Un algorithme d'optimisation par exploration sélective

Didier Rullière, Alaeddine Faleh, Frédéric Planchet

Un algorithme d'optimisation par exploration sélective[1]

Didier Rullière[2] , Alaeddine Faleh[3] et Frédéric Planchet[4]

Résumé

Cet article traite un problème d'optimisation dans le cas où la fonction objectif est estimée à l'aide de simulations. Il présente un algorithme d'optimisation globale d'une fonction non convexe et bruitée. L'algorithme est construit après une étude de critères de compromis entre, d'une part, l'exploration de la fonction objectif en de nouveaux points et d'autre part l'amélioration de la connaissance de celle-ci, par l'augmentation du nombre de tirages en des points déjà explorés. Une application numérique illustre la conformité du comportement de cet algorithme à celui prévu théoriquement. Les performances de l'algorithme sont analysées au moyen de différents critères de convergence. Des zones de confiance sont également proposées. Enfin, une comparaison avec un algorithme classique d'optimisation stochastique est menée.

mots clés : simulation stochastique, optimisation stochastique, optimisation globale, zone de recherche, zone de confiance, exploration, connaissance, allocation optimale, potentiel.

1 Introduction

1.1 Le problème d'optimisation

Nous allons chercher à déterminer l'optimum global d'une fonction réelle définie sur un ensemble Θ, soumise à un bruit, ainsi que les paramètres de Θ conduisant à cet optimum. Ainsi, dans le contexte de l'allocation stratégique optimale d'actifs, une allocation particulière $\theta \in \mathbb{R}^d$, $d \in \mathbb{N}^*$, peut nécessiter de simuler de nombreuses trajectoires de plusieurs actifs, sur une succession de périodes, afin de déterminer un unique indicateur de risque ou de gain, bruité, $F(\theta)$: l'indicateur en question peut être un ratio de financement, un indicateur de gain pénalisé en fonction du risque, ou un indicateur synthétique de compromis risque/gain. Le calcul de chaque $F(\theta)$ est coûteux en terme de temps de calcul, et les allocations conduisant à un optimum de $f(\theta) = \mathrm{E}\left[F(\theta)\right]$ sont recherchées pour l'allocation optimale.

Nous considérons ici une fonction objectif réelle $f(\theta)$ d'un paramètre $\theta \in \Theta$, $\Theta \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$:

$$f : \Theta \to \mathbb{R}, \quad \Theta \subset \mathbb{R}^d,$$

Nous supposerons que f est continue et bornée, non nécessairement dérivable, et que Θ est une union finie de d-simplexes, typiquement un ensemble de pourcentages d'allocation possibles, inclus dans $[0,\ 1]^d$. Enfin, f n'est pas nécessairement une fonction convexe de θ, de sorte que la fonction peut posséder plusieurs optima locaux.

[1] Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-BLAN-0314-01, et d'une aide de l'ANRT portant la référence 177/2008.

[2]Université de Lyon, Université Lyon 1, Laboratoire SAF, Ecole ISFA.
Auteur correspondant : Didier.Rulliere@univ-lyon1.fr, (+33) 4 37 28 74 38, UCBL, ISFA, 50 Avenue Tony Garnier, F-69007 Lyon, France.
[3]Université de Lyon, Université Lyon 1, Laboratoire SAF, Ecole ISFA ; Caisse des Dépôts et Consignations
[4]Université de Lyon, Université Lyon 1, Laboratoire SAF, Ecole ISFA ; Winter & Associés

En outre, nous supposerons que f n'est pas directement connue, mais qu'elle est estimée à l'aide de simulations. En tout point $\theta \in \mathbb{R}^d$, du fait des erreurs d'estimation, l'observateur ne peut accéder qu'à des réalisations d'une variable aléatoire $F(\theta) = f(\theta) + \epsilon(\theta)$, où $\epsilon(\theta)$ représente un bruit d'espérance nulle, dont nous postulerons l'existence d'une variance finie, non nécessairement homogène en θ. On postule que les $\{\epsilon(\theta)\}_{\theta\in\Theta}$ sont mutuellement indépendants. La fonction bruitée F est donc telle que :

$$\forall\theta \in \Theta, \begin{cases} \mathrm{E}\left[F(\theta)\right] = f(\theta) \\ \mathrm{V}\left[F(\theta)\right] < \infty \end{cases} .$$

Nous nous placerons dans le cadre d'une minimisation. Sous ces hypothèses, chercher à optimiser la fonction bornée $f(\theta) = \mathrm{E}\left[F(\theta)\right]$ revient alors à rechercher, à partir de réalisations ponctuelles indépendantes de la fonction bruitée F :

1. L'unique valeur minimale de la fonction objectif f,

$$m^* = \inf_{\theta\in\Theta} \mathrm{E}\left[F(\theta)\right] .$$

2. L'ensemble des paramètres conduisant à une valeur proche de m^* :

$$\mathcal{S}_x = \{\theta \in \Theta, \mathrm{E}\left[F(\theta)\right] \leqslant x\}$$

pour tout réel x donné dans un voisinage de m^*

Le premier point est intéressant, mais s'agissant d'allocation d'actifs, le résultat le plus utile est naturellement le second. En un mot, nous cherchons tous les paramètres θ conduisant à un $f(\theta)$ proche de l'optimum.

En présence d'une incertitude sur l'estimateur de f, la recherche d'un paramètre θ^* conduisant à un optimum global supposé de f nécessite l'exploration de tous les paramètres susceptibles de conduire à un optimum global inférieur. Il sera ainsi nécessaire de connaître l'ensemble $\mathcal{S}_x$ des paramètres conduisant à une valeur estimée de f proche de m^*. En outre, en présence de plusieurs paramètres solutions, seul l'utilisateur de l'algorithme peut décider lequel privilégier. C'est la raison pour laquelle nous rechercherons l'ensemble des paramètres solution, et non un seul point de cet ensemble.

Lorsque la fonction F est déterministe, dans le cas où $f(\theta) = F(\theta)$, différentes méthodes d'optimisation peuvent être proposées, comme les méthodes de descente de gradient, de Newton-Raphson, de Hooke et Jeeves, la méthode de Nelder, Mead (1965), ou des méthodes spécifiquement adaptées à certaines formes de f, comme lorsque f est convexe. Ces méthodes ne garantissent toutefois pas que l'optimum obtenu est un optimum global. S'agissant d'optimisation globale déterministe, les premières recherches datent d'environ trente ans et sont attribuées à Hansen (1979). L'optimisation Lipschitzienne, l'algorithme de Schubert (cf. Schubert, 1972) ou l'algorithme DIRECT (cf. Jones, Pertunen, Stuckman, 1993) sont des méthodes largement utilisées dans un cadre déterministe. D'une façon plus générale, ces algorithmes peuvent s'intégrer dans le cadre d'algorithmes de type Branch and Bound (cf. Lawler, Wood, 1966), où la zone à explorer (ici Θ) est partitionnée en plusieurs zones (*branching*), dont certaines sont exclues de l'analyse selon certains critères (*bounding*). Les critères d'exclusions de ce type d'algorithme reposent notamment sur des propriétés de la fonction objectif et sur une arithmétique d'intervalle (cf. Wolfe, 1996), on parle alors d'optimisation globale déterministe, permettant de garantir avec certitude l'absence d'optimum sur les zones exclues. En considérant que f est le résultat d'une expérience, généralement coûteuse en terme de temps de calcul, les techniques

d'optimisation entrent dans le cadre de *Computer Experiments* et la représentation de la fonction en dehors des points d'observation peut s'appuyer sur différentes techniques de régression, de Krigeage, ou de champs gaussien (cf. Krige, 1951; Jones, Schonlau, Welch, 1998; Santner, Williams, Notz, 2003). Une revue des différentes techniques utilisées dans le champs de l'optimisation globale pourra être trouvée dans Horst, Parlados (1995), ainsi que dans plusieurs thèses récentes (cf. Emmerich, 2005; Ginsbourger, 2009; Villemonteix, 2009).

Lorsque la fonction F est aléatoire, lors de la recherche d'un unique optimum non nécessairement global, une très large littérature existe sur les algorithmes stochastiques pour la recherche d'optimum. Lorsque $d = 1$ des algorithmes tels que celui de Kiefer-Wolfowitz (voir Kiefer, Wolfowitz, 1952) peuvent être utilisés. Dans le cas $d > 1$, une extension due à Blum peut être exploitée (voir Blum, 1954). Strugarek (2006) présente un ensemble d'algorithmes plus récents et plus détaillés portant sur l'optimisation stochastique. S'agissant de la recherche de racines, que nous pourrons également traiter avec l'algorithme ici présenté, des techniques classiques sont celles dérivées de l'algorithme de Robbins-Monro (Robbins, Monro, 1951). De même, l'algorithme de Cohen, Culioli permet de résoudre des problèmes très proches avec une contrainte de contrôle de la probabilité de ruine (voir Cohen, Culioli, 1994). Nous ne détaillerons pas ici l'ensemble des techniques utilisées dans les champs de l'optimisation stochastique (recherche d'optimum), ou de l'approximation stochastique (recherche de racines).

Enfin, lorsque la fonction F est aléatoire, et lorsque l'on cherche un minimum global d'une fonction non nécessairement convexe, le bruit complique encore le problème difficile de l'optimisation globale (voir Bulger, Romeijn, 2005, pour une discussion sur ce sujet). Certaines méthodes stochastiques ont été adaptée à un environnement bruité : on peut citer les méthodes génétiques (voir Alliot, 1996; Mathias et al., 1996, en présence de bruit), ainsi que les méthodes de recuit simulé (voir Aarts, Laarhoven, 1985; Branke, Meisel, Schmidt, 2008, en présence de bruit). Encore peu connues dans le domaine de l'actuariat, quelques extensions d'algorithmes de type Branch and Bound sont également proposées dans un cadre stochastique (cf. par exemple Norkin, Pflug, Ruszczyński, 1996).

Le risque de se tromper d'optimum Lorsque F est aléatoire, on peut envisager l'usage d'algorithmes stochastiques d'optimisation locale. Toutefois, tout comme les méthodes déterministes de descente de gradient, ces algorithmes requièrent de choisir un point de départ suffisamment proche d'un optimum global, et d'éviter les plus évidents optima locaux. Sauf pour quelques fonctions f particulières, il s'agit alors d'appréhender dans un premier temps la forme de la fonction f. Une exploration globale de la fonction s'avère donc souvent nécessaire, le risque étant moins de mal estimer la valeur d'un optimum global, que de se tromper d'optimum en choisissant indûment un mauvais optimum local.

Une seconde piste serait d'établir un grand nombre de tirages de F, afin de se ramener, à une marge d'erreur près, au cas déterministe. Cela autoriserait l'usage d'algorithmes d'optimisation locale déterministe, généralement rapides. Toutefois, la recherche d'un optimum local se ferait au prix d'une exploration préalable de la fonction. De surcroît, l'algorithme utilisé serait lourdement pénalisé par le grand nombre de tirages requis pour F. De la même façon, cette pénalité frapperait les algorithmes d'optimisation globale déterministe.

Enfin, les méthodes génétiques ou le recuit simulé ne visent pas directement à proposer des zones de confiance pour les optimiseurs de la fonction objectif. Il nous a semblé délicat de quantifier, avec ces méthodes, le risque de proposer un optimum local et d'ignorer un optimum global meilleur.

Explorer ou connaître A titre illustratif, plaçons nous un instant dans un cadre simplifié, en présence d'un unique minimiseur $\theta^* = \arg\min_{\theta\in\Theta} f(\theta)$. Supposons que l'observateur puisse réaliser n tirages de F en chaque point d'un ensemble $\Theta_m = \{\theta_1, \ldots, \theta_m\}$. En supposant donnée la valeur n, qui détermine la précision de la connaissance de f, et Θ_m qui détermine l'ampleur de l'exploration de f, un estimateur envisageable $\widehat{\theta}^*$ de l'unique minimiseur θ^* est le suivant :

$$\begin{aligned} \widehat{f}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} F_i(\theta), \quad \theta \in \Theta_m\,, \\ \widehat{\theta}^*_{n,m} &= \arg\min_{\theta\in\Theta_m} \widehat{f}_n(\theta)\,. \end{aligned}$$

La question se pose ici de l'arbitrage entre le choix d'un n élevé ou d'un cardinal de Θ_m élevé. Les algorithmes usuels d'optimisation supposent généralement que l'ensemble des points $F(\theta)$ sont connus, et ignorent, à notre connaissance, la question pratique de l'optimisation du couple de paramètre (n, m). En effet, la détermination d'une réalisation de la variable aléatoire F peut prendre un temps de calcul important. En conséquence, n réalisations de F en chaque point de Θ_m conduisent à un total de $n \cdot m$ tirages de F. Or, les tirages étant coûteux en terme de temps de calcul, le nombre de tirages est nécessairement limité. Pour un budget de tirages fixé, le choix d'un nombre de simulations n élevé réduit la variance de $\widehat{f}_n$, et donc améliore ponctuellement la connaissance de f, mais limite également l'exploration de la fonction en d'autres points.

Cette illustration simplifiée introduit la problématique de l'arbitrage entre l'exploration de la fonction en différents points et la connaissance ponctuelle de celle-ci. Trouver les minimiseurs de f va requérir d'une part d'explorer la fonction en différents points (nous parlerons d'*exploration*), et d'autre part d'opérer également plusieurs tirages de F en chaque point exploré pour obtenir un estimateur non biaisé et aussi peu dispersé que possible de l'espérance de la fonction (nous parlerons de *connaissance ponctuelle*). Nous proposerons dans cet article un algorithme qui vise à répondre à cet arbitrage.

Objectif poursuivi L'objectif que nous poursuivrons ici est l'exploration des zones susceptibles de contenir un minimum global de f : nous chercherons donc à explorer sélectivement la fonction f, de façon à avoir une vision globale de celle-ci, tout en privilégiant certaines zones d'intérêts, comme les minima globaux, dans un souci d'économie du nombre de tirages de la fonction F. L'algorithme vise au final à estimer m^* ainsi que $\mathcal{S}_x$, pour x donné dans un voisinage de m^*.

1.2 Approche retenue

L'approche proposée ici s'apparente à une approche de type Branch and Bound, mais aucune zone n'est jamais définitivement exclue de l'analyse, l'idée étant ici d'ordonner les zones en fonction de la probabilité qu'une zone contienne un optimum, selon un modèle que nous détaillerons. Nous nous placerons dans le cadre de la minimisation d'une fonction, nous chercherons ici à quantifier la probabilité qu'une zone contienne un minimum plus petit qu'un minimum observé, pour au final construire l'ensemble des zones susceptibles de contenir un minimum global.

Grille à pas fixe Une solution simple et très commune est l'utilisation d'une grille de simulation, à pas fixe. Selon cette solution, la fonction f est explorée sur un ensemble Θ_m. Θ_m est une grille de pas δ, toutes les composantes de $\theta \in \Theta_m$ parcourent l'ensemble des valeurs multiples de δ dans Θ : $\Theta_m = \left\{\theta \in \Theta, \frac{1}{\delta}\theta \in \mathbb{N}^d\right\}$. Le nombre n de tirages de $F(\theta)$ est identique

pour chaque $\theta \in \Theta_m$. Cette solution est néanmoins très onéreuse en terme de temps de calcul, dans la mesure où n simulations seront conduites sur chacun des points, y compris ceux très éloignés d'un optimum, pour lesquels la meilleure connaissance de f n'apporte quasiment rien.

Grille à pas variable Nous chercherons donc à mettre en place une grille à pas variable, où les retirages de F se feront principalement dans les zones susceptibles d'accueillir le minimum. Il va s'agir d'une part d'estimer f en de nouveaux points θ (sommets), et d'autre part de répartir de nouvelles simulations entre les anciens sommets et les nouveaux, en fonction de l'intérêt que peut avoir, sur la connaissance du minimum global, l'ajout de simulations en chacun de ces points. Les positions des sommets envisagés ne seront plus régulièrement répartis, et le nombre de simulations conduites en chaque sommet va différer selon les sommets. Une première difficulté est le choix d'une forme convenable pour les zones de recherche, pour les cellules de la grille. Une autre difficulté est que l'on a besoin de deviner où conduire les futures simulations. Pour cela, il faut avoir une idée de comment va évoluer la fonction entre les points explorés : il faut en un sens deviner quel pourra être l'impact de futures simulations avant même de les réaliser.

Algorithmes présentés et structure du document Nous allons aborder dans cet article deux algorithmes réalisant une grille à pas variable. Dans la section 2, nous envisagerons le cas où le nombre de tirages en chaque point est fixe, l'algorithme cherchant alors uniquement à déterminer les prochains points où évaluer la fonction F. Dans la section 3, nous introduirons la possibilité d'opérer des retirages en des points déjà explorés. Enfin, dans une dernière section 4, nous présenterons des illustrations simples et visuelles du comportement de l'algorithme proposé sur une fonction bimodale élémentaire. Une comparaison avec un algorithme classique d'optimisation stochastique sera menée, nous évoquerons également dans cette section le problème de la dimension.

2 Une grille à pas variable par subdivision systématique

2.1 Forme des zones de recherche

Zone initiale de recherche En pratique, par exemple lors de l'optimisation de choix d'investissements de $d+1$ actifs, les proportions investies dans chacun des actifs se somment à un. Il suffit alors de rechercher d pourcentages d'allocation, l'allocation numéro $d+1$ se déduisant des d allocations précédentes. Il s'agit donc d'optimiser une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$.

Nous nommerons $Z_0 \subset \mathbb{R}^d$ la zone initiale de recherche, simplexe standard orthogonal constitué des sommets $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, ... ,$(0, \ldots, 0, 1)$, ainsi que du sommet $(0, \ldots, 0)$. Cela correspondra bien à la situation où la seule contrainte est d'avoir une somme des composantes inférieure à un, par exemple d pourcentages d'allocation d'actifs dont la somme est inférieure à 100%, la différence avec 1 formant l'allocation de l'actif numéro $d+1$:

$$Z_0 = \left\{(x_1, \ldots, x_d) \in \mathbb{R}^d,\ x_1 + \cdots + x_d \leqslant 1, x_1 \geqslant 0, \ldots x_d \geqslant 0\right\} .$$

Dans les situations où l'optimum est à rechercher sur une zone plus complexe soumise à de nombreuses contraintes, l'algorithme proposé sera applicable si cette zone peut être représentée dès l'initialisation par une union finie de simplexes. Dans tous les cas, nous supposerons que l'ensemble des sommets de la ou des zones initiales ont été explorés, par la réalisation de tirages de $F(\theta)$ en chacun des sommets de l'enveloppe de ces zones.

Mécanisme de scission L'idée d'une grille à pas variable est de séparer la zone de recherche de l'optimum en plusieurs zones de différentes tailles. A chaque étape, la scission d'une zone peut se faire en opérant des tirages de F en un ou plusieurs points non encore explorés de Θ.

Certains algorithmes d'optimisation globale fonctionnent, dans un contexte particulier déterministe, par subdivision de la zone à explorer. On peut notamment citer l'algorithme DIRECT (pour DIviding RECtangles, cf. Jones, Pertunen, Stuckman, 1993), qui subdivise un pavé de $\mathbb{R}^d$ en plusieurs sous-pavés. Si la subdivision d'une zone nécessite l'exploration de n^+ nouveaux points, il paraît préférable de choisir $n^+ = 1$. Dans ce seul cas, le choix d'un nouveau point de subdivision se fait alors en connaissance de tous les précédents tirages de la fonction. Nous avons ici choisi un mode de division qui d'une part nous semblait plus adapté aux problèmes définis sur une union de simplexes, et d'autre part ne nécessitait l'exploration que d'un unique point à chaque subdivision.

Nous supposerons que les zones de recherche sont des ensembles convexes, de façon à faciliter d'éventuelles interpolations au cœur de chaque zone. Imaginons un pavé de $\mathbb{R}^d$ dont on connaît les sommets, et à l'intérieur duquel on ajouterait un point θ_c. Comment découper rapidement ce pavé en une partition d'ensembles convexes dont l'enveloppe convexe contiendrait θ_c ? La réponse n'étant pas si évidente, nous opterons pour le choix suivant :

- Les zones considérées seront délimitées par $d+1$ sommets.
- Chaque nouveau point sera ajouté sur un segment de l'enveloppe convexe de la zone.
- A chaque étape, chaque zone sera éventuellement scindée en deux ensembles convexes.

Tout point à l'intérieur d'une zone appartiendra à une zone convexe délimitée par $d+1$ points, et le choix d'un nombre de sommets égal à $d+1$ facilitera par la suite la séparation d'une zone en plusieurs zones. D'autre part, le choix d'un nombre fixe de sommets délimitant chaque zone sera de nature à faciliter l'implémentation de l'algorithme.

Nous appellerons zone un d-simplexe, c'est-à-dire un ensemble convexe inclus dans Θ, avec $\Theta \subset \mathbb{R}^d$, dont l'enveloppe convexe est déterminée par $d+1$ sommets distincts (les sommets formant un repère affine de $\mathbb{R}^d$). Par la suite, nous noterons $S(Z)$ l'ensemble des $d+1$ sommets délimitant une zone convexe Z. La zone initiale de recherche de l'optimum est notée Z_0. A l'issue de l'étape numéro k, la zone de recherche est subdivisée en un ensemble de zones recouvrant Z_0, dont l'intersection est de mesure nulle. Cet ensemble de zones est noté $\mathfrak{Z}_k$.

