ANNALES SCIENTIFIQUES
DE L'UNIVERSITÉ DE CLERMONT-FERRAND 2
Série Mathématiques

J. B. HIRIART-URRUTY

Optimisation stochastique : méthodes de descente

Annales scientifiques de l'Université de Clermont-Ferrand 2, tome 58, série *Mathématiques*, n° 12 (1976), p. 110-131

<http://www.numdam.org/item?id=ASCFM_1976__58_12_110_0>

© Université de Clermont-Ferrand 2, 1976, tous droits réservés.

L'accès aux archives de la revue « Annales scientifiques de l'Université de Clermont-Ferrand 2 » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

OPTIMISATION STOCHASTIQUE : méthodes de descente

J.B. HIRIART-URRUTY , Université de Clermont-Ferrand

Résumé

On développe une version stochastique des méthodes de descente en optimisation. La théorie des multiapplications permet dans le cas déterministe de démontrer la convergence des méthodes de type descente ([1] [2] CH IV). Pour les méthodes stochastiques, on se servira de ce même formalisme pour obtenir des théorèmes de convergence.

Introduction

Le problème est celui de la minimisation d'une fonction f sur un ensemble C ; on ne sait pas calculer f parce que par exemple sa forme analytique est inconnue. Pour tout $x \in C$, on n'a seulement accès à des *observations bruitées* soit de f (x) soit de f' (x). Les algorithmes construits génèrent des variables aléatoires X_n et il s'agit de prouver que des sous-suites de $\{X_n(\omega)\}$ convergent vers des points stationnaires de f.

Dans ce qui suit, le domaine des contraintes C est fixé et connu, c'est-à-dire que si C se met sous la forme $\{x \,/\, f_i(x) \leqslant 0 \quad i = 1, \ldots, m\}$, les fonctions f_i sont supposées certaines.

Pour la construction de l'algorithme de descente, on opère de la façon suivante : supposons que l'on ait calculé les n premiers itérés : $X_1, \ldots, X_n$; il s'agit à présent de choisir une nouvelle direction $\tilde{d}_n$ mais il faut que ce soit une *"direction de descente stochastique"* dans un sens à définir. Le nouvel itéré X_{n+1} doit être pris dans la direction $\tilde{d}_n$ tout en restant dans C et il faut que le gain acquis en passant de X_n à X_{n+1} soit suffisamment grand pour assurer une convergence. En particulier il n'est pas question de calculer exactement le point qui minimise f dans cette direction compte tenu du fait que les observations de f sont bruitées.

Le problème général

C étant un *convexe fermé* de $\mathbb{R}^n$, soit $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ une fonction finie et continue relativement à C en tout point de C, *inf-compacte sur C* (c'est-à-dire que $\{x \in C \ / \ f(x) \leq \lambda\}$ est compact pour tout λ). Considérons le problème d'optimisation :

($\mathcal{P}$) *Trouver* x^* *tel que :* $x^* \in C$ *et* $f(x^*) = \mathrm{Inf}\ \{f(x) \ / \ x \in C\}$

Posons $f^* = \mathrm{Inf}\ \{f(x) \ / \ x \in C\}$

On suppose que : $\forall x, y \in C$ $f'(x ; y - x)$ *existe* ($f'(x ; d)$ désignant la dérivée directionnelle de f en x dans la direction d).

L'inf-compacité de f sur C entraîne que l'ensemble M des solutions de ($\mathcal{P}$) n'est pas vide. Une condition *nécessaire* pour que $\overline{x} \in M$ est que $\overline{x}$ soit un *point stationnaire*, c'est-à-dire :

$$\overline{x} \in C \quad \forall y \in C \qquad f'(\overline{x} ; y - \overline{x}) \geq 0$$

Cette condition est en particulier suffisante lorsque *f est pseudo-convexe sur C*. Désignons par F l'ensemble des points stationnaires.

Procédure déterministe (cf. A. Auslender)

Soit D une multiapplication de C dans $\mathbb{R}^n$, Gr (D) désigne son graphe dans $C \times \mathbb{R}^n$, considérons α : Gr (D) $\to \overline{\mathbb{R}}$ définie par :

$$\forall x \in C,\ \forall d \in D(x) \quad \alpha(x, d) = \mathrm{Sup} \quad \{\alpha \geq 0 \mid x + \alpha d \in C\}.$$

$\{T_n\}_{n \in \mathbb{N}}$ est une famille de multiapplications de Gr (D) dans $\mathbb{R}$. A tout couple $(D, \{T_n\}_{n \in \mathbb{N}})$, on fait correspondre la *méthode d'optimisation suivante*.

Partant d'un élément $x_1 \in C$, on construit la suite $\{x_n\}$ de C de la façon suivante :

$$\text{(ALG)} \quad \begin{cases} x_n \rightsquigarrow \text{direction } d_n \in D(x_n) \\ (x_n, d_n) \rightsquigarrow x_{n+1} = x_n + \alpha_n^* d_n \text{ avec } \alpha_n^* = \mathrm{Min}(\alpha_n, \alpha(x_n, d_n)) \end{cases}$$

$$\text{et } \alpha_n \in T_n(x_n, d_n)$$

On voit que D désigne la multiapplication des *directions de déplacement* possibles tandis que T_n permet d'obtenir la multiapplication des *points de déplacement* sur ces directions.

Suivant qu'il s'agit d'un problème d'optimisation sans contraintes ou avec contraintes, l'ensemble des directions D (x_n) possibles à partir de x_n est connu ou est l'ensemble des solutions d'un programme (linéaire par exemple). Les types de choix de T_n pour construire X_{n+1} conduiront à des méthodes et à des théorèmes différents.

Procédure stochastique

Dans le cas stochastique on n'aura que des évaluations stochastiques des directions ou bien à des solutions de sous-problèmes où les données (fonctions, gradients) sont bruitées. D'autre part, pour les points de déplacement sur la direction choisie, il est préférable de prendre des pas simples ou faciles à calculer comme par exemple :

T_n (x, d) ne dépend pas de (x, ω), ne dépend que de n :

$T_n\ (x, d) = [\lambda_{n,1}, \lambda_{n,2}]$ où $\{\lambda_{n,1}\}$ et $\{\lambda_{n,2}\}$ sont deux suites de réels positifs.

Dans la première famille d'algorithmes étudiée (ALG 1), *on considérera ce choix des T_n.*

I - ALG 1

I.1. Dans ce paragraphe, on suppose que C est *borné*, f est *G-dérivable* sur C (c'est-à-dire dérivable au sens de Gâteaux), *f' continue* sur C relativement à C.

