2

Vue d'ensemble de la méthode du gradient stochastique

La méthode du gradient stochastique est relativement ancienne. Les initiateurs en sont H. Robbins et S. Monro (Robbins et Monro (1951)) d'une part, J. Kiefer et J. Wolfowitz (Kiefer et Wolfowitz (1952)) d'autre part, dans le cadre général de l'approximation stochastique (voir Lai (2003) pour une mise en perspective de ces travaux). Plus récemment, B. T. Polyak (Polyak (1976)–Polyak et Tsypkin (1979)) a donné des conditions de convergence pour ce type d'algorithme, ainsi que des résultats de vitesse de convergence. Sur la base de ces travaux, J. C. Dodu et ses coauteurs (Dodu et collab. (1981)) ont étudié dans certains cas l'optimalité de l'algorithme du gradient stochastique, c'est-à-dire l'efficacité asymptotique de l'estimateur fourni par l'algorithme. Une importante contribution de B. T. Polyak (Polyak (1990)–Polyak et Juditsky (1992)) dans ce domaine a été d'introduire dans l'algorithme du gradient stochastique une technique de moyennisation permettant de garantir en un certain sens son optimalité.

Ces travaux ont aussi été développés dans le cadre de l'approximation stochastique. Le premier livre de référence sur le sujet est celui de H. J. Kushner et D. S. Clark (Kushner et Clark (1978)) présentant, dans le cas non convexe, la méthode de l'équation différentielle moyenne (ODE dans la terminologie anglo-saxonne) permettant l'étude de la convergence locale des algorithmes stochastiques généraux. Plusieurs ouvrages, comme ceux de M. Duflo (Duflo (1996)–Duflo (1997)) et de H. J. Kushner et G. G. Yin (Kushner et Yin (2003)) ont traité de développements importants de cette théorie, comme l'étude de la normalité asymptotique ou la prise en compte de contraintes. On se référera au cours proposé par B. Delyon (Delyon (2000)), disponible sur le site Web de l'auteur et d'une lecture relativement aisée.

Le but de ce chapitre est de décrire l'algorithme du gradient stochastique dans le cas le plus simple, de donner le cadre probabiliste adapté à son étude, et d'énoncer les théorèmes classiques de convergence. On énoncera aussi un théorème de type limite centrale associé à cet algorithme, les principaux résultats concernant l'optimalité de la méthode et enfin l'algorithme du gra-

dient stochastique moyenné et son comportement asymptotique. On conclura en donnant quelques indications pratiques sur la mise en œuvre de la méthode.

2.1 Position du problème

Soit un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et une variable aléatoire $\boldsymbol{W}$ définie sur Ω à valeurs dans un espace probabilisé $(\mathbb{W}, \mathcal{W})$. On note $\mu = \mathbb{P} \circ \boldsymbol{W}^{-1}$ la loi de probabilité résultant du transport de la loi $\mathbb{P}$ par $\boldsymbol{W}$. 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 note $J(u)$ l'espérance de la fonction $j(u, \boldsymbol{W})$ (supposée intégrable pour tout $u \in U^{\mathrm{ad}}$) :

$$J(u) = \mathbb{E}\left(j(u, \boldsymbol{W})\right) = \int_{\Omega} j\big(u, \boldsymbol{W}(\omega)\big) \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{W}} j(u, w) \mathrm{d}\mu(w) \ .$$

On suppose que la fonction j est différentiable par rapport à u, et que les conditions sont remplies pour pouvoir dériver sous le signe somme (résultat classique d'intégration : voir par exemple (Schwartz, 1993, §3, Théorème 6.3.5)). Alors, la fonction J est différentiable et son gradient, noté $\nabla J(u)$, est donné par la relation :

$$\nabla J(u) = \mathbb{E}\left(\nabla_u j(u, \boldsymbol{W})\right) \ , \tag{2.1}$$

où $\nabla_u j$ est le gradient partiel de j par rapport à la variable u. On s'intéresse au problème d'optimisation suivant :

$$\min_{u \in U^{\mathrm{ad}}} J(u) \ . \tag{2.2}$$

Sous les hypothèses classiques de convexité et de différentiabilité, et si l'on est prêt à calculer le gradient $\nabla J(u)$ de J en tout point u, on peut utiliser pour obtenir la solution du problème (2.2) un algorithme de type gradient (gradient conjugué, quasi-Newton, Newton, ...). Le plus simple de ces algorithmes est celui du gradient projeté, dont une itération s'écrit :

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

où ϵ est un scalaire positif représentant la longueur du pas de gradient. Dans cette approche, on s'attaque en fait au problème *déterministe* (2.2) et l'aspect stochastique est caché dans le calcul de $\nabla J(u^{(k)})$ comme une espérance (voir (2.1)). Cette approche peut cependant s'avérer extrêmement coûteuse en temps de calcul, car chaque évaluation du gradient passe par le calcul d'une espérance sur l'espace $\mathbb{W}$ dont la dimension peut être grande.

On considère alors le problème (2.2), dans lequel on remplace $J(u)$ par son expression en fonction de j :

$$\min_{u \in U^{\text{ad}}} \mathbb{E}\left(j(u, \boldsymbol{W})\right) \ . \tag{2.3}$$

Une façon classique de contourner la difficulté liée au calcul de l'espérance est de faire appel à la méthode de Monte Carlo (voir Bouleau (1986)), et donc de remplacer le problème (2.3) par l'approximation suivante :

$$\min_{u \in U^{\text{ad}}} \frac{1}{k} \sum_{l=1}^{k} j(u, w^l) \ , \tag{2.4}$$

où $(w^1, \ldots, w^k)$ est une réalisation d'un k-échantillon [1] de $\boldsymbol{W}$. Alors, le gradient de la fonction coût du problème (2.4) est égal à

$$\frac{1}{k} \sum_{l=1}^{k} \nabla_u j(u, w^l) \ ,$$

et correspond à l'approximation de Monte Carlo du « vrai » gradient $\nabla J(u)$. Cette façon de procéder est connue sous le nom de *Sample Average Approximation* (SAA) (voir (Shapiro et collab., 2009, Chapter 5) pour une présentation détaillée). Son inconvénient principal est que la taille k de l'échantillon doit être fixée *avant* la résolution du problème d'optimisation approximé. Si cette taille s'avère insuffisante, il faut enrichir l'échantillon, puis résoudre un nouveau problème d'optimisation.

La méthode du gradient stochastique a pour ambition de surmonter les deux difficultés évoquées ci-dessus (calcul de la vraie espérance ou choix a priori de la taille de l'échantillon). Comme dans la méthode SAA, elle utilise une approximation du gradient ∇J basée sur un échantillonnage de $\boldsymbol{W}$. Mais, à la différence de SAA, les échantillons sont incorporés un à un dans l'algorithme de manière à produire une suite d'estimateurs convergeant vers la solution du problème (2.3).

2.2 Algorithme du gradient stochastique

2.2.1 Description de l'algorithme

La méthode du *gradient stochastique* consiste à mettre en œuvre un algorithme au cours duquel la variable à optimiser u évolue en fonction du gradient partiel de j par rapport à u évalué pour des réalisations successives de la variable aléatoire $\boldsymbol{W}$, et non en fonction du gradient de J. En fait, on effectue des itérations de type gradient afin de réaliser la tâche d'optimisation, et on utilise en même temps les réalisations de la variable aléatoire $\boldsymbol{W}$ obtenues au cours des itérations afin d'évaluer l'espérance, à la manière de la méthode de Monte Carlo. L'algorithme associé est le suivant.

1. On rappelle qu'un k-échantillon de $\boldsymbol{W}$ est une suite $(\boldsymbol{W}^1, \ldots, \boldsymbol{W}^k)$ de variables aléatoires indépendantes de même loi de probabilité que $\boldsymbol{W}$ (voir Bouleau (1986) pour plus de détails).

Algorithme 2.1. (Algorithme du gradient stochastique)

1. Choisir un $u^{(0)} \in U^{\text{ad}}$ initial, 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 $\boldsymbol{W}$.
3. Calculer le gradient partiel de j en $(u^{(k)}, w^{(k+1)})$ et mettre à jour u : $u^{(k+1)} = \text{proj}_{U^{\text{ad}}}\left(u^{(k)} - \epsilon^{(k)}\nabla_u j(u^{(k)}, w^{(k+1)})\right)$.
4. Incrémenter l'indice k de 1 et retourner à l'étape 2.

On notera que l'on n'a pas donné de *critère d'arrêt* pour l'algorithme de gradient stochastique. Ce point sera discuté au §2.6.

