \ldots, w_d, d)$ est "environ aussi bonne" que celle du (t, m, s)-réseau. Par exemple, en dimension 10, si $m = 12$ et que la règle a une valeur de $\Delta = 0$, alors $\max_{I \in H(w_1,\ldots,w_d,d)} q_I \leq 6 < \tau = 8$. On dit "environ aussi bonne" car les deux valeurs que l'on compare sont des bornes supérieures sur la mesure de qualité $\max_{I \in H(w_1,\ldots,w_d,d)} q_I$; en effet, $\tau \geq t \geq \max_{I \in H(w_1,\ldots,w_d,d)} q_I$.

4.5 Variance des estimateurs construits à partir de règles de réseau polynômiales

Nous présentons ici l'analogue de l'expression pour l'erreur d'intégration donnée dans [116, *Theorem 2.8*] et de la proposition 2.2.1, mais dans le cas polynômial. Tout au long de cette section, nous supposons que $L = \infty$, afin de pouvoir appliquer le lemme 4.2.1. Bien sûr, en pratique, la valeur de L est finie, mais puisque la méthode MC subit le même genre d'approximation, c.-à-d., elle n'est pas implantée en utilisant des points $\mathbf{x}_i$ qui suivent exactement la loi uniforme sur $[0,1)^s$, mais plutôt une loi uniforme discrète sur $[0, 1/2^L, \ldots, (2^L - 1)/2^L]^s$, nous croyons que l'hypothèse que $L = \infty$ ne déforme pas trop la réalité, surtout si notre but est de comparer la variance des règles de réseau polynômiales avec la méthode MC.

L'expression pour l'erreur obtenue en utilisant une règle de réseau polynômiale est donnée sans démonstration dans [79, Proposition 3]. Dans [67], une version plus générale est donnée, mais qui n'utilise pas la notion de réseau dual ; on montre que $\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i) - \mu = \frac{1}{N}\sum_{\mathbf{0}\neq\mathbf{h}} \tilde{f}(\mathbf{h}) \left(\sum_{i=1}^{N} (-1)^{\mathbf{h}\odot\mathbf{x}_i}\right)$.

Proposition 4.5.1 *Soit P_N une règle de réseau polynômiale. Si f a une représentation en série de Walsh absolument convergente, alors*

$$\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i) - \mu = \sum_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s^*} \tilde{f}(\mathbf{h}).$$

Démonstration : on a que

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i) - \mu &= \frac{1}{N}\sum_{i=1}^{N}\sum_{\mathbf{h}\in\mathbf{N}^s} \tilde{f}(\mathbf{h})(-1)^{\mathbf{h}\odot\mathbf{x}_i} - \tilde{f}(\mathbf{0}) \\ &= \frac{1}{N}\sum_{\mathbf{h}\in\mathbf{N}^s} \tilde{f}(\mathbf{h}) \sum_{i=1}^{N}(-1)^{\mathbf{h}\odot\mathbf{x}_i} - \tilde{f}(\mathbf{0})\end{aligned}$$

$$= \sum_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s} \tilde{f}(\mathbf{h}),$$

où la première égalité est obtenue en utilisant la représentation en série de Walsh de f et le fait que $\tilde{f}(\mathbf{0}) = \mu$; la convergence de $\sum|\tilde{f}(\mathbf{h})|$ nous permet d'appliquer le théorème de Fubini pour justifier le changement d'ordre de sommation effectué afin d'obtenir la deuxième égalité; la troisième égalité suit par application du lemme 4.2.1.

∎

Dans [67], on donne des bornes sur l'erreur lorsque les coefficients $\tilde{f}(\mathbf{h})$ décroissent suffisamment rapidement (analogue des fonctions qui sont dans $E_\alpha(c)$ quand on utilise les séries de Fourier) et qu'un (t, m, s)-réseau en base b est utilisé, avec b premier. Dans [65], on fait la même chose, mais pour des (t, m, s)-réseaux digitaux plus généraux. Dans ce qui suit, nous analysons plutôt la *variance* de l'estimateur XOR-translaté

$$\hat{\mu}_{\mathrm{PLR}} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i \oplus \boldsymbol{U}),$$

où $\boldsymbol{U}$ est un vecteur aléatoire uniforme sur $[0,1)^s$,

$$\mathbf{x}_i \oplus \boldsymbol{U} = (x_{i,1} \oplus U_1, \ldots, x_{i,s} \oplus U_s)$$

et

$$x_{i,j} \oplus U_j = \sum_{k=1}^{\infty}((x_{i,j,k} + U_{j,k}) \bmod 2)2^{-k}.$$

Le résultat suivant est donné sans démonstration dans [79, Proposition 3].

Proposition 4.5.2 *Pour $f \in \mathcal{L}^2$ et P_N une règle de réseau polynômiale, on a que*

$$\mathrm{E}(\hat{\mu}_{\mathrm{PLR}}) = \mu \tag{4.12}$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) = \sum_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s} |\tilde{f}(\mathbf{h})|^2. \tag{4.13}$$

Démonstration : puisque $\boldsymbol{U}$ suit la loi uniforme sur $[0,1)^s$, cela signifie que les $U_{j,k}$ sont i.i.d. de loi Bernoulli avec $p = 1/2$. Ainsi, les $(x_{i,j,k} + U_{j,k}) \bmod 2$ sont aussi i.i.d. de loi Bernoulli et donc, chaque $\mathbf{x}_i \oplus \boldsymbol{U}$ suit la loi uniforme sur $[0,1)^s$. Ainsi, $\mathrm{E}(f(\mathbf{x}_i \oplus \boldsymbol{U})) = \mu$ pour chaque i, d'où le résultat (4.12).

Pour démontrer (4.13), on procède de façon similaire à ce qui a été fait dans le cas standard, c.-à-d., on définit la fonction $g(\cdot) : [0,1)^s \to \mathbb{R}$ telle que $g(\boldsymbol{U}) = \sum_{i=1}^{N} f(\mathbf{x}_i \oplus \boldsymbol{U})/N$. Ainsi, $\mathrm{Var}(g(\boldsymbol{U})) = \mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$. Puisque l'égalité de Parseval tient également dans le cas de la décomposition en série de Walsh (voir [36], par exemple), on a

$$\mathrm{Var}(g(\boldsymbol{U})) = \sum_{\mathbf{0} \neq \mathbf{h} \in \mathbb{N}^s} |\tilde{g}(\mathbf{h})|^2. \tag{4.14}$$

Il faut maintenant calculer $\tilde{g}(\mathbf{h})$:

$$\begin{aligned}
\tilde{g}(\mathbf{h}) &= \int_{[0,1)^s} g(\mathbf{u})(-1)^{\mathbf{h}\odot\mathbf{u}} d\mathbf{u} \\
&= \int_{[0,1)^s} \left(\frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i \oplus \mathbf{u}) \right) (-1)^{\mathbf{h}\odot\mathbf{u}} d\mathbf{u} \\
&= \frac{1}{N} \sum_{i=1}^{N} \int_{[0,1)^s} f(\mathbf{x}_i \oplus \mathbf{u})(-1)^{\mathbf{h}\odot\mathbf{u}} d\mathbf{u} \\
&= \frac{1}{N} \sum_{i=1}^{N} \int_{[0,1)^s} f(\mathbf{v}_i)(-1)^{\mathbf{h}\odot(\mathbf{v}_i\oplus\mathbf{x}_i)} d\mathbf{v}_i \\
&= \frac{1}{N} \sum_{i=1}^{N} (-1)^{\mathbf{h}\odot\mathbf{x}_i} \int_{[0,1)^s} f(\mathbf{v}_i)(-1)^{\mathbf{h}\odot\mathbf{v}_i} d\mathbf{v}_i \\
&= \frac{1}{N} \sum_{i=1}^{N} (-1)^{\mathbf{h}\odot\mathbf{x}_i} \tilde{f}(\mathbf{h}) \\
&= \begin{cases} \tilde{f}(\mathbf{h}) & \text{si } \mathbf{h} \in \mathcal{L}_s^*, \\ 0 & \text{sinon.} \end{cases}
\end{aligned}$$

Dans la série d'équations précédentes, la troisième égalité est obtenue en interchangeant la somme et l'intégrale et ce changement d'ordre est légal par le théorème de Fubini, puisque f est de carré-intégrable ; la quatrième égalité est obtenue en appliquant le changement de variable $\mathbf{v}_i = \mathbf{x}_i \oplus \mathbf{u}$, qui nous permet également de réécrire $\mathbf{u}$ comme étant $\mathbf{x}_i \oplus \mathbf{v}_i$ et la cinquième vient du fait que $(-1)^{\mathbf{h}\odot(\mathbf{v}_i\oplus\mathbf{x}_i)} = (-1)^{\mathbf{h}\odot\mathbf{v}_i}(-1)^{\mathbf{h}\odot\mathbf{x}_i}$; la dernière égalité suit par application du lemme 4.2.1.

En remplaçant dans (4.14), on obtient bien que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) = \sum_{\mathbf{0} \neq \mathbf{h} \in \mathcal{L}_s^*} |\tilde{f}(\mathbf{h})|^2.$$

∎

Remarquons que dans cette démonstration comme dans celle de la proposition 2.2.1, donnant la variance de l'estimateur formé à partir d'une règle de réseau standard translatée aléatoirement, on utilise le fait que

$$w(\mathbf{h}, r(\mathbf{u}, \mathbf{x}_i))/w(\mathbf{h}, \mathbf{x}_i) = w(\mathbf{h}, \mathbf{u}), \tag{4.15}$$

où

$$w(\mathbf{h}, \mathbf{y}) = \begin{cases} e^{-2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{y}} & \text{dans le cas standard,} \\ (-1)^{\mathbf{h}\odot\mathbf{y}} & \text{dans le cas polynômial,} \end{cases}$$

est la fonction de base utilisée dans la décomposition choisie et

$$r(\mathbf{u}, \mathbf{x}_i) = \begin{cases} (\mathbf{x}_i + \mathbf{u}) \bmod 1 & \text{dans le cas standard,} \\ \mathbf{x}_i \oplus \mathbf{u} & \text{dans le cas polynômial,} \end{cases}$$

est la randomisation choisie. Dans les deux cas, les propriétés de la fonction de base font en sorte que l'on obtient

$$\sum_{i=1}^{N} w(\mathbf{h}, \mathbf{x}_i) = \begin{cases} N & \text{si } \mathbf{h} \in \mathcal{L}_s^*, \\ 0 & \text{sinon.} \end{cases}$$

Aussi, avec une randomisation différente (par exemple, en utilisant $((\mathbf{x}_i + \mathbf{u}) \bmod 1)$ pour les règles de réseau polynômiales), l'égalité (4.15) ne tiendrait plus nécessairement et donc, la formule pour la variance tomberait. Cela justifie d'une certaine manière le choix de la randomisation dans chaque cas.

4.5.1 Liens entre la décomposition ANOVA et celle en série de Walsh

Toujours en analogie avec ce que nous avons fait dans le cas standard, on montre dans cette sous-section comment les coefficients de Walsh de la fonction peuvent servir à décomposer les variances $\sigma_I^2 = \mathrm{Var}(f_I)$ associées à la décomposition ANOVA de f, pour $I \subseteq S$ non vide. Rappelons que cette technique permet de calculer les σ_I^2 sans connaître explicitement les composantes f_I. Ce résultat, donné à la proposition 4.5.3, nous permettra entre autres de réécrire $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$ en fonction des résolutions ℓ_I à la sous-section suivante, justifiant ainsi notre critère $\Delta_{w_1,\ldots,w_d}$.

Étant donné que les démonstrations de cette sous-section sont similaires à celles des résultats de la section 3.3 (cas standard), elles ont été mises en annexe.

Lemme 4.5.1 *Soit $f \in \mathcal{L}^2$. Alors pour tout $I \subseteq S$ non vide, les coefficients de Walsh des composantes ANOVA f_I (voir définition 3.1.1) sont donnés par :*

$$\tilde{f}_I(\mathbf{h}) = \begin{cases} \tilde{f}(\mathbf{h}) & \text{si } I = I_{\mathbf{h}}, \\ 0 & \text{sinon .} \end{cases}$$

Démonstration : voir l'annexe B, page xxxvi.

Le résultat suivant réécrit la variance σ_I^2 de f_I en fonction des coefficients de Walsh de f et regroupe ces coefficients $\tilde{f}(\mathbf{h})$ selon le vecteur d'indices $I_{\mathbf{h}}$ qui leur est associé, dans l'expression pour la variance de $\hat{\mu}_{\text{PLR}}$.

Proposition 4.5.3 *Soit $f \in \mathcal{L}^2$. Alors pour tout $I \subseteq S$ non vide,*

$$\sigma_I^2 = \sum_{\mathbf{h} \in \mathbb{N}_I^*} |\tilde{f}(\mathbf{h})|^2,$$

où $\mathbb{N}_I^* = \{\mathbf{h} \in \mathbb{N}^s : I_{\mathbf{h}} = I\}$. De plus, si on écrit l'estimateur $\hat{\mu}_{\text{PLR}}$ comme étant

$$\hat{\mu}_{\text{PLR}} = \sum_{I \subseteq S} \left(\frac{1}{N} \sum_{i=1}^{N} f_I(\mathbf{x}_i \oplus \mathbf{u}) \right)$$

et que l'on pose

$$\sigma_{I,\text{PLR}}^2 = \text{Var}\left(\frac{1}{N} \sum_{i=1}^{N} f_I(\mathbf{x}_i \oplus \mathbf{u}) \right),$$

alors

$$\left.\begin{array}{rcl} \text{Var}(\hat{\mu}_{\text{PLR}}) & = & \sum_{\emptyset \neq I \subseteq S} \sigma_{I,\text{PLR}}^2, \\ \text{où } \sigma_{I,\text{PLR}}^2 & = & \sum_{\mathbf{h} \in \mathcal{L}_I^* \cap \mathbb{N}_I^*} |\tilde{f}(\mathbf{h})|^2. \end{array}\right\} \tag{4.16}$$

Démonstration : voir l'annexe B, page xxxvii.

4.5.2 Bornes sur la variance données en fonction du critère de sélection

Nous donnons maintenant des bornes sur la variance de $\hat{\mu}_{\text{PLR}}$ en utilisant soit les résolutions ℓ_I, soit les paramètres q_I. Dans les deux cas, la borne est donnée en décomposant d'abord la variance selon les sous-ensembles $I \subseteq S$, puis en sommant sur les $\mathbf{h}$ dans $\mathbb{N}_I^*$ qui ne sont pas plus petits que le plus court vecteur dans $\mathcal{L}_I^*$, dont la longueur peut est écrite en fonction de ℓ_I ou q_I. Le but de cette étude est d'aller chercher dans ces bornes des arguments supportant la forme de notre critère de sélection.

Puisque l'hypothèse que P_N est stationnaire dans la dimension sera utilisée à plusieurs reprises dans cette sous-section, nous voulons rappeler que les règles de réseau polynômiales que nous utilisons (basées sur la récurrence (4.1)) ont toujours cette propriété. De plus, afin de faciliter la présentation dans le cas de la borne en fonction des q_I, on utilise la notation suivante :

$$K_I(p) = \{\mathbf{k}_I : \sum_{j \in I} k_j \geq p\}. \tag{4.17}$$

Proposition 4.5.4 *Soit $f \in \mathcal{L}^2$ et P_N une règle de réseau polynômiale avec $N = 2^m$ points. Alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \sum_{\emptyset \neq I \subseteq S} \sum_{\mathbf{h} \in \mathbb{N}_I^*} |\tilde{f}(\mathbf{h})|^2 \mathbf{1}_{\{\|\mathbf{h}\| \geq 2^{\ell_I}\}} \tag{4.18}$$

et

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \sum_{\emptyset \neq I \subseteq S} \sum_{\mathbf{k}_I \in K_I(\kappa^*)} \sum_{\mathbf{h} \in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2, \tag{4.19}$$

où $\kappa^* = m - q_I - |I| + 1$ et $H(\mathbf{k}_I) = \{\mathbf{h} \in \mathbb{N}_I^* : |h_j|_p = 2^{k_j},$ pour tout $j \in I\}$.

Démonstration : par définition, si $\mathbf{h} \in \mathbb{N}_I^*$ est tel que $\|\mathbf{h}\| < 2^{\ell_I}$, alors $\mathbf{h} \notin \mathcal{L}_s^*$, puisque $\mathbb{N}_I^* \cap \mathcal{L}_s^* \subseteq \mathcal{L}_I^*$ et que $\min_{\mathbf{0} \neq \mathbf{h} \in \mathcal{L}_I^*} \|\mathbf{h}\| = 2^{\ell_I}$. De même dans le cas où $\mathbf{h} \in H(\mathbf{k}_I)$ et que $\kappa = \sum_{j \in I} k_j < m - q_I - |I| + 1$. Les deux bornes suivent en utilisant l'expression (4.16) pour $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$. ∎

La borne (4.18) justifie d'une certaine façon l'utilisation de notre critère de sélection $\Delta_{w_1,\ldots,w_d}$, puisque plus les ℓ_I sont grands, plus cette borne est petite, car elle contient alors moins de termes. Cependant, pour justifier notre critère de façon satisfaisante, on doit également motiver le choix des paramètres d et $w_1, \ldots, w_d$, en relation avec la variance de $\hat{\mu}_{\mathrm{PLR}}$. L'objectif est de trouver une borne sur la variance de $\hat{\mu}_{\mathrm{PLR}}$ qui soit de la forme $g(\Delta_{w_1,\ldots,w_d}) \cdot \sigma^2$, avec $g(\cdot)$ une fonction qui augmente avec $\Delta_{w_1,\ldots,w_d}$.

Or, la motivation derrière un critère comme $\Delta_{w_1,\ldots,w_d}$ est de s'assurer que les résolutions ℓ_I associées aux sous-ensembles I pour lesquels σ_I^2/σ^2 est grand sont suffisamment grandes et ceci devrait nous permettre de construire des estimateurs ayant une petite variance. Le résultat donné à la proposition 4.5.5 précise cette idée, justifiant ainsi la forme du critère.

Plus précisément, on donne une borne supérieure sur la variance de $\hat{\mu}_{\mathrm{PLR}}$ qui est de la forme voulue, c.-à-d., $g(\Delta_{w_1,\ldots,w_d}) \cdot \sigma^2$, mais pour y arriver, on doit imposer des

conditions *très fortes* sur la fonction. D'abord, on doit faire des hypothèses sur l'importance des composantes f_I correspondant aux sous-ensembles I dans $H(w_1, \ldots, w_d, d)$, sous forme d'une définition s'apparentant à celle de la dimension effective de f. Il faut également faire des hypothèses sur les $\tilde{f}(\mathbf{h})$ qui impliquent une borne supérieure *et* une borne inférieure sur $|\tilde{f}(\mathbf{h})|$ en fonction de $\|\mathbf{h}\|$. On a besoin de cette borne inférieure pour obtenir une borne sur $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$ qui soit en fonction de σ^2. En effet, pour arriver à cela, il n'est pas suffisant de savoir que les coefficients $|\tilde{f}(\mathbf{h})|$ avec $\|\mathbf{h}\| \geq 2^{\ell_I}$ sont petits. Il faut s'assurer qu'ils sont petits relativement à ce que l'on avait au départ, c.-à-d., par rapport à σ_I^2. De façon équivalente, cela signifie que l'on doit vérifier que les coefficients qui ont été éliminés (ceux tels que $\|\mathbf{h}\| < 2^{\ell_I}$) sont grands. Sans la borne inférieure sur $|\tilde{f}(\mathbf{h})|$, rien ne nous garantirait que ces coefficients ne sont pas tous nuls. Voici donc ce résultat :

Proposition 4.5.5 *Supposons que $f \in \mathcal{L}^2$ et qu'il existe une constante positive réelle $k \leq 1$ telle que*

$$\sum_{I \in H(w_1, \ldots, w_d, d)} \sigma_I^2 \geq k\sigma^2, \tag{4.20}$$

et qu'il existe $\alpha_1 \leq \ldots \leq \alpha_{w_1}$ telles que $2\alpha_j > j$ pour tout $j = 1, \ldots, w_1$ et deux constantes positives $c_1 \leq c_2$ telles que

$$|\tilde{f}(\mathbf{h})| \begin{cases} \leq c_2 \|\mathbf{h}\|^{-\alpha_j} & \text{si } \|\mathbf{h}\| \geq 2^{\ell_{I_\mathbf{h}}} \text{ et } I_\mathbf{h} \in H(w_1, \ldots, w_d, d), \\ \geq c_1 \|\mathbf{h}\|^{-\alpha_j} & \text{si } I_\mathbf{h} \in H(w_1, \ldots, w_d, d), \end{cases}$$

où $j = |I_\mathbf{h}|$.

Posons

$$\beta_1 = \max_{1 \leq j \leq w_1} (2\alpha_j - j) > 0$$

et

$$\beta_2 = \min_{1 \leq j \leq w_1} ((2\alpha_j - j)\lfloor m/j \rfloor).$$

Pour une règle de réseau polynômiale comptant $N = 2^m$ points et telle que $\Delta_{w_1, \ldots, w_d} = \Delta$, si $w_1 \geq w_j$ pour tout $j \leq d$, alors on a que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \left(\frac{c_2^2}{c_1^2} 2^{\Delta\beta_1 - \beta_2} 2^{w_1+1} + (1-k) \right) \sigma^2. \tag{4.21}$$

Le lemme suivant est utilisé dans la démonstration de la proposition 4.5.5 :

Lemme 4.5.2 *Si on pose*

$$N_h(n) = \textit{nombre de vecteurs } \mathbf{h} \textit{ dans } \mathbb{N}_I^* \textit{ tels que } \|\mathbf{h}\| = 2^n,$$

pour $n \geq 0$ entier, alors pour tout $r \geq 0$ entier,

$$\frac{N_h(n+r)}{N_h(n)} \leq 2^{(r+1)|I|+1}.$$

Démonstration : voir l'annexe B, page xxxvii.

Démonstration de la proposition 4.5.5 : par hypothèse, si $I \in H(w_1, \ldots, w_d, d)$ et que $|I| = j$, on a que

$$\sigma^2_{I,\mathrm{PLR}} \leq c_2^2 \sum_{\mathbf{h} \in \mathbb{N}_I^*, \|\mathbf{h}\| \geq 2^{\ell_I}} \|\mathbf{h}\|^{-2\alpha_j} = c_2^2 \sum_{n \geq \ell_I} N_h(n) 2^{-2\alpha_j n},$$

en utilisant la notation du lemme 4.5.2.

De plus, par la même procédure, on peut montrer que

$$\sigma_I^2 \geq c_1^2 \sum_{n \geq 0} N_h(n) 2^{-2\alpha_j n}$$

et ainsi,

$$\begin{aligned}
(\sigma^2_{I,\mathrm{PLR}}/\sigma_I^2) &\leq (c_2/c_1)^2 2^{-2\alpha_j \ell_I} \left(\sum_{n \geq 0} 2^{-2\alpha_j n} N_h(n + \ell_I) \right) \left(\sum_{n \geq 0} 2^{-2\alpha_j n} N_h(n) \right)^{-1} \\
&\leq (c_2/c_1)^2 2^{-2\alpha_j \ell_I} \max_{n \geq 0} (N_h(n + \ell_I)/N_h(n)) \\
&\leq (c_2/c_1)^2 2^{-(2\alpha_j - j)\ell_I + j + 1},
\end{aligned}$$

où la dernière inégalité suit par application du lemme 4.5.2. Notons que le rapport

$$\left(\sum_{n \geq 0} 2^{-2\alpha_j n} N_h(n) \right) / \left(\sum_{n \geq 0} 2^{-2\alpha_j n} N_h(n) \right)$$

qu'on laisse tomber à la deuxième ligne est bien défini et vaut 1, car

$$\sum_{n \geq 0} 2^{-2\alpha_j n} N_h(n) \leq \sum_{n \geq 0} 2^{-2\alpha_j n} 2^{(n+1)j} = \sum_{n \geq 0} 2^{-(2\alpha_j - j)n} 2^j < \infty,$$

puisque $2\alpha_j - j > 0$. En combinant, on obtient que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq (c_2/c_1)^2 \sum_{I \in H(w_1, \ldots, w_d, d)} 2^{-(2\alpha_{|I|} - |I|)\ell_I + |I| + 1} \sigma_I^2 + (1-k)\sigma^2.$$

Le résultat (4.21) est obtenu en notant que pour tout $I \in H(w_1, \ldots, w_d, d)$ avec $|I| = j$,

$$2^{-(2\alpha_j - j)\ell_I} = 2^{-(2\alpha_j - j)(\ell_I - \ell_j^* + \ell_j^*)} \leq 2^{(2\alpha_j - j)\Delta} 2^{-(2\alpha_j - j)\ell_j^*},$$

l'inégalité suivant par définition de Δ, par le fait que $2\alpha_j > j$ pour $j = 1, \ldots, w_1$ et par l'hypothèse que P_N est stationnaire dans la dimension. Ainsi, on obtient

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) &\leq (c_2/c_1)^2 \sum_{I \in H(w_1, \ldots, w_d, d)} 2^{(2\alpha_{|I|} - |I|)\Delta} 2^{-((2\alpha_{|I|} - |I|)\ell_{|I|}^* - |I| - 1)} \sigma_I^2 + (1 - k)\sigma^2 \\ &\leq \left((c_2/c_1)^2 \, 2^{\Delta\beta_1 - \beta_2} 2^{w_1 + 1} + (1 - k) \right) \sigma^2, \end{aligned}$$

par définition de β_1, β_2 et en utilisant le fait que $\sum_{I \in H(w_1, \ldots, w_d, d)} \sigma_I^2 \leq \sigma^2$. ∎

La raison pour laquelle nous avons supposé que les paramètres α_j augmentaient avec j est pour refléter l'hypothèse que si $|J| < |I|$, alors $\sigma_J^2 \geq \sigma_I^2$. Notons que cela concorde avec la condition que $2\alpha_j$ soit supérieur à j.

La borne (4.21) peut nous indiquer comment choisir les paramètres $d, w_1, \ldots, w_d$ qui définissent le critère $\Delta = \Delta_{w_1, \ldots, w_d}$. En effet, on sait que Δ croît lorsque d ou les w_j augmentent, car on considère alors plus de projections. Par contre, la quantité k dans la borne (4.20) augmente lorsque d ou les w_j augmentent. Puisque la borne (4.21) diminue lorsque Δ diminue ou que k augmente, cela signifie que les paramètres $d, w_1, \ldots, w_d$ devraient être choisis de façon à ce que si on les augmentait davantage, cela ferait augmenter Δ sans nous permettre d'augmenter k de façon importante. Autrement dit, si f_I n'est pas très importante, on ne veut pas que la projection $P_N(I)$ vienne affecter notre jugement sur P_N.

On peut voir dans le développement qui a mené au résultat précédent qu'en définissant autrement le critère Δ, on aurait pu arriver à une borne plus serrée sur la variance de $\hat{\mu}_{\mathrm{PLR}}$. C'est ce que l'on exploite dans le résultat suivant, en définissant le critère $\hat{\Delta}_{w_1, \ldots, w_d}$:

Proposition 4.5.6 *Soient $w_1, \ldots, w_d$ des entiers tels que $w_1 \geq w_j$, pour $j = 1, \ldots, d$. Supposons que $f \in \mathcal{L}^2$ et qu'il existe une constante positive $k \leq 1$ telle que*

$$\sum_{I \in H(w_1, \ldots, w_d, d)} \sigma_I^2 \geq k\sigma^2,$$

et qu'il existe $\alpha_1 \leq \ldots \leq \alpha_{w_1}$ *telles que* $2\alpha_j > j$ *pour* $j = 1, \ldots, w_1$ *et deux constantes positives* $c_1 \leq c_2$ *telles que*

$$|\tilde{f}(\mathbf{h})| \begin{cases} \leq c_2 \|\mathbf{h}\|^{-\alpha_j} & \text{si } \|\mathbf{h}\| \geq 2^{\ell_{I_\mathbf{h}}} \text{ et } I_\mathbf{h} \in H(w_1, \ldots, w_d, d), \\ \geq c_1 \|\mathbf{h}\|^{-\alpha_j} & \text{si } I_\mathbf{h} \in H(w_1, \ldots, w_d, d), \end{cases}$$

où $j = |I_\mathbf{h}|$.

Pour une règle de réseau polynômiale comptant $N = 2^m$ *points, si on pose*

$$\hat{\Delta}_{w_1,\ldots,w_d} = \min\left(\min_{1\leq j\leq w_1}(2\alpha_j - j)\ell_j, \min_{2\leq j\leq d}\min_{I\in S(w_j,j)}(2\alpha_j - j)\ell_I\right) \geq 0,$$

alors on a que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \left(\frac{c_2^2}{c_1^2} 2^{-\hat{\Delta}} 2^{w_1+1} + (1-k)\right)\sigma^2. \tag{4.22}$$

Démonstration : les hypothèses sur la fonction étant les mêmes qu'à la proposition 4.5.5, on a que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq (c_2/c_1)^2 \sum_{I\in H(w_1,\ldots,w_d,d)} 2^{-(2\alpha_{|I|}-|I|)\ell_I + |I|+1}\sigma_I^2 + (1-k)\sigma^2$$

et le résultat suit en utilisant le fait que P_N est stationnaire dans la dimension et donc, que $\min_{I\in H(w_1,\ldots,w_d,d)}(2\alpha_{|I|} - |I|)\ell_I = \hat{\Delta}_{w_1,\ldots,w_d}$. ∎

On a ici une borne qui décroît avec $\hat{\Delta}_{w_1,\ldots,w_d}$: c'est normal, car plus $\hat{\Delta}$ est grand, meilleure est l'équidistribution de P_N. Cette borne est plus serrée que celle donnée à la proposition 4.5.5, car en remplaçant ℓ_I par $\ell_I - \ell^*_{|I|} + \ell^*_{|I|}$ dans la définition de $\hat{\Delta}$, on obtient que $\hat{\Delta} \geq \min_j(-\Delta(2\alpha_j - j) + (2\alpha_j - j)\lfloor m/j \rfloor) \geq -\Delta\beta_1 + \beta_2$. Comment cette quantité est-elle reliée à notre critère Δ ? On peut voir $\hat{\Delta}_{w_1,\ldots,w_d}$ comme une version de $\Delta_{w_1,\ldots,w_d}$ qui utilise une normalisation basée sur les α_j plutôt que sur les ℓ^*_j et donc, sur la fonction plutôt que sur le nombre de points N. Si on suppose que $(2\alpha_j - j)$ croît avec j, alors les deux critères ont la caractéristique d'être plus sévères envers les ℓ_I lorsque $|I|$ est petit, car on veut que $(2\alpha_{|I|} - |I|)\ell_I$ soit grand et que $\ell^*_{|I|} - \ell_I$ soit petit. Donc, on peut dire que le critère $\hat{\Delta}$ utilise le même genre de normalisation que celle utilisée dans la définition de $\Delta_{w_1,\ldots,w_d}$.

Toutefois, pour utiliser le critère $\hat{\Delta}_{w_1,\ldots,w_d}$ en pratique, il faudrait connaître explicitement la valeur des différents paramètres α_j associés à la fonction, ce qui ne semble

pas être évident. Le critère $\Delta_{w_1,\ldots,w_d}$ ne requiert pas cette connaissance et utilise une normalisation qui va dans le même sens que celle basée sur les α_j. Les deux critères devraient donc *grosso modo* choisir les mêmes règles. L'avantage de $\hat{\Delta}$ est simplement que la borne sur $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$ que l'on obtient en fonction de ce critère est plus serrée que celle définie en fonction de Δ.

À partir des propositions précédentes, on peut donner des conditions sur f et P_N garantissant la réduction de variance par rapport à la méthode MC.

Corollaire 4.5.1 *Soit P_N une règle de réseau polynômiale avec $N = 2^m$. Supposons que f satisfait les conditions énoncées à la proposition 4.5.5 avec $k \geq 1 - c_3/N$, pour c_3 une constante indépendante de N et que $w_1 \geq w_j$ pour $j = 1, \ldots, d$. Si $\hat{\Delta}_{w_1,\ldots,w_d} \geq m$ ou que $\Delta_{w_1,\ldots,w_d}\beta_1 - \beta_2 \leq -m$, où β_1 et β_2 sont définies à la proposition 4.5.5, alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \left(\frac{c_2^2}{c_1^2} 2^{w_1+1} + c_3\right) \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : par hypothèse, on sait que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq N \left(\frac{c_2^2}{c_1^2} 2^{\Delta\beta_1 - \beta_2} 2^{w_1+1} + (1-k)\right) \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$$

en appliquant la proposition 4.5.5, et que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq N \left(\frac{c_2^2}{c_1^2} 2^{-\hat{\Delta}+w_1+1} + (1-k)\right) \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}),$$

par la proposition 4.5.6. Le résultat suit par application des hypothèses sur Δ, $\hat{\Delta}$ et k, puisqu'on obtient alors que $N\left((c_2/c_1)^2 2^{\Delta\beta_1-\beta_2} 2^{w_1+1} + (1-k)\right) \leq (c_2/c_1)^2 2^{w_1+1} + c_3$ et que $N\left((c_2/c_1)^2 2^{-\hat{\Delta}} 2^{w_1+1} + (1-k)\right) \leq (c_2/c_1)^2 2^{w_1+1} + c_3$. ∎

Étant donné toutes les ressemblances que nous avons notées jusqu'à maintenant entre le cas standard et le cas polynômial, on pourrait penser que les bornes développées dans cette sous-section pourraient aussi être démontrées dans le cas standard, avec le critère $M_{t_1,\ldots,t_d}$. Le problème qui survient lorsque l'on essaie de faire cela, c'est qu'au lieu de calculer la quantité $N_h(n)$ définie au lemme 4.5.2, page 158, il faut être capable de compter le nombre de vecteurs $\mathbf{h} \in \mathbb{Z}_I^*$ tels que $\|\mathbf{h}\|_2^2 = k$, pour $k \in \mathbb{N}$ et cela complique beaucoup le calcul, surtout lorsque $s > 8$ [41]. Nous avons essayé d'utiliser des bornes sur ce nombre mais en faisant cela, les conditions requises sur la fonction

devenaient trop irréalistes. L'amélioration de ces résultats dans le cas standard est un des sujets de recherche que nous aimerions explorer prochainement.

4.5.3 Comparaison avec les (t, m, s)-réseaux brouillés

Maintenant, comment les règles de réseau polynômiales XOR-translatées se comparent-elles aux (t, m, s)-réseaux brouillés pour ce qui est de la variance ? Nous croyons qu'il est important de répondre à cette question, car cette méthode due à Owen a reçu beaucoup d'attention parmi la communauté QMC dans les dernières années. Pour y répondre, nous allons d'abord rappeler brièvement les résultats obtenus par Owen pour les (t, m, s)-réseaux brouillés.

Proposition 4.5.7 *([106, Theorem 1]) Soit $\hat{\mu}_{\text{tms}}$ l'estimateur formé à partir des $N = 2^m$ points d'un (t, m, s)-réseau brouillé en base b, $\hat{\mu}_{\text{MC}}$ l'estimateur MC basé sur 2^m points et soit $f \in \mathcal{L}^2$. Alors*

$$\text{Var}(\hat{\mu}_{\text{tms}}) = o(1/N) \text{ quand } N \to \infty \tag{4.23}$$

et

$$\text{Var}(\hat{\mu}_{\text{tms}}) = b^t \left(\frac{b+1}{b-1}\right)^s \text{Var}(\hat{\mu}_{\text{MC}}). \tag{4.24}$$

Si $t = 0$, alors

$$\text{Var}(\hat{\mu}_{\text{tms}}) \leq \left(\frac{b}{b-1}\right)^{s-1} \text{Var}(\hat{\mu}_{\text{MC}}). \tag{4.25}$$

Proposition 4.5.8 *([109, Theorem 2]) Sous les conditions de [106, Theorem 1] (voir la proposition 4.5.7), si f est à variation bornée au sens de Hardy et Krause, alors*

$$\text{Var}(\hat{\mu}_{\text{tms}}) = O(N^{-2}(\log N)^{2(s-1)}). \tag{4.26}$$

Si $\partial^s f/\partial X$ est Lipschitz-continue, c.-à-d., s'il existe $A > 0$ et $\beta \in (0, 1]$ tels que

$$\left|\frac{\partial^s}{\partial \mathbf{x}} f(\mathbf{x}) - \frac{\partial^s}{\partial \mathbf{x}} f(\mathbf{x}^*)\right| \leq A\|\mathbf{x} - \mathbf{x}^*\|_2^\beta,$$

pour tous $\mathbf{x}, \mathbf{x}^* \in [0, 1)^s$, alors

$$\text{Var}(\hat{\mu}_{\text{tms}}) = O(N^{-3}(\log N)^{(s-1)}). \tag{4.27}$$

La borne asymptotique (4.26) est obtenue en utilisant le fait que pour ce type de fonction, on a que $(\hat{\mu}_{\text{tms}} - \mu) = O(N^{-1}(\log N)^{s-1})$ avec probabilité 1, par simple application de l'inégalité de Koksma-Hlawka, puisque les (t, m, s)-réseaux brouillés sont des (t, m, s)-réseaux avec probabilité 1 [104] et donc, leur discrépance-étoile est dans $O(N^{-1}(\log N)^{s-1})$. Le résultat pour la variance suit en utilisant le fait que $\mathrm{E}(\hat{\mu}_{\text{tms}}) = \mu$.

Les autres résultats ont été obtenus en utilisant une décomposition en fonctions de Haar. De cette façon, la variance est donnée par

$$\mathrm{Var}(\hat{\mu}_{\text{tms}}) = \frac{1}{N} \sum_{\emptyset \neq I \subseteq S} \sum_{\mathbf{k}_I} \Gamma_{I,\mathbf{k}_I} \sigma^2_{I,\mathbf{k}_I}, \tag{4.28}$$

où $\sigma^2_{I,\mathbf{k}_I}$ représente la variance de la fonction de base qui est constante sur chaque boîte déterminée par la partition induite par $\mathbf{k}_I$ et $\Gamma_{I,\mathbf{k}_I}$ est déterminé par les propriétés d'équidistribution du (t, m, s)-réseau et par la randomisation choisie.