Considérons un ensemble de sommets E délimitant une zone Z et deux points distincts θ_1 et θ_2 de cet ensemble. Supposons que θ_c soit le barycentre (par exemple équipondéré) entre ces deux points. Lorsque la décision est prise de scinder cette zone en deux zones Z_1 et Z_2, les ensembles de sommets délimitant ces deux zones seront :

$$E_1 = (E \setminus \theta_1) \cup \{\theta_c\} \ ,$$
$$E_2 = (E \setminus \theta_2) \cup \{\theta_c\} \ .$$

Lemme 2.1 (lemme de séparation) *Les zones Z_1 et Z_2 comportent $d+1$ sommets, sont convexes, recouvrent Z, et leur intersection est incluse dans un hyperplan de $\mathbb{R}^d$.*

Une preuve de ce lemme est donnée en appendice.

Par souci de simplicité, nous nommerons par la suite *sommets* d'une zone Z l'ensemble des $d+1$ sommets délimitant l'enveloppe convexe de la zone Z.

La figure 1 illustre en dimension 2 la façon dont une zone peut être progressivement subdivisée.

On peut remarquer que le choix d'une décomposition de l'ensemble Θ en un ensemble de zones contenant des points de tirage $\{\theta_i\}_{i=1,2...,n}$ est un problème de triangulation, classique

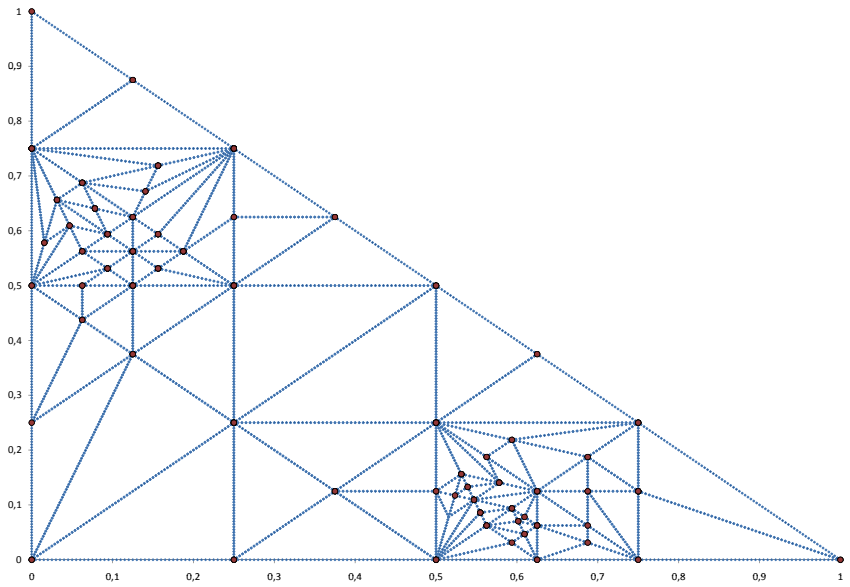

FIG. 1 – Un exemple de partition d'une zone dans $\mathbb{R}^2$ en 60 zones. ($d = 2$)

en géométrie algorithmique. Une technique de triangulation très connue est la triangulation de Delaunay (cf. De Berg, Cheong, van Kreveld, Overmars, 2008). Le problème étant ici de choisir un point de tirage au sein d'une zone et non pas de construire des zones pour séparer des points de tirage existants, les zones seront scindées par la simple bissection présentée. Un des avantages de la bissection présentée est de permettre l'exploration des frontières de la zone initiale. D'autres choix de subdivision de zones sont naturellement possibles, par exemple autour de l'isobarycentre d'une zone. Comme on peut l'observer sur la figure 1 ainsi que sur les différentes illustrations de la section 4, l'union des enveloppes convexes des zones scindées forme un ensemble beaucoup plus vaste que l'enveloppe convexe du simplexe initial, de sorte que les points de tirage ne sont pas condamnés à rester dans l'enveloppe de la zone initiale. Dans le cas de barycentres équipondérés, on peut d'ailleurs montrer par récurrence qu'il est possible d'atteindre en un nombre fini d'étapes tout point de Θ de coordonnées :

$$\left(\frac{i_1}{2^{k_1}}, \ldots, \frac{i_d}{2^{k_d}}\right) \text{ avec pour tout } j \in \{1, \ldots, d\},\ k_j \in \mathbb{N}, i_j \in \mathbb{N}, i_j \leqslant 2^{k_j}\ .$$

2.2 Choix de la zone à explorer ou segmenter

Idée du potentiel d'une zone L'étude de la possibilité pour f d'atteindre un minimum global sur une zone non explorée nécessite de fixer des hypothèses : si f est supposée extrêmement erratique, f pourra franchir un seuil inférieur sur à peu près n'importe quelle zone, et les tirages opérés de F n'apporteront que très peu d'information. Une solution classiquement retenue, dans le domaine de l'optimisation globale déterministe, est le choix d'une forme Lipschitzienne pour f (cf Jones, Pertunen, Stuckman, 1993) : cela revient à dire, dans un cadre déterministe, que f pourra atteindre un minimum global sur une zone si la pente nécessaire pour cette atteinte est inférieure à un certain seuil, ou que f ne pourra pas atteindre le minimum global.

Cette logique binaire permet d'indiquer de façon certaine si f appartient ou non à un in-

tervalle donné, et correspond à une logique d'arithmétique d'intervalle (cf. Wolfe, 1996). En l'absence d'information très précise sur f, nous avons préféré une logique probabiliste, malgré une part de subjectivité qu'elle peut engendrer : plutôt qu'une indicatrice de franchissement possible à valeur dans $\{0, 1\}$, nous allons rechercher une mesure sur $[0,\ 1]$ quantifiant la probabilité, selon la représentation de f par l'observateur, que f franchisse le seuil sur une zone. Si cette solution introduit une nécessaire subjectivité sur le calcul de cette probabilité, elle offre néanmoins l'avantage de maintenir une hiérarchie entre différentes zones, ce qui est notamment utile lorsque beaucoup de zones sont susceptibles de contenir l'optimum, et permet de ne jamais exclure a priori de zone.

Comme le temps de calcul de F est supposé beaucoup plus important que le temps d'exécution de l'algorithme de choix des zones à explorer, nous choisirons à chaque étape de l'algorithme une unique zone à scinder en deux ou à explorer davantage. L'objectif étant de limiter les risques qu'une zone non explorée contienne un optimum, nous piocherons à chaque étape une zone à explorer, avec une probabilité proportionnelle à un coefficient $\boldsymbol{\beta}$ spécifique à la zone.

Pour chaque zone Z, le coefficient $\boldsymbol{\beta}(Z)$ déterminera s'il est plausible (dans un sens que nous préciserons) que la zone contienne un minimum plus petit que le minimum estimé m^*. Afin d'approcher ce coefficient sur toute la zone, nous aurons besoin de le déterminer en un unique point. Selon un modèle probabiliste que nous préciserons, nous nommerons "potentiel" d'un point une mesure d'autant plus grande que la fonction objectif est susceptible, en ce point, d'être plus petite que le minimum global observé. Une représentation probabiliste du potentiel d'un point passe par la modélisation de la connaissance incertaine de la fonction f au vu des tirages opérés en différents points, par exemple au moyen de champs aléatoires conditionnels. Nous présentons ici une approche plus simple partant d'une modélisation de l'incertitude en dimension 1, puis agrégeant différentes incertitudes compte tenu des sommets explorés.

Potentiel d'un point en direction d'un sommet Considérons un sommet θ' déjà exploré d'une zone Z, et un point θ de cette zone, distinct de θ'. Par souci de simplicité, nous étudierons ici l'évolution de la fonction f entre ces deux points θ et θ' en connaissance de la seule exploration en θ'.

Même en l'absence d'erreur d'échantillonnage au sommet θ', l'évolution de la fonction f est naturellement inconnue sur le segment $]\theta',\ \theta]$: nous parlerons *d'erreur de grille*. Nous supposerons que, à partir de la seule connaissance de f au point θ', l'observateur représente $f(\theta)$ par une variable aléatoire $\tilde{f}_{\theta'}(\theta)$, choisie de loi normale, d'espérance $f(\theta')$ et dont l'écart-type est une fonction croissante de la distance $d(\theta, \theta')$. Nous choisirons :

$$\begin{aligned} \tilde{f}_{\theta'}(\theta) - f(\theta') &\sim N\left(0, \sigma_g^2\right) , \\ \sigma_g(\theta, \theta') &= \sigma_K d(\theta, \theta')^\alpha , \qquad \sigma_K > 0, \alpha > 0 . \end{aligned}$$

Cela revient à décrire l'incertitude maximale pesant sur $\tilde{f}$, lorsque σ_g est soumis à une condition de type Hölder (ou de type Lipschitz lorsque $\alpha = 1$). A titre d'illustration, pour $\theta_0 \in [\theta', \theta]$, la représentation $\tilde{f}_{\theta'}(\theta_0) - f(\theta')$ par un mouvement brownien correspond à $\alpha = 1/2$. La représentation de la pente $(\tilde{f}_{\theta'}(\theta_0) - f(\theta'))/d(\theta_0, \theta')$ par un mouvement brownien correspond à $\alpha = 3/2$. Nous évoquerons plus en détail le choix de α et σ_K au paragraphe 2.3, ainsi que dans la section d'application numérique 4.

En présence d'erreur d'échantillonnage, on dispose d'un estimateur $\widehat{f}(\theta')$ de $f(\theta')$. Si $\widehat{f}(\theta') - f(\theta')$ est elle-même une variable aléatoire de loi normale centrée et de variance $\sigma_e^2(\theta')$, indépendante de l'erreur de grille $\tilde{f}_{\theta'}(\theta) - f(\theta')$, alors :

$$\tilde{f}_{\theta'}(\theta) - \widehat{f}(\theta') \sim N\left(0, \sigma_e^2(\theta') + \sigma_g^2(\theta, \theta')\right) .$$

En connaissance d'une réalisation de $\widehat{f}(\theta')$, nous noterons :

$$\frac{1}{2}L(\theta,\theta') = \mathrm{P}\left[\tilde{f}_{\theta'}(\theta) \leqslant m^*\right],$$

où m^* représente la valeur du minimum global de $f(\theta)$.

La quantité $L(\theta,\theta')$ sera nommée potentiel directionnel du point θ en direction de θ'. Le facteur de normalisation $\frac{1}{2}$ n'a aucune incidence sur les comparaisons des potentiels entre eux (comme tout facteur strictement positif), mais permettra par la suite à la quantité L de pouvoir atteindre toutes les valeurs de $[0,\ 1]$, et non pas seulement $[0, \frac{1}{2}]$. Le potentiel directionnel s'interprète alors comme la probabilité que $\tilde{f}_{\theta'}$ franchisse un minimum global connu sachant que $\tilde{f}_{\theta'}$ décroît depuis le point d'accroche θ'.

Il est possible de tenir compte de l'erreur d'estimation du minimum m^* : en supposant que l'estimateur $\widehat{m}^*$ est une variable aléatoire indépendante de $\tilde{f}_{\theta'}(\theta) - \widehat{f}(\theta')$, de loi normale, d'espérance m^* et de variance $\sigma^2_{m^*}$,

$$\begin{aligned}\frac{1}{2}L(\theta,\theta') &= 1 - \Phi\left(\frac{\widehat{f}(\theta') - \widehat{m}^*}{\sigma_T(\theta,\theta')}\right), \quad \sigma_T(\theta,\theta') > 0\\ \text{avec } \sigma^2_T(\theta,\theta') &= \sigma^2_{m^*} + \sigma^2_e(\theta') + \sigma^2_K d(\theta',\theta)^{2\alpha}.\end{aligned}$$

La fonction Φ désigne la fonction de répartition d'une loi normale centrée réduite. Nous prendrons la convention $L(\theta,\theta') = \mathbb{1}_{\widehat{f}(\theta')=\widehat{m}^*}$ dans le cas où $\sigma_T(\theta,\theta') = 0$, en l'absence de bruit et lorsque $\sigma_K = 0$ ou $\theta = \theta'$. Nous n'évoquerons pas en détail la détermination très classique, en un point exploré θ, de l'estimateur $\widehat{f}(\theta)$ de $f(\theta)$, ni de l'estimateur de la variance de $\widehat{f}(\theta)$. L'usage de la moyenne et de la variance empirique non biaisée donnerait, à partir de n_0 observations de $F(\theta)$ notées $F_1(\theta),\ldots,F_{n_0}(\theta)$, $n_0 \geqslant 2$:

$$\begin{aligned}\widehat{f}(\theta) &= \frac{1}{n_0}\sum_{i=1}^{n_0} F_i(\theta),\\ \sigma^2_e(\theta) &= \frac{1}{n_0}\widehat{\sigma}^2_F(\theta), \text{ avec } \widehat{\sigma}^2_F(\theta) = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(F_i(\theta) - \widehat{f}(\theta))^2.\end{aligned}$$

$\sigma^2_F(\theta)$ désigne ici la variance de $F(\theta)$, et $\widehat{\sigma}^2_F(\theta)$ un estimateur de cette variance. Des raffinements peuvent être envisagés afin de tenir compte de l'erreur d'estimation de σ_F. Notons surtout la nécessité d'un paramètre $n_0 \geqslant 2$, nombre de tirages requis pour l'estimation de la variance empirique $\sigma^2_e(\theta)$. Dans le cas déterministe, on peut fixer $n_0 = 1$, $\widehat{f}(\theta) = f(\theta)$ et $\sigma^2_e(\theta) = 0$. Mais dans le cas général, ce paramètre n_0 restera une entrée de l'algorithme. Eventuellement, l'usage d'une hypothèse de répartition gaussienne pour $F(\theta)$ peut permettre de ne pas mémoriser l'intégralité des tirages de $F(\theta)$, mais de simplement mémoriser le nombre de tirages précédents, leur somme et la somme de leurs carrés.

Pour l'estimation de m^* et de $\sigma^2_{m^*}$, nous retiendrons en première approche les valeurs obtenues pour $\widehat{f}(\theta^*)$ et $\sigma^2_e(\theta^*)$ au dernier point optimal θ^* rencontré parmi les tirages réalisés itérativement par l'algorithme, bien que là encore des perfectionnements puissent être suggérés.

Remarquons que le potentiel peut être adapté à d'autres recherches que la recherche d'optimum. Ainsi, si l'on recherchait les paramètres θ tels que $f(\theta)$ appartienne à un ensemble $A \subset \mathbb{R}$, on pourrait utiliser un potentiel proportionnel à $\mathrm{P}\left[\tilde{f}_{\theta'}(\theta) \in A\right]$. L'algorithme peut ainsi être très facilement adapté à la recherche de racines.

Potentiel d'un point Considérons un point θ à l'intérieur d'une zone Z, et cherchons à définir une mesure d'autant plus grande qu'un minimum global de f peut être observé au point θ. Ce point θ est muni de $d+1$ potentiels directionnels en direction des $d+1$ sommets délimitant la zone Z, sommets notés ici θ_i, $i \in \{1, \ldots, d+1\}$. Ces potentiels marquent chacun la probabilité, selon la modélisation de l'observateur, que la fonction f franchisse le seuil m^* sur l'un des segments $[\theta, \theta_i]$. Les potentiels de points de différentes zones sont destinés à être comparés entre eux. La principale exigence est ici de respecter la logique booléenne selon laquelle si le minimum ne peut pas être atteint en direction d'un sommet, alors il est exclu qu'il soit atteint sur la zone. Une mesure commode respectant cette exigence est naturellement le produit, qui correspond bien au ET booléen lorsque les quantités $L(\theta, \theta_i)$ appartiennent à $\{0,\ 1\}$.

Nous appellerons potentiel du point θ dans la zone Z la quantité :

$$\beta_Z(\theta) = \left(\prod_{i=1}^{d+1} L(\theta, \theta_i) \right)^{\gamma} .$$

Nous avons choisi ici d'introduire une fonction lien $g(x) = x^\gamma$, pour $\gamma > 0$, qui préserve les propriétés requises de logique booléenne du produit, et permet de modifier le comportement de l'algorithme. Le choix de γ, régissant la convexité de g, permet de distordre les potentiels et de moduler ainsi l'importance relative accordée à la zone de plus grand potentiel. Au niveau de l'impact de la distorsion, pour un coefficient γ très grand, le tirage d'une zone reviendrait au tirage systématique de la zone dont le potentiel est le plus grand. Un coefficient γ faible conduirait à davantage explorer des zones de potentiel médian. Le facteur γ s'interprète comme un facteur de priorité attribuée aux zones de grand potentiel.

Nous avons constaté que l'effet de γ était modeste pour des valeurs raisonnables de paramètres (cf section 4). Considérant par ailleurs une certaine redondance de ce paramètre avec les paramètres (σ_K, α) du potentiel, nous opterons sauf mention contraire pour

$$\gamma = 1/(d+1)\,,$$

qui a le mérite de clarifier l'interprétation du potentiel : $\beta_Z(\theta)$ correspondra simplement à la moyenne géométrique des potentiels directionnels. Le potentiel d'un point s'interprétera alors comme un potentiel directionnel moyen.

Le choix d'une mesure pour le potentiel d'un point justifierait de nombreuses études. S'agissant d'agrégation de potentiels directionnels et de fonctions liens, d'autres distorsions de probabilités facilement utilisables pourront être trouvées dans Bienvenüe, Rullière (2010). Une autre piste pour cette agrégation est l'usage de copules, ou encore celle de la logique floue (cf. Zadeh, 1966). Nous étudions actuellement d'autres mesures de ce potentiel (cf. Rullière, Faleh, Planchet, 2010), basés sur l'agrégation des variances directionnelles $\sigma_T(\theta, \theta')$, ainsi que la modélisation de $\tilde{f}$ par des champs aléatoires, en lien avec la théorie du Krigeage (cf. par exemple Krige, 1951; Jones, Schonlau, Welch, 1998; Jones, 2001). Ces mesures présentent l'avantage d'éliminer l'agrégation de potentiels directionnels, au prix d'un modèle d'une complexité parfois accrue.

Il faut néanmoins tempérer l'impact du choix d'une mesure de potentiel : les potentiels des différents points serviront à établir une hiérarchie des zones les plus susceptibles de contenir un optimum global, afin de choisir laquelle explorer en priorité. Le choix de la fonction lien g croissante modifiera les priorités d'exploration, mais ne modifiera pas la hiérarchie elle-même, et celle-ci dépendra en très grande partie de l'éloignement de f à la valeur de l'optimum.

Si les potentiels directionnels appartenaient à $\{0, 1\}$, un potentiel permettrait simplement de dire si il est possible ou non que f possède un minimum global sur une zone, compte tenu des

sommets adjacents et de la valeur estimée du minimum. Cela correspond à l'idée développée dans les algorithmes mettant en oeuvre une arithmétique d'intervalle (cf. Wolfe, 1996). Le potentiel proposé ici doit être vu comme une simple mesure, à valeur sur $[0,\ 1]$, permettant d'étendre une logique d'arithmétique d'intervalle qui conduirait à des potentiels définis sur $\{0, 1\}$.

Potentiel d'une zone En pratique, une zone sera d'autant plus susceptible de contenir un minimum plus petit que la valeur estimée de m^* si sa surface est grande et si ses points ont un potentiel élevé. Une mesure logique de la "surface probable" d'une zone Z est donnée par :

$$\bar{\boldsymbol{\beta}}(Z) = \int_{\theta \in Z} \beta_Z(\theta) d\theta \,.$$

Bien que le calcul de cette intégrale puisse être approché par des techniques de simulation, nous avons préféré retenir comme mesure du potentiel d'une zone la mesure suivante :

$$\boldsymbol{\beta}(Z) = V(Z) \cdot \beta_Z(\theta_{B_Z}) \,,$$
$$\theta_{B_Z} \text{ isobarycentre des sommets de la zone } Z \,,$$

où $V(Z)$ est le volume de la zone Z (hypervolume dans le cas $d > 3$).

Cette solution a le mérite de la simplicité, offre l'avantage d'être très rapide et de ne pas être aléatoire. Un calcul d'un $\boldsymbol{\beta}(Z)$ précédemment mené n'aura donc pas à être réitéré si la zone Z n'est pas modifiée. En outre, l'objectif est de comparer les coefficients $\boldsymbol{\beta}$ entre eux : le calcul fin de l'intégrale a peu de raisons explicites de beaucoup perturber la hiérarchie entre les différentes zones. Enfin, l'isobarycentre nous a paru bien rendre compte de l'erreur de grille au sein de la zone, et facilitera l'interprétation future du potentiel d'une zone.

S'agissant du calcul du volume $V(Z)$, dans le cas où $d = 2$, les zones sont des triangles et un volume $V(Z)$ est donné par la formule de Héron. Dans le cas général, le déterminant de Cayley-Menger donne le volume exact de la zone (cf. Sommerville, 1958). Plus simplement, dans le cas d'une séparation d'une zone en deux volumes égaux, ce qui sera ici le cas, la simple mémorisation du volume de la zone à subdiviser permet de déduire immédiatement le demi-volume de chaque zone fille.

