

3 Gradient stochastique généralisé sans contrainte explicite

3.1 Rappel du principe du problème auxiliaire

Dans le cadre de l'optimisation convexe déterministe, le *principe du problème auxiliaire* (PPA) consiste à remplacer le problème :

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

par une suite de problèmes auxiliaires indexés par k . Le k -ème problème auxiliaire s'écrit :

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

et sa solution $u^{(k+1)}$ permet de formuler le problème auxiliaire d'indice $k+1$.

Ce principe n'a d'intérêt que dans la mesure où la résolution de chaque problème auxiliaire $(\mathbf{PA}^{(k)})$ est plus facile que la résolution du problème initial (\mathbf{P}) . La fonction K est *choisie* par la personne mettant en œuvre le PPA, et on l'appelle le *noyau de décomposition*. C'est de ce choix que provient l'intérêt pratique de la méthode. Les principales propriétés de ce principe sont les suivantes.

- Le PPA est **consistant** : si l'on suppose que la suite $\{u^{(k)}\}_{k \in \mathbb{N}}$ des solutions converge vers un élément u^\sharp , le passage à la limite dans les conditions d'optimalité du problème $(\mathbf{PA}^{(k)})$:

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

conduit à la condition du d'optimalité du problème (\mathbf{P}) par élimination des dérivées de K :

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

ce qui montre que u^\sharp est solution du problème (\mathbf{P}) .

- Le PPA permet de **retrouver** les algorithmes classiques d'optimisation : par exemple, avec un choix de noyau quadratique $K(u) = 1/2 \|u\|^2$, le problème $(\mathbf{PA}^{(k)})$ prend la forme :

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

Sa solution s'obtient alors de manière explicite :²¹

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

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

- Le PPA permet de **décomposer** chaque problème auxiliaire : si l'on suppose que l'ensemble admissible U^{ad} se met sous la forme du produit cartésien de N ensembles U_i^{ad} , le vecteur $u \in U^{\text{ad}}$ s'écrivant (u_1, \dots, u_N) avec $u_i \in U_i^{\text{ad}}$, et si l'on choisit un noyau K additif par rapport à cette décomposition de u : $K(u) = \sum_{i=1}^N K_i(u_i)$, le problème $(\mathbf{PA}^{(k)})$ prend la forme :

$$\begin{aligned} & \min_{\substack{u_1 \in U_1^{\text{ad}} \\ \dots \\ u_N \in U_N^{\text{ad}}}} \sum_{i=1}^N \left(K_i(u_i) + \langle \epsilon \nabla J_{u_i}(u^{(k)}) - \nabla K_i(u_i^{(k)}), u_i \rangle \right) , \end{aligned}$$

qui se décompose en N sous-problèmes d'optimisation indépendants, de la forme :

$$\min_{u_i \in U_i^{\text{ad}}} K_i(u_i) + \langle \epsilon \nabla J_{u_i}(u^{(k)}) - \nabla K_i(u_i^{(k)}), u_i \rangle .$$

On pourra consulter le chapitre 11 pour un bref exposé de ce principe. Pour une description complète du PPA et de ses applications, on se reportera au cours [4] de G. Cohen.

²¹Comme on minimise un opérateur sphérique ($\frac{1}{2} \|u\|^2$), on a la propriété que la solution sous la contrainte $u \in U^{\text{ad}}$ est obtenue en projetant la solution $u^{(k)} - \epsilon \nabla J(u^{(k)})$ du problème sans contraintes sur l'ensemble U^{ad} . Cette propriété est bien sûr fautive dans le cas d'un opérateur quadratique quelconque.