Cette expression pour la variance peut être comparée facilement avec la variance de $\hat{\mu}_{\text{MC}}$, qui est obtenue en posant $\Gamma_{I,\mathbf{k}_I} = 1$ pour tout $I, \mathbf{k}_I$. Pour les (t, m, s)-réseaux, ces $\Gamma_{I,\mathbf{k}_I}$ valent 0 lorsque les boîtes associées ont toutes le même nombre de points. Les inégalités (4.24) et (4.25) sont donc obtenues selon la borne (indépendante de N) qui peut être trouvée sur $\Gamma_{I,\mathbf{k}_I}$ pour les autres valeurs de $\mathbf{k}_I$. Le résultat (4.23) vient du fait que les $\Gamma_{I,\mathbf{k}_I}$ sont tous bornés par des quantités indépendantes de N. Finalement, la borne asymptotique (4.27) est obtenue en démontrant que pour ce type de fonction, $\sigma^2_{I,\mathbf{k}_I} = O(b^{-2\kappa})$ [106], où $\kappa = \sum_{j \in I} k_j$.

On peut établir une correspondance entre la décomposition en fonctions de Haar et celle en série de Walsh, que l'on a utilisée dans l'expression pour la variance de $\hat{\mu}_{\text{PLR}}$ à la proposition 4.5.2 :

Lemme 4.5.3 *Soit $f \in \mathcal{L}^2$. On a que*

$$\sigma^2_{I,\mathbf{k}_I} = \sum_{\mathbf{h} \in H(\mathbf{k}_I)} |\bar{f}(\mathbf{h})|^2,$$

où $H(\mathbf{k}_I) = \{\mathbf{h} \in \mathbb{N}_I^* : |h_j|_p = 2^{k_j}$, pour tout $j \in I\}$.

Démonstration : voir l'annexe B, page xxxviii.

Ce que ce résultat signifie, c'est que dans le cas de l'estimateur $\hat{\mu}_{\mathrm{PLR}}$, le facteur $\Gamma_{I,\mathbf{k}_I}$ dans (4.28) correspond au rapport

$$\frac{N \sum_{\mathbf{h} \in \mathcal{L}_s^* \cap H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2}{\sum_{\mathbf{h} \in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2}. \tag{4.29}$$

On sait que ce rapport vaut 0 lorsque $\mathbf{k}_I$ induit une équidistribution, car il n'y a alors aucun $\mathbf{h}$ dans $\mathcal{L}_s^*$. Mais pour les autres $\mathbf{k}_I$, il semble difficile de borner cette quantité. Le problème vient du fait que même si on sait que $\mathcal{L}_s^* \cap H(\mathbf{k}_I) \subset H(\mathbf{k}_I)$, on peut toujours tomber sur une fonction telle que $\tilde{f}(\mathbf{h})$ est nul lorsque $\mathbf{h} \notin \mathcal{L}_s^*$ et dans ce cas, le rapport (4.29) vaut N. Pour cette raison, on ne peut arriver à démontrer l'équivalent du résultat (4.24) pour les règles de réseau XOR-translatées. Si on se rappelle bien, on avait le même problème avec la translation aléatoire dans le cas standard, c.-à-d., pour une règle donnée, on peut toujours trouver une fonction qui va nous donner $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = N\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$.

Cela nous mène à la question suivante : à quel endroit dans la méthode employée par Owen pour démontrer que (4.28) tient avec $\Gamma_{I,\mathbf{k}_I}$ indépendant de N utilise-t-on un argument qui ne tient pas dans le cas des règles de réseau polynômiales XOR-translatées? Pour répondre à cela, notons que les propriétés de l'ensemble de points randomisé $\tilde{P}_N$ requises dans le développement d'Owen sont [105, 49] :

(1) chaque point $\tilde{\mathbf{x}}_i$, $i = 1, \ldots, N$ de $\tilde{P}_N$ suit la loi uniforme sur $[0,1)^s$:

(2) $\tilde{P}_N$ est un (t, m, s)-réseau (avec probabilité 1) ;

(3) [49, *Algorithm 1*] pour $1 \leq i, j, \leq N$ et $1 \leq r \leq s$, si $x_{i,r,l} = x_{j,r,l}$ pour $l = 1, \ldots, k$ mais que $x_{i,r,k+1} \neq x_{j,r,k+1}$, alors $\tilde{x}_{i,r,k+1} \neq \tilde{x}_{j,r,k+1}$ et $\tilde{x}_{i,r,k+2}, \tilde{x}_{i,r,k+3}, \ldots, \tilde{x}_{j,r,k+2}, \tilde{x}_{j,r,k+3}, \ldots$ ne sont pas corrélés.

Pour une règle de réseau polynômiale XOR-translatée, la première condition est respectée. En ce qui concerne la deuxième condition, pour n'importe quelle partition $\mathbf{k}_I$ de l'hypercube, la XOR-translation a seulement pour effet de permuter les boîtes selon les k_j premiers bits de $\boldsymbol{U}$ dans chaque dimension $j \in I$. Ainsi, les propriétés d'équidistribution de la règle XOR-translatée sont exactement les mêmes que celles de la règle non translatée : la deuxième condition est donc respectée. Par contre, la troisième condition n'est pas respectée, car pour une règle de réseau XOR-translatée,

si on pose $\mathbf{y} = \mathbf{x}_j \oplus \boldsymbol{U}$, alors $(\mathbf{x}_i \oplus \boldsymbol{U} \mid \mathbf{y}) = \mathbf{x}_j \oplus \mathbf{x}_i \oplus \mathbf{y}$ et donc, $\mathbf{x}_i \oplus \boldsymbol{U}$ est complètement déterminé une fois que l'on connaît $\mathbf{x}_j \oplus \boldsymbol{U}$. Autrement dit, la XOR-translation injecte dans P_N moins d'aléatoire que le "brouillage". À cause de cela, le facteur $\Gamma_{I,\mathbf{k}_I}$ associé aux règles de réseau polynômiales n'a pas les mêmes propriétés que celui associé aux (t, m, s)-réseaux brouillés. C'est pour cette raison que l'on n'arrive pas à borner le rapport (4.29) par une quantité indépendante de N. Rappelons que dans le cas des règles de réseau standard, nous avions le même problème (c.-à-d., la translation modulo 1 n'était pas assez aléatoire) et il a fallu randomiser davantage pour démontrer des résultats de réduction de variance par rapport à MC.

Puisque dans le cas où N est fini, on ne peut arriver à démontrer des résultats bornant la variance de $\hat{\mu}_{\mathrm{PLR}}$ en fonction de la variance de l'estimateur MC pour n'importe quelle fonction dans $\mathcal{L}^2$, regardons maintenant quels résultats obtient-on pour la variance de $\hat{\mu}_{\mathrm{PLR}}$ sous les conditions menant à (4.26) et (4.27) dans le cas des (t, m, s)-réseaux brouillés.

Pour obtenir l'équivalent de (4.26) pour les règles de réseau polynômiales, il suffit de connaître la discrépance pour ce type d'ensemble de points. Or, dans [95, 98, 64], on démontre qu'il existe des polynômes $P(z)$ permettant de construire des générateurs de Tausworthe tels que la discrépance-étoile associée à P_N est dans $O(N^{-1}(\log N)^s \log\log N)$. Cela signifie que pour les fonctions à variation bornée, il existe une règle de réseau polynômiale telle que

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) = O(N^{-2}(\log N)^{2s}(\log\log N)^2).$$

Remarquons que de façon générale, pour donner des résultats asymptotiques dans le cas des règles de réseau polynômiales, on doit faire l'hypothèse supplémentaire que l'on peut construire une suite de règles de réseau telle que chaque règle dans la suite contient deux fois plus de points que la précédente et dont les paramètres t_I sont uniformément bornés par une quantité indépendante de m sur toutes les règles de la suite. Nous croyons que cela peut être fait avec des règles de réseau polynômiales, mais c'est un sujet que nous n'explorons pas ici et qui sera abordé dans des travaux futurs. Cependant, par souci de clarté, nous préférons énoncer les résultats sur la variance de $\hat{\mu}_{\mathrm{PLR}}$ sans avoir recours à la notation asymptotique, en utilisant plutôt des bornes de

la forme (constante $\times\, g(N)$).

Dans le cas des fonctions *lisses* (dans [106], on réfère de cette façon aux fonctions ayant une dérivée partielle mixte d'ordre s satisfaisant une condition de Lipschitz, tel que requis dans l'énoncé de la proposition 4.5.8), on a le résultat intermédiaire énoncé au lemme 4.5.4. L'équivalent est donné dans [106, *Lemma 2*], mais en utilisant la décomposition de f en fonctions de Haar. Ici, on passe par la décomposition en série de Walsh, mais notre démonstration est bien sûr inspirée de [106].

Lemme 4.5.4 *Soit $f \in \mathcal{L}^2$ lisse (telle que $\partial^s f/\partial X$ est Lipschitz-continue). Soit $\mathbf{k}_I = (k_j)_{j\in I}$ un vecteur d'entiers $k_j \geq 0$ et $\kappa = \sum_{j\in I} k_j$. Alors*

$$\sum_{\mathbf{h}\in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2 \leq \left(c_f(\beta) 2^{-4|I|}\right) 2^{-2\kappa},$$

où $H(\mathbf{k}_I) = \{\mathbf{h} \in \mathbb{N}_I^* : |h_j|_p = 2^{k_j},$ pour tout $j \in I\}$ et

$$c_f(\beta) = \left[\int_{[0,1)^{|I|}} \left(\frac{\partial f(\mathbf{x}_I)}{\partial \mathbf{x}_I}\right)^2 d\mathbf{x}_I\right] \left(1 + O(2^{-\beta \min_{j\in I} k_j})\right).$$

Démonstration : voir l'annexe B, page xxxix.

Remarquons que cette borne est exprimée en utilisant la norme $\|\mathbf{h}\|_\pi$. En effet, tous les $\mathbf{h}$ dans $H(\mathbf{k}_I)$ sont tels que $\|\mathbf{h}\|_\pi = 2^\kappa$. Ainsi, tout comme dans l'étude de la variance de $\hat{\mu}_{\mathrm{LR}}$ lorsque la fonction est polynômiale (voir à la section 3.4), nous avons été conduits naturellement à un choix de norme bien précis pour mesurer les vecteurs $\mathbf{h}$ dans le réseau dual, sans que l'hypothèse sur la fonction ait été donnée en faisant intervenir $\|\mathbf{h}\|$ où $\tilde{f}(\mathbf{h})$.

À ce point-ci, pour arriver au même résultat qu'Owen, il faudrait être capable de démontrer que :

$$\frac{\sum_{\mathbf{h}\in\mathcal{L}_s^*\cap H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2}{\sum_{\mathbf{h}\in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2} \leq \frac{c}{2^{m-q_I}}, \tag{4.30}$$

où c est une quantité indépendante de m. Or, pour les raisons mentionnées précédemment, on ne peut y arriver et ce, même si nous avons fait l'hypothèse que f était lisse (c.-à-d., $\partial^s f/\partial X$ était Lipschitz-continue). En effet, même si cette condition sur f nous permet de démontrer que les coefficients $|\tilde{f}(\mathbf{h})|^2$ décroissent assez rapidement, ce n'est pas suffisant pour démontrer (4.30) puisque cela n'empêche pas que l'on ait $|\tilde{f}(\mathbf{h})|^2 = 0$ si $\mathbf{h} \notin \mathcal{L}_s^*$. Pour l'instant, voici le mieux que l'on puisse faire :

Corollaire 4.5.2 *Soit P_N une règle de réseau polynômiale qui satisfait la propriété de (τ, m, s)-réseau avec $\tau = t$. Soit $f \in \mathcal{L}^2$ lisse (c.-à-d., telle que $\partial^s f/\partial \mathbf{x}$ est Lipschitz-continue). Alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \left(c_f(\beta)(4/3)^s 2^{2t}\right) N^{-2} (\lg N)^{s-1},$$

où $c_f(\beta)$ est défini au lemme 4.5.4.

Nous avons besoin du lemme suivant pour la démonstration :

Lemme 4.5.5 *Si $j \leq m-u+1$ et que $b > 1$, avec $j, u, m \in \mathbb{N}$, $u, m \geq 1$, $j \geq 0$, alors*

$$\sum_{k \geq 1} b^{-k} \binom{k+m-j-1}{u-1} \leq (1-b^{-1})^{-u} m^{u-1}. \tag{4.31}$$

Démonstration : voir l'annexe B, page xl..

Démonstration du corollaire 4.5.2 : on utilise la borne (4.19), le lemme 4.5.4 et le fait que pour toute valeur de κ, il existe $\binom{\kappa+|I|-1}{|I|-1}$ vecteurs $\mathbf{k}_I$ tels que $\sum_{j \in I} k_j = \kappa$. Ainsi, on a que

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) &\leq \sum_{\emptyset \neq I \subseteq S} \sum_{\kappa \geq m - q_I - |I| + 1} 2^{-2\kappa - 4|I|} \binom{\kappa + |I| - 1}{|I| - 1} c_f(\beta) \\
&= \sum_{\emptyset \neq I \subseteq S} 2^{-2|I|} 2^{-2m+2q_I} \sum_{\kappa \geq 1} 2^{-2\kappa} \binom{\kappa + m - q_I - 1}{|I| - 1} c_f(\beta) \\
&\leq c_f(\beta) 2^{-2m} \sum_{\emptyset \neq I \subseteq S} 2^{-2|I| + 2q_I} (4/3)^{|I|} m^{|I|-1} \\
&\leq c_f(\beta) 2^{2t} N^{-2} (\lg N)^{s-1} \sum_{\emptyset \neq I \subseteq S} 3^{-|I|} \\
&\leq c_f(\beta) (4/3)^s 2^{2t} N^{-2} (\lg N)^{s-1},
\end{aligned}$$

où la deuxième inégalité découle du lemme 4.5.5. ∎

4.5.4 Bornes sur la variance : autres cas

La comparaison avec les (t, m, s)-réseaux brouillés nous a permis d'étudier la variance de l'estimateur $\hat{\mu}_{\mathrm{PLR}}$ lorsque la fonction est à variation bornée ou lisse. Dans ce qui suit, nous établissons des bornes sur $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$ lorsque les coefficients de Walsh satisfont certaines conditions. Ce genre d'hypothèses facilite cette étude, puisque l'on a l'expression (4.13) pour $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}})$ en termes de ces coefficients, mais cela a le

désavantage qu'il n'est pas toujours évident de traduire ces hypothèses en conditions sur la fonction qui puissent être facilement vérifiées. Le résultat suivant sera utile dans ce qui suit, car il nous permet de borner le nombre de vecteurs $\mathbf{h}$ dans $H(\mathbf{k}_I) = \{\mathbf{h} \in \mathbb{N}_I^* : |h_j|_p = 2^{k_j}, \text{ pour tout } j \in I\}$ qui sont dans le réseau dual $\mathcal{L}_s^*$:

Lemme 4.5.6 *Soit $\mathbf{k}_I = (k_j)_{j\in I}$ un vecteur d'entiers $k_j \geq 0$ et $\kappa = \sum_{j\in I} k_j$. Si P_N est une règle de réseau polynômiale avec $N = 2^m$ points, alors*

$$|H(\mathbf{k}_I) \cap \mathcal{L}_s^*| \begin{cases} \leq & 2^{\kappa-(m-q_I)} & \text{si } \kappa \geq m - q_I + 1, \\ \leq & 1 & \text{si } m - q_I - |I| + 1 \leq \kappa < m - q_I + 1 \\ = & 0 & \text{sinon.} \end{cases}$$

Démonstration : voir l'annexe B, page xli.

Remarque 4.5.1 *Nous avons expliqué à la page 167 que pour arriver à une borne de type constante $\times N^{-3}\log^{s-1} N$ dans le cas des fonctions lisses, il faudrait avoir*

$$\frac{\sum_{\mathbf{h}\in\mathcal{L}_s^*\cap H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2}{\sum_{\mathbf{h}\in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2} \leq \frac{c}{2^{m-q_I}}.$$

Le résultat précédent nous montre que le rapport du nombre de termes au numérateur sur le nombre de termes au dénominateur est bien égal à 2^{-m+q_I}, puisque $|H(\mathbf{k}_I)| = 2^\kappa$. Par contre, pour que cela nous aide, il faudrait montrer que chaque $|\tilde{f}(\mathbf{h})|^2$ est dans un intervalle $[c_1, c_2]$, avec $c_1 > 0$.

Nous allons séparer les résultats selon le type de norme choisie pour mesurer les vecteurs $\mathbf{h}$, en commençant avec la norme produit et ensuite la norme *sup*.

Proposition 4.5.9 *Soit P_N une règle de réseau polynômiale avec $N = 2^m$ qui satisfait la propriété de (τ, m, s)-réseau avec $\tau = t$. Soit $f \in \mathcal{L}^2$ telle qu'il existe $c > 0$ et $\alpha \geq 1$ tels que*

$$|\tilde{f}(\mathbf{h})| \leq c\|\mathbf{h}\|_\pi^{-\alpha}$$

pour tout $\mathbf{h}$ tel que $\|\mathbf{h}\|_\pi \geq \min_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s^*} \|\mathbf{h}\|_\pi$. Alors

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \left(c^2 2^{2\alpha t}(1 + 2^{4\alpha-2})^s\right) N^{-2\alpha}(\lg N)^{s-1}.$$

Démonstration : posons $\kappa^* = m - q_I - |I| + 1$ et rappelons que $K_I(\cdot)$ a été défini en (4.17), à la page 157. On sait que

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) &= \sum_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s^*} |\tilde{f}(\mathbf{h})|^2 \\
&= \sum_{\emptyset\neq I\subseteq S} \sum_{\mathbf{k}_I\in K_I(\kappa^*)} \sum_{\mathbf{h}\in\mathcal{L}_s^*\cap H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2 \\
&\leq c^2 \sum_{\emptyset\neq I\subseteq S} \sum_{\mathbf{k}_I\in K_I(\kappa^*)} 2^{-2\alpha\kappa} |H(\mathbf{k}_I)\cap\mathcal{L}_s^*| \\
&\leq c^2 \sum_{\emptyset\neq I\subseteq S} \sum_{\mathbf{k}_I\in K_I(\kappa^*)} 2^{-2\alpha\kappa} 2^{\kappa-m+q_I} \\
&= c^2 \sum_{\emptyset\neq I\subseteq S} \sum_{\kappa\geq\kappa^*} 2^{-(2\alpha-1)\kappa} 2^{-m+q_I} \binom{\kappa+|I|-1}{|I|-1} \\
&= c^2 \sum_{\emptyset\neq I\subseteq S} 2^{-m+q_I} 2^{-(2\alpha-1)(\kappa^*-1)} \sum_{\kappa\geq 1} 2^{-(2\alpha-1)\kappa} \binom{\kappa+m-q_I-1}{|I|-1} \\
&\leq c^2 \sum_{\emptyset\neq I\subseteq S} 2^{-2\alpha m} 2^{2\alpha q_I+(2\alpha-1)|I|} 2^{(2\alpha-1)|I|} m^{|I|-1} \\
&= \left(c^2\, 2^{2\alpha t}(1+2^{4\alpha-2})^s\right) N^{-2\alpha} (\lg N)^{s-1}.
\end{aligned}$$

Dans ce qui précède, la deuxième égalité est obtenue en utilisant le fait que si $\|\mathbf{h}\|_\pi < 2^{\kappa^*}$, alors $\mathbf{h}$ n'est pas dans $\mathcal{L}_s^*$; la première inégalité est obtenue en appliquant l'hypothèse sur les $|\tilde{f}(\mathbf{h})|$; la deuxième vient du lemme 4.5.6 ; l'égalité à la cinquième ligne vient du fait que le nombre de $\mathbf{k}_I$ tels que $\sum_{j\in I} k_j = \kappa$ vaut $\binom{\kappa+|I|-1}{|I|-1}$; la dernière inégalité est obtenue en appliquant le lemme 4.5.5 et en notant que $(1-2^{-(2\alpha-1)})^{-|I|} \leq 2^{2(\alpha-1)|I|}$, puisque $\alpha \geq 1$. ∎

Cela signifie que pour arriver à un ordre de convergence de $O(N^{-3}(\log N)^{s-1})$, il faut que la vitesse de décroissance des coefficients $|\tilde{f}(\mathbf{h})|$ soit de $\alpha = 3/2$. Le résultat que nous avons obtenu au corollaire 4.5.2 pour les fonctions lisses est un cas particulier du résultat précédent, car on a montré au lemme 4.5.4 que pour ce type de fonction, la condition requise dans la proposition 4.5.9 était satisfaite avec $\alpha = 1$.

La proposition précédente ressemble au résultat donné dans [65, *Theorem 1*], qui dit que si les coefficients de Walsh de f satisfont $|\tilde{f}(\mathbf{h})| \leq c\prod_{j=1}^s (\max(1,|h_j|))^{-\alpha}$ pour tout $\mathbf{h}$, alors l'erreur d'intégration obtenue par un (t,m,s)-réseau digital est dans $O(N^{-\alpha}\log^{s-1} N)$. La différence est qu'ils doivent supposer que la fonction f a une représentation en série de Walsh qui est absolument convergente afin d'utiliser

l'expression pour l'erreur donnée à la proposition 4.5.1. Ainsi, ils doivent supposer que $\alpha > 1$, car avec $\alpha = 1$, la condition sur la décroissance des $|\tilde{f}(\mathbf{h})|$ n'est pas suffisante pour assurer la convergence absolue de la série $\sum_{\mathbf{h}} |\tilde{f}(\mathbf{h})|$. Cela signifie que leur résultat ne permet pas de traiter le cas des fonctions lisses en général.

Regardons maintenant ce qui se passe dans le cas où on utilise la norme *sup*. Dans ce cas, si on veut que l'hypothèse sur le taux de décroissance des $|\tilde{f}(\mathbf{h})|$ soit à peu près équivalente à celle que nous avons utilisée dans la proposition précédente, il faut supposer qu'il y a un paramètre α qui est associé à chaque valeur de $|I_{\mathbf{h}}| = 1, \ldots, s$. En effet, avec la norme $\|\cdot\|_\pi$, les vecteurs $\mathbf{h}$ deviennent "plus longs" quand $|I|$ augmente, car on somme les puissances k_j sur plus de dimensions. Avec la norme *sup*, ce n'est pas le cas puisque l'on prend le maximum $|h_j|_{\mathrm{p}}$ sur chaque dimension. Les paramètres α_j servent à donc à faire en sorte que $\|\mathbf{h}\|^{\alpha_j} \approx \|\mathbf{h}\|_\pi^\alpha$ et ce, en supposant que $\alpha_1 \leq \ldots \leq \alpha_s$.

Proposition 4.5.10 *Soit P_N une règle de réseau polynômiale avec $N = 2^m$ et $f \in \mathcal{L}^2$. Supposons qu'il existe des constantes $c > 0$ et $\alpha_1 \leq \ldots \leq \alpha_s$ satisfaisant $2\alpha_j - j \geq 1$ pour $j = 1, \ldots, s$ et telles que, si on pose $j = |I_{\mathbf{h}}|$, alors*

$$|\tilde{f}(\mathbf{h})| \leq c\|\mathbf{h}\|^{-\alpha_j}$$

pour tout $\mathbf{h}$ tel que $\|\mathbf{h}\| \geq \min_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s^*} \|\mathbf{h}\|$. Posons

$$\beta_1 = max_{1\leq j\leq s}(2\alpha_j - j) \geq 1$$

(comme dans la proposition 4.5.5),

$$\beta_3 = min_{1\leq j\leq s}(2\alpha_j/j) > 1$$

et $\Delta = \Delta_{s,\ldots,s} = \max_{\emptyset\neq I\subseteq S}(\ell^*_{|I|} - \ell_I)$. Alors

$$\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) \leq \left(c^2\, 2^{s+\beta_1(\Delta+2)}\right) N^{-\beta_3+1}.$$

Démonstration : on a que

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) &= \sum_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s^*} |\tilde{f}(\mathbf{h})|^2 \\
&= \sum_{\emptyset\neq I\subseteq S} \sum_{k\geq\ell_I} \sum_{\mathbf{h}\in\mathcal{L}_s^*\cap\mathbf{N}_I^*} |\tilde{f}(\mathbf{h})|^2 \mathbf{1}_{\|\mathbf{h}\|=2^k}
\end{aligned}$$

$$
\begin{aligned}
&\leq c^2 \sum_{\emptyset \neq I \subseteq S} \sum_{k \geq \ell_I} 2^{-2\alpha_{|I|}k} 2^{(k+1)|I|} \\
&= c^2 \sum_{\emptyset \neq I \subseteq S} 2^{-(2\alpha_{|I|}-|I|)\ell_I + |I|} \sum_{k \geq 0} 2^{-(2\alpha_{|I|}-|I|)k} \\
&\leq c^2 2^s \sum_{\emptyset \neq I \subseteq S} 2^{-(2\alpha_{|I|}-|I|)(\ell_I - 1)} \\
&= c^2 2^s \sum_{\emptyset \neq I \subseteq S} 2^{-(2\alpha_{|I|}-|I|)(\ell_I - \ell^*_{|I|} + \ell^*_{|I|} - 1)} \\
&\leq c^2 2^s 2^{\beta_1 \Delta} 2^{-(2\alpha_{|I|}-|I|)((m/|I|)-(|I|-1)/|I|-1)} \\
&\leq c^2 2^s 2^{\beta_1(\Delta+2)} 2^{-(\beta_3-1)m},
\end{aligned}
$$

la première inégalité vient du fait que le nombre de $\mathbf{h} \in \mathbb{N}_I^*$ tels que $\|\mathbf{h}\| = 2^k$ est égal à $(2^{k+1}-1)^{|I|} - (2^k-1)^{|I|} \leq 2^{(k+1)|I|}$; la deuxième inégalité suit par hypothèse que $2\alpha_j - j \geq 1$ pour $j = 1, \ldots, s$ et donc, la somme sur k à la ligne précédente est bornée par $2^{2\alpha_{|I|}-|I|}$. L'avant-dernière inégalité vient du fait que $\ell^*_{|I|} = \lfloor m/|I| \rfloor \geq m/|I| - (|I|-1)/|I|$. ∎

Ce qui nous a empêché d'arriver à une borne plus serrée dans le cas de cette norme est relié au fait qu'il est difficile de compter le nombre de vecteurs $\mathbf{h}$ dans le réseau dual qui sont tels que $\|\mathbf{h}\| = 2^k$ et pour cette raison, nous avons inclus tous les $\mathbf{h}$ ayant une norme supérieure ou égale à 2^{ℓ_I} dans le calcul de la borne, à la proposition 4.5.10. Ce calcul est plus facile dans le cas de la norme $\|\cdot\|_\pi$, car ce nombre de vecteurs est relié au nombre de points par boîte, qui lui-même est relié au volume de la boîte, qui lui dépend de $\sum_{j \in I} k_j$ et non de $\max_{j \in I} k_j$. Voici un autre exemple du fait que le choix de la norme $\|\cdot\|_\pi$ semble plus naturel lorsque l'on étudie les règles de réseau polynômiales. Néanmoins, tout comme c'était le cas pour les règles de réseau standard (choix entre la norme produit et la norme euclidienne), nous préférons utiliser la norme *sup* pour définir les critères de sélection, car cela permet un calcul plus rapide et ainsi, les recherches peuvent être faites en plus grande dimension.

4.6 Résultats numériques

Nous présentons d'abord à la sous-section 4.6.1 les résultats obtenus en faisant des recherches à l'aide du critère $\Delta_{w_1,\ldots,w_d}$. Puis, nous utilisons les règles de réseau ainsi obtenues pour construire des estimateurs sur le problème des options asiatiques, à la

sous-section 4.6.2, et sur une fonction-test, à la sous-section 4.6.3.

4.6.1 Résultats des recherches

Nous avons effectué des recherches parmi tous les générateurs de Tausworthe combinés à deux ou trois composantes et dont les polynômes caractéristiques satisfont les conditions d'implantation mentionnées à la sous-section 4.1.2, afin de trouver les meilleurs par rapport au critère $\Delta_{10,10,10}$. Les résultats de ces recherches sont donnés au tableau 4.2. Dans ce tableau, $\delta_{w_u,u}$ est défini par

$$\delta_{w_u,u} = \begin{cases} \max_{I \in S(w_u,u)} (\ell_u^* - \ell_I) & \text{si } u > 1, \\ \max_{1 \le j \le w_u} \left(\ell_j^* - \ell_j\right) & \text{si } u = 1 \end{cases}$$

et donc, on a que $\Delta_{w_1,\ldots,w_d} = \max_{1 \le u \le d} \delta_{w_u,u}$. Nous avons décomposé le critère de cette façon afin de voir, pour chaque type de sous-ensemble I (indices successifs, paires ou triplets), quelle était la qualité des projections $P_N(I)$ correspondantes. Les paramètres (m_j, q_j, v_j) de chaque composante du générateur combiné sont donnés dans la deuxième colonne du tableau.

TABLEAU 4.2: Meilleurs générateurs combinés avec leur $\delta_{10,u}$

m	(m_j, q_j, v_j)	$\delta_{10,1}$	$\delta_{10,2}$	$\delta_{10,3}$	$\Delta_{10,10,10}$
10	(7,1,4) (3,1,1)	1	1	1	1
12	(5,2,3) (4,1,2) (3,1,1)	0	1	2	2
14	(9,4,5) (5,2,2)	1	1	1	1
16	(9,4,3) (7,1,4)	1	1	1	1

Nous avons par la suite fait d'autres recherches parmi le même ensemble de générateurs de Tausworthe combinés, mais en utilisant le critère $\Delta_{32,32,32}$. Notons que pour être exact, on devrait peut-être plutôt dire que l'on a utilisé le critère $\Delta_{m,32,32}$, car lorsque $w_1 > m$, alors $\ell_j^* - \ell_j = 0 - 0 = 0$ pour tout $m < j \le w_1$ et donc, $\delta_{j,1} = \delta_{m,1}$ pour tout $j > m$. Les résultats de ces recherches sont donnés au tableau 4.3, qui a la même structure que le tableau 4.2.

TABLEAU 4.3: Meilleurs générateurs combinés avec leur $\delta_{32,u}$

m	(m_j, q_j, v_j)	$\delta_{32,1}$	$\delta_{32,2}$	$\delta_{32,3}$	$\Delta_{32,32,32}$
10	(7,1,3) (3,1,2)	0	2	2	2
12	(7,3,4) (5,2,2)	1	2	3	3
14	(7,1,4) (4,1,1) (3,1,1)	1	2	2	2
16	(9,4,1) (7,1,2)	2	2	2	2

4.6.2 Options asiatiques

Dans ce qui suit, nous étudions la performance des règles de réseau polynômiales données à la sous-section précédente sur le problème des options asiatiques, qui a été expliqué à la sous-section 2.5.2.

Dans les tableaux 4.4 et 4.5, nous donnons les facteurs de réduction de variance obtenus par les règles qui se trouvent au tableau 4.2 (PLR). Le tableau 4.5 contient les résultats pour l'estimateur "ACV", qui est celui utilisant des variables antithétiques et une variable de contrôle, soit le prix de l'option sur la moyenne géométrique. Sans ces deux techniques de réduction de variance, on a l'estimateur "naïf" et les résultats associés se trouvent au tableau 4.4. La variance des estimateurs formés à partir de ces règles est estimée à l'aide de 100 XOR-translations indépendantes. L'estimateur MC correspondant est basé sur $100N$ répétitions i.i.d., afin que la comparaison soit juste. Les paramètres de l'option sont $S(0) = 100$, $r = \ln 1.09$, $\sigma = 0.2$, $T = 120$ jours et $T_1 = T - s$ jours. Le paramètre L vaut 32 dans la définition de la fonction de sortie (4.2). Les estimateurs basés sur des règles de réseau polynômiales sont plus rapides à calculer que ceux provenant de la méthode MC : en fait, le temps de calcul est environ le même que celui requis dans le cas des règles standard.

Nous donnons aussi les résultats obtenus en utilisant une règle de réseau standard (LR) choisie à l'aide du critère $M_{10,10,10}$. Les deux types de règles sont ainsi choisies de façon équivalente, c.-à-d., en considérant les mêmes projections. On utilise 100

translations aléatoires afin d'estimer la variance de l'estimateur.

TABLEAU 4.4: Facteurs de réduction de variance, estimateur naïf

s	Méth.	N	$K = 90$	$K = 100$	$K = 110$
10	PLR	1024	3400	670	810
	LR	1021(143)	200	130	69
	PLR	4096	3200	1600	730
	LR	4093(465)	610	470	200
	PLR	16384	23000	5100	4400
	LR	16381(357)	2460	1400	600
	PLR	65536	40000	6600	1300
	LR	65521(4398)	4030	2280	750
60	PLR	1024	260	52	14
	LR	1021(143)	89	23	10
	PLR	4096	200	88	7.4
	LR	4093(465)	32	13	2.6
	PLR	16384	840	105	48
	LR	16381(357)	67	22	6.1
	PLR	65536	360	200	34
	LR	65521(4398)	120	43	11

Pour ce problème, les règles de réseau polynômiales réduisent la variance par des facteurs qui varient entre 2 et 40000, en comparaison avec MC. Comme prévu, les facteurs de réduction augmentent habituellement avec N et diminuent avec s. L'amélioration par rapport à MC est plus importante avec l'estimateur naïf qu'avec l'estimateur ACV et c'est le cas aussi pour les règles de réseau standard, tel que mentionné dans [82, 84]. De plus, les facteurs de réduction de variance amenés par les règles de réseau polynômiales diminuent avec K et la raison pour cela est similaire à celle prévalant dans le cas standard : la bonne équidistribution des règles de réseau polynômiales n'est pas très utile lorsque K est grand puisque dans ce cas, la fonction f que l'on intègre vaut 0 sur la plupart du domaine $[0,1)^s$.

Notons que la plupart des règles utilisées ne sont pas ME (voir au tableau 4.2). Par exemple, pour $m = 16$, nous avons constaté que parmi les générateurs ME considérés dans notre recherche, le meilleur par rapport à $\Delta_{10,10,10}$ n'obtenait pas mieux qu'une valeur de 3 pour ce critère et cette valeur était causée par une mauvaise projection en dimension 3 (c.-à-d., on avait $\delta_{10,3} = 3$). Ce générateur s'est avéré être plutôt mauvais

TABLEAU 4.5: Facteurs de réduction de variance, estimateur ACV

s	Méth.	N	$K = 90$	$K = 100$	$K = 110$
10	PLR	1024	22	12	3.5
	LR	1021(143)	5.0	3.8	3.0
	PLR	4096	64	22	7.4
	LR	4093(465)	9.4	7.5	4.9
	PLR	16384	89	36	9.0
	LR	16381(357)	22	12	10
	PLR	65536	114	32	13
	LR	65521(4398)	19	10	5.0
60	PLR	1024	9.0	6.0	2.4
	LR	1021(143)	2.8	2.5	1.4
	PLR	4096	16	8.4	2.7
	LR	4093(465)	0.8	0.9	1.5
	PLR	16384	16	14	3.6
	LR	16381(357)	5.5	2.8	2.1
	PLR	65536	34	9.5	3.6
	LR	65521(4398)	4.6	2.3	2.3

TABLEAU 4.6: Règle de réseau polynômiale ME, $m = 16$

Méth.	s	$K = 90$	$K = 100$	$K = 110$
		estimateur naïf		
PLR (ME)	10	15800	810	1500
PLR	10	40000	6600	1300
PLR (ME)	60	650	78	43
PLR	60	360	200	34
		estimateur ACV		
PLR (ME)	10	4.4	15	0.23
PLR	10	114	32	13
PLR (ME)	60	18	12	3.3
PLR	60	34	9.5	3.6

quand $s = 10$ pour le problème des options asiatiques, donnant des estimateurs ayant parfois une variance supérieure à celle de l'estimateur MC, tel qu'illustré au tableau 4.6. Celui qui se trouve au tableau 4.2 fournit des estimateurs ayant généralement une plus petite variance, surtout quand $s = 10$. Nous avons recopié ses facteurs dans le tableau 4.6, afin de faciliter la comparaison. Cela nous montre à nouveau qu'il est important de regarder les projections $P_N(I)$ sur des sous-ensembles I qui ne sont pas seulement à indices successifs pour ce type de problème.

En ce qui concerne la comparaison entre les deux types de règles, de façon générale, il semble que les règles de réseau polynômiales obtiennent de meilleurs résultats que les règles standard pour cet exemple. On peut même dire que dans plusieurs cas, les règles de réseau polynômiales font significativement mieux que les règles standard, par des facteurs allant jusqu'à 10. En dimension 60 et pour l'estimateur ACV, on retrouve un cas où la règle de réseau standard est légèrement moins bonne que l'estimateur MC. Trouvant cela curieux, nous avons vérifié la valeur de $M_{32,24,16,12}$ pour cette règle et avons trouvé qu'elle valait 0.05245 : cette petite valeur explique les mauvais résultats pour cette règle en dimension 60.

Pour donner une idée de la précision des facteurs donnés au tableau 4.4, les intervalles de confiance au niveau 98% pour le rapport des variances théoriques associées au cas où $N = 65536$ et $s = 10$ sont (28141, 54729), (4641, 9025) et (891, 1733) pour $\hat{\mu}_{\text{PLR}}$ et (2815, 5476), (1595, 3101) et (524, 1018) pour $\hat{\mu}_{\text{LR}}$, avec $N = 65521$. Les intervalles associés au tableau 4.5 pour les mêmes valeurs de N et s sont (80, 155), (22, 44) et (9.0, 18) pour $\hat{\mu}_{\text{PLR}}$ et (13, 25), (7.1, 14) et (3.5, 6.8) pour $\hat{\mu}_{\text{LR}}$.

Nous avons également testé les générateurs de Tausworthe donnés au tableau 4.3, dans le cas où $s = 60$, afin de voir si le fait de considérer plus de projections améliorait par beaucoup la qualité des estimateurs.

TABLEAU 4.7: Règles choisies avec $\Delta_{32,32,32}$, $s = 60$

Méth.	N	$K = 90$	$K = 100$	$K = 110$
		estimateur naïf		
PLR	1024	78	55	9.3
PLR	4096	200	120	10
PLR	16384	850	210	46
PLR	65536	1100	200	53
		estimateur ACV		
PLR	1024	16	8.0	2.2
PLR	4096	19	7.4	3.7
PLR	16384	25	13	3.1
PLR	65536	24	10	3.4

Les résultats du tableau 4.7 nous indiquent que l'utilisation du critère $\Delta_{32,32,32}$ améliore habituellement la qualité des estimateurs, mais pas de façon remarquable.

4.6.3 Fonction-test

Dans cette sous-section, nous comparons les règles de réseau polynômiales et standard considérées dans l'exemple précédent, mais en utilisant une fonction-test tirée de [111] et utilisée dans d'autres articles [123, 133]. Cette fonction est de la forme :

$$f(\mathbf{x}) = \frac{1}{\mu} \prod_{j=1}^{s} \frac{|4x_j - 2| + c_j}{1 + c_j}.$$

L'intégrale de cette fonction vaut 1 et selon la valeur des paramètres c_j, $j = 1, \ldots, s$, le poids des différentes composantes ANOVA varient et donc, la dimension effective varie. On considère les cas suivants [111] :

(1) $c_j = 0.01, j = 1, \ldots, s$;

(2) $c_j = 1, j = 1, \ldots, s$;

(3) $c_j = j, j = 1, \ldots, s$;

(4) $c_j = j^2, j = 1, \ldots, s$.