Choix d'une zone en fonction des potentiels Comme nous l'avons évoqué, le choix d'une zone Z^+ parmi un ensemble $\mathfrak{Z}$ de zones se fera de la façon suivante : la probabilité de piocher une zone sera proportionnelle au potentiel de chaque zone.

Notons $\mathfrak{Z} = \{Z_1, \ldots, Z_n\}$ l'ensemble des zones dans lequel doit être piochée Z^+. Si $\{U_\nu\}_{\nu=1,2\ldots}$ désigne une suite de variables aléatoires de loi uniforme sur $[0, 1]$, mutuellement indépendantes, piocher une zone parmi n avec une probabilité fixée au prorata de son potentiel revient à choisir, à une étape ν de l'algorithme :

$$Z^+ = Z_{k^*(U_\nu)} \,,$$
$$\text{avec } \left\{ \begin{array}{l} k^*(u) = \min \{k \in \{1, \ldots, n\} \,,\, B_k \geqslant u \cdot B_n\} \,, \\ B_k = \sum_{i=1}^{k} \boldsymbol{\beta}(Z_i) \,, k \in \{1, \ldots, n\} \,. \end{array} \right.$$

Ce choix découle de plusieurs idées. D'une part, l'idée d'explorer de façon uniforme la zone de recherche lorsque les potentiels sont égaux, d'autre part l'idée de préserver, à la façon d'un recuit simulé, la possibilité d'exploration de zones a priori peu prometteuses. On imagine, en

présence d'un très grande nombre de très petites zones, que cette solution limite le risque de confiner les points de tirage dans le voisinage d'un unique minimiseur, et favorise ainsi une certaine prudence dans l'exploration de la fonction. D'autres choix possibles sont évoqués dans la section d'applications numériques (section 4).

2.3 Choix des paramètres (σ_k, α) du potentiel

Au moyen du potentiel que nous avons défini, nous avons transféré une part de la subjectivité du choix du prochain point de tirage sur le choix de quelques paramètres, au premier rang desquels se trouvent les coefficients de type Hölder α et σ_K. Le choix ou l'estimation de ces paramètres est un problème délicat qui nécessiterait à lui seul une étude poussée, et que nous présentons ici de façon simplifiée.
Le choix des paramètres α et σ_K dépend de la connaissance de la fonction f considérée, ainsi que, lorsque celle-ci s'avère insuffisante, de la prudence de l'observateur. Le choix du paramètre α est un choix de modèle : envisage-t-on que f puisse varier brutalement sur un très petit intervalle comme le ferait une trajectoire de mouvement brownien, ou de façon plus régulière, comme une fonction lipschitzienne ?

Considérons deux points distincts θ et θ', en ignorant dans un premier temps l'erreur d'échantillonnage aux sommets explorés. Si seul θ' est exploré, l'observateur suppose a priori que $Y_{\theta,\theta'} = \tilde{f}_{\theta'}(\theta) - f(\theta')$ est distribué selon une loi normale centrée, d'écart-type $\sigma_K d^\alpha$, où $d = d(\theta, \theta')$, $d > 0$. Après exploration des deux points, l'observateur dispose d'une observation (d, y) de ce couple (d, Y). En supposant que sont collectées n réalisations (d_i, y_i), en les supposant de surcroît mutuellement indépendantes, une estimation maximum de vraisemblance de σ_K en connaissance de α conduirait à :

$$\sigma_K^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i^2}{d_i^{2\alpha}} ,$$

et une estimation maximum de vraisemblance de α en connaissance de σ_K conduirait à α tel que :

$$\sum_{i=1}^{n} \frac{y_i^2 \ln(d_i)}{\sigma_K^2 d_i^{2\alpha}} = \sum_{i=1}^{n} \ln(d_i) .$$

Si l'on considère que, en tout point exploré θ, $\widehat{f}(\theta)$ subit une erreur d'estimation, on peut observer $Y'_{\theta,\theta'} = (\tilde{f}_{\theta'}(\theta) + N_\theta) - (f(\theta') + N_{\theta'})$, où N_θ et $N_{\theta'}$ sont des variables aléatoires indépendantes, de lois normales, de variances respectives σ_θ^2 et $\sigma_{\theta'}^2$. Alors $Y'_{\theta,\theta'}$ est supposée distribuée selon une loi normale centrée, de variance $\sigma_e^2 + \sigma_K^2 d^{2\alpha}$, où $\sigma_e^2 = \sigma_\theta^2 + \sigma_{\theta'}^2$ représente une erreur d'estimation, connue après exploration des deux points puisque mesurée aux points d'observation. L'estimation maximum de vraisemblance de σ_K (sachant α) et de α (sachant σ_K), conduit respectivement à σ_K et α tels que :

$$\sum_{i=1}^{n} \frac{d_i^{2\alpha}}{\sigma_{e_i}^2 + \sigma_K^2 d_i^{2\alpha}} = \sum_{i=1}^{n} \frac{y_i^2 d_i^{2\alpha}}{(\sigma_{e_i}^2 + \sigma_K^2 d_i^{2\alpha})^2} ,$$
$$\sum_{i=1}^{n} \frac{d_i^{2\alpha} \ln(d_i) y_i^2}{(\sigma_{e_i}^2 + \sigma_K^2 d_i^{2\alpha})^2} = \sum_{i=1}^{n} \frac{d_i^{2\alpha} \ln(d_i)}{\sigma_{e_i}^2 + \sigma_K^2 d_i^{2\alpha}} .$$

Dans la pratique, la répétition des étapes d'estimation de σ_K sachant α puis de α sachant σ_K a rapidement convergé dans tous les cas que nous avons testé (cf section 4). Cette estimation

nous à conduit à des résultats très proches de ceux obtenus en maximisant directement la log-vraisemblance de l'échantillon :

$$\ln V(\sigma_K, \alpha) = -\frac{n}{2} \ln 2\pi - \sum_{i=1}^{n} \frac{1}{2} \ln(\sigma_{e_i}^2 + \sigma_K^2 d_i^{2\alpha}) - \sum_{i=1}^{n} \frac{y_i^2}{2(\sigma_{e_i}^2 + \sigma_K^2 d_i^{2\alpha})} \,.$$

Au fur et à mesure des nouveaux tirages, chaque nouveau point θ_0 exploré conduit, en direction des $d+1$ sommets de la zone, à $d+1$ nouvelles réalisations de variables aléatoires de loi normale centrée, d'écart-type fonction de la distance, et il est alors possible de corriger à chaque étape une valeur a priori de α et de σ_K.

Il faut ici noter que rien n'interdit de faire varier les coefficients α et σ_K en fonction de la zone considérée, et l'estimation des coefficients pourrait se faire sur chaque zone Z en affectant chaque réalisation de la variable aléatoire $Y'_{\theta,\theta'}$ (sur d'autres zones) de poids d'autant plus élevés que la θ et θ' sont proches de la zone considérée Z. L'estimation de coefficients de Hölder à partir d'observations de tirages de F est un vaste sujet (cf. Blanke, 2002, pour un article traitant de ce type d'estimation sur des processus). Les choix numériques concrets des paramètres de l'algorithme seront détaillés dans la section d'application numérique 4.

Enfin, le raisonnement tenu jusqu'à présent se basait sur le fait que pour deux points d'une zone, l'observateur représentait la variation des pentes entre ces deux points par une variable aléatoire dont l'écart-type était une fonction croissante de la distance. Même si cela n'est pas requis pour l'implémentation de l'algorithme, un élément susceptible de fournir a posteriori des informations sur l'hypothèse prise est le suivant : pour une zone Z, on définit une variable aléatoire IR_α :

$$\mathrm{IR}_\alpha(Z) = \frac{f(\theta_1) - f(\theta_2)}{d(\theta_1, \theta_2)^\alpha} \,,$$

θ_1, θ_2 vecteurs aléatoires distincts issus de tirages indépendants, uniformes sur Z .

Précisons ici les modalités des tirages. D'une part, le tirage uniforme d'un vecteur θ sur une zone Z peut s'opérer facilement en choisissant des coordonnées barycentriques de façon uniforme dans un simplexe unité : ainsi, si ω_j est la $j^{\text{ème}}$ coordonnée barycentrique de θ dans la zone Z, $j \in \{1, \ldots, d+1\}$, on pourra prendre $\omega_j = e_j / e_{tot}$, avec $e_{tot} = \sum_{j=1}^{d+1} e_j$, pour un ensemble $\{e_j\}_{j \in \{1,\ldots,d+1\}}$ de variables aléatoires mutuellement indépendantes, de loi exponentielle de paramètre 1 (ce qui revient à prendre les intervalles successifs de statistiques d'ordre d'un vecteur uniforme). Le lecteur intéressé par ce type de tirages pourra consulter notamment Smith, Tromble (2004). D'autre part, on conviendra que (θ_1, θ_2) est le premier couple distinct issu de tirages uniformes (du fait de la précision arithmétique finie des ordinateurs, la probabilité que les deux points soient confondus peut être non nulle en pratique, bien qu'extrêmement faible). Selon le modèle présenté, à défaut d'autres informations, l'observateur suppose qu'il existe une quantité $\alpha \geqslant 0$ telle que la variation de f entre θ_1 et θ_2 ne s'explique que par la distance $d(\theta_1, \theta_2)^\alpha$, et pour lequel l'écart-type de $\mathrm{IR}_\alpha(Z)$ serait fini (ce qui n'est pas acquis pour tout α bien sûr). Sous l'hypothèse prise, cet écart-type fournira une information sur les variations de pentes auxquelles l'utilisateur pourra s'attendre. En particulier, si $\mathrm{IR}_\alpha(Z)$ est supposé distribué de façon gaussienne, alors son écart-type empirique correspondra à un estimateur de σ_K en connaissance de α. Des distributions de IR_α seront illustrées dans la section 4.

2.4 Choix du sommet de scission

Une fois une zone Z^+ choisie parmi un ensemble de zones $\mathfrak{Z}$, il va s'agir de scinder la zone en ajoutant un point à l'intérieur de celle-ci. De par nos choix précédents, ce point se situera

sur l'un des segments de l'enveloppe convexe de la zone.

Pour la zone Z^+, nous noterons $C = \{(\theta_{j_1}, \theta_{j_2})\}_{j_1,j_2}$ l'ensemble des couples de sommets distincts de Z^+. Le sommet de séparation retenu peut être choisi de nombreuses façons.

La solution que nous retiendrons consiste ici à sélectionner, parmi l'ensemble des couples possibles, l'un de ceux qui maximisent la longueur du segment. Si $\theta_1, \cdots, \theta_{d+1}$ désignent les sommets de la zone Z^+, nous choisirons comme sommet de séparation de la zone le sommet θ^+, isobarycentre du segment $(\theta_{j_1}, \theta_{j_2})$ tel que :

$$\begin{aligned} (\theta_{j_1}, \theta_{j_2}) &= \arg \max_{(\theta_i,\theta_j)\in C} d(\theta_i, \theta_j)\,, \\ C &= \{(\theta_i, \theta_j)\}_{i\in\{1,\ldots,d\}, j\in\{i+1,\ldots,d+1\}}\,. \end{aligned}$$

Dans le cas de plusieurs segments de longueur maximale, nous conviendrons que arg max désignera un segment choisi de façon uniforme parmi l'ensemble (fini) des segments de longueur maximale.

Choisir le point de segmentation sur un segment de longueur maximale a le mérite de limiter l'apparition de simplexes très déséquilibrés dans leurs longueurs de segment : en effet, la connaissance de l'appartenance de l'optimum à une zone est moins utile si cette zone est très allongée dans certaines directions. Le choix d'un sommet conduisant également à la partition des zones adjacentes, l'apparition de simplexes très déséquilibrés, si elle est ainsi limitée, n'est toutefois pas absolument exclue. Une idée de la forme des zones ainsi obtenues en dimension 2 est illustrée par la figure 1, qui a été construite par scission du segment de plus grande longueur de la zone choisie.

Lorsque des simulations sont opérées au point θ^+ du segment choisi de séparation, la scission des zones se fait ainsi :

- L'ensemble des zones $\mathfrak{Z}^+$ contenant ce segment est déterminé (celles qui contiennent les deux sommets du segment à la fois).
- Chacune des zones Z de l'ensemble $\mathfrak{Z}^+$ est scindée en deux zones Z_1 et Z_2 comme indiqué dans le paragraphe 2.1.

2.5 Schéma de l'algorithme de scission systématique

Les données en entrée de l'algorithme sont les paramètres de régularité α et σ_K, le nombre de simulations n_0 à opérer en chaque point (nombre de tirages requis pour l'estimation de la variance empirique $\sigma_e^2(\theta)$), le nombre d'étapes n de l'algorithme. On suppose également que la zone de recherche Z_0 est connue et que les premiers tirages ont été réalisés aux sommets de cette zone.

L'algorithme 1 récapitule le procédé général de scission des zones. L'algorithme permet de construire à chaque étape j l'ensemble des zones $\mathfrak{Z}_j$ formant une partition de la zone de recherche initiale Z_0. Par construction, les nouveaux sommets de scission sont toujours explorés, de sorte que l'on a exploré, en fin d'algorithme, l'ensemble des sommets de la partition finale $\mathfrak{Z}_n$. On rappelle ici que l'ensemble des sommets d'une zone Z est noté $S(Z)$.

L'algorithme est ici présenté de façon synthétique. Chacune des étapes de l'algorithme est détaillée dans les sections précédentes.

L'exploitation des nombreuses données disponibles en sortie de l'algorithme est détaillée dans la section 2.6 ci-après.

Algorithme 1 algorithme à scission systématique

Entrée: σ_K, α, n_0, n
Entrée: $\mathfrak{Z}_0 = \{Z_0\}$
pour $j = 0$ à $n-1$
choix d'une zone Z^+ de $\mathfrak{Z}_j$ (cf. §2.2)
calculer $\widehat{m}^*$ et σ_{m^*}
$\forall Z_i \in \mathfrak{Z}_j$, calculer $\boldsymbol{\beta}(Z_i)$
piocher une zone Z^+ en fonction des $\{\boldsymbol{\beta}(Z_i)\}_{Z_i \in \mathfrak{Z}_j}$
choix d'un sommet de scission θ^+ de Z^+ (cf. §2.4)
piocher un sommet de scission θ^+
scinder les zones contenant θ^+
calculer $\mathfrak{Z}_{j+1}$, ensemble des nouvelles zones
tirages au sommet $\theta^+ \in Z^+$
calculer n_0 tirages de $F(\theta^+)$
mise à jour facultative du couple (σ_K, α), éventuellement par zone (cf. §2.3)
fin pour
Sortie: $\widehat{m}^*$, σ_{m^*}
Sortie: $\forall Z \in \mathfrak{Z}_n$, $V(Z)$, $\boldsymbol{\beta}(Z)$
Sortie: $\forall Z \in \mathfrak{Z}_n, \forall \theta \in S(Z)$, $\widehat{f}(\theta)$, $\sigma_e(\theta)$, $\beta_Z(\theta)$

2.6 Résultat final et critère de convergence

Zones de confiance Nous supposons ici que l'algorithme de scission systématique est terminé, par atteinte du critère d'arrêt proposé (nombre suffisant de tirages ici) : plus aucun tirage de F ne sera donc opéré.

Un estimateur de la valeur m^* de l'unique minimum global de f est fourni par l'algorithme. Pour autant, nous recherchons essentiellement à agir sur les paramètres, c'est-à-dire à obtenir l'ensemble des paramètres susceptibles de conduire à ce minimum, ce qui nous donnera également une indication sur la fiabilité du résultat obtenu et sur les investigations futures à opérer, par exemple pour départager deux candidats potentiels. Nous cherchons donc un ensemble discret (car il doit être traité numériquement) s'approchant (selon une mesure qui sera définie ultérieurement) de :

$$\mathcal{S}_x = \{\theta \in \Theta, \mathrm{E}\left[F(\theta)\right] \leqslant x\}$$

pour tout x appartenant à un voisinage de m^*.

A l'issue de l'algorithme, le domaine de recherche initial Z_0 est scindé en un ensemble de zones $\mathfrak{Z}$. La définition des zones de confiance sera facilitée si nous définissons le potentiel d'un point y compris sur les frontières qui peuvent appartenir à plusieurs zones à la fois (sur les faces des simplexes). Les points appartenant à plus d'une zone formeront un ensemble de volume nul. Toutefois, afin de lever toute ambiguïté pour ces points particuliers, nous définirons pour chaque point une unique valeur de potentiel :

$$\beta_{\mathfrak{Z}}(\theta) = \max_{Z \in \mathfrak{Z}, \theta \in Z} \beta_Z(\theta)\,.$$

L'ensemble des paramètres admissibles, que nous nommerons également *zone de confiance*, sera défini pour un seuil $s \in [0,1]$ comme l'ensemble des points candidats θ pour lesquels $f(\theta)$ est potentiellement inférieur à m^* :

$$\widehat{\mathcal{S}}_{m^*,s} = \{\theta \in \Theta_0, \beta_{\mathfrak{Z}}(\theta) \geqslant s\}$$

L'ensemble Θ_0 des points candidats pourra être l'ensemble des points pour lesquels des tirages ont été opérés, ou bien, lorsque cela facilite l'usage futur de $\widehat{\mathcal{S}}_{m^*,s}$, un ensemble de points régulièrement espacés, pour un pas donné strictement positif δ :

$$\Theta_0 = \left\{ \theta \in \Theta, \frac{1}{\delta}\theta \in \mathbb{N}^d \right\}$$

Lorsque s est égal à 0, tous les points de Θ_0 sont retenus. Lorsque s augmente, l'ensemble $\widehat{\mathcal{S}}_{m^*,s}$ se restreint à l'ensemble des paramètres pour lesquels $\widehat{f}(\theta)$ est très proche de m^*.

En résumé, on obtient finalement :

– la valeur m^* de l'optimum de f,

– l'ensemble des paramètres $\widehat{\mathcal{S}}_{m^*,s}$ susceptibles de conduire à cet optimum.

Critère de convergence en connaissance du résultat Nous allons dans un premier temps chercher un critère de convergence de l'algorithme lorsque l'on connaît l'ensemble solution $\mathcal{S}_{m^*}$. Un tel critère facilitera la compréhension de l'algorithme sur des fonctions de test, et constituera un outil de comparaison de différents algorithmes.

Nous cherchons à produire un ensemble $\widehat{\mathcal{S}}_{m^*}$ qui donne une représentation fidèle de $\mathcal{S}_{m^*}$. Par fidèle, on imagine d'une part que tout point du veritable ensemble solution $\mathcal{S}_{m^*}$ doit être proche d'un point de l'ensemble solution proposé $\widehat{\mathcal{S}}_{m^*}$: $\widehat{\mathcal{S}}_{m^*}$ doit être suffisamment grand (condition n^o 1). D'autre part, l'ensemble proposé ne doit pas non plus contenir de points trop éloignés des véritables solutions, et tout point de l'ensemble proposé $\widehat{\mathcal{S}}_{m^*}$ doit être proche d'un point de $\mathcal{S}_m$: $\widehat{\mathcal{S}}_{m^*}$ ne doit pas être trop grand (condition n^o 2). Cela nous conduira à proposer, comme distance entre les deux ensembles $\mathcal{S}_{m^*}$ et $\widehat{\mathcal{S}}_{m^*}$ la distance de Hausdorff, pour $X \subset \Theta, Y \subset \Theta$:

$$d^H(X,Y) = \max \left\{ \sup_{y\in Y} \inf_{x\in X} d(x,y), \sup_{x\in X} \inf_{y\in Y} d(x,y) \right\} .$$

Si $X = \widehat{\mathcal{S}}_{m^*,s}$ est l'ensemble proposé et $Y = \mathcal{S}_{m^*}$ est l'ensemble cible, alors majorer le premier terme du max indique que tout point de la cible Y est proche d'un point de X (condition n^o 1). Majorer le second terme du max indique que tout point de X est proche d'un point de Y (condition n^o 2). Finalement, un point que l'on propose comme solution ne doit pas être trop éloigné d'une solution réelle, et une solution réelle ne doit pas être trop éloignée d'un point proposé comme solution. La distance de Hausdorff est donc parfaitement adaptée à l'objectif recherché. Cette distance représente, pour le pire point de l'un des ensembles, la distance de ce point à l'autre des deux ensembles : elle fournit directement une idée de l'incertitude sur l'ensemble des paramètres conduisant à l'optimum. Nous obtenons donc un critère de convergence en connaissance du résultat recherché $\mathcal{S}_{m^*}$:

$$\rho_1(s) = d^H(\widehat{\mathcal{S}}_{m^*,s}, \mathcal{S}_{m^*}) .$$

Le critère précédant dépendant de la mesure choisie pour le potentiel, nous proposons également l'usage d'une distance de Hausdorff partielle :

$$\rho_2 = \sup_{y\in\mathcal{S}_{m^*}} \inf_{x\in\Theta_e} d(x,y) ,$$

Où Θ_e représente l'ensemble des points explorés. Ce critère fournit la pire distance d'un point solution au plus proche point exploré. Il indique donc si tous les points solutions ont bien été explorés, et serait naturellement très bon si Θ_e recouvrait Θ. Ce critère n'a de sens que dans

la mesure où le nombre de tirages total de F est limité : il pourra notamment servir à comparer différents algorithmes pour un même budget de tirages. Son avantage est de ne requérir que l'ensemble des points de tirage successifs de F. Une limite de ce critère est qu'un algorithme peut converger très vite vers un optimum global sans avoir exploré les zones candidates, et donc en ayant pris un risque important : trouver rapidement le vrai optimum global d'une fonction n'indique pas si la méthode est prudente ou non, et seule la considération de la variabilité de la fonction entre les points de tirage permet de trancher cette question (en un sens l'usage d'un potentiel). Une autre limite de ce critère est qu'il ne tient pas compte de la précision de l'estimation de f au points explorés. Nous avons néanmoins tenu à le présenter dans la mesure où il s'agit d'un des critères permettant de comparer l'algorithme proposé à d'autres algorithmes ne permettant pas le calcul de potentiels.

