Méthode de recherche directe avec descente probabiliste

Clément W. Royer
ENSEEIHT-IRIT, Toulouse, France

Co-auteurs: S. Gratton, L. N. Vicente, Z. Zhang

2 octobre 2014
Journée des doctorants APO

Sommaire

Problème

On cherche à résoudre le problème d'optimisation sans contraintes suivant :

$$\min_{x \in \mathbb{R}^n} f(x).$$

Hypothèses sur f

- f lisse ($\mathcal{C}^1$), minorée ;
- ∇f lipschitzien.

Problème

On cherche à résoudre le problème d'optimisation sans contraintes suivant :

$$\min_{x \in \mathbb{R}^n} f(x).$$

Hypothèses sur f

- f lisse ($\mathcal{C}^1$), minorée ;
- ∇f lipschitzien.

Minimisation utilisant le gradient

Partant de $x \in \mathbb{R}^n$, un déplacement dans la direction $-\nabla f(x)$ peut conduire à une décroissance de f !

- Méthode de la plus grande pente ;
- Méthodes de type gradient.

La classe de méthodes

Optimisation sans dérivées

- Le gradient est supposé indisponible (Ex : code de simulation) ;
- Deux grandes catégories de méthodes :
  - basées sur des modèles (Régions de confiance, etc) ;
  - Recherche directe.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

La classe de méthodes

Optimisation sans dérivées

- Le gradient est supposé indisponible (Ex : code de simulation) ;
- Deux grandes catégories de méthodes :
  - basées sur des modèles (Régions de confiance, etc) ;
  - Recherche directe.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

Nous nous concentrerons sur les méthodes de recherche directe directionnelle.

Optimization by Direct Search : New Perspectives on Some Classical and Modern Methods
T.G. Kolda, R.M. Lewis, V. Torczon (2003).

Sommaire

Algorithme basique de recherche directe

1. **Initialisation :** Choisir $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, \ldots$
   - Choisir un ensemble D_k de m vecteurs unitaires.
   - Si $\exists\ d_k \in D_k$ tel que

     $$f(x_k + \alpha_k\, d_k) < f(x_k) - \alpha_k^2,$$

     alors l'itération k est *réussie* ; poser $x_{k+1} := x_k + \alpha_k\, d_k$ et $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Sinon l'itération est *non réussie* ; poser $x_{k+1} := x_k$ et $\alpha_{k+1} := \theta\, \alpha_k$.

Algorithme basique de recherche directe

1. **Initialisation :** Choisir $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, \ldots$
   - Choisir un ensemble D_k de m vecteurs unitaires.
   - Si $\exists\ d_k \in D_k$ tel que

     $$f(x_k + \alpha_k\, d_k) < f(x_k) - \alpha_k^2,$$

     alors l'itération k est *réussie* ; poser $x_{k+1} := x_k + \alpha_k\, d_k$ et $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Sinon l'itération est *non réussie* ; poser $x_{k+1} := x_k$ et $\alpha_{k+1} := \theta\, \alpha_k$.

Procédé de sondage en recherche directe déterministe

Un bon algorithme doit choisir des ensembles de sondage D_k qui peuvent assurer la convergence :

- Comment quantifier la qualité d'un ensemble ?
- Comment utiliser cette qualité pour prouver la convergence ?

Procédé de sondage en recherche directe déterministe

Un bon algorithme doit choisir des ensembles de sondage D_k qui peuvent assurer la convergence :

- Comment quantifier la qualité d'un ensemble ?
- Comment utiliser cette qualité pour prouver la convergence ?

Une mesure de qualité d'un ensemble

Soit D un ensemble de vecteurs unitaires. Alors

$$\mathrm{cm}(D) = \min_{\|v\|=1} \max_{d \in D} d^T v$$

s'appelle la mesure cosinus de D.

Procédé de sondage en recherche directe déterministe

Un bon algorithme doit choisir des ensembles de sondage D_k qui peuvent assurer la convergence :

- Comment quantifier la qualité d'un ensemble ?
- Comment utiliser cette qualité pour prouver la convergence ?

Une mesure de qualité d'un ensemble

Soit D un ensemble de vecteurs unitaires. Alors

$$\mathrm{cm}(D) = \min_{\|v\|=1} \max_{d \in D} d^T v$$

s'appelle la mesure cosinus de D.

Assumption

Il existe $\kappa > 0$ *tel que* $\forall k,\ \mathrm{cm}(D_k) \geq \kappa$.
Tout vecteur (ex : $-\nabla f(x_k)$*) est alors proche d'un élément de* D_k.