Ainsi, la dimension effective de f diminue en allant du cas 1 au cas 4. Pour cette raison, nous trouvons que ce genre de test est intéressant, car la structure de la fonction permet de facilement contrôler la dimension effective et de voir de façon précise comment nos estimateurs réagissent à ces divers changements.

Nous avons décidé de mener l'expérience comme dans [133]. Ainsi, plutôt que de calculer l'erreur relative (c'est ce qui est souvent fait avec les fonctions-tests), nous effectuons des randomisations sur les règles de réseau, comme dans l'exemple des options asiatiques, et donnons les facteurs de réduction de variance par rapport à MC.

On utilise 100 répétitions i.i.d. afin d'estimer la variance des estimateurs. Les facteurs sont donnés par rapport à la méthode MC qui utilise $100N$ répétitions i.i.d., où N est le nombre de points de la règle de réseau polynômiale. Nous utilisons les mêmes règles qu'à la sous-section précédente, que ce soit dans le cas standard ou polynômial. Les résultats donnés au tableau 4.8 sont pour les règles choisies avec les critères $\Delta_{10,10,10}$ et $M_{10,10,10}$.

Étant donné que la dimension effective de f diminue en allant du cas 1 au cas 4, on s'attend à ce que les facteurs de réduction de variance des méthodes QMC augmentent

TABLEAU 4.8: Fonction-test, règles choisies avec $\Delta_{10,10,10}$ et $M_{10,10,10}$

Méth.	c_j	N	$s=5$	$s=10$	$s=15$	$s=20$	$s=25$	$s=30$
LR	0.01	1021	7.77e2	2.15	1.22	0.67	0.57	0.65
PLR	0.01	1024	42.8	2.90	1.43	1.04	1.05	0.19
LR	1	1021	1.84e4	58.7	19.9	5.02	4.32	2.68
PLR	1	1024	3.41e3	68.4	24.3	8.46	4.60	2.79
LR	j	1021	1.04e5	4.69e3	3.04e3	9.86e2	1.06e3	1.13e3
PLR	j	1024	1.21e5	1.09e4	5.31e3	2.82e3	4.76e2	6.73e2
LR	j^2	1021	5.17e5	2.83e5	2.67e5	3.69e5	2.49e5	2.81e5
PLR	j^2	1024	1.75e5	1.61e5	1.68e5	2.03e5	8.09e4	1.49e5
LR	0.01	4093	18.3	9.77	2.75	1.64	2.12	0.35
PLR	0.01	4096	1.24e4	19.1	0.29	0.62	0.64	0.68
LR	1	4093	7.39e2	3.37e2	77.4	18.3	9.57	2.73
PLR	1	4096	1.86e5	2.79e2	0.60	0.48	0.47	0.33
LR	j	4093	3.54e4	3.32e4	2.65e4	2.85e4	1.70e4	2.05e4
PLR	j	4096	1.99e5	6.21e3	2.55e2	2.11e2	2.15e2	2.30e2
LR	j^2	4093	5.54e6	4.35e6	3.69e6	4.41e6	3.33e6	3.56e6
PLR	j^2	4096	1.81e5	1.27e5	1.26e5	1.25e5	1.27e5	1.01e5
LR	0.01	16381	3.49e4	11.0	3.39	1.76	6.50	2.72
PLR	0.01	16384	4.26e4	6.11	1.37	1.57	4.64	0.41
LR	1	16381	1.10e6	2.16e3	95.0	41.2	16.2	6.37
PLR	1	16384	3.45e6	3.38e2	32.4	15.6	11.9	5.59
LR	j	16381	2.11e7	5.49e6	3.30e3	5.40e3	5.02e3	3.82e3
PLR	j	16384	1.20e8	2.75e6	1.97e4	1.81e4	1.78e4	3.08e3
LR	j^2	16381	1.70e8	1.81e8	8.05e5	1.21e6	1.49e6	1.09e6
PLR	j^2	16384	3.28e8	2.75e8	8.90e6	1.30e7	1.59e7	3.37e6

en allant du cas 1 au cas 4. C'est bien ce qui se produit, que ce soit dans le cas standard ou polynômial. En fait, on peut voir que pour les types 1 et 2, les facteurs diminuent avec s de façon évidente, alors que pour le type 4, les facteurs restent à peu près constants quand s augmente.

Quand N est petit (1024) et que la dimension est de 20 ou plus, les deux types de règles font moins bien que l'estimateur MC pour la fonction de type 1. Plus précisément, les intervalles de confiance au niveau 98% pour le rapport théorique des variances dans le cas de la fonction de type 1 et lorsque $N = 1021$ sont (543, 1057), (1.50, 2.92), (0.85, 1.66), (0.47, 0.91), (0.40, 0.78) et (0.45, 0.88) pour LR et (30, 58), (2.03, 3.94), (1.00, 1.94), (0.73, 1.41), (0.73, 1.43) et (0.13, 0.26) pour PLR.

Ce problème surgit aussi pour la fonction de type 2, même quand on passe à $N = 4096$ points. Par contre, avec 16384 points, nos deux estimateurs font presque toujours mieux que MC, pour les quatre types de fonctions. Évidemment, en sélectionnant les règles à l'aide d'un critère qui mesure la qualité des projections en plus grande dimension que 10 (c.-à-d., en prenant $\Delta_{w_1,\ldots,w_d}$ et $M_{w_1,\ldots,w_d}$ avec les $w_j > 10$), on obtiendrait sans doute de meilleurs résultats, du moins dans le cas où $s > 10$. Afin de vérifier cela, nous avons refait l'expérience, mais avec les générateurs de Tausworthe donnés au tableau 4.3 et les GCL sélectionnés à l'aide du critère $M_{32,32,32}$, tels que donnés au tableau C.1.

TABLEAU 4.9: Fonction-test, règles choisies avec $\Delta_{32,32,32}$ et $M_{32,32,32}$

Méth.	c_j	N	$s = 5$	$s = 10$	$s = 15$	$s = 20$	$s = 25$	$s = 30$
LR	0.01	1021	2.76e3	6.22	0.70	0.57	0.73	0.08
PLR	0.01	1024	2.95e2	3.29	0.90	1.28	0.37	0.84
LR	1	1021	7.44e3	1.51e2	13.6	7.45	8.60	3.28
PLR	1	1024	1.85e3	31.7	8.22	6.16	2.74	2.89
LR	j	1021	2.76e4	2.63e3	1.80e3	1.17e3	1.81e3	1.46e3
PLR	j	1024	5.30e3	4.86e2	4.40e2	4.00e2	2.84e2	3.64e2
LR	j^2	1021	2.80e5	8.85e4	8.93e4	8.80e4	1.10e5	8.63e4
PLR	j^2	1024	7.13e3	4.13e3	7.02e3	7.63e3	5.91e3	6.62e3
LR	0.01	4093	3.50e4	13.8	3.65	1.46	0.79	0.33
PLR	0.01	4096	3.09e3	5.57	1.63	1.18	0.77	0.47
LR	1	4093	3.89e5	9.11e2	1.30e2	16.5	6.96	2.93
PLR	1	4096	2.26e5	2.22e2	9.99	5.44	3.37	1.85
LR	j	4093	6.39e6	1.43e5	7.39e4	1.46e4	7.97e3	8.72e3
PLR	j	4096	6.89e6	4.27e4	4.54e3	3.35e3	2.84e3	2.64e3
LR	j^2	4093	2.81e8	5.93e7	4.85e7	5.24e7	4.00e7	3.84e7
PLR	j^2	4096	7.32e7	3.42e7	1.78e7	1.67e7	1.37e7	1.13e7
LR	0.01	16381	5.02e5	12.2	4.26	1.31	2.41	0.36
PLR	0.01	16384	1.87e6	76.2	3.77	1.61	5.07	0.96
LR	1	16381	8.91e6	7.39e2	1.94e2	64.7	34.1	3.94
PLR	1	16384	1.79e6	3.90e3	1.83e2	26.9	17.1	7.31
LR	j	16381	4.83e7	9.40e4	5.70e4	8.42e4	8.08e4	2.84e4
PLR	j	16384	9.87e5	5.43e5	2.55e5	5.78e4	6.70e4	2.61e4
LR	j^2	16381	1.23e8	2.26e7	2.25e7	2.38e7	3.22e7	2.81e7
PLR	j^2	16384	7.94e5	8.94e5	5.98e5	7.67e5	1.03e6	7.50e5

On peut dire que le fait d'utiliser le critère $\Delta_{32,32,32}$ permet généralement d'obtenir de plus grands facteurs de réduction de variance par rapport à MC. En particulier, dans

le cas des règles de réseau polynômiales, lorsque $N = 4096$, l'utilisation de $\Delta_{32,32,32}$ au lieu de $\Delta_{10,10,10}$ fait en sorte que la variance est maintenant inférieure à celle de l'estimateur MC pour la fonction de type 1, quand $s = 15, 20$.

Les résultats que nous avons obtenus pour ces deux types de règles sont aussi bons (sinon meilleurs) que ceux rapportés dans [133], où l'on a utilisé différentes randomisations d'ensembles de points basés sur la suite de Halton [39]. Par exemple, en dimension 20, quand $N = 1024$, les facteurs obtenus par leur meilleure méthode sont de 1.58, 0.95, 41 et 1135 pour les types 1 à 4, respectivement et donc, ce n'est que pour le type 1 que leur méthode fait mieux que nos règles de réseau.

En ce qui concerne la différence entre les deux types de règles, il semble qu'il y ait plus de cas où les règles de réseau standard ont une plus petite variance. Cependant, nous ne croyons pas que cela devrait servir à tirer quelque conclusion qu'il soit au sujet de la supériorité des règles de type standard. En effet, on travaille ici avec des fonctions artificielles et donc, il se peut qu'*a priori*, la structure de f favorise une des deux constructions.

Chapitre 5

Règles de type ν^r-copie

Dans ce chapitre, nous considérons les règles de type ν^r-copie, telles que définies à la section 2.1.2. Ces règles constituent un cas particulier de règles de rang supérieur à 1 et sont étudiées dans [120, 23, 55, 54]. Un résumé des principaux résultats émanant de ces travaux est donné dans [116]. Le but de ce chapitre est de fournir des évidences théoriques et empiriques visant à montrer que ce type de règle devrait être utilisé avec beaucoup de prudence. Nous expliquons que l'aspect négatif de ces règles se trouve au niveau de leurs projections en basse dimension. Nous croyons que ces résultats sont importants, car ce type de règle est en quelque sorte recommandé par Sloan et Joe dans leur livre [116] : par exemple, les tableaux de règles fournis dans leur annexe A contiennent exclusivement des règles de type 2^s-copie. Ainsi, nous pensons que notre point de vue pourra permettre aux utilisateurs de pouvoir faire un choix plus éclairé quant au choix du type de règle à utiliser.

Nous rappelons d'abord brièvement à la section 5.1 les avantages de ces règles sur les règles de rang 1, tel que décrit dans [116]. Ensuite, nous regardons à la section 5.2 les projections de ces règles sur les sous-espaces de $[0,1)^s$ et voyons ainsi quels sont leurs désavantages. Nous montrons aussi que ces règles n'ont pas une aussi petite variance que les règles de rang 1 quand on les utilise pour intégrer des fonctions linéaires. Des tableaux comparant les règles de type ν^s-copie avec celles de Korobov selon divers critères de sélection illustrent ces différents aspects à la section 5.3. Finalement, nous comparons à la section 5.4 ces deux types de règles sur le problème des options

asiatiques et sur une fonction-test.

5.1 Rappels sur les avantages des règles de type ν^r-copie

Rappelons d'abord la forme de ces règles, telle que donnée à l'équation (2.3). Une règle ν^r-copie contient $N = n \cdot \nu^r$ points et est définie par

$$P_N = \bigcup_{m_1=0}^{\nu-1} \cdots \bigcup_{m_r=0}^{\nu-1} \bigcup_{i=1}^{n} \{((m_1/\nu, \ldots, m_r/\nu, \underbrace{0, \ldots, 0}_{s-r \text{ fois}}) + \mathbf{x}_i) \bmod 1\},$$

où l'ensemble $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ correspond aux n points de la règle de rang 1 qui est copiée. Pour se rafraîchir la mémoire, le lecteur peut retourner voir à la page 28 une illustration d'une règle de type 2^2-copie, avec $n = 5$ pour la règle de rang 1 qui est copiée. Dans [116], on recommande d'utiliser $\nu = 2$: nous verrons pourquoi au prochain paragraphe. Dans le cas où $r = s$, on remplace parfois le point $\mathbf{x}_i$ par $\mathbf{x}_i/\nu$ (dans [116, *Definition 6.1*], par exemple), mais il est démontré dans [55] que les deux représentations sont équivalentes. Tout au long de ce chapitre, nous supposons que $\text{pgcd}(n, \nu) = 1$ et cela nous garantit que la règle a bien $N = n\nu^r$ points distincts.

Les avantages de ces règles sur les règles de rang 1 sont reliés au critère P_α^s et peuvent se résumer ainsi : 1) pour un ordre N donné, le P_α^s moyen des règles de type ν^r-copie est plus petit que le P_α^s moyen des règles de rang 1 ; 2) pour un ordre N donné, le calcul du P_α^s est plus rapide pour les règles de type ν^r-copie que pour les règles de rang 1. Plus précisément, en ce qui concerne le premier point, on montre dans [23, 55] (voir aussi [116, proposition 6.8, théorèmes 6.9 et 7.5]) que le ratio du P_α^s moyen des règles ν^r-copie d'un ordre donné sur le P_α^s moyen des règles de rang 1 de même ordre est borné par une quantité λ^r, où λ est minimisé en prenant $\nu = 2$ et est inférieur à 1 pour cette valeur de ν. Même si cela ne tient que pour le P_α^s moyen, les résultats numériques donnés dans [120, 54, 116] semblent indiquer que cela tient aussi lorsque l'on compare la meilleure règle ν^r-copie par rapport à P_2^s avec la meilleure règle de rang 1 et ce, pour différents rangs r. Les résultats numériques de la sous-section 5.3.1 vont dans le même sens.

Le deuxième avantage vient du fait que pour les règles de type ν^r-copie, le P_α^s est

donné par (voir [116, *Theorem 7.3*] et [46, page 150]) :

$$-1+\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{r}\left[1-\frac{(-1)^{\alpha/2}(2\pi)^{\alpha}}{\nu^{\alpha}\alpha!}B_{\alpha}\left(\nu x_{ij}\right)\right]\prod_{j=r+1}^{s}\left[1-\frac{(-1)^{\alpha/2}(2\pi)^{\alpha}}{\alpha!}B_{\alpha}\left(x_{ij}\right)\right],$$

où $\{\mathbf{x}_i, i = 1,\ldots,n\} = \{((i-1)\mathbf{z}/n) \bmod 1, i = 1,\ldots,n\}$ est la règle de rang 1 qui est copiée. Autrement dit, le P_α^s d'une règle ν^r-copie correspond au $\tilde{P}_\alpha^s$ (défini en (3.24), page 112) d'une règle de rang 1 dont le vecteur générateur est $\tilde{\mathbf{z}} = (\nu z_1,\ldots,\nu z_r, z_{r+1},\ldots,z_s)$, en prenant les poids $\beta_1 = \ldots = \beta_r = 1/\nu$, $\beta_{r+1} = \ldots = \beta_s = 1$. Ainsi, le calcul du P_α^s se fait en un temps dans $O(ns) = O(Ns\nu^{-r})$, plutôt que dans $O(Ns)$, comme c'est le cas généralement, en particulier pour les règles de rang 1. Il n'est pas surprenant que ce soit ainsi, car les propriétés d'une règle de type ν^r-copie dépendent seulement de la règle de rang 1 qui est copiée et le reste est déterminé par la forme générale qu'a une telle règle. Il est donc normal que le calcul de P_α^s ne prenne pas plus de temps que le calcul du P_α^s de la règle de rang 1 qui est copiée.

Le revers de la médaille, comme nous le verrons dans les deux prochaines sections, c'est que si on utilise un autre critère que P_α^s, les choses changent complètement. En fait, n'importe quel critère qui accorde plus d'importance aux projections $P_N(I)$ sur des sous-ensembles I contenant moins de variables détectera facilement les défauts des règles de type ν^r-copie. Par exemple, dans [46], on montre qu'en remplaçant P_α^s par $\tilde{P}_\alpha^s$ avec des poids $\beta_1 = \ldots = \beta_s = \beta < 1$, si β est suffisamment petit, alors on a que le ν optimal mentionné précédemment est 1, c.-à-d., il est alors préférable de ne pas faire de copie. Nous verrons aux tableaux 5.3 et 5.4 une illustration numérique de ce phénomène et la section qui suit vise à expliquer plus en détail cet état de fait.

5.2 Projections des règles de type ν^r-copie

Le premier problème des règles de type ν^r-copie d'ordre N en ce qui concerne leurs projections, c'est qu'elles ne contiennent pas toujours N points distincts, même si la règle de rang 1 qui est copiée est complètement projection-régulière. En particulier, si $I = \{i_1,\ldots,i_t\} \subseteq \{1,\ldots,r\}$, alors $P_N(I)$ contient seulement N/ν^{r-t} points distincts. Par exemple, si $\nu = 2$, $N = 2^{10}$ et $r = s = 8$, alors les projections unidimensionnelles $P_N(\{j\})$ contiennent chacune $n\nu = 8$ points distincts qui sont répétés $2^7 = 128$ fois

chacun. Les projections bidimensionnelles contiennent chacune 16 points distincts, répétés 64 fois chacun. En fait, seule la projection $P_N(\{1,2,\ldots,10\})$ contient $N = 2^{10}$ points distincts.

Ce problème avait déjà été noté dans le cas des grilles rectangulaires à l'exemple 1.1.2. En fait, une grille rectangulaire correspond à une règle ν^s-copie avec $n = 1$ pour la règle de rang 1 qui est copiée, ce qui est la pire combinaison possible de ν et r par rapport au nombre de points contenus dans les projections. De façon plus générale, on a que :

Proposition 5.2.1 *Soit $I \subseteq S$ et posons $\eta_I(r) = \sum_{j\in I} \mathbf{1}_{\{j\leq r\}}$, le nombre d'indices dans I qui sont inférieurs ou égaux à r. Si P_N est une règle ν^r-copie et que la règle de rang 1 qui est copiée est complètement projection-régulière, alors on a que*

$$|P_N(I)| = N\nu^{\eta_I(r)-r}.$$

Démonstration : par définition d'une règle ν^r-copie, pour $I = \{j_1,\ldots,j_t\}$, on a que

$$P_N(I) = \bigcup_{m_1=0}^{\nu-1} \cdots \bigcup_{m_{\eta_I(r)}=0}^{\nu-1} \bigcup_{i=1}^{n} \left\{ \left(\left(\frac{m_1}{\nu}, \ldots, \frac{m_{\eta_I(r)}}{\nu}, \underbrace{0,\ldots,0}_{|I|-\eta_I(r) \text{ fois}} \right) + (x_{ij_1},\ldots,x_{ij_t}) \right) \bmod 1 \right\}.$$

Il suffit de vérifier que les $n\nu^{\eta_I(r)} = (N/\nu^r)\nu^{\eta_I(r)} = N/\nu^{r-\eta_I(r)}$ points de $P_N(I)$ sont distincts. Or, s'il existe deux points de $P_N(I)$ qui sont identiques, cela implique que pour chaque $j \in I$, l'égalité

$$\frac{m_j}{\nu} + x_{ij} = \frac{m_j + b_j}{\nu} + x_{lj} \qquad (\text{mod } 1)$$

tient avec soit $i = l$ et $0 < b_j \leq \nu - 1 - m_j$, soit avec $i \neq l$ et $0 \leq b_j \leq \nu - 1 - m_j$. Supposons que $x_{ij} = a_i/n$ et $x_{lj} = a_l/n$ ($a_i \neq a_l \bmod n$ si $i \neq l$, puisque l'on a supposé que la règle de rang 1 était complètement projection-régulière), avec $0 \leq a_i, a_l \leq n-1$. On doit alors avoir que

$$\nu(a_i - a_l) = nb_j \qquad (\text{mod } n\nu),$$

pour tout $j \in I$. Or, cette congruence a une solution si et seulement si $\text{pgcd}(\nu, n\nu) = \nu$ divise nb_j et ceci ne peut arriver que si $b_j = 0$, puisque $b_j \leq \nu - 1$ et $\text{pgcd}(n,\nu) = 1$.

Donc, on doit avoir $a_i = a_l$, ce qui implique que $i = l$, contredisant ainsi l'hypothèse de départ à savoir que les points de $P_N(I)$ ne sont pas distincts. ∎

De plus, excepté lorsque le rang r est égal à s, les règles de type ν^r-copie ne sont pas stationnaires dans la dimension. Cela est facilement compris lorsque l'on remarque que le nombre de points dans $P_N(I)$ dépend du premier indice dans I, par l'intermédiaire de $\eta_I(r)$.

Proposition 5.2.2 *Une règle ν^r-copie est stationnaire dans la dimension si et seulement si $r = s$ et la règle de rang 1 qui est copiée est stationnaire dans la dimension.*

Démonstration : si $r = s$, alors pour $I = \{j_1, \ldots, j_t\}$, on a que

$$P_N(I) = \bigcup_{m_{j_1}=0}^{\nu-1} \cdots \bigcup_{m_{j_t}=0}^{\nu-1} \bigcup_{i=1}^{n} \left\{ \left(\left(\frac{m_{j_1}}{\nu}, \ldots, \frac{m_{j_t}}{\nu} \right) + (x_{ij_1}, \ldots, x_{ij_t}) \right) \bmod 1 \right\}.$$

Si on a $J = \{k_1, \ldots, k_t\}$ tel que $\tilde{I} = \tilde{J}$ (rappelons que $\tilde{I} = \{1, j_2 - j_1 + 1, \ldots, j_t - j_1 + 1\}$), puisque

$$\left\{ \left(\frac{m_{j_1}}{\nu}, \ldots, \frac{m_{j_t}}{\nu} \right), m_{j_t} = 0, \ldots, \nu - 1, 1 \leq l \leq t \right\} =$$
$$\left\{ \left(\frac{m_{k_1}}{\nu}, \ldots, \frac{m_{k_t}}{\nu} \right), m_{k_t} = 0, \ldots, \nu - 1, 1 \leq l \leq t \right\},$$

on a $P_N(I) = P_N(J)$ si et seulement si

$$\{(x_{ij_1}, \ldots, x_{ij_t}), i = 1, \ldots, n\} = \{(x_{ik_1}, \ldots, x_{ik_t}), i = 1, \ldots, n\},$$

et cette égalité tient pour tout I, J tels que $\tilde{I} = \tilde{J}$ si et seulement si la règle de rang 1 qui est copiée est stationnaire dans la dimension.

Si $r < s$, alors on peut poser $J = \{s - r + 1, \ldots, s\}$ qui est tel que $\tilde{J} = \tilde{I}$, où $I = \{1, \ldots, r\}$, mais avec $\eta_J(r) < \eta_I(r)$: en effet, $\eta_I(r) = r$ et $\eta_J(r) < r$ puisque $|J| = r$ et l'indice $s > r$ est dans J. Or, par la proposition 5.2.1, on sait que $|P_N(I)| = N\nu^{\eta_I(r)-r} \neq N\nu^{\eta_J(r)-r} = |P_N(J)|$ et donc, on ne peut avoir $P_N(I) = P_N(J)$. ∎

Au chapitre 3, nous avons supposé à plusieurs reprises que les règles utilisées étaient stationnaires dans la dimension, de rang 1 et complètement projection-régulières. Ces deux dernières propriétés nous permettaient de supposer que les projections unidimensionnelles étaient toujours données par $\{0, 1/N, \ldots, (N-1)/N\}$. En fait, sans ces

propriétés, les résultats obtenus dans ce chapitre tomberaient ou devraient être affaiblis. De plus, si $1 < r < s$, le calcul des critères de la forme $M_{t_1,\ldots,t_d}$ doit être modifié et inclure le calcul de l_I pour des projections $P_N(I)$ pour lesquelles $i_1 > 1$, si on veut mesurer autant d'information que dans le cas où P_N est stationnaire dans la dimension. Ensuite, l'intégration des polynômes de degré 1 et 2 est beaucoup moins efficace avec les règles de type ν^r-copie, puisque les projections $P_N(I)$ ne contiennent pas toujours N points distincts. Ceci nous laisse envisager que ce type de règle n'obtiendra pas de très bons résultats lorsque la fonction à intégrer possède des composantes f_I importantes pour $|I|$ petit.

Avant d'énoncer les résultats concernant l'intégration des polynômes, nous avons besoin du résultat intermédiaire suivant. C'est un lemme qui exprime $P^r_{\alpha_I}(I,N)$, la valeur de $P_{\alpha_I}(I)$ (voir la définition 3.4.1, page 92) pour une règle ν^r-copie d'ordre N, en fonction de $P^1_{\alpha_I}(I,n)$, la valeur de $P_{\alpha_I}(I)$ de la règle de rang 1 basée sur le vecteur générateur $\tilde{\mathbf{z}} = (\nu z_1, \ldots, \nu z_r, z_{r+1}, \ldots, z_s)$, où $\mathbf{z}$ est le vecteur générant la règle de rang 1 qui est copiée. La signification de $\mathbf{z}$ et de $\tilde{\mathbf{z}}$ sera fixée de cette façon à partir d'ici.

Lemme 5.2.1 *Soit P_N une règle ν^r-copie d'ordre $N = n\nu^r$. Soit $I \subseteq S$ et posons $\eta = \eta_I(r)$. Alors,*

$$P^r_{\alpha_I}(I,N) = \begin{cases} \nu^{-(\alpha_{i_1}+\cdots+\alpha_{i_\eta})} P^1_{\alpha_I}(I,n) & \text{si } \eta \geq 1, \\ P^1_{\alpha_I}(I,n) & \text{si } \eta = 0. \end{cases}$$

Démonstration : par définition, on a que

$$P^r_{\alpha_I}(I,N) = \sum_{\mathbf{h} \in L^\perp \cap \mathbf{Z}_I^*} \prod_{j \in I} h_j^{-\alpha_j}.$$

Dénotons par $L_n^\perp$ le réseau dual associé à la règle de rang 1 qui est copiée et par $\tilde{L}_n^\perp$, celui associé à la règle de rang 1 générée par $\tilde{\mathbf{z}}$. Par définition de la règle ν^r-copie, si on suppose que $I = \{i_1, \ldots, i_t\}$, alors $L^\perp = \{(\nu h_{i_1}, \ldots, \nu h_{i_\eta}, h_{i_{\eta+1}}, \ldots, h_{i_t}) : \mathbf{h} \in \tilde{L}_n^\perp\}$. En effet, $\mathbf{h} \in L^\perp$ si et seulement si $\sum_{i=1}^N e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i} = N$, où $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ est la règle ν^r-copie. Mais

$$\sum_{i=1}^{N} e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{x}_i} = \sum_{i=1}^{n} e^{(i2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{z})/n} \prod_{j=1}^{r} \sum_{m_j=0}^{\nu-1} e^{2\pi\sqrt{-1}h_j m_j/\nu}$$

et donc, $\mathbf{h} \in L^\perp$ si et seulement si $h_j = 0 \pmod{\nu}$ pour $j = 1, \ldots, r$ et $\mathbf{h} \in L_n^\perp$. Or,

$$\{\mathbf{h} : \mathbf{h} \in L_n^\perp, h_j = 0 \pmod{\nu}, j = 1, \ldots, r\} = \{(\nu h_1, \ldots, \nu h_r, h_{r+1}, \ldots, h_s) : \mathbf{h} \in \tilde{L}_n^\perp\}.$$

En effet, si $\mathbf{h} \in L_n^\perp$, alors $\mathbf{h} \cdot \mathbf{z} = 0 \pmod{n}$ et donc, $(h_1/\nu, \ldots, h_r/\nu, h_{r+1}, \ldots, h_s) \in \tilde{L}_n^\perp$, par définition de $\tilde{\mathbf{z}}$ et par hypothèse que $h_j = 0 \pmod{\nu}$, $j = 1, \ldots, r$. Dans l'autre sens, si $\mathbf{h} \in \tilde{L}_n^\perp$, alors $\mathbf{h} \cdot \tilde{\mathbf{z}} = 0 \pmod{n}$ et donc, $(\nu h_1, \ldots, \nu h_r, h_{r+1}, \ldots, h_s) \cdot \mathbf{z} = 0 \pmod{n}$, ce qui implique que $(\nu h_1, \ldots, \nu h_r, h_{r+1}, \ldots, h_s) \in L_n^\perp$, avec $\nu h_j = 0 \pmod{\nu}$ pour tout $j = 1, \ldots, r$.

Ainsi,

$$\begin{aligned} P_{\alpha_l}^r(I, N) &= \sum_{\mathbf{h} \in \tilde{L}_n^\perp \cap \mathbf{Z}_l^*} \prod_{j \in I, j \le r} (\nu h_j)^{-\alpha_j} \prod_{j \in I, j > r} h_j^{-\alpha_j} \\ &= \nu^{-(\alpha_{i_1} + \ldots + \alpha_{i_\eta})} P_{\alpha_l}^1(I, n), \end{aligned}$$

qui vaut bien $P_{\alpha_l}^1(I, n)$ dans le cas où $\eta = 0$. ∎

Avec cette expression pour $P_{\alpha_l}^r(I, N)$ qui est donnée en fonction de la règle de rang 1 qui est générée par $\tilde{\mathbf{z}}$, on peut utiliser les résultats de la section 3.4 afin de calculer la variance de l'estimateur associé à la règle de type ν^r-copie et la comparer avec la variance de l'estimateur MC, dans le cas d'un polynôme de degré 1 (fonction linéaire). C'est ce que nous faisons dans la proposition qui suit. Plus précisément, on donne une borne inférieure et une borne supérieure pour la variance de l'estimateur associé à la règle de type ν^r-copie qui sont toutes deux de la forme $g(n, \nu, r) \cdot \mathrm{Var}(\hat{\mu}_{MC})$, où $g(n, \nu, r)$ est un facteur dépendant de n, ν et r. Puisque, pour les deux bornes, ce facteur qui multiplie $\mathrm{Var}(\hat{\mu}_{MC})$ est supérieur à $1/N$ (le facteur que l'on retrouve dans le cas des règles de rang 1), ce résultat nous indique que ce type de règle ne réduit pas autant la variance que les règles de rang 1 de même ordre.

Dans ce qui suit, on dénote par $\hat{\mu}_{\mathrm{LR},r,\nu,n}$ l'estimateur obtenu en translatant aléatoirement (modulo 1) une règle de type ν^r-copie d'ordre $N = n\nu^r$.

Proposition 5.2.3 *Soit P_N une règle ν^r-copie d'ordre $N = n\nu^r$ pour laquelle la règle de rang 1 qui est copiée est complètement projection-régulière. Soit f un polynôme de degré 1. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC obtenu en utilisant N points. Alors*

$$\frac{\nu^{r-2}}{n}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) \le \mathrm{Var}(\hat{\mu}_{\mathrm{LR},r,\nu,n}) \le \frac{\nu^r}{n}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$$

et la borne inférieure est atteinte lorsque $r = s$.

Démonstration : d'abord, notons que $\text{pgcd}(\nu z_j, n) = 1$ si $\text{pgcd}(\nu, n) = 1$ et $\text{pgcd}(z_j, n) = 1$ et donc, la règle générée par $\tilde{\mathbf{z}}$ est complètement projection-régulière si celle basée sur $\mathbf{z}$ l'est.

En utilisant la notation $\sigma^2_{I,\text{LR}}$ de la proposition 3.3.1, page 88, on a que

$$\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) = \sum_{j=1}^{s} \sigma^2_{\{j\},\text{LR},r,\nu,n}$$

et par la proposition 3.4.7, on obtien

$$\sigma^2_{\{j\},\text{LR},r,\nu,n} = \gamma_2(\{j\}) P_2^r(\{j\}, N) = c^2(\mathbf{e}_j)(2\pi)^{-2} P_2^r(\{j\}, N).$$

Or, en utilisant les lemmes 5.2.1 et 3.4.4 (le dernier sert à calculer $P_2^1(\{j\}, n)$ et s'applique car la règle générée par $\tilde{\mathbf{z}}$ est complètement projection-régulière), on obtient que

$$P_2^r(\{j\}, N) = \begin{cases} P_2^1(\{j\}, n)/\nu^2 = 2\zeta(2)/(\nu n)^2 & \text{si } j \le r, \\ P_2^1(\{j\}, n) = 2\zeta(2)/n^2 & \text{sinon,} \end{cases}$$

car $j \le r$ implique que $\eta_{\{j\}}(r) = 1$ et si $j > r$, alors $\eta_{\{j\}}(r) = 0$. Ainsi,

$$\sigma^2_{\{j\},\text{LR},r,\nu,n} = \begin{cases} \sigma^2_{\{j\}}/(\nu n)^2 & \text{si } j \le r, \\ \sigma^2_{\{j\}}/n^2 & \text{sinon.} \end{cases}$$

Puisque

$$\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) = \sum_{j=1}^{s} \sigma^2_{\{j\},\text{LR},r,\nu,n} = \sum_{j=1}^{r} \sigma^2_{\{j\}}/(\nu n)^2 + \sum_{j=r+1}^{s} \sigma^2_{\{j\}}/n^2,$$

et que $\nu \ge 1$, cela signifie que

$$\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) \le \frac{1}{n^2} \sum_{j=1}^{s} \sigma^2_{\{j\}} = \frac{N}{n^2} \text{Var}(\hat{\mu}_{\text{MC}}) = \frac{\nu^r}{n} \text{Var}(\hat{\mu}_{\text{MC}})$$

et que

$$\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) \ge \frac{1}{(\nu n)^2} \sum_{j=1}^{s} \sigma^2_{\{j\}} = \frac{N}{(\nu n)^2} \text{Var}(\hat{\mu}_{\text{MC}}) = \frac{\nu^{r-2}}{n} \text{Var}(\hat{\mu}_{\text{MC}}).$$

Dans le cas où $r = s$, $\sigma^2_{\{j\},\text{LR},r,\nu,n} = \sigma^2_{\{j\}}/(\nu n)^2$ pour tout $j = 1, \ldots, s$ et donc $\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) = (N/(\nu n)^2)\text{Var}(\hat{\mu}_{\text{MC}})$. ∎

Le facteur devant $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ est entre ν^{r-2}/n et ν^r/n, au lieu d'être $1/N$, comme c'était le cas pour les règles de rang 1 complètement projection-régulières. Si on dénote par $\hat{\mu}_{\mathrm{LR},1,N}$ l'estimateur obtenu à partir d'une règle de rang 1 et d'ordre N, alors on a le résultat suivant, qui compare la variance de $\hat{\mu}_{\mathrm{LR},r,\nu,n}$ avec celle de $\hat{\mu}_{\mathrm{LR},1,N}$:

Corollaire 5.2.1 *Soit une règle de type ν^r-copie d'ordre $N = n\nu^r$ pour laquelle la règle de rang 1 qui est copiée est complètement projection-régulière. Pour toute règle de rang 1 d'ordre N qui est complètement projection-régulière, on a que*

$$\nu^{2(r-1)} \leq \frac{\mathrm{Var}(\hat{\mu}_{\mathrm{LR},r,\nu,n})}{\mathrm{Var}(\hat{\mu}_{\mathrm{LR},1,N})} \leq \nu^{2r}.$$

Démonstration : le corollaire 3.4.3 s'applique puisqu'on a une règle de rang 1 complètement projection-régulière et donc,

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR},1,N}) = \frac{1}{N}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$$

et la proposition 5.2.3 nous indique que

$$\frac{N\nu^{r-2}}{n} \leq \frac{\mathrm{Var}(\hat{\mu}_{\mathrm{LR},r,\nu,n})}{\mathrm{Var}(\hat{\mu}_{\mathrm{LR},1,N})} \leq \frac{N\nu^r}{n}.$$

Le résultat suit en remplaçant N par $n\nu^r$. ∎

Autrement dit, le rapport entre les deux variances croît exponentiellement avec r. Les règles de type ν^r-copie ne sont donc pas recommandées pour intégrer ce type de fonction.

Le résultat suivant donne des conditions suffisantes pour qu'une règle ν^r-copie réduise la variance par rapport à l'estimateur MC, lorsque l'on intègre un polynôme de degré 2. Dans le cas des règles de rang 1, le résultat équivalent se trouve au corollaire 3.4.4 et nous allons voir qu'elles étaient moins fortes dans ce cas.

Proposition 5.2.4 *Soit P_N une règle ν^r-copie d'ordre $N = n\nu^r$ pour laquelle la règle de rang 1 qui est copiée est complètement projection-régulière. Soit f un polynôme de degré 2. Soit $\hat{\mu}_{\mathrm{MC}}$ l'estimateur MC obtenu en utilisant N points.*

Si $n \geq \nu^r$ et

$$P_{2,2}^1(\{i,j\},n) \leq \begin{cases} \nu^{-r}\pi^4/9n & \text{si } i,j > r \\ \nu^{2\eta(i,j)(r)-r}\pi^4/9n & \text{sinon,} \end{cases}$$

alors

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : suivant la démonstration de la proposition 3.4.4, on doit montrer que

$$P_\alpha^r(\{j\}, N) \leq \frac{2\zeta(\alpha)}{N} \tag{5.1}$$

pour $\alpha = 2, 4$, $j = 1, \ldots, s$ et que

$$P_{2,2}^r(\{i,j\}, N) \leq \frac{(2\zeta(2))^2}{N}$$

pour tout $1 \leq i < j \leq s$. Si $j > r$, alors pour $\alpha = 2, 4$, on a

$$P_\alpha^r(\{j\}, N) = P_\alpha^1(\{j\}, n) = \frac{2\zeta(\alpha)}{n^\alpha} = \frac{\nu^r}{n^{\alpha-1}} \frac{2\zeta(\alpha)}{N} \leq \frac{2\zeta(\alpha)}{N},$$

la deuxième égalité vient du lemme 3.4.4, qui s'applique puisque la règle générée par $\tilde{\mathbf{z}}$ est complètement projection-régulière (comme nous l'avons vu dans la démonstration de la proposition 5.2.3) et l'inégalité suit par hypothèse que $n \geq \nu^r$ et que $\alpha \geq 2$. Le fait que $P_\alpha^r(\{j\}, N) = \nu^{-\alpha} P_\alpha^1(\{j\}, n) \leq P_\alpha^1(\{j\}, n)$ pour $1 \leq j \leq r$ nous assure que la condition (5.1) est respectée pour tout j. Ensuite, puisque $\eta_{\{i,j\}}(r) = 0$ équivaut à avoir $i, j > r$, on a que

$$P_{2,2}^r(\{i,j\}, N) = \begin{cases} P_{2,2}^1(\{i,j\}, n) & \text{si } i, j > r \\ \nu^{-2\eta_{\{i,j\}}(r)} P_{2,2}^1(\{i,j\}, n) & \text{sinon,} \end{cases}$$

par le lemme 5.2.1 et l'hypothèse de départ sur $P_{2,2}^1(\{i,j\}, n)$ nous assure que

$$P_{2,2}^r(\{i,j\}, N) \leq \pi^4/9N = (2\zeta(2))^2/N,$$

puisque $N = n\nu^r$. ∎

Rappelons que dans le cas des règles de rang 1 et d'ordre N, la condition sur $P_{2,2}^1(\{i,j\}, N)$ était de demander qu'il soit inférieur à $\pi^4/9N$: pour les règles ν^r-copie, ce facteur est multiplié par $\nu^{-r} < 1$ si $i, j > r$ et par $\nu^{2\eta_{\{i,j\}}(r)-r}$, sinon. Donc, la règle de rang 1 générée par $\tilde{\mathbf{z}}$ doit satisfaire des conditions plus fortes afin que l'on réduise la variance par rapport à MC que celles qui sont exigées lorsque l'on construit directement l'estimateur à partir d'une règle de rang 1.

