2. Algorithmes stochastiques en optimisation globale

Laurent Dumas

Laboratoire de Mathématiques de Versailles
Université de Versailles Saint Quentin en Yvelines

Ecole d'été Mathématiques et Interactions, Agadir, 17 mai 2011

Problème d'optimisation à résoudre

- On s'intéresse ici au problème de l'optimisation globale d'une fonction $J : \mathcal{O} \rightarrow \mathbb{R}$ dont le calcul peut éventuellement être complexe et coûteux.
- Dans le cas où $\mathcal{O} \subset \mathbb{R}^n$, le choix doit se porter vers une méthode d'optimisation sans gradient.
- La plupart des méthodes déterministes existantes (Nelder Mead, NEWUOA, etc...) ne permettent de rechercher qu'un minimum local.

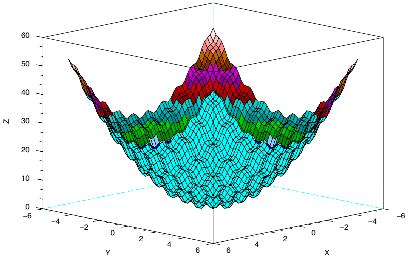
Algorithmes stochastiques : historique

- Les méthodes de type Monte Carlo et leurs variantes (recuit simulé) sont les premiers algorithmes stochastiques ayant été introduits en optimisation.
- Développés plus récemment, les algorithmes évolutionnaires (algorithmes génétiques, stratégies d'évolution, PSO, etc...) sont des algorithmes stochastiques d'optimisation qui tirent leur nom d'une analogie avec la théorie de l'évolution des espèces de Darwin.
- Références : *Holland* (1976), *Goldberg* (1989), *Cerf* (1994), *Schoenaueur* (1996), *Hansen* (2001), etc...

Algorithmes stochastiques

- On présente ici 4 méthodes (SA, GA, ES, PSO) qui seront testées sur la fonction de Rastrigin :

$$Rast(x_1, ..., x_n) = \sum_{i=1}^{n} \left(x_i^2 - \cos(2\pi x_i)\right) + n$$



- Celle-ci possède un grand nombre de minimas locaux (10^n sur $[-5, 5]$) et un seul minimum global.

Principe général du recuit simulé (SA)

- Choix d'un élément initial $X^1 \in \mathcal{O}$
- for n from 1 to $Ngen$
- Mutation : remplacer X^n par Y^n, choisi aléatoirement dans un voisinage.
- Evaluation de $J(Y^n)$.
  - si $J(Y^n) < J(X^n)$, alors $X^{n+1} = Y^n$.
  - si $J(Y^n) \geq J(X^n)$, alors $X^{n+1} = Y^n$ avec une probabilité $exp(-\frac{J(Y^n)-J(X^n)}{T})$ et $X^{n+1} = X^n$ sinon.
- Mise à jour du paramètre T ($T \rightarrow 0$ lentement)
- end for

Principe général d'un algorithme génétique (GA)

- Choix d'une population initiale $P_1 = \{X_i^1 \in \mathcal{O}, 1 \leq i \leq N_p\}$
- for n from 1 to $Ngen$
- Evaluation de $\{J(X_i^n), 1 \leq i \leq N_p\}$.
- Creation d'une population de N_p individus par :
  - Selection de (X_α^n, X_β^n) en fonction de leur facteur de santé.
  - Croisement : remplacer (X_α^n, X_β^n) par (Y_α^n, Y_β^n).
  - Mutation : remplacer (Y_α^n, Y_β^n) par (Z_α^n, Z_β^n).
- end for
- Generation de la nouvelle population P_{n+1}.
- end for

Principe général d'une stratégie d'évolution (ES)

- Choix d'une population initiale de μ parents :
  $P_1 = \{X_i^1 \in \mathcal{O}, \quad 1 \leq i \leq \mu\}$
- for n from 1 to $Ngen$
- Creation d'une population de $\lambda \geq \mu$ enfants O_n par :
  - Croisement à ω parents $Y_i^n = \frac{1}{\omega}(\sum_{j=1}^{\omega} X_j^n)$
  - Mutation : remplacer Y_i^n par Z_i^n
- Evaluation de $\{J(Z_i^n), 1 \leq i \leq \lambda\}$
- Selection des meilleurs μ parents dans la population $P_n \cup O_n$.
- end for

