

Le PPA stochastique sur un ensemble admissible

Dans le cadre de l'optimisation convexe déterministe, on a présenté au §B.1 le *Principe du Problème Auxiliaire* (PPA), qui consiste à remplacer le problème :

$$\min_{u \in U^{\text{ad}}} J(u) , \quad (3.1)$$

par une *suite* de problèmes auxiliaires, le k -ème problème auxiliaire s'écrivant :

$$\min_{u \in U^{\text{ad}}} K(u) + \langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}) , u \rangle , \quad (3.2)$$

et sa solution $u^{(k+1)}$ permet de formuler le problème auxiliaire d'indice $k+1$. La fonction K est *choisie* par celui qui met en œuvre le PPA, et on l'appelle le *noyau de décomposition*. On rappelle les principales propriétés de ce principe.

- Le PPA permet de **retrouver** les algorithmes classiques d'optimisation : ainsi, avec un choix de noyau $K(u) = \|u\|^2/2$, le problème (3.2) prend la forme :

$$\min_{u \in U^{\text{ad}}} \frac{1}{2} \|u\|^2 + \langle \epsilon \nabla J(u^{(k)}) - u^{(k)} , u \rangle .$$

dont la solution s'obtient de manière analytique :

$$u^{(k+1)} = \text{proj}_{U^{\text{ad}}} (u^{(k)} - \epsilon \nabla J(u^{(k)})) ,$$

et correspond à un algorithme de gradient avec un pas fixe ϵ .

- Étant donnée une décomposition de u sous la forme $u = (u_1, \dots, u_N)$, le PPA permet de **décomposer** chaque problème auxiliaire (3.2) en N sous-problèmes indépendants, pourvu que la contrainte $u \in U^{\text{ad}}$ se mette sous la forme de N contraintes $u_i \in U_i^{\text{ad}}$ et pourvu que l'on choisisse un noyau K additif par rapport à la décomposition de u .

On va maintenant étendre ce principe au cas stochastique en le mélangeant l'idée du gradient stochastique.

3.1 Algorithme du gradient stochastique généralisé

On se donne un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et une variable aléatoire $\mathbf{W}$ définie sur Ω à valeurs dans un espace $\mathbb{W}$ muni de sa tribu $\mathcal{W}$. On se donne un espace de Hilbert $\mathbb{U}$ (dont le produit scalaire et la norme sont notés $\langle \cdot, \cdot \rangle$ et $\|\cdot\|$), une partie convexe fermée non vide U^{ad} de $\mathbb{U}$ et une fonction j définie sur $\mathbb{U} \times \mathbb{W}$ à valeurs dans $\overline{\mathbb{R}}$.

On considère le problème de minimisation (2.3) :

$$\min_{u \in U^{\text{ad}}} J(u), \quad (3.3)$$

avec $J(u) = \mathbb{E}(j(u, \mathbf{W}))$, et on remplace ce problème par la suite de problèmes auxiliaires issue de l'application du PPA :

$$\min_{u \in U^{\text{ad}}} K(u) + \langle \epsilon \nabla J(u^{(k)}) - \nabla K(u^{(k)}), u \rangle.$$

Puis, dans chaque problème auxiliaire, on remplace le gradient de la fonction J par le gradient partiel par rapport à u de la fonction j , évaluée en des tirages de la variable aléatoire $\mathbf{W}$. De plus, on remplace le « grand » coefficient ϵ par des « petits » coefficients $\epsilon^{(k)}$ (tendant vers 0 quand k tend vers l'infini). Le k -ème problème auxiliaire stochastique s'écrit alors :

$$\min_{u \in U^{\text{ad}}} K(u) + \langle \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)}) - \nabla K(u^{(k)}), u \rangle, \quad (3.4)$$

où $w^{(k+1)}$ est un tirage de la variable aléatoire $\mathbf{W}$. On en déduit l'algorithme suivant.

Algorithme 3.1. (Algorithme du gradient stochastique généralisé)

1. Choisir un point initial $u^{(0)} \in U^{\text{ad}}$, et une suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ de réels positifs.
2. À l'itération k , effectuer un tirage $w^{(k+1)}$ de la variable aléatoire $\mathbf{W}$.
3. Résoudre le problème auxiliaire (3.4), dont la solution est notée $u^{(k+1)}$.
4. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Comme dans le cas de l'algorithme 2.1, une hypothèse fondamentale pour la convergence de cet algorithme du gradient stochastique généralisé est que les tirages $(w^{(1)}, \dots, w^{(k)})$ correspondent à une réalisation d'un échantillon de taille k de la variable aléatoire $\mathbf{W}$, c'est-à-dire la réalisation d'une suite de k variables aléatoires $(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)})$ indépendantes de même loi que $\mathbf{W}$.

Remarque 3.2. Avec le choix de noyau $K(u) = \|u\|^2/2$, le problème auxiliaire (3.4) se met sous la forme :

$$\min_{u \in U^{\text{ad}}} \frac{1}{2} \|u\|^2 - \langle u^{(k)} - \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)}), u \rangle,$$

dont la solution $u^{(k+1)}$ est donnée par

$$u^{(k+1)} = \text{proj}_{U^{\text{ad}}} \left(u^{(k)} - \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)}) \right),$$

ce qui est précisément l'algorithme 2.1 du gradient stochastique.