THEOREME 1

Sous les hypothèses suivantes :

a) hypothèses relatives à la multiapplication D

($\mathcal{H}_1$) *D est une multiapplication bornée (i.e. transforme les bornés en bornés)*

($\mathcal{H}_2$) *D est une multiapplication fermée en $\bar{x}$ pour tout $\bar{x} \notin F$*

(i.e., si $x_k \to \bar{x}$ $\quad y_k \in D\ (x_k)$ $\quad y_k \to \bar{y}$, alors $\bar{y} \in D\ (\bar{x})$)

($\mathcal{H}_3$) $\forall x \in C \quad \forall d \in D\ (x) < f'\ (x), d > \ \leq 0$

$\forall x \notin F\ , \forall d \in D\ (x) < f'\ (x), d > \ < 0.$

b) hypothèses sur le choix de la direction $\tilde{d}_n$

Soient $X_1, X_2, \ldots, X_n$ *les n premiers itérés stochastiques ;* $\tilde{d}_1, \ldots, \tilde{d}_{n-1}$ *les n-1 directions utilisées. Appelons* $\mathcal{B}^n_{n-1}$ *la tribu engendrée par* $\{(X_i)_{i \in <1,n>}, (\tilde{d}_i)_{i \in <1,n-1>}\}$ *; le problème est de choisir une nouvelle direction* $\tilde{d}_n$.

$(\mathcal{D}_1)$ $\tilde{d}_n$ *est p.s. bornée et* $E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n > \leqslant \pi_n$ *où* π_n *est une variable aléatoire positive telle que* $\sum_{n=1}^{\infty} \lambda_{n,2}\, E(\pi_n) < +\infty$

$(\mathcal{D}_2)$ $\exists\, d_n \in D(X_n)$, $\mathcal{B}^n_{n-1}$*-mesurable,* $\exists\, \varphi_1 : \mathbb{R}^{-*} \to \mathbb{R}^{-*}$ *telle que :*

$$\forall \varepsilon > 0, \text{ pour } n \geqslant N(\varepsilon) < f'(X_n(\omega), d_n(\omega) > \leqslant -\varepsilon \Rightarrow (E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n >)(\omega) \leqslant \varphi_1(-\varepsilon)$$

$(\mathcal{D}_3)$ *En désignant par* $\alpha(X_n, \tilde{d}_n) = \mathrm{Sup}\ \{\alpha \geqslant 0 \mid X_n + \alpha_n \tilde{d}_n \in C\}$

$$\alpha(X_n, d_n) \geqslant \bar{\alpha} > 0$$

c) hypothèses sur les pas de descente :

$$T_n(X_n, \tilde{d}_n) = [\lambda_{n,1}, \lambda_{n,2}] \text{ avec } \lambda_{n,1} > 0 \quad \sum_{n=1}^{\infty} \lambda_{n,1} = +\infty$$

$$\lambda_{n,2} > 0 \quad \lambda_{n,2} \xrightarrow[n\to\infty]{} 0$$

Dans ces conditions, presque sûrement, il existe une sous-suite de $\{X_n(\omega)\}$ *qui converge vers un point de F.*

<u>Remarques sur les différentes hypothèses</u>

a) Les hypothèses relatives à D sont faites, comme dans le cas déterministe, pour assurer que :

si $x_n \to \bar{x}$, $d_n \to \bar{d}$ et $< f'(x_n), d_n > \to 0$, alors $\bar{x} \in F$.

b) Les hypothèses sur $\tilde{d}_n$ sont faites pour exprimer que :

En moyenne conditionnelle, $\tilde{d}_n$ est une direction de descente, moyennant un terme résiduel π_n dû aux effets du bruit dans la recherche de $\tilde{d}_n$.

Le fait que $\tilde{d}_n$ soit une *"direction de descente stochastique"* est précisé par l'hypothèse $(\mathcal{D}_2)$ qui relie $\tilde{d}_n$ à un élément d_n de $D(X_n)$.

c) Le choix du pas de descente a l'avantage d'être simple et de ne pas nécessiter de calculs. Les hypothèses expriment que le pas doit être suffisamment grand pour avoir une amélioration sensible dans la direction donnée et qu'il doit tendre vers 0 pour obtenir une convergence. Pour que, à partir d'un certain rang ce soit effectivement un pas $\lambda_n \in T_n\,(X_n,\ \tilde{d}_n)$ qui soit utilisé, nous sommes amenés à faire l'hypothèse $(\mathcal{D}_3)$: pour les choix de $\tilde{d}_n$ que nous explorerons, cette hypothèse sera facile à vérifier. Remarquons que si $\tilde{d}_n\,(\omega) = 0$ pour $\omega \in \Omega_o$, alors $X_{n+1}\,(\omega) = X_n\,(\omega)$ pour $\omega \in \Omega_o$.

Démonstration

$$X_{n+1} = X_n + \alpha_n^* \tilde{d}_n \text{ avec } \alpha_n^* = \text{Min}\,(\lambda_n,\ \alpha\,(X_n,\ \tilde{d}_n)) \text{ et } \lambda_n \in [\lambda_{n,1},\ \lambda_{n,2}]$$

f étant G-dérivable de G-dérivée f', on a le développement suivant :

$$(1)\quad f\,(X_{n+1}) = f\,(X_n) + \alpha_n^* \left[< f'\,(X_n),\ \tilde{d}_n > + < f'\,(\overline{X}_n) - f'\,(X_n),\ \tilde{d}_n >\right]$$

ou $\overline{X}_n$ est un élément du segment $[X_n,\ X_n + \alpha_n^* \tilde{d}_n]$

Appelons δ_n l'expression $< f'\,(\overline{X}_n) - f'\,(X_n),\ \tilde{d}_n >$.

$||\overline{X}_n - X_n|| = \alpha_n^*\ ||\tilde{d}_n||$. $\tilde{d}_n$ étant borné p.s., $||\overline{X}_n - X_n|| \xrightarrow[n\to\infty]{} 0$ p.s. ; de même, puisque f' est continue, $f'\,(\overline{X}_n) - f'\,(X_n)$ est borné p.s. et converge vers 0 p.s. Il s'en suit que :

δ_n est borné p.s. et $\delta_n \xrightarrow[n\to\infty]{} 0$ p.s. D'où $E\,(\delta_n) \xrightarrow[n\to\infty]{} 0$

Dans l'égalité (1), prenons l'espérance conditionnelle par rapport à la tribu $\mathcal{B}^n_{n-1}$:

$$E^{\mathcal{B}^n_{n-1}}\ f\,(X_{n+1}) - f\,(X_n) = \alpha_n^* \left[E^{\mathcal{B}^n_{n-1}} < f'\,(X_n),\ \tilde{d}_n > + E^{\mathcal{B}^n_{n-1}}\ \delta_n\right]$$

D'après l'hypothèse $(\mathcal{D}_1)$, nous avons l'inégalité :

$$(2)\quad E^{\mathcal{B}^n_{n-1}}\ f\,(X_{n+1}) - f\,(X_n) \leqslant \alpha_n^*\ E^{\mathcal{B}^n_{n-1}}\ \pi_n + \alpha_n^*\ E^{\mathcal{B}^n_{n-1}}\ \delta_n .$$

De plus, $\forall \varepsilon > 0$ nous avons :

$$(3)\quad E^{\mathcal{B}^n_{n-1}}\ f\,(X_{n+1}) - f(X_n) \leqslant \alpha_n^*\ \varphi_1(-\varepsilon)\ I_{\{E^{\mathcal{B}^n_{n-1}} < f'(X_n),\tilde{d}_n > \leqslant\ \varphi_1(-\varepsilon)\}} + \alpha_n^*\ E^{\mathcal{B}^n_{n-1}}\ \pi_n + \alpha_n^*\ E^{\mathcal{B}^n_{n-1}}\ \delta_n .$$