3.2 Algorithme du gradient stochastique généralisé

Partant du problème de minimisation (10) :

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

où $J(u) = \mathbb{E}(j(u, \mathbf{W}))$, le mélange de l'idée du principe du problème auxiliaire et de l'idée du gradient stochastique conduit à remplacer la résolution du problème (P) par la résolution d'une suite de problèmes auxiliaires dont la k -ème instance ($\mathbf{PA}^{(k)}$) est :

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

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

Algorithme 1. (*Algorithme du gradient stochastique généralisé*)

1. Choisir un point initial $u^{(0)} \in U^{\text{ad}}$ ainsi qu'une suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ de réels positifs.
2. À l'itération k de l'algorithme, effectuer un tirage $w^{(k+1)}$ de la variable aléatoire \mathbf{W} suivant sa loi, indépendamment des tirages $(w^{(1)}, \dots, w^{(k)})$ des itérations précédentes.
3. Résoudre le problème auxiliaire ($\mathbf{PA}^{(k)}$) et noter $u^{(k+1)}$ une solution de ce problème.
4. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Remarque 6. Avec le choix de noyau $K(u) = \frac{1}{2} \|u\|^2$, le problème auxiliaire ($\mathbf{PA}^{(k)}$) 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 $u^{(k+1)}$ se calcule explicitement :

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

et correspond exactement à un pas de l'algorithme 1 du gradient stochastique. \square

On va s'intéresser dans ce chapitre à la convergence de cet algorithme du gradient stochastique généralisé. Les résultats qui suivent ont été obtenus par J. C. Culioli dans le cadre de sa thèse [5], et exposés dans [6]. On a pris le parti de présenter les résultats dans le cadre différentiable, et non dans le cadre sous-différentiable plus général comme cela est fait dans [5] et [6].

3.3 Théorème de convergence

On introduit pour commencer la notion de σ -suite, qui caractérise la longueur des pas de gradients que l'on doit effectuer dans un algorithme de type gradient stochastique.

Définition 1. Soit $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ une suite de nombres réels strictement positifs. On dit que $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est une σ -suite si elle vérifie :

$$\sum_{k \in \mathbb{N}} \epsilon^{(k)} = +\infty \quad \text{et} \quad \sum_{k \in \mathbb{N}} (\epsilon^{(k)})^2 < +\infty .$$

Comme au § 2.2, on utilise la description de l'algorithme en terme de variables aléatoires, et on suppose que l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ contient une suite $\{\mathbf{W}^{(k)}\}_{k \in \mathbb{N}^*}$ de variables aléatoires indépendantes de même loi que \mathbf{W} . La question de la convergence de l'algorithme 1 et celle du lien de cet algorithme avec le problème de départ (P) sont réglées par le théorème suivant.

Théorème 1. (Convergence presque sûre du gradient stochastique généralisé)

On suppose que les hypothèses suivantes sont vérifiées.

- H0** U^{ad} est une partie convexe fermée non vide de \mathbb{U} , espace de Hilbert de dimension finie ;
 \mathbf{W} est une variable aléatoire définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans un espace $(\mathbb{W}, \mathcal{W})$, de loi μ .
- H1** La variable aléatoire $j(u, \mathbf{W}) : \Omega \rightarrow \mathbb{R}$ est mesurable et son espérance existe pour tout $u \in U^{\text{ad}}$.
- H2** La fonction $j(\cdot, w) : \mathbb{U} \rightarrow \mathbb{R}$ est propre, convexe, semi-continue inférieurement, différentiable pour tout $w \in \mathbb{W}$, et son gradient partiel en u est noté $\nabla_u j(u, w)$.
- H3** La fonction $j(\cdot, w)$ est à gradient linéairement borné (GLB) uniformément en w :
- $$\exists c_1 > 0, \exists c_2 > 0, \forall w \in \mathbb{W}, \forall u \in U^{\text{ad}}, \|\nabla_u j(u, w)\| \leq c_1 \|u\| + c_2 .$$
- H4** La fonction J est coercive sur l'ensemble U^{ad} .
- H5** La fonction K est propre, fortement convexe de module b , différentiable.
- H6** La suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est une σ -suite.