L'algorithme 3.1 correspond à la mise en œuvre numérique de la méthode du gradient stochastique généralisé. Pour étudier les propriétés de cet algorithme, il est nécessaire de la décrire en terme de variables aléatoires. Pour cela, on considère un échantillon $\{\mathbf{W}^{(k)}\}_{k \in \mathbb{N}}$ de taille infinie de la variable aléatoire $\mathbf{W}$, et l'étape 3 de l'algorithme 3.1 prend la forme :

$$\min_{u \in U^{\text{ad}}} K(u) + \langle \epsilon^{(k)} \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) - \nabla K(\mathbf{U}^{(k)}), u \rangle. \quad (3.5)$$

Il est sous-entendu que la minimisation dans (3.5) est faite ω par ω , de telle sorte que le résultat de la minimisation dépend de ω .

On va s'intéresser à la convergence de l'algorithme du gradient stochastique généralisé. Les résultats qui suivent ont été obtenus par J. C. Culioli dans le cadre de sa thèse CULIOLI (1987), et exposés dans CULIOLI et COHEN (1990).

3.2 Théorème de convergence

Comme au § 2.2, on utilise la description de l'algorithme 3.1 en terme de variables aléatoires. On considère la multi-application définie sur Ω à valeurs dans $\mathbb{U}$ qui à chaque ω associe l'arg min obtenu par résolution du problème (3.5), et on suppose que cette multi-application admet une sélection mesurable¹ notée $\mathbf{U}^{(k+1)}$. La question de la convergence de l'algorithme 3.1 et celle du lien de cet algorithme avec le problème de départ (3.3) sont réglées par le théorème suivant.

Théorème 3.3. (Convergence du gradient stochastique généralisé)

On suppose que les hypothèses suivantes sont vérifiées.

1. U^{ad} est une partie convexe fermée non vide de l'espace de Hilbert $\mathbb{U}$.
2. La fonction $j : \mathbb{U} \times \mathbb{W} \rightarrow \mathbb{R}$ est une intégrande normale², et l'espérance de $j(u, \mathbf{W})$ existe pour tout $u \in U^{\text{ad}}$.
3. Pour tout $w \in \mathbb{W}$, la fonction $j(\cdot, w) : \mathbb{U} \rightarrow \mathbb{R}$ est propre, convexe, semi-continue inférieurement, différentiable sur un sous-ensemble ouvert contenant U^{ad} , et son gradient partiel par rapport à u est noté $\nabla_u j(u, w)$.

1. Voir (ROCKAFELLAR et WETS, 1998, Chapter 14) pour les questions de mesurabilité; une sélection mesurable d'une multi-application $\Phi : \Omega \rightrightarrows \mathbb{U}$ est une application $\varphi : \Omega \rightarrow \mathbb{U}$ qui est mesurable et telle que $\varphi(\omega) \in \Phi(\omega)$ pour tout $\omega \in \Omega$.

2. voir §3.5 pour cette notion

4. La fonction $j(\cdot, w)$ est à gradient linéairement borné (GLB), uniformément en w :

$$\exists c_1 > 0, \exists c_2 > 0, \forall w \in \mathbb{W}, \forall u \in U^{\text{ad}}, \|\nabla_u j(u, w)\| \leq c_1 \|u\| + c_2.$$

5. La fonction J est Lipschitzienne, coercive sur l'ensemble U^{ad} .

6. La fonction K est propre, fortement convexe de module b , semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant U^{ad} .

7. La suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est une σ -suite.

On a alors les conclusions suivantes.

1. Le problème (3.3) admet un ensemble de solutions $U^\sharp$ non vide.
2. Le problème (3.5) admet une solution $U^{(k+1)}$ unique.
3. La suite de variables aléatoires $\{J(U^{(k)})\}_{k \in \mathbb{N}}$ converge vers $\min_{u \in U^{\text{ad}}} J(u)$, presque sûrement.
4. La suite des solutions $\{U^{(k)}\}_{k \in \mathbb{N}}$ engendrée par l'algorithme 3.1 est bornée presque sûrement, et tout point d'accumulation d'une réalisation de cette suite appartient à l'ensemble optimal $U^\sharp$.

Si l'on fait l'hypothèse supplémentaire que la fonction J est fortement convexe, alors la suite $\{U^{(k)}\}_{k \in \mathbb{N}}$ engendrée par l'algorithme 3.1 converge presque sûrement vers l'unique solution $u^\sharp$ du problème (3.3).