Ensembles positivement générateurs

Certains ensembles de vecteurs sont connus pour être de bonne qualité.

Ensemble Générateur Positif (EGP)

D est in EGP si il engendre $\mathbb{R}^n$ par combinaisons linéaires positives.

- D est un EGP **ssi** $\text{cm}(D) > 0$;
- un EGP contient au moins $n+1$ vecteurs.

Ensembles positivement générateurs

Certains ensembles de vecteurs sont connus pour être de bonne qualité.

Ensemble Générateur Positif (EGP)

D est in EGP si il engendre $\mathbb{R}^n$ par combinaisons linéaires positives.

- D est un EGP **ssi** $\text{cm}(D) > 0$;
- un EGP contient au moins $n+1$ vecteurs.

Exemple

$D_{\oplus} = [I \ -I]$ est un EGP tel que

$$\text{cm}\,(D_{\oplus}) \ = \ \frac{1}{\sqrt{n}}.$$

Convergence de la recherche directe déterministe

Lemma

$$\lim_{k \to \infty} \alpha_k = 0.$$

Proposition

Si la k-ème itération est réussie et que $\mathrm{cm}(D_k) \geq \kappa > 0$, *on a*

$$\mathcal{O}(\alpha_k) \geq \|\nabla f(x_k)\|.$$

Résultat de convergence

Si $\forall k,\ \mathrm{cm}(D_k) \geq \kappa$, alors

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Convergence globale et complexité au pire cas

On a prouvé que $\liminf \|\nabla f(x_k)\| \to 0$, mais pas à quel prix.

Complexité au pire cas (méthodes sans dérivées)

Estimer le nombre maximum d'appels à f nécessaires pour obtenir

$$\inf_{0 \leq l \leq k} \|\nabla f(x_l)\| \leq \epsilon.$$

Worst-case complexity of direct search
L. N. Vicente (2013)

Complexité (déterministe) au pire cas

Théorème (Vicente - 2013)

Soit N_ϵ le nombre d'appels à f requis pour que la norme du gradient soit plus petite que $\epsilon \in (0,1)$; alors

$$N_\epsilon \leq \mathcal{O}\left(m\left(\kappa\,\epsilon\right)^{-2}\right).$$

avec $m \geq n+1$.

Corollary

Avec $D_k = D_\oplus$, on obtient $\kappa = 1/\sqrt{n}, m = 2n$, et la borne devient

$$N_\epsilon \leq \mathcal{O}\left(n^2\,\epsilon^{-2}\right).$$

Sommaire

De l'aléatoire en recherche directe

Idée de base (Gratton, Vicente - 2013)

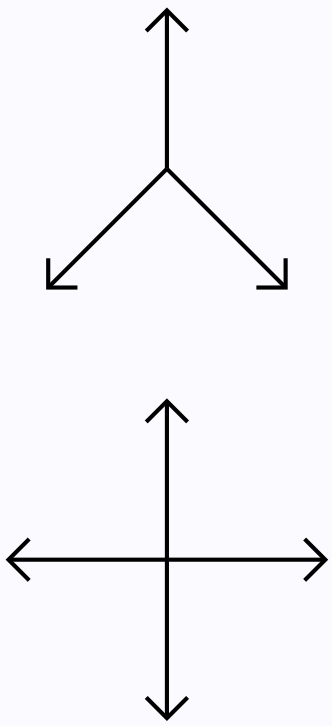
Générer les ensembles de sondage de façon aléatoire et indépendante, possiblement avec moins de $n+1$ éléments !

De l'aléatoire en recherche directe

Idée de base (Gratton, Vicente - 2013)

Générer les ensembles de sondage de façon aléatoire et indépendante, possiblement avec moins de $n+1$ éléments !

Des EPG...

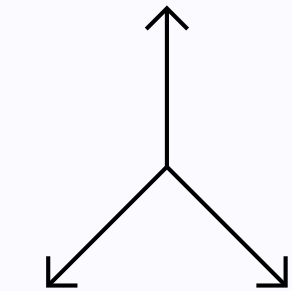


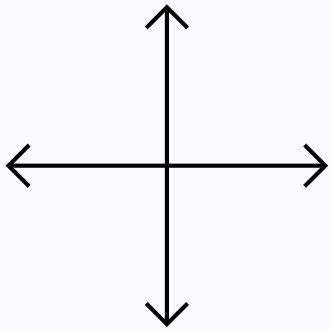
De l'aléatoire en recherche directe

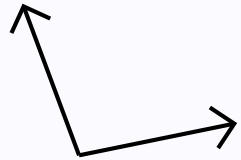
Idée de base (Gratton, Vicente - 2013)

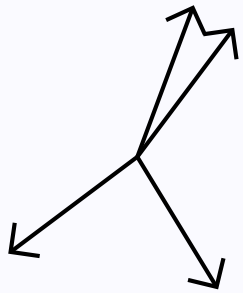
Générer les ensembles de sondage de façon aléatoire et indépendante, possiblement avec moins de $n+1$ éléments !