On a alors les conclusions suivantes.

- C1** Le problème (\mathbf{P}) admet un ensemble de solutions $U^\#$ non vide.
- C2** Le problème $(\mathbf{PA}^{(k)})$ admet une solution $\mathbf{U}^{(k+1)}$ unique.
- C3** La suite de variables aléatoires $\{J(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ converge presque sûrement vers $\min_{u \in U^{\text{ad}}} J(u)$.
- C4** La suite des solutions $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ engendrée par l'algorithme 1 est bornée presque sûrement, tout point d'accumulation d'une réalisation de cette suite appartenant alors à l'ensemble optimal $U^\#$.

Si l'on suppose de plus que la fonction J est fortement convexe, la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ engendrée par l'algorithme 1 converge presque sûrement vers l'unique solution $u^\#$ du problème (\mathbf{P}) .

Preuve. La démonstration des points **C1** et **C2** découle des théorèmes généraux relatifs à l'optimisation sans contraintes explicites. La démonstration des points **C3** et **C4** se fait en quatre étapes, à savoir :

- choix d'une fonction de Lyapunov,
- majoration de la variation de cette fonction,
- convergence par un argument de type martingale,
- analyse des limites des suites.

1. **Fonction de Lyapunov.** Soit $u^\# \in U^\#$ une solution de (\mathbf{P}) . On choisit la 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 directement :

$$\frac{b}{2} \|u - u^\#\|^2 \leq \Lambda(u) , \tag{33}$$

ce qui prouve que Λ est bornée inférieurement et coercive.

2. **Majoration.** 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 1 pour une réalisation $(w^{(1)}, \dots, w^{(k)}, \dots)$ des variables aléatoires du problème.

$$\Delta^{(k)} = \underbrace{K(u^{(k)}) - K(u^{(k+1)}) - \left\langle \nabla K(u^{(k)}), u^{(k)} - u^{(k+1)} \right\rangle}_{T_1} + \underbrace{\left\langle \nabla K(u^{(k)}) - \nabla K(u^{(k+1)}), u^\# - u^{(k+1)} \right\rangle}_{T_2},$$

- Par convexité de K , on a :

$$T_1 \leq 0.$$

- Avec la notation $r^{(k)} = \nabla_u j(u^{(k)}, w^{(k+1)})$, les conditions d'optimalité du problème $(\mathbf{PA}^{(k)})$ évaluées au point $u = u^\#$ se mettent sous la forme :

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

- Par convexité de $j(\cdot, w^{(k+1)})$, on a :

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

- Les conditions d'optimalité du problème $(\mathbf{PA}^{(k)})$ évaluées au point $u = u^{(k)}$ et la propriété de forte monotonie de K' impliquent :

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

L'inégalité de Schwartz appliquée à cette inégalité donne :

$$\left\| u^{(k+1)} - u^{(k)} \right\| \leq \frac{\epsilon^{(k)}}{b} \left\| r^{(k)} \right\|, \quad (34)$$

d'où, en appliquant de nouveau l'inégalité de Schwartz au terme T_4 :

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

La condition GLB se met sous la forme équivalente : $\left\| r^{(k)} \right\| \leq c_3 \left\| u^{(k)} - u^\# \right\| + c_4$; élevant cette inégalité au carré, utilisant la propriété générale $(a+b)^2 \leq 2(a^2 + b^2)$ ainsi que la relation (33), on obtient :

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

d'où la nouvelle majoration :

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

Regroupant les majorations sur T_1, T_2, T_3 et T_4 , il vient :

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

ce qui, en terme de variables aléatoires, s'écrit :

$$\Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) \leq \epsilon^{(k)} \left(j(u^\#, \mathbf{W}^{(k+1)}) - j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) \right) + \frac{(\epsilon^{(k)})^2}{b} \left(\alpha \Lambda(\mathbf{U}^{(k)}) + \beta \right).$$