D'après l'hypothèse $(\mathcal{D}_2)$, il existe $d_n \in D(X_n)$ tel que, pour $n \geqslant N(\varepsilon)$

$$I_{\{<f'(X_n), d_n> \leqslant -\varepsilon\}} \leqslant I_{\{E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n> \leqslant \varphi_1(-\varepsilon)\}}$$

Il découle ainsi de la relation (3) :

$$(4) \quad E^{\mathcal{B}^n_{n-1}} f(X_{n+1}) - f(X_n) \leqslant \alpha^*_n \ \varphi_1(-\varepsilon) \ I_{\{<f'(X_n), d_n> \leqslant -\varepsilon\}} + \alpha^*_n \ E^{\mathcal{B}^n_{n-1}} \pi_n + \alpha^*_n \ E^{\mathcal{B}^n_{n-1}} \delta_n$$

$\alpha^*_n = \mathrm{Min}\,(\lambda_n, \ \alpha\,(X_n, \tilde{d}_n))$ avec $\lambda_n \in [\lambda_{n,1}, \lambda_{n,2}]$.

Comme $\lambda_{n,2} \xrightarrow[n\to\infty]{} 0$ et que $\alpha\,(X_n, \tilde{d}_n) \geqslant \overline{\alpha} > 0$, à partir d'un certain rang n_o ; $\alpha^*_n = \lambda_n$

En sommant à partir d'un certain rang $n_1 = \mathrm{Max}\,(n_o, N(\varepsilon))$

$$E(f(X_n)) - E(f(X_{n_1})) \leqslant \sum_{m=n_1}^{n} \lambda_m \varphi_1(-\varepsilon) \ P(<f'(X_m), d_m> \leqslant -\varepsilon) + \lambda_m E(\delta_m) + \lambda_m E(\pi_m)$$

Le premier membre de cette inégalité étant borné uniformément en n, on en déduit que :

$$\sum_{n=n_1}^{\infty} -\lambda_m \left[\varphi_1(-\varepsilon) \ P(<f'(X_m), dm> \leqslant -\varepsilon) + E(\delta_m)\right] < +\infty$$

Comme la série de terme général λ_m est divergente et que $E(\delta_m) \xrightarrow[m\to\infty]{} 0$, il s'en suit que :

$$\forall \varepsilon > 0 \qquad \liminf_{n\to\infty} P(<f'(X_n), d_n> \leqslant -\varepsilon) = 0.$$

D'après ce qui précède, pour tout entier k, il existe n_k tel que :

$$P(<f'(X_{n_k}), d_{n_k}> \leqslant -\frac{1}{k}) < \frac{1}{2^k}$$

$$\Longrightarrow \sum_{k=1}^{\infty} P(<f'(X_{n_k}), d_{n_k}) > \leqslant -\frac{1}{k}) < +\infty$$

$$\Longrightarrow P(\limsup_{k\to\infty} (<f'(X_{n_k}), d_{n_k}> \leqslant -\frac{1}{k})) = 1$$ d'après le lemme de Borel-Cantelli

(5) D'où $<f'(X_{n_k}), d_{n_k}> \xrightarrow[k\to\infty]{} 0$ p.s.

C étant borné, D (C) est un borné (puisque D est une multiapplication bornée) $\{X_{n_k}(\omega)\}_{k=1}^{\infty}$ est une suite bornée, de même $\{d_{n_k}(\omega)\}_{k=1}^{\infty}$; d'après la relation (5). p.s. il existe une sous suite $(X_{n_l}(\omega), d_{n_l}(\omega))$ telle que :

$$X_{n_l}(\omega) \xrightarrow[l\to\infty]{} \overline{X}(\omega)$$

$$d_{n_l}(\omega) \xrightarrow[l\to\infty]{} \overline{d}(\omega) \quad \text{et} \quad < f'(\overline{X}(\omega), \overline{d}(\omega) > = 0.$$

Si $\overline{X}(\omega) \notin F$, $\overline{d}(\omega) \in D(\overline{X}(\omega))$ (puisque D est une multiapplication fermée en $\overline{x} \notin F$).

On aurait alors :

$$< f'(\overline{X}(\omega)), \overline{d}(\omega) > \;<\; 0 \quad \text{d'où la contradiction.}$$

Ainsi, avec une probabilité 1, il existe une sous-suite de $\{X_n(\omega)\}$ qui converge vers un point de F.

Concernant la convergence de la suite de variables aléatoires $\{f(X_n)\}$, nous avons le résultat suivant :

THEOREME 1 bis

Outre les hypothèses du théorème précédent, supposons que :

f' est Höldérienne d'ordre p > 0 sur C, c'est-à-dire que :

$$\forall x, y \in C \qquad ||f'(x) - f'(y)|| \leq k\, ||x - y||^p$$

et que :

$$\sum_{n=1}^{\infty} (\lambda_{n,2})^{p+1} < +\infty$$

Dans ces conditions $f(X_n)$ converge p.s. vers une variable aléatoire f_∞ à valeurs dans $f(F)$. Si de plus $F = M$, $f_\infty = f^*$ p.s. et p.s. toute valeur d'adhérence de $\{X_n(\omega)\}$ est une solution du problème ($\mathcal{P}$).

Démonstration

D'après l'inégalité (2) de la démonstration du théorème 1, nous avons :

$$E^{\mathcal{B}^n_{n-1}} f(X_{n+1}) \leq f(X_n) + \alpha_n^* E^{\mathcal{B}^n_{n-1}} \pi_n + \alpha_n^* E^{\mathcal{B}^n_{n-1}} \delta_n .$$

On a : $\delta_n = < f'(\overline{X}_n) - f'(X_n), \tilde{d}_n >$.

Par suite $|\delta_n| \leq k\, (\alpha_n^*)^p \, ||\tilde{d}_n||^{p+1} \leq k\, (\alpha_n^*)^p \, \tilde{d}^{p+1}$

D'où, pour $n \geq n_o$

$$E^{\mathcal{B}^n_{n-1}} f(X_{n+1}) \leq f(X_n) + \lambda_{n,2}\, E^{\mathcal{B}^n_{n-1}} \pi_n + k\,(\lambda_{n,2})^{p+1}\, \tilde{d}^{p+1}$$

$f(X_n)$ est $\mathcal{B}^n_{n-1}$ -mesurable, bornée inférieurement. $\{f(X_n)\}_{n \in \mathbb{N}}$ est "presque une surmartingale" et comme $\sum_{n=1}^{\infty} \lambda_{n,2}\, E(\pi_n) < \infty$ et que $\sum_{n=1}^{\infty} (\lambda_{n,2})^{p+1}\, \tilde{d}^{p+1} < +\infty$, il est classique alors que $f(X_n)$ converge vers une variable aléatoire f_∞ avec une probabilité 1. D'après le théorème 1, il est facile de voir que $f_\infty(\omega) \in f(F)$ p.s. Supposons maintenant que $F = M$. On a vu que p.s. il existe une sous-suite $\{X_{n_k}(\omega)\}$ de $\{X_n(\omega)\}$ qui converge vers un point $\overline{X}(\omega)$ de F, donc de M.

f étant continue, $f(X_{n_k}(\omega)) \xrightarrow{k\to\infty} f(\overline{X}(\omega)) = f^*$.