Pour terminer cette section, nous donnons un résultat qui est l'équivalent du corollaire 3.4.1, mais dans le cas particulier où la règle est utilisée est de type ν^r-copie. Ce résultat nous indique que pour un polynôme de degré d, la variance de l'estimateur obtenu en utilisant une règle de type ν^r-copie est bornée par une certaine quantité dépendant de la fonction (la même que dans le cas général) multipliée par le $P_2^{\text{sup}}(d)$ de la règle de rang 1 générée par $\tilde{\mathbf{z}}$. Puisque $P_2^{\text{sup}}(d,n)$ diminue avec n, cela veut dire que pour un polynôme donné, les règles de type ν^r-copie ont une borne supérieure plus grande sur leur variance que les règles de rang 1 d'ordre équivalent.

Corollaire 5.2.2 *Soit f un polynôme de degré d. Pour une règle de type ν^r-copie d'ordre $N = n\nu^r$, on a que*

$$\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) \leq P_2^{\text{sup}}(d,n) \max_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \right)$$

où $P_2^{\text{sup}}(d,n)$ est le $P_2^{\text{sup}}(d)$ de la règle de rang 1 générée par $\tilde{\mathbf{z}}$.

Démonstration : par le corollaire 3.4.1, on a que

$$\text{Var}(\hat{\mu}_{\text{LR},r,\nu,n}) \leq P_2^{\text{sup}}(d,N) \max_{\emptyset \neq I \subseteq S, |I| \leq d} \left(\sum_{\alpha_I \in A_{I,d}} |\gamma_{\alpha_I}(I)| \right)$$

où $P_2^{\text{sup}}(d,N)$ est le $P_2^{\text{sup}}(d)$ de la règle de type ν^r-copie. Or,

$$P_2^{\text{sup}}(d,N) = \sum_{\emptyset \neq I \subseteq S, |I| \leq d} P^r_{\{2,\ldots,2\}}(I,N) \leq \sum_{\emptyset \neq I \subseteq S, |I| \leq d} P^1_{\{2,\ldots,2\}}(I,n) = P_2^{\text{sup}}(d,n),$$

l'inégalité suivant par le lemme 5.2.1. ∎

5.3 Performance par rapport à différents critères de sélection

Dans cette section, nous donnons les résultats des recherches que nous avons faites pour trouver les meilleures règles, soit de type 2^s-copie, soit de type Korobov, par rapport à différents critères. Cette étude empirique nous permet de voir quantitativement quelle est l'importance des défauts mentionnés à la section précédente. Nous présentons d'abord les résultats où le critère P_α^s a été utilisé, puis à la sous-section 5.3.2, nous refaisons les mêmes recherches, mais avec le critère $\tilde{P}_\alpha^s$. Finalement, à la sous-section 5.3.3, nous regardons plus en détail comment les règles de type ν^r-copie performent par rapport au critère $M_{s,s,s}$.

5.3.1 Utilisation du critère P_α^s

Dans les tableaux 5.1 et 5.2, nous comparons les meilleures règles (par rapport à P_2^s) de type 2^s-copie pour $s = 6, 12$ qui sont données dans [116, *Tables A.6, A.11*] avec les meilleures règles de Korobov d'ordre correspondant que nous avons trouvées en utilisant les critères P_2^s, $S_s = l_s/l_s^*$, M_s et $M_{s,s,s}$. Pour chaque règle, nous donnons le nombre total de points N, la valeur de a qui définit la règle (celle de rang 1 qui est copiée, dans le cas des règles de type 2^s-copie) et la valeur de chaque critère. Dans chaque colonne, les étoiles (*) indiquent la meilleure valeur obtenue pour le critère associé à la colonne. Nous voulons rappeler que parmi les critères considérés, il n'y a que P_2^s que l'on veut minimiser : les critères basés sur le test spectral doivent au contraire être maximisés et sont toujours entre 0 et 1. Le fait de considérer une règle 2^s-copie plutôt qu'une règle 2^r-copie avec $r < s$ nous permet de travailler avec des règles qui sont toutes stationnaires dans la dimension et ainsi, nous n'avons pas à modifier la façon de calculer le critère $M_{t_1,\ldots,t_d}$ afin de considérer les projections $P_N(I)$ pour lesquelles $i_1 > 1$.

Pour calculer $l_I/l_{|I|}^*(N)$, nous utilisons le fait que pour une règle ν^s-copie, $l_I = \nu l_I(n)$ et $l_{|I|}^*(N) = c_{|I|}(n\nu^s)^{1/|I|} = \nu^{s/|I|} l_{|I|}^*(n)$, où $l_{|I|}^*(n)$ et $l_I(n)$ sont les valeurs de $l_{|I|}^*$ et l_I pour la règle de rang 1 qui est copiée. Cela nous donne que $l_I/l_{|I|}^*(N) = \nu^{1-s/|I|} l_I(n)/l_{|I|}^*(n)$. Autrement dit, la qualité de la projection $P_N(I)$ de la règle ν^s-copie est donnée par la qualité de la projection $P_n(I)$ de la règle copiée, que l'on multiplie par un facteur d'ajustement $\nu^{1-s/|I|}$. Or, ce facteur d'ajustement diminue lorsque s augmente ou que $|I|$ diminue. Donc, on peut s'attendre à ce que plus s soit grand, plus les projections sur des sous-ensembles I contenant peu d'indices soient mauvaises. Par exemple, pour $s = 10$, même si la règle de rang 1 qui est copiée a des projections bidimensionnelles satisfaisant $\ell_{\{i,j\}}(n)/\ell_2^*(n) = 0.96$ (ce qui est très bon), les projections de la règle 2^{10}-copie ont une valeur de $\ell_{\{i,j\}}/\ell_2^*(N)$ égale à $2^{-4} \times 0.96 = 0.06 \ll 1$.

On aurait pu normaliser l_I en utilisant $\nu l_I^*(n)$: ceci correspond à normaliser l_I en utilisant le meilleure valeur possible que l'on peut obtenir pour l_I parmi toutes les règles de type ν^s-copie qui sont basées sur n'importe quelle sorte de réseau (et non pas

nécessairement un réseau d'intégration). Étant donné que nous comparons les règles de type ν^s-copie avec des règles de rang 1, nous avons choisi de normaliser l_I en utilisant plutôt la meilleure valeur possible ($l_I^*(N)$) que l'on peut obtenir pour une règle d'ordre N basée sur n'importe quelle sorte de réseau, afin que les deux types de règles soient normalisés de la même façon. En général, si $s > |I|+1$, alors $\nu l_I^*(n)$ est beaucoup plus petite que $l_I^*(N)$, ce qui signifie que nous avons choisie une normalisation plus sévère pour les règles de type ν^s-copie.

TABLEAU 5.1: Règles 2^s-copie et règles de Korobov, $s = 6$

rang	critère	n	N	a	P_2^s	S_6	M_6	$M_{6,6,6}$
6	P_2^s	79	5056	27	0.279*	0.7482	0.0944	0.0944
1	P_2^s	5051	5051	2486	0.671	0.7713	0.4784	0.2983
1	S_6	5051	5051	4916	0.738	0.9163*	0.4571	0.2380
1	M_6	5051	5051	4413	0.721	0.7245	0.7019*	0.2907
1	$M_{6,6,6}$	5051	5051	2254	0.902	0.7483	0.4987	0.4987*
6	P_2^s	157	10048	18	0.126*	0.7460	0.1912	0.1497
1	P_2^s	10039	10039	4578	0.256	0.7646	0.2408	0.0542
1	S_6	10039	10039	9617	0.378	0.8669*	0.4677	0.2831
1	M_6	10039	10039	9844	0.408	0.7272	0.7272*	0.4486
1	$M_{6,6,6}$	10039	10039	2304	0.534	0.5780	0.5669	0.5669*
6	P_2^s	313	20232	80	0.051*	0.6650	0.1060	0.1060
1	P_2^s	20229	20229	13912	0.109	0.7435	0.2520	0.2022
1	S_6	20229	20229	19193	0.113	0.8656*	0.4358	0.3032
1	M_6	20229	20229	13217	0.135	0.7435	0.7435*	0.1885
1	$M_{6,6,6}$	20229	20229	5363	0.193	0.5564	0.5564	0.5564*

Nos résultats concordent avec ceux de [116] : les règles de type 2^s-copie ont des plus petites valeurs de P_α^s que les meilleures règles de Korobov. Des expériences similaires faites sur $s = 4, 8$ nous ont permis de constater que plus s est grand, plus le rapport du P_2^s de la meilleure règle de Korobov avec celui de la meilleure 2^s-copie est grand : il vaut environ 1.5 lorsque $s = 4$, 2.1 lorsque $s = 6$, 3.2 lorsque $s = 8$ et 6.5 pour $s = 12$. Cette constatation n'est pas surprenante si on se rappelle que pour le P_α^s moyen, l'inverse de ce rapport est borné supérieurement par λ^s, où $\lambda < 1$ (voir à la section 5.1). Les meilleures règles de type 2^s-copie ont également de très bonnes valeurs pour S_s. En fait, quand $s = 12$, cette valeur est parfois égale à celle de la meilleure règle de Korobov trouvée par rapport à ce critère. Par contre, les règles de type 2^s-copie

TABLEAU 5.2: Règles 2^s-copie et règles de Korobov, $s = 12$

rang	critère	n	N	a	P_2^s	S_{12}	M_{12}	$M_{12,12,12}$
12	P_2^s	3	12288	1	447*	0.8097*	0.0237	0.0237
1	P_2^s	12281	12281	3636	2930	0.6401	0.0863	0.0187
1	S_{12}	12281	12281	12201	3180	0.8097*	0.4924	0.1158
1	M_{12}	12281	12281	9948	3160	0.7012	0.6683*	0.1202
1	$M_{12,12,12}$	12281	12281	4076	3160	0.7013	0.4458	0.3581*
12	P_2^s	5	20480	2	268*	0.7759	0.0291	0.0184
1	P_2^s	20479	20479	11077	1730	0.6134	0.0728	0.0145
1	S_{12}	20479	20479	18860	1890	0.8230*	0.4938	0.1080
1	M_{12}	20479	20479	14700	1900	0.7258	0.6915*	0.1927
1	$M_{12,12,12}$	20479	20479	1753	1900	0.5487	0.5013	0.3507*
12	P_2^s	11	45056	3	121*	0.7266	0.0277	0.0124
1	P_2^s	45053	45053	4928	806	0.6293	0.2334	0.0613
1	S_{12}	45053	45053	39426	866	0.8124*	0.3541	0.1798
1	M_{12}	45053	45053	26149	853	0.7266	0.6874*	0.1053
1	$M_{12,12,12}$	45053	45053	661	859	0.7266	0.4293	0.3967*

ne font pas très bien par rapport aux critères M_s et $M_{s,s,s}$, comme prévu et c'est pire pour $s = 12$ comparativement au cas où $s = 6$. Une autre remarque intéressante est qu'en dimension 12, lorsque l'on compare les meilleures règles de Korobov par rapport à P_2^s avec les meilleures par rapport à S_{12}, ces dernières ont généralement de meilleurs résultats par rapport aux critères M_{12} et $M_{12,12,12}$.

5.3.2 Utilisation du critère $\tilde{P}_\alpha^s$

Nous avons également effectué des recherches pour comparer les meilleures règles de Korobov avec les meilleures de type 2^s-copie, mais par rapport au critère $\tilde{P}_2^s$ avec des poids $\beta_1 = \ldots = \beta_s = \sqrt{3/8\pi^2} = 0.1949242$ (voir à la sous-section 3.6.4 pour la justification de ce choix de poids). Suivant les résultats mentionnés à la fin de la section 5.1, on s'attend à ce que les règles de Korobov obtiennent de meilleurs résultats que celles de type 2^s-copie. La recherche pour les règles de type 2^s-copie se fait sur toutes les règles de Korobov (à copier) d'ordre n.

Les tableaux 5.3 et 5.4 résument les résultats de ces recherches, pour $s = 6$ et $s = 12$, respectivement. On y donne, pour chacun des deux types de règles qui sont les meilleures par rapport à $\tilde{P}_2^s$, les valeurs qu'elles obtiennent pour les différents critères

utilisés aux tableaux 5.1 et 5.2. De plus, nous redonnons les meilleures règles par rapport à S_s, M_s et $M_{s,s,s}$, afin de voir comment leur valeur de $\tilde{P}_2^s$ se compare avec celle des autres.

TABLEAU 5.3: Résultats pour $s = 6$, avec $\beta_1 = \ldots = \beta_s = \sqrt{3/(8\pi^2)}$

rang	critère	n	N	a	$\tilde{P}_2^s$	S_6	M_6	$M_{6,6,6}$
6	$\tilde{P}_2^s$	79	5056	15	8.08e-5	0.6479	0.1676	0.0944
1	$\tilde{P}_2^s$	5051	5051	3564	2.10e-6*	0.8154*	0.3851	0.3851
1	S_6	5051	5051	4916	8.62e-6	0.9163*	0.4571	0.2380
1	M_6	5051	5051	4413	4.06e-6	0.7245	0.7019*	0.2907
1	$M_{6,6,6}$	5051	5051	2254	5.15e-6	0.7483	0.4987	0.4987*
6	$\tilde{P}_2^s$	157	10048	48	2.59e-5	0.7460	0.2477	0.0928
1	$\tilde{P}_2^s$	10039	10039	651	6.20e-7*	0.5533	0.5444	0.4429
1	S_6	10039	10039	9617	2.16e-6	0.8669*	0.4677	0.2831
1	M_6	10039	10039	9844	1.77e-6	0.7272	0.7272*	0.4486
1	$M_{6,6,6}$	10039	10039	2304	2.09e-6	0.5780	0.5669	0.5669*

TABLEAU 5.4: Résultats pour $s = 12$, avec $\beta_1 = \ldots = \beta_s = \sqrt{3/(8\pi^2)}$

rang	critère	n	N	a	$\tilde{P}_2^s$	S_{12}	M_{12}	$M_{12,12,12}$
12	$\tilde{P}_2^s$	3	12288	1	7.02e-2	0.8097*	0.0237	0.0237
1	$\tilde{P}_2^s$	12281	12281	11754	1.47e-5*	0.7013	0.4930	0.1811
1	S_{12}	12281	12281	12201	5.71e-5	0.8097*	0.4924	0.1158
1	M_{12}	12281	12281	9948	3.08e-5	0.7013	0.6683*	0.1202
1	$M_{12,12,12}$	12281	12281	4076	2.05e-5	0.7013	0.4458	0.3581*
12	$\tilde{P}_2^s$	5	20480	2	3.11e-2	0.7760*	0.0291	0.0184
1	$\tilde{P}_2^s$	20479	20479	6348	7.36e-6*	0.7258	0.4732	0.2209
1	S_{12}	20479	20479	18860	2.46e-5	0.8230*	0.4938	0.1080
1	M_{12}	20479	20479	14700	2.01e-5	0.7258	0.6915*	0.1927
1	$M_{12,12,12}$	20479	20479	1753	1.69e-5	0.5487	0.5013	0.3507*

Dans tous les cas, la règle de Korobov choisie avec $\tilde{P}_2^s$ obtient une plus petite valeur de $\tilde{P}_2^s$ que la règle 2^s-copie (lignes 2 et 1, respectivement) et la différence entre les deux augmente quand on passe de $s = 6$ à $s = 12$. Même les règles de Korobov choisies à l'aide d'un critère basé sur le test spectral (S_s, M_s ou $M_{s,s,s}$) obtiennent des plus petites valeurs de $\tilde{P}_2^s$ que la règle 2^s-copie. Le critère S_s est le seul où la règle 2^s-copie réussit parfois à faire mieux que les autres : ce n'est pas surprenant, car c'est celui qui

accorde le moins d'importance aux projections de petite dimension. Le fait d'utiliser le critère $\tilde{P}_2^s$ n'améliore pas les résultats des règles de type 2^s-copie par rapport au critère $M_{s,s,s}$. Par contre, les règles de Korobov choisies avec $\tilde{P}_2^s$ ont de meilleures valeurs de $M_{s,s,s}$ que celles sélectionnées en utilisant P_2^s. Ceci peut être expliqué par le fait que, tout comme $M_{s,s,s}$, $\tilde{P}_2^s$ accorde plus de poids aux projections $P_N(I)$ telles que $|I|$ est petit que ne le fait P_2^s. Or, on pourrait se demander pourquoi ce phénomène ne s'étend pas aux règles de type 2^s-copie. Nous tentons de répondre à cela dans la prochaine sous-section.

5.3.3 Critère $M_{s,s,s}$ et règles de type ν^r-copie

Nous avons fait des recherches pour trouver les meilleures règles de type 2^s-copie par rapport au critère $M_{s,s,s}$ en dimension $s = 6$ et $s = 12$, afin de voir quelle est la meilleure valeur possible que l'on peut obtenir pour ce type de règle. En dimension 12, nous avons constaté pour $n = 3$ et 5 que les règles données au tableau 5.4 étaient également les meilleures par rapport à $M_{s,s,s}$. En dimension 6, nous avons trouvé des règles ayant des valeurs de $M_{6,6,6}$ légèrement plus élevées que celles qui se trouvent au tableau 5.3. Le résultat de ces recherches est donné au tableau 5.5, où nous donnons également les valeurs de $\tilde{P}_2^6$, S_6 et M_6 pour les règles ainsi trouvées.

TABLEAU 5.5: Meilleures règles de type 2^6-copie par rapport à $M_{6,6,6}$

n	N	a	$\tilde{P}_2^s$	S_6	M_6	$M_{6,6,6}$
79	5056	18	8.78e-5	0.6479	0.2110	0.1676
157	10048	18	3.18e-5	0.7460	0.1912	0.1497
313	20232	15	1.17e-5	0.7868	0.1977	0.1539

La valeur des différents critères ne change pas beaucoup en comparaison avec les règles données au tableau 5.3. Autrement dit, le fait de choisir une règle 2^s-copie avec $\tilde{P}_2^s$ ou $M_{s,s,s}$ ne semble pas tellement affecter la qualité de la règle. Remarquons qu'étant donné que $l_I/l_{|I|}^*(N) = \nu^{1-s/|I|} l_I(n)/l_{|I|}^*(n)$, pour une règle de type 2^s-copie, si $s \geq 2$ alors

$$M_s \leq l_{\{1,2\}}/l_2^*(N) \leq 2^{1-s/2} \tag{5.2}$$

et donc, en dimension 6, on ne peut trouver de règles ayant $M_6 > 0.25$ et en dimension 12, on ne peut avoir $M_{12} > 1/32 = 0.03125$. Ceci s'applique aussi à $M_{s,s,s}$ puisque $M_{s,s,s} \leq M_s$. Donc, peu importe le critère utilisé pour choisir les règles de type 2^s-copie, on ne pourra jamais arriver à dépasser le seuil donné en (5.2) pour la valeur de $M_{s,s,s}$. Autrement dit, on obtiendra toujours de mauvaises valeurs pour ce type de règle.

Tout comme à la section précédente, ces résultats nous laissent entrevoir (et cela concorde avec ce qu'Hickernell décrit dans [46, section 6.4], en utilisant une approche différente) que pour les fonctions ayant des composantes f_I importantes lorsque $|I|$ est petit, les règles de Korobov devraient faire mieux que celles de type ν^r-copie. Par contre, pour les fonctions dont les composantes f_I avec $|I|$ près de s sont les plus importantes, alors il est peut-être préférable d'utiliser une règle de type ν^r-copie. La prochaine section explore numériquement cette hypothèse.

5.4 Exemples numériques

Nous étudions d'abord la performance des règles de type 2^s-copie en comparaison avec celles de Korobov sur le problème des options asiatiques. Puis, nous regardons ce qui se passe sur une fonction-test du même genre que celle utilisée au chapitre 4.

5.4.1 Options asiatiques

Nous comparons dans ce qui suit les facteurs de réduction de variance induits par trois types de règles sur le problème de l'évaluation des options asiatiques. Les facteurs sont donnés par rapport à la méthode MC. On compare : 1) la meilleure règle de type 2^s-copie par rapport à $\tilde{P}_2^s$; 2) la meilleure règle de Korobov par rapport au critère $\tilde{P}_2^s$; 3) la meilleure règle de Korobov par rapport à $M_{s,s,s}$, telles que données aux tableaux 5.3 et 5.4. Nous avons vu à la sous-section 3.6.3 que la qualité des projections $P_N(I)$ pour lesquelles $|I|$ est petit était importante pour le problème des options asiatiques. Pour cette raison, nous avons choisi d'utiliser le critère $\tilde{P}_2^s$ plutôt que P_2^s, car le premier donne plus de poids que le deuxième aux projections en petite dimension.

Les paramètres de l'option sont $T_1 = 0$, $T = s$ semaines, $r = 0.05$, $\sigma = 0.2$ et $S(0) = 100$. L'estimateur "ACV" est celui qui utilise des variables antithétiques et une variable de contrôle, qui est le prix de l'option sur la moyenne géométrique. Sans ces deux techniques de réduction de variance, on a l'estimateur "naïf". On utilise 500 translations aléatoires pour estimer la variance des estimateurs basés sur les règles de réseau et les valeurs rapportées aux tableaux 5.6 et 5.7 sont les facteurs de réduction de variance par rapport à l'estimateur MC correspondant, qui utilise $500N$ répétitions.

TABLEAU 5.6: Facteurs de réduction de variance, estimateur naïf

s	N	Méth.	$K = 90$	$K = 100$	$K = 110$
6	5056	copie $\tilde{P}_2^s$	1.3	0.74	0.21
	5051	$M_{6,6,6}$	601	135	2.6
	5051	$\tilde{P}_2^s$	530	29	2.5
	10048	copie $\tilde{P}_2^s$	2.4	1.0	0.23
	10039	$M_{6,6,6}$	621	33	3.0
	10039	$\tilde{P}_2^s$	1058	199	6.6
12	12288	copie $\tilde{P}_2^s$	1.7e-3	1.8e-3	1.6e-3
	12281	$M_{12,12,12}$	286	28	4.4
	12281	$\tilde{P}_2^s$	386	41	5.2
	20480	copie $\tilde{P}_2^s$	2.3e-3	2.0e-3	2.6e-3
	20479	$M_{12,12,12}$	484	58	7.5
	20479	$\tilde{P}_2^s$	554	67	9.0

Avant de commenter les résultats, nous voulons mentionner qu'étant donné que les facteurs de réduction de variance ont été estimés à l'aide de 500 répétitions, si le facteur estimé est dénoté par $\hat{R}$, alors l'intervalle de confiance au niveau 98% est donné par $(1.19^{-1}\hat{R}, 1.15\hat{R})$, puisque $F_{0.99,\infty,500} = 1.19$ et $F_{0.99,500,\infty} = 1.15$.

Pour cet exemple, les règles de type 2^s-copie sont très mauvaises en comparaison avec les règles de Korobov choisies selon $\tilde{P}_2^s$ ou $M_{s,s,s}$. En dimension 12, la variance de l'estimateur qui leur est associé est significativement plus grande que celle de l'estimateur MC, par des facteurs allant jusqu'à 500, alors que les règles de Korobov réduisent toujours la variance par rapport à MC. Les mauvaises projections des règles de type 2^s-copie expliquent leurs piètres résultats pour ce genre de problème. En ce qui concerne la différence entre les deux types de règles de Korobov, on peut dire que les

TABLEAU 5.7: Facteurs de réduction de variance, estimateur ACV

s	N	Méth.	$K = 90$	$K = 100$	$K = 110$
6	5056	copie $\tilde{P}_2^s$	0.34	0.35	0.31
	5051	$M_{6,6,6}$	8.5	18	1.6
	5051	$\tilde{P}_2^s$	15	7.4	1.8
	10048	copie $\tilde{P}_2^s$	0.36	0.32	0.32
	10039	$M_{6,6,6}$	5.8	5.1	2.3
	10039	$\tilde{P}_2^s$	20	20	2.4
12	12288	copie $\tilde{P}_2^s$	2.5e-3	2.3e-3	5.2e-3
	12281	$M_{12,12,12}$	5.1	4.5	1.8
	12281	$\tilde{P}_2^s$	11	8.8	1.5
	20480	copie $\tilde{P}_2^s$	2.6e-3	2.3e-3	5.6e-3
	20479	$M_{12,12,12}$	12	12	2.7
	20479	$\tilde{P}_2^s$	13	13	2.7

règles choisies avec $\tilde{P}_2^s$ font souvent mieux que celles choisies avec $M_{s,s,s}$, en particulier lorsque $s = 6$ et $N = 10039$. Mais le contraire est vrai dans d'autres cas, par exemple, lorsque $s = 6$ et $N = 5051$. N'oublions pas que les règles de Korobov choisies à l'aide du critère $\tilde{P}_2^s$ ont obtenu d'assez bonnes valeurs pour $M_{s,s,s}$ (voir aux tableaux 5.3 et 5.4) et donc, ce n'est pas surprenant qu'elles fassent aussi bien que ces dernières sur un problème pour lequel les projections bi- et tridimensionnelles sont importantes (c.-à-d., un problème pour lequel les composantes f_I importantes correspondent *grosso modo* aux projections $P_N(I)$ qui sont mesurées dans le critère de sélection).

Pour donner une idée de la précision de ces résultats, voici les intervalles de confiance au niveau 98% pour les rapports théoriques des variances associés aux trois dernières lignes du tableau 5.6 : (1.9e-3, 2.6e-3), (1.7e-3, 2.3e-3) et (2.2e-3, 3.0e-3) pour la règle 2^{12}-copie, (406, 556), (48.7, 66.6) et (6.33, 8.66) pour la règle choisie avec $M_{12,12,12}$ et (465, 637), (56.5, 77.4) et (7.53, 10.3) pour la règle choisie avec $\tilde{P}_2^s$. Pour cette combinaison de (N, s), il n'y pas de différence significative entre les variances des estimateurs basés sur les deux règles de Korobov. Par contre, si on regarde le cas où $s = 6$ et $N = 10048$ (ou 10039) dans le tableau 5.7, les intervalles de confiance au niveau 98% sont (0.30, 0.41), (0.27, 0.37) et (0.27, 0.37) pour la règle 2^6-copie, (4.8, 6.6), (4.3, 5.9) et (1.9, 2.6) pour la règle choisie avec $M_{6,6,6}$ et (16, 23), (16, 22) et (2.0,

2.8) pour la règle choisie avec $\tilde{P}_2^s$. Dans ce cas, il y a une différence significative entre les deux règles de Korobov lorsque K vaut 90 ou 100.

5.4.2 Fonction-test

Pour tenter d'avantager les règles de type 2^s-copie, nous comparons dans cette sous-section ces règles avec des règles de Korobov, mais sur une fonction-test (artificielle) ressemblant à celle que nous avons utilisée au chapitre précédent (section 4.6.3) et qui vient de [111]. Nous avons dû modifier la fonction, car sous sa forme originale, les règles de type 2^s-copie l'intégraient parfaitement. En effet, la symétrie de la fonction originale fait en sorte que les coefficients de Fourier $\hat{f}(\mathbf{h})$ valent 0 lorsque chaque h_j est pair. Or, on a pu voir dans la démonstration du lemme 5.2.1 que les vecteurs $\mathbf{h}$ dans le réseau dual d'une règle de type ν^s-copie sont des multiples de ν, donc si $\nu = 2$, tous les coefficients de Fourier associés au réseau dual valent 0 et ainsi, l'erreur et la variance sont nuls. Pour éviter cela, la fonction considérée ici est de la forme :

$$f(\mathbf{x}) = \frac{1}{\mu} \prod_{j=1}^{s} \frac{|4x_j - 1| + c_j}{1 + c_j},$$

où $\mu = \prod_{j=1}^{s} \frac{1.25+c_j}{1+c_j}$ et les c_j peuvent être de quatre types différents, comme dans [111] :

(1) $c_j = 0.01,\ j = 1, \ldots, s$;

(2) $c_j = 1,\ j = 1, \ldots, s$;

(3) $c_j = j,\ j = 1, \ldots, s$;

(4) $c_j = j^2,\ j = 1, \ldots, s$.

Comme pour la fonction-test utilisée à la section 4.6.3, la dimension effective de f diminue quand on passe du type 1 au type 4. Ainsi, on s'attend à ce que les règles de type 2^s-copie soient bonnes pour le type 1 et 2, quand la dimension s n'est pas trop grande.

Nous comparons d'abord les meilleures règles de type 2^s-copie par rapport au critère P_2^s avec les meilleures règles de type Korobov par rapport aux critères P_2^s, M_s et $M_{s,s,s}$, pour $s = 4, 6$ et 12. En dimension 6 et 12, ces règles se trouvent aux tableaux 5.1 et 5.2 et en dimension 4, elles se trouvent au tableau 5.8. Les tableaux 5.9, 5.10 et

5.11 contiennent les facteurs de réduction de variance par rapport à la méthode MC utilisant le même nombre de points. La variance des estimateurs basés sur les règles de réseau est estimée à l'aide de 500 translations aléatoires. Les étoiles (*) indiquent les meilleurs résultats.

TABLEAU 5.8: Règles 2^s-copie et règles de Korobov, $s = 4$

rang	critère	n	N	a	P_2^s	M_4	$M_{4,4,4}$
4	copie P_2^s	313	5008	51	5.45e-3*	0.2425	0.2425
1	P_2^s	5003	5003	962	8.33e-3	0.3444	0.3444
1	M_4	5003	5003	242	9.70e-3	0.8061*	0.2531
1	$M_{4,4,4}$	5003	5003	1633	2.02e-2	0.7415	0.7255*
4	copie P_2^s	619	9904	73	1.98e-3*	0.3229	0.2957
1	P_2^s	9901	9901	2622	2.87e-3	0.4839	0.3371
1	M_4	9901	9901	3336	3.59e-3	0.8345*	0.6651
1	$M_{4,4,4}$	9901	9901	1731	5.13e-3	0.7298	0.7210*
4	copie P_2^s	1249	19984	197	6.47e-4*	0.2556	0.2344
1	P_2^s	19979	19979	16675	8.97e-4	0.6721	0.3678
1	M_4	19979	19979	1205	1.67e-2	0.8663*	0.0858
1	$M_{4,4,4}$	19979	19979	126	1.63e-3	0.7812	0.7275*

TABLEAU 5.9: Facteurs de réduction de variance, $s = 4$

n	Méth.	$c_j = 0.01$	$c_j = 1$	$c_j = j$	$c_j = j^2$
5008	copie P_2^s	94.8	129	151	192
5003	$M_{4,4,4}$	64.1	175	413	1850
5003	M_4	139*	358*	882*	2140*
5003	P_2^s	79.3	181	387	1100
9904	copie P_2^s	144	208	275	309
9901	$M_{4,4,4}$	88.0	462	1920*	8930*
9901	M_4	132	317	746	2940
9901	P_2^s	155*	487*	1030	2420
19984	copie P_2^s	192	319	465	629
19979	$M_{4,4,4}$	33.9	74.5	156	323
19979	M_4	1.15	2.24	5.34	35.0
19979	P_2^s	715*	2420*	6380*	16200*

Comme on peut le voir au tableau 5.9, en dimension 4, les règles de type 2^s-copie obtiennent d'assez bons résultats pour la fonction de type 1 et sont meilleures que la règle de rang 1 choisie à l'aide de $M_{4,4,4}$ pour les quatre types de paramètres, lorsque

TABLEAU 5.10: Facteurs de réduction de variance, $s = 6$

n	Méth.	$c_j = 0.01$	$c_j = 1$	$c_j = j$	$c_j = j^2$
5056	copie P_2^s	6.51	7.24	8.69	9.83
5051	$M_{6,6,6}$	10.2	49.5	450	4190*
5051	M_6	23.8	141	636	2720
5051	P_2^s	26.2*	146*	969*	2800
10048	copie P_2^s	8.90	11.0	16.1	22.3
10039	$M_{6,6,6}$	9.10	81.8	488	1220
10039	M_6	10.7*	96.8*	508*	1390*
10039	P_2^s	5.43	7.68	20.1	104
20232	copie P_2^s	18.7	22.1	30.7	38.7
20229	$M_{6,6,6}$	4.66	54.3	693*	11700*
20229	M_6	32.2	122	566	2970
20229	P_2^s	33.3*	157*	647	3400

TABLEAU 5.11: Facteurs de réduction de variance, $s = 12$

n	Méth.	$c_j = 0.01$	$c_j = 1$	$c_j = j$	$c_j = j^2$
12288	copie P_2^s	4.51e-2	9.44e-3	6.59e-3	6.23e-3
12281	$M_{12,12,12}$	2.32	18.4*	373*	1770
12281	M_{12}	2.49*	15.7	283	4790*
12281	P_2^s	7.04e-2	3.50e-2	7.88e-2	0.34
20480	copie P_2^s	5.92e-2	1.33e-2	1.11e-2	1.15e-2
20479	$M_{12,12,12}$	0.64	3.08	311	10700
20479	M_{12}	2.19*	10.9*	468*	7770*
20479	P_2^s	3.95e-2	2.04e-2	5.01e-2	0.20

$N = 19979$. La règle choisie avec M_4 obtient de mauvais résultats pour cette valeur de N. Or, on voit au tableau 5.8 que sa valeur de $M_{4,4,4}$ est plutôt petite, ce qui explique ces piètres résultats. Pour avoir une idée de la précision, les intervalles de confiance au niveau 98% associés au cas où N=19984 (ou 19979) sont (161, 221), (268, 367), (391, 535) et (529, 723) pour la règle 2^4-copie (première ligne) et (601, 822), (2034, 2783), (5361, 7337) et (13613, 18630) pour la règle de Korobov choisie à l'aide de P_2^s (dernière ligne).

En dimension 6, il n'y a qu'un cas où les règles de type 2^s-copie font mieux que les règles choisies à l'aide du critère $M_{6,6,6}$ et elles sont généralement moins bonnes que les règles de Korobov choisies avec P_2^s. Quand la dimension augmente, les règles de type 2^s-copie deviennent de moins en moins bonnes, même qu'en dimension 12, leur

variance est significativement plus grande que celle de l'estimateur MC.

En dimension 12, les règles de Korobov choisies à l'aide du critère P_2^s n'obtiennent pas de très bons résultats. Nous avons répété l'expérience pour $s = 6$ et 12, mais en utilisant les règles de type 2^s-copie et de Korobov choisies à l'aide du critère $\tilde{P}_2^s$ (ces règles sont données au tableau 5.3 et 5.4). Les résultats ainsi obtenus sont donnés au tableau 5.12.

TABLEAU 5.12: Facteurs de réduction de variance, critère $\tilde{P}_2$

s	n	Méth.	$c_j = 0.01$	$c_j = 1$	$c_j = j$	$c_j = j^2$
6	5056	copie $\tilde{P}_2^s$	6.24	7.75	10.62	10.5
	5051	$\tilde{P}_2^s$	24.2	192	1410	5120
	10048	copie $\tilde{P}_2^s$	9.59	12.7	17.6	22.2
	10039	$\tilde{P}_2^s$	38.4	350	2090	6090
12	12288	copie $\tilde{P}_2^s$	4.51e-2	1.18e-2	6.51e-3	6.14e-3
	12281	$\tilde{P}_2^s$	2.21	28.7	846	4540
	20480	copie $\tilde{P}_2^s$	5.92e-2	1.74e-2	1.01e-2	1.11e-2
	20479	$\tilde{P}_2^s$	3.43	33.2	1420	17800

Pour les règles de Korobov, en dimension 12, on observe une nette amélioration en comparaison avec les résultats obtenus au tableau 5.11, alors qu'en dimension 6, il n'y a pas beaucoup de différence avec les résultats du tableau 5.10 pour $N = 5051$, mais il y en a quand $N = 10039$. Cela s'explique facilement lorsque l'on retourne aux tableaux 5.1, 5.2, 5.3 et 5.4. On constate alors qu'en dimension 12, les meilleures règles de Korobov par rapport au critère P_2^s ont de très petites valeurs pour $M_{12,12,12}$ lorsque $N = 12281, 20479$ et c'est le cas également en dimension 6, lorsque $N = 10039$. Dans le cas des règles de type 2^s-copie, le fait d'utiliser les meilleures par rapport à $\tilde{P}_2^s$ n'améliore pas vraiment les résultats.

De façon plus générale, on peut dire que lorsqu'une règle a une petite valeur pour $M_{s,s,s}$, les facteurs de réduction de variance sont moins bons. En plus des règles mentionnées au paragraphe précédent, nous avons également observé ce phénomène en dimension 4 quand $N = 19979$, avec la règle choisie selon le critère M_4. Donc, on peut dire que la valeur de $M_{s,s,s}$ d'une règle est un bon indicateur de la qualité de

son estimateur associé pour cette fonction-test. Plus précisément, la meilleure règle par rapport à $M_{s,s,s}$ n'est pas nécessairement celle ayant la plus petite variance, mais elle semble faire généralement mieux que MC et si une règle a une très petite valeur de $M_{s,s,s}$, alors sa variance associée risque d'être grande, voire supérieure à celle de l'estimateur MC. Ceci s'applique aussi aux règles de type 2^s-copie, car leur valeur de $M_{s,s,s}$ devient pire à mesure que s augmente, tout comme leurs facteurs de réduction de variance.