Preuve. La démonstration des 2 premières conclusions du théorème découle des théorèmes généraux relatifs à l'optimisation convexe. Le fait que $U^{(k+1)}$, unique solution du problème (3.5) soit une variable aléatoire, et donc une application mesurable, provient de ce que l'on a supposé que j est une intégrande normale (voir le théorème 3.13 de l'annexe 3.5). Plus précisément, supposons que $U^{(k)}$ soit mesurable. On considère le problème (3.5) et on note :

$$\begin{aligned} \varphi^{(k)} : \omega &\mapsto \epsilon^{(k)} \nabla_u j(U^{(k)}(\omega), W^{(k+1)}(\omega)) - \nabla K(U^{(k)}(\omega)), \\ \Phi^{(k)} : (\omega, u) &\mapsto K(u) + \langle \varphi^{(k)}(\omega), u \rangle. \end{aligned}$$

La propriété de semi-continuité inférieure du noyau K fait que l'intégrande $(u, w) \mapsto K(u)$ est une intégrande autonome, et donc une intégrande normale (ROCKAFELLAR et WETS, 1998, Exemple 14.30). Comme on a supposé que j est une intégrande normale, on en déduit par (ROCKAFELLAR et WETS, 1998, Theorem 14.56) que $\nabla_u j$ et ∇K sont aussi des intégrandes normales³. Par composition avec les applications mesurables $U^{(k)}$ et $W^{(k+1)}$, on en déduit que l'application $\varphi^{(k)}$ est mesurable. L'intégrande $(\omega, u) \mapsto \langle \varphi^{(k)}(\omega), u \rangle$, étant

3. Les résultats sur les intégrandes dans ROCKAFELLAR et WETS (1998) sont donnés dans le cas où l'espace $\mathbb{U}$ est de dimension finie. On consultera HESS (1995) dans le cas où $\mathbb{U}$ est un espace de Banach séparable.

mesurable en ω et continue en u , est une intégrande de Carathéodory (ROCKAFELLAR et WETS, 1998, Exemple 14.29). En tant que somme de deux intégrandes normales, $\Phi^{(k)}$ est elle aussi une intégrande normale. Par (ROCKAFELLAR et WETS, 1998, Theorem 14.37), on en déduit que l'application $U^{(k+1)}$, unique solution du problème (3.5) est une application mesurable.

La démonstration des 2 dernières conclusions du théorème se fait en suivant le déroulement « classique » en quatre étapes d'un algorithme d'optimisation, à savoir :

1. choix d'une fonction de Lyapunov pour mesurer la distance à l'optimum,
2. majoration de sa variation d'une itération de l'algorithme sur l'autre,
3. convergence de l'algorithme, à l'aide d'un argument de type martingale,
4. analyse des limites des suites permettant de caractériser la solution.

Pour alléger les écritures, on utilisera la notation :

$$g^{(k)} = \nabla_u j(u^{(k)}, w^{(k+1)}) .$$

Le fait que $u^{(k+1)}$ soit solution du problème (3.4) est caractérisé par la condition d'optimalité suivante :

$$\forall u \in U^{\text{ad}}, \langle \nabla K(u^{(k+1)}) - \nabla K(u^{(k)}) + \epsilon^{(k)} g^{(k)}, u - u^{(k+1)} \rangle \geq 0 . \quad (3.6)$$

1. **Choix de la fonction de Lyapunov.** Soit $u^\# \in U^\#$ une solution du problème d'optimisation (3.3). On se donne une fonction de Lyapunov Λ de la forme :

$$\Lambda(u) = K(u^\#) - K(u) - \langle \nabla K(u), u^\# - u \rangle .$$

La forte convexité de K implique la majoration :

$$\frac{b}{2} \|u - u^\#\|^2 \leq \Lambda(u) , \quad (3.7)$$

qui montre que la fonction Λ est bornée inférieurement et coercive.

2. **Majoration de la variation de Λ .** On forme la différence :

$$\Delta^{(k)} = \Lambda(u^{(k+1)}) - \Lambda(u^{(k)}) ,$$

où $\{u^{(k)}\}_{k \in \mathbb{N}}$ est la suite des solutions engendrée par l'algorithme 3.1 pour une réalisation $(w^{(1)}, \dots, w^{(k)}, \dots)$ d'un échantillon de taille infinie de la variable $\mathbf{W}$.

$$\begin{aligned} \Delta^{(k)} = & \underbrace{K(u^{(k)}) - K(u^{(k+1)}) - \langle \nabla K(u^{(k)}), u^{(k)} - u^{(k+1)} \rangle}_{T_1} \\ & + \underbrace{\langle \nabla K(u^{(k)}) - \nabla K(u^{(k+1)}), u^\# - u^{(k+1)} \rangle}_{T_2} . \end{aligned}$$

– Par convexité de K , on a :

$$T_1 \leq 0 .$$

– L'écriture de la condition d'optimalité (3.6) au point $u = u^\sharp$ implique :

$$\begin{aligned} T_2 &\leq \epsilon^{(k)} \langle g^{(k)}, u^\sharp - u^{(k+1)} \rangle \\ &\leq \underbrace{\epsilon^{(k)} \langle g^{(k)}, u^\sharp - u^{(k)} \rangle}_{T_3} + \underbrace{\epsilon^{(k)} \langle g^{(k)}, u^{(k)} - u^{(k+1)} \rangle}_{T_4} . \end{aligned}$$

