

2 Vue d'ensemble de la méthode du gradient stochastique

2.1 Présentation générale

La méthode du gradient stochastique telle qu'on l'a esquissée au §1.2.1 est relativement ancienne. Les initiateurs en sont H. Robbins et S. Monro [36] d'une part, J. Kiefer et J. Wolfowitz [21] d'autre part, dans le cadre général de l'approximation stochastique (voir [24] pour une mise en perspective de ces travaux). Plus récemment, B. T. Polyak [31]–[34] 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 [14] 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 [32]–[33] 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 [22] présentant, dans le cas non convexe, la méthode de l'équation différentielle moyenne (méthode de l'ODE dans la terminologie anglo-saxonne) permettant l'étude de la convergence locale des algorithmes stochastiques généraux. Plusieurs ouvrages, comme ceux de M. Duflo [15]–[16] et de H. J. Kushner et G. G. Yin [23] 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 [12], disponible sur le site Web de l'auteur et d'une lecture relativement aisée.

Le but de ce chapitre est de donner un panorama relativement complet des méthodes dont on dispose pour analyser le comportement d'un algorithme de gradient stochastique. On présentera :

- l'algorithme du gradient stochastique dans le cas le plus simple et le cadre probabiliste adapté à son étude ;
- une preuve élémentaire de convergence de cet algorithme (sous des hypothèses trop fortes), ainsi que des indications sur les théorèmes généraux de convergence ;
- le théorème de la limite centrale (TCL) associé à cet algorithme ainsi que ses variantes ;
- les principaux résultats concernant l'optimalité de la méthode, et le lien avec la borne de Cramer-Rao (voir [18] pour ce qui concerne l'estimation statistique) ;
- l'algorithme moyenné et son comportement asymptotique.

2.2 Algorithme du gradient stochastique

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

$$J(u) = \mathbb{E}(j(u, \mathbf{W})) = \int_{\Omega} j(u, \mathbf{W}(\omega)) d\mathbb{P}(\omega) = \int_{\mathbb{W}} j(u, w) d\mu(w),$$

et on s'intéresse au problème suivant :

$$\min_{u \in U^{\text{ad}}} J(u). \quad (10)$$

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 (10) un

algorithme de type gradient (à pas optimal, gradient conjugué, quasi-Newton, Newton, ...). 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. En fait, la structure probabiliste du problème n'est pas utilisée dans l'optimisation proprement dite, car totalement prise en compte lors du calcul du gradient.

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 de j évalué pour une valeur particulière w de la variable aléatoire \mathbf{W} , et non en fonction du gradient de l'espérance de j . C'est là une idée nouvelle par rapport aux méthodes d'optimisation classiques, car on cherche *simultanément* à faire progresser la méthode de gradient et à effectuer un calcul d'espérance. L'algorithme que l'on propose alors est le suivant.

Algorithme 1. (Algorithme du gradient stochastique)

1. Choisir une valeur initiale $u^{(0)} \in U^{\text{ad}}$ 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. Calculer le gradient de j par rapport à u au point $(u^{(k)}, w^{(k+1)})$ et remettre à jour u par :

$$u^{(k+1)} = \text{proj}_{U^{\text{ad}}} (u^{(k)} - \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)})) .$$
4. Incrémenter l'indice k de 1 et retourner à l'étape 2.

On notera que l'on n'a pas précisé de *critère d'arrêt* pour l'algorithme de gradient stochastique. Ce point sera précisé ultérieurement.

Il est commode, lorsque l'on étudie les propriétés de l'algorithme ci-dessus, de le décrire en terme de variables aléatoires. Il est en effet équivalent de considérer les tirages $(w^{(1)}, \dots, w^{(k)})$ comme étant k réalisations indépendantes de la même variable aléatoire \mathbf{W} , ou comme étant une réalisation de k variables aléatoires $(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)})$ indépendantes, de même loi que \mathbf{W} . L'étape 3 de remise à jour de u dans l'algorithme 1 peut alors être interprétée comme une relation de récurrence sur des variables aléatoires $\mathbf{U}^{(k)}$ à valeurs dans l'espace \mathbb{U} :

$$\mathbf{U}^{(k+1)} = \text{proj}_{U^{\text{ad}}} (\mathbf{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)})) , \quad (11)$$

les valeurs $u^{(k)}$ correspondant simplement à une réalisation des variables aléatoires $\mathbf{U}^{(k)}$:

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

Remarque 1. 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 $\{\mathbf{W}^{(k)}\}_{k \in \mathbb{N}^*}$ de variables aléatoires indépendantes et de même loi μ que \mathbf{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 $\mathbf{W}^{(k)}$ sont alors définies sur l'espace de probabilité $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$ comme étant les *applications coordonnées* :⁹

$$\mathbf{W}^{(k)}(w^{(1)}, \dots, w^{(k)}, \dots) = w^{(k)} .$$

Il faut faire attention à ce que l'on est appelé à manipuler par la suite *deux* espaces de probabilité, à savoir $(\Omega, \mathcal{A}, \mathbb{P})$ pour tout ce qui concerne la variable aléatoire \mathbf{W} , et $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$ pour ce qui concerne les variables aléatoires $\mathbf{W}^{(k)}$ et $\mathbf{U}^{(k)}$, mais ceci ne génère en pratique aucune difficulté.