L'algorithme 2.1 correspond à la mise en œuvre numérique de la méthode du gradient stochastique. Il est commode, lorsque l'on étudie les propriétés de cet algorithme, de le décrire en terme de variables aléatoires. Les tirages $w^{(k)}$ qui y apparaissent sont des réalisations de la variable aléatoire $\boldsymbol{W}$, et une hypothèse fondamentale pour que l'algorithme 2.1 converge vers la solution du problème initial est que ces tirages $(w^{(1)}, \ldots, w^{(k)})$ correspondent à une réalisation d'un échantillon de taille k de la variable aléatoire $\boldsymbol{W}$, c'est-à-dire la réalisation d'une suite de k variables aléatoires $(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(k)})$ indépendantes, de même loi que $\boldsymbol{W}$.

On considère donc un échantillon $\{\boldsymbol{W}^{(k)}\}_{k\in\mathbb{N}}$ de taille infinie de la variable aléatoire $\boldsymbol{W}$, et l'étape 3 de remise à jour de u dans l'algorithme 2.1 peut être interprétée comme une relation de récurrence sur des variables aléatoires $\boldsymbol{U}^{(k)}$ à valeurs dans l'espace $\mathbb{U}$:

$$\boldsymbol{U}^{(k+1)} = \text{proj}_{U^{\text{ad}}}\left(\boldsymbol{U}^{(k)} - \epsilon^{(k)}\nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})\right) , \tag{2.5}$$

chaque valeur $u^{(k)}$ calculée par l'algorithme 2.1 correspondant alors à une réalisation de la variable aléatoire $\boldsymbol{U}^{(k)}$:

$$\exists\omega \in \Omega\,,\ \ \forall k \in \mathbb{N}\,,\ \ u^{(k)} = \boldsymbol{U}^{(k)}(\omega)\,.$$

L'opérateur de projection dans l'équation (2.5) doit être interprété comme une projection $\ll \omega$ par $\omega \gg$ dans l'espace $\mathbb{U}$.

Remarque 2.2. Pour que la description en terme de variables aléatoires de l'algorithme de gradient stochastique soit valide, il faut être capable de construire une suite $\{\boldsymbol{W}^{(k)}\}_{k\in\mathbb{N}^\star}$ de variables aléatoires indépendantes et de même loi μ que $\boldsymbol{W}$. Un moyen classique pour réaliser cela est de considérer l'espace de suites $\widetilde{\mathbb{W}} = \mathbb{W}^{\mathbb{N}}$ muni de la tribu $\widetilde{\mathcal{W}} = \mathcal{W}^{\otimes\mathbb{N}}$ avec la loi de probabilité $\widetilde{\mu} = \mu^{\otimes\mathbb{N}}$. Les variables aléatoires $\boldsymbol{W}^{(k)}$ sont alors définies sur l'espace de probabilité $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$ comme étant les *applications coordonnées*[2] :

$$\boldsymbol{W}^{(k)}(w^{(1)}, \ldots, w^{(k)}, \ldots) = w^{(k)}\,.$$

2. Voir (Bouleau, 1986, Ch. VII) pour plus de détails sur cette construction.

On est donc conduit à manipuler *deux* espaces de probabilité, à savoir l'espace canonique $(\Omega, \mathcal{A}, \mathbb{P})$ pour ce qui concerne la variable aléatoire $\boldsymbol{W}$, et l'espace produit $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$ pour ce qui concerne les variables aléatoires $\boldsymbol{W}^{(k)}$ et $\boldsymbol{U}^{(k)}$. Comme $\boldsymbol{W}$ peut elle-même être définie sur l'espace produit, toutes les variables aléatoires du problème peuvent en fait être définies sur l'espace $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$. Dans toute la suite, pour simplifier les notations, cet espace produit sera noté $(\Omega, \mathcal{A}, \mathbb{P})$. Avec ce changement de notation, on suppose implicitement que l'espace Ω est « assez gros » pour qu'une suite $\{\boldsymbol{W}^{(k)}\}_{k\in\mathbb{N}^\star}$ de variables aléatoires indépendantes de même loi que $\boldsymbol{W}$ puisse y exister.

2.2.2 Exemple

Cet algorithme a déjà été vu, sous une forme différente, dans le cadre de l'estimation. Donnons en un exemple lié à l'application de la méthode de Monte-Carlo. Soit $\boldsymbol{W} : \Omega \longrightarrow \mathbb{R}$ une variable aléatoire intégrable définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, dont on veut estimer l'espérance :

$$\mathbb{E}\left(\boldsymbol{W}\right) = \int_{\Omega} \boldsymbol{W}(\omega)\mathrm{d}\mathbb{P}\left(\omega\right) \ .$$

Une manière de calculer cette espérance est d'effectuer un tirage d'un k-échantillon $(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(k)})$ de la variable aléatoire $\boldsymbol{W}$, et d'en faire la moyenne arithmétique. En termes de variables aléatoires, cette moyenne s'écrit

$$\boldsymbol{U}^{(k)} = \frac{1}{k}\sum_{l=1}^{k} \boldsymbol{W}^{(l)} \ . \tag{2.6}$$

On sait par la loi forte des grands nombres (voir Bouleau (1986)) que la variable aléatoire $\boldsymbol{U}^{(k)}$ converge presque sûrement vers l'espérance de $\boldsymbol{W}$.

Par la relation (2.6), on a :

$$\begin{aligned}
\boldsymbol{U}^{(k+1)} &= \frac{1}{k+1}\sum_{l=1}^{k} \boldsymbol{W}^{(l)} \ + \ \frac{\boldsymbol{W}^{(k+1)}}{k+1} \\
&= \frac{1}{k}\sum_{l=1}^{k} \boldsymbol{W}^{(l)} \ - \ \frac{1}{k+1}\left(\frac{1}{k}\sum_{l=1}^{k} \boldsymbol{W}^{(l)} - \boldsymbol{W}^{(k+1)}\right) \\
&= \boldsymbol{U}^{(k)} - \frac{1}{k+1}\left(\boldsymbol{U}^{(k)} - \boldsymbol{W}^{(k+1)}\right) \ .
\end{aligned}$$

Posant $\epsilon^{(k)} = 1/(k+1)$ et $j(u,w) = \frac{1}{2}\big(u-w\big)^2$, cette dernière expression de $\boldsymbol{U}^{(k+1)}$ se met sous la forme :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} - \epsilon^{(k)}\nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \ . \tag{2.7}$$

Si l'on se rappelle que l'espérance de la variable aléatoire $\boldsymbol{W}$ correspond à la valeur autour de laquelle la dispersion de cette variable est minimale :

$$\mathbb{E}\left(\boldsymbol{W}\right) = \arg\min_{u\in\mathbb{R}} \frac{1}{2}\mathbb{E}\left((u-\boldsymbol{W})^2\right) \ ,$$

alors le calcul de l'espérance de $\boldsymbol{W}$ par la méthode de Monte-Carlo donné par la relation (2.7) s'interprète comme l'algorithme du gradient stochastique 2.1 appliqué à ce problème d'optimisation, l'ensemble U^{ad} étant l'espace $\mathbb{R}$ tout entier et la projection associée étant donc égale à l'identité.

Sur ce petit exemple, on notera les quelques points suivants :

- le pas de gradient stochastique $\epsilon^{(k)}$ tend vers zéro lorsque k tend vers l'infini, alors que le pas d'un algorithme de gradient classique est constant ; cependant, $\epsilon^{(k)}$ ne doit pas tendre trop vite vers zéro : il correspond ici au terme d'une série divergente [3] ;
- la convergence de l'algorithme de gradient stochastique est celle de la loi des grands nombres, c'est-à-dire la convergence presque-sûre ; c'est donc la notion de convergence à laquelle on peut s'attendre dans l'étude théorique du gradient stochastique ;
- on trouve en statistique, en plus de la loi des grands nombres qui renseigne sur la convergence, le théorème de la limite centrale qui donne des indications sur la vitesse de convergence de l'estimation ; on peut donc aussi espérer obtenir un résultat de ce type dans le cadre du gradient stochastique.

2.2.3 Cadre probabiliste

Une itération de la méthode du gradient stochastique (2.5) peut se mettre sous la forme générale suivante :

$$\boldsymbol{U}^{(k+1)} = \mathcal{R}^{(k)}\left(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}\right) \ . \tag{2.8}$$

On suppose que la variable aléatoire $\boldsymbol{U}^{(0)}$ est constante, égale à $u^{(0)} \in U^{\mathrm{ad}}$.