Chapitre 6

Conclusion

Dans cet ouvrage, nous nous sommes intéressés à une méthodologie pour la simulation qui est basée sur des règles de réseau. Nous avons étudié la variance théorique des estimateurs obtenus à l'aide de cette méthode et l'avons comparée avec celle provenant de la méthode Monte Carlo. Deux familles de règles ont été considérées : celles de type standard et celles de type polynômial. Plusieurs parallèles ont été établis entre ces deux constructions et aussi, entre les règles de réseau polynômiales et les (t, m, s)-réseaux. De nouveaux critères de sélection ont été définis, dans le but de choisir des règles de réseau permettant généralement de réduire la variance par rapport à la méthode Monte Carlo. Des résultats numériques provenant de différents problèmes nous ont permis de constater empiriquement cette réalité, en plus de nous laisser voir la simplicité de la méthode et l'étendue de son champ d'application.

Nos résultats théoriques nous ont avertis qu'il existait des fonctions pour les lesquelles les règles de réseau translatées aléatoirement feraient pire que la méthode Monte Carlo. Or, parmi les résultats numériques présentés dans cette thèse, les seuls cas où cela s'est produit étaient dus soit à un mauvais choix de la règle (c.-à-d., en utilisant un critère ne donnant pas assez de poids aux projections en basse dimension), soit au fait que l'on intègrait une fonction-test spécialement conçue pour avoir des composantes en haute dimension qui soient importantes. Aussi, nous croyons que pour la plupart des problèmes rencontrés en pratique, ce sont surtout les composantes en basse dimension de la fonction qui sont importantes. Ainsi, on peut dire que le succès de cette méthode

repose sur la capacité de construire des ensembles de points qui exploitent la structure des fonctions sur lesquelles on travaille.

Plusieurs questions ont été laissées de côté dans ce travail et mériteraient d'être étudiées. Dans le cas des règles de réseau standard, les résultats obtenus sur les polynômes pourraient être généralisés à d'autres types de fonctions. Aussi, certains résultats ont été obtenus que pour un type de règle (standard ou polynômiale) : il serait intéressant d'obtenir les résultats équivalents pour l'autre type de règle et ainsi, avoir une correspondance complète entre les deux cas. Par exemple, nous avons travaillé avec Fred Hickernell [48] à la construction de règles de réseau standard imbriquées et nous aimerions étendre ces résultats au cas polynômial, afin de pouvoir fournir une alternative aux (τ, s)-suites.

Du côté des applications, on pourrait comparer empiriquement les règles de réseau polynômiales XOR-translatées avec des (t, m, s)-réseaux brouillés, entre autres afin d'établir les différences en ce qui a trait à leur temps de calcul et à leur performance sur des problèmes en grande dimension.

Nous croyons aussi qu'il y aurait lieu de regarder plus en détail les liens entre la variance et les critères de sélection, ainsi que ceux entre la randomisation choisie et la réduction de variance obtenue. Par exemple, on pourrait tenter de répondre à la question suivante : "jusqu'à quel point doit-on ajouter de l'aléatoire dans une règle de réseau afin d'obtenir une réduction de variance garantie par rapport à la méthode Monte Carlo ?"

Le développement de méthodes d'estimation pour calculer la dimension effective d'une fonction et plus précisément, pour établir l'importance des différentes composantes ANOVA constitue une autre avenue de recherche dont l'étude pourrait nous permettre d'améliorer la performance de cette méthode. Cela pourrait peut-être aussi nous permettre de déterminer, pour une fonction donnée, si un critère de la forme $M_{t_1,\ldots,t_d}$ est préférable à un de type $\tilde{P}_\alpha^s$ pour choisir la règle de réseau. De plus, il serait intéressant de mieux comprendre comment la combinaison des techniques pour réduire la dimension effective (par exemple, le pont brownien [13] et l'analyse en composantes principales [1]) et la variance (variable de contrôle, variables antithétiques, méthode

de Monte Carlo conditionnelle, etc.) affecte la structure de la fonction que l'on intègre et quelles sont les conséquences de ces changements sur la variance des estimateurs basés sur les règles de réseau.

Dans ce travail, l'approche que nous avons privilégiée en ce qui concerne le choix des règles a été de proposer des critères de sélection, puis de faire une recherche exhaustive pour un nombre de points N donné et d'utiliser la règle obtenant la meilleure valeur pour le critère choisi. Plus généralement, ces critères peuvent servir à construire des tableaux de "bonnes règles" desquelles l'utilisateur éventuel n'a qu'à choisir la règle contenant le nombre de points désiré. La forme du critère nous permet de prédire que dans la plupart des cas, l'estimateur ainsi obtenu devrait être plus efficace que celui obtenu à l'aide de la méthode Monte Carlo. De plus, comme nous l'avons mentionné à la fin du chapitre 3, il serait intéressant de voir si la combinaison des critères $M_{t_1,\ldots,t_d}$ et $\tilde{P}_\alpha^s$ pourrait choisir des règles permettant de construire des estimateurs plus précis que ceux obtenus en prenant l'un ou l'autre de ces critères. Une approche différente de celle pour laquelle nous avons opté dans ce travail serait de recueillir de l'information sur la fonction associée au problème qui nous intéresse et d'utiliser cela pour construire une règle plus appropriée, conçue spécialement pour le problème en question. L'implantation d'une telle méthode *adaptative* constitue un autre sujet dont nous croyons que l'étude serait intéressante.

Bibliographie

[1] P. Acworth, M. Broadie, and P. Glasserman. A comparison of some Monte Carlo and quasi-Monte Carlo techniques for option pricing. In P. Hellekalek and H. Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, number 127 in Lecture Notes in Statistics, pages 1–18. Springer-Verlag, 1997.

[2] S. Asmussen. Conjugate processes and the simulation of ruin problems. *Stochastic Processes and their Applications*, 20 :213–229, 1985.

[3] S. Asmussen. Ruin probabilities expressed in terms of storage process. *Advances in Applied Probability*, 20 :913–916, 1988.

[4] S. Asmussen and R. Rubinstein. Complexity properties of steady-state rare events simulation in queueing models. In J. Dshalalow, editor, *Advances in Queueing : Theory, Methods, and Open Problems*, pages 429–462. CRC Press, 1995.

[5] A. N. Avramidis and J. R. Wilson. Integrated variance reduction strategies for simulation. *Operations Research*, 44 :327–346, 1996.

[6] A. N. Avramidis and J. R. Wilson. Correlation-induction techniques for estimating quantiles in simulation experiments. *Operations Research*, 46(4) :574–591, 1998.

[7] N. S. Bahvalov. On approximate calculation of multiple integrals. *Vestnik Moskovskogo Universiteta, Seriya Matematiki, Mehaniki, Astronomi, Fiziki, Himii*, 4 :3–18, 1959. En russe.

[8] K. G. Beauchamp. *Walsh Functions and their Applications.* Academic Press, London, 1975.

[9] H. Ben Ameur, P. L'Ecuyer, and C. Lemieux. Variance reduction of Monte Carlo and randomized quasi-Monte Carlo estimators for stochastic volatility models in finance. In *Proceedings of the 1999 Winter Simulation Conference*, pages 632–639. IEEE Press, December 1999.

[10] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81 :637–654, 1973.

[11] P. Boyle, M. Broadie, and P. Glasserman. Monte Carlo methods for security pricing. *Journal of Economic Dynamics and Control*, 21 :1267–1321, June 1997.

[12] P. Bratley, B. L. Fox, and L. E. Schrage. *A Guide to Simulation.* Springer-Verlag, New York, second edition, 1987.

[13] R. E. Caflisch and B. Moskowitz. Modified Monte Carlo methods using quasi-random sequences. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, number 106 in Lecture Notes in Statistics, pages 1–16, New York, 1995. Springer-Verlag.

[14] R. E. Caflish, W. Morokoff, and A. Owen. Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *The Journal of Computational Finance*, 1(1) :27–46, 1997.

[15] W. G. Cochran. *Sampling Techniques.* John Wiley and Sons, New York, second edition, 1977.

[16] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups.* Grundlehren der Mathematischen Wissenschaften 290. Springer-Verlag, New York, 1988.

[17] R. Couture and P. L'Ecuyer. Lattice computations for random numbers. *Mathematics of Computation*, 69(230) :757–765, 2000.

[18] R. Couture, P. L'Ecuyer, and S. Tezuka. On the distribution of k-dimensional vectors for simple and combined Tausworthe sequences. *Mathematics of Computation*, 60(202) :749–761, S11–S16, 1993.

[19] R. R. Coveyou and R. D. MacPherson. Fourier analysis of uniform random number generators. *Journal of the ACM*, 14 :100–119, 1967.

[20] R. Cranley and T. N. L. Patterson. Randomization of number theoretic methods for multiple integration. *SIAM Journal on Numerical Analysis*, 13(6) :904–914, 1976.

[21] U. Dieter. How to calculate shortest vectors in a lattice. *Mathematics of Computation*, 29(131) :827–833, 1975.

[22] S. A. R. Disney and I. H. Sloan. Error bounds for the method of good lattice points. *Mathematics of Computation*, 56 :257–266, 1991.

[23] S.A.R. Disney and I.H. Sloan. Lattice integration rules of maximal rank formed by copying rank 1 rules. *SIAM Journal on Numerical Analysis*, 29 :566–577, 1992.

[24] D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, second edition, 1996.

[25] B. Efron and C. Stein. The jackknife estimator of variance. *Annals of Statistics*, 9 :586–596, 1981.

[26] K. Entacher. Quasi-Monte Carlo methods for numerical integration of multivariate Haar series. *BIT*, 37 :846–861, 1997.

[27] K. Entacher, P. Hellekalek, and P. L'Ecuyer. Quasi-Monte Carlo node sets from linear congruential generators. In H. Niederreiter and J. Spanier, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pages 188–198, Berlin, 2000. Springer.

[28] H. Faure. Discrépance des suites associées à un système de numération. *Acta Arithmetica*, 61 :337–351, 1982.

[29] G. S. Fishman. *Monte Carlo : Concepts, Algorithms, and Applications*. Springer Series in Operations Research. Springer-Verlag, New York, 1996.

[30] G. S. Fishman and B. D. Wang. Antithetic variates revisited. *Communications of the ACM*, 26 :964–971, 1983.

[31] B. L. Fox. *Stategies for Quasi-Monte Carlo.* Kluwer Academic, Boston, MA, 1999.

[32] M. Fushimi, K. Kobayashi, and H. Morohosi. Error estimation for quasi-Monte Carlo methods : Empirical results in financial engineering. *Mathematics and Computers in Simulation*, 1999. À paraître.

[33] M. Fushimi and S. Tezuka. The k-distribution of generalized feedback shift register pseudorandom numbers. *Communications of the ACM*, 26(7) :516–523, 1983.

[34] H. Gerber. *An Introduction To Risk Theory.* Monograph 8, University of Pennsylvania, Philadelphia. Huebner Foundation, 1979.

[35] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Asymptotically optimal importance sampling and stratification for pricing path dependent options. *Mathematical Finance*, 9(2) :117–152, 1999.

[36] B. Golubov, A. Efimov, and V. Skvortsov. *Walsh Series and Transforms : Theory and Applications*, volume 64 of *Mathematics and Applications : Soviet Series.* Kluwer Academic Publishers, Boston, 1991.

[37] A. Graham. *Kronecker Products and Matrix Calculus : with Applications.* Ellis Horwood series in mathematics and its applications. Halsted Press, Chichester, 1981.

[38] S. Haber. Parameters for integrating periodic functions of several variables. *Mathematics of Computation*, 41 :115–129, 1983.

[39] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2 :84–90, 1960.

[40] J. M. Hammersley and D. C. Handscomb. A new Monte Carlo technique : Antithetic variates. *Proceedings of the Cambridge Philosophical Society*, 52 :449–475, 1956.

[41] G. H. Hardy. On the representation of a number as the sum of any number of squares, and in particular of five. *Transactions of the American Mathematical Society*, 21 :255–284, 1920.

[42] P. Hellekalek. On the assessment of random and quasi-random point sets. In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*, volume 138 of *Lecture Notes in Statistics*, pages 49–108. Springer, New York, 1998.

[43] P. Hellekalek, G. Larcher, H. Niederreiter, and P. Zinterhof, editors. *Monte Carlo and Quasi-Monte Carlo Methods 1996*, volume 127 of *Lecture Notes in Statistics*. Springer, New York, 1998.

[44] F. J. Hickernell. Quadrature error bounds with applications to lattice rules. *SIAM Journal on Numerical Analysis*, 33 :1995–2016, 1996.

[45] F. J. Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67 :299–322, 1998.

[46] F. J. Hickernell. Lattice rules : How well do they measure up? In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*, volume 138 of *Lecture Notes in Statistics*, pages 109–166. Springer, New York, 1998.

[47] F. J. Hickernell. What affects accuracy of quasi-Monte Carlo quadrature? In H. Niederreiter and J. Spanier, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pages 16–55, Berlin, 2000. Springer.

[48] F. J. Hickernell, H. S. Hong, P. L'Ecuyer, and C. Lemieux. Extensible lattice sequences for quasi-Monte Carlo quadrature. *SIAM Journal on Scientific Computing*, 2000. À paraître.

[49] F. J. Hickernell and R. Yue. The mean-square discrepancy of scrambled (t, s)-sequences. Soumis, 1999.

[50] E. Hlawka. Zur angenäherten berechnung mehrfacher integrale. *Monatshefte für Mathematik*, 66 :140–151, 1962.

[51] W. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19 :293–325, 1948.

[52] J. Hoogland, F. James, and R. Kleiss. Quasi-Monte Carlo, discrepancies and error estimates. In P. Hellekalek, G. Larcher, H. Niederreiter, and P. Zinte-

rhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1996*, volume 127 of *Lecture Notes in Statistics*, pages 266–276. Springer-Verlag, New York, 1998.

[53] S. Joe. Randomization of lattice rules for numerical multiple integration. *Journal of Computational and Applied Mathematics*, 31 :299–304, 1990.

[54] S. Joe and S.A.R. Disney. Intermediate rank lattice rules for multidimensional integration. *SIAM Journal on Numerical Analysis*, 30 :569–582, 1993.

[55] S. Joe and I.H. Sloan. Imbedded lattice rules for multidimensional integration. *SIAM Journal on Numerical Analysis*, 29 :1119–1135, 1992.

[56] I. Karatzas. *Lectures on the Mathematics of Finance.* Centre de Recherches Mathématiques, Université de Montréal, 1996.

[57] A. G. Z. Kemna and A. C. F. Vorst. A pricing method for options based on average asset values. *Journal of Banking and Finance*, 14 :113–129, 1990.

[58] D. E. Knuth. *The Art of Computer Programming, Volume 2 : Seminumerical Algorithms.* Addison-Wesley, Reading, Mass., second edition, 1981.

[59] D. E. Knuth. *The Art of Computer Programming, Volume 2 : Seminumerical Algorithms.* Addison-Wesley, Reading, Mass., third edition, 1998.

[60] N.M. Korobov. The approximate computation of multiple integrals. *Dokl. Akad. Nauk. SSSR*, 124 :1207–1210. 1959.

[61] N.M. Korobov. Properties and calculation of optimal coefficients. *Soviet Math. Dokl.*, 1 :696–700. 1960.

[62] L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences.* John Wiley, New York, 1974.

[63] T. N. Langtry. An application of Diophantine approximation to the construction of rank-1 lattice quadrature rules. *Mathematics of Computation*, 65 :1635–1662, 1996.

[64] G. Larcher. Digital point sets : Analysis and applications. In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*, volume 138 of *Lecture Notes in Statistics*, pages 167–222. Springer, New York, 1998.

[65] G. Larcher, H. Niederreiter, and W. Ch. Schmid. Digital nets and sequences constructed over finite rings and their application to quasi-Monte Carlo integration. *Monatshefte für Mathematik*, 121(3) :231–253, 1996.

[66] G. Larcher and W. C. Schmid. Multivariate Walsh series, digital nets and quasi-Monte Carlo integration. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Note in Statistics*, pages 252–262, New York, 1995. Springer-Verlag.

[67] G. Larcher and C. Traunfellner. The numerical integration of Walsh series. *Mathematics of Computation*, 63 :277–291, 1994.

[68] H. J. Larson. *Introduction to probability theory and statistical inference.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, third edition, 1982.

[69] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis.* McGraw-Hill, New York, second edition, 1991.

[70] P. L'Ecuyer. Efficiency improvement via variance reduction. In *Proceedings of the 1994 Winter Simulation Conference*, pages 122–132. IEEE Press, 1994.

[71] P. L'Ecuyer. Uniform random number generation. *Annals of Operations Research*, 53 :77–120, 1994.

[72] P. L'Ecuyer. Maximally equidistributed combined Tausworthe generators. *Mathematics of Computation*, 65(213) :203–213, 1996.

[73] P. L'Ecuyer. Uniform random number generators. In *Proceedings of the 1998 Winter Simulation Conference*, pages 97–104. IEEE Press, Dec 1998.

[74] P. L'Ecuyer. Good parameters and implementations for combined multiple recursive random number generators. *Operations Research*, 47(1) :159–164, 1999.

[75] P. L'Ecuyer. Tables of linear congruential generators of different sizes and good lattice structure. *Mathematics of Computation*, 68(225) :249–260, 1999.

[76] P. L'Ecuyer and R. Couture. LatMRG user's guide, a toolkit for theoretical testing of linear congruential and multiple recursive generators. Rapport technique, Montréal, Canada, 1996. En préparation.

[77] P. L'Ecuyer and R. Couture. An implementation of the lattice and spectral tests for multiple recursive linear random number generators. *INFORMS Journal on Computing*, 9(2) :206–217, 1997.

[78] P. L'Ecuyer and P. Hellekalek. Random number generators : Selection criteria and testing. In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*, volume 138 of *Lecture Notes in Statistics*, pages 223–265. Springer, New York, 1998.

[79] P. L'Ecuyer and C. Lemieux. Quasi-Monte Carlo via linear shift-register sequences. In *Proceedings of the 1999 Winter Simulation Conference*, pages 336–343. IEEE Press, 1999.

[80] P. L'Ecuyer and C. Lemieux. Variance reduction via lattice rules. *Management Science*, 2000. À paraître.

[81] E. L. Lehmann. Some concepts of dependence. *Annals of Mathematical Statistics*, 37 :1137–1153, 1966.

[82] C. Lemieux and P. L'Ecuyer. Efficiency improvement by lattice rules for pricing asian options. In D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 579–586, Piscataway, NJ, 1998. IEEE Press.

[83] C. Lemieux and P. L'Ecuyer. Lattice rules for the simulation of ruin problems. In *Proceedings of the 1999 European Simulation Multiconference*, volume 2, pages 533–537, Ghent, Belgium, 1999. The Society for Computer Simulation.

[84] C. Lemieux and P. L'Ecuyer. A comparison of Monte Carlo, lattice rules and other low-discrepancy point sets. In H. Niederreiter and J. Spanier, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pages 326–340, Berlin, 2000. Springer.

[85] C. Lemieux and P. L'Ecuyer. Selection criteria for lattice rules and other low-discrepancy point sets. *Mathematics and Computers in Simulation*, 2000. À paraître.

[86] R. Lidl and H. Niederreiter. *Introduction to Finite Fields and Their Applications.* Cambridge University Press, Cambridge, 1986.

[87] D. V. Lindley. The theory of queues with a single server. *Proceedings of the Cambridge Philosophical Society*, 43 :277–289, 1952.

[88] J. N. Lyness and I. H. Sloan. Some properties of rank-2 lattice rules. *Mathematics of Computation*, 53 :627–637, 1989.

[89] D. Maisonneuve. Recherche et utilisation des "bons treillis", programmation et résultats numériques. In S. K. Zaremba, editor, *Applications of Number Theory to Numerical Analysis*, pages 121–201. Academic Press, New York, 1972.

[90] M. D. Mckay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21 :239–245, 1979.

[91] N. Metropolis and S.M. Ulam. The Monte Carlo method. *Journal of the Amer. Stat. Assoc.*, 44 :335–341, 1949.

[92] F. Michaud. Estimating the probability of ruin for variable premiums by simulation. *Astin Bulletin*, 26 :93–105, 1996.

[93] H. Morohosi and M. Fushimi. A practical approach to the error estimation of quasi-Monte Carlo integration. Technical Report METR 98-10, The University of Tokyo, Dept. of Math. Engineering and Information Physics, 1998.

[94] H. Niederreiter. Quasi-Monte Carlo methods and pseudorandom numbers. *Bulletin of the American Mathematical Society*, 84(6) :957–1041, 1978.

[95] H. Niederreiter. Multidimensional numerical integration using pseudorandom numbers. *Mathematical Programming Study*, 27 :17–38, 1986.

[96] H. Niederreiter. Point sets and sequences with small discrepancy. *Monatshefte für Mathematik*, 104 :273–337, 1987.

[97] H. Niederreiter. New methods for pseudorandom number and pseudorandom vector generation. In *Proceedings of the 1992 Winter Simulation Conference*, pages 264–269. IEEE Press, 1992.

[98] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63 of *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1992.

[99] H. Niederreiter. Improved error bounds for lattice rules. *Journal of Complexity*, 9 :60–75, 1993.

[100] H. Niederreiter and J. Spanier, editors. *Monte Carlo and Quasi-Monte Carlo Methods 1998*. Springer, Berlin, 2000.

[101] H. Niederreiter and C. Xing. Quasirandom points and global function fields. In *Finite fields and applications (Glasgow, 1995)*, volume 233 of *London Math. Soc. Lecture Note Ser.*, pages 269–296. Cambridge Univ. Press, Cambridge, 1996.

[102] H. Niederreiter and C. Xing. The algebraic-geometry approach to low-discrepancy sequences. In P. Hellekalek, G. Larcher, H. Niederreiter, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 127 of *Lecture Notes in Statistics*, pages 139–160, New York, 1997. Springer-Verlag.

[103] H. Niederreiter and C. Xing. Nets, (t, s)-sequences, and algebraic geometry. In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*, volume 138 of *Lecture Notes in Statistics*, pages 267–302. Springer, New York, 1998.

[104] A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, number 106 in Lecture Notes in Statistics, pages 299–317. Springer-Verlag, 1995.

[105] A. B. Owen. Monte Carlo variance of scrambled equidistribution quadrature. *SIAM Journal on Numerical Analysis*, 34(5) :1884–1910, 1997.

[106] A B. Owen. Scrambled net variance for integrals of smooth functions. *Annals of Statistics*, 25(4) :1541–1562, 1997.

[107] A. B. Owen. Latin supercube sampling for very high-dimensional simulations. *ACM Transactions of Modeling and Computer Simulation*, 8(1) :71–102, 1998.

[108] A. B. Owen and D. A. Tavella. Scrambled nets for value-at-risk calculations. In S. Grayling, editor, *VAR Understanding and Applying Value-At-Risk*, pages 257–273, London, 1997. Risk Publications.

[109] A.B. Owen. Scrambling Sobol and Niederreiter-Xing points. Technical report, Department of Statistics, Stanford University, Palo Alto, California, U.S.A., 1997.

[110] S. Paskov and J. Traub. Faster valuation of financial derivatives. *Journal of Portfolio Management*, 22 :113–120, 1995.

[111] I. Radovic, I. M. Sobol', and R. F. Tichy. Quasi-Monte Carlo methods for numerical integration : Comparison of different low-discrepancy sequences. *Monte Carlo Methods and Applications*, 2 :1–14, 1996.

[112] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, New York, second edition, 1974.

[113] W. Ch. Schmid. Shift-nets : a new class of binary digital (t, m, s)-nets. In P. Hellekalek, G. Larcher, H. Niederreiter, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 127 of *Lecture Notes in Statistics*, pages 369–381. New York, 1997. Springer-Verlag.

[114] W. Ch. Schmid. Projections of digital nets and sequences. *Mathematics and Computers in Simulation*, 1999. À paraître.

[115] J. E. H. Shaw. A quasirandom approach to bayesian statistics. *Annals of Statistics*, 16 :895–914, 1988.

[116] I. H. Sloan and S. Joe. *Lattice Methods for Multiple Integration.* Clarendon Press, Oxford, 1994.

[117] I.H. Sloan and P.J. Kachoyan. Lattice methods for multiple integration : Theory, error analysis and examples. *SIAM Journal on Numerical Analysis*, 24 :116–128, 1987.

[118] I.H. Sloan and J.N. Lyness. The representation of lattice quadrature rules as multiple sums. *Mathematics of Computation*, 52 :81–94, 1989.

[119] I.H. Sloan and L. Walsh. Lattice rules - classification and searches. In H. Brass and G. Hämmerlin, editors, *Numerical Integration III*, volume 85, pages 251–260. ISNM, Birkhäuser, Basel, 1988.

[120] I.H. Sloan and L. Walsh. A computer search of rank 2 lattice rules for multidimensional quadrature. *Mathematics of Computation*, 54 :281–302, 1990.

[121] I. M. Sobol'. The distribution of points in a cube and the approximate evaluation of integrals. *U.S.S.R. Comput. Math. and Math. Phys.*, 7 :86–112, 1967.

[122] I. M. Sobol'. *Multidimensional Quadrature Formulas and Haar Functions.* Nauka, Moskow, 1969. En russe.

[123] I. M. Sobol'. On quasi-Monte Carlo integration. *Mathematics and Computers in Simulation*, 47 :103–112, 1998.

[124] J. Spanier. Quasi-Monte Carlo methods for particle transport problems. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Notes in Statistics*, pages 121–148, New York, 1995. Springer-Verlag.

[125] J. Spanier and E. H. Maize. Quasi-random methods for estimating integrals using relatively small samples. *SIAM Review*, 36 :18–44, 1994.

[126] M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29 :143–151, 1987.

[127] S. Tezuka. *Uniform Random Numbers : Theory and Practice.* Kluwer Academic Publishers, Norwell, Mass., 1995.

[128] J. P. R. Tootill, W. D. Robinson, and D. J. Eagle. An asymptotically random Tausworthe sequence. *Journal of the ACM*, 20 :469–481, 1973.

[129] B. Tuffin. On the use of low-discrepancy sequences in Monte Carlo methods. Technical Report No. 1060, I.R.I.S.A., Rennes, France, 1996.

[130] B. Tuffin. Variance reduction order using good lattice points in Monte Carlo methods. *Computing*, 61 :371–378, 1998.

[131] F. J. Vázquez-Abad. RPA pathwise derivative estimation of ruin probabilities. *Insurance : Mathematics and Economics*, 2000. À paraître.

[132] F. J. Vázquez-Abad and D. Dufresne. Accelerated simulation for pricing Asian options. In D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 1493–1500, Piscataway, NJ, 1998. IEEE Press.

[133] X. Wang and F. J. Hickernell. Randomized Halton sequences. À paraître, 2001.

[134] G. W. Wasilkowski. Average case complexity. *Journal of Complexity*, 12 :257–272, 1996.

[135] G. A. Willard. Calculating prices and sensitivities for path-dependent derivatives securities in multifactor models. *Journal of derivatives*, 5 :45–61, Fall 1997.

[136] H. Woźniakowski. Average case complexity of multivariate integration. *Bulletin (New Series) of the American Mathematical Society*, 24 :185–194, 1991.

[137] P. Zinterhof. Über einige Abschätzungen bei der Approximation von Funktionen mit Gleichverteilungsmethoden. *Sitzungsber. Österr. Akad. Wiss. Math.-Natur. Kl. II*, 185 :121–132, 1976.

Annexe A

Contre-exemple

Proposition : *Soit la règle de réseau* $P_N = \{((i-1)/N, (i-1)/N), i = 1, \ldots, N\}$. *Soit* $\epsilon < (1 + 2(N-1))/2N^2$ *et* $b_\epsilon = (1 - \sqrt{1-2\epsilon})$. *Soient* $c_1 < c_2 < c_3$ *tels que* $c_2 = (c_1 + c_3)/2$. *Si* f *est définie par*

$$f(x,y) = \begin{cases} c_1 & si\ y < x - b_\epsilon \\ c_2 & si\ x - b_\epsilon \leq y < x + b_\epsilon \\ c_3 & si\ y \geq x + b_\epsilon, \end{cases}$$

et que $N > 8$, *alors*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) > \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Démonstration : d'abord, remarquons que $b_\epsilon < 1/N$ et que $\epsilon = b_\epsilon - b_\epsilon^2/2$. Posons

$$\begin{aligned} C_1 &= \{(x,y) : f(x,y) = c_1\}; \\ C_2 &= \{(x,y) : f(x,y) = c_2\}; \\ C_3 &= \{(x,y) : f(x,y) = c_3\}. \end{aligned}$$

Ces trois régions sont illustrées à la figure A.1.

Dans le cas de l'estimateur MC, la variance est donnée par $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) = \sigma^2/N$, où $\sigma^2 = \mathrm{Var}(f(X,Y)) = \int_0^1 \int_0^1 f^2(x,y)dxdy - \left[\int_0^1 \int_0^1 f(x,y)dxdy\right]^2$. Or, on a que $\mathrm{Vol}(C_1) = \mathrm{Vol}(C_3) = 0.5 - \epsilon$ et $\mathrm{Vol}(C_2) = 2\epsilon$, car C_2 est constitué de deux parallélogrammes symétriques par rapport à l'axe $y = x$ et d'aire b_ϵ auxquels on a enlevé un triangle-

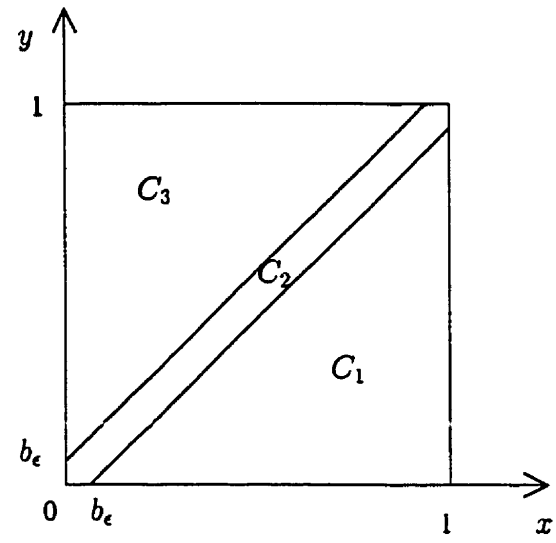


FIGURE A.1: $f(x, y)$

rectangle d'aire $b_\epsilon^2/2$. Donc, on a que

$$\mathrm{E}(\hat{\mu}_{\mathrm{MC}}) = \mu = (0.5 - \epsilon)(c_1 + c_3) + 2\epsilon c_2 = c_2 = (c_1 + c_3)/2$$

et

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) &= \frac{1}{N}\left\{(0.5 - \epsilon)\left[(c_1 - \mu)^2 + (c_3 - \mu)^2\right] + 2\epsilon(c_2 - \mu)^2\right\} \\ &= (1 - 2\epsilon)(c_1 - \mu)^2/N, \end{aligned}$$

puisque $(c_2 - \mu) = 0$ et $(c_1 - \mu) = -(c_3 - \mu)$. Remarquons que l'hypothèse que $c_1 < c_2 < c_3$ nous assure que $c_1, c_3 \neq \mu$ et donc, que $\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) > 0$.

Maintenant, pour calculer $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$, on doit procéder différemment. On sait que les points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ se trouvent sur la diagonale $y = x$, distancés de la même longueur $\sqrt{2}/N$. Après la translation aléatoire donnée par $\Delta = (\Delta_1, \Delta_2)$, on doit voir dans quelle région les N points se trouveront. Pour faire cela, nous définissons les ensembles suivants :

$$\begin{aligned} L_{i,j}^H &= \{(x, y) : j_x = i, j_y = j,\ y - j_y > x - j_x\} \\ L_{i,j}^B &= \{(x, y) : j_x = i, j_y = j,\ y - j_y \leq x - j_x\} \\ L_{i,j} &= L_{i,j}^H \cup L_{i,j}^B, \end{aligned}$$

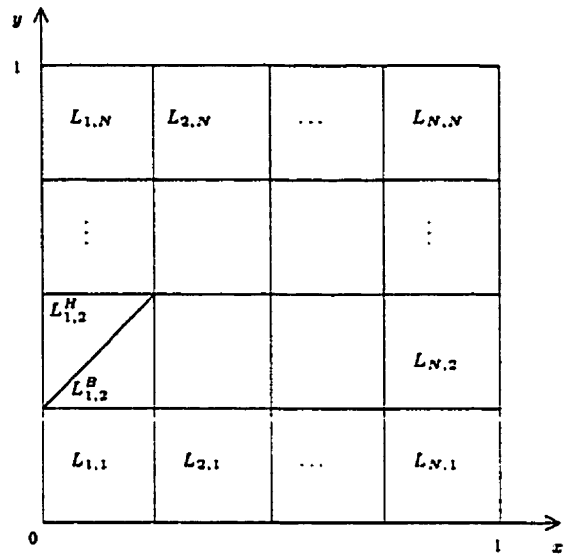


FIGURE A.2: Ensembles $L_{i,j}$

pour $1 \leq i, j \leq N$ et où $j_x = \lfloor Nx \rfloor + 1$ et $j_y = \lfloor Ny \rfloor + 1$. Donc, les ensembles $L_{i,j}$ sont des carrés de côté $1/N$ qui partitionnent l'hypercube $[0,1)^2$ et $L^H_{i,j}$ et $L^B_{i,j}$ représentent respectivement la partie au-dessus et au-dessous de la diagonale de longueur $\sqrt{2}/N$ qui part du coin inférieur gauche du carré. La figure A.2 illustre comment ces régions sont définies.

Supposons que $\Delta \in L_{i,j}$ et posons $d_\Delta = j - i \bmod N$. On peut voir que chaque boîte $\{L_{1,d_\Delta+1}, \ldots, L_{N,d_\Delta}\}$ contient un point parmi $\{(\mathbf{x}_w + \Delta) \bmod 1,\ w = 1, \ldots, N\}$ (Δ se trouve dans $L_{i,j}$, $((1/N + \Delta_1, 1/N + \Delta_2) \bmod 1)$ se trouve dans $L_{\alpha,\beta}$, avec $\alpha = (i+1) \bmod N$ et $\beta = (j+1) \bmod N$, etc.). Nous allons donc regarder, selon la valeur de Δ, dans quelle région (parmi C_1, C_2 et C_3) se trouvent chacune de ces boîtes $\{L_{1,d_\Delta+1}, \ldots, L_{N,d_\Delta}\}$ et cela va nous indiquer combien de points translatés trouve-t-on dans chaque région C_1, C_2 et C_3. Il y a six cas possibles, qui sont illustrés à la figure A.3.

(1) Lorsque $d_\Delta \in \{2, \ldots, N-2\}$, on trouvera d_Δ points dans C_1 et $N - d_\Delta$ points dans C_3 : ceci vient du fait que pour (i,j) tel que $j - i \in \{2, \ldots, N-2\}$, on a

$$L_{i,j} \subseteq \begin{cases} C_1 & \text{si } i > j \\ C_3 & \text{sinon,} \end{cases}$$

puisque $b_\epsilon < 1/N$ implique que ces boîtes $L_{i,j}$ n'ont pas de point en commun

avec C_2 et que parmi les N couples (i,j) tels que $j-i = d_\Delta \bmod N$, il y en a d_Δ pour lesquels $i > j$ (quand $(i,j) = (N, d_\Delta), (N-1, d_\Delta - 1), \ldots, (d_\Delta + 1, 1)$). Pour chaque valeur de k dans $\{2, \ldots, N-2\}$, on a que

$$P(\Delta \in L_{i,j} \text{ et } d_\Delta = k) = N(1/N^2) = 1/N.$$

(2) $\Delta \in L^H_{i,j}$, avec $d_\Delta = 1$. Alors sur les N couples (i,j) tels que $j - i = 1 \bmod N$, il n'y en a qu'un $((i,j) = (N,1))$ pour lequel on a $j < i$ et donc, un seul des $L^H_{i,j}$ associés se trouve dans C_1. Donc, si $\Delta \in L^H_{i,j}$ et $j - i = 1 \pmod N$, alors on a un point dans C_1 et $N-1$ points dans C_3 et la probabilité de cet événement est donnée par

$$P(\Delta \in L^H_{i,j} \text{ et } j - i = 1 \bmod N) = N(1/2N^2),$$

car chacun des N ensembles $L^H_{i,j}$ tel que $j - i = 1 \bmod N$ a une aire de $1/(2N^2)$.

(3) $\Delta \in L^B_{i,j}$, avec $d_\Delta = N-1$. Alors sur les N couples (i,j) tels que $j - i = N - 1 \pmod N$, il n'y en a qu'un $((i,j) = (1,N))$ pour lesquels on a $j > i$ et donc, un seul des $L^B_{i,j}$ associés se trouve dans C_3. Donc, si $\Delta \in L^B_{i,j}$ et $j - i = N - 1 \pmod N$, alors on a $N-1$ points dans C_1 et un point dans C_3 et la probabilité de cet événement est donnée par

$$P(\Delta \in L^B_{i,j} \text{ et } j - i = N - 1 \pmod N) = N(1/2N^2),$$

car chacun des N ensembles $L^B_{i,j}$ tel que $j - i = N - 1 \pmod N$ a une aire de $1/(2N^2)$.

(4) Si $\Delta \in L^B_{i,j}$ et $d_\Delta = 1$, alors il y a deux cas possibles :

(a) si $1/N - b_\epsilon \leq \Delta_1 - i < 1/N$ et $0 \leq \Delta_2 - j < b_\epsilon$, alors $N-1$ des points $\mathbf{x}_w + \Delta$ se trouvent dans C_2 et un se trouve dans C_1. La probabilité que Δ soit dans un de ces ensembles est de $N\,(1/N - (1/N - b_\epsilon))\, b_\epsilon/2 = N b_\epsilon^2/2$.

(b) Sinon, $N-1$ points se trouvent dans C_3 et un dans C_1. Ceci se produit avec une probabilité de $N\,((1/2N^2) - b_\epsilon^2/2)$.

(5) Si $\Delta \in L^H_{i,j}$ et $d_\Delta = N-1$, alors il y a deux cas possibles :

(a) si $0 \leq \Delta_1 - i < b_\epsilon$ et $1/N - b_\epsilon \leq \Delta_2 - j < 1/N$ alors parmi les N points $\{(\mathbf{x}_w + \Delta) \bmod 1, w = 1, \ldots, N\}$, il y en a $N-1$ qui se trouvent dans C_2 et l'autre point se trouve dans C_3. Ceci se produit avec probabilité de $Nb_\epsilon^2/2$.

(b) Sinon, $N-1$ se trouvent dans C_1 et un seul se trouve dans C_3. Ceci se produit avec une probabilité de $N\left((1/2N^2) - b_\epsilon^2/2\right)$.