Il faut noter que du fait du caractère aléatoire de la fonction observée et de l'algorithme proposé, ces critères devraient varier selon les exécutions de l'algorithme. En toute rigueur, pour un critère ρ, la comparaison de plusieurs algorithmes sur une fonction test, pour un même budget de tirages autorisé, devrait requérir l'obtention des distributions de ρ pour chaque algorithme, puis l'usage d'un indicateur de risque, comme un quantile de ρ. Une piste pour le choix opérationnel des paramètres de l'algorithme est l'étude approfondie du comportement du critère choisi en fonction de ces paramètres.

Critère de convergence hors connaissance du résultat Dans la pratique, on ne connaît pas l'ensemble cible $\mathcal{S}_{m^*}$, puisque l'on cherche précisément à en cerner les contours.

Si l'on cherche à assurer que la fonction est suffisamment bien connue sur chaque zone, un indicateur de convergence est simplement la quantité :

$$\rho_3 = \max_{Z \in \mathfrak{Z}} \boldsymbol{\beta}(Z)\,,$$

avec le choix ici opéré $\boldsymbol{\beta}(Z) = V(Z) \cdot \beta_Z(\theta_{B_Z})$, θ_{B_Z} isobarycentre de la zone Z (cf. section 2). Majorer cet indicateur garantira en effet que pour toute zone $Z \in \mathfrak{Z}$:

- Soit que la présence d'un minimum global au point central est très peu probable, $\beta_Z(\theta_{B_Z})$ étant suffisamment faible.
- Soit que la zone est suffisamment petite et que le minimum potentiel a bien été exploré, $V(Z)$ étant suffisamment faible.

Enfin, si l'on cherche à assurer que chaque point ne conduira pas à un minimum plus petit que celui connu, pour un écart η donné, un critère pourra être :

$$\rho_4 = \max_{\theta \in \Theta_0} \beta_{\mathfrak{Z}}^{(m^*-\eta)}(\theta)\,,$$

Nous notons ici $\beta_{\mathfrak{Z}}^{(m^*)}(\theta)$ le potentiel au point θ obtenu pour un minimum global estimé m^* (l'indice supérieur $^{(m^*)}$ était resté implicite jusqu'à présent pour ne pas alourdir les notations).

D'autres critères peuvent naturellement être envisagés. Un élément incontournable dans la définition d'un critère est la modélisation de $f(\theta)$ ou de l'incertitude de $\widehat{f}(\theta)$ en dehors des points θ d'observation : si f *est* ou *est supposée* extrêmement erratique, toute zone sera susceptible de contenir un point conduisant à un optimum global, et le critère devra être très différent de celui obtenu en supposant f très régulière. C'est un avantage des critères proposés ici, à l'exception de ρ_2, mais c'est également une limitation puisque ces critères ne s'appliqueront qu'aux algortihmes proposant une partition de l'espace en zones affectées d'un potentiel.

3 Une grille à pas variable avec retirage possible

Jusqu'à présent, chaque zone était systématiquement subdivisée en deux. Or, il peut être plus avantageux, plutôt que de subdiviser une zone, d'ajouter des simulations en des points déjà explorés. Ce sera l'objet de cette section.

3.1 Critères de scission

Supposons que l'on estime une espérance et un intervalle de confiance de la fonction $f(\theta)$, aux points explorés θ_1 et θ_2. Supposons que θ_c soit le barycentre équipondéré (isobarycentre) entre ces deux points.

La question que l'on se pose est la suivante : au vu des intervalles de confiance de $f(\theta)$, connus aux points θ_1 et θ_2, et connaissant la variabilité de l'évolution de $f(\theta)$ sur ce segment, mesurée par les paramètres (α, σ_K), nous cherchons un critère permettant de déterminer la décision à prendre entre les suivantes :

– diminuer l'incertitude sur $\widehat{f}(\theta_i)$ par de nouveaux tirages de F en θ_i, $i \in \{1, 2\}$.

– scinder le segment en deux parties, en opérant des tirages de F au nouveau point θ_c.

Plusieurs critères peuvent être retenus pour opérer ce choix. Le choix du critère retenu au final dépend essentiellement de l'objectif poursuivi : s'assurer que le minimum ne peut être présent dans une zone sous l'hypothèse que la fonction objectif ne varie pas trop brusquement, minimiser le maximum des potentiels d'une zone, etc. Ce choix pouvant dépendre du problème considéré, nous détaillerons ci-dessous deux critères dont nous étudierons la performance dans la section 4.

Comparaison des erreurs de grille et d'estimation Considérons une zone Z. En un point θ de Z, le potentiel est la mesure retenue pour quantifier la probabilité que $\tilde{f}(\theta)$ soit plus petite que l'optimum global $\widehat{m}^*$ estimé. Cette mesure dépend à la fois de l'erreur d'estimation aux sommets explorés adjacents, et à la fois de l'erreur de grille, liée à l'éloignement entre θ et les sommets adjacents (et dépendant de la régularité supposée de $\tilde{f}$, mesurée par σ_K et α). Par souci de simplicité, l'incertitude liée à l'estimation du minimum global m^* ne sera pas ici scindée en erreur d'estimation et erreur de grille.

Considérons l'*erreur d'échantillonnage*. Cette erreur est liée à la mauvaise connaissance de la fonction f aux point explorés, du fait du nombre réduit de tirages et du bruit frappant f. Elle dépend essentiellement de la probabilité de présence du minimum aux sommets θ_i de Z du fait de trop larges intervalles de confiance aux points explorés. En l'absence de doute lié à l'évolution de la fonction entre ces points, cette probabilité peut être mesurée par le potentiel $\beta(\theta)$ calculé en l'absence d'erreur de grille :

$$\beta_e(\theta) = \beta(\theta)|_{\sigma_K=0}\,.$$

Considérons l'*erreur de grille*. Cette erreur dépend essentiellement de la possibilité que la fonction évolue brusquement entre un sommet de Z et le point θ considéré. En l'absence de doute lié à la mauvaise connaissance de la fonction aux extrémités du segment, compte tenu de la seule mauvaise exploration de la fonction, cette erreur peut être estimée au point θ par le potentiel en θ, calculé en l'absence d'erreur d'estimation aux $d+1$ sommets θ_i de la zone Z :

$$\beta_g(\theta) = \beta(\theta)|_{\sigma_{e_i}=0,\, i=1,\ldots,d+1}\,.$$

Envisageons de scinder un segment autour d'un point barycentre non encore exploré θ_c. Si le potentiel en ce point est grand essentiellement à cause de l'erreur de grille, il apparaîtra

raisonnable de scinder le segment et d'explorer θ_c. Si le potentiel est grand essentiellement à cause de l'erreur d'estimation, il paraîtra plus raisonnable au contraire de réduire cette dernière en explorant davantage les sommets du segments. Le critère de scission que nous retiendrons sera donc le suivant :

$$\text{Critère n}^{\text{o}}\ 1 : \text{Scission si } \beta_g(\theta_c) \geqslant \beta_e(\theta_c)\,.$$

En dimension $d = 1$, si θ_c est l'isobarycentre de θ_1 et θ_2, alors on peut noter $d_0 = d(\theta_c, \theta_1) = d(\theta_c, \theta_2)$. Si l'on suppose $\sigma_e = \sigma_e(\theta_1) = \sigma_e(\theta_2)$, compte tenu de la définition retenue ici pour le potentiel d'une zone, on montre facilement que l'on a l'équivalence suivante :

$$(\beta_g(\theta_c) \geqslant \beta_e(\theta_c)) \;\Leftrightarrow\; (\sigma_e \leqslant \sigma_K d_0^{\alpha})\ .$$

Ce dernier critère s'interprète alors très aisément : la scission est opérée si l'erreur de grille ($\sigma_g = \sigma_K d_0^{\alpha}$) est supérieure à l'erreur d'estimation (σ_e). En dimension $d > 1$, le critère $\beta_g(\theta) \geqslant \beta_e(\theta)$ permet de tenir compte en un point θ des différences éventuelles de distances aux sommets adjacents, et des différences d'erreurs d'estimation aux différents sommets de la zone Z : Ce critère revient à opérer une moyenne particulière entre les erreurs d'estimation σ_e et les erreurs de grille σ_g en direction des différents sommets.

En toute logique, si F est aléatoire et si $\sigma_K = 0$, alors la supposition de la connaissance parfaite de f entre les points de tirage conduit à seulement mieux estimer $\widehat{f}$ aux sommets de la zone Z. Par ailleurs, en l'absence d'erreur d'estimation (par exemple si F est déterministe), le critère conduit bien à une scission systématique, il devient naturellement inutile d'effectuer un retirage.

Meilleur potentiel après ajout de n_0 simulations Nous envisageons ici le cas où la répartition se fait en fonction de l'amélioration du potentiel maximal de la zone, scindée ou non. On suppose que n_0 simulations seront ajoutés sur un sommet de la zone choisie Z^+. Les erreurs d'estimation correspondent à des écarts-type de variables aléatoires supposées de loi normales. Il est donc simple d'estimer quelle sera la réduction de la variance empirique en cas d'ajout de n_0 simulations, avant même la réalisation de ces simulations. Ainsi, en un point déjà exploré θ, si n_θ tirages de F ont déjà eu lieu et si $\sigma_F(\theta)$ représente la variance empirique de F en ce point, on pourra proposer $\sigma_e(\theta) = \sigma_F(\theta)/\sqrt{n_\theta + n_0}$.

- Si l'on envisage d'ajouter n_0 simulations en un sommet existant θ de la zone choisie, il est donc aisé d'estimer la nouvelle erreur d'échantillonnage $\sigma_e(\theta)$ en cas d'ajout.
- Si l'on envisage d'ajouter n_0 points sur le barycentre θ_c du segment choisi, on peut alors estimer $\sigma_F(\theta_c)$ par interpolation linéaire, avant l'exploration du point θ_c, et en déduire une estimation de l'erreur d'échantillonnage $\sigma_e(\theta_c) = \sigma_F(\theta)/\sqrt{n_0}$ en cas d'exploration.

Quel que soit le sommet choisi pour l'ajout parmi les $d+1$ sommets existants, on peut donc donner un nouvel estimateur $\widehat{\boldsymbol{\beta}}(Z^+)$ du potentiel de la zone choisie après ajout de n_0 simulations. Si un sommet existant doit être privilégié pour l'ajout, il semble logique de convenir d'un ajout sur celui de ces $d+1$ points qui conduit à minimiser cette valeur estimée, ce minimum étant noté $\widehat{\boldsymbol{\beta}}_{\min}$:

$$\widehat{\boldsymbol{\beta}}_{\min} = \min_{\theta_i \in S(Z^+)} \left\{ \widehat{\boldsymbol{\beta}}(Z^+)|_{n_0 \text{ ajouts en } \theta_i} \right\}\ .$$

Si au contraire une scission est envisagée, deux zones Z_1 et Z_2 seront formées. La scission pourra être privilégiée, par exemple, si les potentiels estimés sur ces deux zones sont tous deux inférieurs à $\boldsymbol{\beta}_{\min}$:

$$\text{Critère n}^{\text{o}}\ 2 : \text{Scission si } \max\left(\widehat{\boldsymbol{\beta}}(Z_1), \widehat{\boldsymbol{\beta}}(Z_2)\right) < \widehat{\boldsymbol{\beta}}_{\min}\,,$$

l'esprit étant alors qu'une scission ne doit pas conduire à augmenter le maximum des $\boldsymbol{\beta}(Z_i)$ sur l'ensemble des zones $\{Z_i\}$ à l'étape considérée : l'ajout se porte ainsi sur le sommet dont on estime qu'il minimise le maximum des potentiels sur l'ensemble des zones.

Cette répartition est également pleinement cohérente avec la définition choisie des coefficients $\boldsymbol{\beta}$, et elle conduit, par construction, à une diminution du potentiel maximum estimé sur l'ensemble des zones, ce qui est de nature à faciliter les futures démonstrations de la convergence de l'algorithme, notamment si le critère de convergence ρ_3 est utilisé.

3.2 Schéma de l'algorithme à retirage possible

Algorithme 2 algorithme à retirage possible

Entrée: σ_K, α, n_0, n
Entrée: $\mathfrak{Z}_0 = \{Z_0\}$
pour $j = 0$ à $n-1$
choix d'une zone Z^+ de $\mathfrak{Z}_j$ (cf. §2.2)
calculer $\widehat{m}^*$ et σ_{m^*}
$\forall Z_i \in \mathfrak{Z}_j$, calculer $\boldsymbol{\beta}(Z_i)$
piocher une zone Z^+ en fonction des $\{\boldsymbol{\beta}(Z_i)\}_{Z_i \in \mathfrak{Z}_j}$
si critère scission (Z^+) vrai (cf. §3.1) **alors**
choix d'un sommet de scission θ^+ de Z^+ (cf. §2.4)
piocher un sommet de scission θ^+
scinder les zones contenant θ^+
calculer $\mathfrak{Z}_{j+1}$, ensemble des nouvelles zones
sinon
choix d'un sommet de retirage θ^+ de Z^+ (cf. §3.1)
piocher un sommet déjà exploré $\theta^+ \in S(Z^+)$
$\mathfrak{Z}_{j+1} = \mathfrak{Z}_j$, les zones sont inchangées
fin si
tirages au sommet $\theta^+ \in Z^+$
calculer n_0 tirages de $F(\theta^+)$
mise à jour facultative du couple (σ_K, α), éventuellement par zone (cf. §2.3)
fin pour
Sortie: $\widehat{m}^*$, σ_{m^*}
Sortie: $\forall Z \in \mathfrak{Z}_n$, $V(Z)$, $\boldsymbol{\beta}(Z)$
Sortie: $\forall Z \in \mathfrak{Z}_n, \forall \theta \in S(Z)$, $\widehat{f}(\theta)$, $\sigma_e(\theta)$, $\beta_Z(\theta)$

L'algorithme 2 récapitule le procédé général de scission des zones. L'algorithme permet de construire à chaque étape j l'ensemble des zones $\mathfrak{Z}_j$ formant une partition de la zone de recherche initiale Z_0. Cet algorithme diffère de l'algorithme 1 par sa faculté d'opérer un nouveau tirage en un sommet déjà exploré de la zone à explorer Z^+, plutôt que de systématiquement scinder Z^+.

4 Applications numériques

4.1 Fonction test utilisée

Nous présenterons ici une application partant d'une fonction test connue, afin notamment de vérifier le bon comportement de l'algorithme. Par souci de lisibilité, cette application sera tout d'abord présentée en dimension $d = 2$ (mais l'algorithme présenté dans ce papier est proposé pour toute dimension $d \in \mathbb{N}^*$, et le problème de la montée en dimension sera évoqué par la suite).

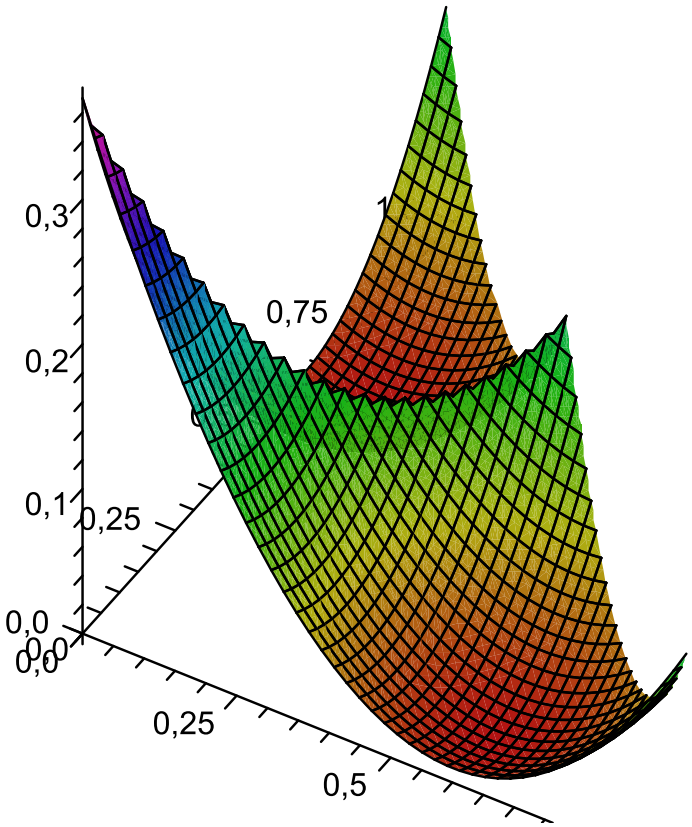

FIG. 2 – Allure générale de la fonction test $f(\theta)$ pour $\theta \in Z_0$

Nous considérerons ici la fonction suivante dans le cas $d = 2$ (correspondant par exemple à 3 poids d'allocation d'actifs se sommant à 1).

$$\begin{aligned} f(\theta) &= (\min(x,y) - 0.1)^2 + (\max(x,y) - 0.6)^2 \ , \\ F(\theta) &= f(\theta) + \sigma_B(U - 0.5) \, , \\ \theta &= (x,y) \, , \end{aligned}$$

avec U une variable aléatoire uniforme sur $[0,1]$. L'espérance f de la fonction F admet deux minima, l'un en $\theta_1^* = (0.1, 0.6)$, l'autre en $\theta_2^* = (0.6, 0.1)$, la valeur de f étant alors 0. A titre indicatif, afin d'imaginer les variations possibles de cette fonction, la fonction f atteint son maximum sur Z_0 en $\theta = (0,0)$ et l'on a alors $f(\theta) = 0.37$ (mais l'on cherche ici le minimum, non le maximum). Un bruit pouvant conduire à des variations d'amplitude de 0.1 entre deux

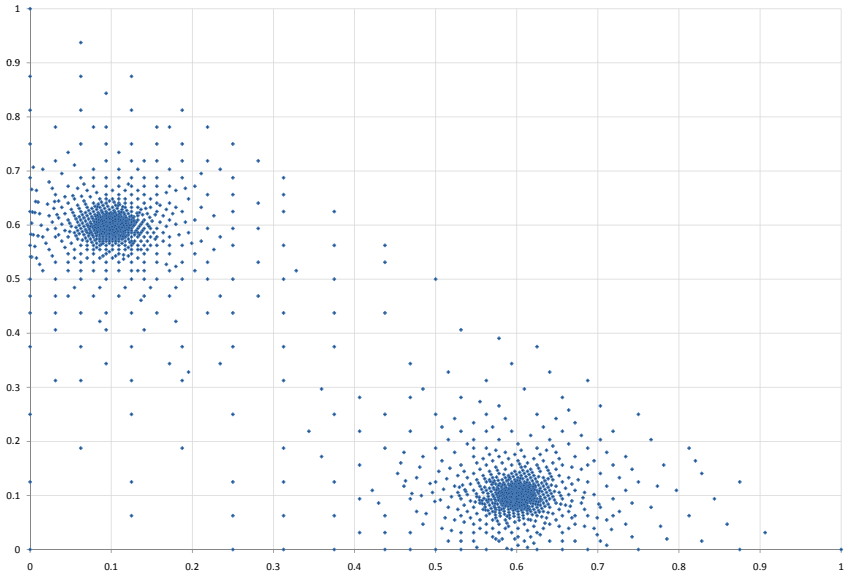

FIG. 3 – Ensemble des points de simulations de F pour une variabilité $\sigma_K = 2$ et un niveau de bruit $\sigma_B = 0$ (scission systématique)

points proches est donc assez élevé au regard du domaine de variation $[0, 0.37]$ de la fonction sur la zone initiale de recherche. Un aperçu de la fonction f est donné dans la figure 2.

Dans le cadre de l'algorithme à scission systématique, nous utiliserons ici toujours les paramètres suivants :

- le nombre n de tirages réalisés, sera $n = 2000$,
- le nombre de tirages en chaque point sera $n_0 = 10$,
- le coefficient de priorité sera $\gamma = 1/(d+1) = 1/3$, conduisant à une moyenne géométrique des potentiels directionnels pour le potentiel d'un point,
- le seuil η du critère de convergence ρ_4 sera $\eta = 0.01$.

Par ailleurs, pour chaque illustration, nous préciserons

- le niveau du bruit σ_B frappant la fonction f,
- les coefficients de variabilité σ_K et α,
- le seuil s du critère de convergence ρ_1.

A titre de remarque, la figure 1 présentée précédemment a été obtenue avec la fonction F évoquée ci-dessus, à partir des paramètres $(\sigma_K, \alpha) = (1, 1.5)$ et $\sigma_B = 0.1$ (et $\gamma = 1$ dans ce seul cas).