- On définit les sous-tribus $\mathcal{F}^{(k)}$ de la tribu $\mathcal{A}$ engendrées par la collection des variables aléatoires $\boldsymbol{W}^{(k)}$:

$$\mathcal{F}^{(0)} = \{\emptyset, \Omega\} \quad , \quad \mathcal{F}^{(k)} = \sigma\left(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(k)}\right) \ .$$

  La suite $\{\mathcal{F}^{(k)}\}_{k\in\mathbb{N}}$ vérifie la propriété d'inclusion $\mathcal{F}^{(k)} \subset \mathcal{F}^{(k+1)}$ et est donc une filtration.
- L'utilisation récursive de la relation (2.8) montre que la variable aléatoire $\boldsymbol{U}^{(k)}$ ne dépend que des variables aléatoires $\boldsymbol{W}^{(l)}$, avec $l \leq k$. Supposant cette dépendance mesurable, on en déduit que chaque variable aléatoire $\boldsymbol{U}^{(k)}$ est $\mathcal{F}^{(k)}$-mesurable, et on a donc :

$$\mathbb{E}\left(\boldsymbol{U}^{(k)} \mid \mathcal{F}^{(k)}\right) = \boldsymbol{U}^{(k)} \ .$$

3. Demander que $\epsilon^{(k)}$ soit le terme d'une série *convergente* serait irréaliste car on construirait alors très facilement des exemples pour lesquels l'algorithme convergerait vers une valeur dépendant de la suite $\epsilon^{(k)}$ et du point initial $u^{(0)}$.

– Définissant la fonction $\varphi^{(k)}$ de la manière suivante :

$$\varphi^{(k)}(u) = \mathbb{E}\left(\mathcal{R}^{(k)}(u, \boldsymbol{W})\right) ,$$

utilisant le fait que les variables aléatoires $\boldsymbol{W}^{(k)}$ sont indépendantes et que les variables aléatoires $\boldsymbol{U}^{(k)}$ sont $\mathcal{F}^{(k)}$-mesurables, on a que :

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{U}^{(k+1)} \mid \mathcal{F}^{(k)}\right) &= \mathbb{E}\left(\mathcal{R}^{(k)}(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \mid \mathcal{F}^{(k)}\right) \\ &= \varphi^{(k)}(\boldsymbol{U}^{(k)}) ,\end{aligned}$$

ce qui s'écrit encore pour presque tout $\omega \in \Omega$:

$$\mathbb{E}\left(\boldsymbol{U}^{(k+1)} \mid \mathcal{F}^{(k)}\right)(\omega) = \int_{\Omega} \mathcal{R}^{(k)}(\boldsymbol{U}^{(k)}(\omega), \boldsymbol{W}(\omega'))\mathrm{d}\mathbb{P}\omega' .$$

Cette dernière relation traduit le fait que l'espérance conditionnelle de $\boldsymbol{U}^{(k+1)}$ par rapport à $\mathcal{F}^{(k)}$ se calcule en fait comme une simple espérance.

– Comme on l'a noté dans l'exemple page 21, la notion de convergence adaptée à l'étude de la suite générée par la relation (2.8) est celle de la convergence presque sûre :

$$\lim_{k \to +\infty} \mathbb{P}\left(\sup_{m \geq k} \left\| \boldsymbol{U}^{(m)} - u^{\sharp} \right\| > \epsilon\right) = 0 \quad \forall \epsilon > 0 .$$

On rappelle que la convergence presque sûre d'une suite de variables aléatoires $\{\boldsymbol{U}^{(k)}\}_{k \in \mathbb{N}}$ vers une valeur constante $u^{\sharp}$ s'interprète intuitivement de la manière suivante : presque toutes les fois que l'on applique l'algorithme (i.e. pour tout $\omega \in \Omega$ à l'exception d'un ensemble de mesure nulle), la suite des valeurs $\boldsymbol{U}^{(k)}(\omega)$ engendrée par l'algorithme converge vers $u^{\sharp}$.

De nombreux ouvrages présentent les outils probabilistes utilisés dans le cadre de ce cours. On consultera par exemple le livre de Bouleau (1986), ou encore celui de Dacunha-Castelle et Duflo (1994).

2.3 Premiers résultats

On rappelle que le problème que l'on veut résoudre par l'algorithme du gradient stochastique est :

$$\min_{u \in U^{\mathrm{ad}}} \mathbb{E}\left(j(u, \boldsymbol{W})\right) , \tag{2.9}$$

où $\boldsymbol{W}$ est une variable aléatoire définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{W}$ et où U^{ad} est une partie convexe fermée d'un espace de Hilbert $\mathbb{U}$,

Reprenant les travaux de Polyak (1976) et Dodu et collab. (1981), on dispose d'un premier théorème de convergence de l'algorithme du gradient stochastique, qui a l'avantage de pouvoir être démontré avec des arguments élémentaires. On donne pour commencer la définition suivante.

Définition 2.3. *On dit qu'une suite de réels positifs* $\left\{\epsilon^{(k)}\right\}_{k\in\mathbb{N}}$ *est une* σ*-suite si la série qu'elle engendre est divergente, la série de ses carrés étant quant à elle convergente :*

$$\sum_{k\in\mathbb{N}} \epsilon^{(k)} = +\infty \quad , \quad \sum_{k\in\mathbb{N}} \left(\epsilon^{(k)}\right)^2 < +\infty \; . \tag{2.10}$$

On fait alors les hypothèses suivantes.

Hypothèses 2.4.

1. La variable aléatoire $j(u, \boldsymbol{W}) : \Omega \to \mathbb{R}$ est mesurable et son espérance existe pour tout $u \in U^{\mathrm{ad}}$.
2. La fonction $j(\cdot, w) : \mathbb{U} \to \mathbb{R}$ est propre, convexe, s.c.i. (semi-continue inférieurement), différentiable pour tout $w \in \mathbb{W}$.
3. Le gradient partiel de j par rapport à u est borné uniformément en u et en w :
$$\exists m > 0,\ \forall u \in U^{\mathrm{ad}},\ \forall w \in \mathbb{W},\ \left\|\nabla_u j(u, w)\right\| \leq m \; .$$
4. Le problème (2.9) admet un ensemble de solutions $U^\sharp$ non vide, qui vérifie la relation :
$$\forall u \in U^{\mathrm{ad}},\ J(u) - J^\sharp \geq c \left(\mathrm{dist}_{U^\sharp}(u)\right)^2 ,$$
où $J^\sharp$ est la valeur du minimum de (2.9) et où $\mathrm{dist}_{U^\sharp}(.)$ est la fonction distance à l'ensemble $U^\sharp$.
5. La suite $\left\{\epsilon^{(k)}\right\}_{k\in\mathbb{N}}$ est une σ-suite décroissante.

2.3.1 Convergence

On dispose d'un premier résultat de convergence en moyenne quadratique de l'algorithme du gradient stochastique.

Théorème 2.5. ***(Convergence en moyenne quadratique)***
Sous les hypothèses 2.4, la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ *de variables aléatoires générée par l'algorithme 2.1 converge en moyenne quadratique vers l'ensemble* $U^\sharp$ *:*

$$\lim_{k\to+\infty} \mathbb{E}\left(\mathrm{dist}_{U^\sharp}(\boldsymbol{U}^{(k)})^2\right) = 0 \; .$$

Preuve. L'ensemble $U^\sharp$ étant convexe fermé, la projection sur cet ensemble est bien définie. Soit $\{u^{(k)}\}_{k\in\mathbb{N}}$ une réalisation de l'algorithme 2.1 et soit $\overline{u}^{(k)}$ la projection de $u^{(k)}$ sur $U^\sharp$:

$$\mathrm{dist}_{U^\sharp}\big(u^{(k)}\big)^2 = \big\|u^{(k)} - \overline{u}^{(k)}\big\|^2 .$$

Notant $d^{(k)} = \mathrm{dist}_{U^\sharp}\big(u^{(k)}\big)^2$, utilisant le fait que la projection est contractante et l'hypothèse 2.4-3, on a :

$$\begin{aligned} d^{(k+1)} &\leq \big\|u^{(k+1)} - \overline{u}^{(k)}\big\|^2 \\ &\leq \big\|\mathrm{proj}_{U^{\mathrm{ad}}}\big(u^{(k)} - \epsilon^{(k)}\nabla_u j(u^{(k)}, w^{(k+1)})\big) - \overline{u}^{(k)}\big\|^2 \\ &\leq \big\|u^{(k)} - \epsilon^{(k)}\nabla_u j(u^{(k)}, w^{(k+1)}) - \overline{u}^{(k)}\big\|^2 \\ &\leq d^{(k)} + {\epsilon^{(k)}}^2 m^2 - 2\epsilon^{(k)}\big\langle u^{(k)} - \overline{u}^{(k)}\,, \nabla_u j(u^{(k)}, w^{(k+1)})\big\rangle . \end{aligned}$$