(6) Si $\Delta \in L_{i,j}$ et que $d_\Delta = 0$, alors il y a trois cas possibles :

(a) Si $b_\epsilon \leq \Delta_1 - i < 1/N$ et que $0 \leq \Delta_2 - i < 1/N - b_\epsilon$, alors les N points du réseau translaté se trouvent dans C_1. L'aire de cette région est donnée par $N(1/N - b_\epsilon)^2/2 = N(1/N^2 - 2b_\epsilon/N + b_\epsilon^2)/2$.

(b) Si $0 \leq \Delta_1 - i < 1/N - b_\epsilon$ et que $b_\epsilon \leq \Delta_2 - i < 1/N$, alors les N points du réseau translaté se trouvent dans C_3. L'aire de cette région est donnée par $N(1/N - b_\epsilon)^2/2 = N(1/N^2 - 2b_\epsilon/N + b_\epsilon^2)/2$.

(c) Sinon, les N points se trouvent dans C_2. L'aire de cette région est donnée par $2N(1/2N^2 - (1/N - b_\epsilon)^2/2) = 2b_\epsilon - Nb_\epsilon^2$.

Récapitulons en donnant tous les cas possibles pour (N_1, N_2, N_3) avec leur probabilité associée, où

$$N_k = \sum_{i=1}^{N} \mathbf{1}_{\{((\mathbf{x}_i + \Delta) \bmod 1) \in C_k\}}, \quad k = 1, 2, 3.$$

$$\begin{aligned}
P(N_1 = j, N_2 = 0, N_3 = N - j) &= 1/N, \qquad j = 2, \ldots, N-2 \qquad \text{(voir item 1)}, \\
P(N_1 = 1, N_2 = 0, N_3 = N - 1) &= N(1/2N^2) + N(1/2N^2 - b_\epsilon^2/2) \\
&= 1/N - Nb_\epsilon^2/2 \qquad \text{(voir items 2 et 4(b))}, \\
P(N_1 = N - 1, N_2 = 0, N_3 = 1) &= N(1/2N^2) + N(1/2N^2 - b_\epsilon^2/2) \\
&= 1/N - Nb_\epsilon^2/2 \qquad \text{(voir items 3 et 5(b))}, \\
P(N_1 = 0, N_2 = 0, N_3 = N) &= N(1/N^2 - 2b_\epsilon/N + b_\epsilon^2)/2 \\
&= 1/2N - b_\epsilon + Nb_\epsilon^2/2 \qquad \text{(voir item 6(b))}, \\
P(N_1 = N, N_2 = 0, N_3 = 0) &= N(1/N^2 - 2b_\epsilon/N + b_\epsilon^2)/2 \\
&= 1/2N - b_\epsilon + Nb_\epsilon^2/2 \qquad \text{(voir item 6(a))}, \\
P(N_1 = 0, N_2 = N, N_3 = 0) &= 2b_\epsilon - Nb_\epsilon^2 \qquad \text{(voir item 6(c))},
\end{aligned}$$

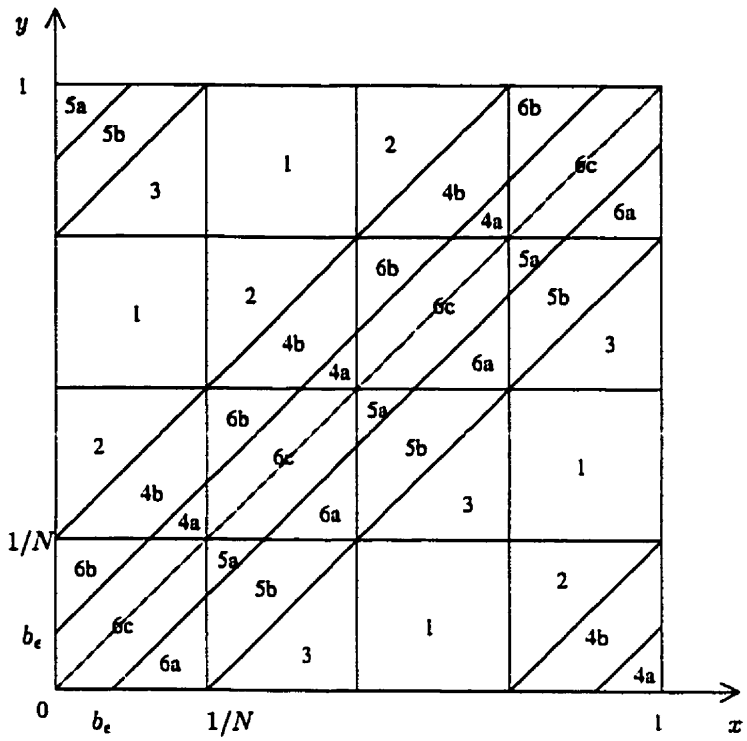


FIGURE A.3: Régions déterminées par les cas 1 à 5

$$P(N_1 = 1, N_2 = N-1, N_3 = 0) = Nb_\epsilon^2/2 \qquad \text{(voir item 4(a))},$$
$$P(N_1 = 0, N_2 = N-1, N_3 = 1) = Nb_\epsilon^2/2 \qquad \text{(voir item 5(a))}.$$

En combinant ces différents cas, on peut maintenant calculer la variance de $\hat{\mu}_{\mathrm{LR}}$:

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) &= \mathrm{E}\left[\frac{1}{N}\left(c_1N_1 + c_2N_2 + c_3N_3 - N\mu\right)\right]^2 \\
&= \mathrm{E}\left[(c_1-\mu)\frac{N_1}{N} + (c_2-\mu)\frac{N_2}{N} + (c_3-\mu)\frac{N_3}{N}\right]^2 \\
&= \frac{1}{N}\sum_{i=2}^{N-2}\left[\frac{i}{N}(c_1-\mu) + \frac{N-i}{N}(c_3-\mu)\right]^2 \\
&\quad + \left(1/N - Nb_\epsilon^2/2\right)\left[\frac{1}{N}(c_1-\mu) + \frac{N-1}{N}(c_3-\mu)\right]^2 \\
&\quad + \left(1/N - Nb_\epsilon^2/2\right)\left[\frac{N-1}{N}(c_1-\mu) + \frac{1}{N}(c_3-\mu)\right]^2 \\
&\quad + \left(1/2N - b_\epsilon + Nb_\epsilon^2/2\right)\left[(c_1-\mu)^2 + (c_3-\mu)^2\right] \\
&\quad + \left(2b_\epsilon - Nb_\epsilon^2\right)(c_2-\mu)^2 + Nb_\epsilon^2/2\left[\frac{N-1}{N}(c_2-\mu) + \frac{1}{N}(c_1-\mu)\right]^2 \\
&\quad + Nb_\epsilon^2/2\left[\frac{N-1}{N}(c_2-\mu) + \frac{1}{N}(c_3-\mu)\right]^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}(c_1-\mu)^2\sum_{i=1}^{N-1}\left(1-\frac{2i}{N}\right)^2+(-Nb_\epsilon^2)\left(1-\frac{2}{N}\right)^2(c_1-\mu)^2\\
&\quad+\left(\frac{1}{N}-2b_\epsilon+Nb_\epsilon^2\right)(c_1-\mu)^2+\frac{b_\epsilon^2}{N}(c_1-\mu)^2\\
&= (c_1-\mu)^2\left[\frac{N-1}{N}-\frac{4N(N-1)}{2N^2}+\frac{4N(N-1)(2N-1)}{6N^3}\right.\\
&\quad\left.+b_\epsilon^2\left(-N\left(1-\frac{4}{N}+\frac{4}{N^2}\right)+N+\frac{1}{N}\right)+\left(\frac{1}{N}-2b_\epsilon\right)\right]\\
&= (c_1-\mu)^2\left[\frac{N^2+2}{3N^2}+b_\epsilon^2\left(4-\frac{3}{N}\right)-2b_\epsilon\right].
\end{aligned}
$$

Ainsi, une condition suffisante pour avoir $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) > \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ est que

$$
\left[\left(\frac{1}{3}+\frac{2}{3N^2}\right)-2b_\epsilon+b_\epsilon^2\left(4-\frac{3}{N}\right)\right] > \frac{(1-2\epsilon)}{N},
$$

puisque $c_1 \neq \mu$. Or, ceci est équivalent à avoir

$$
\begin{aligned}
\left[\left(\frac{1}{3}+\frac{2}{3N^2}\right)-2b_\epsilon+b_\epsilon^2\left(4-\frac{3}{N}\right)\right] &> \frac{(1-2b_\epsilon+b_\epsilon^2)}{N}\\
\iff \left[\left(\frac{1}{3}-\frac{1}{N}+\frac{2}{3N^2}\right)-2b_\epsilon(1-\frac{1}{N})+b_\epsilon^2\left(4-\frac{3}{N}-\frac{1}{N}\right)\right] &> 0\\
\iff \left(\frac{(N^2-3N+2)}{3N^2}-\frac{2b_\epsilon(N-1)}{N}+\frac{4b_\epsilon^2(N-1)}{N}\right) &> 0\\
\iff \left(\frac{N-2}{3N}-2b_\epsilon+4b_\epsilon^2\right) &> 0, \text{ si } N>1.
\end{aligned}
$$

Mais $b_\epsilon < 1/N$ implique que

$$
\begin{aligned}
\left(\frac{N-2}{3N}-2b_\epsilon+4b_\epsilon^2\right) &> \frac{N-2}{3N}-\frac{2}{N}+4b_\epsilon^2\\
&= \frac{N-8}{3N}+4b_\epsilon^2
\end{aligned}
$$

et $(N-8)/3N + 4b_\epsilon^2 > 0$ si $N > 8$. Donc, en autant que $N > 8$, on a bien que $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) > \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. ∎

Annexe B

Démonstrations

Démonstration du lemme 2.2.1 : on procède par induction généralisée sur s. Pour simplifier la présentation, posons $P(j) =$ (le résultat est vrai en dimension j). Remarquons d'abord que peu importe la valeur de s, lorsque $\eta(\mathbf{h}) = 1$, on a que

$$\frac{(N-1)^s}{N}\left(1+(-1)^{\eta(\mathbf{h})}(N-1)^{-\eta(\mathbf{h})+1}\right) = \frac{(N-1)^s}{N}(1-1) = 0.$$

Commençons par démontrer que $P(1)$ et $P(2)$ sont vraies :

- Si $s = 1$, alors pour $a \in [1, \ldots, N-1]$, $h \in L^{\perp}(a)$ si et seulement si $ha = 0 \pmod N$. Si $\eta(h) = 1$, alors $\text{pgcd}(h, N) = 1$ et aucun a ne peut satisfaire la condition. Ainsi, on a bien que $M_N^1(h) = 0$. Si $\eta(h) = 0$, alors $h = 0 \pmod N$ et n'importe quel $a \in [1, \ldots, N-1]$ satisfait la condition. Ainsi, on a bien $M_N^1(h) = (N-1)(1+N-1)/N = N-1$. Donc, $P(1)$ est vraie.
- Si $s = 2$, alors si $\mathbf{h} = \mathbf{0} \pmod N$, pour tout $\mathbf{a} \in [1, \ldots, N-1]^2$, on a que $\mathbf{h}\cdot\mathbf{a} = 0 \pmod N$, c.-à-d., $\mathbf{h} \in L^{\perp}(\mathbf{a})$ et on a bien que $M_N^2(\mathbf{h}) = (N-1)^2(1+(N-1))/N = (N-1)^2$. Si $\mathbf{h} = (h_1, h_2)$ avec $h_1 = 0 \pmod N$ et $h_2 \neq 0 \pmod N$, alors $\eta(\mathbf{h}) = 1$ et pour $\mathbf{a} \in [1, \ldots, N-1]^2$, on a que $\mathbf{h}\cdot\mathbf{a} = 0 \pmod N$ si et seulement si $h_2 a_2 = 0 \pmod N$, ce qui est impossible pour tout $\mathbf{a} \in [1, \ldots, N-1]^2$. Le résultat est le même si $h_1 \neq 0 \pmod N$ et $h_2 = 0 \pmod N$ et donc, $M_N^2(\mathbf{h}) = 0$ si $\eta(\mathbf{h}) = 1$. Si (h_1, h_2) est tel que $h_1 \neq 0 \pmod N$ et $h_2 \neq 0 \pmod N$, alors $\mathbf{h}\cdot\mathbf{a} = 0 \pmod N$ si et seulement si $h_2 a_2 = -h_1 a_1 \pmod N$. Mais puisque $h_1 a_1 \neq 0 \pmod N$ pour chaque $a_1 \in [1, \ldots, N-1]$ et que $\text{pgcd}(h_2, N) = 1$ (car

N est premier), cela signifie qu'il existe un et un seul $a_2 \in [1,\ldots,N-1]$ tel que $h_2a_2 = -h_1a_1 \pmod N$ et donc, $M_N^2(\mathbf{h}) = (N-1)^2(1+1/(N-1))/N = (N-1)$. Donc, $P(2)$ est vraie.

Pour le pas d'induction, on doit montrer que si $P(1), P(2), \ldots, P(s-1)$ sont vraies, alors $P(s)$ est vraie. Examinons les différents cas possibles :

(1) Si $\mathbf{h} = \mathbf{0} \pmod N$, alors pour tout $\mathbf{a} \in [1,\ldots,N-1]^s$, on a que $\mathbf{h}\cdot\mathbf{a} = 0 \pmod N$. Donc, pour ce type de $\mathbf{h}$, $M_N^s(\mathbf{h}) = (N-1)^s = (N-1)^s(1+N-1)/N$.

(2) Supposons que $\eta(\mathbf{h}) = v$, avec $1 \le v < s$. Posons $\tilde{\mathbf{h}} = (h_{i_1},\ldots,h_{i_v})$ avec $h_{i_j} \not\equiv 0 \pmod N$ pour tout $1 \le j \le v$. Pour qu'un vecteur $\mathbf{a} \in [1,\ldots,N-1]^s$ soit tel que $\mathbf{h} \in L^\perp(\mathbf{a})$, une condition nécessaire et suffisante est que les a_{i_j} correspondant aux indices i_j soient tels que $\sum_{j=1}^v h_{i_j}a_{i_j} = 0 \pmod N$ et les autres a_j peuvent valoir n'importe quelle valeur, puisqu'ils seront multipliés par 0. Il y a $M_N^v(\tilde{\mathbf{h}})$ combinaisons possibles pour $(a_{i_1},\ldots,a_{i_v})$. Ainsi, en appliquant $P(v)$, qui est vraie par induction généralisée, on obtient

$$\begin{aligned} M_N^s(\mathbf{h}) &= (N-1)^{s-v}M_N^v(\tilde{\mathbf{h}}) \\ &= (N-1)^{s-v}\frac{(N-1)^v}{N}(1+(-1)^v(N-1)^{-v+1}) \\ &= \frac{(N-1)^s}{N}(1+(-1)^v(N-1)^{-v+1}), \end{aligned}$$

où le terme $(N-1)^{s-v}$ à la première ligne représente le nombre de combinaisons possibles pour $(a_j)_{j:h_j=0 \pmod N}$.

(3) Finalement, si $\eta(\mathbf{h}) = s$, alors on doit avoir

$$h_sa_s = -\sum_{j=1}^{s-1} h_ja_j \pmod N. \tag{B.1}$$

Si $\sum_{j=1}^{s-1} h_ja_j = 0 \pmod N$, alors aucun a_s ne peut satisfaire (B.1). Sinon, exactement un a_s satisfait (B.1). Donc, si on pose $\tilde{\mathbf{h}} = (h_1,\ldots,h_{s-1})$

$$\begin{aligned} M_N^s(\mathbf{h}) &= (N-1)^{s-1} - M_N^{s-1}(\tilde{\mathbf{h}}) \\ &= (N-1)^{s-1} - \frac{(N-1)^{s-1}}{N}\left(1+(-1)^{s-1}(N-1)^{-s+2}\right) \\ &= (N-1)^{s-1}\left[1-\left(\frac{1}{N}+(-1)^{s-1}\frac{(N-1)^{-s+2}}{N}\right)\right] \end{aligned}$$

$$= (N-1)^{s-1}\left(\frac{N-1}{N} - (-1)^{s-1}\frac{(N-1)^{-s+2}}{N}\right)$$
$$= \frac{(N-1)^s}{N}\left(1+(-1)^s(N-1)^{-s+1}\right).$$

Dans ce qui précède, la première égalité nous donne le nombre de vecteurs $(a_1,\ldots,a_{s-1})$ pour lesquels $\sum_{j=1}^{s-1} h_j a_j \not\equiv 0 \pmod N$ et la deuxième suit par hypothèse d'induction, en notant que $\eta(\tilde{\mathbf{h}}) = s-1$.

∎

Démonstration du corollaire 2.2.3 : posons

$$\mathbf{H}_0 = \{\mathbf{h} \in \mathbb{Z}^s : \eta(\mathbf{h}) = 0,\ \mathbf{h} \neq \mathbf{0}\},$$
$$\mathbf{H}_1 = \{\mathbf{h} \in \mathbb{Z}^s : \eta(\mathbf{h}) = 1\},$$

et pour $\mathbf{h} \in \mathbf{H}_0$, $1 \le j \le s$, posons

$$\mathbb{Z}_j^s(\mathbf{h}) = \{(h_1,\ldots,h_{j-1},h_j - l\cdot\mathrm{sgn}(h_j),h_{j+1},\ldots,h_s) : 1 \le l \le N-1\}.$$

L'ensemble $\mathbb{Z}_j^s(\mathbf{h})$ contient les $N-1$ vecteurs obtenus en enlevant (ou additionnant, selon le signe de h_j) l à la composante h_j du vecteur $\mathbf{h} \in \mathbf{H}_0$ qui doit être différente de 0, pour $l = 1,\ldots,N-1$.

On doit d'abord montrer que l'ensemble des $\mathbb{Z}_j^s(\mathbf{h})$, pour tous les $\mathbf{h} \in \mathbf{H}_0$ et les $j \in [1,\ldots,s]$ tels que $h_j \neq 0$ forme une partition de $\mathbf{H}_1$.

Trivialement, on a que pour tout $\mathbf{h} \in \mathbf{H}_0$ et pour tout j t.q. $h_j \neq 0$, chaque élément de $\mathbb{Z}_j^s(\mathbf{h})$ est dans $\mathbf{H}_1$.

Soit $\mathbf{h} = (h_1,\ldots,h_s) \in \mathbf{H}_1$. On doit montrer qu'il existe un unique $\hat{\mathbf{h}} \in \mathbf{H}_0$ et un unique $j \in [1,\ldots,s]$ tels que $\mathbf{h} \in \mathbb{Z}_j^s(\hat{\mathbf{h}})$ et $\hat{h}_j \neq 0$. Prenons j l'unique indice dans $\{1,\ldots,s\}$ tel que $h_j = w \bmod N$, avec $1 \le w \le N-1$.

Si $h_j > 0$, alors posons $\hat{\mathbf{h}} = (h_1,\ldots,h_{j-1},h_j+N-w,h_{j+1},\ldots,h_s)$. Alors $\hat{\mathbf{h}} \in \mathbf{H}_0$ et $\mathbf{h} \in \mathbb{Z}_j^s(\hat{\mathbf{h}})$, en prenant $l = N-w \in \{1,\ldots,N-1\}$. Supposons qu'il existe $\tilde{\mathbf{h}} \in \mathbf{H}_0$ tel que $\mathbf{h} \in \mathbb{Z}_j^s(\tilde{\mathbf{h}})$ et que $\tilde{\mathbf{h}} \neq \hat{\mathbf{h}}$. Alors on doit avoir $\tilde{h}_k = h_k = \hat{h}_k$, si $1 \le k \le s, k \neq j$. Donc, on doit avoir $\tilde{h}_j \neq \hat{h}_j$. Mais alors, on a que pour un certain $m \in \mathbb{Z}$, $m \neq 0$, $\tilde{h}_j = \hat{h}_j + mN$ puisque $\tilde{h}_j = \hat{h}_j = 0 \pmod N$ et donc, $\tilde{h}_j - h_j = \tilde{h}_j - \hat{h}_j + \hat{h}_j - h_j =$

$mN + N - w = N(m+1) - w$. Donc, $\tilde{h}_j - h_j$ est soit inférieur à 0, soit supérieur à N et ainsi, on ne peut trouver $l \in \{1, \ldots, N-1\}$ tel que $\tilde{h}_j - l = h_j$, ce qui contredit que $\mathbf{h} \in \mathbb{Z}_j^s(\tilde{\mathbf{h}})$.

Le cas $h_j < 0$ est traité de façon similaire au cas $h_j > 0$, mais en choisissant $\hat{\mathbf{h}} = (h_1, \ldots, h_{j-1}, h_j - w, h_{j+1}, \ldots, h_s)$.

Ainsi, on a que

$$\begin{aligned} \frac{1}{N-2} \sum_{\mathbf{h} \in \mathbf{H}_1} |\hat{f}(\mathbf{h})|^2 &= \frac{1}{N-2} \sum_{\mathbf{h} \in \mathbf{H}_0} \sum_{\substack{j=1 \\ h_j \neq 0}}^{s} \sum_{l=1}^{N-1} |\hat{f}(\mathbf{h} - l \cdot \mathrm{sgn}(h_j)\mathbf{e}_j)|^2 \\ &\geq \sum_{\mathbf{h} \in \mathbf{H}_0} |\hat{f}(\mathbf{h})|^2, \end{aligned} \tag{B.2}$$

par hypothèse. En utilisant l'inégalité (2.24) que l'on retrouve dans la démonstration de la proposition 2.2.3, on obtient que

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{LR}}(\mathbf{A})) &\leq \frac{1}{N-1} \sum_{\mathbf{0} \neq \mathbf{h} \in \mathbb{Z}^s} |\hat{f}(\mathbf{h})|^2 + \frac{N-2}{N-1} \sum_{\substack{\mathbf{0} \neq \mathbf{h} \in \mathbb{Z}^s \\ \eta(\mathbf{h}) = 0}} |\hat{f}(\mathbf{h})|^2 \\ &\quad - \frac{1}{N-1} \sum_{\substack{\mathbf{0} \neq \mathbf{h} \in \mathbb{Z}^s \\ \eta(\mathbf{h}) = 1}} |\hat{f}(\mathbf{h})|^2 \\ &\leq \frac{1}{N-1} \sum_{\mathbf{0} \neq \mathbf{h} \in \mathbb{Z}^s} |\hat{f}(\mathbf{h})|^2, \end{aligned}$$

en utilisant (B.2). ∎

Démonstration du lemme 3.4.1 : d'abord, si $|I_{\mathbf{h}}| = 0$, alors $\hat{f}(\mathbf{h}) = \int_{[0,1)^s} f(\mathbf{x}) d\mathbf{x} = \mu$. Sinon, on a que

$$\hat{f}(\mathbf{h}) = \hat{f}_{I_{\mathbf{h}}}(\mathbf{h}),$$

par le lemme 3.3.1. Mais

$$\begin{aligned} f_{I_{\mathbf{h}}}(\mathbf{x}) &= \int_{[0,1)^{I_{\mathbf{h}}^c}} f(\mathbf{x}) d\mathbf{x}_{I_{\mathbf{h}}^c} - \sum_{J \subset I_{\mathbf{h}}} f_J(\mathbf{x}) \\ &= \sum_{\mathbf{d} \in D(d)} c(\mathbf{d}) \prod_{j \in I_{\mathbf{h}}} x_j^{d_j} \prod_{k \notin I_{\mathbf{h}}} \frac{1}{d_k + 1} - \sum_{J \subset I_{\mathbf{h}}} f_J(\mathbf{x}), \end{aligned}$$

par simple intégration de f, où $I_{\mathbf{h}}^c$ est le complément de $I_{\mathbf{h}}$ dans S. Donc,

$$\hat{f}_{I_{\mathbf{h}}}(\mathbf{h}) = \sum_{\mathbf{d} \in D(d)} c(\mathbf{d}) \prod_{k \notin I_{\mathbf{h}}} \frac{1}{d_k + 1} \prod_{j \in I_{\mathbf{h}}} \int_0^1 x_j^{d_j} e^{-2\pi\sqrt{-1} h_j x_j} dx_j - \sum_{J \subset I_{\mathbf{h}}} \hat{f}_J(\mathbf{h})$$

$$= \sum_{\mathbf{d}\in D(I_{\mathbf{h}},d)} c(\mathbf{d}) \prod_{k\notin I_{\mathbf{h}}} \frac{1}{d_k+1} \prod_{j\in I_{\mathbf{h}}} \int_0^1 x_j^{d_j} e^{-2\pi\sqrt{-1}h_j x_j} dx_j$$

$$= \sum_{\mathbf{d}\in D(I_{\mathbf{h}},d)} c(\mathbf{d}) \prod_{k\notin I_{\mathbf{h}}} \frac{1}{d_k+1} \prod_{j\in I_{\mathbf{h}}} F(h_j,d_j),$$

la deuxième égalité venant du fait que $\int_0^1 x_j^{d_j} e^{-2\pi\sqrt{-1}h_j x_j} dx_j$ vaut 0 si $d_j = 0$ et que $\hat{f}_J(\mathbf{h}) = 0$ si $J \neq I_{\mathbf{h}}$.

Il reste à montrer que

$$F(h,d) = \sum_{w=1}^{d} \left(\frac{\sqrt{-1}}{2\pi h}\right)^w (-1)^{w+1} \prod_{l=0}^{w-2} (d-l).$$

On peut procéder par induction sur d. Pour $d = 1$, on a que

$$\begin{aligned} F(h,1) &= \int_0^1 x e^{-2\pi\sqrt{-1}hx} dx \\ &= \left[x\frac{\sin(2\pi hx)}{2\pi h}\Big|_0^1 + \frac{\cos(2\pi hx)}{(2\pi h)^2}\Big|_0^1 - \sqrt{-1}\left(-x\frac{\cos(2\pi hx)}{2\pi h}\Big|_0^1 \right.\right. \\ &\quad \left.\left. + \frac{\sin(2\pi hx)}{(2\pi h)^2}\right)\Big|_0^1 \right] \\ &= \frac{\sqrt{-1}}{2\pi h}. \end{aligned}$$

Supposons que le résultat est vrai pour $d - 1 \geq 1$. En intégrant par parties, on obtient

$$\begin{aligned} F(h,d) &= -d\sqrt{-1}\frac{F(h,d-1)}{2\pi h} + x^d \frac{\sin(2\pi hx)}{2\pi h}\Big|_0^1 + \sqrt{-1}\, x^d \frac{\cos(2\pi hx)}{2\pi h}\Big|_0^1 \\ &= -\frac{d\sqrt{-1}}{2\pi h}\left[\sum_{w=1}^{d-1}\left(\frac{\sqrt{-1}}{2\pi h}\right)^w (-1)^{w+1}\prod_{l=0}^{w-2}(d-1-l)\right] + \frac{\sqrt{-1}}{2\pi h} \\ &= \left[\sum_{w=1}^{d-1}\left(\frac{\sqrt{-1}}{2\pi h}\right)^{w+1} (-1)^{w+2}\prod_{l=0}^{w-1}(d-l)\right] + \frac{\sqrt{-1}}{2\pi h} \\ &= \left[\sum_{w=1}^{d}\left(\frac{\sqrt{-1}}{2\pi h}\right)^{w} (-1)^{w+1}\prod_{l=0}^{w-2}(d-l)\right]. \end{aligned}$$

Finalement, remarquons que si $|I_{\mathbf{h}}| > d$, alors aucun coefficient $c(\mathbf{d})$ ne peut être dans $D(I_{\mathbf{h}}, d)$ et ainsi $\hat{f}(\mathbf{h}) = 0$. ∎

Démonstration de la proposition 3.4.1 : supposons que $I = \{i_1, \ldots, i_t\}$. On doit écrire σ_I^2 sous une forme où $\mathbf{h}$ est isolé :

$$\sigma_I^2 = \sum_{\mathbf{h}\in\mathbf{Z}_I^*} |\hat{f}(\mathbf{h})|^2$$

$$
\begin{aligned}
&= \sum_{\mathbf{h}\in\mathbf{Z}_I^*} \left| \sum_{\mathbf{d}\in D(I,d)} c(\mathbf{d}) \left(\prod_{k\notin I} \frac{1}{d_k+1} \right) \prod_{j\in I} F(h_j,d_j) \right|^2 \\
&= \sum_{\mathbf{h}\in\mathbf{Z}_I^*} \left| \sum_{\mathbf{d}\in D(I,d)} c(\mathbf{d}) \left(\prod_{k\notin I} \frac{1}{d_k+1} \right) \left[\sum_{v_1=1}^{d_{i_1}} \cdots \sum_{v_t=1}^{d_{i_t}} (2\pi)^{-\sum_{j=1}^t v_j} (-1)^{3\sum_{j=1}^t v_j/2} \right.\right. \\
&\quad \left.\left. \prod_{j=1}^t h_{i_j}^{-v_j} \prod_{l=0}^{v_1-2} (d_{i_1}-l) \cdots \prod_{l=0}^{v_t-2} (d_{i_t}-l) \right] \right|^2 \\
&= \sum_{\mathbf{h}\in\mathbf{Z}_I^*} \left[\sum_{z=0}^{1} \left(\sum_{\mathbf{d},\mathbf{d}'\in D(I,d)} c(\mathbf{d})c(\mathbf{d}') \left(\prod_{k\notin I} \frac{1}{(d_k+1)(d_k'+1)} \right) \sum_{v_1=1}^{d_{i_1}} \cdots \sum_{\substack{v_t=1 \\ \sum_{j=1}^t v_j = z \bmod 2}}^{d_{i_t}} \right.\right. \\
&\quad \sum_{w_1=1}^{d'_{i_1}} \cdots \sum_{\substack{w_t=1 \\ \sum_{j=1}^t w_j = z \bmod 2}}^{d'_{i_t}} (2\pi)^{-\sum_{j=1}^t v_j+w_j} (-1)^{\lfloor \sum_{j=1}^t \frac{v_j}{2} \rfloor + \lfloor \sum_{j=1}^t \frac{w_j}{2} \rfloor + \sum_{j=1}^t v_j + w_j} \\
&\quad \left.\left. \left. \prod_{j=1}^t h_{i_j}^{-v_j-w_j} \prod_{j=1}^t \left(\prod_{l=0}^{v_j-2} (d_{i_j}-l) \prod_{l=0}^{w_j-2} (d'_{i_j}-l) \right) \right) \right] \\
&= \sum_{\mathbf{d},\mathbf{d}'\in D(I,d)} c(\mathbf{d})c(\mathbf{d}') \left(\prod_{k\notin I} \frac{1}{(d_k+1)(d_k'+1)} \right) \qquad \text{(B.3)} \\
&\quad \sum_{\alpha_{i_1}=2}^{d_{i_1}+d'_{i_1}} \cdots \sum_{\substack{\alpha_{i_t}=2 \\ \sum_{j=1}^t \alpha_{i_j} = 0 \bmod 2}}^{d_{i_t}+d'_{i_t}} g(\alpha_I,\mathbf{d},\mathbf{d}') \left(\sum_{\mathbf{h}\in\mathbf{Z}_I^*} \prod_{j=1}^t h_{i_j}^{-\alpha_{i_j}} \right) \\
&= \sum_{\alpha_I\in A_{I,d}} \gamma_{\alpha_I}(I) \left(\sum_{\mathbf{h}\in\mathbf{Z}_I^*} \prod_{j\in I} h_j^{-\alpha_j} \right) \qquad \text{(B.4)} \\
&= \sum_{\alpha_I\in A_{I,d}^+} \gamma_{\alpha_I}(I) 2^t \prod_{j\in I} \zeta(\alpha_j)
\end{aligned}
$$

où

$$
\begin{aligned}
\gamma_{\alpha_I}(I) &= \sum_{\substack{\mathbf{d},\mathbf{d}'\in D(I,d) \\ d_{i_j}+d'_{i_j}\geq \alpha_{i_j},\ j=1,\ldots,t}} c(\mathbf{d})c(\mathbf{d}') \left(\prod_{k\notin I} \frac{1}{(d_k+1)(d_k'+1)} \right) g(\alpha_I,\mathbf{d},\mathbf{d}') \\
g(\alpha_I,\mathbf{d},\mathbf{d}') &= \prod_{j=1}^t \left[\left(\frac{\sqrt{-1}}{2\pi} \right)^{\alpha_{i_j}} \sum_{v_j=\max(1,\alpha_{i_j}-d'_{i_j})}^{\min(\alpha_{i_j}-1,d_{i_j})} (-1)^{v_j} \prod_{l=0}^{v_j-2} (d_{i_j}-l) \prod_{l=0}^{\alpha_{i_j}-v_j-2} (d'_{i_j}-l) \right].
\end{aligned}
$$

Dans la série d'égalités précédentes, la deuxième vient du lemme 3.4.1 ; la troisième est obtenue en réécrivant le produit des $F(h_j,d_j)$ de façon à ce que les h_j soient regroupés dans un seul produit ; dans la quatrième, $z=0$ correspond à $|\mathrm{Re}(\hat{f}(\mathbf{h}))|^2$ et $z=1$

à $|\mathrm{Im}(\hat{f}(\mathbf{h}))|^2$ et pour chacun de ces deux termes, le carré est développé de façon à ce que les h_{i_j} soient regroupés dans un même produit ; la cinquième est obtenue en posant $\alpha_{i_j} = v_j + w_j$ et en passant la somme sur les $\mathbf{h}$ à l'intérieur ; pour obtenir la fonction $g(\cdot)$, on regroupe les termes à l'intérieur de la somme sur les v_j et les w_j, en remarquant que

$$\{(v_j, w_j)_{j=1}^t : v_j + w_j = \alpha_{i_j}, 1 \le v_j \le d_{i_j}, 1 \le w_j \le d'_{i_j}, (\sum_{j=1}^t v_j = \sum_{j=1}^t w_j) \bmod 2\}$$

$$= \{(v_j, \alpha_{i_j} - v_j)_{j=1}^t : 1 \le v_{i_j} \le d_{i_j}, 1 \le \alpha_{i_j} - v_j \le d'_{i_j}, \sum_{j=1}^t \alpha_{i_j} = 0 \bmod 2\}$$

et que

$$(-1)^{\lfloor \sum_{j=1}^t \frac{v_j}{2} \rfloor + \lfloor \sum_{j=1}^t \frac{w_j}{2} \rfloor + \sum_{j=1}^t v_j + w_j} = \sqrt{-1}^{\sum_{j=1}^t \alpha_{i_j}} (-1)^{\sum_{j=1}^t v_j},$$

lorsque $\sum_{j=1}^t \alpha_{i_j} = 0 \bmod 2$; la sixième égalité est obtenue en interchangeant l'ordre de sommation afin de regrouper les coefficients associés à un vecteur α_I donné et en remarquant que $\alpha_{i_j} \le d_{i_j} + d'_{i_j} \le 2(d - |I| + 1)$ puisque $\mathbf{d}$ et $\mathbf{d}'$ sont dans $D(I, d)$; la septième tient puisque dès qu'un des α_{i_j} est impair, la somme sur les $\mathbf{h}$ s'annule.