Principe général d'un essaim de particules (PSO)

- Choix d'une population initiale $P_1 = \{(X_i^1, v_i^1, p_i^1), 1 \leq i \leq N_p\}$ de particules ayant la position actuelle $X_i \in \mathcal{O}$, la vitesse v_i et une meilleure position p_i.
- for n from 1 to $Ngen$
- Evaluation de $\{J(X_i^n), 1 \leq i \leq N_p\}$.
- Actualisation de la meilleure position individuelle et globale(p_g^n)
  - Calcul des nouvelles vitesses de chaque particule :

$$v_i^{n+1} = \omega v_i^n + c_1 \rho_1 (p_i^n - x_i^n) + c_2 \rho_2 (p_g^n - x_i^n)$$

  - Calcul des nouvelles positions de chaque particule :

$$X_i^{n+1} = X_i^n + v_i^{n+1}$$

- end for

Principaux avantages

- Grâce à leur caractère stochastique, il s'agit de méthodes d'optimisation globale.
- Grâce à l'utilisation d'une population, il s'agit de méthodes facilement parallélisables.
- Toutes ces méthodes permettent de gérer les contraintes de manière relativement simple et efficace par un principe de pénalisation statique ou dynamique ou de stochastic ranking.
- La plupart de ces méthodes possèdent une version multi-objectif permettant de déterminer un front de Pareto.
- Ces méthodes sont également très stables par rapport aux erreurs numériques commises sur la fonction J.

Principaux inconvénients

- Le principal inconvénient de ces méthodes est leur coût de calcul important lié au grand nombre d'évaluations de la fonction J à effectuer.
- La vitesse de convergence est également très lente comparée à une méthode déterministe.
- Le choix des paramètres est très influent sur la qualité des résultats obtenus.
- Il existe très peu de résultats de convergence théorique de ces méthodes.

Un résultat de convergence pour les ES

- La plupart des résultats de convergence concernent des fonctions sphériques : $J(x) = g(||x||^2)$ avec g croissante.
- Soit par exemple une stratégie d'évolution $(1, 1)$ représentée par la suite de vecteurs aléatoires (X_n) et ayant une mutation gaussienne de variance $\sigma||X_n||$. Sous certaines hypothèses techniques et pour une fonction sphérique, X_n converge presque surement vers le minimum de J et :

$$\lim_{n\to+\infty} \frac{1}{n} \ln\left(\frac{||X_n||}{||X_0||}\right) = c < 0$$

Modèles approchés

- Pour rendre plus performants les algorithmes évolutionnaires, l'incorporation d'un modèle approché, affiné au cours des itérations, permet d'améliorer grandement leur efficacité.
- De manière générale, l'objectif consiste à construire une fonction approchée $\tilde{J}$ (*surrogate ou metamodèle*) de la fonction exacte J à partir d'un certain nombre de points $(X_i, J(X_i))_{1 \leq i \leq N}$ où la fonction exacte est supposée connue.
- Références : *Giannakoglou* (2001), *Jin* (2005), etc...

Modèles approchés : méthode RBF

- Une méthode possible, appelée méthode RBF, est construite comme une combinaison linéaire de fonctions radiales centrées en chacun des points X_i.
- L'approximation de la fonction coût en un point $X \in \mathbb{R}^n$ s'écrit alors :

$$\tilde{J}(X) = \sum_{i=1}^{N} w_i h(||X - X_i||)$$

où h désigne une fonction $r \mapsto h(r)$ dite fonction de base radiale.

Modèles approchés : méthode RBF

- Les poids $(w_i)_{1 \leq i \leq N}$ sont calculés par résolution de l'équation matricielle $Aw = z$ traduisant l'exactitude du réseau sur les points $(X_i)_{1 \leq i \leq N}$, où la matrice $A \in \mathcal{M}_N(\mathbb{R})$ a pour terme général $a_{i,j} = h(||X_i - X_j||)$ et le second membre a pour terme général $z_i = J(X_i)$.
- On a donc ici :

$$\tilde{J}(X) = R^T A^{-1} z$$

où R est le vecteur colonne de terme général $h(||X - X_i||)$.
- Pour des fonctions h bien choisies, on peut montrer que la matrice A est toujours inversible voire définie positive.