Avec des notations évidentes, cette relation s'écrit en terme de variables aléatoires :

$$\boldsymbol{D}^{(k+1)} \leq \boldsymbol{D}^{(k)} + {\epsilon^{(k)}}^2 m^2 - 2\epsilon^{(k)}\big\langle \boldsymbol{U}^{(k)} - \overline{\boldsymbol{U}}^{(k)}\,, \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})\big\rangle .$$

Prenant de part et d'autre de cette inégalité l'espérance conditionnelle par rapport à la sous-tribu $\mathcal{F}^{(k)} = \sigma(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(k)})$, utilisant ensuite les propriétés de mesurabilité des variables aléatoires ainsi que le fait que l'on ait $\mathbb{E}\left(\nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \mid \mathcal{F}^{(k)}\right) = \nabla J(\boldsymbol{U}^{(k)})$, et enfin la convexité de J ainsi que l'hypothèse 2.4-4, on obtient [4] :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{D}^{(k+1)} \;\big|\; \mathcal{F}^{(k)}\big) &\leq \boldsymbol{D}^{(k)} + \big(\epsilon^{(k)}\big)^2 m^2 - 2\epsilon^{(k)}\big\langle \boldsymbol{U}^{(k)} - \overline{\boldsymbol{U}}^{(k)}\,, \nabla J(\boldsymbol{U}^{(k)})\big\rangle \\ &\leq \boldsymbol{D}^{(k)} + \big(\epsilon^{(k)}\big)^2 m^2 - 2\epsilon^{(k)}\big(J(\boldsymbol{U}^{(k)}) - J^\sharp\big) \\ &\leq \big(1 - 2\epsilon^{(k)}c\big)\boldsymbol{D}^{(k)} + {\epsilon^{(k)}}^2 m^2 . \end{aligned}$$

Prenant l'espérance de cette dernière inégalité, il vient :

$$\mathbb{E}\big(\boldsymbol{D}^{(k+1)}\big) \leq \big(1 - 2\epsilon^{(k)}c\big)\mathbb{E}\big(\boldsymbol{D}^{(k)}\big) + \big(\epsilon^{(k)}\big)^2 m^2 . \tag{2.11}$$

On montre alors par récurrence que, pour k donné suffisamment grand, on a :

$$\mathbb{E}\big(\boldsymbol{D}^{(k+n+1)}\big) \leq \left(\prod_{l=0}^{n}\big(1 - 2\epsilon^{(k+l)}c\big)\right)\mathbb{E}\big(\boldsymbol{D}^{(k)}\big) + \left(\sum_{l=0}^{n}\big(\epsilon^{(k+l)}\big)^2\right) m^2 \quad \forall n \in \mathbb{N} .$$

Comme la suite de terme général $\displaystyle\prod_{l=0}^{k}\big(1 - 2\epsilon^{(l)}c\big)$ converge vers zéro (voir la proposition 2.8 page 28) et que la suite de terme général $\displaystyle\sum_{l=0}^{k}\big(\epsilon^{(l)}\big)^2$ converge (hypothèse 2.4-5), on en déduit le résultat annoncé.

4. L'hypothèse 2.4-3 et le fait que $j(u, \boldsymbol{W}(.))$ soit intégrable impliquent, par un argument de convergence dominée, que le gradient de j par rapport à u est lui aussi intégrable.

2.3.2 Vitesse de convergence

On dispose en fait d'un résultat plus précis concernant la vitesse de décroissance en moyenne de la distance $\boldsymbol{D}^{(k)}$.

Théorème 2.6. ***(Vitesse de convergence en moyenne quadratique)*** *Sous les hypothèses du théorème 2.5, et en choisissant une suite* $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$ *de la forme :*

$$\epsilon^{(k)} = \frac{1}{c\, k + \frac{m^2}{c\, d^{(0)}}} \,,$$

(avec $d^{(0)} = \mathrm{dist}_{U^\sharp}\left(u^{(0)}\right)^2$*), on obtient la borne suivante sur la vitesse de convergence :*

$$\mathbb{E}\big(\mathrm{dist}_{U^\sharp}\big(\boldsymbol{U}^{(k)}\big)^2\big) \leq \frac{1}{\frac{c^2}{m^2}k + \frac{1}{d^{(0)}}} \qquad \forall k \in \mathbb{N}\,.$$

Preuve. On repart de l'inégalité (2.11) :

$$\mathbb{E}\big(\boldsymbol{D}^{(k+1)}\big) \leq \big(1 - 2\epsilon^{(k)}c\big)\mathbb{E}\big(\boldsymbol{D}^{(k)}\big) + {\epsilon^{(k)}}^2 m^2 \,,$$

et on choisit une suite $\epsilon^{(k)}$ de la forme :

$$\epsilon^{(k)} = \frac{\gamma}{\alpha k + \beta} \,,$$

On montre alors par récurrence que l'inégalité :

$$(\alpha k + \beta)\,\mathbb{E}\big(\boldsymbol{D}^{(k)}\big) \leq 1 \,,$$

est vérifiée avec les choix $\alpha = \dfrac{c^2}{m^2}$, $\beta = \dfrac{1}{d^{(0)}}$ et $\gamma = \dfrac{c}{m^2}$ (voir Dodu et collab. (1981) pour plus de détails).

2.3.3 Interprétation

On constate que, dans le cas où la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ converge vers un point $u^\sharp$, l'erreur quadratique moyenne $\mathbb{E}\big(\|\boldsymbol{U}^{(k)} - u^\sharp\|^2\big)$ est asymptotiquement bornée, l'expression de la borne étant :

$$\frac{1}{k}\left(\frac{m}{c}\right)^2 \,. \tag{2.12}$$

D'après les hypothèses 2.4-3 et 2.4-4, les deux constantes m et c peuvent être considérées comme représentant respectivement une borne supérieure de la variance de la norme du gradient de la fonction j et une borne inférieure de la constante de forte convexité de la fonction J. Cette interprétation sera utilisée §2.5 pour la comparaison de différentes versions de l'algorithme du gradient stochastique.

2.3.4 Discussion

Ces résultats de convergence se trouvent dans Dodu et collab. (1981). On a choisi de les présenter car ils sont représentatifs des résultats dont on dispose sur le gradient stochastique (convergence et vitesse). Ils ne sont cependant pas entièrement satisfaisants, pour les raisons suivantes.

- Tout d'abord, l'hypothèse 2.4-3 de gradient uniformément borné n'est pas raisonnable dès que l'ensemble U^{ad} n'est pas lui-même borné, puisque qu'elle exclut par exemple le cas des fonctions j quadratiques en u.
- De plus, l'interprétation faite au §2.2 du calcul d'une espérance en tant qu'algorithme de gradient stochastique suggère que le type de convergence que l'on doit obtenir est la convergence presque sûre plutôt que la convergence en moyenne quadratique.

On trouve bien dans Dodu et collab. (1981) un théorème de convergence presque sûre ainsi que l'estimation de vitesse de convergence associée, mais ces résultats sont obtenus sous l'hypothèse 2.4-3.

On donnera au §3.2 un théorème de convergence presque sûre très général pour une famille d'algorithmes incluant l'algorithme 2.1. Dans ce théorème, établi dans le cadre convexe, l'hypothèse 2.4-3 de gradient de j par rapport à u borné uniformément en w sera remplacée par une hypothèse de gradient *linéairement* borné en u uniformément en w :

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

Une telle hypothèse n'est pas surprenante dans une méthode de type gradient et constitue en fait une extension au cadre stochastique de l'hypothèse classique de gradient lipschitzien. La démonstration du théorème correspondant fait appel à des outils probabilistes évolués comme la théorie des quasi-martingales et sera détaillée au §3.

2.3.5 Lemmes techniques

On a utilisé dans la preuve du théorème de convergence en moyenne quadratique deux propriétés, que l'on démontre maintenant.