Pour obtenir $\sigma^2_{I,\mathrm{LR}}$, on utilise la formule (B.4), mais la somme sur les $\mathbf{h}$ est restreinte aux éléments du dual, c.-à-d., on obtient que

$$\sigma^2_{I,\mathrm{LR}} = \sum_{\alpha_I \in A_{I,d}} \gamma_{\alpha_I}(I) P_{\alpha_I}(I).$$

∎

Démonstration du lemme 3.4.8 : on a que

$$\frac{1}{(N-1)^s} \sum_{\mathbf{a} \in [1 \ldots N-1]^s} P_{\alpha_I}(I, \mathbf{a}) = \frac{1}{(N-1)^s} \sum_{\mathbf{h} \in \mathbf{Z}_I^*} M_N^s(\mathbf{h}) \prod_{j \in I} h_j^{-\alpha_j},$$

où, rappelons-le, $M_N^s(\mathbf{h})$ est le nombre de vecteurs $\mathbf{a}$ dans $[1, \ldots, N-1]^s$ tels que $\mathbf{h} \in L^\perp(\mathbf{a})$ et $L^\perp(\mathbf{a})$ est le réseau dual à la règle dont le vecteur générateur est $\mathbf{a}$. Or, par le lemme 2.2.1 du chapitre 2, on sait que

$$M_N^s(\mathbf{h}) = \frac{(N-1)^s}{N} \left(1 + (-1)^{\eta(\mathbf{h})} (N-1)^{-\eta(\mathbf{h})+1}\right)$$

où

$$\eta(\mathbf{h}) = \sum_{j=1}^s \mathbf{1}_{\{h_j \neq 0 \bmod N\}}.$$

Posons $\Gamma(\mathbf{h}) = \{j : h_j \neq 0 \bmod N\}$ (donc, $\eta(\mathbf{h}) = |\Gamma(\mathbf{h})|$). Ainsi,

$$
\begin{aligned}
&\sum_{\mathbf{h}\in\mathbb{Z}_I^*} M_N^s(\mathbf{h}) \prod_{j\in I} h_j^{-\alpha_j} \\
&= \sum_{k=0}^{t} \left(\sum_{\substack{\mathbf{h}\in\mathbb{Z}_I^* \\ \eta(\mathbf{h})=k}} \frac{(N-1)^s}{N} \left(1 + (-1)^k (N-1)^{-k+1}\right) \prod_{j\in I} h_j^{-\alpha_j} \right) \\
&= \frac{(N-1)^s}{N} \sum_{k=0}^{t} \sum_{J\subseteq I, |J|=k} \sum_{\substack{\mathbf{h}\in\mathbb{Z}_I^* \\ \Gamma(\mathbf{h})=J}} \left[\left(1 + (-1)^k (N-1)^{-k+1}\right) \prod_{j\in I} h_j^{-\alpha_j} \right] \\
&= \frac{(N-1)^s}{N-1} \sum_{k=0}^{t} \sum_{J\subseteq I, |J|=k} \sum_{\substack{\mathbf{h}\in\mathbb{Z}_I^* \\ \Gamma(\mathbf{h})=J}} \left[\left(1 - \frac{1}{N} + \frac{(-1)^k}{N(N-1)^{k-2}}\right) \prod_{j\in I} h_j^{-\alpha_j} \right] \\
&= \frac{(N-1)^s}{N-1} \sum_{k=0}^{t} \sum_{J\subseteq I, |J|=k} \sum_{\substack{\mathbf{h}\in\mathbb{Z}_I^* \\ \Gamma(\mathbf{h})=J}} \prod_{j\in I} h_j^{-\alpha_j} \\
&\quad + \frac{(N-1)^s}{N-1} \sum_{\mathbf{h}\in\mathbb{Z}_I^*, \eta(\mathbf{h})=0} \left(-\frac{1}{N} + \frac{(N-1)^2}{N}\right) \prod_{j\in I} h_j^{-\alpha_j} \\
&\quad - \frac{(N-1)^s}{N-1} \sum_{J\subseteq I, |J|=1} \sum_{\mathbf{h}\in\mathbb{Z}_I^*, \Gamma(\mathbf{h})=J} \left(\frac{1}{N} + \frac{N-1}{N}\right) \prod_{j\in I} h_j^{-\alpha_j} \\
&\quad + \sum_{k=2}^{t} \sum_{J\subseteq I, |J|=k} \sum_{\mathbf{h}\in\mathbb{Z}_I^*, \Gamma(\mathbf{h})=J} \left(-\frac{1}{N} + \frac{(-1)^k}{N(N-1)^{k-2}}\right) \prod_{j\in I} h_j^{-\alpha_j} \\
&\leq \frac{(N-1)^s}{N-1} \left(1 + \left(\frac{-1 + (N-1)^2}{N^{1+\sum_{j\in I}\alpha_j}}\right)\right) \sum_{\mathbf{h}\in\mathbb{Z}_I^*} \prod_{j\in I} h_j^{-\alpha_j}, \\
&\leq \frac{(N-1)^s}{N-1} \left(1 + \left(\frac{N-2}{N^{2t}}\right)\right) 2^t \prod_{j\in I} \zeta(\alpha_j).
\end{aligned}
$$

Pour obtenir la première égalité, on partitionne $\mathbb{Z}_I^*$ selon le nombre de composantes de $\mathbf{h} \in \mathbb{Z}_I^*$ qui sont différentes de 0, modulo N ; la deuxième est obtenue en partitionnant $\{\mathbf{h} \in \mathbb{Z}_I^* : \eta(\mathbf{h}) = k\}$ selon les sous-ensembles $\Gamma(\mathbf{h})$; on multiplie par $(N-1)/(N-1)$ pour arriver à la troisième ; la première inégalité est obtenue en remarquant que 1) le quatrième terme de l'expression précédente (comme le troisième) est négatif, puisque $1/(N(N-1)^{k-2}) \leq 1/N$ pour tout $k \geq 2$; 2) la somme sur $\{\mathbf{h} \in \mathbb{Z}_I^* : \eta(\mathbf{h}) = 0\}$ que l'on retrouve dans le deuxième terme de l'expression précédente revient à une somme sur $\{(Nh_{i_1}, \ldots, Nh_{i_t}) : \mathbf{h} \in \mathbb{Z}_I^*\}$. ∎

Démonstration de la proposition 3.5.1 : pour simplifier la notation, nous allons poser $l_I^*(N) = l_I^*$ et $l_t^*(N) = l_t^*$. En définissant les $w(\mathbf{h})$ comme dans l'énoncé de la

proposition, on a que

$$D'(P_n) = \sup_{\mathbf{0}\neq\mathbf{h}\in L^\perp} w(\mathbf{h}) = \max\left(\sup_{\substack{\mathbf{h}\in L^\perp, \\ I_\mathbf{h}\in H(t_1,\ldots,t_d,d)}} l^*_{|I_\mathbf{h}|}/\|\mathbf{h}\|_2, \sup_{\substack{\mathbf{h}\in L^\perp \\ I_\mathbf{h}\notin H(t_1,\ldots,t_d,d), r(I_\mathbf{h})\leq t_1}} l^*_{r(I_\mathbf{h})}/\|\mathbf{h}\|_2 \right).$$

Or,

$$\begin{aligned} \sup_{\mathbf{h}\in L^\perp: I_\mathbf{h}\in H(t_1,\ldots,t_d,d)} l^*_{|I_\mathbf{h}|}/\|\mathbf{h}\|_2 &= \sup_{\mathbf{h}\in L^\perp: I_\mathbf{h}\in H(t_1,\ldots,t_d,d)} l^*_{|I_\mathbf{h}|}/l_{I_\mathbf{h}} \\ &= \max_{I\in H(t_1,\ldots,t_d,d)} l^*_I/l_I \\ &= \max\left(\max_{2\leq u\leq d} \max_{I\in S(t_u,u)} l^*_I/l_I, \max_{2\leq j\leq t_1} l^*_j/l_j \right). \end{aligned} \quad \text{(B.5)}$$

Dans ce qui précède, la première égalité vient du fait que $\|\mathbf{h}\|_2 = \|\mathbf{h}_{I_\mathbf{h}}\|_2 \geq l_{I_\mathbf{h}}$ (par définition, puisque $\mathbf{h}_{I_\mathbf{h}} \in L^\perp_{I_\mathbf{h}}$), où $\mathbf{h}_{I_\mathbf{h}}$ représente la projection de $\mathbf{h}$ sur les dimensions où ses composantes sont non nulles. La troisième égalité découle du fait que P_N est stationnaire dans la dimension.

Maintenant, si $\mathbf{h} \in L^\perp$ avec $I_\mathbf{h} = \{i_1, \ldots, i_t\}$ n'est pas dans $H(t_1, \ldots, t_d, d)$, mais que $r(I_\mathbf{h}) \leq t_1$, alors $\mathbf{h}_J \in L^\perp_J$, où $J = \{i_1, i_1+1, \ldots, i_1 + r(I_\mathbf{h}) - 1\}$ et donc,

$$\begin{aligned} \sup_{\mathbf{h}\in L^\perp: I_\mathbf{h}\notin H(t_1,\ldots,t_d,d), r(I_\mathbf{h})\leq t_1} l^*_{r(I_\mathbf{h})}/\|\mathbf{h}\|_2 &= \sup_{\mathbf{h}\in L^\perp: I_\mathbf{h}\notin H(t_1,\ldots,t_d,d), r(I_\mathbf{h})\leq t_1} l^*_{r(I_\mathbf{h})}/l_{r(I_\mathbf{h})}, \\ &= \max_{1\leq j\leq t_1} l_j/l^*_j, \\ &= \max_{2\leq j\leq t_1} l_j/l^*_j, \end{aligned} \quad \text{(B.6)}$$

où la première égalité vient du fait que $\|\mathbf{h}\|_2 = \|\mathbf{h}_J\|_2 \geq l_J = l_{r(I_\mathbf{h})}$, puisque P_N est stationnaire dans la dimension ; la deuxième vient du fait que $r(I_\mathbf{h}) \leq t_1$ et la troisième tient puisque $l^*_1/l_1 = 1 \leq l^*_j/l_j$ pour tout j dans $\{2, \ldots, t_1\}$. En combinant (B.5) et (B.6), on obtient bien que

$$D'(P_N) = \left[\min\left(\min_{2\leq u\leq d} \min_{I\in S(t_u,u)} l_I/l^*_I, \min_{2\leq j\leq t_1} l_j/l^*_j \right) \right]^{-1}.$$

∎

Démonstration du lemme 4.2.1 : si on travaille dans l'espace des polynômes $\mathbb{F}_2[z]/(P)$ tel qu'expliqué à la sous-section 4.1.4 et qu'on utilise la représentation (4.3) pour P_N, alors la version polynômiale de chaque $\mathbf{x}_i$ est définie par

$$\mathbf{x}_i(z) = \left(\frac{p^i_0(z)}{P(z)}, \frac{p^i_v(z)}{P(z)} \ldots, \frac{p^i_{v(s-1)}(z)}{P(z)} \right)$$

où $p_0^i(z)$ est le polynôme $p_0(z) = \sum_{w=1}^{m} c_{0,w} z^{m-w}$ obtenu en posant le vecteur $\mathbf{s}_0 = (\xi_w)_{w=0}^{m-1}$ égal à $\mathbf{s}_0^i$ (le i^e élément de $\mathbb{F}_2^m$) dans l'équation (4.4) définissant les $c_{0,w}$, pour $1 \le w \le m$, et $p_{(vj)}^i(z) = z^{vj} p_0^i(z) \bmod (P(z), 2)$, pour $j = 0, \ldots, s-1$, en utilisant la récurrence (4.5).

Ainsi,

$$\sum_{j=1}^{s} h_j(z) x_{ij}(z) = \frac{p_0^i(z)}{P(z)} \sum_{j=1}^{s} h_j(z) z^{v(j-1)} \bmod (P(z), 2)$$
$$= \begin{cases} 0 & \text{si } \mathbf{h}(z) \in \mathcal{L}_s^*, \\ (p_0^i(z) q(z))/P(z) & \text{sinon,} \end{cases}$$

où $q(z) = \sum_{j=1}^{s} h_j(z) z^{(j-1)v} \bmod (P(z), 2) \neq 0$, lorsque $\mathbf{h}(z) \notin \mathcal{L}_s^*$.

Étant donné que le résultat est présenté dans ce contexte, revenons maintenant à la représentation de $\mathbf{h}$ dans $\mathbb{N}^s$ et de $\mathbf{x}$ dans $[0,1)^s$. Il suffit de montrer que $p_0^i(2) = 0 \pmod 2$ pour exactement la moitié des éléments dans $\mathbb{F}_2^m$. Or, par définition de p_n, on a que $p_0(2) = c_{0,m} \pmod 2$ et $c_{0,m} = \sum_{w=1}^{m} a_{m-w} \xi_{w-1}$, où les a_j sont les coefficients de la récurrence (4.1). Posons $J = \{w : a_{m-w} = 1, 1 \le w \le m\}$ et définissons $\alpha_J(\mathbf{s}_0^i) = \sum_{w=1}^{m} \mathbf{1}_{(w \in J) \cap (\xi_{w-1}^i = 1)}$, où $\mathbf{s}_0^i = (\xi_w^i)_{w=0}^{m-1}$ représente le i^e élément de $\mathbb{F}_2^m$. On a donc que $p_0^i(2) = 0 \pmod 2$ si et seulement si $\alpha_J(\mathbf{s}_0^i) = 0 \pmod 2$. Or, les ξ_{w-1}^i avec $w \notin J$ sont indépendants de $p_0^i(2)$ et parmi les 2^J combinaisons possibles pour $(\xi_{w-1}^i)_{w \in J}$, la moitié est telle que $\alpha_J(\mathbf{s}_0^i) = 0 \pmod 2$. Ainsi, en posant $G = \{i : p_0^i(2) = 0 \pmod 2\}$, on a que $|G| = 2^{|J|-1} 2^{m-|J|} = 2^{m-1}$, et donc,

$$\sum_{i=1}^{N} (-1)^{\mathbf{h} \odot \mathbf{x}_i} = \begin{cases} \sum_{i=1}^{N} (-1)^0 = N & \text{si } \mathbf{h} \in \mathcal{L}_s^* \\ \sum_{i : i \in G} 1 + \sum_{i : i \notin G} (-1) = 0 & \text{sinon,} \end{cases}$$

puisque $|\{i : p_0^i(2) = 1 \bmod 2\}| = |\{i : i \notin G\}| = 2^m - |G| = 2^{m-1}$. ∎

Démonstration du lemme 4.3.1 : par hypothèse que $\tilde{\mathbf{k}}_u$ induit une équidistribution pour tout $u \in J$, on a que

$$\sum_{i=1}^{N} \mathbf{1}_{\{\mathbf{x}_i(\mathbf{g}_J) = b\}} = \frac{N}{2^\gamma},$$

pour toute chaîne b de γ bits. Il suffit donc de montrer que si $\mathbf{x}_{i_1}(\mathbf{g}_J) = \mathbf{x}_{i_2}(\mathbf{g}_J) = b$, pour $1 \le i_1, i_2 \le N$, alors $(\sum_{j \in J} x_{i_1,j,k_j} = \sum_{j \in J} x_{i_2,j,k_j}) \bmod 2$, c.-à-d., $(x_{i_1,j,k_j})_{j \in J}$ et

$(x_{i_2,j,k_j})_{j\in J}$ ont la même parité. Or, par hypothèse que $\mathbf{k}_I$ n'induit pas d'équidistribution, on peut trouver une chaîne $\mathbf{y} = y_1 \dots y_{|J|}$ de $|J|$ bits telle qu'aucun $\mathbf{x}_i$ ne satisfait

$$\mathbf{x}_i(\mathbf{g}_J) = b \tag{B.7}$$

$$x_{i,i_t,k_{i_t}} = z_t \text{ pour tout } i_t \in J, \tag{B.8}$$

lorsque $\mathbf{z} = \mathbf{y}$. Cela implique que pour $t = 1, \dots, |J|$, les chaînes de la forme $\mathbf{z} = y_1 \dots 1 - y_t \dots y_{|J|}$, doivent être telles que $N/2^{\gamma+|J|-1}$ points $\mathbf{x}_i$ respectent (B.7) et (B.8) : sinon, $\tilde{\mathbf{k}}_{i_t}$ n'induirait pas une équidistribution. De façon générale, nous allons montrer que l'ensemble des chaînes de $|J|$ bits peut être séparé en deux sous-ensembles comptant $2^{|J|-1}$ chaînes chacun et tels que 1) un sous-ensemble contient toutes les chaînes $\mathbf{z}$ telles que $N/2^{\gamma+|J|-1}$ points $\mathbf{x}_i$ satisfont (B.7) et (B.8) pour chaque $\mathbf{z}$ et l'autre sous-ensemble contient les chaînes $\mathbf{z}$ pour lesquelles aucun $\mathbf{x}_i$ ne satisfait ces conditions ; 2) les chaînes dans chaque sous-ensemble ont la même parité. Pour faire cela, nous allons démontrer par induction que pour une chaîne $\mathbf{z}$ de $|J|$ bits donnée, si on pose

$$n_{\mathbf{z}} = \sum_{j\in J} \mathbf{1}_{\{z_j = y_j\}},$$

alors $n_{\mathbf{z}} = 0 \bmod 2$ implique qu'aucun point $\mathbf{x}_i$ ne satisfait (B.7) et (B.8) et que $n_{\mathbf{z}} = 1 \bmod 2$ implique qu'exactement $N/2^{\gamma+|J|-1}$ points $\mathbf{x}_i$ satisfont (B.7) et (B.8). Nous avons déjà traité le cas où $n_{\mathbf{z}} = 0, 1$. Supposons que le résultat tient lorsque $n_{\mathbf{z}} = j - 1$, pour $j \geq 2$. Si $n_{\mathbf{z}} = j$, considérons une chaîne $\tilde{\mathbf{z}}$ qui diffère de $\mathbf{z}$ par un bit, et ce bit $\tilde{z}_p$ est tel que $\tilde{z}_p \neq z_p = y_p$. Donc, $n_{\tilde{\mathbf{z}}} = j - 1$ et si $n_{\tilde{\mathbf{z}}} = 0 \bmod 2$, alors par hypothèse d'induction, cela signifie que $\mathbf{z}$ doit être tel que $N/2^{\gamma+|J|-1}$ points $\mathbf{x}_i$ satisfont (B.7) et (B.8), car sinon, $\tilde{\mathbf{k}}_p$ n'induirait pas une équidistribution. Si $n_{\tilde{\mathbf{z}}} = 1 \bmod 2$, alors par hypothèse d'induction, $\mathbf{z}$ doit être tel qu'aucun point $\mathbf{x}_i$ ne satisfait (B.7) et (B.8), car sinon, on aurait une boîte de volume $2^{-\gamma-|J|+1}$ contenant $2N/2^{\gamma+|J|-1}$ points, contredisant le fait que chaque $\tilde{\mathbf{k}}_u$ induit une équidistribution. Finalement, on a bien que $\{\mathbf{z} : n_{\mathbf{z}} = 0 \bmod 2\} = 2^{|J|-1}$. ∎

Démonstration de la proposition 4.4.1 : en définissant les $w(\mathbf{h})$ par (4.11), on a que

$$D^*(P_n) = \sup_{\mathbf{0}\neq\mathbf{h}\in\mathcal{L}_s^*} w(\mathbf{h})$$

$$= \max\left(\sup_{\substack{\mathbf{h}\in\mathcal{L}_s^*,\\ I_\mathbf{h}\in H(w_1,\ldots,w_d,d)}} \ell^*_{|I_\mathbf{h}|} - \lg\|\mathbf{h}\|, \sup_{\substack{\mathbf{h}\in\mathcal{L}_s^*, r(I_\mathbf{h})\le w_1\\ I_\mathbf{h}\notin H(w_1,\ldots,w_d,d)}} l^*_{r(I_\mathbf{h})} - \lg\|\mathbf{h}\| \right).$$

Or,

$$\sup_{\mathbf{h}\in\mathcal{L}_s^*: I_\mathbf{h}\in H(w_1,\ldots,w_d,d)} \ell^*_{|I_\mathbf{h}|} - \lg\|\mathbf{h}\| \tag{B.9}$$

$$= \sup_{\mathbf{h}\in\mathcal{L}_s^*: I_\mathbf{h}\in H(w_1,\ldots,w_d,d)} \ell^*_{|I_\mathbf{h}|} - \ell_{|I_\mathbf{h}|},$$

$$= \max_{I\in H(w_1,\ldots,w_d,d)} \ell^*_{|I|} - \ell_I,$$

$$= \max\left(\max_{2\le u\le d} \max_{I\in S(w_u,u)} \ell^*_{|I|} - \ell_I, \max_{1\le j\le w_1} \ell^*_j - \ell_j \right). \tag{B.10}$$

Dans ce qui précède, la première égalité vient du fait que par définition de $\ell_{I_\mathbf{h}}$, $\|\mathbf{h}\| \ge 2^{\ell_{I_\mathbf{h}}}$ et la troisième découle du fait que P_N est stationnaire dans la dimension.

Maintenant, si $\mathbf{h} \in \mathcal{L}_s^*$ avec $I_\mathbf{h} = \{i_1, \ldots, i_t\}$ n'est pas dans $H(w_1, \ldots, w_d, d)$, mais que $r(I_\mathbf{h}) \le w_1$, alors $\mathbf{h}_J \in \mathcal{L}_j^*$, où $J = \{i_1, i_1+1, \ldots, i_1 + r(I_\mathbf{h}) - 1\}$ et donc,

$$\sup_{\mathbf{h}\in\mathcal{L}_s^*: I_\mathbf{h}\notin H(w_1,\ldots,w_d,d), r(I_\mathbf{h})\le w_1} \ell^*_{r(I_\mathbf{h})} - \lg\|\mathbf{h}\| = \sup_{\mathbf{h}\in\mathcal{L}_s^*: I_\mathbf{h}\notin H(w_1,\ldots,w_d,d), r(I_\mathbf{h})\le w_1} \ell^*_{r(I_\mathbf{h})} - \ell_{r(I_\mathbf{h})},$$

$$= \max_{1\le j\le w_1} \ell^*_j - \ell_j, \tag{B.11}$$

où la première égalité vient du fait que $\|\mathbf{h}\| \ge 2^{\ell_J} = 2^{\ell_{r(I_\mathbf{h})}}$, puisque P_N est stationnaire dans la dimension et la deuxième, du fait que $r(I_\mathbf{h}) \le w_1$. En combinant (B.10) et (B.11), on obtient bien que

$$D'(P_N) = \max\left(\max_{1\le j\le w_1} \ell^*_j - \ell_j, \max_{2\le u\le d} \max_{I\in S(w_u,u)} \ell^*_{|I|} - \ell_I \right).$$

■

Démonstration de la proposition 4.4.2 : on sait que

$$\max_{I\in H(w_1,\ldots,w_d,d)} (\ell^*_I - \ell_I) = \max\left(\max_{1\le u\le w_1} (\ell^*_u - \ell_u), \max_{2\le u\le d} \max_{I\in S(w_u,u)} (\ell^*_I - \ell_I) \right) = \Delta, \tag{B.12}$$

car P_N est stationnaire dans la dimension. De plus, on sait par la proposition 4.3.3 que

$$q_I \le m - |I| + 1 - \ell_I \le m - |I| + 1 - \ell^*_I + (\ell^*_I - \ell_I). \tag{B.13}$$

En combinant (B.13) et (B.12), on obtient bien que

$$\begin{aligned}\max_{I\in H(w_1,\ldots,w_d,d)} q_I &\leq \max\Big[\max_{1\leq u\leq w_1}(m-u+1-\lfloor m/u\rfloor+\Delta),\\ &\qquad \max_{2\leq u\leq d}\max_{I\in S(w_u,u)}(m-u+1-\lfloor m/u\rfloor+\Delta)\Big],\\ &= m+1+\Delta+\max_{1\leq u\leq w_1}(-u-\lfloor m/u\rfloor),\end{aligned}$$

car $w_1 \geq w_u$ implique que $d \leq w_1$ puisque d doit être inférieur à w_d par définition.

Donc, on doit calculer

$$\max_{1\leq p\leq w_1} g(p),$$

où $g(p) = -p - \lfloor m/p \rfloor$ est définie sur $\mathbb{N}$. Or,

$$g(p) = \lceil -p - m/p \rceil = \lceil \tilde{g}(p) \rceil,$$

en posant $\tilde{g}(p) = -p - m/p$, qui est définie sur $[0,\infty)$. Soit $p^* \in [0,\infty)$ tel que $p^* = \text{argmax}\, \tilde{g}(p)$. On peut montrer que p^* est l'unique maximum de $\tilde{g}(p)$. En effet, on a que

$$\frac{\partial \tilde{g}(p)}{\partial p} = -1 + \frac{m}{p^2} = 0 \Leftrightarrow \frac{m}{p^2} = 1 \Leftrightarrow p^2 = m \Leftrightarrow p = \sqrt{m}$$

et si $m > 1$, alors

$$\begin{aligned}-1+\frac{m}{p^2} &> -1+1=0, \qquad \text{si } 1 \leq p < \sqrt{m},\\ -1+\frac{m}{p^2} &< -1+1=0, \qquad \text{si } p > \sqrt{m}.\end{aligned}$$

Donc, $\tilde{g}$ est monotone croissante sur $[1,\sqrt{m})$, atteint son maximum en $p^* = \sqrt{m}$, puis est monotone décroissante sur $(\sqrt{m},\infty)$.

Ainsi, $\text{argmax}\, g(p) \in \{\lfloor p^* \rfloor, \lfloor p^* \rfloor + 1\}$, car s'il existe $v \in \mathbb{N}$ tel que $g(v) > g(\lfloor p^* \rfloor)$ et $g(v) > g(\lfloor p^* \rfloor + 1)$, cela est équivalent à dire que

$$\begin{aligned}\lceil \tilde{g}(v) \rceil &> \max(\lceil \tilde{g}(\lfloor p^* \rfloor) \rceil, \lceil \tilde{g}(\lfloor p^* \rfloor + 1) \rceil)\\ \Rightarrow \tilde{g}(v) &> \max(\tilde{g}(\lfloor p^* \rfloor), \tilde{g}(\lfloor p^* \rfloor + 1)), \qquad \text{(B.14)}\end{aligned}$$

mais soit $v < \lfloor p^* \rfloor$, soit $v > \lfloor p^* \rfloor + 1$ et $\tilde{g}$ est monotone croissante sur $[1, p^*)$ et monotone décroissante sur $(p^*, \infty]$, ce qui contredit (B.14).

Puisque $w_1 \geq \lfloor\sqrt{m}\rfloor + 1$ par hypothèse, on a que

$$\begin{aligned}
\max_{1\leq p\leq w_1} g(p) &= \max(g(\lfloor p^*\rfloor), g(\lfloor p^*\rfloor + 1)) \\
&= \max(g(\lfloor\sqrt{m}\rfloor), g(\lfloor\sqrt{m}\rfloor + 1)) \\
&= \max\left(-\lfloor\sqrt{m}\rfloor - \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor}\right\rfloor, -\lfloor\sqrt{m}\rfloor - 1 - \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor + 1}\right\rfloor\right) \\
&= -\min\left(\lfloor\sqrt{m}\rfloor + \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor}\right\rfloor, \lfloor\sqrt{m}\rfloor + 1 + \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor + 1}\right\rfloor\right) \\
&= -\lfloor\sqrt{m}\rfloor - \min\left(\left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor}\right\rfloor, 1 + \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor + 1}\right\rfloor\right),
\end{aligned}$$

et donc

$$m + 1 + \Delta + \max_{1\leq p\leq w_1} g(p) = m + 1 + \Delta - \lfloor\sqrt{m}\rfloor - \min\left(\left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor}\right\rfloor, 1 + \left\lfloor\frac{m}{\lfloor\sqrt{m}\rfloor + 1}\right\rfloor\right).$$

∎

Démonstration du lemme 4.5.1 : dénotons par I^c le complément de I dans S. On a que

$$\begin{aligned}
\tilde{f}_I(\mathbf{h}) &= \int_{[0,1)^s} f_I(\mathbf{x})(-1)^{\mathbf{h}\odot\mathbf{x}} d\mathbf{x} \\
&= \int_{[0,1)^I} f_I(\mathbf{x}_I)(-1)^{\mathbf{h}_I\odot\mathbf{x}_I} d\mathbf{x}_I \left(\int_{[0,1)^{I^c}} (-1)^{\mathbf{h}_{I^c}\odot\mathbf{x}_{I^c}} d\mathbf{x}_{I^c}\right) \\
&= \begin{cases} 0 & \text{si } h_j \neq 0 \text{ pour au moins un } j \in I^c, \\ \int_{[0,1)^I} f_I(\mathbf{x}_I)(-1)^{\mathbf{h}_I\odot\mathbf{x}_I} d\mathbf{x}_I & \text{sinon,} \end{cases}
\end{aligned}$$

car $\int_{[0,1)^{I^c}} (-1)^{\mathbf{h}_{I^c}\odot\mathbf{x}_{I^c}} d\mathbf{x}_{I^c}$ vaut 0 si $\mathbf{h}_{I^c} \neq \mathbf{0}$ et 1 sinon. Maintenant, supposons que $h_j = 0$ pour tout $j \in I^c$ et que $h_j = 0$ pour au moins un $j \in I$. Posons $I_0 = \{j \in I : h_j = 0\}$ et I_0^c le complément de I_0 dans I. Alors

$$\tilde{f}_I(\mathbf{h}) = \int_{[0,1)^{I_0^c}} (-1)^{\mathbf{h}_{I_0^c}\odot\mathbf{x}_{I_0^c}} \left(\int_{[0,1)^{I_0}} f_I(\mathbf{x}_I) d\mathbf{x}_{I_0}\right) d\mathbf{x}_{I_0^c} = 0,$$

puisque $I_0 \neq \emptyset$ et donc, $\int_{[0,1)^{I_0}} f_I(\mathbf{x}_I) d\mathbf{x}_{I_0} = 0$ (voir la propriété donnée en (3.1)).

Étant donné que $I \neq I_{\mathbf{h}}$ est équivalent à ($h_j = 0$ pour au moins un $j \in I$ ou $h_j \neq 0$ pour au moins un $j \in I^c$) et que l'on a montré que $\tilde{f}_I(\mathbf{h}) = 0$ si une de ces deux conditions est satisfaite, il reste à montrer que si $I = I_{\mathbf{h}}$, alors

$$\tilde{f}_I(\mathbf{h}) = \tilde{f}(\mathbf{h}).$$

Or,

$$\begin{aligned}\tilde{f}(\mathbf{h}) &= \int_{[0,1)^s} \left(\sum_{J\subseteq S} f_J(\mathbf{x})\right)(-1)^{\mathbf{h}\odot\mathbf{x}} d\mathbf{x}, \\ &= \sum_{J\subseteq S} \tilde{f}_J(\mathbf{h}) = \tilde{f}_{I_\mathbf{h}}(\mathbf{h}),\end{aligned}$$

la troisième égalité venant du fait que pour tout $J \neq I_\mathbf{h}$, $\tilde{f}_J(\mathbf{h}) = 0$. ∎

Démonstration de la proposition 4.5.3 : d'abord, on a que

$$\sigma_I^2 = \int_{[0,1)^s} f_I^2(\mathbf{x})d\mathbf{x} = \sum_{\mathbf{h}\in\mathbb{N}^s} |\tilde{f}_I(\mathbf{h})|^2,$$

la première égalité venant du fait que $\int_{[0,1)^s} f_I(\mathbf{x})d\mathbf{x} = 0$ pour tout $\emptyset \neq I \subseteq S$, et la deuxième, de l'égalité de Parseval. Or, par le résultat du lemme 4.5.1, on a que

$$\sum_{\mathbf{h}\in\mathbb{N}^s} |\tilde{f}_I(\mathbf{h})|^2 = \sum_{\mathbf{h}\in\mathbb{N}_I^*} |\tilde{f}(\mathbf{h})|^2,$$

puisque $I = I_\mathbf{h}$ pour $\mathbf{h} \in \mathbb{N}^s$ est équivalent à avoir $\mathbf{h} \in \mathbb{N}_I^*$. Par définition de $\sigma_{I,\mathrm{PLR}}^2$, on a que $\mathrm{Var}(\hat{\mu}_{\mathrm{PLR}}) = \sum_{\emptyset\neq I\subseteq S} \sigma_{I,\mathrm{PLR}}^2$. Finalement, on peut calculer

$$\sigma_{I,\mathrm{PLR}}^2 = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} f_I(\mathbf{x}_i \oplus \mathbf{u})\right) = \sum_{0\neq\mathbf{h}\in\mathcal{L}_\mathbf{I}^*} |\tilde{f}_I(\mathbf{h})|^2 = \sum_{\mathbf{h}\in\mathcal{L}_\mathbf{I}^*\cap\mathbb{N}_I^*} |\tilde{f}(\mathbf{h})|^2,$$

où la deuxième égalité est obtenue par application de la proposition 4.5.2 à la fonction f_I et la troisième, par le lemme 4.5.1. ∎

Démonstration du lemme 4.5.2 : pour $n \geq 0$, on a que

$$N_h(n) = (2^{n+1}-1)^{|I|} - (2^n-1)^{|I|},$$

en soustrayant du nombre de vecteurs dans $\mathbb{N}_I^*$ tels que $\|\mathbf{h}\| \leq 2^n$ le nombre de ces vecteurs pour lesquels $\|\mathbf{h}\| < 2^n$. Ainsi,

$$\begin{aligned}\frac{N_h(n+r)}{N_h(n)} &= \frac{(2^{n+r+1}-1)^{|I|} - (2^{n+r}-1)^{|I|}}{(2^{n+1}-1)^{|I|} - (2^n-1)^{|I|}} \\ &\leq \frac{(2^{n+r+1}-1)^{|I|}}{(2^{n+1}-1)^{|I|}/2} \\ &= 2\left(\frac{2^{n+r+1}-1}{2^{n+1}-1}\right)^{|I|} \\ &\leq 2\left(2^{r+1}-1\right)^{|I|} \leq 2^{(r+1)|I|+1},\end{aligned}$$

la première inégalité vient du fait que

$$(2^{n+1}-1)^{|I|}-(2^n-1)^{|I|}=(2^{n+1}-1)^{|I|}\left(1-(2^n-1)^{|I|}/(2^{n+1}-1)^{|I|}\right)\geq(2^{n+1}-1)^{|I|}/2$$

pour tout $n \geq 0$ et la deuxième, du fait que le rapport $(2^{n+r+1}-1)/(2^{n+1}-1)$ est maximisé lorsque $n=0$. ∎

Démonstration du lemme 4.5.3 : pour cette démonstration, nous suivons la notation utilisée dans [105]. On a que

$$\sigma^2_{I,\mathbf{k}_I}=\int_{[0,1)^s}\nu^2_{I,\mathbf{k}_I}(\mathbf{x})d\mathbf{x},$$

où

$$\nu_{I,\mathbf{k}}(\mathbf{x})=\sum_{\tau(I,\mathbf{k})}\sum_{\gamma(I)}\langle f,\psi_{I,\mathbf{k},\tau,\gamma}\rangle\psi_{I,\mathbf{k},\tau,\gamma}(\mathbf{x}).$$

Nous ne voulons pas expliquer en détail ces sommes et comment sont définies exactement les fonctions $\psi(\cdot)$: en gros, les ψ sont des ondelettes utilisées dans la décomposition, les $\tau(\cdot)$ vont chercher les boîtes dans la partition induite par $\mathbf{k}_I$ et les γ sont des paramètres associés à ψ. Pour plus d'explications, voir [105].

Par l'égalité de Parseval, on a

$$\sigma^2_{I,\mathbf{k}_I}=\sum_{\mathbf{0}\neq\mathbf{h}\in\mathbf{N}^s}|\tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h})|^2.$$

Il faut donc calculer $\tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h})$:

$$\begin{aligned}\tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h}) &= \int_{[0,1)^s}\sum_{\tau(I,\mathbf{k})}\sum_{\gamma(I)}\langle f,\psi_{I,\mathbf{k},\tau,\gamma}\rangle\psi_{I,\mathbf{k},\tau,\gamma}(\mathbf{x})(-1)^{\mathbf{h}\odot\mathbf{x}}d\mathbf{x}\\ &= \sum_{\tau(I,\mathbf{k})}\sum_{\gamma(I)}\langle f,\psi_{I,\mathbf{k},\tau,\gamma}\rangle\int_{[0,1)^s}\psi_{I,\mathbf{k},\tau,\gamma}(\mathbf{x})(-1)^{\mathbf{h}\odot\mathbf{x}}d\mathbf{x}.\end{aligned}$$

Or, $\int_{[0,1)^s}\psi_{I,\mathbf{k},\tau,\gamma}(\mathbf{x})(-1)^{\mathbf{h}\odot\mathbf{x}}d\mathbf{x}=0$ si $\mathbf{h}\notin H(\mathbf{k}_I)$, par les propriétés de la fonction $\psi(\cdot)$. En effet, la fonction $\psi_{I,\mathbf{k},\tau,\gamma}(\cdot)$ est constante sur les boîtes déterminées par $(k_j+1)_{j\in I}$ et est intégrée à 0 sur celles déterminées par $\mathbf{k}_I$ [105]. S'il existe $j\in I$ tel que $|h_j|_{\mathrm{p}}=2^{k_j+1}$, alors

$$\int_{[0,1)^s}\psi_{I,\mathbf{k},\tau,\gamma}(\mathbf{x})(-1)^{\mathbf{h}\odot\mathbf{x}}d\mathbf{x}=\prod_{j\in I}\int_0^1\psi_{k_j,\tau_j,\gamma_j}(x_j)(-1)^{h_j\odot x_j}dx_j=0,$$

car $(-1)^{h_j\odot x_j}$ vaut -1 sur la moitié de chaque intervalle sur lequel $\psi_{k_j,\tau_j,\gamma_j}$ est constante et 1 sur l'autre moitié. Ensuite, s'il existe $r\in I$ tel que $|h_r|_{\mathrm{p}}=2^{k_r-1}$, alors $(-1)^{h_r\odot x_r}$

est constante sur les intervalles de la forme $[(n_r-1)2^{-k_r}, n_r 2^{-k_r})$, et $\psi_{k_r,\tau_r,\gamma_r}$ est intégrée à 0 sur ce type d'intervalle.

Ainsi,

$$\sigma^2_{I,\mathbf{k}_I} = \sum_{\mathbf{h}\in H(\mathbf{k}_I)} |\tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h})|^2.$$

Il reste donc à montrer que si $\mathbf{h} \in H(\mathbf{k}_I)$, alors $\tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h}) = \tilde{f}(\mathbf{h})$. Or, on a que

$$\tilde{f}(\mathbf{h}) = \int_{[0,1)^s} \sum_I \sum_{\mathbf{k}_I} \sum_{\tau(I,\mathbf{k})} \sum_{\gamma(I)} \langle f, \psi_{I,\mathbf{k},\tau,\gamma}\rangle \psi_{I,\mathbf{k},\tau,\gamma}(\mathbf{x})(-1)^{\mathbf{h}\odot\mathbf{x}} d\mathbf{x} = \sum_I \sum_{\mathbf{k}_I} \tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h}).$$

Dans la somme sur I, il n'y a que le terme $I = I_{\mathbf{h}}$ qui est non nul, par le lemme 4.5.1 appliqué à la fonction $\psi_{I,\mathbf{k},\tau,\gamma}$. De même, de la somme sur les $\mathbf{k}_I$, il ne reste que le vecteur $\mathbf{k}_I$ tel que $\mathbf{h} \in H(\mathbf{k}_I)$, puisque $\tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h}) = 0$ si $\mathbf{h} \notin H(\mathbf{k}_I)$. Donc, on a bien que $\tilde{f}(\mathbf{h}) = \tilde{\nu}_{I,\mathbf{k}_I}(\mathbf{h})$. ∎

Démonstration du lemme 4.5.4 : nous allons utiliser les deux règles suivantes :

(1) Règle 1, [106, *Rule C1*] : supposons qu'il existe $A > 0$ fini et $\beta \in (0, 1]$ tels que $|f'(x) - f'(x^*)| \leq A|x - x^*|^\beta$ pour tous x, x^* réels. Alors

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + C|x - x^*|^{1+\beta},$$

où $|C| \leq A(1 + \beta)^{-1} \leq A$.

(2) Règle 2, [106, *Rule C3*] : sous les mêmes conditions que celles de la règle 1, on a que

$$\frac{1}{n}\sum_{i=1}^{n} f'(\bar{x}_t)^2 = \int_0^1 f'(x)^2 dx + O(n^{-\beta}),$$

où $\bar{x}_t = (t - 0.5)/n$.

De plus, comme l'explique Owen dans [106], si $\partial^s f/\partial\mathbf{x}$ est Lipschitz-continue, alors toutes les dérivées partielles mixtes d'ordre $|I|$ des composantes ANOVA composantes f_I sont également Lipschitz-continues. Pour simplifier la notation, supposons que $I = \{1, \ldots, t\}$. En utilisant le lemme E.1 combiné avec le lemme 4.5.1, la première étape consiste à calculer

$$w_n = \int_{v_1/2^{k_1}}^{(v_1+1)/2^{k_1}} \cdots \int_{v_t/2^{k_t}}^{(v_t+1)/2^{k_t}} f_I(\mathbf{x}_I)(-1)^{\sum_{j=1}^t x_{j,k_j+1}} d\mathbf{x}_I, \tag{B.15}$$

où chaque v_j est entre 0 et $2^{k_j} - 1$, inclusivement.