Prenant de part et d'autre de cette dernière inégalité l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^{(k)}$ engendrée par les k premières variables aléatoires $(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)})$, on obtient :

$$\mathbb{E} \left(\Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) \mid \mathcal{F}^{(k)} \right) \leq \underbrace{\epsilon^{(k)} \left(J(u^\#) - J(\mathbf{U}^{(k)}) \right)}_{T_5} + \left(\alpha^{(k)} \mathbb{E} \left(\Lambda(\mathbf{U}^{(k)}) \mid \mathcal{F}^{(k)} \right) + \beta^{(k)} \right), \quad (35)$$

où $\alpha^{(k)} = (\epsilon^{(k)})^2 \alpha / b$ et $\beta^{(k)} = (\epsilon^{(k)})^2 \beta / b$ sont les termes de deux séries convergentes. De par l'optimalité de $u^\#$, le terme T_5 est toujours négatif. Comme la variable aléatoire $\Lambda(\mathbf{U}^{(k)})$ est mesurable par rapport à la tribu $\mathcal{F}^{(k)}$, et donc égale à son espérance conditionnelle, on a :

$$\mathbb{E} \left(\Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) \mid \mathcal{F}^{(k)} \right) \leq \alpha^{(k)} \Lambda(\mathbf{U}^{(k)}) + \beta^{(k)}. \quad (36)$$

Prenant l'espérance de part et d'autre de l'inégalité dans la relation (36) et appliquant le lemme 1 (voir l'annexe de ce chapitre), on obtient que la suite de terme général $\mathbb{E}(\Lambda(\mathbf{U}^{(k)}))$ est bornée par un réel $M > 0$.²²

3. **Convergence.** Le processus $\{\Lambda(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$, qui est adapté à la filtration $\{\mathcal{F}^{(k)}\}_{k \in \mathbb{N}}$, est une *quasi-martingale*. En effet :

- $\inf_{k \in \mathbb{N}} \mathbb{E}(\Lambda(\mathbf{U}^{(k)})) > -\infty$ car $\Lambda(\mathbf{U}^{(k)}) \geq 0$ d'après la relation (33) ;
- notant $G^{(k)}$ l'ensemble $\{\omega \in \Omega, \mathbb{E}(\Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) \mid \mathcal{F}^{(k)}) > 0\}$, remarquant que la fonction indicatrice $\mathbf{1}_{G^{(k)}}$ de cet ensemble est mesurable par rapport à la tribu $\mathcal{F}^{(k)}$, et utilisant la relation (36), il vient :

$$\begin{aligned} \sum_{k \in \mathbb{N}} \mathbb{E} \left(\mathbf{1}_{G^{(k)}} \cdot \left(\Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) \right) \right) &= \sum_{k \in \mathbb{N}} \mathbb{E} \left(\mathbf{1}_{G^{(k)}} \cdot \mathbb{E} \left(\Lambda(\mathbf{U}^{(k+1)}) - \Lambda(\mathbf{U}^{(k)}) \mid \mathcal{F}^{(k)} \right) \right) \\ &\leq \sum_{k \in \mathbb{N}} \mathbb{E} \left(\mathbf{1}_{G^{(k)}} \cdot \left(\alpha^{(k)} \Lambda(\mathbf{U}^{(k)}) + \beta^{(k)} \right) \right) \\ &\leq \sum_{k \in \mathbb{N}} \alpha^{(k)} \mathbb{E} \left(\Lambda(\mathbf{U}^{(k)}) \right) + \beta^{(k)} \\ &\leq \sum_{k \in \mathbb{N}} \alpha^{(k)} M + \beta^{(k)}, \end{aligned}$$

et cette dernière quantité est bornée car $\alpha^{(k)}$ et $\beta^{(k)}$ sont les termes de séries convergentes.

La propriété de quasi-martingale découle alors directement de la proposition 12 (voir § 13.2).