Des EPG...









...aux ensembles aléatoires

Motivations expérimentales

- Quelques résultats avec $n = 40$ et $\epsilon = 10^{-3}$:

Problème	$[I\ -I]$	$[Q\ -Q]$	$2\,n$	$n+1$	$n/2$	2	1
arglina	3.42	8.44	10.30	6.01	3.21	1.00	–
arglinb	20.50	10.35	7.38	2.81	2.35	1.00	2.04
broydn3d	4.33	6.55	6.54	3.59	2.04	1.00	–
dqrtic	7.16	9.37	9.10	4.56	2.77	1.00	–
engval1	10.53	20.89	11.90	6.48	3.55	1.00	2.08
freuroth	56.00	6.33	1.00	1.67	1.33	1.00	4.00
integreq	16.04	16.29	12.44	6.76	3.52	1.00	–
nondquar	6.90	30.23	7.56	4.23	2.76	1.00	–
sinquad	–	–	1.31	1.00	1.60	1.23	–
vardim	1.00	3.80	1.80	2.40	2.30	1.80	4.30

Table : Moyenne relative du nombre d'appels à f pour différents choix d'ensembles

Algorithme de recherche directe probabiliste

Notations : réalisations et variables aléatoires

- Ensembles de sondage : $D_k \to \mathfrak{D}_k$;
- Itérés : $x_k \to X_k$;
- Longueurs de pas : $\alpha_k \to \mathcal{A}_k$.

Algorithme de recherche directe probabiliste

Notations : réalisations et variables aléatoires

- Ensembles de sondage : $D_k \to \mathfrak{D}_k$;
- Itérés : $x_k \to X_k$;
- Longueurs de pas : $\alpha_k \to \mathcal{A}_k$.

1. **Initialisation :** Choisir $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, \ldots,$
   - Choisir un ensemble $\mathfrak{D}_k$ de m vecteurs unitaires aléatoires et indépendants.
   - Si $\exists\, \mathfrak{d}_k \in \mathfrak{D}_k$ tel que

$$f(X_k + \mathcal{A}_k \mathfrak{d}_k) < f(X_k) - \mathcal{A}_k^2,$$

     alors l'itération k est réussie ; poser $X_{k+1} := X_k + \mathcal{A}_k \mathfrak{d}_k$ et $\mathcal{A}_{k+1} := \gamma \mathcal{A}_k$.
   - Sinon, l'itération est non réussie ; poser $X_{k+1} := X_k$ et $\mathcal{A}_{k+1} := \theta \mathcal{A}_k$.

Sommaire

Analyse de l'algorithme probabiliste

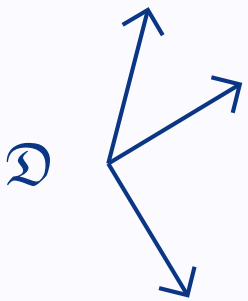
Ce qui nous intéresse

- Peut-on prouver qu'il converge quel que soit le point initial ?
  Convergence Globale
- Peut-on borner les appels à f effectués pour atteindre une tolérance ϵ ?
  Complexité Au Pire Cas

Toute la difficulté consiste à trouver (et utiliser) de bons outils probabilistes.

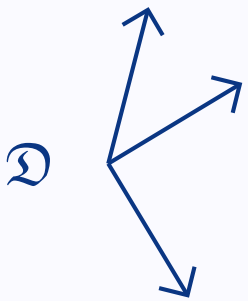
Nouvelle mesure de qualité

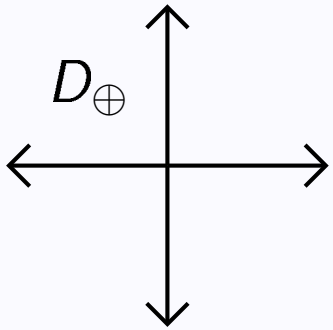
- $\mathfrak{D}$ n'est pas un EPG...