– Par la convexité de $j(\cdot, w^{(k+1)})$, on obtient :

$$T_3 \leq j(u^\sharp, w^{(k+1)}) - j(u^{(k)}, w^{(k+1)}) .$$

– L'écriture de la condition (3.6) au point $u = u^{(k)}$ et la propriété de forte monotonie de ∇K impliquent :

$$b \|u^{(k+1)} - u^{(k)}\|^2 \leq \epsilon^{(k)} \langle g^{(k)}, u^{(k)} - u^{(k+1)} \rangle .$$

L'inégalité de Schwartz appliquée à cette inégalité conduit à :

$$\|u^{(k+1)} - u^{(k)}\| \leq \frac{\epsilon^{(k)}}{b} \|g^{(k)}\| , \quad (3.8)$$

d'où, en appliquant une nouvelle fois l'inégalité de Schwartz au terme T_4 :

$$T_4 \leq \frac{\epsilon^{(k)}}{b} \|g^{(k)}\|^2 .$$

Une forme équivalente de la condition GLB est qu'il existe des constantes c_3 et c_4 telles que l'on ait : $\|g^{(k)}\| \leq c_3 \|u^{(k)} - u^\sharp\| + c_4$. Élevant cette inégalité au carré, utilisant $(a+b)^2 \leq 2(a^2 + b^2)$ ainsi que la relation (3.7), on obtient :

$$\exists \alpha > 0, \exists \beta > 0, \forall k \in \mathbb{N}, \|g^{(k)}\|^2 \leq \alpha \Lambda(u^{(k)}) + \beta ,$$

d'où la nouvelle majoration :

$$T_4 \leq \frac{\epsilon^{(k)}}{b} (\alpha \Lambda(u^{(k)}) + \beta) .$$

Collectant les majorations obtenues pour les termes T_1 , T_3 et T_4 , on obtient la majoration suivante sur la variation de la fonction de Lyapunov :

$$\Delta^{(k)} \leq \epsilon^{(k)} (j(u^\sharp, w^{(k+1)}) - j(u^{(k)}, w^{(k+1)})) + \frac{(\epsilon^{(k)})^2}{b} (\alpha \Lambda(u^{(k)}) + \beta) .$$

Cette inégalité, écrite sur des réalisations $u^{(k)}$ et $w^{(k+1)}$ des variables aléatoires $\mathbf{U}^{(k)}$ et $\mathbf{W}^{(k+1)}$, peut s'écrire sur les variables aléatoires elles-mêmes, ce qui donne :

$$\begin{aligned} \Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) &\leq \epsilon^{(k)} (j(u^\sharp, \mathbf{W}^{(k+1)}) - j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)})) \\ &\quad + \frac{(\epsilon^{(k)})^2}{b} (\alpha \Lambda(\mathbf{U}^{(k)}) + \beta) . \end{aligned}$$

On prend alors de part et d'autre de cette inégalité l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^{(k)}$ engendrée par les k variables aléatoires $(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)})$. Comme la variable aléatoire $\mathbf{U}^{(k)}$ est par construction mesurable par rapport à la tribu $\mathcal{F}^{(k)}$, on en déduit que $\mathbb{E}(\Lambda(\mathbf{U}^{(k)}) \mid \mathcal{F}^{(k)}) = \Lambda(\mathbf{U}^{(k)})$. Comme la variable aléatoire $\mathbf{W}^{(k+1)}$ est indépendante des $\mathbf{W}^{(l)}$ précédentes et donc de $\mathbf{U}^{(k)}$, on en déduit que $\mathbb{E}(j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) \mid \mathcal{F}^{(k)}) = J(\mathbf{U}^{(k)})$. On obtient finalement :

$$\begin{aligned} \mathbb{E}(\Lambda(\mathbf{U}^{(k+1)}) \mid \mathcal{F}^{(k)}) - \Lambda(\mathbf{U}^{(k)}) &\leq \alpha^{(k)} \Lambda(\mathbf{U}^{(k)}) + \beta^{(k)} \\ &\quad + \epsilon^{(k)} (J(u^\sharp) - J(\mathbf{U}^{(k)})) , \quad (3.9) \end{aligned}$$

où $\alpha^{(k)} = (\alpha/b)(\epsilon^{(k)})^2$ et $\beta^{(k)} = (\beta/b)(\epsilon^{(k)})^2$ sont les termes de deux séries convergentes. On rappelle que le terme $J(u^\sharp) - J(\mathbf{U}^{(k)})$ est toujours négatif ou nul (optimalité de $u^\sharp$).