Enfin, comme $\Lambda(\mathbf{U}^{(k)}) \geq 0$, il est clair que $\sup_{k \in \mathbb{N}} \mathbb{E}((\Lambda(\mathbf{U}^{(k)}))_-) < +\infty$. Par le théorème 16 (§ 13.2), on obtient que la quasi-martingale $\{\Lambda(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ converge presque sûrement vers une variable aléatoire intégrable.

4. **Limites des suites.** De la convergence presque sûre précédente, on déduit que la suite $\{\Lambda(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ est bornée presque sûrement²³. À l'aide de la relation (33), on obtient que la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ est aussi bornée presque sûrement : l'ensemble $\Omega_0 \subset \Omega$ des aléas, tels que la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ ne soit pas bornée, est de mesure nulle. Soit alors $\omega \notin \Omega_0$. La suite des réalisations $\{u^{(k)}\}_{k \in \mathbb{N}}$ associée à cet ω est bornée et chaque $u^{(k)}$ appartient à U^{ad} , partie fermée de \mathbb{U} . Par compacité (car \mathbb{U} est de dimension finie), on déduit qu'il existe une sous-suite convergente $\{u^{(\Phi(k))}\}_{k \in \mathbb{N}}$ de la suite $\{u^{(k)}\}_{k \in \mathbb{N}}$. On remarquera que cette sous-suite *dépend* de l'aléa $\omega \in \Omega_0$, ce qui signifie que la fonction Φ correspond à la réalisation en ω d'une fonction aléatoire Φ :

$$u^{(\Phi(k))} = \left(\mathbf{U}^{(\Phi(k))} \right) (\omega), \text{ avec } \Phi(k) = (\Phi(\omega))(k).$$

Pour montrer la convergence de la suite $\{J(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$, on revient à la majoration (35), qui se met sous la forme :

$$\epsilon^{(k)} \left(J(\mathbf{U}^{(k)}) - J(u^\sharp) \right) \leq \mathbb{E} \left(\Lambda(\mathbf{U}^{(k)}) - \Lambda(\mathbf{U}^{(k+1)}) \mid \mathcal{F}^{(k)} \right) + \alpha^{(k)} \Lambda(\mathbf{U}^{(k)}) + \beta^{(k)}.$$

Prenant l'espérance de part et d'autre de l'inégalité et sommant pour k variant de 0 à n , il vient :

$$\begin{aligned} \sum_{k=0}^n \epsilon^{(k)} \mathbb{E} \left(J(\mathbf{U}^{(k)}) - J(u^\sharp) \right) &\leq \mathbb{E} \left(\Lambda(u^{(0)}) \right) - \mathbb{E} \left(\Lambda(\mathbf{U}^{(n+1)}) \right) + \sum_{k=0}^n \left(\alpha^{(k)} \mathbb{E} \left(\Lambda(\mathbf{U}^{(k)}) \right) + \beta^{(k)} \right) \\ &\leq \mathbb{E} \left(\Lambda(u^{(0)}) \right) + \sum_{k=0}^n \left(\alpha^{(k)} M + \beta^{(k)} \right), \end{aligned}$$

car $\mathbb{E}(\Lambda(\mathbf{U}^{(k)}))$ est borné par la constante $M > 0$ pour tout k . Comme $\alpha^{(k)}$ et $\beta^{(k)}$ sont les termes de séries convergentes, on obtient en passant à la limite en n dans l'expression précédente :

$$\sum_{k \in \mathbb{N}} \epsilon^{(k)} \mathbb{E} \left(J(\mathbf{U}^{(k)}) - J(u^\sharp) \right) < +\infty.$$

²²Une utilisation directe du théorème de Robbins-Siegmund montre même que la suite de variables aléatoires $\mathbb{E}(\Lambda(\mathbf{U}^{(k)}))$ converge presque sûrement vers une variable aléatoire intégrable.