Par suite :

p.s. $f(X_n(\omega)) \xrightarrow[n\to\infty]{} f_\infty(\omega) = f^*$.

Soit $\tilde{X}(\omega)$ une valeur d'adhérence de la suite $\{X_n(\omega)\}$. Appelons $\{X_{n_l}(\omega)\}$ une sous-suite de $\{X_n(\omega)\}$ qui converge vers $\tilde{X}(\omega)$. Supposons que $\tilde{X}(\omega) \notin M$; alors : $f(\tilde{X}(\omega)) > f^*$. Mais f étant continue, on aurait :

$$\lim_{l\to\infty} f(X_{n_l}(\omega)) = f(\tilde{X}(\omega)) > f^*$$

d'où la contradiction avec le fait que $f(X_n(\omega)) \xrightarrow[n\to\infty]{} f^*$.

<u>Remarques</u>

a) on peut prendre par exemple pour intervalle $[\lambda_{n,1}, \lambda_{n,2}]$ l'intervalle suivant : $\left[\frac{1}{n}, \frac{1}{n^\theta}\right]$ avec $\theta \in]0, 1]$. Dans le cadre du théorème 1.bis, si f' est Lipchitzienne sur C, les hypothèses sur le pas sont satisfaites avec $\theta \in \left]\frac{1}{2}, 1\right]$.

b) A chaque itération n, on peut prendre un pas aléatoire Λ_n pourvu que Λ_n soit $\mathcal{B}^n_{n-1}$-mesurable et que $\Lambda_n(\omega) \in [\lambda_{n,1}, \lambda_{n,2}]$ p.s. ; les théorèmes énoncés restent encore vrais.

I.2. Les résultats précédents sont établis dans le cas où C est borné. Considérons maintenant le cas où C est un convexe *fermé* (en particulier pour $C = \mathbb{R}^n$, on envisage le problème de l'optimisation sans contraintes).

Soit donc C un convexe *fermé*, f G-dérivable sur C, *inf-compact* sur C. La direction $\tilde{d}_n$ que l'on déterminera ne sera pas nécessairement p.s. bornée car C n'est pas borné ; on fera donc une hypothèse différente sur $\tilde{d}_n$.

THEOREME 2

Outre les hypothèses du théorème 1, nous supposons que :

. *f' est Höldérienne d'ordre $p > 0$ sur C*

et que

. $\sum_{n=1}^{\infty} (\lambda_{n,2})^{p+1} E(||\tilde{d}_n||^{p+1}) < +\infty$

Alors :

. *$f(X_n)$ converge p.s. vers une variable aléatoire f_∞.*

. *$\exists$ une sous-suite $d_{n_k} \in D(X_{n_k})$ telle que $< f(X_{n_k}), d_{n_k} > \xrightarrow[k\to\infty]{} 0$ p.s.*

. *p.s. $\exists$ une sous-suite de $\{X_n(\omega)\}$ qui converge vers un point de F.*

Démonstration

Les démonstrations des deux premiers points sont analogues à celles faites pour C borné. Pour conclure, il ne reste plus qu'à montrer que p.s. la suite $\{X_n(\omega)\}$ est bornée. Soit $\Omega_\infty = \{\omega \mid \limsup_{n\to\infty} ||X_n(\omega)|| = +\infty\}$ et supposons que $P(\Omega_\infty) > 0$ (hypothèse dont on veut montrer l'absurdité).

Pour $\omega \in \Omega_\infty$ $\exists$ une sous suite de $\{X_n(\omega)\}$ telle que $||X_{n_k}(\omega)|| \xrightarrow[k\to\infty]{} +\infty$. Cela entraîne, d'après l'inf-compacité de f, que $f(X_{n_k}(\omega)) \xrightarrow[k\to\infty]{} +\infty$, ce qui est absurde puisque $f(X_{n_k}(\omega)) \xrightarrow[k\to\infty]{} f_\infty(\omega)$.

Remarques

a) les hypothèses du théorème précédent sont vérifiées en particulier dans le contexte suivant :

f Lipschitzienne sur C , $||\tilde{d}_n||_{L^2} < +\infty$ et $\sum_{n=1}^{\infty} (\lambda_{n,2})^2 < +\infty$.

b) Naturellement, comme dans le cas du théorème 1.bis, si F = M, $f_\infty = f^*$ p.s. et p.s. toute sous-suite convergente de $\{X_n(\omega)\}$ converge vers un point solution du problème ($\mathcal{P}$).

I.3. Nous allons à présent calculer effectivement une direction de descente et vérifier ainsi les hypothèses des théorèmes de convergence : examinons d'abord le cas où C est *borné* (théorèmes 1 et 1.bis).

A. <u>Méthode de FRANK et WOLFE</u>

C est un convexe compact non vide ; la multiapplication D des directions de descente est alors définie par :

$$\forall x \in C \qquad D(x) = \{\bar{x} - x \;/\; \bar{x} \in C < f'(x), \bar{x} > = \operatorname*{Min}_{z \in C} < f'(x), z >\}$$

Les hypothèses déterministes relatives à D sont vérifiées. D est une multiapplication bornée et :

$$\forall x \in C \qquad \forall d \in D(x) \qquad < f'(x), d > \; \leq 0.$$

Si $x \notin F$, $\exists\, \tilde{x} \in C$ tel que $< f'(x), \tilde{x} - x > \; < 0 \Longrightarrow \forall d \in D(x) <f'(x),d>< 0.$

Désignons par δ_C^* la fonction d'appui de C, c'est-à-dire : $\delta_C^*(x^*) = \operatorname*{Sup}_{x \in C} < x, x^* >$

$$d_n \in D(x_n) \iff d_n = \bar{x}_n - x_n \text{ avec } < f'(x_n), \bar{x}_n > = -\delta_C^*(-f'(x_n))$$
$$< -f'(x_n), \bar{x}_n > = \delta_C^*(-f'(x_n)).$$

Ainsi $D(x_n) = \partial \delta_C^*(-f'(x_n)) - x_n$ (*) et si on suppose que $x_n \to \bar{x}$, on a :

(**) $$\underline{\lim}\, D(x_n) \subset \partial \delta_C^*(-f'(\bar{x})) - \bar{x} = D(\bar{x}).$$

La multiapplication D est donc fermée.

(*) $\partial \delta_C^*(u)$ est le sous-différentiel de δ_C^* en u .
$\partial \delta_C^*(u) = \{x \in C \;/\; < x, u > = \operatorname*{Sup}_{z \in C} < z, u >\}$

(**) $\underline{\lim}\, A_n = \{x \;/\; x = \lim_{n \to \infty} x_n \quad \text{ou } x_n \in A_n\}$

En fait, la résolution du programme pour calculer une direction de descente se fera avec une donnée bruitée Y_n au lieu de $f'(X_n)$; la v.a. Y_n est calculée *à partir d'évaluations stochastiques du gradient de f en* $x = X_n$.