3. **Analyse de convergence.** Une application directe du théorème 3.5 de Robbins-Siegmund permet de montrer que la suite de variables aléatoires $\{\Lambda(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ converge presque sûrement vers une variable aléatoire Λ^∞ bornée presque sûrement⁴, et que l'on a :

$$\sum_{k=0}^{+\infty} \epsilon^{(k)} (J(\mathbf{U}^{(k)}) - J(u^\sharp)) < +\infty , \quad \mathbb{P}\text{-p.s.} . \quad (3.10)$$

4. **Limites des suites.** Comme la suite $\{\Lambda(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ est bornée presque sûrement, on déduit de la relation (3.7) que la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ est elle aussi bornée presque sûrement. L'hypothèse de gradient linéairement borné sur j implique alors que la suite $\{\nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)})\}_{k \in \mathbb{N}}$ est elle aussi bornée presque sûrement. Enfin, on déduit de la relation (3.8) que la même conclusion est vraie pour la suite $\{\|\mathbf{U}^{(k+1)} - \mathbf{U}^{(k)}\|/\epsilon^{(k)}\}_{k \in \mathbb{N}}$. Cette dernière propriété, associée à la relation (3.10) et au fait que la fonction J soit Lipschitzienne, permet d'utiliser le lemme 3.7 et d'obtenir que la suite $\{J(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ converge presque sûrement vers $J(u^\sharp)$, valeur optimale du problème (3.3).

On note alors Ω_0 le sous-ensemble (de mesure nulle) de Ω sur lequel la suite $\{\Lambda(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ n'est pas bornée, et Ω_1 le sous-ensemble (de mesure nulle lui aussi) de Ω sur lequel la relation (3.10) n'est pas vérifiée. On a donc : $\mathbb{P}(\Omega_0 \cup \Omega_1) = 0$.

4. i.e. l'ensemble $\{\omega \in \Omega, \lim_{k \rightarrow +\infty} \Lambda(\mathbf{U}^{(k)})(\omega) = +\infty\}$ est de mesure nulle

Soit $\omega \notin \Omega_0 \cup \Omega_1$. La suite des réalisations $\{u^{(k)}\}_{k \in \mathbb{N}}$ associée à cet élément ω est bornée et chaque $u^{(k)}$ appartient à U^{ad} , partie fermée de $\mathbb{U}$. Par un argument de compacité⁵, on conclut que l'on peut extraire de la suite $\{u^{(k)}\}_{k \in \mathbb{N}}$ une sous-suite convergente $\{u^{(\Phi(k))}\}_{k \in \mathbb{N}}$ (on remarquera que cette sous-suite convergente Φ dépend de l'aléa ω et est donc aléatoire). Soit $\bar{u}$ la limite de la suite $\{u^{(\Phi(k))}\}_{k \in \mathbb{N}}$. Par la semi-continuité inférieure de la fonction J , on a :

$$J(\bar{u}) \leq \liminf_{k \rightarrow +\infty} J(u^{(\Phi(k))}) = J(u^\#).$$

On en déduit que $\bar{u} \in U^\#$.

Pour conclure, on considère le cas où la fonction J est fortement convexe de module a . Alors, le problème 3.3 a une unique solution $u^\#$, caractérisée par l'inéquation variationnelle :

$$\forall u \in U^{\text{ad}}, \quad \langle \nabla J(u^\#), u - u^\# \rangle \geq 0.$$

La condition de forte convexité sur J conduit à :

$$\begin{aligned} J(\mathbf{U}^{(k)}) - J(u^\#) &\geq \langle \nabla J(u^\#), \mathbf{U}^{(k)} - u^\# \rangle + \frac{a}{2} \|\mathbf{U}^{(k)} - u^\#\|^2 \\ &\geq \frac{a}{2} \|\mathbf{U}^{(k)} - u^\#\|^2. \end{aligned}$$

La convergence presque sûre de $J(\mathbf{U}^{(k)})$ vers $J(u^\#)$ entraîne donc la convergence presque sûre de $\|\mathbf{U}^{(k)} - u^\#\|$ vers zéro, ce qui achève la démonstration.

Remarque 3.4. Il existe de nombreuses applications dans lesquelles la fonction $j(\cdot, w)$ n'est pas convexe, alors que son espérance J l'est. On notera que, dans le théorème précédent, l'hypothèse de convexité faite sur la fonction j n'est pas nécessaire et peut être remplacée par une hypothèse de convexité sur son espérance J . Il suffit dans la preuve de garder le terme $T_3 = \langle g^{(k)}, u^\# - u^{(k)} \rangle$, d'en prendre l'espérance conditionnelle par rapport à $\mathcal{F}^{(k)}$, ce qui produit $\langle \nabla J(\mathbf{U}^{(k)}), u^\# - \mathbf{U}^{(k)} \rangle$, et d'utiliser l'hypothèse de convexité sur J pour majorer ce terme par $J(u^\#) - J(\mathbf{U}^{(k)})$.

3.3 Conclusions

On a énoncé et démontré un théorème très général de convergence pour l'algorithme issu du principe du problème auxiliaire dans le cas stochastique.

5. On rappelle qu'un sous-ensemble fermé borné d'un espace de Hilbert $\mathbb{U}$ de dimension finie est compact. Si l'espace de Hilbert est de dimension infinie, une telle propriété existe encore pourvu que l'on se place dans la topologie faible ; les propriétés de fermeture U^{ad} et de semi-continuité de J dans la topologie induite par la norme restent vraies dans la topologie faible sous les hypothèses que U^{ad} soit un ensemble convexe et que J soit une fonction convexe (voir [EKELAND et TEMAM \(1999\)](#) pour plus de détails).