²³c'est-à-dire que l'ensemble $\left\{ \omega \in \Omega, \lim_{k \rightarrow +\infty} \Lambda(\mathbf{U}^{(k)})(\omega) = +\infty \right\}$ est de mesure nulle

De plus, chaque terme $J(\mathbf{U}^{(k)}) - J(u^\sharp)$ est positif ou nul par définition de u^\sharp ; on en déduit donc que, presque sûrement, on a :

$$\sum_{k \in \mathbb{N}} \epsilon^{(k)} \left(J(\mathbf{U}^{(k)}) - J(u^\sharp) \right) < +\infty .$$

Comme la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ est bornée presque sûrement, on déduit de l'hypothèse GLB que la suite $\{\mathbf{R}^{(k)}\}_{k \in \mathbb{N}}$ est elle aussi bornée presque sûrement. Par application de la relation (34), on sait qu'il existe un réel δ vérifiant $\|\mathbf{U}^{(k+1)} - \mathbf{U}^{(k)}\| \leq \delta \epsilon^{(k)}$. On conclut à l'aide du lemme 2 (en annexe de ce chapitre) que la suite $\{J(\mathbf{U}^{(k)})\}_{k \in \mathbb{N}}$ converge vers $J(u^\sharp)$ presque sûrement.

On considère enfin un point d'accumulation \bar{u} de la suite de réalisations $\{u^{(k)}\}_{k \in \mathbb{N}}$ pour un $\omega \in \Omega_0$ (cette suite est bornée par définition de Ω_0). Le point \bar{u} est donc la limite d'une sous-suite $\{u^{(\Phi(k))}\}_{k \in \mathbb{N}}$. La semi-continuité inférieure de la fonction J permet d'écrire :

$$J(\bar{u}) \leq \lim_{k \rightarrow +\infty} J(u^{(\Phi(k))}) = J(u^\sharp) ,$$

d'où l'on déduit que $\bar{u} \in U^\sharp$.

Pour conclure, on considère le cas où J est fortement convexe de module a . La solution u^\sharp du problème (P) est alors unique. Elle est caractérisée par l'inéquation variationnelle :

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

L'écriture de la condition de forte convexité sur J conduit à :

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

Comme $J(\mathbf{U}^{(k)})$ converge vers $J(u^\sharp)$ presque sûrement, on en déduit que $\|\mathbf{U}^{(k)} - u^\sharp\|$ converge vers zéro presque sûrement, ce qui achève la démonstration. \square

3.4 Conclusions

On a donc énoncé et démontré un théorème très général de convergence pour l'algorithme du gradient stochastique. Ce théorème inclut bien évidemment le cas du gradient stochastique standard (avec un noyau de décomposition de la forme $K(u) = \frac{1}{2} \|u\|^2$), et il inclut aussi l'algorithme de Newton stochastique (avec pour noyau de décomposition $\bar{K}(u) = \frac{1}{2} \langle u, Au \rangle$).

D'un point de vue théorique, les hypothèses sous lesquelles le résultat est obtenu sont assez faibles, et en particulier semblent tout à fait "raisonnables" dans le cadre de l'optimisation convexe. On notera qu'un résultat de convergence de même nature, mais ne faisant pas directement appel à des hypothèses de convexité existe : ce résultat fait référence à la théorie de l'approximation stochastique, et on le trouvera par exemple dans [12].

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 (10), cette décomposition 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 l'algorithme de gradient stochastique standard, qui, comme tout algorithme de gradient, est une méthode de décomposition !

Annexe : lemmes techniques

Lors de l'étude des propriétés de l'algorithme du gradient stochastique généralisé, on a utilisé les deux lemmes suivants.