Ecrivons $Y_n = f'(X_n) + \Xi_n$ avec $E^{\mathcal{B}^n_{n-1}} \Xi_n = 0$.

Ainsi : $\tilde{d}_n = \tilde{X}_n - X_n$ avec $\langle Y_n, \tilde{X}_n \rangle = - \delta^*_c(- Y_n)$

Il est clair que $\tilde{d}_n$ est borné. De plus, si nous considérons $\alpha(X_n, \tilde{d}_n)$, nous avons :

$$\alpha(X_n(\omega), \tilde{d}_n(\omega)) = 1 \text{ si } \tilde{d}_n(\omega) \neq 0$$
$$\alpha(X_n(\omega), \tilde{d}_n(\omega)) = +\infty \text{ si } \tilde{d}_n(\omega) = 0 \Longrightarrow \alpha(X_n, \tilde{d}_n) \geqslant 1.$$

$D(X_n) = \partial \delta^*_C(- f'(X_n)) - X_n$ et $\omega \rightarrow D(X_n(\omega))$ est une multiapplication $\mathcal{B}^n_{n-1}$-mesurable à valeurs convexes compactes. On peut donc choisir $d_n \in D(X_n)$ telle que d_n soit $\mathcal{B}^n_{n-1}$-mesurable (d'après le théorème de sélection mesurable de KURATOWSKI-RYLL-NARDZEWSKI)

$$d_n = \overline{X}_n - X_n \quad \text{avec} \quad \langle f'(X_n), \overline{X}_n \rangle = - \delta^*_C(- f'(X_n))$$
$$\tilde{d}_n = \tilde{X}_n - X_n \quad \text{avec} \quad \langle Y_n, \tilde{X}_n \rangle = - \delta^*_C(- Y_n)$$

Comme $E^{\mathcal{B}^n_{n-1}} Y_n = - f'(X_n)$ et que δ^*_C est convexe, l'inégalité de Jensen sur les espérances conditionnelles entraîne que :

$$E^{\mathcal{B}^n_{n-1}} \langle Y_n, \tilde{d}_n \rangle \leqslant \langle f'(X_n), d_n \rangle$$
$$\Longrightarrow E^{\mathcal{B}^n_{n-1}} \langle f'(X_n), \tilde{d}_n \rangle \leqslant \langle f'(X_n), d_n \rangle - E^{\mathcal{B}^n_{n-1}} \langle \Xi_n, \tilde{d}_n \rangle$$

Comme $E^{\mathcal{B}^n_{n-1}} \Xi_n = 0$, on a aussi :

$$(1) \quad E^{\mathcal{B}^n_{n-1}} \langle f'(X_n), \tilde{d}_n \rangle \leqslant \langle f'(X_n), d_n \rangle - E^{\mathcal{B}^n_{n-1}} \langle \Xi_n, \tilde{d}_n - d_n \rangle$$

D'autre part δ^*_C est Lipschitzienne de coefficient de Lipschitz $k = \text{Sup} \{ ||x|| \ / \ x \in C \}$. Par conséquent :

(2) $\quad E^{\mathcal{B}^n_{n-1}} < f'(X_n),\ \tilde{d}_n > \leqslant\ < f'(X_n),\ d_n > + k\, E^{\mathcal{B}^n_{n-1}}\ ||\Xi_n||^2$

Soient $\{Y^i_n\}_{i=1,\dots,p_n}$ une suite de v.a. d'observation du gradient de f en $x = X_n$; on suppose que les Y^i_n *sont conditionnellement indépendantes* (le conditionnement étant relatif à la tribu $\mathcal{B}^n_{n-1}$)

Plus précisément $Y^i_n = f'(X_n) + \xi^i_n$ avec $E^{\mathcal{B}^n_{n-1}}\ \xi^i_n = 0$ et $E^{\mathcal{B}^n_{n-1}}\ ||\xi^i_n||^2 \leqslant \sigma^2$

Posons $Y_n = \dfrac{1}{p_n} \sum\limits_{i=1}^{p_n} Y^i_n$, soit $\Xi_n = \dfrac{1}{p_n} \sum\limits_{i=1}^{p_n} \xi^i_n$.

On a : $E^{\mathcal{B}^n_{n-1}}\ \Xi_n = 0$ et $E^{\mathcal{B}^n_{n-1}}\ ||\Xi_n||^2 \leqslant \dfrac{\sigma^2}{p_n}$

Pour vérifier les hypothèses $(\mathcal{D}_1)$ et $(\mathcal{D}_2)$ du théorème 1, il suffit que :

$$\sum_{n=1}^{\infty} \frac{\lambda_{n,2}}{p_n} < +\infty \quad \text{et que} \quad p_n \xrightarrow[n\to\infty]{} +\infty .$$

En effet, dans ces conditions, $\pi_n = k\, E^{\mathcal{B}^n_{n-1}}\ ||\Xi_n||^2$ et l'inégalité (2) implique que :

Pour $n \geqslant N_1(\varepsilon)$, sur $< f'(X_n),\ d_n > \leqslant -\varepsilon$, $E^{\mathcal{B}^n_{n-1}} < f'(X_n),\ \tilde{d}_n > \leqslant \varphi_1(-\varepsilon)$

C.Q.F.D.

B. Méthode du GRADIENT PROJETE

La méthode étudiée ici n'est pas la méthode du gradient projeté telle qu'elle a été définie par ROSEN ; il s'agit d'une adaptation à l'optimisation avec contraintes de la méthode classique du gradient.

Définition de la multiapplication D :

$$\forall x \in C \qquad D(x) = \{\bar{x}_\beta - x\} \quad \text{ou} \quad \bar{x}_\beta = P_C(x - \beta f'(x))$$

On suppose que $\beta \in [\beta_1, \beta_2]$ avec $\beta_2 \geqslant \beta_1 > 0$.

Vérifions les hypothèses relatives à D :

D'après les propriétés de l'opérateur de projection P_C, on a :

$$\forall y \in C \qquad < y - \bar{x}_\beta\ ,\ x - \beta f'(x) - \bar{x}_\beta > \ \leqslant 0$$

Puisque $x \in C$, on a alors :

$$< x - \bar{x}_\beta \, , x - \beta f'(x) - \bar{x}_\beta > \leq 0$$

$$- \beta < f'(x), x - \bar{x}_\beta > + ||x - \bar{x}_\beta||^2 \leq 0$$

$$\Longrightarrow \quad < f'(x), \bar{x}_\beta - x > \leq - \frac{1}{\beta} ||x - \bar{x}_\beta||^2 \leq - \frac{1}{\beta_2} ||x - \bar{x}_\beta||^2$$

Par suite :

$$\forall x \in C \, , \forall d \in D(x) \quad < f'(x), d > \leq 0.$$

D'autre part : $x \in F \iff x = \bar{x}_\beta \quad \forall \beta \in [\beta_1, \beta_2]$

$$\Longrightarrow \quad \forall x \in F^c \quad \forall d \in D(x) \quad < f'(x), d > < 0$$

Montrons que la multiapplication D est fermée en $\bar{x} \in F^c$.