Ce théorème inclut bien évidemment le cas du gradient stochastique standard (avec un noyau de décomposition de la forme $K(u) = \|u\|^2/2$) ; il inclut aussi l'algorithme de Newton stochastique (avec pour noyau de décomposition $K(u) = \langle u, Au \rangle/2$).

D'un point de vue théorique, les hypothèses sous lesquelles le théorème 3.3 a été énoncé sont « raisonnables » dans le cadre de l'optimisation convexe. On notera en particulier que l'on ne fait pas d'hypothèse de forte convexité sur la fonction de critère du problème d'optimisation. On a pris le parti dans ce chapitre de présenter les résultats dans le cadre différentiable. On consultera CULIOLI (1987) et CULIOLI et COHEN (1990) pour des résultats analogues dans le cadre plus général des fonctions sous-différentiables.

Pour ce qui concerne la décomposition, si l'utilisation du principe du problème auxiliaire ouvre la voie à la décomposition des problèmes d'optimisation stochastique de type (3.3), cette possibilité a en pratique moins d'impact que dans le cadre déterministe : en effet, la vitesse de l'algorithme du gradient stochastique généralisé est conditionnée par les pas $\epsilon^{(k)}$, et n'est donc pas très différente de la vitesse de la méthode du gradient stochastique standard, qui, comme tout algorithme de gradient, est elle-même une méthode de décomposition !

3.4 Annexe : théorème de Robbins-Siegmund et lemme

Le théorème suivant est l'un des résultats-clés de la théorie de l'approximation stochastique.

Théorème 3.5. (Robbins-Siegmund) *On considère quatre suites de variables aléatoires à valeurs réelles positives $\{\Lambda^{(k)}\}_{k \in \mathbb{N}}$, $\{\alpha^{(k)}\}_{k \in \mathbb{N}}$, $\{\beta^{(k)}\}_{k \in \mathbb{N}}$ et $\{\eta^{(k)}\}_{k \in \mathbb{N}}$, que l'on suppose toutes les quatre adaptées à une filtration $\{\mathcal{F}^{(k)}\}_{k \in \mathbb{N}}$ donnée. On suppose de plus que la relation suivante est vérifiée :*

$$\mathbb{E}(\Lambda^{(k+1)} \mid \mathcal{F}^{(k)}) \leq (1 + \alpha^{(k)})\Lambda^{(k)} + \beta^{(k)} - \eta^{(k)}, \quad \forall k \in \mathbb{N},$$

et que l'on a :

$$\sum_{k \in \mathbb{N}} \alpha^{(k)} < +\infty, \quad \sum_{k \in \mathbb{N}} \beta^{(k)} < +\infty, \quad \mathbb{P}\text{-p.s. .}$$

Alors, la suite de variables aléatoires $\{\Lambda^{(k)}\}_{k \in \mathbb{N}}$ converge presque sûrement vers Λ^∞ , variable aléatoire presque sûrement bornée⁶, et on a de plus que :

$$\sum_{k \in \mathbb{N}} \eta^{(k)} < +\infty, \quad \mathbb{P}\text{-p.s. .}$$

6. Une variable aléatoire $\mathbf{X}$ est presque sûrement bornée si elle telle que l'on ait : $\mathbb{P}(\{\omega \in \Omega, \mathbf{X}(\omega) = +\infty\}) = 0$.

Preuve. Voir (DUFLO, 1997, Theorem 1.3.12).

Une extension du théorème de Robbins-Siegmund, que l'on utilisera par la suite dans ce cours, est donnée par le corollaire suivant.

Corollaire 3.6. *On considère cinq suites de variables aléatoires à valeurs réelles positives $\{\mathbf{A}^{(k)}\}_{k \in \mathbb{N}}$, $\{\boldsymbol{\alpha}^{(k)}\}_{k \in \mathbb{N}}$, $\{\boldsymbol{\beta}^{(k)}\}_{k \in \mathbb{N}}$, $\{\boldsymbol{\gamma}^{(k)}\}_{k \in \mathbb{N}}$, et $\{\boldsymbol{\eta}^{(k)}\}_{k \in \mathbb{N}}$, que l'on suppose adaptées à une filtration $\{\mathcal{F}^{(k)}\}_{k \in \mathbb{N}}$. On suppose de plus que*

$$\mathbb{E}(\mathbf{A}^{(k+1)} \mid \mathcal{F}^{(k)}) \leq (1 + \boldsymbol{\alpha}^{(k)})\mathbf{A}^{(k)} + \boldsymbol{\beta}^{(k)} + \boldsymbol{\gamma}^{(k)}\mathbb{E}(\mathbf{A}^{(k+1)} \mid \mathcal{F}^{(k)}) - \boldsymbol{\eta}^{(k)},$$

et que l'on a :

$$\sum_{k \in \mathbb{N}} \boldsymbol{\alpha}^{(k)} < +\infty, \quad \sum_{k \in \mathbb{N}} \boldsymbol{\beta}^{(k)} < +\infty, \quad \sum_{k \in \mathbb{N}} \boldsymbol{\gamma}^{(k)} < +\infty, \quad \mathbb{P}\text{-p.s. .}$$