Lemme 1. Soit $\{x^{(k)}\}_{k \in \mathbb{N}}$ une suite de nombres réels positifs. Soit $\{\alpha^{(k)}\}_{k \in \mathbb{N}}$ et $\{\beta^{(k)}\}_{k \in \mathbb{N}}$ deux suites de nombres réels positifs, telles que $\alpha^{(k)}$ et $\beta^{(k)}$ soient les termes de deux séries convergentes. Si la relation :

$$x^{(k+1)} - x^{(k)} \leq \alpha^{(k)}x^{(k)} + \beta^{(k)},$$

est vérifiée pour tout k , alors la suite $\{x^{(k)}\}_{k \in \mathbb{N}}$ est bornée.

Preuve. Par sommation de l'inégalité entre $k = 0$ et $k = n$, on obtient :

$$x^{(n+1)} \leq \left(\sum_{k=0}^n \alpha^{(k)}x^{(k)} \right) + \left(x^{(0)} + \sum_{k=0}^n \beta^{(k)} \right).$$

Notant $m^{(k)} = \max_{0 \leq l \leq k} x^{(l)}$, la positivité des termes de l'expression précédente implique :

$$m^{(n+1)} \leq \left(\sum_{k=0}^n \alpha^{(k)}m^{(k)} \right) + \left(x^{(0)} + \sum_{k=0}^n \beta^{(k)} \right).$$

Pour tout $n_0 \in \mathbb{N}$ et tout $n \geq n_0$, on a alors :

$$\begin{aligned} m^{(n+1)} &\leq \left(\sum_{k=0}^{n_0-1} \alpha^{(k)}m^{(k)} \right) + \left(x^{(0)} + \sum_{k=0}^n \beta^{(k)} \right) + \left(\sum_{k=n_0}^n \alpha^{(k)}m^{(k)} \right) \\ &\leq \left(\sum_{k=0}^{n_0-1} \alpha^{(k)}m^{(k)} \right) + \left(x^{(0)} + \sum_{k=0}^n \beta^{(k)} \right) + m^{(n+1)} \left(\sum_{k=n_0}^n \alpha^{(k)} \right), \end{aligned}$$

et donc :

$$m^{(n+1)} \leq \frac{\left(\sum_{k=0}^{n_0-1} \alpha^{(k)}m^{(k)} \right) + \left(x^{(0)} + \sum_{k=0}^n \beta^{(k)} \right)}{1 - \sum_{k=n_0}^n \alpha^{(k)}}.$$

Or, d'après les propriétés des suites $\{\alpha^{(k)}\}_{k \in \mathbb{N}}$ et $\{\beta^{(k)}\}_{k \in \mathbb{N}}$, on sait que :

- $\gamma^{(n)} = x^{(0)} + \sum_{k=0}^n \beta^{(k)}$ est le terme d'une suite croissante convergente, de limite notée γ ,
- $\forall \epsilon > 0, \exists n_0 \in \mathbb{N}, \forall n \geq n_0, \sum_{k=n_0}^n \alpha^{(k)} \leq \epsilon$,

et donc que, étant donné un réel positif ϵ , il existe un entier n_0 tel que :

$$\forall n \geq n_0, m^{(n+1)} \leq \frac{\left(\sum_{k=0}^{n_0-1} \alpha^{(k)}m^{(k)} \right) + \gamma}{1 - \epsilon}.$$

On en déduit que la suite $\{m^{(k)}\}_{k \in \mathbb{N}}$, et donc a fortiori la suite $\{x^{(k)}\}_{k \in \mathbb{N}}$, est bornée. □

Lemme 2. Soit J une fonction définie sur un espace de Hilbert \mathbb{U} à valeurs dans \mathbb{R} , lipschitzienne. Soit $\{u^{(k)}\}_{k \in \mathbb{N}}$ une suite d'éléments de \mathbb{U} et $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ 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 :

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

- (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) :

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

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 $\alpha = \frac{\epsilon}{2}$, $\beta = \frac{\epsilon}{2L\delta}$ (où L est la constante de Lipschitz de J) et 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 :

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

- 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 :

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

d'où le résultat. □