4.2 Comportement de l'algorithme à scission systématique

Dans cette section, nous allons observer le comportement de l'algorithme n° 1 à scission systématique, lorsque le niveau de variabilité σ_K varie, pour différents niveaux de bruit σ_B. Dans ce paragraphe, le paramètre α est ici fixé, égal à 1.5, et correspond à un modèle où les pentes sont supposées suivre un mouvement brownien depuis un point d'accroche (cf. définition du potentiel directionnel dans les sections 2.2 et 2.3).

A titre indicatif, nous mentionnerons dans la légende de certains graphiques la valeur obtenue pour le vecteur critère de convergence, $\vec{\rho} = (\rho_1(s), \rho_2, \rho_3, \rho_4(\eta))$, avec s égal à 10% du

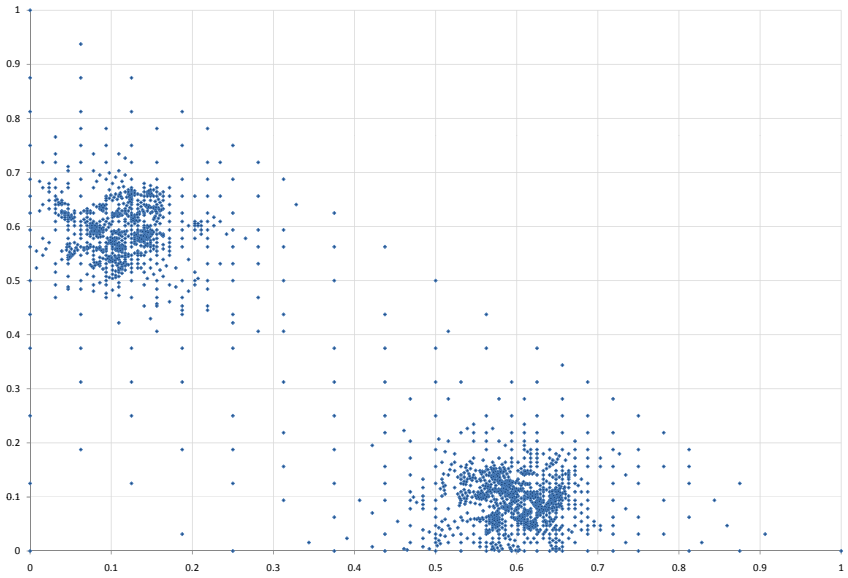

FIG. 4 – Ensemble des points de simulations de F pour une variabilité $\sigma_K = 2$ et un niveau de bruit $\sigma_B = 0.1$ (scission systématique), $\vec{\rho} =$ (0.10, 3.5E-03, 6.8E-07, 1.9E-2)

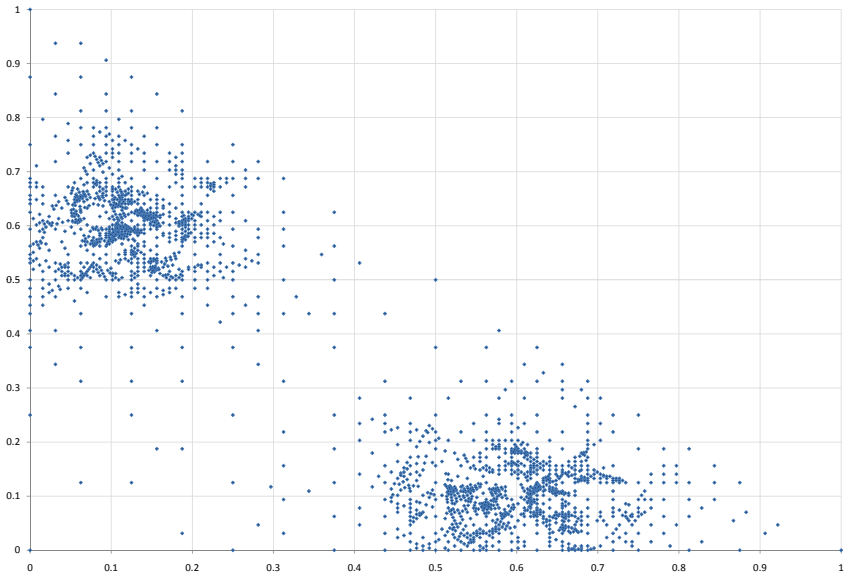

FIG. 5 – Ensemble des points de simulations de F pour une variabilité $\sigma_K = 2$ et un niveau de bruit $\sigma_B = 0.3$ (scission systématique), $\vec{\rho} =$ (0.13, 5.9E-03, 1.6E-06, 5.8E-2)

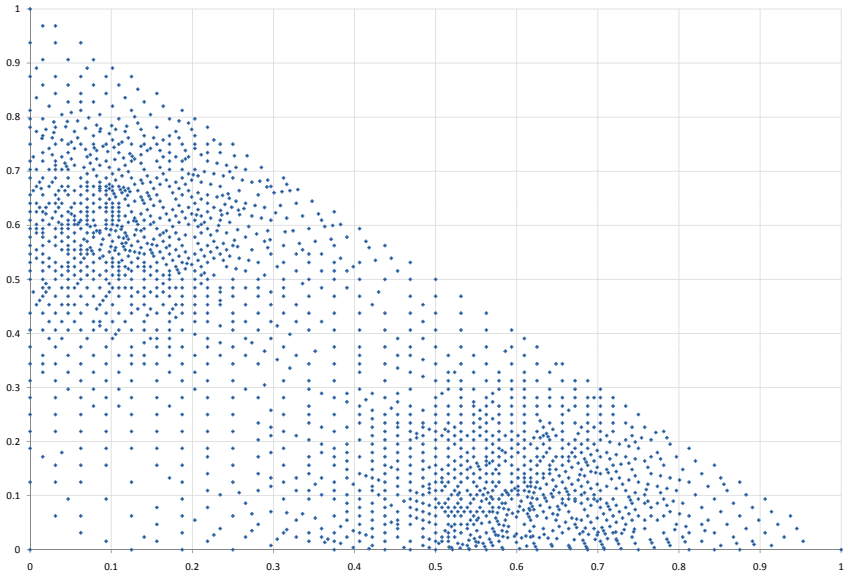

FIG. 6 – Ensemble des points de simulations de F pour une variabilité $\sigma_K = 20$ et un niveau de bruit $\sigma_B = 0.1$ (scission systématique) , $\vec{\rho} =$ (0.45, 2.2E-03, 7.3E-05, 0.72)

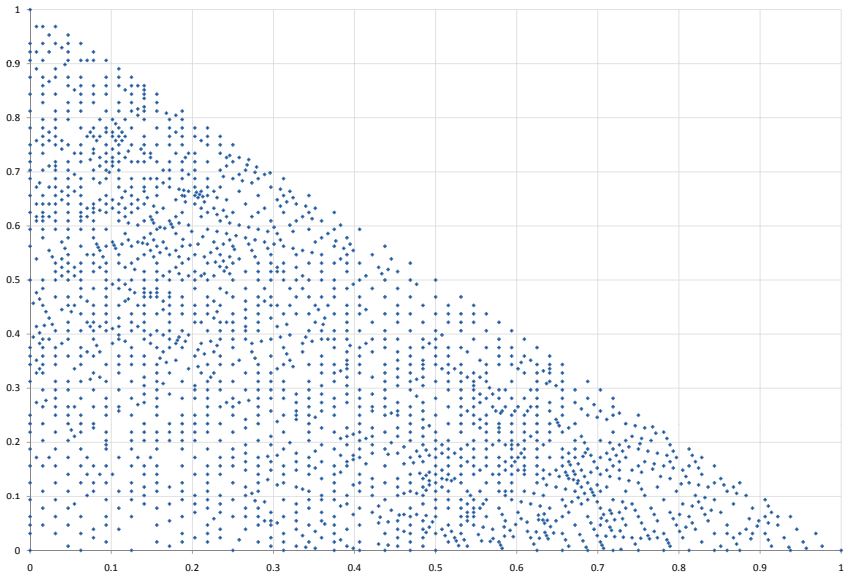

FIG. 7 – Ensemble des points de simulations de F pour une variabilité $\sigma_K = 100$ et un niveau de bruit $\sigma_B = 0.1$ (scission systématique), $\vec{\rho} =$ (0.60, 8.8E-03, 5.9E-04, 0.97). La recherche de l'optimum d'une fonction constante bruitée conduit à un nuage de points d'allure proche de celui-ci.

potentiel maximal observé, et $\eta = 0.01$. Le détail du calcul de ces critères ainsi qu'une analyse plus détaillée des valeurs numériques obtenues pour ceux-ci seront abordés dans la section 4.5.

Pour un niveau de variabilité supposée de la fonction $\sigma_K = 2$, illustré dans les figures 3, 4 et 5, lorsque le niveau de bruit passe respectivement par $\sigma_B = 0$, $\sigma_B = 0.1$ et $\sigma_B = 0.3$, on observe bien l'exploration préférentielle des zones à proximité des minimiseurs de f. Bien entendu, cette exploration est plus étendue lorsque le niveau de bruit augmente, puisque du fait de ce bruit, l'assurance de l'absence d'autres optima locaux nécessite l'exploration d'une zone plus vaste. Il faut noter que dans la pratique, ce niveau de bruit de F est généralement une donnée exogène sur laquelle il n'est pas possible d'agir.

Pour un niveau de bruit fixé à $\sigma_B = 0.1$, illustré dans les figures 4, 6 et 7, lorsque la variabilité supposée de la fonction passe successivement par $\sigma_K = 2$, $\sigma_K = 20$ et $\sigma_K = 100$, on observe un résultat parfaitement logique : si le comportement attendu de la fonction entre les points de tirage est supposé peu variable, l'algorithme se concentre sur les zones proches des minimiseurs supposés. Dans le cas inverse, une variabilité extrême $\sigma_K = 100$ conduit à une répartition quasi uniforme des points d'exploration : avec une telle variabilité, l'optimum global de la fonction peut en effet se trouver dans n'importe quelle zone. Il faut toutefois remarquer que la fixation d'un seuil de variabilité σ_K trop faible conduirait par définition à supposer que le comportement de la fonction est globalement connu entre les points explorés, ce qui conduit à négliger des zones pourtant susceptibles de contenir un optimum. L'arbitrage entre la rapidité de convergence vers les optima globaux et le risque d'avoir négligé une zone dépendra donc largement de ce paramètre σ_K.

Pour α fixé, la fixation du paramètre σ_K est donc un problème délicat, que nous aborderons dans la section 4.4. Elle pourra se baser en particulier sur les explorations précédentes de la fonction, sur d'autres connaissances de celles-ci, ainsi que sur les contraintes d'implémentation et de niveau de risque accepté. Il faut noter à ce sujet que rien dans le modèle n'oblige à fixer cette valeur constante sur l'ensemble de la zone de recherche, et qu'il est également possible de tenir compte d'une variabilité particulière de la fonction sur certaines zones.

En résumé, pour l'algorithme n° 1 à scission systématique, l'algorithme vise essentiellement à répartir des points d'exploration, en délaissant temporairement certaines zones non susceptibles de contenir le minimum. Selon les paramètres choisis et la fonction optimisée, le comportement attendu est le suivant, conforme à ce que nous avons pu observer :

- Lorsque le bruit est important devant les variations de la fonction, ou lorsque l'incertitude sur la variabilité de la fonction σ_K est élevée, l'algorithme explore assez uniformément la fonction, de façon similaire aux traditionnelles grilles de recherche à pas fixe évoquées précédemment.
- Lorsque le bruit n'est pas trop élevé et lorsque la variabilité de la fonction σ_K est faible, les points de tirage se concentrent sur les zones contenant les points solutions supposés, le comportement de la fonction étant supposé sans surprise entre les points déjà explorés.

Le choix du coefficient σ_K de variabilité de la fonction f dépend donc du but poursuivi : trop élevé, l'exploration de la fonction sera très poussée, et le temps de calcul pourra être élevé si l'on vise une bonne connaissance d'un optimum de la fonction. Pour un σ_K trop faible, l'algorithme se focalisera très vite sur un optimum local, donc permettra un gain en terme de temps de calcul, à précision égale, mais le risque de ne pas explorer une zone susceptible de contenir un autre optimum, éventuellement meilleur, sera plus élevé.

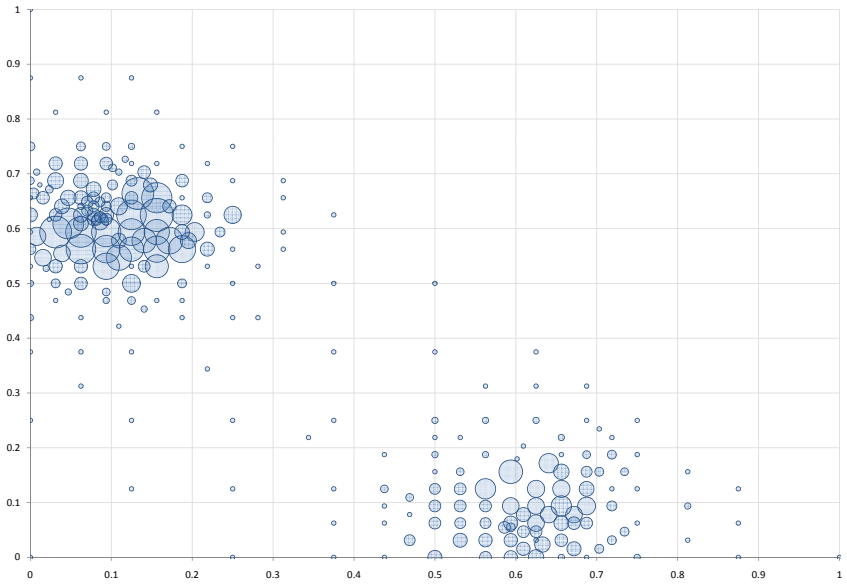

FIG. 8 – Ensemble des points de simulation de F pour un niveau de bruit $\sigma_B = 0.1$, critère de scission n° 2, pour une variabilité $(\sigma_K, \alpha) = (1, 1.5)$, et un nombre minimal de tirage $n_0 = 10$. La surface des bulles est proportionnelle au nombre de tirages de F en chaque point. $\vec{\rho} =$ (0.085, 8.8E-03, 3.1E-05, 8.2E-04)

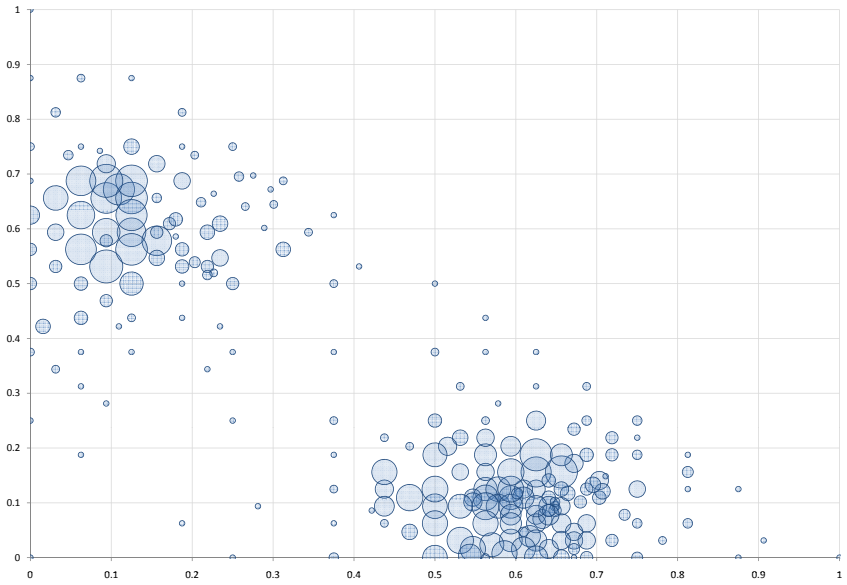

FIG. 9 – Ensemble des points de simulation de F pour un niveau de bruit $\sigma_B = 0.2$, critère de scission n° 2, pour une variabilité $(\sigma_K, \alpha) = (1, 1.5)$, et un nombre minimal de tirage $n_0 = 10$. La surface des bulles est proportionnelle au nombre de tirages de F en chaque point. $\vec{\rho} =$ (0.10, 8.8E-03, 6.7E-05, 6.2E-03)

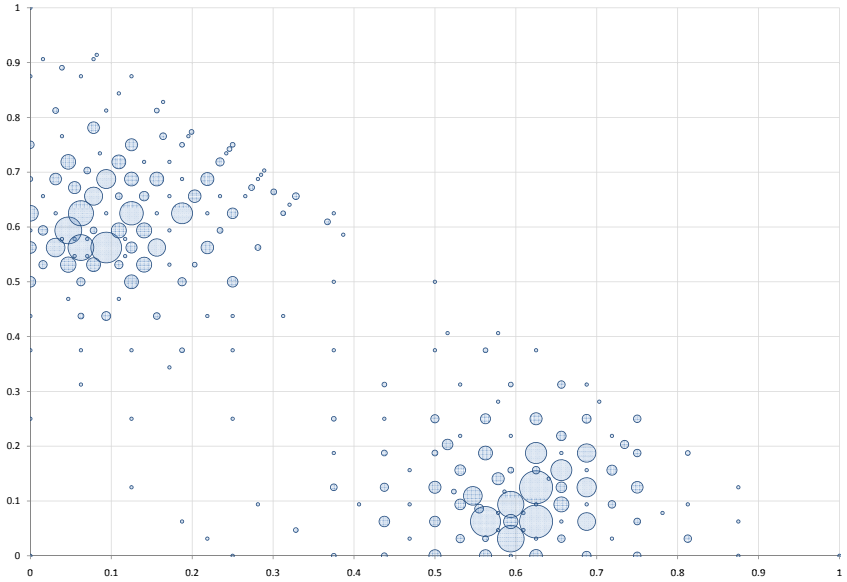

FIG. 10 – Ensemble des points de simulation de F pour un niveau de bruit $\sigma_B = 0.2$, critère de scission n° 1, pour une variabilité $(\sigma_K, \alpha) = (1, 1.5)$, et un nombre minimal de tirage $n_0 = 10$. La surface des bulles est proportionnelle au nombre de tirages de F en chaque point. $\vec{\rho} =$ (0.10, 0.011, 1.5E-05, 1.6E-02)

4.3 Comportement de l'algorithme avec retirages possibles

Les deux figures 8 et 9 illustrent les retirages qui peuvent s'opérer sur des sommets déjà explorés, afin d'améliorer la connaissance de la fonction f en ces points. Pour ces figures, nous avons utilisé le critère de scission n° 2 envisageant l'amélioration du potentiel après ajout de n_0 simulations (cf. §3.1). Lorsque le niveau de bruit est $\sigma_B = 0.1$, la figure 8 montre que les retirages ont essentiellement lieu autour des points solutions supposés, lorsqu'il n'est plus seulement nécessaire d'avoir une idée de la zone où se situe un minimiseur, mais qu'il faut également estimer avec précision la valeur m^* de l'optimum atteint. Lorsque le bruit augmente, cela tend naturellement à disperser un peu les points d'exploration, et favorise le retirage, dans la mesure où l'incertitude en certains points, du fait du bruit, devient plus forte que l'incertitude liée à la variation de la fonction entre les points de la grille.

La figure 10 illustre un choix de scission sur critère n° 1 comparant directement erreur de grille et erreur d'estimation (cf. §3.1) : elle donne une idée de l'impact du choix du critère de scission, qui conduit à un comportement proche pour les deux critères. Les retirages sur des sommets existants se font lorsque la présence du minimum est suffisamment vraisemblable et que l'incertitude pesant sur les sommets de la zone doit être diminuée.

Pour les deux critères abordés ici, il faut enfin noter que pour un niveau de bruit nul, aucun retirage n'intervient (les graphiques illustrant ce résultat, qui conduisent à des figures similaires à la figure 3, ont toutefois été omis). Cela est logique dans la mesure où ces retirages n'apporteraient rien à la connaissance de la fonction f.

Pour l'algorithme avec retirages possibles, le comportement attendu est le suivant, conforme à ce qui a été observé :

– Lorsque le bruit est faible, ou à plus forte raison lorsque F est déterministe, la seule

	$\sigma_B = 0$ scission syst.		$\sigma_B = 0.1$, $\sigma_e = 0$ scission syst.		$\sigma_B = 0.1$, σ_e estimé scission syst.		$\sigma_B = 0.1$, σ_e estimé avec retirages	
	σ_K	α	σ_K	α	σ_K	α	σ_K	α
$n = 10$	0.6349	1.5052	*0.4755*	*1.2121*	0.4745	1.2146	0.4489	1.2769
$n = 100$	0.9460	1.5924	*0.3921*	*1.0636*	0.6457	1.3743	0.9469	1.5845
$n = 1000$	1.0743	1.5665	*0.1124*	*0.5150*	0.5545	1.2209	0.5598	1.2035

TAB. 1 – Coefficients σ_K et α estimées par maximum de vraisemblance, en présence ou non d'un bruit σ_B, à partir de n points explorés.

incertitude réside dans le comportement de la fonction entre les points, et l'algorithme crée systématiquement de nouveaux points de tirage.
– En présence d'un bruit, les points de retirage se concentrent dans les zones petites, généralement proches d'un point solution. Pour ces points, la réduction de l'erreur de grille, faible sur de courtes distances, est moins prioritaire que la réduction de l'erreur d'estimation.