Proposition 2.7. *L'opération de projection sur U^{ad} est contractante.*

Preuve. Soit u et v deux points quelconques. Par définition de la projection, on a :

$$\mathrm{proj}_{U^{\mathrm{ad}}}(u) = \min_{w \in U^{\mathrm{ad}}} \|w - u\|^2\ .$$

La condition d'optimalité de ce problème, évaluée au point $\mathrm{proj}_{U^{\mathrm{ad}}}(v)$, s'écrit :

$$\langle \mathrm{proj}_{U^{\mathrm{ad}}}(u) - u\ , \mathrm{proj}_{U^{\mathrm{ad}}}(v) - \mathrm{proj}_{U^{\mathrm{ad}}}(u) \rangle \geq 0\ .$$

Intervertissant les rôles de u et v, on obtient :

$$\langle \mathrm{proj}_{U^{\mathrm{ad}}}(v) - v \, , \mathrm{proj}_{U^{\mathrm{ad}}}(u) - \mathrm{proj}_{U^{\mathrm{ad}}}(v) \rangle \geq 0 \, .$$

Additionnant ces deux dernières inégalités, il vient :

$$\left\| \mathrm{proj}_{U^{\mathrm{ad}}}(u) - \mathrm{proj}_{U^{\mathrm{ad}}}(v) \right\|^2 \leq \left\langle u - v \, , \mathrm{proj}_{U^{\mathrm{ad}}}(u) - \mathrm{proj}_{U^{\mathrm{ad}}}(v) \right\rangle ,$$

ce qui permet, par application de l'inégalité de Schwartz, de conclure.

Proposition 2.8. *Soit* $\left\{\epsilon^{(k)}\right\}_{k\in\mathbb{N}}$ *une suite décroissante de réels positifs telle que* $\epsilon^{(k)} \to 0$ *et* $\sum \epsilon^{(k)} = +\infty$. *Alors, pour tout* $\alpha > 0$, *la suite de terme général* $\left\{\rho^{(k)}\right\}_{k\in\mathbb{N}}$ *avec :*

$$\rho^{(k)} = \prod_{l=1}^{k} (1 - \alpha\epsilon^{(l)}) \, ,$$

converge vers zéro.

Preuve. Notant k_0 le premier indice tel que l'on ait $0 \leq 1 - \alpha\epsilon^{(l)} \leq 1$ pour tout $l \geq k_0$ (cet indice existe car $\epsilon^{(k)} \to 0$), on se ramène, à une constante multiplicative près, au cas où le produit définissant le terme $\rho^{(k)}$ est pris entre k_0 et k. La suite $\left\{\rho^{(k)}\right\}_{k\in\mathbb{N}}$ est alors positive décroissante, et donc convergente. De plus, on a :

$$\log(\rho^{(k)}) = \sum_{l=k_0}^{k} \log(1 - \alpha\epsilon^{(l)}) \leq -\alpha \sum_{l=k_0}^{k} \epsilon^{(l)} \, .$$

Par l'hypothèse de série divergente sur $\epsilon^{(k)}$, on conclut que la suite $\left\{\rho^{(k)}\right\}_{k\in\mathbb{N}}$ converge vers zéro.

2.4 Lien avec l'approximation stochastique

Un problème classique étudié dans le cadre de l'approximation stochastique (*Stochastic Approximation* ou SA en anglais) est de déterminer les zéros d'une fonction lorsque l'on ne dispose que d'évaluations *bruitées* de cette fonction. Dans ce cadre, on note $\mathbb{U}$ l'espace de Hilbert $\mathbb{R}^n$ et on considère une fonction $h : \mathbb{U} \to \mathbb{U}$, dont l'observation est perturbée de manière additive par une variable aléatoire $\boldsymbol{\xi}$. La méthode de l'approximation stochastique consiste à déterminer un zéro de la fonction h en utilisant la formule itérative suivante :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} + \epsilon^{(k)} \left(h(\boldsymbol{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)} \right) . \tag{2.13}$$

Cet algorithme est très fortement lié à celui du gradient stochastique : dans le cas où l'ensemble U^{ad} est l'espace $\mathbb{U}$ tout entier, la projection sur U^{ad} correspond à l'identité et l'algorithme du gradient stochastique 2.1 s'écrit alors :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \,. \tag{2.14}$$

Définissant la fonction h et les variables aléatoires $\boldsymbol{\xi}^{(k+1)}$ par

$$h(u) = -\nabla J(u) \,, \tag{2.15a}$$
$$\boldsymbol{\xi}^{(k+1)} = \nabla J(\boldsymbol{U}^{(k)}) - \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \,, \tag{2.15b}$$

la formule de mise à jour du gradient stochastique (2.14) est identique à celle de l'approximation stochastique (2.13). On notera que trouver un point $u^\sharp \in \mathbb{U}$ tel que $h(u^\sharp) = 0$ est équivalent à résoudre l'équation $\nabla J(u^\sharp) = 0$ et donc revient à résoudre la condition nécessaire d'optimalité du problème (2.2).

On va présenter deux résultats classiques de la théorie de l'approximation stochastique concernant la convergence et la vitesse de convergence de la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ engendrée par (2.13). Dans ce cadre, la suite $\{\boldsymbol{\xi}^{(k)}\}_{k\in\mathbb{N}}$ des variables aléatoires bruitant l'observation de h constitue une donnée du problème, et on se donne de plus une filtration $\{\mathcal{F}^{(k)}\}_{k\in\mathbb{N}}$.

2.4.1 Théorème de Robbins-Monro

On s'intéresse d'abord à la convergence de la suite de variables aléatoires $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ générée par (2.13). Pour cela, on fait les hypothèses suivantes.

Hypothèses 2.9.

1. La variable aléatoire $\boldsymbol{U}^{(0)}$ est $\mathcal{F}^{(0)}$-mesurable.
2. La fonction $h : \mathbb{U} \to \mathbb{U}$ est continue et vérifie les propriétés suivantes :
   – $\exists\, u^\sharp \in \mathbb{U},\ h(u^\sharp) = 0$ et $\langle h(u)\,, u - u^\sharp \rangle < 0,\ \forall u \neq u^\sharp$;
   – $\exists\, a > 0,\ \forall u \in \mathbb{U},\ \|h(u)\|^2 \leq a\big(1 + \|u\|^2\big)$.
3. La variable aléatoire $\boldsymbol{\xi}^{(k)}$ est $\mathcal{F}^{(k)}$-mesurable quel que soit k, et l'on a :
   – $\mathbb{E}\left(\boldsymbol{\xi}^{(k+1)} \mid \mathcal{F}^{(k)}\right) = 0$,
   – $\exists\, d > 0,\ \mathbb{E}\big(\left\|\boldsymbol{\xi}^{(k+1)}\right\|^2 \mid \mathcal{F}^{(k)}\big) \leq d\big(1 + \left\|\boldsymbol{U}^{(k)}\right\|^2\big)$.
4. La suite $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$ est une σ-suite.

On remarquera que l'hypothèse 2.9-2 implique que $u^\sharp$ est l'unique zéro de la fonction h.

Théorème 2.10. *Sous les hypothèses 2.9, la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ de variables aléatoires engendrées par (2.13) converge presque sûrement vers $u^\sharp$.*

Preuve. Voir (Duflo, 1997, §1.4).

On peut faire le lien entre les hypothèses du théorème 2.10 et celles que l'on pourrait formuler pour obtenir la solution du problème (2.3) dans le cadre de l'optimisation convexe. On suppose d'abord que chaque σ-algèbre $\mathcal{F}^{(k)}$ est engendrée par $(\boldsymbol{W}^{(0)}, \ldots, \boldsymbol{W}^{(k)})$, et l'on déduit alors de (2.15) que chaque variable aléatoire $\boldsymbol{\xi}^{(k)}$ est $\mathcal{F}^{(k)}$-mesurable. Si l'on suppose que la fonction j

est strictement convexe, coercive, continûment différentiable par rapport à u, et mesurable par rapport à w, alors la fonction J est strictement convexe, coercive et continûment différentiable. La première partie de l'hypothèse 2.9-2 en découle (existence et unicité de la solution du problème (2.3)). De même, la première partie de l'hypothèse 2.9-3 est une conséquence immédiate de (2.15). Pour ce qui concerne la seconde partie des hypothèses 2.9-2 et 2.9-3, elles sont reliées à l'hypothèse de *gradient linéairement borné* (GLB) sur la fonction j, à savoir :

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

hypothèse qui implique (par la propriété $(a+b)^2 \leq 2(a^2+b^2)$)

$$\begin{aligned} \exists\, c_3 > 0,\ c_4 > 0,\ \forall u \in \mathbb{R}^n,\ \forall w \in \mathbb{W},\ \|\nabla_u j(u,w)\|^2 &\leq c_3 \|u\|^2 + c_4 \ , \\ \|\nabla J(u)\|^2 &\leq c_3 \|u\|^2 + c_4 \ . \end{aligned}$$

On notera que les hypothèses faites sur la fonction j semblent naturelles dans le cadre de l'optimisation convexe. On donnera un résultat de convergence plus général de l'algorithme de gradient stochastique au §3.