Soit $x_n \to \bar{x}$ et $d_n \in D(x_n)$ tel que $d_n \to \bar{d}$. Posons $d_n = \bar{x}_{\beta_n} - x_n$ où

$\bar{x}_{\beta_n} = P_C(x_n - \beta_n f'(x_n))$ avec $\beta_n \in [\beta_1, \beta_2]$. Il existe une sous-suite

$\{\beta_{n_k}\}$ de $\{\beta_n\}$ telle que $\beta_{n_k} \xrightarrow[k \to \infty]{} \bar{\beta}$ et $\bar{\beta} \in [\beta_1, \beta_2]$;

$$d_{n_k} = P_C(x_{n_k} - \beta_{n_k} f'(x_{n_k})) - x_{n_k} \xrightarrow[k \to \infty]{} P_C(\bar{x} - \bar{\beta} f'(\bar{x})) - \bar{x}$$

Ainsi $\bar{d} \in D(\bar{x})$.

<u>Vérifions maintenant les hypothèses relatives à $\tilde{d}_n$</u>

Y_n étant construit à partir d'observations du gradient de f en $x = X_n$, on construit $\tilde{d}_n$ de la façon suivante :

$$\tilde{d}_n = \tilde{X}_n - X_n \quad \text{avec} \quad \tilde{X}_n = P_C(X_n - \beta_n Y_n) \text{ ou } \beta_n \in [\beta_1, \beta_2]$$

Soit $d_n \in D(X_n)$ définie par $d_n = \bar{X}_n - X_n$ où $\bar{X}_n = P_C(X_n - \beta_n f'(X_n))$

Il est clair que $\tilde{d}_n$ est bornée, $\mathcal{B}^n_{n-1}$-mesurable et que $\alpha(X_n(\omega), \tilde{d}_n(\omega)) \geq 1$.

Par ailleurs :

$$< f'(X_n), d_n > = < f'(X_n), \bar{X}_n - X_n > \leq - \frac{1}{\beta_n} ||\bar{X}_n - X_n||^2 \qquad (1)$$

De même :

$$< Y_n, \tilde{d}_n > \leq - \frac{1}{\beta_n} ||\tilde{X}_n - X_n||^2$$

En décomposant $Y_n = f'(X_n) + \Xi_n$, on a :

$$< f'(X_n), \tilde{d}_n > \leq - \frac{1}{\beta_n} ||\tilde{X}_n - X_n||^2 + < \Xi_n, X_n - \bar{X}_n > + < \Xi_n, \bar{X}_n - \tilde{X}_n >$$

Comme $\|\tilde{X}_n - \overline{X}_n\| \leq \|\Xi_n\|$, il s'en suit que :

$$< f'(X_n), \tilde{d}_n > \leq -\frac{1}{\beta_2} \|\tilde{X}_n - X_n\|^2 + < \Xi_n, X_n - \overline{X}_n > + \beta_2 \|\Xi_n\|^2 \quad (2)$$

Nous faisons les mêmes hypothèses que dans la méthode de Frank et Wolfe en ce qui concerne Ξ_n, à savoir :

$$\Xi_n = \frac{1}{p_n} \sum_{i=1}^{p_n} \xi_n^i \qquad \sum_{n=1}^{\infty} \frac{\lambda_{n,2}}{p_n} < +\infty \qquad p_n \underset{n\to\infty}{\to} +\infty$$

Notons $K = \text{Sup} \{ \|f'(x)\| \,/\, x \in C \}$

Sur $\{< f'(X_n), d_n > \leq -\varepsilon\}$, on a $\|d_n\| \geq \frac{\varepsilon}{K}$

La relation $\|\overline{X}_n - X_n\|^2 \leq 2 (\|\tilde{X}_n - \overline{X}_n\|^2 + \|X_n - \tilde{X}_n\|^2)$ entraîne que :

sur $\{< f'(X_n), d_n > \leq -\varepsilon\}$

$$- \|\tilde{X}_n - X_n\|^2 \leq - \frac{\|\overline{X}_n - X_n\|^2}{2} + \|\tilde{X}_n - \overline{X}_n\|^2 \leq - \frac{\varepsilon^2}{2K^2} + \|\Xi_n\|^2$$

La relation (2) entraîne alors que :

sur $\{< f'(X_n), d_n > \leq -\varepsilon\}$, $E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n > \leq - \frac{\varepsilon^2}{2\beta_2 K^2} + (\frac{1}{\beta_2} + \beta_2) \frac{\sigma^2}{p_n}$

$$\leq \varphi_1(-\varepsilon) \text{ pour } n \geq N_1(\varepsilon)$$

D'autre part, d'après l'inégalité (2)

$$E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n > \leq \beta_2 \, E^{\mathcal{B}^n_{n-1}} \|\Xi_n\|^2$$

$$\pi_n = \beta_2 \, E^{\mathcal{B}^n_{n-1}} \|\Xi_n\|^2 \quad \text{et} \quad \sum_{n=1}^{\infty} \lambda_{n,2} \, E(\pi_n) < +\infty$$

Remarques

1°) Dans les méthodes de calcul de $\tilde{d}_n$ précédentes, à la $n^{\text{ième}}$ itération, on considère p_n observations. Par exemple, soit :

$$[\lambda_{n,1}, \lambda_{n,2}] = \{\frac{1}{n}\} \qquad p_n = [(\log n)^{1+p}] \quad \text{où} \quad p > 0 \text{ et } [x] \text{ désigne}$$

la partie entière de x. Dans ce cas, $\lambda_{n,2}$ et p_n vérifient les hypothèses requises.

2°) On peut utiliser les méthodes précédentes dans le cas où on ne peut estimer directement le gradient de f en $x = X_n$ mais où l'on estime une "approximation" notée $Df\,(X_n)$ de $f'\,(X_n)$. Soit $||Df\,(X_n) - f'\,(X_n)|| \leq h_n$; il suffit alors que h_n vérifie les conditions suivantes :

$$\lim_{n\to\infty} h_n = 0 \quad \text{et} \quad \sum_{n=1}^{\infty} \lambda_{n,2}\, h_n < +\infty.$$

3°) La méthode du gradient projeté est susceptible de multiples variantes. On peut prendre β_n v.a. $\mathcal{B}^n_{n-1}$-mesurable à valeurs dans $[\beta_1, \beta_2]$. D'autre part, la projection sur C peut se faire suivant une *métrique variable*.

Plus précisément :

$$\tilde{d}_n = P_C\,(X_n - \beta_n A_n Y_n) - X_n$$

où A_n est une matrice $\mathcal{B}^n_{n-1}$-mesurable prise dans $\mathcal{A}$ où $\mathcal{A}$ désigne

$$\mathcal{A} = \{A \text{ réelle symétrique définie positive } / \ m\,I \leq A \leq m\,I^{(*)}\}.$$

La projection s'effectue donc au sens de la métrique $||-||_n = <\, .,\ A_n^{-1}(.)\, >^{\frac{1}{2}}$

I.4. Nous allons donner dans ce sous-paragraphe des exemples de direction $\tilde{d}_n$ vérifiant les hypothèses du théorème 2. Considérons, par exemple, le cas où $C = \mathbb{R}^n$ (optimisation sans contraintes).