4.4 Estimation des paramètres (σ_K, α)

Le tableau 1 représente les estimations des coefficients σ_K et α, opérées à partir d'un nombre n de points explorés, pour des données soumises à un bruit σ_B. L'estimation est réalisée par maximum de vraisemblance, comme indiqué au paragraphe 2.3, à l'aide d'un point fixe qui a convergé dans toutes les situations testées. Ici, seuls les segments délimitant les zones ont été retenus pour l'estimation, et non pas les segments joignant deux points explorés de zones distinctes. Les zones en question ont été obtenues par l'algorithme avec des paramètres initiaux $(\sigma_K, \alpha) = (0.3, 0.9)$.

Dans la première colonne de ce tableau 1, lorsque $\sigma_B = 0$ (et en conséquence $\sigma_e = 0$), on peut constater une relative stabilité des paramètres estimés, même dans le cas où l'estimation se base sur un nombre restreint de segments (lorsque $n = 10$ et en l'absence de considération de segments inter-zones). Dans la deuxième colonne, en italique, lorsque $\sigma_B = 0.1$ et pour tout point $\sigma_e(\theta) = 0$, l'estimation est menée comme si l'estimateur $\widehat{f}$ correspondait à la fonction f aux points explorés. Cette colonne 2 est donnée à titre indicatif, dans la mesure où il est parfaitement possible de tenir compte des erreurs d'estimation aux points explorés (cf. colonnes 3 et 4). La non prise en compte des erreurs d'estimation explique le coefficient α inférieur dans le cas bruité, qui correspond bien à une incertitude accrue sur les petites distances, du fait du bruit supporté par $\widehat{f}$. Ce phénomène est d'autant plus marqué que les zones sont petites (ici lorsque n est grand). Dans le cas bruité, le coefficient α inférieur est toutefois compensé par une pente σ_K inférieure. Enfin, dans les troisième et quatrième colonnes, l'estimation est menée sur des données bruitée, en tenant compte cette fois du bruit σ_e aux points explorés. La troisième colonne présente une estimation issue d'un découpage par scission systématique des zones, tandis que la quatrième présente une estimation issue de l'algorithme avec retirages possibles. Les paramètres observés sur ces colonnes 3 et 4 peuvent varier un peu, notamment lorsque le nombre de segments utilisés pour l'estimation est réduit. Toutefois, dans les observations que nous avons pu mener, une valeur moins élevée de α (volatilité supérieure sur de courtes distances) est systématiquement compensée par une valeur moins élevée de σ_K (volatilité globale inférieure) : nous aborderons ce point par l'observation des écarts-types de la variable aléatoire IR_α dans la figure 12. Le comportement de l'algorithme initialisé avec les différents paramètres optimaux

obtenus est resté assez stable dans les exemples que nous avons testés. C'est ce qu'indique la figure 11, qui présente les points d'exploration obtenus, lorsque $\sigma_B = 0.1$, pour l'algorithme initialisé avec les paramètres estimés hors bruit pour $n = 1000$, $(\sigma_K, \alpha) = (1.0743, 1.5665)$ (à gauche) ou estimés en présence de bruit $(\sigma_K, \alpha) = (0.5545, 1.2209)$ (à droite). Un coefficient α supérieur permet de focaliser les recherches un peu plus rapidement sur de petites zones, mais la différence de valeur entre les coefficients α testés est ici trop ténue pour que l'effet global sur le comportement de l'algorithme en soit beaucoup affecté.

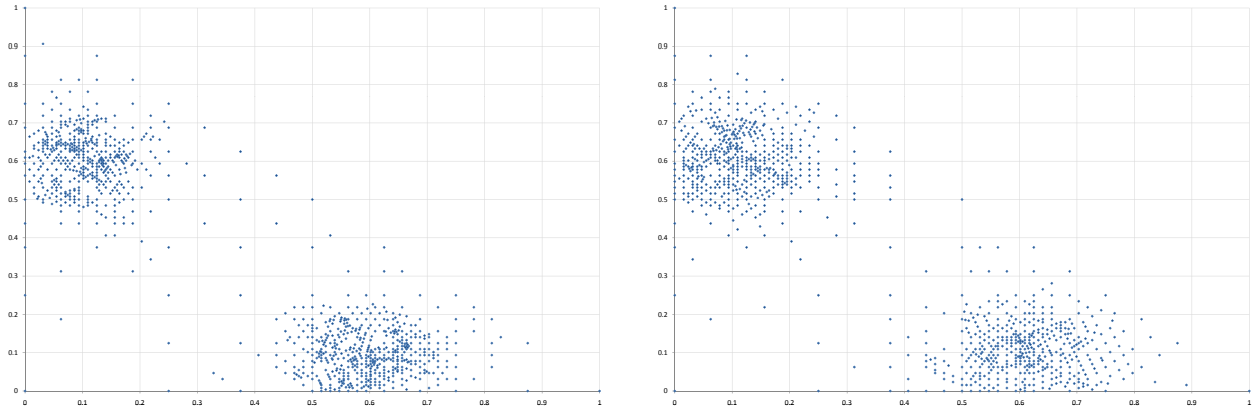

FIG. 11 – Points d'explorations obtenus, lorsque $\sigma_B = 0.1$, avec les paramètres estimés hors bruit, $(\sigma_K, \alpha) = (1.0743, 1.5665)$ (à gauche) ou estimés en présence de bruit $(\sigma_K, \alpha) = (0.5545, 1.2209)$ (à droite)

Enfin, nous avons déterminé, indépendamment de l'algorithme proposé, l'écart-type empirique de la variable aléatoire IR_α :

$$\mathrm{IR}_\alpha(Z) = \frac{f(\theta_1) - f(\theta_2)}{d(\theta_1, \theta_2)^\alpha},$$

θ_1, θ_2 vecteurs aléatoires distincts issus de tirages indépendants, uniformes sur Z.

Le détail de la cette variable aléatoire et des tirages uniformes sur une zone Z est donné dans la section 2.3. L'écart-type de cette variable aléatoire fournit des informations sur la variation des pentes observées sur la zone en fonction de la distance considérée. Comme on peut le voir dans le graphique 12, ces écarts-type sont légèrement supérieurs dans le cas de la demi-zone Z_1 correspondant à la partie de Θ située au dessus de la droite d'équation $y = x$. Cela est logique dans la mesure ou f a la forme d'une cuvette sur cette zone et non plus de deux cuvettes, les pentes sont ici en moyenne un peu plus abruptes que lorsque l'on relie deux points d'une même cuvette, plutôt que deux points de cuvettes différentes. On peut remarquer que les ordres de grandeurs trouvés par maximum de vraisemblance sont tout à fait conformes à l'écart-type empirique de IR_α, notamment de celui de $\mathrm{IR}_\alpha(Z_1)$, dans la mesure où dès la première itération de l'algorithme, le simplexe initial est scindé en une zone Z_1 et sa zone complémentaire.

La variable IR_α n'est pas utilisée par l'algorithme, mais fournit *une* indication sur l'erreur faite a priori sur la régularité de la fonction objectif selon le modèle proposé. Même il ne s'agit pas ici de fournir une étude exhaustive de la distribution de IR_α, qui dépend naturellement de la fonction objectif utilisée, il nous a paru intéressant de montrer quelle pouvait être la nature de cette erreur de modèle sur les données ici testées. Le résultat de cette analyse apparaît dans la figure 13. Sur notre fonction objectif, la distribution a posteriori de IR_α s'est révélée finalement assez proche de la distribution a priori supposée gaussienne, notamment pour $\alpha = 1.2$. Pour

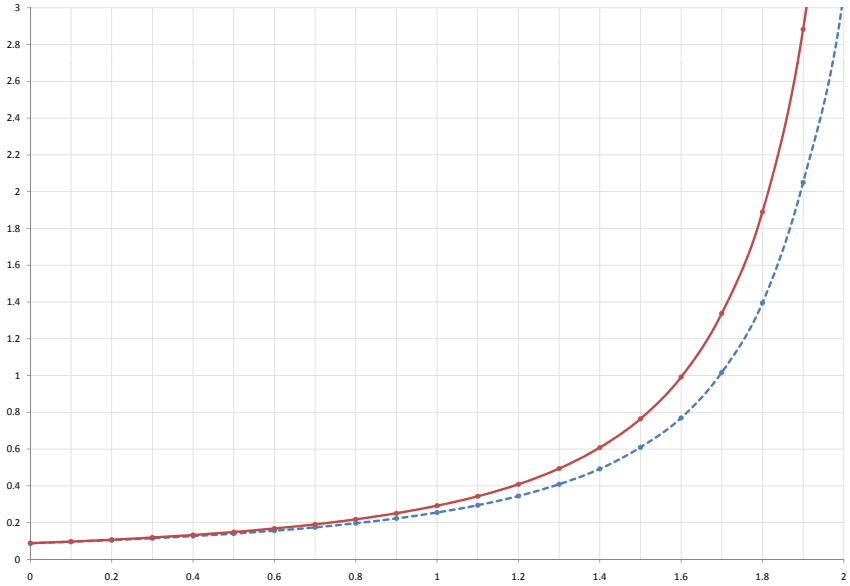

FIG. 12 – Ecart-type empirique de $\mathrm{IR}_\alpha(Z)$ obtenu pour la zone Z_0 correspondant au simplexe orthogonal standard initial (en pointillé) et pour la demi-zone Z_1 correspondant à la partie de Z_0 située au dessus de la première bissectrice (trait plein).

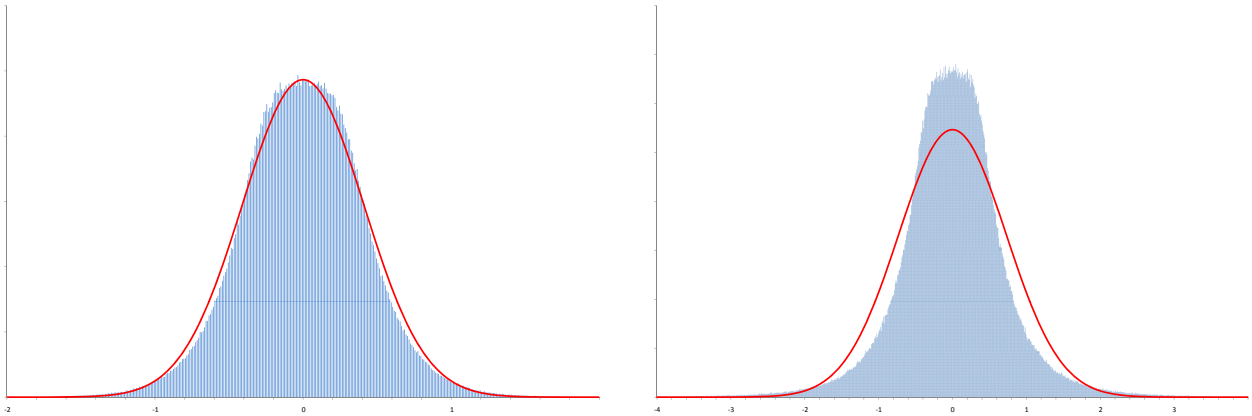

FIG. 13 – Distribution a priori de IR_α (courbe gaussienne continue au premier plan) et histogramme a posteriori, pour la demi-zone Z_1 obtenue après la première itération de l'algorithme. Cas $\alpha = 1.2$ (à gauche) et $\alpha = 1.5$ (à droite). Les écarts-type empiriques obtenus à partir de l'histogramme sont respectivement $\sigma = 0.41$ et $\sigma = 0.73$.

des α plus élevés, la présence de nombreuses valeurs très élevées se traduit par une légère sous-évaluation de la fréquence des pentes très élevées, et une sur-évaluation de la fréquence des pentes moyennes : la distribution observée a posteriori semble leptokurtique. Il est également possible que la distribution de IR_α évolue au fur et à mesure des découpages. Ces résultats indiquent qu'il appartient à l'observateur d'intégrer une éventuelle prudence, en surestimant par exemple le coefficient σ_K ou en sous-estimant le coefficient α, et qu'il est aussi envisageable de modifier a posteriori la distribution supposée des pentes, ou le kurtosis de la distribution, l'hypothèse gaussienne pouvant servir de distribution a priori lors d'une inférence bayesienne.

Un écart important peut également indiquer une erreur dans le choix du coefficient α, le modèle devant conduire à un IR_α proche d'une loi normale pour un α à déterminer, non pour tous. Remarquons toutefois que, convolués avec l'erreur d'estimation en chaque point, les écarts observés peuvent en partie s'estomper. D'autre part, si la hiérarchie entre les zones les plus susceptibles de contenir un optimum global peut se trouver affectée par le décalage observé, elle n'est pas non plus radicalement remise en cause, le coefficient d'aplatissement de la distribution utilisé ayant peu d'incidence pour des valeurs de f proches de la valeur estimée de l'optimum global (pour une zone dont les sommets conduiraient à $\widehat{f}(\theta_i) = \widehat{m}^*$, le potentiel en tout point de la zone serait égal à 1, quel que soit le coefficient d'aplatissement de la distribution utilisée) : le potentiel est utilisé pour opérer un arbitrage entre différents points d'exploration possibles, non pour prévoir précisément le comportement de la fonction en ces points.

4.5 Critères de convergence

critères de convergence calculés Le critère $\rho_1(s)$ est défini dans la section 2.6. Il s'appuie sur une zone de confiance proposée pour l'ensemble des solutions, zone comprenant un ensemble de points de potentiel supérieur à un seuil s. Le critère $\rho_1(s)$ indique la plus grande distance possible entre un point de cette zone et le point solution le plus proche : il peut donc s'interpréter comme un rayon maximal de la zone de confiance depuis chaque point solution.

Le critère ρ_4, dépend quant-à-lui d'un seuil η qui sera, dans les applications numériques présentées, toujours fixé ainsi : $\eta = 0.01$. Ce critère indique la plus grande probabilité (selon le modèle choisi), qu'un minimum plus petit que $m^* - \eta$ soit observé en un point θ (cf. section 2.6 pour une définition précise).

Dans les applications numériques, pour le calcul des critères ρ_1 et ρ_4, l'ensemble des points candidats Θ_0 a été construit sans opérer de nouveaux tirages de F, en piochant a posteriori 50 points dans chaque zone (de façon uniforme, cf. Smith, Tromble, 2004), et en calculant leur potentiel en fonction des points explorés aux sommets de chaque zone. Pour le critère ρ_1, l'ensemble des sommets proposés $\widehat{\mathcal{S}}_{m^*,s}$ a été extrait de Θ_0 en ne retenant que les points dont le potentiel était supérieur au seuil s fixé. Il est enfin rappelé que le critère ρ_2 est construit sur l'ensemble Θ_e des sommets explorés, et le critère ρ_3 sur l'ensemble des zones construites (cf. §2.6) :

$$\begin{aligned}
\widehat{\mathcal{S}}_{m^*,s} &= \{\theta \in \Theta_0, \beta_3(\theta) \geqslant s\}\ , \\
\rho_1(s) &= d^H(\widehat{\mathcal{S}}_{m^*,s}, \mathcal{S}_{m^*})\,, \\
\rho_2 &= \sup_{y \in \mathcal{S}_{m^*}} \inf_{x \in \Theta_e} d(x,y)\,, \\
\rho_3 &= \max_{Z \in \mathfrak{Z}_n} \boldsymbol{\beta}(Z)\,, \\
\rho_4 &= \max_{\theta \in \Theta_0} \beta_3^{(m^*-\eta)}(\theta)\,.
\end{aligned}$$

Fonction objectif non bruitée Bien que cette situation idéale ne soit pas celle nous désirons traiter dans la pratique, nous avons tout d'abord vérifié numériquement la convergence de l'algorithme sur la fonction test non bruitée, dans le cas $\sigma_B = 0$.

Les résultats numériques obtenus pour les critères de convergence apparaissent dans la table 2. Du fait de l'absence de bruit, le minimum estimé $\widehat{m}^*$ était supposé ne pas souffrir d'erreur d'estimation. Nous avons à titre indicatif ajouté dans la table 2 une ligne indiquant les résultats obtenus tenant compte d'une erreur de grille σ_{m^*} pour ce minimum, fonction du diamètre d_* de la zone sur laquelle il était présent, $\sigma_{m^*} = \sigma d_*^\alpha$. La localisation du minimum est

	$\rho_1(s)$	ρ_2	ρ_3	ρ_4
$s = 90\%$, $\sigma_{m^*} = 0$	2.53E-04	3.45E-05	2.43E-08	0
$s = 5\%$, $\sigma_{m^*} = 0$	4.32E-03	3.45E-05	2.43E-08	0
$s = 5\%$, $\sigma_{m^*} > 0$ (erreur de grille)	6.46E-03	1.38E-04	2.28E-07	0

TAB. 2 – Critères de convergence obtenus par l'algorithme lorsque $\sigma_B = 0$, pour $\sigma_K = 1$ et $\alpha = 1.6$, avec ou sans prise en compte d'erreur de grille pour m^*

	$\rho_1(s)$	ρ_2	ρ_3	ρ_4
cas a, $\sigma_B = 0.1$, sans retirage	7.83E-02	2.21E-03	1.32E-06	3.24E-02
cas b, $\sigma_B = 0.1$, retirages	0.11	8.84E-03	1.82E-05	3.99E-02
cas c, $\sigma_B = 0.1$, retirages	5.92E-02	2.21E-03	6.14E-06	4.58E-02

TAB. 3 – Critères de convergence obtenus pour les algorithmes avec ou sans retirage

alors plus dispersée, conduisant alors à une exploration accrue dans une zone plus étalée autour des minima trouvés, au détriment de la vitesse de convergence.

On peut notamment constater dans la table 2 le très bon comportement de l'algorithme en l'absence de bruit, avec des points explorés à une distance d'ordre 10^{-5} de chacune des solutions, et des zones de confiance proposées dans un rayon d'ordre $\rho_1(5\%) \simeq 4 \cdot 10^{-3}$ ou $\rho_1(90\%) \simeq 3 \cdot 10^{-4}$ autour de ces solutions. La zone de confiance proposée est naturellement plus réduite dans le cas où s est elevé.