2.4.2 Normalité asymptotique

On donne maintenant un résultat de type « théorème de la limite centrale » précisant la normalité asymptotique des itérées $\boldsymbol{U}^{(k)}$ de l'algorithme défini par (2.13). Ce résultat permettra de comparer la vitesse de convergence de différentes mises en œuvre des algorithmes de type gradient stochastique. On a alors besoin de donner une définition plus précise de la notion de σ-suite.

Définition 2.11. *Une suite de réels positifs* $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$ *est une* $\sigma(\alpha,\beta,\gamma)$-suite *si elle est telle que :*

$$\epsilon^{(k)} = \frac{\alpha}{k^\gamma + \beta} \ ,$$

avec $\alpha > 0$, $\beta \geq 0$ *and* $1/2 < \gamma \leq 1$.

Une conséquence immédiate de cette définition est qu'une $\sigma(\alpha,\beta,\gamma)$-suite est aussi une σ-suite.

Pour l'étude de vitesse de convergence, on ajoute aux hypothèses 2.9 déjà faites pour l'étude de la convergence les nouvelles hypothèses suivantes.

Hypothèses 2.12.

1. La fonction h est continûment différentiable et s'exprime sous la forme suivante dans un voisinage du point $u^\sharp$:

$$h(u) = -H(u-u^\sharp) + \mathrm{O}(\left\|u-u^\sharp\right\|^2) \ ,$$

où la matrice H est symétrique définie positive[5].

2. La suite $\{\mathbb{E}(\boldsymbol{\xi}^{(k+1)}(\boldsymbol{\xi}^{(k+1)})^\top \mid \mathcal{F}^{(k)})\}_{k\in\mathbb{N}}$ des matrices de covariance conditionnelle des $\boldsymbol{\xi}^{(k)}$ converge presque sûrement vers une matrice symétrique définie positive déterministe Γ.
3. Il existe $\delta > 0$ tel que $\sup_{k\in\mathbb{N}} \mathbb{E}(\|\boldsymbol{\xi}^{(k+1)}\|^{2+\delta} \mid \mathcal{F}^{(k)}) < +\infty$.
4. La suite $\{\epsilon^{(k)}\}_{k\in\mathbb{N}}$ est une $\sigma(\alpha,\beta,\gamma)$-suite.
5. La matrice $H - \lambda I$ est définie positive, le coefficient λ étant défini par :

$$\lambda = \begin{cases} 0 & \text{si } \gamma < 1 \\ \frac{1}{2\alpha} & \text{si } \gamma = 1 \end{cases} \tag{2.16}$$

On notera que, dans le cadre du problème d'optimisation initial (2.3) pour lequel on a $h = -\nabla J$, la matrice H correspond à la matrice hessienne de J au point $u^\sharp$:

$$H = \nabla^2 J(u^\sharp) \, .$$

De plus, puisque l'on a alors $\mathbb{E}\left(\nabla_u j(u^\sharp, \boldsymbol{W})\right) = 0$, la matrice Γ de l'hypothèse 2.12-2 correspond quand à elle à la matrice de covariance du gradient partiel de la fonction j au point $u^\sharp$:

$$\Gamma = \mathbb{E}\big(\nabla_u j(u^\sharp, \boldsymbol{W})(\nabla_u j(u^\sharp, \boldsymbol{W}))^\top\big) \, .$$

On a alors le théorème suivant précisant la vitesse à laquelle les itérées $\boldsymbol{U}^{(k)}$ générées par (2.13) convergent vers $u^\sharp$.

Théorème 2.13. ***(Théorème de la limite centrale)***
Sous les hypothèses 2.9 et 2.12, la suite $\big\{(\epsilon^{(k)})^{-\frac{1}{2}}\big(\boldsymbol{U}^{(k)} - u^\sharp\big)\big\}_{k\in\mathbb{N}}$ *converge en distribution vers la loi normale centrée de matrice de covariance* Σ *:*

$$\frac{1}{\sqrt{\epsilon^{(k)}}}\Big(\boldsymbol{U}^{(k)} - u^\sharp\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \Sigma\big) \, , \tag{2.17}$$

la matrice de covariance Σ *étant solution de l'équation de Lyapunov :*

$$\big(H - \lambda I\big)\Sigma + \Sigma\big(H - \lambda I\big) = \Gamma \, . \tag{2.18}$$

Preuve. Voir (Duflo, 1996, Ch. 4).

On rappelle le résultat classique caractérisant la solution d'une équation de Lyapunov. Ce résultat peut être trouvé dans (Khalil, 2002, Theorem 4.6).

5. Le symbole O correspond à la notation « Grand O » : $f(x) = \mathrm{O}\big(g(x)\big)$ quand $x \to x_0$ si et seulement si il existe une constante positive α et un voisinage V de x_0 tels que $|f(x)| \leq \alpha\,|g(x)|$, $\forall x \in V$.

Proposition 2.14. *Soit H une matrice définie positive et Γ une matrice symétrique définie positive de même dimension. Alors, il existe une matrice Σ symétrique définie positive, solution unique de l'équation de Lyapunov :*

$$H\Sigma + \Sigma H^\top = \Gamma \, ,$$

et cette solution a pour expression :

$$\Sigma = \int_0^{+\infty} e^{-tH} \Gamma e^{-tH^\top} \mathrm{d}t \, . \tag{2.19}$$

Remarque 2.15. Ce résultat reste vrai si la matrice Γ est symétrique semi-définie positive : dans ce cas, la matrice Σ donnée par (2.19) est elle aussi semi-définie positive, et est solution de l'équation (2.19).

Utilisant explicitement le fait que, dans le théorème 2.13, les pas $\epsilon^{(k)}$ forment une $\sigma(\alpha, \beta, \gamma)$-suite, on tire les conclusions suivantes quant à l'influence des coefficients α, β et γ sur la convergence de l'algorithme.

1. Le résultat de convergence donné par le théorème 2.13 se réécrit sous la forme :
$$k^{\frac{\gamma}{2}} \left(\boldsymbol{U}^{(k)} - u^\sharp \right) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \alpha\Sigma\big) \, . \tag{2.20}$$
On constate que le coefficient β n'a aucune influence sur le comportement asymptotique de l'algorithme [6].
2. De la relation (2.20), on déduit que le choix de γ qui conduit à la vitesse de convergence la plus élevée est $\gamma = 1$. On retrouve ainsi la vitesse classique en $1/\sqrt{k}$ d'un estimateur de type Monte Carlo.
3. Le coefficient α doit être choisi de telle sorte que la matrice de covariance $\alpha\Sigma$ soit aussi petite que possible (au sens de l'ordre sur les matrices définies positives). Le raisonnement simpliste consistant à prendre un α aussi petit que possible pour diminuer la covariance asymptotique dans la relation (2.20) ne tient pas. En effet, la solution Σ de l'équation de Lyapunov (2.18) dépend de λ, et donc de α, de telle sorte que la matrice de covariance $\alpha\Sigma$ ne varie ni linéairement, ni de façon monotone, avec le coefficient α. Ainsi, dans le cas scalaire ($\mathbb{U} = \mathbb{R}$), H et Γ sont des réels et la solution de l'équation de Lyapunov (2.18) est :
$$\Sigma = \frac{\alpha\Gamma}{2\alpha H - 1} \, .$$
On peut facilement minimiser la variance $\alpha\Sigma$ par rapport à α, le minimum étant atteint pour la valeur $\alpha^\sharp = 1/H$. Cette valeur vérifie bien la condition : $2\alpha^\sharp H - 1 > 0$ imposée par l'hypothèse 2.12-5.

6. Ce coefficient β a bien sûr une influence dans la phase transitoire de l'algorithme...

On se place donc dans le cas optimal $\gamma = 1$. Il reste maintenant à rendre aussi petite que possible (au sens des matrices définies positives) la matrice de covariance $\alpha\Sigma$ dans la relation (2.20). On verra au prochain paragraphe qu'une manière de réduire la variance de l'algorithme du gradient stochastique est de considérer des algorithmes à gain matriciel plutôt qu'à gain scalaire.

2.5 Efficacité asymptotique et moyennisation

En optimisation déterministe, il est bien connu qu'une amélioration du comportement des algorithmes à direction de descente est obtenue en pré-multipliant le gradient par une matrice bien choisie ; dans le cas où cette matrice est identique à l'inverse du Hessien, on obtient l'algorithme de Newton, dont la vitesse de convergence est quadratique dans un voisinage de la solution optimale. On ne peut bien sûr pas espérer un tel résultat dans le cadre du gradient stochastique car on a vu que les pas $\epsilon^{(k)}$ devaient tendre vers zéro avec l'indice k. On peut cependant espérer une amélioration de la méthode si l'on effectue un pré-conditionnement du gradient.