Nouvelle mesure de qualité

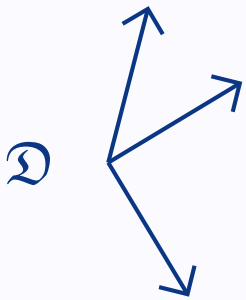
- $\mathfrak{D}$ n'est pas un EPG... ...$D_{\oplus}$ en est un...

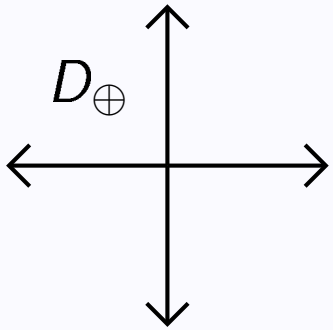


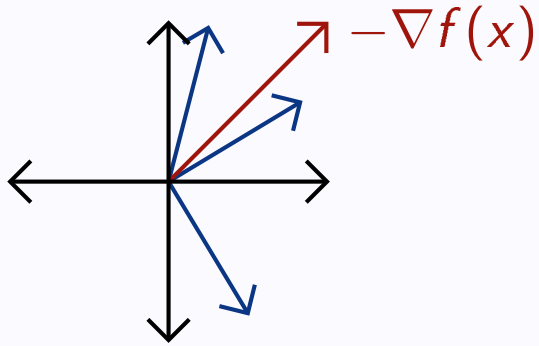


Nouvelle mesure de qualité

$\mathfrak{D}$ n'est pas un EPG... ...$D_\oplus$ en est un... ...mais mieux vaut choisir $\mathfrak{D}$!







Nouvelle mesure de qualité

Hypothèse dans le cas déterministe

- On suppose que :

$$\mathrm{cm}(D_k) = \min_{\|v\|=1} \max_{d \in D_k} d^T v > \kappa.$$

- Mais en réalité, on se sert juste de :

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) \overset{d}{=} \max_{d \in D_k} d^T \frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|} > \kappa.$$

Dans le cas aléatoire, la seconde propriété peut être vraie en probabilité pour un ensemble sans que celui-ci soit un EGP.

Descente-(p, κ)

- On veut étudier $\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) > \kappa\right)$,
  mais X_k dépend de $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$.

Convergence on trust-region methods based on probabilistic models
A.S. Bandeira, K. Scheinberg, L.N. Vicente. (2014)

Descente-(p, κ)

- On veut étudier $\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) > \kappa\right)$,
  mais X_k dépend de $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$.

Convergence on trust-region methods based on probabilistic models
A.S. Bandeira, K. Scheinberg, L.N. Vicente. (2014)

Propriété de descente probabiliste

Une suite d'ensembles aléatoires $\{\mathfrak{D}_k\}$ est appelée suite de descente-(p, κ) si :

$$\forall k,\ \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) > \kappa \,\middle|\, \mathfrak{S}_{k-1}^{\mathfrak{D}}\right) \geq p,$$

avec $\mathfrak{S}_{k-1}^{\mathfrak{D}} = \sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1})$.

Convergence avec descente probabiliste

Lemma

Pour toute réalisation $\{\alpha_k\}$ *de* $\{\mathcal{A}_k\}$ *:*

$$\lim_{k\to\infty} \alpha_k = 0.$$

Théorème de convergence

Si $\{\mathfrak{D}_k\}$ est de descente-(p,κ) où $p \geq \ln(\theta)\,\ln(\theta/\gamma)^{-1}$, alors

$$\mathbb{P}\left(\liminf_{k\to\infty} \|\nabla f(X_k)\| = 0 \right) = 1.$$

Eléments de preuve

Deux arguments essentiels :

Lemma

Si k est non réussie, alors

$$\{\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) > \kappa\} \subset \{\mathcal{O}\,(\mathcal{A}_k) \geq \|\nabla f(X_k)\|\}\,.$$

Lemma

Soit $Z_k = \mathbf{1}\,(\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) > \kappa)$; alors

$$S_k = \sum_{i=0}^{k-1} \left[\left(1 - \frac{\ln \gamma}{\ln \theta}\right) \cdot Z_i - 1\right].$$

est une sous-martingale et $\mathbb{P}\,(\limsup S_k = \infty) = 1$.

Sur la probabilité p

Pour assurer la convergence, on doit supposer :

$$p \geq p_0 = \frac{\ln(\theta)}{\ln(\theta/\gamma)},$$

ce qui permet de borner $m = |\mathfrak{D}_k|$.