Alors, $\{\mathbf{A}^{(k)}\}_{k \in \mathbb{N}}$ converge presque sûrement vers une variable aléatoire $\mathbf{A}^\infty$ bornée presque sûrement, et on a de plus :

$$\sum_{k \in \mathbb{N}} \boldsymbol{\eta}^{(k)} < +\infty, \quad \mathbb{P}\text{-p.s. .}$$

Preuve. On considère une réalisation des différentes suites vérifiant les hypothèses du corollaire, et on définit (avec des notations évidentes) les trois suites $\{\tilde{\alpha}^{(k)}\}_{k \in \mathbb{N}}$, $\{\tilde{\beta}^{(k)}\}_{k \in \mathbb{N}}$ et $\{\tilde{\eta}^{(k)}\}_{k \in \mathbb{N}}$ telles que :

$$1 + \tilde{\alpha}^{(k)} = \frac{1 + \alpha^{(k)}}{1 - \gamma^{(k)}}, \quad \tilde{\beta}^{(k)} = \frac{\beta^{(k)}}{1 - \gamma^{(k)}}, \quad \tilde{\eta}^{(k)} = \frac{\eta^{(k)}}{1 - \gamma^{(k)}}.$$

Comme la suite de terme général $\gamma^{(k)}$ tend vers zéro, elle est, à partir d'un certain rang, inférieure ou égale à $1/2$. De $\gamma^{(k)} \in [0, 1/2]$, on obtient que :

$$\frac{1}{1 - \gamma^{(k)}} \leq 1 + 2\gamma^{(k)} \quad \text{et} \quad 1 \leq \frac{1}{1 - \gamma^{(k)}} \leq 2.$$

On en déduit alors que $\tilde{\alpha}^{(k)} \leq 2(\alpha^{(k)} + \gamma^{(k)})$, $\tilde{\beta}^{(k)} \leq 2\beta^{(k)}$ et $\tilde{\eta}^{(k)} \geq \eta^{(k)}$. Les conclusions du corollaire découlent alors d'une application directe du théorème de Robbins-Siegmund.

Lors de l'étude des propriétés de l'algorithme du gradient stochastique généralisé, on a utilisé le lemme suivant.

Lemme 3.7. *Soit J une fonction définie sur un espace de Hilbert $\mathbb{U}$ à valeurs dans $\mathbb{R}$, lipschitzienne de constante L . Soit $\{u^{(k)}\}_{k \in \mathbb{N}}$ une suite d'éléments de $\mathbb{U}$ et soit $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ une suite de nombres réels positifs vérifiant :*

- (a) $\sum_{k \in \mathbb{N}} \epsilon^{(k)} = +\infty$,
 (b) $\exists \mu \in \mathbb{R}, \sum_{k \in \mathbb{N}} \epsilon^{(k)} |J(u^{(k)}) - \mu| < +\infty$,
 (c) $\exists \delta > 0, \forall k \in \mathbb{N}, \|u^{(k+1)} - u^{(k)}\| \leq \delta \epsilon^{(k)}$.

Alors, la suite $\{J(u^{(k)})\}_{k \in \mathbb{N}}$ converge vers μ .

Preuve. Pour tout réel positif α , on définit le sous-ensemble N_α de $\mathbb{N}$ et son complémentaire N_α^c par :

$$N_\alpha = \{k \in \mathbb{N}, |J(u^{(k)}) - \mu| \leq \alpha\}, \quad N_\alpha^c = \mathbb{N} \setminus N_\alpha.$$

- (i) Les hypothèses (a) et (b) impliquent que le cardinal de l'ensemble N_α n'est pas fini.
 (ii) D'après l'hypothèse (b) :

$$+\infty > \sum_{k \in \mathbb{N}} \epsilon^{(k)} |J(u^{(k)}) - \mu| \geq \sum_{k \in N_\alpha^c} \epsilon^{(k)} |J(u^{(k)}) - \mu| \geq \alpha \sum_{k \in N_\alpha^c} \epsilon^{(k)};$$

on en déduit donc que :

$$\forall \beta > 0, \exists n_\beta \in \mathbb{N} \text{ tel que } \sum_{k \geq n_\beta, k \in N_\alpha^c} \epsilon^{(k)} \leq \beta.$$

Pour tout $\epsilon > 0$, on choisit $\alpha = \epsilon/2$ et $\beta = \epsilon/(2L\delta)$ (où L est la constante de Lipschitz de J). Soit n_β l'entier défini en (ii). Pour $k \geq n_\beta$ quelconque, deux cas sont possibles :

- ou $k \in N_\alpha$: on a alors par définition :

$$|J(u^{(k)}) - \mu| \leq \alpha < \epsilon,$$

- ou $k \in N_\alpha^c$: soit alors m le plus petit élément de N_α tel que $m \geq k$ (cet élément existe d'après (i)) ; utilisant le fait que J soit lipschitzienne et la condition (ii), il vient :

$$\begin{aligned} |J(u^{(k)}) - \mu| &\leq |J(u^{(k)}) - J(u^{(m)})| + |J(u^{(m)}) - \mu| \\ &\leq L \|u^{(k)} - u^{(m)}\| + \alpha \\ &\leq L\delta \left(\sum_{l=k}^{m-1} \epsilon^{(l)} \right) + \alpha \\ &\leq L\delta \left(\sum_{l \geq n_\beta, l \in N_\alpha^c} \epsilon^{(l)} \right) + \alpha \\ &\leq \epsilon, \end{aligned}$$

d'où le résultat.