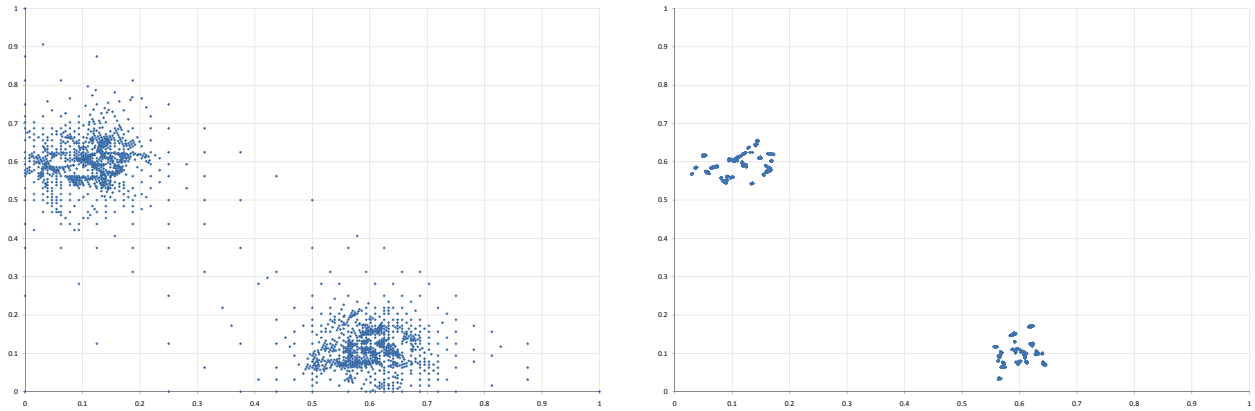

FIG. 14 – Points d'explorations obtenus pour l'algorithme à scission systématique, cas a, lorsque $\sigma_B = 0.1$, avec les paramètres $(\sigma_K, \alpha) = (0.6, 1.2)$ (à gauche). Zone finalement retenue $\mathcal{S}_{m^*,5\%}$ avec un seuil de 5 % (à droite)

Fonction objectif bruitée Nous avons comparé les résultats obtenus par l'algorithme à scission systématique, ainsi que par l'algorithme à retirage pour les critères de scission n° 1 et n° 2. Les figures 14, 15 et 16 illustrent cette comparaison. Ces figures ont été obtenues pour un niveau de bruit $\sigma_B = 0.1$, des paramètres de régularités $(\sigma_K, \alpha) = (0.6, 1.2)$, pour $n = 2000$ points de tirage ou retirage, avec un seuil minimal de $n_0 = 10$ tirages par points. Dans ces figures apparaissent les nuages de points de tirage, ainsi que les zones de confiance proposées

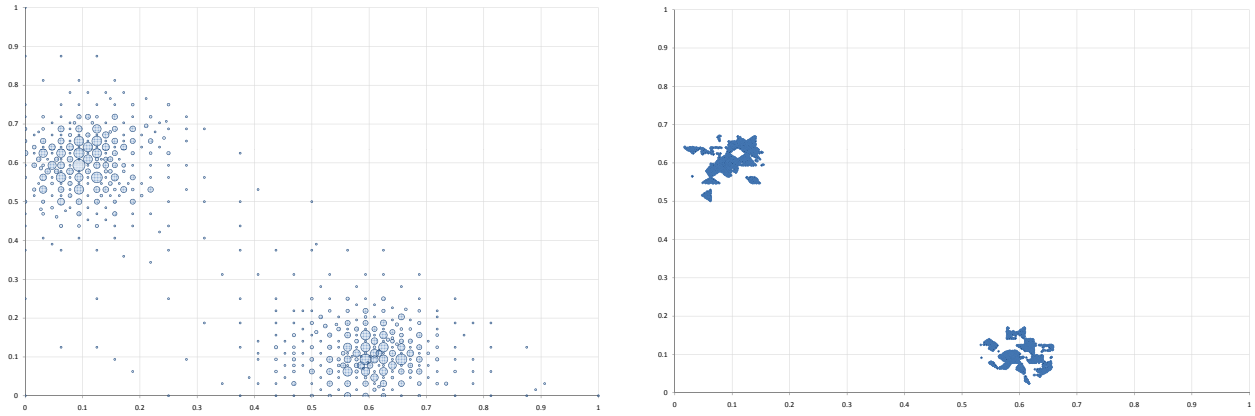

FIG. 15 – Points d'explorations obtenus pour l'algorithme avec retirage, cas b, critère de scission n° 1 de comparaison des erreurs de grille et d'estimation, lorsque $\sigma_B = 0.1$, avec les paramètres $(\sigma_K, \alpha) = (0.6, 1.2)$ (à gauche). Zone finalement retenue $\mathcal{S}_{m^*,5\%}$ avec un seuil de 5 % (à droite)

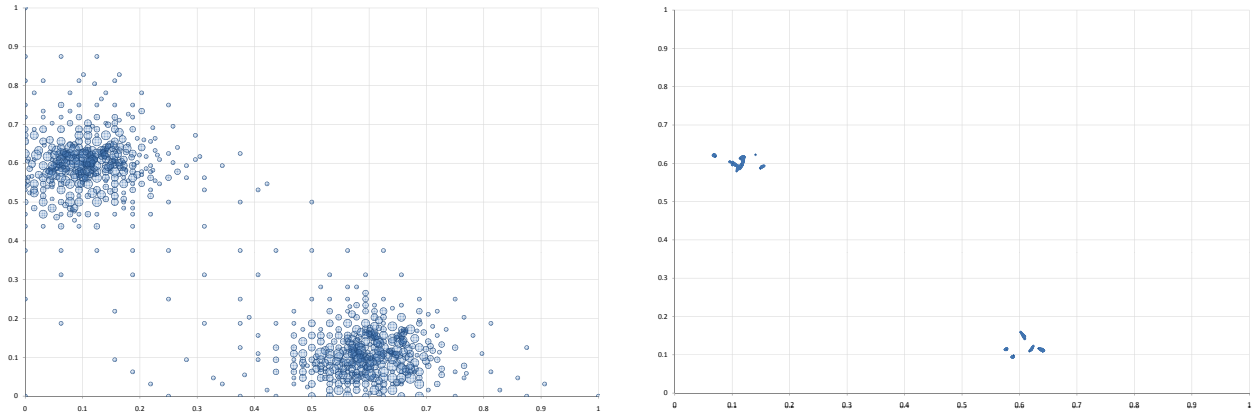

FIG. 16 – Points d'explorations obtenus pour l'algorithme avec retirage, cas c, lorsque $\sigma_B = 0.1$, critère de scission n° 2 considérant l'ajout de n_0 observations, avec les paramètres $(\sigma_K, \alpha) = (0.6, 1.2)$ (à gauche). Zone finalement retenue $\mathcal{S}_{m^*,5\%}$ avec un seuil de 5 % (à droite)

$\mathcal{S}_{m^*,5\%}$. Du fait de l'absence de prise en compte d'erreur de grille pour m^*, et du fait du seuil s utilisé, il est possible qu'un point solution soit très proche mais en dehors d'une zone de confiance.

Les trois algorithmes se comportent correctement, et conduisent à la proposition de zones de confiance dans le voisinage direct des points solutions. Du fait du bruit frappant la fonction f, on observe que des zones très proches peuvent être tantôt exclues, tantôt englobées dans $\mathcal{S}_{m^*,s}$. Les zones de confiance ainsi bâties sont donc très fractionnées, mais bien localisées. Le critère n° 2 basé sur l'estimation des potentiels après ajout de n_0 tirages, a conduit à une zone de confiance légèrement moins dispersée autour des solutions connues.

Les résultats sont récapitulés pour ces mêmes tirages dans le tableau 3. Au regard du critère $\rho_1(5\%)$, l'algorithme à retirage fournit sur ces données les meilleurs résultats pour le critère n° 2. Ces résultats peuvent néanmoins varier d'une exécution à l'autre du fait du caractère aléatoire du bruit pesant sur f et du caractère stochastique de l'algorithme. L'étude de la distribution de $\rho_1(s)$ constitue à cet égard un champ d'investigation intéressant.

priorité	bruit σ_B	(σ_K, α)	$\rho_1(s)$	ρ_2	ρ_3	$\rho_4(\eta)$
absolue	0.1	(0.6, 1.2)	9.23E-02	2.21E-03	6.46E-08	1.08E-02
relative	0.1	(0.6, 1.2)	8.87E-02	2.76E-03	3.78E-07	9.57E-03
absolue	0	(0.6, 1.2)	3.61E-02	5.52E-04	1.17E-06	0
relative	0	(0.6, 1.2)	6.18E-02	5.52E-04	6.77E-06	8.05E-05
absolue	0.1	(1, 1.6)	9.79E-02	2.21E-03	2.04E-07	4.72E-02
relative	0.1	(1, 1.6)	8.69E-02	1.61E-03	1.01E-06	1.79E-02
absolue	0	(1, 1.6)	1.26E-03	3.45E-05	3.17E-09	0
relative	0	(1, 1.6)	2.57E-03	3.45E-05	2.57E-08	0

TAB. 4 – Comparaison des critères de convergence pour les choix *priorité relative* (exploration d'une zone avec une probabilité proportionnelle à son potentiel) ou *priorité absolue*(exploration de la zone de meilleur potentiel)

Impact de la priorité Nous avons omis ici les graphiques qui permettaient de discuter du choix de γ, paramètre présenté dans la section 2.2. Nos observations nous conduisaient à un impact limité de ce paramètre. Dans les lignes suivantes, nous allons néanmoins évoquer plus précisément le choix de la priorité accordée à la zone de meilleur potentiel.

Jusqu'à présent, la zone à subdiviser ou explorer était choisie avec une probabilité proportionnelle au potentiel de la zone (choix qualifié ici "*priorité relative*"). Il est également envisageable de ne plus choisir la prochaine zone de façon stochastique, mais de choisir de façon déterministe zone de meilleur potentiel (choix qualifié ici "*priorité absolue*"). Cela revient à traiter une question déjà abordée : choisir entre explorer davantage la fonction, ou privilégier la vitesse en explorant en priorité une zone prometteuse, au risque d'ignorer un optimum global. Les paramètres (σ_K, α) permettent justement d'opérer cet arbitrage. Si ces paramètres sont en partie estimés, ce dernier choix de priorité *relative* versus *absolue* conduit à un arbitrage naturellement différent. Dans les applications numériques opérées, ce choix n'a pas eu beaucoup d'incidence, notamment en présence de bruit.

Le tableau 4 récapitule les résultats obtenus avec le choix de priorité *relative* ou le choix *absolue*. Nous nous sommes placé pour cette application dans un cadre de scission systématique (pas de retirages), avec un nombre de tirages $n = 2000$. Pour le critère ρ_1, nous avons bâti l'ensemble solution proposé à l'aide d'un seuil s égal à 10% du meilleur potentiel observé, en conservant ainsi une part notable des solutions possibles. Rappelons que le choix d'un seuil supérieur conduit à rétrécir les zones de confiance autour des optima trouvés, et peut améliorer de façon importante le critère de convergence ρ_1, comme on peut le constater en comparant ces résultats avec ceux du tableau 3. Ce critère ρ_1 n'est donc pas comparable à d'autres critères qui seraient obtenus avec d'autres seuils. Par ailleurs l'augmentation du seuil s augmenterait le risque d'ignorer un optimum global, et le critère ρ_1 pourrait alors être brutalement dégradé.

Remarquons tout d'abord dans le tableau 4 la bonne exploration des voisinages de chacun des points solutions : dans tous les cas, après 2000 itérations, des points ont été tirés à une distance ρ_2 inférieure à $3 \cdot 10^{-03}$ de chacune des solutions, avec des zones de confiance d'un rayon ρ_1 raisonnable autour des solutions (qu'il est possible de réduire en augmentant le seuil s). En l'absence de bruit, des tirages sont obtenus, avec $(\sigma_K, \alpha) = (1, 1.6)$, à une distance très réduite d'environ $3 \cdot 10^{-05}$ de chaque point solution.

Au vu des mesures effectuées dans ce tableau 4, le choix *priorité relative* ou *priorité absolue* a surtout un impact dans les situations non bruitées ou les paramètres $(\sigma_K, \alpha) = (1, 1.6)$ privilégient une moindre exploration. Dans ce cas, la convergence est assez rapide, et le choix

priorité absolue conduit à un zone de confiance de diamètre faible $\rho_1 = 1.2 \cdot 10^{-03}$, plus faible que dans le cas *priorité relative*. Dans les autres cas, il apparaît que les mesures ne sont pas radicalement perturbées par ce choix, notamment dans le cas d'un bruit σ_B non nul.

4.6 Comparaison avec l'algorithme de Kiefer-Wolfowitz-Blum

Notre algorithme vise à opérer une optimisation à partir d'une fonction F bruitée, et il n'est pas en pratique possible d'éliminer ce bruit. Pour cette raison, nous ne comparerons pas notre algorithme avec les algorithmes d'optimisation classiques en l'absence de bruit. Nous avons donc choisi de comparer les résultats obtenus avec ceux que donnerait un autre algorithme d'optimisation de fonction bruitée.

L'algorithme utilisé pour la comparaison est un des algorithmes d'optimisation stochastique parmi les plus classiques : il s'agit de l'extension multidimensionnelle, proposée par Blum (1954) de l'algorithme de Kiefer, Wolfowitz (1952). Rappelons tout d'abord que l'algorithme de Kiefer-Wolfowitz vise l'obtention d'un unique optimum, non la garantie d'absence d'autres optima. Tel que décrite dans Broadie, Cicek, Zeevi (2009), la version de l'algorithme de Blum consiste à déterminer une suite de points $\theta^{(1)}, \theta^{(2)}, \ldots$ convergeant vers la solution θ^*, supposée unique, minimisant $f(\theta) = \mathrm{E}\left[F(\theta)\right]$. En dimension $d = 2$, si l'on note $\theta^{(k)} = (\theta_x^{(k)}, \theta_y^{(k)})$, $k \in \mathbb{N}$, la suite de points est telle que :

$$\begin{aligned} \theta_x^{(n+1)} &= \theta_x^{(n)} - a_n \frac{F(\theta^{(n)} + c_n e_x) - F(\theta^{(n)})}{c_n}, \\ \theta_y^{(n+1)} &= \theta_y^{(n)} - a_n \frac{F(\theta^{(n)} + c_n e_y) - F(\theta^{(n)})}{c_n}, \end{aligned}$$

où $e_x = (1, 0)$ et $e_y = (0, 1)$, et où $\{a_n\}_{n\in\mathbb{N}}$ et $\{c_n\}_{n\in\mathbb{N}}$ sont des suites réelles décroissantes en n. Les contraintes auxquelles sont soumises ces constantes, ainsi que les conditions générales d'application et de convergence de l'algorithme, sont détaillées dans Broadie, Cicek, Zeevi (2009).

Remarquons qu'à chaque étape de l'algorithme, le budget de tirage de F est amputé de $d + 1 = 3$ tirages : d tirages pour estimer le gradient, ici en $\theta^{(n)} + c_n e_x$ et $\theta^{(n)} + c_n e_y$, et un tirage pour le nouveau point $\theta^{(n+1)}$. Pour un budget de n points de tirage de la fonction F supposée coûteuse, l'algorithme de pourra pas faire appel à un nombre d'étapes supérieur à $n/3$. A cet égard, cette version de l'algorithme, faisant appel à des différences finies sur un seul côté, dépense moins de tirages que d'autres versions pour l'estimation du seul gradient.

Par ailleurs, les suites $\{a_n\}$ et $\{c_n\}$ semblent délicates à choisir, et les écrits de Broadie, Cicek, Zeevi (2009) mentionnent une grande sensibilité du comportement de l'algorithme en fonction de ces choix. En l'absence supposée d'autres informations sur f, nous avons opté pour un choix par défaut classique de $a_n = 1/n$ et $c_n = 1/n^{1/3}$, choix proposé en première page dans l'article originel Kiefer, Wolfowitz (1952). Les paramètres $a_n = a_0/n$, $c_n = c_0/n^{1/4}$, $n \geqslant 1$, présentés dans Broadie, Cicek, Zeevi (2009) conduisaient ici à des résultats décevants pour $a_0 = c_0 = 1$ et les quelques choix alternatifs testés, mais nous n'avons pas cherché spécifiquement à trouver a posteriori les meilleurs paramètres, pour une fonction f supposée inconnue. En conséquence, les résultats numériques obtenus ici sont représentatifs d'un type de trajectoire classiquement obtenue lors d'une descente stochastique de gradient, mais certainement améliorables en terme de vitesse de convergence, les suites $\{a_n\}$ et $\{c_n\}$ pouvant être modifiées à cette fin.

Dans la plupart des illustrations précédentes, pour l'algorithme à scission systématique, $n_0 = 10$ tirages étaient opérés en chacun des points explorés, pour un budget de 2000 points d'exploration. Le budget global de tirage de F était donc de $2000 n_0$ tirages.

La figure 17 rend compte des points de tirage obtenus par l'algorithme de Kiefer-Wolfowitz-Blum, pour 2000 points de tirage d'une fonction F soumise à un bruit $\sigma_B/\sqrt{n_0}$ (figure de gauche), ou pour $2000n_0$ tirages d'une fonction F soumise à un bruit σ_B (figure de droite). Le nombre global de tirage correspond ainsi à celui des illustrations précédentes. Le point initial a été tiré aléatoirement, de façon uniforme, sur le simplexe initial (cf. Smith, Tromble, 2004).

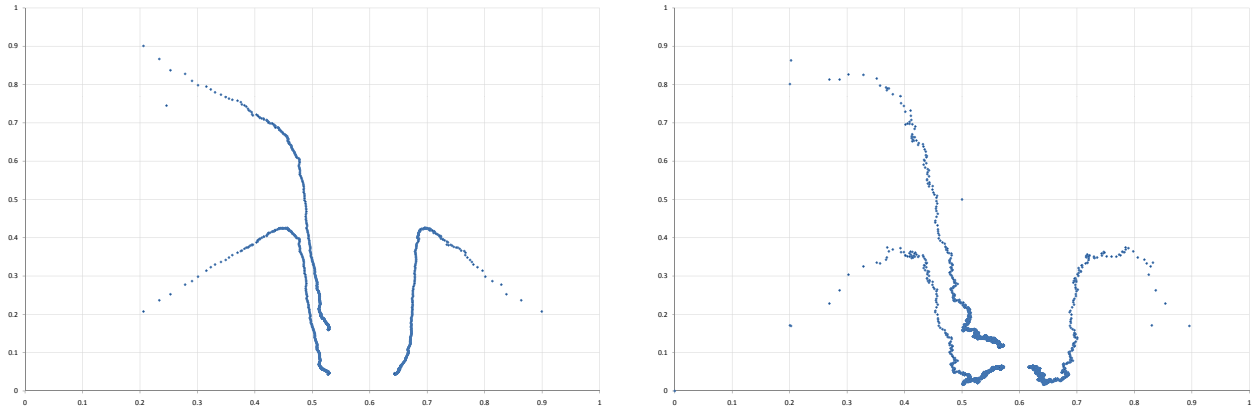

FIG. 17 – Points d'explorations obtenus pour l'algorithme Kiefer-Wolfowitz-Blum, pour 2000 tirages avec $\sigma_B = 0.1/\sqrt{10}$ (à gauche), ou 20000 tirages avec $\sigma_B = 0.1$(à droite)

Sur la figure 17, on constate tout d'abord la présence de trois séries de points, correspondant à l'ensemble des points suggérés, ainsi qu'aux points décalés verticalement et horizontalement pour l'estimation du gradient (points décalés respectivement vers le haut et vers la droite). Les points sont initialement assez espacés (a_n grand) puis sont de plus en plus proches au fil des tirages (a_n petits, lorsque l'on est en principe proche d'une solution). Sur les données moins bruitées (figure de gauche), les gradients estimés sont moins erratiques, les directions choisies d'un point à l'autre étant plus stables.

Cette figure 17 est essentiellement intéressante dans la mesure où elle marque bien les différences d'approches et de finalité entre les algorithmes de descente stochastique de gradient et l'algorithme proposé dans le présent article. Concernant l'algorithme de Kiefer-Wolfowitz-Blum, on peut notamment faire les constats suivants :

- En dehors de l'espace occupé par les trois trajectoires qui se dessinent, la fonction F est très peu explorée sur le reste de la zone de recherche, et est donc susceptible d'ignorer des optima globaux. Cela est logique dans la mesure où l'on n'exige normalement pas d'un algorithme d'optimisation locale qu'il fournisse un optimum global.
- Par ailleurs, même en présence supposée d'un unique optimum, la construction de zones de confiance pour l'optimum, à partir des trois trajectoires obtenues, semble ici délicate.
- Des points à l'exterieur du simplexe initial Z_0 ont pu être tirés, dans la mesure où $f(\theta)$ pouvait prendre ici des valeurs en dehors du simplexe initial, mais l'exploitation de cet algorithme supposerait l'introduction préalable de contraintes.
- L'algorithme se comporte ici de façon cohérente dans la mesure où la fonction f est suffisamment régulière sur une grande partie du simplexe initial (en dehors de la première diagonale). La situation serait bien différente en cas de non-dérivabilité de f sur l'essentiel du domaine Z_0.

S'agissant des critères de convergence, l'algorithme de Kiefer-Wolfowitz-Blum ne permet malheureusement pas a priori de partitioner la zone de recherche initiale Z_0 en plusieurs zones.

	ρ_2	$\tilde{\rho}_2$
cas $n = 2000$, $\sigma_B = 0.1/\sqrt{10}$	0.2058	6.35E-2
cas $n = 20000$, $\sigma_B = 0.1$	0.2248	3.342E-2

TAB. 5 – Critères de convergence obtenus avec l'algorithme de Kiefer-Wolfowitz (pour des séquences $\{a_n\}$ et $\{c_n\}$ non spécifiquement optimisées)

En conséquence, nous ne calculerons pas ici de potentiel. En supposant que Θ_e désigne l'ensemble des points de tirage obtenus, nous réutiliserons donc le critère ρ_2 :

$$\rho_2 = \sup_{y \in \mathcal{S}_{m^*}} \inf_{x \in \Theta_e} d(x, y)\,,$$

Ce critère permet de savoir si il existe des points de tirage proches de chacune des solutions.

L'algorithme de Kiefer-Wolfowitz-Blum recherchant une unique solution est très fortement pénalisé par ce critère, car il peut converger vers une solution tout en ayant des points d'exploration très éloignés des autres solutions. Un second critère retenu spécifiquement pour cet algorithme sera :

$$\tilde{\rho}_2 = \inf_{y \in \mathcal{S}_{m^*}} \inf_{x \in \Theta_e} d(x, y)\,.$$

Ce dernier critère indique juste si il existe des points de tirage proches de l'une des solutions.

Les résultats obtenus sont indiqués dans le tableau 5, qui correspond aux graphiques de la figure 17.

D'une part, l'algorithme de Kiefer-Wolfowitz-Blum ne recherche qu'un optimum local. Nous devrions donc obtenir un critère ρ_2 de convergence vers un optimum global très dégradé par rapport à l'algorithme de scission systématique, ce qui est bien le cas.

D'autre part, la recherche d'un optimum local doit intuitivement être moins coûteuse que la recherche d'un optimum global. Les points explorés près d'une unique solution devraient être plus proches de la solution pour un algorithme d'optimisation locale et nous devrions obtenir un critère $\tilde{\rho}_2$ meilleur. Ce n'est pas le cas ici : les mesures de ρ_2 obtenues avec l'algorithme à scission systématique ont conduit (cf. tableau 4) à des points toujours tirés à une distance ρ_2 inférieure à $3 \cdot 10^{-3}$ de chaque solution. Or, la distance $\tilde{\rho}_2$ à la solution la mieux explorée est encore inférieure à ρ_2. L'algorithme de Kiefer-Wolfowitz Blum a conduit quant-à-lui à des résultats de l'ordre de 0.2 pour la distance ρ_2, et $3 \cdot 10^{-2}$ pour la distance $\tilde{\rho}_2$. Notre algorithme a donc été ici localement et globalement plus performant. Ce dernier constat est sans doute lié à l'absence d'optimisation spécifique des séquences $\{a_n\}$ et $\{c_n\}$, qui reste toutefois délicate à opérer a priori sur une fonction supposée inconnue.