Calculons l'intégrale par rapport à x_t en utilisant la règle 1 avec $x_t^* = v_t/2^{k_t} + 1/2^{k_t+1}$, le point milieu de l'intervalle sur lequel on intègre. On obtient

$$\begin{aligned}
&\frac{\partial f}{\partial x_t} f(x_1, \ldots, x_{t-1}, x_t^*) \left(\frac{(x - x_t^*)^2}{2} \Big|_{v_t/2^{k_t}}^{x_t^*} - \frac{(x - x_t^*)^2}{2} \Big|_{x_t^*}^{(v_t+1)/2^{k_t}} \right) \\
&+C \left(\frac{(x - x_t^*)^{2+\beta}}{2+\beta} \Big|_{v_t/2^{k_t}}^{x_t^*} + \frac{(x - x_t^*)^{2+\beta}}{2+\beta} \Big|_{x_t^*}^{(v_t+1)/2^{k_t}} \right) \\
= \; & -2^{-2(k_t+1)} \frac{\partial f}{\partial x_t} f(x_1, \ldots, x_{t-1}, x_t^*),
\end{aligned}$$

puisque le premier terme venant de la règle 1 s'annule, étant donné que $(-1)^{x_t, k_{t+1}}$ vaut 1 sur la moitié de l'intégrale et -1 sur l'autre moitié. Ainsi, en appliquant successivement la règle 1 sur les $t-1$ autres intégrales, on obtient que (B.15) est de la forme

$$\frac{\partial f}{\partial \mathbf{x}_I} f(\mathbf{x}_I^*)(-1)^{|I|} 2^{-\sum_{j \in I} 2(k_j+1)}.$$

Pour calculer $\sum_{\mathbf{h} \in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2$, on utilise le lemme E.1, qui nous dit que

$$\sum_{\mathbf{h} \in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2 = 2^\kappa \sum_{n=1}^{2^\kappa} w_n^2$$

et

$$\begin{aligned}
\sum_{n=1}^{2^\kappa} w_n^2 &= \sum_{v_1=0}^{2^{k_1}-1} \cdots \sum_{v_t=0}^{2^{k_t}-1} \left(\frac{\partial f}{\partial \mathbf{x}_I} f(\mathbf{x}_I^*) \right)^2 (-1)^{2|I|} 2^{-\sum_{j \in I} 4(k_j+1)} \\
&= 2^{-3\kappa - 4|I|} \left[\int_{[0,1)^{|I|}} \left(\frac{\partial f(\mathbf{x}_I)}{\partial \mathbf{x}_I} \right)^2 d\mathbf{x}_I \right] \left(1 + O(2^{-\beta \min_{j \in I} k_j}) \right),
\end{aligned}$$

en appliquant successivement la règle 2 sur les t sommes et donc,

$$\sum_{\mathbf{h} \in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2 = 2^{-2\kappa - 4|I|} \left[\int_{[0,1)^{|I|}} \left(\frac{\partial f(\mathbf{x}_I)}{\partial \mathbf{x}_I} \right)^2 d\mathbf{x}_I \right] \left(1 + O(2^{-\beta \min_{j \in I} k_j}) \right).$$

∎

Démonstration du lemme 4.5.5 : on sait que

$$\sum_{k \geq 0} b^{-k} (1 - b^{-1})^u \binom{k+u-1}{u-1} = 1,$$

car chaque terme de cette somme représente la probabilité qu'une variable aléatoire binômiale négative de paramètres $p = 1/b$ et $r = u$ soit égale à k. Pour pouvoir

utiliser cela afin de démontrer (4.31), on doit borner le rapport $\left(\binom{k+m-j-1}{u-1}/\binom{k+u-1}{u-1}\right)$. Si $j = m-u+1$, alors ce rapport est inférieur ou égal à 1 et donc

$$\sum_{k\geq 1} b^{-k}\binom{k+m-j-1}{u-1} \leq (1-b^{-1})^{-u}.$$

Sinon, ce rapport vaut $\prod_{l=0}^{u-2} g(l)$, avec

$$g(l) = \frac{k+m-j-1-l}{k+u-1-l}$$

et cette fonction croît avec l si $k+m-j-1 \geq k+u-1$, qui équivaut à avoir $m-j \geq u$, qui tient par hypothèse. Donc,

$$\begin{aligned}
\prod_{l=0}^{u-2} g(l) &\leq (g(u-2))^{u-1} \\
&= \left(\frac{k+m-j-1-u+2}{k+u-1-u+2}\right)^{u-1} \\
&= \left(\frac{k+m-j+1-u}{k+1}\right)^{u-1} \\
&\leq (m-j+1-u)^{u-1}, \text{ pour tout } k \geq 0, \\
&\leq m^{u-1},
\end{aligned}$$

puisque $u+j \geq 1$. Ainsi, on obtient que si $j \leq m-u$, alors

$$\begin{aligned}
\sum_{k\geq 1} b^{-k}\binom{k+m-j-1}{u-1} &\leq m^{u-1}\sum_{k\geq 1} b^{-k}\binom{k+u-1}{u-1} \\
&\leq m^{u-1}(1-b^{-1})^{-u}.
\end{aligned}$$

Donc, pour toute valeur de $j \leq m-u+1$, on a que

$$\sum_{k\geq 1} b^{-k}\binom{k+m-j-1}{u-1} \leq m^{u-1}(1-b^{-1})^{-u},$$

puisque $m^{u-1} \geq 1$ et donc, l'inégalité tient aussi pour $j = m-u+1$. ∎

Démonstration du lemme 4.5.6 : on a besoin du le lemme suivant :

Lemme B.1 *Soient $x_1, \ldots, x_n$ des entiers non négatifs pour lesquels on sait que $x_i \leq c$ et $\sum_{i=1}^n x_i \leq M \leq nc$. Alors*

$$\sum_{i=1}^n x_i^2 \leq Mc.$$

Démonstration : posons $n(j)$ le nombre de x_i qui sont égaux à j, pour $j = 1, \ldots, c$. On a donc que $\sum_{i=1}^{n} x_i = \sum_{j=1}^{c} jn(j) \leq M$ et on veut borner $\sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{c} j^2 n(j)$. On n'a qu'à utiliser l'inégalité de Cauchy-Schwarz, ce qui nous donne

$$\begin{aligned}\sum_{j=1}^{c} j^2 n(j) &\leq \left(\sum_{j=1}^{c} jn(j)\right)^{1/2} \left(\sum_{j=1}^{c} j^3 n(j)\right)^{1/2} \\ &\leq \sqrt{M}\sqrt{c}\left(\sum_{j=1}^{c} j^2 n(j)\right)^{1/2},\end{aligned}$$

ce qui implique que $\left(\sum_{j=1}^{c} j^2 n(j)\right)^{1/2} \leq \sqrt{Mc}$ et donc, que $\sum_{i=1}^{n} x_i^2 \leq Mc$. ∎

Le cas où $\kappa \leq m - q_I - |I|$ découle directement de la proposition 4.3.2. Ensuite, par le lemme E.2, on sait que

$$|H(\mathbf{k}_I) \cap \mathcal{L}_s^*| = \frac{2^\kappa}{N^2} \sum_{i=1}^{2^\kappa} y_n^2,$$

où $y_n = \sum_{i=1}^{N} \mathbf{1}_{\mathbf{x}_i(\mathbf{k}_I)=b_n} (-1)^{\sum_{j\in I} x_{i,j,k_j+1}}$.

Si $\kappa \geq m - q_I + 1$, alors $|y_n| \leq \sum_{i=1}^{N} \mathbf{1}_{\mathbf{x}_i(\mathbf{k}_I)=b_n} \leq 2^{q_I}$, puisqu'en réunissant $2^{q_I - m + \kappa}$ boîtes b_n ayant les mêmes $\tilde{k}_j$ premiers bits pour chaque $j \in I$, où $\sum_{j\in I} \tilde{k}_j = m - q_I$, on obtient une boîte de volume $2^{q_I - m}$ (car l'ensemble des $2^{q_I - m + \kappa}$ boîtes représente toutes les combinaisons possibles pour les $(k_j - \tilde{k}_j)$ derniers bits), qui par définition de q_I, doit contenir exactement 2^{q_I} points et donc, chacune des boîtes b_n ne peut contenir plus de 2^{q_I} points. En utilisant le lemme B.1 avec $c = 2^{q_I}$, $n = 2^\kappa$ et $M = 2^m \leq 2^{q_I + \kappa}$, on a que

$$\sum_{i=1}^{2^\kappa} y_n^2 \leq 2^{m+q_I}$$

et ainsi

$$|H(\mathbf{k}_I) \cap \mathcal{L}_s^*| \leq 2^{\kappa - (m - q_I)}.$$

Si $m - q_I - |I| + 1 \leq \kappa < m - q_I + 1$, alors toutes les boîtes b_n dans la définition de y_n contiennent exactement $2^{m-\kappa}$ points et donc, $|y_n| \leq 2^{m-\kappa}$, ce qui implique que

$$\sum_{i=1}^{2^\kappa} y_n^2 \leq 2^\kappa 2^{2(m-\kappa)} = 2^{2m-\kappa}.$$

Ainsi,

$$|H(\mathbf{k}_I) \cap \mathcal{L}_s^*| = \frac{2^\kappa}{N^2} 2^{2m-\kappa} = 1.$$

∎

Annexe C

Expériences avec le critère $\bar{M}_{t,t,t}$

Nous donnons au tableau les résultats de recherches utilisant le critère $\bar{M}_{32,32,32}$ (défini en (3.26)) afin de choisir des GCL à période maximale. Nous donnons également les meilleurs GCL par rapport aux critères M_{32} et $M_{32,32,32}$.

TABLEAU C.1: Meilleurs a par rapport à M_{32}, $\bar{M}_{32,32,32}$ et $M_{32,32,32}$

N	a	M_{32}	$\bar{M}_{32,32,32}$	$M_{32,32,32}$
1021	65	0.61872*	0.61872	0.15324
	65	0.61872	0.61872*	0.15324
	178	0.57338	0.57338	0.20594*
2039	393	0.65283*	0.56586	0.15695
	208	0.61885	0.61302*	0.17209
	131	0.39572	0.39572	0.20712*
4093	3593	0.64259*	0.54584	0.05245
	118	0.64051	0.61044*	0.13642
	1074	0.55470	0.55470	0.18472*
8191	2685	0.64854*	0.59667	0.10334
	147	0.62502	0.62502*	0.10826
	1043	0.60655	0.60655	0.18751*
16381	12957	0.65508*	0.59230	0.10524
	1480	0.62497	0.62497*	0.10524
	319	0.54457	0.54457	0.18891*
32749	9515	0.67356*	0.61307	0.08354
	8762	0.65808	0.62020*	0.06821
	1339	0.54134	0.54134	0.19690*

Comme on peut le voir dans ce tableau, les GCL choisis selon $\bar{M}_{32,32,32}$ ou $M_{32,32,32}$

ont généralement une valeur de $\bar{M}_{32,32,32}$ très près de leur valeur pour M_{32}. Aussi, les meilleurs GCL par rapport à M_{32} ont de bons résultats par rapport au critère $\bar{M}_{32,32,32}$. Cela semble indiquer que le fait de considérer la moyenne cache les mauvaises projections non successives. En fait, les différentes règles obtiennent probablement des valeurs assez semblables pour les moyennes sur les paires et les triplets. Ainsi, c'est ce qui se passe sur les projections successives qui détermine la valeur du critère. En effet, on constate en regardant le tableau C.1 que généralement, seules les règles ayant de mauvaises projections non successives ($M_{32,32,32} < 0.10$, disons) ont une valeur de $\bar{M}_{32,32,32}$ inférieure à leur valeur de M_{32} et même dans ce cas, la différence n'est pas très grande entre ces deux quantités. Par exemple, on peut voir que pour $N = 32749$, le meilleur GCL par rapport à $\bar{M}_{32,32,32}$ n'obtient pas une très bonne valeur pour $M_{32,32,32}$, alors que sa valeur de $\bar{M}_{32,32,32}$ n'est pas très loin de celle de M_{32}. Cela veut dire qu'il y a une projection sur une paire ou un triplet qui est mauvaise, mais que le fait de regarder ce qui se passe en moyenne cache ce défaut. En conclusion, on peut dire que l'utilisation du critère $\bar{M}_{t,t,t}$ n'offre pas vraiment d'avantages, car si on peut se contenter d'un ensemble de points dont les projections moyennes sont bonnes, il semble que le critère M_t nous permet de choisir des règles appropriées et si on ne veut pas avoir de mauvaises projections, le critère M_{t_1,t_2,t_3} est plus sûr.

Annexe D

Expériences additionnelles sur une fonction-test

Puisque la dimension s va jusqu'à 30 dans l'exemple de la sous-section 3.6.2, on pourrait penser qu'un a choisi à partir de M_{t_1,t_2,t_3} avec $t_2, t_3 > 8$ serait préférable et c'est pourquoi dans le tableau D.2, on refait la même expérience que dans le tableau 3.3, mais on compare maintenant le meilleur a par rapport à M_{30} avec les meilleurs par rapport à $M_{30,24,12}$ et $M_{30,30,30}$, qui sont donnés au tableau D.1. En particulier, nous voulons voir si les règles choisies avec $M_{30,30,30}$ obtiennent des résultats significativement meilleurs que celles choisies avec $M_{30,24,12}$.

TABLEAU D.1: Meilleurs a par rapport à M_{30}, $M_{30,24,12}$ et $M_{30,30,30}$.

N	a	M_{30}	$M_{30,24,12}$	$M_{30,30,30}$
32749	9515	0.67356*	0.33761*	0.08354
	9515	0.67356*	0.33761*	0.08354
	1339	0.54134	0.21023	0.19690*
65521	2469	0.63900*	0.17455	0.06630
	8950	0.55678	0.34307*	0.07329
	8753	0.39716	0.22859	0.21311*
131071	29803	0.66230*	0.03138	0.03138
	3244	0.42390	0.34857*	0.15087
	82893	0.45522	0.20565	0.20565*
Nb. de projections		29	105	462

En gros, on observe le même phénomène que dans le tableau 3.3, c.-à-d., lorsque le

TABLEAU D.2: Facteurs de réduction de variance moyens, $\alpha = 3$

	N	$s = 5$	$s = 10$	$s = 15$	$s = 20$	$s = 25$	$s = 30$
M_{30}	32749	2160	1100	1240	1610*	1860*	2370*
$M_{30,24,12}$	32749	2160	1100	1240	1610*	1860*	2370*
$M_{30,30,30}$	32749	2970*	3250*	1350*	1490	1680	2130
M_{30}	65521	256	32.3	42.1	60.0	115	124
$M_{30,24,12}$	65521	1680	2020	4110*	4320*	4260*	4310*
$M_{30,30,30}$	65521	4150*	2170*	2170	2000	2690	2850
M_{30}	131071	9020*	2200	137	94.3	116	159
$M_{30,24,12}$	131071	5620	4380*	6730*	6360*	10500*	8920*
$M_{30,30,30}$	131071	4790	635	505	519	899	1590

meilleur a par rapport à M_{30} a une petite valeur de $M_{30,30,30}$, en comparaison avec le meilleur a par rapport à ce critère (c'est ce qui se produit pour les trois valeurs de N), alors la variance de la règle associée est habituellement supérieure à celle des autres a, surtout quand s est grand. Le cas où $N = 32749$ n'obéit pas à ce comportement, malgré une assez petite valeur de $M_{30,30,30}$. Nous croyons que cela s'explique par le fait que la meilleure règle par rapport à M_{30} pour cette valeur de N est également la meilleure par rapport à $M_{30,24,12}$. Cela signifie que pour les 105 projections considérées dans $M_{30,24,12}$, cette règle est excellente. Si les projections de f qui sont les plus importantes sont parmi ces 105, alors ce qui se passe sur le restant des projections ne réussit pas à amoindrir de façon significative la qualité de la règle pour ce problème.

Remarquons également que les règles choisies selon $M_{30,24,12}$ ont une variance du même ordre que celles choisies avec $M_{30,30,30}$, même qu'elle est parfois inférieure par des facteurs non négligeables, par exemple lorsque $N = 131071$. En ce qui concerne la différence entre les règles choisies avec $M_{8,8,8}$ (voir tableau 3.3) et celles choisies avec $M_{30,24,12}$ ou $M_{30,30,30}$, on observe que pour $N = 65521$, les règles $M_{8,8,8}$ obtiennent souvent de meilleurs résultats et pour $N = 131071$, $M_{30,24,12}$ est la meilleure. Comment expliquer cela ? N'oublions pas que notre fonction-test a des paramètres c_j choisis aléatoirement. Il est ainsi difficile de savoir à l'avance quelles projections f_I sont importantes et donc, quel critère est-il préférable d'utiliser.

Annexe E

Représentation des fonctions de base de la série de Walsh à l'aide d'une matrice d'Hadamard

Dans cette annexe, nous expliquons comment représenter les fonctions de base de la série de Walsh à l'aide d'une *matrice d'Hadamard*, dont nous rappellerons la définition à la page xlviii. Cela nous sert ensuite à démontrer deux résultats intermédiaires qui sont utiles dans le chapitre 4. Le premier nous aide à calculer la variance de l'estimateur XOR-translaté dans le cas des fonctions *lisses* (lemme 4.5.4) et le deuxième nous permet de borner le nombre de vecteurs $\mathbf{h}$ dans $\mathcal{L}_s^*$ respectant

$$|h_j|_{\mathrm{p}} = 2^{k_j}, \text{ pour tout } j \in I,$$

pour un vecteur $(k_j)_{j\in I}$ donné, où la norme $|\cdot|_{\mathrm{p}}$ a été définie en (4.8), page 139. Ce résultat, combiné avec le lemme 4.2.1, sert à manipuler des bornes sur la variance de l'estimateur XOR-translaté à la section 4.5 et est également utilisé dès la section 4.3, lorsque nous analysons les critères de qualité pour les règles de réseau polynômiales.

Présentons d'abord la notation :

$$\begin{aligned} \mathbf{k}_I &= (k_j)_{j\in I}, \text{ où les } k_j \text{ sont des entiers non négatifs,} \\ \kappa &= \sum_{j\in I} k_j, \end{aligned}$$

$\mathbf{x}_I(\mathbf{k}_I)$ = chaîne de κ bits représentant la troncation de $\mathbf{x}_I$ à ses k_j premiers bits dans chaque dimension $j \in I$,

b_n = chaîne de κ bits correspondant à la représentation binaire de $n-1$, pour $n = 1, \ldots, 2^\kappa$,

$H(\mathbf{k}_I)$ = ensemble des vecteurs $\mathbf{h} \in \mathbb{N}_I^*$ tels que $|h_j|_p = 2^{k_j}$ pour tout $j \in I$, où $\mathbb{N}_I^* = \{\mathbf{h} \in \mathbb{N}^s : I_h = I\}$.

Le vecteur $\mathbf{k}_I$ va servir à déterminer une partition de $\prod_{j \in I}[0,1)$ en 2^κ boîtes de volume $2^{-\kappa}$, qui peuvent être identifiées à l'ensemble $\{b_1, \ldots, b_{2^\kappa}\}$ des chaînes comptant κ bits. Pour déterminer dans quelle boîte un point $\mathbf{x}$ se trouve, il suffit de considérer ses k_j premiers bits dans chaque dimension $j \in I$, d'où la notation $\mathbf{x}_I(\mathbf{k}_I)$. Finalement, les $H(\mathbf{k}_I)$ sont des sous-ensembles de vecteurs $\mathbf{h}$ qui ont tous la même norme, c.-à-d., tels que $|h_j|_p = 2^{k_j}$ pour tout $j \in I$.

Définition E.1 *[8, 37] Une* matrice d'Hadamard *est une matrice carrée dont les entrées sont des 1 et des -1 et qui a la propriété que ses lignes (et ses colonnes) sont orthogonales l'une à l'autre.*

Par exemple, la matrice

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

est une matrice d'Hadamard d'ordre 2.

La représentation des fonctions de base $(-1)^{\mathbf{h} \odot \mathbf{x}}$ à l'aide d'une matrice d'Hadamard est expliquée dans [8] lorsque $\mathbf{h}$ et $\mathbf{x}$ sont unidimensionnels. Nous expliquons ici comment appliquer cela pour représenter $(-1)^{\mathbf{h} \odot \mathbf{x}}$ lorsque $\mathbf{h}$ et $\mathbf{x}$ sont multidimensionnels.

Nous rappelons d'abord la définition du *produit de Kronecker* (ou produit direct) de deux matrices [37], dénoté par $\mathbf{A} \otimes \mathbf{B}$: si $\mathbf{A}$ est de dimension i par j et $\mathbf{B}$ est de dimension k par l, alors $\mathbf{A} \otimes \mathbf{B}$ est de dimension ik par jl et peut être vue comme une matrice de dimension i par j dont chaque élément correspond à la matrice que l'on

obtient en multipliant chaque élément de $\mathbf{A}$ par la matrice $\mathbf{B}$. Par exemple, on a que

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

En une dimension, le lien avec les séries de Walsh est introduit en définissant une matrice que l'on dénote par $\mathbf{W}_k$ et qui est telle que chaque ligne est associée à un entier non-négatif $h \leq 2^k$. Sur cette ligne, on trouve les valeurs successives que prend $(-1)^{h \odot x}$ à mesure que x passe d'un intervalle de longueur 2^{-k} à l'autre. Plus précisément :

Définition E.2 *La matrice $\mathbf{W}_k$ de dimension 2^k par 2^k est définie comme ayant en position (h, n) l'élément*

$$\mathbf{W}_{k,h,n} = (-1)^{h \odot ((n-1)2^{-k})}, \tag{E.1}$$

pour $0 \leq h, n < 2^k$.

En fait, l'élément $\mathbf{W}_{k,h,n}$ nous donne la valeur de $(-1)^{h \odot x}$ lorsque $x \in [(n-1)2^{-k}, n2^{-k})$. On peut facilement vérifier que la matrice $\mathbf{W}_k$ est une matrice d'Hadamard, ce qui signifie que

$$\mathbf{W}_k^T \mathbf{W}_k = 2^k \mathbf{I}_k,$$

où $\mathbf{I}_k$ est la matrice identité d'ordre 2^k.

On veut maintenant généraliser la définition de $\mathbf{W}_k$ au cas multidimensionnel, afin de représenter les fonctions de base $(-1)^{\mathbf{h} \odot \mathbf{x}}$ pour les vecteurs $\mathbf{h}$ ayant la propriété que $|h_j|_p < 2^{k_j}$ pour tout $j \in I$, pour un vecteur $\mathbf{k}_I$ et un sous-ensemble $I \subseteq S$ donnés. La raison derrière cela est que nous aurons besoin de travailler avec cet ensemble de fonctions de base afin de démontrer les lemmes de cette sous-section, qui à leur tour seront utilisés plus loin dans le chapitre 4.

Pour faire cela, nous allons définir une matrice $\mathbf{H}_\kappa$ de dimension 2^κ par 2^κ. Les lignes de $\mathbf{H}_\kappa$ sont indexées par les 2^κ différentes valeurs que peut prendre $\mathbf{h} = (h_j)_{j \in I}$ (si on restreint chaque h_j à respecter $|h_j|_p < 2^{k_j}$) et ses colonnes sont indexées par les 2^κ différentes valeurs que peut prendre le vecteur $(n_j)_{j \in I}$ (qui nous indique quelle boîte dans $[0,1)^s$ est considérée, c.-à-d., $\prod_{j \in I} [(n_j - 1)2^{-k_j}, n_j 2^{-k_j}))$.

Définition E.3 *La matrice* $\mathbf{H}_\kappa$ *de dimension* 2^κ *par* 2^κ *est définie comme ayant en position* $((h_j)_{j\in I},(n_j)_{j\in I})$ *l'élément*

$$\mathbf{H}_{\kappa,(h_j)_{j\in I},(n_j)_{j\in I}} = \prod_{j\in I}(-1)^{h_j\odot((n_j-1)2^{-k_j})}.$$

L'élément $\mathbf{H}_{\kappa,(h_j)_{j\in I},(n_j)_{j\in I}}$ nous donne la valeur que prend $(-1)^{\mathbf{h}\odot\mathbf{x}}$ lorsque $\mathbf{x}$ est dans $\prod_{j\in I}[(n_j-1)2^{-k_j},n_j2^{k_j})$.

Si on suppose que le sous-ensemble I est donné par $\{i_1,\ldots,i_t\}$, alors en utilisant la définition du produit de Kronecker entre deux matrices, on obtient que

$$\mathbf{H}_\kappa = \mathbf{W}_{k_{i_1}}\otimes\ldots\otimes\mathbf{W}_{k_{i_t}},$$

où les $\mathbf{W}_{k_{i_j}}$ proviennent de la définition E.2. De plus, la matrice $\mathbf{H}_\kappa$ est aussi une matrice d'Hadamard, puisqu'elle contient des 1 et des -1 et que

$$\begin{aligned}
\mathbf{H}_\kappa^T\mathbf{H}_\kappa &= (\mathbf{W}_{k_{i_1}}\otimes\ldots\otimes\mathbf{W}_{k_{i_t}})^T(\mathbf{W}_{k_{i_1}}\otimes\ldots\otimes\mathbf{W}_{k_{i_t}})\\
&= ((\mathbf{W}_{k_{i_t}}^T\otimes\ldots\otimes\mathbf{W}_{k_{i_2}}^T)\otimes\mathbf{W}_{k_{i_1}}^T)(\mathbf{W}_{k_{i_1}}\otimes(\mathbf{W}_{k_{i_2}}\otimes\ldots\otimes\mathbf{W}_{k_{i_t}}))\\
&= 2^{k_{i_t}}((\mathbf{W}_{k_{i_t}}^T\otimes\ldots\mathbf{W}_{k_{i_3}}^T)\otimes\mathbf{W}_{k_{i_2}}^T)(\mathbf{W}_{k_{i_2}}\otimes(\mathbf{W}_{k_{i_3}}\otimes\ldots\otimes\mathbf{W}_{k_{i_t}}))\\
&= 2^\kappa\mathbf{I}_\kappa.
\end{aligned}$$

Dans ce qui précède, la deuxième égalité vient du fait que $(\mathbf{A}\otimes\mathbf{B})^T = \mathbf{A}^T\otimes\mathbf{B}^T$ [37] ; la troisième, du fait que $\mathbf{W}_{k_{i_1}}^T\mathbf{W}_{k_{i_1}} = 2^{k_{i_1}}\mathbf{I}_{k_{i_1}}$ et le résultat suit en appliquant successivement cette identité aux $t-1$ produits suivants $\mathbf{W}_{k_{i_2}}^T\mathbf{W}_{k_{i_2}},\ldots,\mathbf{W}_{k_{i_t}}^T\mathbf{W}_{k_{i_t}}$.

En résumé, pour un vecteur $\mathbf{k}_I$ donné, si on considère tous les vecteurs $\mathbf{h}$ tels que $|h_j|_p < 2^{k_j}$ pour tout $j\in I$, alors on peut associer à chaque ligne de la matrice $\mathbf{H}_\kappa$ un de ces vecteurs $\mathbf{h}$ et les entrées de cette ligne nous donnent les valeurs successives que prend $(-1)^{\mathbf{h}\odot\mathbf{x}}$ lorsque $\mathbf{x}$ passe d'une boîte de la forme

$$\prod_{j\in I}[2^{-k_j}(n_j-1),2^{-k_j}n_j)$$

à l'autre, avec $n_j = 1,\ldots,2^{k_j}$.

Comment cette représentation va-t-elle nous aider ? Voici en gros l'idée : le fait de pouvoir utiliser la matrice $\mathbf{H}_\kappa$ pour représenter les fonctions de base qui sont dans l'ensemble $\{(-1)^{\mathbf{h}\odot\mathbf{x}} : |h_j|_p < 2^{k_j}, j\in I\}$ pour un vecteur $\mathbf{k}_I$ donné va nous être

utile afin de calculer la norme euclidienne des vecteurs $\mathbf{z}$ pouvant s'écrire sous la forme $\mathbf{z} = \mathbf{H}_\kappa \mathbf{y}$. Ceci vient du fait que $\mathbf{H}_\kappa$ est une matrice d'Hadamard et donc, que $\|\mathbf{z}\|_2^2 = \mathbf{y}^T \mathbf{H}_\kappa^T \mathbf{H}_\kappa \mathbf{y} = 2^\kappa \|\mathbf{y}\|_2^2$. Cette représentation est donc utile si la norme $\|\mathbf{y}\|_2^2$ est plus facile à calculer que $\|\mathbf{z}\|_2^2$.

Les deux résultats qui suivent exploitent précisément cette idée, c.-à-d., dans les deux cas, on donne, pour un vecteur $\mathbf{z}$ donné, quel est le $\mathbf{y}$ qui fait en sorte que $\mathbf{z} = \mathbf{H}_\kappa \mathbf{y}$. Au lemme E.1, le vecteur $\mathbf{z}$ auquel on s'intéresse est le vecteur de coefficients de Walsh associés aux $\mathbf{h}$ qui sont dans $H(\mathbf{k}_I)$, pour un $\mathbf{k}_I$ donné et la formule que l'on donne pour $\|\mathbf{z}\|_2^2$ est utilisée dans la démonstration du lemme 4.5.4. Au lemme E.2, le vecteur $\mathbf{z}$ est carrément le vecteur de fonctions de base $(-1)^{\mathbf{h}\odot\mathbf{x}}$ pour tous les $\mathbf{h}$ dans $H(\mathbf{k}_I)$ et dans ce cas, la quantité $\|\mathbf{z}\|_2^2$ nous permet de déterminer combien de vecteurs dans $H(\mathbf{k}_I)$ font partie du réseau dual $\mathcal{L}_s^*$, en utilisant le lemme 4.2.1. Voici le premier de ces deux résultats :

Lemme E.1 *Soit $\mathbf{k}_I = (k_j)_{j\in I}$ un vecteur d'entiers $k_j \geq 0$. Pour $f \in \mathcal{L}^2$, on a que*

$$\sum_{\mathbf{h}\in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2 = \mathbf{w}^T \mathbf{H}_\kappa^T \mathbf{H}_\kappa \mathbf{w} = 2^\kappa \sum_{n=1}^{2^\kappa} w_n^2,$$

où $\mathbf{w} = (w_1, \ldots, w_{2^\kappa})^T$, $w_n = \int_{\mathbf{x}_I:\mathbf{x}_I(\mathbf{k}_I)=b_n} f(\mathbf{x})(-1)^{\sum_{j\in I} x_{j,k_j+1}} d\mathbf{x}$ et b_n est une chaîne de κ bits correspondant à la représentation binaire de $n-1$.

Démonstration : suivant la définition E.3, pour une valeur de $\mathbf{x} \in [0,1)^s$ donnée, l'ensemble des valeurs de $(-1)^{\mathbf{h}\odot\mathbf{x}}$ correspondant aux vecteurs $\tilde{\mathbf{h}}_n, n = 1, \ldots, 2^\kappa$ qui sont formés de k_j bits dans chaque dimension $j \in I$ (c.-à-d., $|h_j|_p < 2^{k_j}$ pour tout $j \in I$) peut être obtenu en calculant le produit matriciel

$$\mathbf{H}_\kappa \left(\mathbf{1}_{\mathbf{x}_I(\mathbf{k}_I)=b_1}, \ldots, \mathbf{1}_{\mathbf{x}_I(\mathbf{k}_I)=b_{2^\kappa}}\right)^T. \tag{E.2}$$

En effet, chaque colonne de $\mathbf{H}_\kappa$ correspond à une des 2^κ boîtes résultant de la partition induite par $\mathbf{k}_I$ et on a que

$$\mathbf{x}_I \in \prod_{j\in I} [2^{-k_j}(n_j - 1), 2^{-k_j} n_j)$$

si et seulement si $\mathbf{x}_I(\mathbf{k}_I) = b_m$, où $n_{i_1,0} \ldots n_{i_1,k_{i_1}-1} \ldots n_{i_t,0} \ldots n_{i_t,k_{i_t}-1} = b_m$, pour $I = \{i_1, \ldots, i_t\}$, et les $n_{j,l}$ pour $l = 0, \ldots, k_j - 1$ sont les coefficients de l'expansion binaire

de n_j, $j \in I$. Il y a donc bijection entre les 2^κ boîtes de type $\prod_{j\in I}[2^{-k_j}(n_j-1), 2^{-k_j}n_j)$ et les chaînes b_m, $m = 1, \ldots, 2^\kappa$. Pour récapituler, le produit qui se trouve à l'équation (E.2) nous donne, dans un ordre qu'il n'est pas important de préciser (car notre but final est de calculer une somme dépendant de ces termes), la valeur que prend chaque fonction de base au point $\mathbf{x}$, c.-à-d., le vecteur

$$\left((-1)^{\tilde{\mathbf{h}}_1 \odot \mathbf{x}}, \ldots, (-1)^{\tilde{\mathbf{h}}_{2^\kappa} \odot \mathbf{x}}\right)^T.$$

Pour obtenir l'ensemble des $\tilde{f}(\mathbf{h})$ pour $\mathbf{h} \in H(\mathbf{k}_I)$ à partir de cela, remarquons que les 2^κ vecteurs dans $H(\mathbf{k}_I)$ peuvent être obtenus en insérant un bit égal à 1 après chaque sous-chaîne de longueur k_j dans les $\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_{2^\kappa}$, c.-à-d., si on a

$$\tilde{\mathbf{h}}_n = h_{i_1,0}h_{i_1,1}\ldots h_{i_1,k_{i_1}-1}\ldots h_{i_\ell,0}h_{i_\ell,1}\ldots h_{i_\ell,k_{i_\ell}-1},$$

alors on pose

$$\mathbf{h}_n = h_{i_1,0}h_{i_1,1}\ldots h_{i_1,k_{i_1}-1}1\ldots h_{i_\ell,0}h_{i_\ell,1}\ldots h_{i_\ell,k_{i_\ell}-1}1,$$

pour $n = 1, \ldots, 2^\kappa$. Donc, pour $\mathbf{x}$ donné,

$$(-1)^{\mathbf{h}_n \odot \mathbf{x}} = (-1)^{\tilde{\mathbf{h}}_n \odot \mathbf{x}}(-1)^{\sum_{j\in I} x_{j,k_j+1}}, \qquad n = 1, \ldots, 2^\kappa.$$

On veut calculer

$$\begin{aligned}
\tilde{f}(\mathbf{h}_n) &= \int f(\mathbf{x})(-1)^{\mathbf{h}_n \odot \mathbf{x}} d\mathbf{x} \\
&= \int f(\mathbf{x})(-1)^{\tilde{\mathbf{h}}_n \odot \mathbf{x}}(-1)^{\sum_{j\in I} x_{j,k_j+1}} d\mathbf{x} \\
&= \sum_{m=1}^{2^\kappa} \mathbf{H}_{\kappa,n,m} \int_{\mathbf{x}_I(\mathbf{k}_I)=b_m} f(\mathbf{x})(-1)^{\sum_{j\in I} x_{j,k_j+1}} d\mathbf{x} \\
&= \sum_{m=1}^{2^\kappa} \mathbf{H}_{\kappa,n,m} w_m, \qquad \text{(E.3)}
\end{aligned}$$

où $\mathbf{H}_{\kappa,n,m}$ représente l'élément en position (n, m) dans la matrice $\mathbf{H}_\kappa$.

Par (E.3), on a que

$$\sum_{\mathbf{h}\in H(\mathbf{k}_I)} |\tilde{f}(\mathbf{h})|^2 = \mathbf{w}^T \mathbf{H}_\kappa^T \mathbf{H}_\kappa \mathbf{w} = 2^\kappa \sum_{n=1}^{2^\kappa} w_n^2,$$

la deuxième égalité suivant par le fait que $\mathbf{H}_\kappa^T = 2^\kappa \mathbf{H}_\kappa^{-1}$. ∎

Le lemme suivant sert à compter combien de vecteurs dans $H(\mathbf{k}_I)$ font partie du réseau dual. On utilise pour cela le lemme 4.2.1 et la représentation de la fonction $(-1)^{\mathbf{h}\odot\mathbf{x}}$ à l'aide de la matrice d'Hadamard $\mathbf{H}_\kappa$.

Lemme E.2 *Soit $\mathbf{k}_I = (k_j)_{j\in I}$ un vecteur d'entiers $k_j \geq 0$. Soit P_N une règle de réseau polynômiale avec $N = 2^m$. Si on pose $\mathbf{y} = (y_1, \ldots, y_{2^\kappa})^T$ avec*

$$y_n = \sum_{i=1}^{N} \mathbf{1}_{\mathbf{x}_i(\kappa)=b_n} (-1)^{\sum_{j\in I} x_{i,j,k_j+1}},$$

alors le nombre de vecteurs dans $H(\mathbf{k}_I)$ qui font partie du réseau dual $\mathcal{L}_s^*$ est donné par

$$|H(\mathbf{k}_I) \cap \mathcal{L}_s^*| = 2^{-2m} \mathbf{y}^T \mathbf{H}_\kappa^T \mathbf{H}_\kappa \mathbf{y} = 2^{\kappa-2m} \sum_{n=1}^{2^\kappa} y_n^2.$$

Démonstration : pour chaque $\mathbf{h}_n \in H(\mathbf{k}_I)$, posons

$$s_n = \sum_{i=1}^{N} (-1)^{\mathbf{h}_n\odot\mathbf{x}_i}, \qquad n = 1, \ldots, 2^\kappa.$$

Par le lemme 4.2.1, on a que

$$\sum_{n=1}^{2^\kappa} s_n^2 = N^2 |H(\mathbf{k}_I) \cap \mathcal{L}_s^*|.$$

Or, si $\mathbf{h}_n(\mathbf{k}_I)$ représente la troncation de $\mathbf{h}_n$ à des k_j premiers bits dans chaque dimension j, alors

$$\begin{aligned}
\sum_{i=1}^{N} (-1)^{\mathbf{h}_n\odot\mathbf{x}_i} &= \sum_{i=1}^{N} (-1)^{\mathbf{h}_n(\mathbf{k}_I)\odot\mathbf{x}_i(\mathbf{k}_I)} (-1)^{\sum_{j\in I} x_{i,j,k_j+1}} \\
&= \sum_{l=1}^{2^\kappa} \mathbf{H}_{\kappa,n,l} \sum_{i=1}^{N} \mathbf{1}_{\{\mathbf{x}_i(\mathbf{k}_I)=b_l\}} (-1)^{\sum_{j\in I} x_{i,j,k_j+1}},
\end{aligned}$$

en utilisant le fait que $(-1)^{\mathbf{h}_n(\mathbf{k}_I)\odot\mathbf{x}_i(\mathbf{k}_I)} = \sum_{l=1}^{2^\kappa} \mathbf{H}_{\kappa,n,l} \mathbf{1}_{\{\mathbf{x}_i(\mathbf{k}_I)=b_l\}}$ et donc,

$$(s_1, \ldots, s_{2^\kappa})^T = \mathbf{H}_\kappa \mathbf{y}^T.$$

Le résultat suit en utilisant le fait que

$$\sum_{n=1}^{2^\kappa} s_n^2 = \mathbf{y}^T \mathbf{H}_\kappa^T \mathbf{H}_\kappa \mathbf{y} = 2^\kappa \sum_{n=1}^{2^\kappa} y_n^2,$$

puisque $\mathbf{H}_\kappa^T = 2^\kappa \mathbf{H}_\kappa^{-1}$. ∎