Pour appliquer cette idée à l'algorithme du gradient stochastique, on se donne une matrice A carrée de dimension d symétrique définie positive. On garde dans la forme des pas $\epsilon^{(k)}$ le coefficient $\gamma = 1$ conduisant à la vitesse optimale, mais on remplace le gain scalaire α par le gain matriciel A, ce qui conduit à substituer dans l'algorithme (2.14) l'itération courante de gradient stochastique par la nouvelle relation :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} - \frac{A}{k+\beta}\, \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \,,$$

ou encore, dans le formalisme de l'approximation stochastique :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} + \frac{A}{k+\beta}\big(h(\boldsymbol{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)}\big) \,. \tag{2.21}$$

On est donc dans le cadre de l'approximation stochastique, avec un champ de vecteurs Ah, des bruits $A\boldsymbol{\xi}^{(k)}$ et des pas de taille $1/(k+\beta)$. Dans ce contexte, l'hypothèse 2.12-5 devient :

Hypothèses 2.16.

La matrice $AH - \dfrac{I}{2}$ est définie positive.

Le théorème 2.13 s'applique alors et implique pour la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ générée par l'algorithme à gain matriciel (2.21) le résultat de convergence suivant :

$$\sqrt{k}\big(\boldsymbol{U}^{(k)} - u^\sharp\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \Sigma_A\big) \,, \tag{2.22}$$

la matrice de covariance asymptotique Σ_A étant donnée par :

$$\big(AH - \frac{I}{2}\big)\Sigma_A + \Sigma_A\big(HA - \frac{I}{2}\big) = A\Gamma A\,. \tag{2.23}$$

Soit $\mathcal{C}_H$ l'ensemble des matrices A symétriques définies positives telles que la matrice $AH - I/2$ soit elle aussi définie positive. Le théorème suivant caractérise le choix optimal du gain matriciel dans l'algorithme de gradient stochastique.

Théorème 2.17. ***(Algorithme de Newton stochastique)*** *Le choix $A^\sharp = H^{-1}$ comme gain dans la relation (2.21) minimise la variance asymptotique Σ_A définie par (2.23) sur l'ensemble $\mathcal{C}_H$, l'expression de la covariance optimale étant alors :*

$$\Sigma_{A^\sharp} = H^{-1}\Gamma H^{-1}\,.$$

Preuve. La matrice de covariance Σ_A du théorème 2.13 correspondant à l'algorithme (2.21) peut toujours se mettre sous la forme :

$$\Sigma_A = \Delta_A + H^{-1}\Gamma H^{-1}\,.$$

Reportant cette expression dans (2.23), on obtient :

$$\big(AH - \frac{I}{2}\big)\Delta_A + \Delta_A\big(HA - \frac{I}{2}\big) = \big(A - H^{-1}\big)\Gamma\big(A - H^{-1}\big)\,.$$

La matrice Δ_A vérifie donc une équation de Lyapunov et est, d'après la proposition 2.14 et la remarque 2.15, semi définie positive pour tout $A \in \mathcal{C}_H$. Comme $\Delta_A = 0$ pour $A = H^{-1}$, on en déduit que $\Sigma_A \geq H^{-1}\Gamma H^{-1}$ pour tout $A \in \mathcal{C}_H$, l'égalité étant obtenue pour la valeur $A^\sharp = H^{-1}$.

Remarque 2.18. Le gain H^{-1} correspond à l'inverse de la matrice hessienne de la fonction J évaluée au point $u^\sharp$ dans le cas du gradient stochastique, d'où le nom « algorithme de Newton stochastique » donné à l'algorithme (2.21) avec ce choix optimal de gain. Bien sûr, les pas utilisés dans l'algorithme stochastique doivent être de longueur $1/k$ alors qu'il sont de longueur 1 dans l'algorithme de Newton déterministe. Dans le cas stochastique, les méthodes à gain scalaire et à gain matriciel ont toutes les deux une vitesse de convergence de type $a/\sqrt{k}$. L'amélioration apportée par le gain matriciel est due à la constante multiplicative (i.e. la matrice de covariance) et non à la vitesse en $1/\sqrt{k}$. Si l'on note c la plus petite valeur propre de la matrice H et m un majorant de la norme du gradient de j, la plus grande valeur propre de la matrice $H^{-1}\Gamma H^{-1}$ est de l'ordre du quotient $(m/c)^2$, qui correspond au coefficient sur la vitesse donnée par le théorème 2.6 pour l'algorithme à gain scalaire.

On donne alors la définition suivante pour caractériser les algorithmes ayant le même comportement asymptotique que l'algorithme de Newton stochastique.

Définition 2.19. *Un algorithme de gradient stochastique est dit Newton-efficace si la suite* $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ *qu'il engendre a la même vitesse de convergence asymptotique que celle de l'algorithme de Newton stochastique, à savoir :*

$$\sqrt{k}\big(\boldsymbol{U}^{(k)} - u^\sharp\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, H^{-1}\Gamma H^{-1}\big) \ .$$

Comme on vient de le voir, l'algorithme de Newton stochastique est en un certain sens optimal dans la classe des algorithmes de type gradient. La question qui se pose alors est : *comment mettre en œuvre un algorithme de Newton-efficace ?* La difficulté vient du fait que l'algorithme à gain matriciel optimal ne peut pas être directement mis en œuvre car son H^{-1} dépend du point $u^\sharp$ que l'on cherche ! Plutôt que de proposer des algorithmes approximant la matrice H^{-1} au cours des itérations, on va donner au paragraphe suivant une technique de moyennisation permettant d'obtenir un algorithme Newton-efficace.

2.5.1 Moyennisation

Afin de contourner la difficulté de mise en œuvre d'un algorithme stochastique Newton-efficace, B. T. Polyak a proposé dans Polyak (1990) de modifier l'algorithme standard en lui ajoutant une étape de *moyennisation*. Cette modification consiste à remplacer, dans le cas où l'ensemble U^{ad} est l'espace $\mathbb{U}$ tout entier, la phase de mise à jour classique :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \ .$$

par le calcul en deux étapes suivant :

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) \ , \tag{2.24a}$$

$$\boldsymbol{U}_{\mathrm{M}}^{(k+1)} = \frac{1}{k+1} \sum_{l=1}^{k+1} \boldsymbol{U}^{(l)} \ , \tag{2.24b}$$

dans lequel la première étape (2.24a) est identique à celle du gradient stochastique classique, la deuxième étape (2.24b) consistant à former la *moyenne arithmétique* des variables aléatoires obtenues à la première étape. Lors de la mise en œuvre de l'algorithme, on utilise plutôt la forme récursive de l'étape (2.24b) :

$$\boldsymbol{U}_{\mathrm{M}}^{(k+1)} = \boldsymbol{U}_{\mathrm{M}}^{(k)} + \frac{1}{k+1}\big(\boldsymbol{U}^{(k+1)} - \boldsymbol{U}_{\mathrm{M}}^{(k)}\big) \ . \tag{2.24c}$$

On remarquera que, par le théorème de Césaro, la convergence presque sûre de la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ implique la convergence de la suite moyennée $\{\boldsymbol{U}_{\mathrm{M}}^{(k)}\}_{k\in\mathbb{N}}$. Sous les conditions du théorème 2.10, et en particulier avec des pas $\epsilon^{(k)}$ de la forme :

$$\epsilon^{(k)} = \frac{\alpha}{k^\gamma + \beta}\ , \quad \text{avec} \quad \frac{1}{2} < \gamma \leq 1\ ,$$

on sait que la suite $\{\boldsymbol{U}_{\mathrm{M}}^{(k)}\}_{k\in\mathbb{N}}$ converge vers la solution $u^\sharp$ du problème.

L'intérêt essentiel de l'algorithme moyenné (2.24) tient à ses propriétés asymptotiques. Les hypothèses que l'on fait alors sont semblables à celles ayant permis d'établir le théorème 2.13 de la limite centrale, mais on restreint l'hypothèse 2.12-4 au cas où le coefficient γ est *strictement* inférieur à 1.

Hypothèses 2.20.

La suite $\left\{\epsilon^{(k)}\right\}_{k\in\mathbb{N}}$ une une $\sigma(\alpha,\beta,\gamma)$-suite, avec $1/2 < \gamma < 1$.

Avec l'hypothèse $\gamma < 1$, la vitesse de convergence de la suite $\{\boldsymbol{U}^{(k)}\}_{k\in\mathbb{N}}$ est inférieure strictement à $1/\sqrt{k}$ d'après le théorème 2.13 et donc non optimale. C'est avec la suite $\{\boldsymbol{U}_{\mathrm{M}}^{(k)}\}_{k\in\mathbb{N}}$ obtenue *après moyennisation* que l'on obtient des propriétés de convergence intéressantes, comme le montre le théorème suivant.