3.5 Annexe : intégrande normale et sélection mesurable

Les résultats donnés dans cette annexe concernant les questions de mesurabilité des multi-applications et de leurs sélections, ainsi que les propriétés des intégrandes, se trouvent dans (ROCKAFELLAR et WETS, 1998, Chapter 14).

On se donne un espace Ω muni de sa tribu $\mathcal{A}$, ainsi qu'un espace de Hilbert $\mathbb{U}$ de dimension finie.

Définition 3.8. *On dit que la multi-application $S : \Omega \rightrightarrows \mathbb{U}$ est mesurable si, pour tout ouvert $\mathcal{O} \subset \mathbb{U}$, l'ensemble $S^{-1}(\mathcal{O}) = \{\omega \in \Omega, S(\omega) \cap \mathcal{O} \neq \emptyset\}$ est mesurable, c'est-à-dire si $S^{-1}(\mathcal{O}) \in \mathcal{A}$.*

On notera que cette définition correspond à la définition classique de la mesurabilité dans le cas où S est une application. Le résultat fondamental suivant (ROCKAFELLAR et WETS, 1998, Corollary 14.6) définit et caractérise les sélections mesurables d'une multi-application.

Théorème 3.9. *Une multi-application $S : \Omega \rightrightarrows \mathbb{U}$ mesurable et à valeur fermée admet une sélection mesurable : il existe une application $s : \Omega \rightarrow \mathbb{U}$ telle que $s(\omega) \in S(\omega)$ pour tout $\omega \in \Omega$.*

On s'intéresse maintenant à une application f dépendant de deux variables ω et $u : f : \Omega \times \mathbb{U} \rightarrow \mathbb{R}$. Une question fondamentale est de déterminer les conditions qui garantissent la mesurabilité de l'application $\omega \mapsto f(\omega, U(\omega))$ lorsque l'application $\omega \mapsto U(\omega)$ est mesurable. On introduit pour cela les définitions suivantes

Définition 3.10. *Une application $f : \Omega \times \mathbb{U} \rightarrow \overline{\mathbb{R}}$ qui est mesurable par rapport à ω pour tout $u \in \mathbb{U}$ est appelée une intégrande.*

Dans le contexte de l'optimisation, on associe à une intégrande f la multi-application épigraphique S_f définie par :

$$S_f : \omega \mapsto \text{epi} f(\omega, \cdot) = \{(u, \alpha) \in \mathbb{U} \times \mathbb{R}, f(\omega, u) \leq \alpha\}.$$

La notion d'intégrande normale (ROCKAFELLAR et WETS, 1998, Definition 14.27) est alors définie comme suit.

Définition 3.11. *Une intégrande normale f est une intégrande telle que la multi-application épigraphique S_f associée soit mesurable et à valeur fermée.*

Une première conséquence de cette dernière définition est le résultat suivant (ROCKAFELLAR et WETS, 1998, Theorem 14.37).

Théorème 3.12. *Soit $f : \Omega \times \mathbb{U} \rightarrow \overline{\mathbb{R}}$ une intégrande normale. Alors, l'application $u \mapsto f(\omega, u)$ est semi-continue inférieurement pour tout $\omega \in \Omega$, l'application $\omega \mapsto f(\omega, u)$ est mesurable pour tout $u \in \mathbb{U}$, et l'application $\omega \mapsto f(\omega, U(\omega))$ est mesurable pour toute application U mesurable.*

On notera que la réciproque de ce théorème est en général fausse : toute intégrande semi-continue inférieurement en u et mesurable en ω n'est pas une intégrande normale...

L'un des intérêts majeurs de la notion d'intégrande normale est donné par le résultat suivant (ROCKAFELLAR et WETS, 1998, Theorem 14.37).

Théorème 3.13. *Soit une intégrande normale $f : \Omega \times \mathbb{U} \rightarrow \overline{\mathbb{R}}$, à laquelle on associe l'application p définie par :*

$$p(\omega) = \inf_u f(\omega, u) ,$$

et la multi-application P définie par :

$$P(\omega) = \arg \min_u f(\omega, u) .$$

Alors, la fonction $p : \Omega \rightarrow \overline{\mathbb{R}}$ est mesurable. La multi-application $P : \Omega \rightrightarrows \mathbb{U}$ est mesurable à valeurs fermées et admet donc une sélection mesurable.