Exemple : distribution uniforme sur la sphère unité

Dans ce cas, $\{\mathfrak{D}_k\}_k$ est de descente-(p_0, κ) si

$$m \geq \ln\left(1 - \frac{\ln\theta}{\ln(\theta/\gamma)}\right) \ln\left(1 - \frac{1}{2} B_{1-\kappa^2}\left(\frac{n-1}{2}, \frac{1}{2}\right)\right)^{-1}.$$

où $B_x(a, b)$ est la fonction Bêta incomplète.

Complexité avec descente probabiliste

Intuition

Soit $G_k \overset{n}{=} \nabla f(X_k)$ et $Z_k \overset{n}{=} \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) > \kappa\right)$.

Complexité avec descente probabiliste

Intuition

Soit $G_k \overset{n}{=} \nabla f(X_k)$ et $Z_k \overset{n}{=} \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) > \kappa\right)$.

- Si $Z_k = 1$ et k est non réussie, alors $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...

Complexité avec descente probabiliste

Intuition

Soit $G_k \overset{n}{=} \nabla f(X_k)$ et $Z_k \overset{n}{=} \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) > \kappa\right)$.

- Si $Z_k = 1$ et k est non réussie, alors $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...donc si $\inf_{0 \leq l \leq k} \|G_l\|$ n'a pas trop décru, $\sum_{l=0}^{k} Z_l$ ne doit pas être trop grand.

Complexité avec descente probabiliste

Intuition

Soit $G_k \overset{n}{=} \nabla f(X_k)$ et $Z_k \overset{n}{=} \mathbf{1}\left(\text{cm}(\mathfrak{D}_k, -G_k) > \kappa\right)$.

- Si $Z_k = 1$ et k est non réussie, alors $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...donc si $\inf_{0 \leq l \leq k} \|G_l\|$ n'a pas trop décru, $\sum_{l=0}^{k} Z_l$ ne doit pas être trop grand.

Une borne utile

Pour toute réalisation, on a

$$\sum_{l=0}^{k} z_l \;\leq\; \mathcal{O}\left(\frac{1}{\kappa^2 \, \|\tilde{g}_k\|^2}\right) + p_0 \, k,$$

avec $\|\tilde{g}_k\| = \inf_{0 \leq l \leq k} \|g_l\|$.

Complexité avec descente probabiliste (2)

Complexité probabiliste

Soit $\{\mathfrak{D}_k\}$ de descente-(p, κ), $\epsilon \in (0, 1)$ et N_ϵ le nombre d'appels à f nécessaires pour que $\|\tilde{G}_k\| \leq \epsilon$. Alors

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}(m\,(\kappa\epsilon)^{-2})\right) \geq 1 - \exp\left(-\mathcal{O}\left(\epsilon^{-2}\right)\right).$$

Complexité avec descente probabiliste (2)

Complexité probabiliste

Soit $\{\mathfrak{D}_k\}$ de descente-(p, κ), $\epsilon \in (0,1)$ et N_ϵ le nombre d'appels à f nécessaires pour que $\|\tilde{G}_k\| \leq \epsilon$. Alors

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}(m\,(\kappa\epsilon)^{-2})\right) \geq 1 - \exp\left(-\mathcal{O}\left(\epsilon^{-2}\right)\right).$$

- En prenant $\mathfrak{D}_k = D_\oplus$, on a $\kappa = 1/\sqrt{n}$, $m = 2\,n$ et $p = 1$; on retrouve

$$\mathcal{O}(n^2\,\epsilon^{-2}).$$

- Avec une distribution uniforme, la borne devient $\mathcal{O}(m\,n\,\epsilon^{-2})$, et on peut avoir $m \ll n + 1$!

Résumé

Que ressort-il de cette étude ?

- Un nouvel algorithme convergent qui se passe des EGP ;

Résumé

Que ressort-il de cette étude ?

- Un nouvel algorithme convergent qui se passe des EGP ;
- Une nouvelle démonstration de complexité probabiliste, qui s'adapte à d'autres méthodes (ex : Régions de Confiance) ;

Résumé

Que ressort-il de cette étude ?

- Un nouvel algorithme convergent qui se passe des EGP ;
- Une nouvelle démonstration de complexité probabiliste, qui s'adapte à d'autres méthodes (ex : Régions de Confiance) ;
- Une amélioration sur le plan numérique.

Et pour la suite...

L'article

Direct Search based on Probabilistic Descent.
S. Gratton, C. W. Royer, L. N. Vicente, Z. Zhang.
Soumis et consultable sur www.optimization-online.org.

Quels sont les développements prévus ?

- Extension aux problèmes non lisses et avec contraintes ;
- Preuve probabiliste de résultats d'ordre 2.

Merci de votre attention !