4.7 une réflexion sur la montée en dimension

Les applications ont ici été présentées dans le cas $d = 2$. Dans le cas plus général où d devient grand, on peut s'interroger sur la rapidité de convergence de l'algorithme, du fait du problème de la montée en dimension.

le problème de la dimension Notons $Z_0^{(d)}$ le domaine de recherche initial Z_0, de façon à rendre explicite la dimension d du problème. Imaginons un cas où l'ensemble de la zone initiale $Z_0^{(d)}$ doit être exploré ; ce cas survient en particulier si l'on cherche les minimiseurs d'une fonction constante bruitée. Supposons que l'on souhaite que chaque solution potentielle

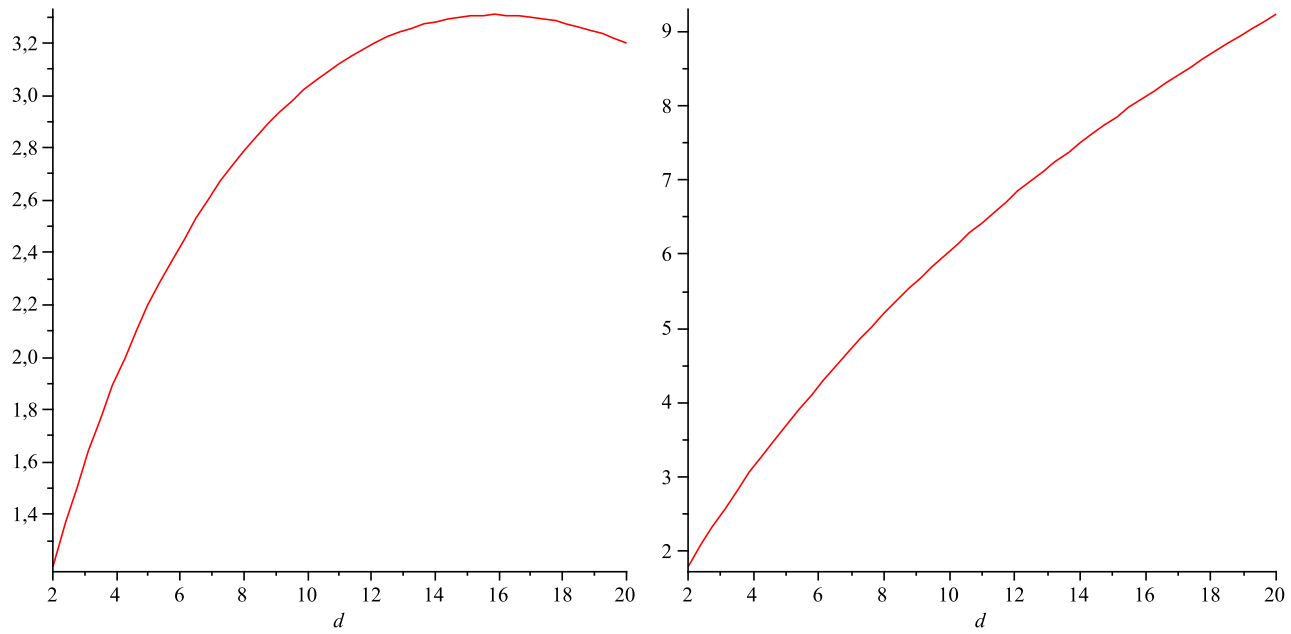

FIG. 18 – $\log_{10}(n_\rho(d))$ en fonction de d, pour $\rho = 0.1$ (à gauche) et $\rho = 0.05$ (à droite)

se trouve à une distance maximale ρ d'un point exploré, ρ étant fixé. Sans entrer dans le détail des calculs, on peut constater que, en dimension d, le nombre minimal de $d-$sphères $S_\rho^{(d)}$ de rayon ρ nécessaires pour recouvrir $Z_0^{(d)}$ est donné par le ratio des deux volumes :

$$n_\rho(d) \quad = \quad \frac{V(Z_0^{(d)})}{V(S_\rho^{(d)})}\,.$$

La figure 18 montre l'évolution du logarithme en base 10 de n_ρ en fonction de la dimension d du problème. Cela donne une indication du nombre de tirages à réaliser pour espérer obtenir des tirages à une distance ρ de chacune des solutions, lors d'une exploration *exhaustive* du domaine de recherche initial. On voit alors qu'il peut être envisageable d'exiger une précision de l'ordre de $\rho = 0.1$ pour l'exploration, mais que simplement diviser par deux cette précision conduit à une explosion du nombre de tirages requis lorsque la dimension augmente. Ce constat se base toutefois sur le cas d'une exploration *exhaustive* du domaine initial, comme lorsque l'on cherche les minimiseurs d'une fonction constante bruitée. Dans la pratique, le recouvrement des minimiseurs par des sphères de rayon ρ occupe un volume souvent très restreint de l'ensemble de recherche initial. Il devient alors possible, selon la fonction f considérée, d'espérer atteindre une distance ρ inférieure pour un budget de tirages raisonnable. Si le potentiel de $\tilde{f}$ décroît suffisamment entre différentes zones à une distance $\rho = 0.1$ explorable, on peut espérer exclure temporairement suffisamment de zones de l'analyse pour atteindre des distances ρ inférieures.

Le problème de la montée en dimension est un problème épineux qui n'est pas spécifique à l'algorithme présenté. Ce problème est évoqué en détail notamment dans Bellman (1957) : lorsque la dimension d du problème augmente, il devient très coûteux d'explorer exhaustivement un domaine $[0,1]^d$. Ce constat est ici tempéré par le fait que le volume $V(Z_0^{(d)})$ diminue rapidement en fonction de la dimension. D'autre part, ce constat peut ne plus être valable si l'exploration du domaine initial n'est pas exhaustive.

Par ailleurs, l'implémentation pratique de l'algorithme en grande dimension conduit assez rapidement à une augmentation importante du nombre de zones de scission, malgré un nombre de points d'exploration restreint. L'algorithme nécessite alors une implémentation spécifique

	$\rho_1(s)$	ρ_2	ρ_3	ρ_4	ρ_2^{dir}	ρ_2^U	$\tilde{\rho}_2^U$
$d=2$	2.21E-02	**1.38E-04**	8.30E-07	0	9.77E-05	**1.31E-02**	2.83E-03
$d=3$	9.09E-02	**1.17E-02**	3.43E-08	6.18E-03	6.75E-03	**3.24E-02**	2.29E-02
$d=4$	0.30	**0.15**	1.13E-07	0.21	7.53E-02	**0.17**	0.17
$d=5$	0.37	**0.27**	1.05E-06	0.16	0.12	**0.32**	0.29

TAB. 6 – Critères de convergence obtenus par l'algorithme lorsque $\sigma_B = 0.1$, pour $\sigma_K = 1$ et $\alpha = 1.5$, $s = 5\%$, $\sigma_{m^*} > 0$ en fonction de la dimension d

pour gérer ce grand nombre de zones, ou une adaptation du mécanisme de scission pour limiter leur nombre. L'étude et l'adaptation du comportement de l'algorithme à de grande dimensions reste un champ d'investigation très large.

résultats numériques pour les premières dimensions Nous allons ici tester l'algorithme à scission systématique dans le cas où la dimension augmente. Nous considérons ici le cas où $f(\theta)$ est définie en dimension $d > 2$ par la formule :

$$\begin{aligned} f(\theta) &= (\min(x_1, x_2) - 0.1)^2 + (\max(x_1, x_2) - 0.6)^2 + \sum_{i=3}^{d} (x_i - 0.3)^2 \,, \\ F(\theta) &= f(\theta) + \sigma_B(U - 0.5)\,, \\ \theta &= (x_1, \ldots, x_d)\,, \end{aligned}$$

avec U une variable aléatoire uniforme sur $[0, 1]$. Cela revient à ajouter un paraboloïde à la fonction test précédemment utilisée. L'espérance f de la fonction F admet deux minima :

$$\theta_1^* = (0.1, 0.6, 0.3, \ldots, 0.3)\,, \theta_2^* = (0.6, 0.1, 0.3, \ldots, 0.3)\,,$$

La valeur de f en ces points θ_1^* et θ_2^* est 0.

Afin d'avoir une idée de l'erreur opérée sur chaque composante pour une distance ρ_2 donnée en dimension d, nous noterons $\rho^{dir} = \rho_2/\sqrt{d}$: la distance euclidienne entre deux points translatés d'un vecteur $(\rho^{dir}, \ldots, \rho^{dir})$ est alors égale à ρ_2. Nous avons également indiqué dans le tableau 4.7 les distances ρ_2^U et $\tilde{\rho}_2^U$ obtenues à partir de n points piochés de façon uniforme sur le simplexe initial Z_0 :

$$\begin{aligned} \rho_2^U &= \sup_{y \in \mathcal{S}_{m^*}} \inf_{x \in \Theta_U} d(x, y)\,, \\ \tilde{\rho}_2^U &= \inf_{y \in \mathcal{S}_{m^*}} \inf_{x \in \Theta_U} d(x, y)\,. \end{aligned}$$

où Θ_U est l'ensemble des n points piochés de façon uniforme sur Z_0, comme indiqué dans Smith, Tromble (2004). Ces distances ρ_2 et $\tilde{\rho}_2$ donnent une indication de la précision obtenue pour une exploration du domaine Z_0 ne privilégiant aucune zone, de façon assez semblable à une grille à pas fixe. Les distances ρ_2 et ρ_2^U étant comparables entre elles, nous les avons placées en caractères gras dans le tableau 4.7.

En dimension $d = 3$, les résultats sont donnés de la table 4.7 sur la ligne $d = 3$. Pour ces mêmes données, les points explorés les plus proches de chacune des deux solutions sont respectivement :

$$\begin{gathered} \widehat{\theta_1^*} = (0.09375, 0.59375, 0.296875)\,, \\ \widehat{\theta_2^*} = (0.59375, 0.109375, 0.296875)\,, \\ \text{ici } \rho_2 = d(\theta_2, \theta_2^*) = 1.169 \cdot 10^{-2},\ \tilde{\rho}_2 = d(\theta_1, \theta_1^*) = 9.375 \cdot 10^{-3}\,. \end{gathered}$$

Les points solutions proposés sont très proches des solutions réelles dans la dimension $d = 3$, malgré la présence d'un bruit $\sigma_B = 0.1$. En dimension $d = 5$, le comportement de l'algorithme a été décevant dans le cas $\sigma_B = 0.1$, les résultats obtenus pour les critères de convergence sont indiqués dans la dernière ligne du tableau 4.7, et sont dégradés par rapport aux résultats des dimensions inférieures.

Toutefois, en présence d'un bruit $\sigma_B = 0.1$, l'algorithme proposé a toujours conduit à des distances ρ_2 inférieures à celles obtenues pour des tirages uniformes. La précision atteinte est bien meilleure dans les petites dimensions, mais s'estompe au fur et à mesure que la dimension augmente : cela traduit le fait que lorsque le domaine de recherche est trop vaste, l'algorithme se voit contraint d'explorer sans préférence un grand nombre de zones, conduisant à un comportement proche de l'exploration par une grille à pas fixe.

Enfin, nous n'avons pas placé les résultats obtenus en utilisant une *priorité absolue* plutôt qu'une *priorité relative*. Les résultats obtenus étaient très proches de ceux présentés ici. L'impact du bruit est également atténué en grande dimension : du fait de l'effet de dilution lié à la montée en dimension, l'algorithme se focalise peu sur les solutions supposées, et adopte alors un comportement similaire à celui adopté en environnement très bruité.

5 Conclusion

Nous avons présenté un algorithme permettant de rechercher des paramètres de $\mathbb{R}^d$ conduisant à minimiser, globalement, une fonction réelle bruitée. L'algorithme permet également de construire des zones de confiance autour des paramètres solutions supposés.

L'approche retenue s'appuie sur la définition du potentiel d'une fonction en un point, et sur la mesure de l'incertitude frappant la fonction objectif entre les points déjà explorés. Cette incertitude découle d'une part des *erreurs d'estimation* aux point explorés, d'autre part des *erreurs de grille* liées à la distance entre un point considéré et les points explorés alentours. L'algorithme permet de moduler facilement, au moyen des deux paramètres σ_K et α, la réitération de tirages dans un voisinage des points solutions découverts, ou à l'inverse l'exploration de la fonction dans des zones encore peu explorées. En outre, une procédure d'estimation de la régularité fonction objectif f permet d'ajuster les valeurs de ces deux paramètres.

L'algorithme présenté ici vise simplement à montrer l'intérêt et la faisabilité technique d'une grille à pas variable, susceptible de répondre au problème du choix entre exploration et connaissance d'une fonction aléatoire. L'approche proposée ici se voulait simple, avec notamment la définition d'un potentiel directionnel pour une unique dimension, puis l'agrégation de potentiels directionnels sur l'ensemble des dimensions. Il est clair que de nombreux autres choix peuvent être faits pour la mesure du potentiel d'un point et d'une zone. Les techniques d'agrégation et de tirages peuvent également être amendées de façon à améliorer l'efficacité de l'algorithme sur des architectures parallèles.

Pour des dimensions pas trop élevées, le comportement de l'algorithme est assez convaincant sur de nombreux points, comme sa capacité d'exploration globale du domaine, sa capacité à privilégier les zones proches des optima, sa faible exigence sur la régularité de la fonction objectif, son adaptation à un bruit nul comme à un bruit élevé. Toutefois, lorsque la dimension du problème devient grande, l'algorithme conduit à des résultats proches de ceux obtenus par une exploration uniforme du domaine initial.

L'analyse détaillée de la vitesse de convergence de l'algorithme et l'amélioration éventuelle de cette vitesse en grande dimension constituent une suite logique de cette étude.

Remerciements Nous tenons à exprimer notre sincère gratitude envers le referee anonyme qui, par ses très nombreuses remarques et suggestions, a beaucoup contribué à améliorer la qualité de ce papier. Nous remercions également Areski Cousin pour sa relecture et ses remarques.

Références

Aarts, E.H.L., Laarhoven V. (1985) *Statistical cooling : a general approach to combinatorial optimization problems,* Philips J.Res, 40 (4), 193-226.

Alliot, J.M. (1996) *Techniques d'optimisation stochastique appliquées aux problèmes du contrôle aérien.* INPT, Habilitation à Diriger des Recherches.

Bellman, R.E. (1957) *Dynamic Programming.* Princeton University Press, Princeton, NJ.

Bienvenüe, A., Rullière, D. (2010) *Iterative adjustment of survival functions by composed probability distortions,* submitted.

Blanke, D. (2002) *Estimation du coefficient de régularité locale d'une trajectoire de processus,* C.R. Acad. Sci. Paris, Ser. I, 334, 145-148.

Blum, J.R. (1954) *Multidimensional stochastic approximation methods.* Annals of Mathematical Statistics, 25, 737-744.

Branke, J., Meisel, S., Schmidt, C. (2008) *Simulated annealing in the presence of noise.* Journal of Heuristics, vol. 14, n° 6, pp.627-654.

Broadie, M., Cicek, D.M., Zeevi, A. (2009) *An adaptative multidimensional version of the kiefer-Wolfowitz stochastic approximation algorithm,* Proceeding of the 2009 Winter Simulation Conference. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin and R.G. Ingalls, eds.

Bulger, D.W., Romeijn, H.E. (2005) *Optimising noisy objective functions,* Journal of Global Optimization, 31 : 599-600.

Cohen, G., Culioli, J.-C. (1994) *Optimisation stochastique sous contraintes en espérance.* Rapport interne Centre Automatique et Systèmes, Ecole des Mines de Paris, no. A–288.

De Berg, M., Cheong, O., van Kreveld, M., Overmars, M. (2008) *Computational Geometry : Algorithms and Applications,* Springer-Verlag.

Emmerich, M.T.M. (2005) *Single and Multi-objective evolutionary design optimization assisted by Gaussian Random Field Metamodels.* Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften der Universität Dortmund, Dortmund.

Ginsbourger, D. (2009) *Multiples métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariables,* Thèse de doctorat de mathématiques appliquées, Ecole nationale supérieure des mines de Saint-Etienne, n° 519MA.

Horst, R., Parlados, P.M. (1995) *Handbook of Global Optimization,* Kluwer Academic Publishers, Dordrecht Boston London.

Hansen, E.R. (1979) *Global optimization using interval analysis : the one dimensional case,* JOTA 29 :331-344.

Jones, D.R., Pertunen, C.D., Stuckman (1993) *Lipschitzian optimization without the Lipschitz constant.* Journal of Optimization Theory and Applications, 79(1), 157-181.

Jones, D.R., Schonlau, M., Welch, W.J. (1998) *Efficient global optimization of expensive black-box functions,* Journal of Global Optimization, 13, 455-492.

Jones, D.R. (2001) *A taxonomy of global optimization methods based on response surface (2001).* Journal of Global Optimization, 21 :345-383.

Kiefer, J., Wolfowitz, J. (1952) *Stochastic estimation of the maximum of a regression function.* Annals of Mathematical Statistics, 23, 462-466.

Krige, D.G. (1951) *A statistical approach to some mine valuations and allied problems at the Witwatersrand,* Master's thesis of the university of Witwatersrand.

Lawler, E.L., Wood, D.E. (1966) *Branch and Bound methods : a survey,* Operations Research, Vol. 14, n° 4, pp 699-719.

Mathias, K., Whitley, D., Kusuma, A., Stork, C. (1996) An empirical evaluation of genetic algorithms on noisy objective functions.

Nelder, J., Mead, R. (1965) *A simplex method for function minimization,* Computer Journal, vol. 7, n° 4, p.308-313.

Norkin, V., Pflug, G.Ch., Ruszczyński, A. (1996) *A branch and bound method for stochastic global optimization,* Mathematical Programming, vol 83, n° 1-3, pp 452-450.

Robbins, H., Monro, S. (1951) *A Stochastic approximation method.* Annals of Mathematical Statistics, 22, 400-407.

Rullière, D., Faleh, A., Planchet, F. (2010) *Fonction d'information et optimisation globale d'une fonction bruitée,* preprint.

Santner, T., Williams, B., Notz, W. (2003) *The design and analysis of computer experiments,* Springer Verlag, New York.

Schubert, B. (1972) *A sequential method seeking the global maximum of a function.* SIAM J. Numer. Anal., 9 :379-388.

Smith, N.A., Tromble, R.W. (2004) *Sampling uniformly from the unit simplex,* Technical Report, Johns Hopkins University.

Sommerville, D.M.Y. (1958) *An introduction to the Geometry of n dimensions,* New York, Dover, p.124.

Strugarek, C. (2006) *Approches variationnelles et autres contributions en optimisation stochastique.* ENPC, Thèse de doctorat.

Villemonteix, J. (2009) *Optimisation de fonctions coûteuses,* Thèse de doctorat de physique, Université Paris Sud 11, Faculté des sciences d'Orsay, n° 9278.

Wolfe, M.A. (1996) *Interval Methods for Global Optimization,* Applied Mathematics and Computation, Vol. 75, Issues 2-3, pp. 179-206.

Zadeh, L.A. (1965) *Fuzzy Set,* Information and Control, 8 (3) : 338-353.

6 Appendice

Preuve du Lemme 2.1. Soit P un point de Z, en notant $J = \{1, \ldots, d+1\}$ et $\{\theta_j\}_{j\in J}$ l'ensemble des sommets de Z, on peut écrire P à l'aide de coordonnées barycentriques $P = \sum_{j\in J} \omega_j\theta_j = (\sum_{j\in J,\ j\neq j_1,\ j\neq j_2} \omega_j\theta_j) + \omega_{j_1}\theta_{j_1} + \omega_{j_2}\theta_{j_2}$. En posant $\theta_c = \frac{1}{2}(\theta_{j_1} + \theta_{j_2})$, on peut écrire

- Si $\omega_{j_1} \geqslant \omega_{j_2}$, alors nécessairement $\omega_{j_2} < \frac{1}{2}$ et

$$P = \sum_{j\in J} \omega_j\theta_j = \sum_{j\in J,\ j\neq j_1,\ j\neq j_2} \omega_j\theta_j + (\omega_{j_1} - \omega_{j_2})\theta_{j_1} + 2\omega_{j_2}\theta_c.$$

  Alors $P \in Z_2$.
- Si $\omega_{j_2} \geqslant \omega_{j_1}$, alors nécessairement $\omega_{j_1} < \frac{1}{2}$ et

$$P = \sum_{j\in J} \omega_j\theta_j = \sum_{j\in J,\ j\neq j_1,\ j\neq j_2} \omega_j\theta_j + (\omega_{j_2} - \omega_{j_1})\theta_{j_2} + 2\omega_{j_1}\theta_c.$$

  Alors $P \in Z_1$.

Il s'ensuit que $Z \subset Z_1 \cup Z_2$. Par ailleurs, il est clair que $Z_1 \cup Z_2 \subset Z$. Enfin, on établit que $Z_1 \cap Z_2$ est l'ensemble des points P tels que $\omega_{j_1} = \omega_{j_2}$, ce qui implique l'inclusion de $Z_1 \cap Z_2$ dans un hyperplan de $\mathbb{R}^d$. □

version de ce document en date du 14 juin 2010, 9:56.