A. Exemple

Soit f convexe, inf-compact, G-dérivable de G-dérivée f' vérifiant :

(L) $\qquad \exists\, \gamma > 0 \qquad ||f'\,(x) - f'\,(y)|| \leq \gamma\ ||x - y||$

L'hypothèse de convexité fait que l'ensemble F des points stationnaires est l'ensemble M des points solutions du problème $(\mathcal{P})$

(*)

$A \leq B$ au sens de la relation d'ordre dans l'ensemble des matrices réelles symétriques, définies positives

$$A \leq B \iff <x,\ Ax> \ \leq\ <x,\ Bx> \qquad \forall x$$

$X_{n+1} = X_n + \Lambda_n \tilde{d}_n$ où Λ_n est $\mathcal{B}^n_{n-1}$-mesurable, à valeurs dans $[\lambda_{n,1}, \lambda_{n,2}]$

et $\sum_{n=1}^{\infty} \lambda_{n,1} = +\infty$ $\sum_{n=1}^{\infty} (\lambda_{n,2})^2 < +\infty$. La multiapplication D est définie par

$D(x) = \{-f'(x)\}$ et la direction $\tilde{d}_n$ est donnée par :

$$\tilde{d}_n = -Y_n = -[f'(X_n) + \xi_n] \quad \text{avec} \quad E^{\mathcal{B}^n_{n-1}} \xi_n = 0 \text{ et } E^{\mathcal{B}^n_{n-1}} ||\xi_n||^2 \leq \sigma^2$$

Les hypothèses relatives à D ont été vérifiées dans l'étude de la méthode du gradient projeté (§ I.3.B). De plus :

$$E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n > = -||f'(X_n)||^2 \leq 0$$

Dans ce cas, $\pi_n = 0$ et $\varphi_1(-\varepsilon) = -\varepsilon$ (hypothèse $(\mathcal{D}_2)$)

Soit $x^* \in M$; on a $f'(x^*) = 0$

$$(1) \quad ||X_{n+1} - x^*||^2 = ||X_n - x^*||^2 + 2 < \Lambda_n \tilde{d}_n, X_n - x^* > + ||\Lambda_n \tilde{d}_n||^2$$

$$= ||X_n - x^*||^2 - 2 < \Lambda_n f'(X_n), X_n - x^* > - 2 < \Lambda_n \xi_n, X_n - x^* >$$

$$+ ||\Lambda_n f'(X_n)||^2 + ||\Lambda_n \xi_n||^2 + 2 < \Lambda_n f'(X_n), \xi_n >$$

La convexité de f, l'hypothèse (L) sur la G-dérivée font que :

$$(2) \quad E^{\mathcal{B}^n_{n-1}} ||X_{n+1} - x^*|| \leq ||X_n - x^*||^2 (1 + \gamma \lambda^2_{n,2}) + \lambda^2_{n,2} \sigma^2$$

D'où :

$$(3) \quad E(||X_{n+1} - x^*||^2) \leq E(||X_n - x^*||^2)(1 + \gamma \lambda^2_{n,2}) + \lambda^2_{n,2} \sigma^2 .$$

On en déduit que $\{X_n\}$ est uniformément bornée dans L^2 et par conséquent que : $\sup_n ||\tilde{d}_n||_{L^2} < +\infty$. On a donc vérifié l'hypothèse relative à $\tilde{d}_n$ du théorème 2, pour p = 1, à savoir que :

$$\sum_{n=1}^{\infty} (\lambda_{n,2})^2 \, E(||\tilde{d}_n||^2) < +\infty .$$

On peut donc déduire du théorème 2 (+ remarque b) que p.s. toute sous-suite convergente de $\{X_n(\omega)\}$ converge vers un point de M.

Dans le cas présent, on peut améliorer ce résultat car il découle de l'inégalité (2) que $||X_n - x^*||$ converge p.s. Par suite si $M = \{x^*\}$, on a :

$$X_n \text{ converge p.s. vers } x^*.$$

Comme dans le cas de la méthode du gradient projeté, on peut prendre comme direction :

$\tilde{d}_n = - A_n Y_n$ où A_n, $\mathcal{B}^n_{n-1}$-mesurable, peut être un estimateur de $\left[f''(X_n)\right]^{-1}$ dans la mesure où ce dernier existe ou bien A_n est fonction des itérés et des directions précédentes.

II. ALG 2

Dans l'étude de la famille d'algorithmes ALG 1, nous avons utilisé un choix particulier de multiapplications $\{T_n\}_{n \in \mathbb{N}}$. Un autre choix fondamental est défini comme suit :

X_n étant le $n^{\text{ième}}$ itéré stochastique et $\tilde{d}_n$ la direction choisie, posons :

$$\varphi_n(\alpha) = f(X_n + \alpha \tilde{d}_n) \text{ pour } \alpha \in [0, \alpha(X_n, \tilde{d}_n)].$$

$$\forall p \in \mathbb{N} \quad T_p \equiv T \quad \text{et} \quad T(X_n, \tilde{d}_n) = \{\alpha^* \in [0, \alpha(X_n, \tilde{d}_n)] \;/\; \varphi(\alpha^*) = \inf_{\alpha \in [0, \alpha(X_n, \tilde{d}_n)]} \varphi(\alpha)\}$$

En fait, on n'a pas accès à $\alpha^* \in T(X_n, \tilde{d}_n)$ et dans le cas stochastique où l'on n'a que des évaluations aléatoires de f en $x = X_n$ ou de f' en $x = X_n$, on est amené à modifier le choix de T. Nous ferons une hypothèse analogue à celle de Kushner et Gavin ([3] , [4]) pour une méthode de direction admissible (hypothèse vérifiée, du moins dans le cas convexe pour des procédures tronquées de recherche dans une direction donnée).

$$\forall p \quad T_p \equiv T \quad et \quad T(X_n, \tilde{d}_n) = \{\alpha^* \in [0, \alpha(X_n, \tilde{d}_n)] \;/\; \alpha^* \text{ vérifiant } (\mathcal{C})\}$$

la condition $(\mathcal{C})$ explicitée dans l'énoncé du théorème de convergence.

Le nouvel itéré X_{n+1} est donc obtenu par :

$X_{n+1} = X_n + \alpha^* \tilde{d}_n$ ou $\alpha^* \in T(X_n, \tilde{d}_n)$.

α^* mesurable.

Soit donc C <u>convexe fermé</u>, f <u>G-dérivable sur C</u>, <u>inf-compact sur C</u>.