Modèles approchés : méthode RBF

- Une fonction continue f définie sur $\mathbb{R}_+^*$ est dite (définie) positive si pour toute famille de points distincts $X_1, ..., X_N$ de $\mathbb{R}^n$, la forme quadratique

$$q(c_1, ..., c_N) = \sum \sum_{1 \leq i,j \leq N} c_i c_j f(||X_i - X_j||)$$

  est (définie) positive.
- **Théorème** (Schoenberg) : Une fonction f est totalement monotone sur $\mathbb{R}_+^*$ si et seulement si la fonction $r \to f(r^2)$ est positive.
- Ainsi, les fonctions $r \mapsto e^{-r^2}$ et $r \mapsto (1 + r^2)^{-\alpha}$ avec $\alpha > 0$ peuvent être utilisées comme fonctions de base dans les réseaux RBF.

Modèles approchés : méthode RBF

- **Théorème** (Micchelli) : soit h une fonction dérivable sur $\mathbb{R}_+$, strictement positive sur $\mathbb{R}_+^*$. Si la première dérivée de h est totalement monotone et non constante sur $\mathbb{R}_+^*$, alors pour toute famille de points distincts $X_1, ..., X_N$ de $\mathbb{R}^n$:

$$(-1)^{N-1} det\left(\left[h(||X_i - X_j||^2)\right]\right) > 0$$

- Ainsi, les fonctions $r \to (c^2 + r^2)^\alpha$ avec $c \in \mathbb{R}$ et $0 < \alpha < 1$ peuvent être utilisées comme fonction de base radiale dans les réseaux RBF.

Modèles approchés : méthode RBF

- Afin de déterminer les paramètres optimaux des fonctions de base du modèle RBF, une méthode de type 'leave-one out' peut être utilisée.
- Cette méthode consiste à entrainer le réseau sur tous les points sauf un et à tester l'erreur commise sur ce point. En répétant ce procédé sur tous les points, on aboutit à une erreur globale qu'il s'agit de rendre minimale.

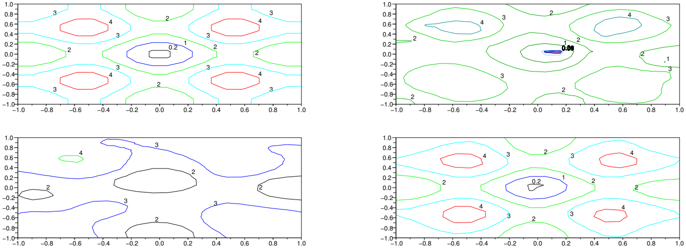
Modèles approchés : méthode RBF

- Dans le cas où le nombre de points N est trés grand, la matrice A peut être mal conditionnée. Afin d'éviter ce problème, deux choix sont possibles.
- Soit un procédé de régularisation de Tychonov est ajouté permettant de réduire le conditionnement de A. Dans ce cas, la méthode RBF cesse d'être une méthode d'interpolation.
- Soit le nombre de points N est réduit à m en ne considérant que les plus proches points du point X à calculer. Dans ce cas, le réseau construit devient local.

Modèles approchés : méthode RBF

- Exemple sur la fonction de Rastrigin :

$$Rast(x_1, ..., x_n) = \sum_{i=1}^{n} \left(x_i^2 - \cos(2\pi x_i)\right) + n$$



- La figure ci-dessus compare les contours de la fonction de Rastrigin et trois modèles approchés, construit à partir d'un réseau RBF avec 40 ou 200 points d'exemples.

Incorporation dans un algorithme évolutionnaire

- A partir d'un algorithme évolutionnaire de type Algorithme Génétique, un algorithme plus performant peut être construit.
- Il consiste à introduire le modèle approché de type RBF en l'améliorant au fil des itérations à l'aide de nouveaux points d'exemples.
- Ces nouveaux points correspondent aux éléments les plus performants,au sens de la fonction approchée.
- Le nombre de nouveaux points d'exemples décroit au cours des générations pour ne concerner plus que quelques éléments lors de la dernière génération de l'algorithme.

Exemple d'utilisation de l'algorithme AGA

- Sur la fonction de Rastrigin, le gain comparé à un AG classique se situe entre un facteur 2 et 10 (pour $n = 6$ et $n = 20$ représentés ci dessous) :