Théorème 2.21. ***(Optimalité du gradient stochastique moyenné)*** *Sous les hypothèses 2.9 et 2.12, dans laquelle on remplace 2.12-4 par l'hypothèse 2.20, l'algorithme du gradient stochastique moyenné est Newton-efficace :*

$$\sqrt{k}\big(\boldsymbol{U}_{\mathrm{M}}^{(k)} - u^\sharp\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, H^{-1}\Gamma H^{-1}\big)\ .$$

Preuve. Voir (Duflo, 1996, Ch. 4).

Ce théorème présente un intérêt pratique certain puisqu'il montre que l'algorithme de gradient stochastique moyenné permet d'atteindre la matrice de covariance optimale de l'algorithme de Newton sans pour autant avoir à connaître à l'avance le gain optimal H^{-1}.

2.6 Considérations pratiques

La mise en œuvre d'un algorithme du gradient stochastique pose un certain nombre de difficultés pratiques, qu'il est essentiel de résoudre pour que la résolution du problème soit effectuée de manière satisfaisante.

2.6.1 Critère d'arrêt

Une première question porte sur les conditions d'arrêt de l'algorithme. Il est clair que le critère d'arrêt ne peut pas être basé sur l'écart $\left\|u^{(k+1)} - u^{(k)}\right\|$ car cette différence converge mécaniquement vers zéro par le biais des hypothèses faites sur les pas $\epsilon^{(k)}$. Par ailleurs, la norme du gradient partiel $\nabla_u j(u^{(k)}, w^{(k+1)})$ n'a elle non plus aucune bonne propriété de convergence. Par contre, l'espérance de la variable aléatoire $\nabla_u j(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)})$ converge

vers $\nabla J(u^\sharp)$ et peut donc servir pour effectuer un test de convergence. Comme on peut approximer cette espérance par :

$$\left(\sum_{l=1}^{k} \epsilon^{(l)}\right)^{-1} \left(\sum_{l=1}^{k} \epsilon^{(l)} \nabla_u j(u^{(l)}, w^{(l+1)})\right) ,$$

on doit être capable de construire un test d'arrêt raisonnable. En pratique, on se contentera souvent de fixer un nombre d'itérations suffisamment grand et de vérifier « visuellement » sur des graphiques représentant les itérées de l'algorithme que ce dernier converge de manière correcte.

2.6.2 Réglage de l'algorithme standard

La deuxième question porte sur la forme de la suite de pas $\left\{\epsilon^{(k)}\right\}_{k\in\mathbb{N}}$. On a vu lors de l'études de la vitesse de convergence qu'il est raisonnable de prendre des pas $\epsilon^{(k)}$ de la forme $\frac{1}{k^\gamma}$, avec $\frac{1}{2} < \gamma \leq 1$. D'après le théorème 2.13, la vitesse de convergence optimale est obtenue pour $\gamma = 1$. Mais le réglage du coefficient γ ne prend en compte qu'une partie du comportement asymptotique de l'algorithme. Le choix d'une $\sigma(\alpha, \beta, \gamma)$-suite permet alors de préciser le reste du comportement de l'algorithme. Les coefficients α et β paramétrant de telles suites sont choisis suivant les règles suivantes.

- le coefficient α a une influence sur la vitesse asymptotique de l'algorithme : son effet *multiplicatif* fait, d'une part que la matrice de covariance $\alpha\Sigma$ croît avec α, et d'autre part que choisir un α trop petit va réduire la taille du pas de gradient et donc ralentir la convergence de l'algorithme. Le choix de α doit donc résulter d'un compromis entre stabilité et précision.
- le coefficient β permet de régler les problèmes dans la phase transitoire de l'algorithme : au cours des premières itérations, si le terme k^γ est petit devant le terme *additif* β, le pas $\epsilon^{(k)}$ est approximativement égal au ratio α/β ; ce rapport sert donc à déterminer un pas « acceptable » en début d'algorithme : un pas trop petit pénalise la vitesse de convergence, alors qu'un pas trop grand provoque des explosions numériques durant les premières itérations.

En pratique, sur un ordinateur, les considérations précédentes sont plus utilisées en terme de ligne de conduite qu'en terme de règles. On trouve d'ailleurs un grand nombre d'articles décrivant des stratégies de mises à jour des pas $\epsilon^{(k)}$. On citera :

1. la méthode de projection de Chen Chen et collab. (1988) qui, en plus d'être un outil théorique permettant d'affaiblir les hypothèses nécessaires à la convergence des approximations stochastiques, permet d'un point de vue pratique d'éviter le phénomène d'explosion numérique dans la phase transitoire de l'algorithme en projetant les itérées $u^{(k)}$ sur des compacts formant une suite croissante dans l'espace $\mathcal{U}$;

2. l'algorithme de Kesten KESTEN (1958), dont l'idée générale est de faire décroître le pas du gradient stochastique seulement lorsque les directions de deux gradients successifs sont opposées ; pour cela, on définit la suite de variables aléatoires de nombres entiers $\boldsymbol{N}^k$ par la relation :

$$\boldsymbol{N}^{k+1} = \boldsymbol{N}^k + \mathbf{1}_{\left\{\left\langle \nabla_u j(\boldsymbol{U}^{(k-1)},\boldsymbol{W}^{(k)}),\nabla_u j(\boldsymbol{U}^{(k)},\boldsymbol{W}^{(k+1)})\right\rangle<0\right\}},$$

le dernier terme de la somme prenant la valeur 1 si le produit scalaire des deux gradients successifs est négatif et 0 sinon ; le pas de l'algorithme est alors défini par :

$$\boldsymbol{\epsilon}^{(k)} = \frac{\alpha}{\boldsymbol{N}_k^{\gamma} + \beta} ;$$

3. une règle multiplicative d'adaptation du pas PLAKHOV et CRUZ (2005), qui autorise une convergence rapide des itérées de l'algorithme, mais vers un point qui est alors une approximation de la solution recherchée.

En conclusion, on peut dire que la mise en œuvre d'un algorithme de gradient stochastique nécessite un certain nombre d'expérimentations numériques avant de donner des résultats satisfaisants. Une erreur classique est de penser que l'algorithme a convergé alors que la stabilisation est en fait due à une suite de pas $\epsilon^{(k)}$ mal choisie. Une bonne règle de conduite consiste à ne diminuer le pas $\epsilon^{(k)}$ que lorsque cela est nécessaire. Signalons enfin qu'il existe toute une littérature concernant les algorithmes stochastiques *à pas constant* (voir par exemple BENVENISTE et collab. (1990) ou VAZQUEZ-ABAD (2006)).

2.6.3 Réglage de l'algorithme moyenné

On a montré l'algorithme du gradient stochastique moyenné était Newton-efficace à la condition de choisir des pas $\epsilon^{(k)}$ formant une $\sigma(\alpha, \beta, \gamma)$-suite avec la condition $1/2 < \gamma < 1$. Le choix des coefficients α, β et γ se fait suivant les considérations suivantes.

- La valeur $\gamma = \frac{2}{3}$ est présentée par certains auteurs comme étant un bon choix pour l'exposant dans la formule des pas $\epsilon^{(k)}$ (voir B. Delyon DELYON (2000) pour plus de détails).
- Le réglage des paramètres α et β est beaucoup moins critique pour la « bonne convergence » dans l'algorithme moyenné que dans l'algorithme de gradient stochastique standard. Il faut cependant éviter les explosions numériques durant les premières itérations de l'algorithme.
- Plutôt que de moyenner dès la première itération, ce qui ralentit sensiblement l'algorithme durant sa phase transitoire, il est préférable de ne commencer le processus de moyennisation que lorsque le gradient stochastique (2.24a) s'est approché de la zone de convergence.

2.6.4 Illustration numérique

Ajouter l'exemple illustrant la convergence de l'algorithme du gradient stochastique et de la moyennisation ?

2.7 Conclusions

On a dans ce chapitre présenté brièvement les principales caractéristiques de l'algorithme du gradient stochastique, à savoir :

- sa convergence,
- son efficacité,
- sa moyennisation.

Dans les deux chapitres suivants (§3 et §4), on va étudier en détail la convergence du gradient stochastique, d'abord dans le cas sans contraintes, puis dans le cas des contraintes déterministes.

Puis, au dernier chapitre (§5), on présentera les extensions de la méthode du gradient stochastique au cas des contraintes en espérance, ainsi que son utilisation dans le cadre du Lagrangien augmenté.

Parler de tout le courant de littérature autour du "Machine Learning" concernant les algorithmes stochastiques (E. Moulines, F. Bach) ?