Rappelons que l'ensemble F des points stationnaires est définie par :

$$F = \{\bar{x} \in C \;/\; \langle f'(\bar{x}), y - \bar{x} \rangle \geqslant 0 \quad \forall y \in C\}.$$

THEOREME 3

Sous les hypothèses suivantes :

a) hypothèses relatives à la multiapplication D : $(\mathcal{H}_1) - (\mathcal{H}_2) - (\mathcal{H}_3)$ *(théorème 1)*

b) hypothèse sur la direction $\tilde{d}_n$:

$$\exists\, d_n \in D(X_n),\ \mathcal{B}^n_{n-1}\text{-mesurable}, \qquad \exists\, \varphi_1 : \mathbb{R}^{-*} \to \mathbb{R}^{-*} \text{ tel que :}$$

$$\forall \varepsilon > 0,\ \text{pour } n \geq N(\varepsilon) \qquad < f'(X_n(\omega)),\ d_n(\omega) > \leq -\varepsilon \Rightarrow E^{\mathcal{B}^n_{n-1}} < f'(X_n), \tilde{d}_n > \leq \varphi_1(-\varepsilon)$$

c) hypothèse sur le pas de descente :

Appelons $\mathcal{B}^n_n$ *la tribu engendrée par* $\{(X_i)_{i \in <1,n>}, (\tilde{d}_i)_{i \in <1,n>}\}$.

$X_{n+1} = X_n + \alpha^* \tilde{d}_n$ *avec* α^* *mesurable,* $\alpha^* \in [0, \alpha(X_n, \tilde{d}_n)]$ *vérifiant la condition* $(\mathcal{C})$

$$(\mathcal{C}) \quad \left\{ \begin{array}{l} E^{\mathcal{B}^n_n} f(X_{n+1}) \leq f(X_n) + s_n \quad \text{avec} \quad s_n \geq 0 \text{ et } \sum\limits_{n=1}^{\infty} E(s_n) < +\infty \\ \exists\ \varphi_2 : \mathbb{R}^{-*} \to \mathbb{R}^{-*} \ \text{tel que :} \\ E^{\mathcal{B}^n_n} f(X_{n+1}) \leq f(X_n) + \varphi_2(-\eta) + s_n \quad \text{p.s. sur} \\ \qquad \{< f'(X_n), \tilde{d}_n > \leq -\eta\} \end{array} \right.$$

Dans ces conditions, p.s. toute valeur d'adhérence de $\{X_n(\omega)\}$ *- et il en existe au moins une - est un point de F.*

Remarques sur les hypothèses

Comparativement au théorème 1 (et au théorème 1.bis), nous remarquons que les hypothèses sur $\{\tilde{d}_n\}$ sont affaiblies grâce à une "meilleure" recherche du nouvel itéré stochastique dans la direction $\tilde{d}_n$.

D'autre part, l'hypothèse $(\mathcal{D}_3)$: $\alpha(X_n, \tilde{d}_n) \geq \bar{\alpha}$, si elle est vérifiée sur certains exemples (§ I.3), ne l'est pas toujours : dans les exemples ou elle n'est pas vérifiée, c'est le choix de ALG 2 qu'il faut prendre.

La démonstration du théorème 3 est parallèle à celle de Kushner-Gavin ([3], [4] Elle est adaptée à notre formalisme et à nos choix de direction.

Démonstration

D'après la condition $(\mathcal{C})$, nous avons :

(1) $$E^{\mathcal{B}_n^n} f(X_{n+1}) \leqslant f(X_n) + s_n \quad \text{et} \quad \sum_{n=1}^{\infty} E(s_n) < +\infty .$$

Soit X_1 le premier itéré ; supposons le non aléatoire $X_1 = x_1 \in C$; f étant bornée inférieurement, il s'en suit que : $f(X_n)$ converge p.s. vers une v.a. notée f_∞ (la convergence a également lieu dans L^1).

De plus ; pour $n \geqslant M(\varepsilon)$.

(2) $$E^{\mathcal{B}_n^n} f(X_{n+1}) - f(X_n) \leqslant (\varphi_2 \circ \varphi_1)(-\varepsilon)\, I_{\{<f'(X_n), \tilde{d}_n> \leqslant \varphi_1(-\varepsilon)\}} + s_n$$

D'après l'hypothèse (b), pour $n \geqslant N(\varepsilon)$, sur $\{ < f'(X_n), d_n > \leqslant -\varepsilon\}$, on a $E^{\mathcal{B}_{n-1}^n} < f'(X_n), \tilde{d}_n > \leqslant \varphi_1(-\varepsilon)$. En conséquence ; $\exists\, \varphi_3 : \mathbb{R}^{-*} \to \mathbb{R}^{+*}$ tel que :

$$P^{\mathcal{B}_{n-1}^n} \{< f'(X_n), \tilde{d}_n > \leqslant \varphi_1(-\varepsilon)\} \geqslant \varphi_3(-\varepsilon) \text{ sur } \{< f'(X_n), d_n > \leqslant -\varepsilon\}$$

$f(X_n)$ converge dans L^1 ; $\{f(X_n)\}$ est uniformément bornée dans L^1. L'inégalité (2) implique que :

$$E^{\mathcal{B}_{n-1}^n} f(X_{n+1}) - f(X_n) \leqslant (\varphi_2 \circ \varphi_1)(-\varepsilon)\, I_{\{<f'(X_n), d_n> \leqslant -\varepsilon\}} \varphi_3(-\varepsilon) + E^{\mathcal{B}_{n-1}^n} s_n$$

D'où, en sommant à partir d'un certain rang $n_0(\varepsilon) \geqslant \text{Max}(N(\varepsilon), M(\varepsilon))$, on a :

$$\forall \varepsilon > 0 \quad \sum_{n=1}^{\infty} P\{< f'(X_n), d_n > \leqslant -\varepsilon\} < +\infty$$

$$\Longrightarrow \qquad < f'(X_n), d_n > \xrightarrow[n\to\infty]{\text{p.s.}} 0$$

On termine ensuite comme dans le cas des théorèmes 1 et 1.bis, l'inf-compacité de f et la convergence p.s. de $f(X_n)$ (d'après l'inégalité (1)) font que p.s. la suite $\{X_n(\omega)\}$ est bornée. Ainsi, p.s. toute sous-suite convergente de $\{X_n(\omega)\}$ et il en existe, converge vers un point stationnaire.

Conclusion

Dans le cadre des problèmes d'optimisation stochastique avec contraintes, nous avons donné un formalisme des méthodes de descentes stochastiques. Lorsque les contraintes sont connues (et c'est le cas dans notre approche), on peut aussi définir d'autres méthodes permettant de résoudre ces problèmes d'optimisation stochastique ; par exemple : l'analogue stochastique de la méthode d'accumulation des contraintes.

Dans le cas où les contraintes sont *stochastiques*, c'est-à-dire lorsque les fonctions définissant les contraintes sont bruitées, les méthodes précédentes sont inapplicables. Pour résoudre de tels problèmes, des méthodes du type pénalisation et du type dual sont en cours d'étude.

Références

[1] A. AUSLENDER, Résolution numérique d'inégalités variationnelles C.R.A.S. Série A, t. 276, Avril 73

[2] A. AUSLENDER, Optimisation convexe : méthodes numériques, Masson 1975 (à paraître)

[3] T. GAVIN, Stochastic approximation type methods for unconstrained and constrained optimization problems, Ph. D. Thesis, Brown University (1974)

[4] H.J. KUSHNER, Stochastic approximation algorithms for constrained optimization problems. The Annals of Statistics 1974, vol. 2 n° 4, p. 713-723