Dans toute la suite, on supposera que l'espace Ω est "assez gros" pour qu'une telle suite $\{\mathbf{W}^{(k)}\}_{k \in \mathbb{N}^*}$ puisse y exister, et *toutes* les variables aléatoires seront définies sur le même espace $(\Omega, \mathcal{A}, \mathbb{P})$. \square

⁹voir [2, Ch. VII] pour plus de détails sur cette construction

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 $\mathbf{W} : \Omega \rightarrow \mathbb{R}$ une variable aléatoire intégrable définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, dont on veut estimer l'espérance :

$$\mathbb{E}(\mathbf{W}) = \int_{\Omega} \mathbf{W}(\omega) d\mathbb{P}(\omega) .$$

Une manière d'approximer cette espérance est d'effectuer $k+1$ tirages $(w^{(1)}, \dots, w^{(k+1)})$ de la variable aléatoire \mathbf{W} , indépendamment les uns des autres, et de calculer :

$$u^{(k+1)} = \frac{1}{k+1} \sum_{l=1}^{k+1} w^{(l)} .$$

En termes de variables aléatoires, cette dernière relation s'écrit

$$\mathbf{U}^{(k+1)} = \frac{1}{k+1} \sum_{l=1}^{k+1} \mathbf{W}^{(l)} ,$$

où les $(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k+1)})$ sont des variables aléatoires indépendantes de même loi que \mathbf{W} . On sait alors par la loi forte des grands nombres que la variable aléatoire $\mathbf{U}^{(k+1)}$ converge presque sûrement vers l'espérance de \mathbf{W} . Or l'expression de $\mathbf{U}^{(k+1)}$ se met sous la forme :

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

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

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) . \quad (12)$$

On se rappelle alors que l'espérance de la variable \mathbf{W} correspond à la valeur autour de laquelle la dispersion de cette variable est minimale :

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

et le calcul de l'espérance de \mathbf{W} par la méthode de Monte-Carlo donné par la relation (12) s'interprète comme l'algorithme du gradient stochastique 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 ;¹⁰

¹⁰Demander 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 $u^{(0)}$.

- 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 précision de l'estimation ; on peut donc aussi espérer obtenir un résultat de ce type dans le cadre du gradient stochastique.

Cadre probabiliste. Notant $(\Omega, \mathcal{A}, \mathbb{P})$ l'espace de probabilité, l'étape de remise à jour de u dans l'algorithme de gradient stochastique peut se mettre sous la forme générale suivante :

$$u^{(k+1)} = \mathcal{R}^{(k)} (u^{(k)}, w^{(k+1)}) , \quad (13)$$

ou encore, en terme de variables aléatoires :

$$\mathbf{U}^{(k+1)} = \mathcal{R}^{(k)} (\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) . \quad (14)$$

On note \mathbb{U} l'espace dans lequel les variables $\mathbf{U}^{(k)}$ prennent leurs valeurs et \mathcal{U} la tribu associée. On suppose que la variable aléatoire $\mathbf{U}^{(0)}$ est constante et prend la valeur $u^{(0)} \in U^{\text{ad}}$.

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

$$\mathcal{F}^{(0)} = \{\emptyset, \Omega\} \quad , \quad \mathcal{F}^{(k)} = \sigma (\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)}) .$$

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.

- De la structure triangulaire (car $\mathbf{U}^{(k)}$ ne dépend que des variables aléatoires $\mathbf{W}^{(l)}$, avec $l \leq k$) de la relation (14), on déduit que la variable aléatoire $\mathbf{U}^{(k)}$ est $\mathcal{F}^{(k)}$ -mesurable quelque soit k ; le processus stochastique $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ est donc adapté à la filtration $\{\mathcal{F}^{(k)}\}_{k \in \mathbb{N}}$.
- Des considérations précédentes et de l'indépendance des variables aléatoires $\mathbf{W}^{(k)}$, on déduit les règles de calcul suivantes concernant l'espérance conditionnelle :

$$\mathbb{E} (\mathbf{U}^{(k)} \mid \mathcal{F}^{(k)}) (\omega) = \mathbf{U}^{(k)}(\omega) ,$$

$$\begin{aligned} \mathbb{E} (\mathbf{U}^{(k+1)} \mid \mathcal{F}^{(k)}) (\omega) &= \mathbb{E} (\mathcal{R}^{(k)} (\mathbf{U}^{(k)}(\omega), \mathbf{W}^{(k+1)})) \\ &= \int_{\Omega} \mathcal{R}^{(k)} (\mathbf{U}^{(k)}(\omega), \mathbf{W}(\xi)) d\mathbb{P} (\xi) . \end{aligned}$$

- Enfin, on rappelle les différentes notions de convergence d'une suite de variables aléatoires $\mathbf{U}^{(k)}$ vers une constante $u^{\#}$:

- en probabilité : $\lim_{k \rightarrow +\infty} \mathbb{P} (\|\mathbf{U}^{(k)} - u^{\#}\| > \epsilon) = 0 \quad \forall \epsilon > 0,$
- presque sûre : $\lim_{k \rightarrow +\infty} \mathbb{P} (\sup_{m \geq k} \|\mathbf{U}^{(m)} - u^{\#}\| > \epsilon) = 0 \quad \forall \epsilon > 0,$
- en moyenne quadratique : $\lim_{k \rightarrow +\infty} \mathbb{E} (\|\mathbf{U}^{(k)} - u^{\#}\|^2) = 0.$

Pour l'algorithme 1, la convergence presque sûre s'interprète 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 $\mathbf{U}^{(k)}(\omega)$ engendrée par l'algorithme converge vers $u^{\#}$.

Remarque 2. Le cadre précédent présente le défaut de ne pas faire ressortir le fait que, dans un algorithme de gradient stochastique, on manipule essentiellement des *trajectoires* (indexées par les itérations) des variables w et u . Pour illustrer cet aspect trajectorien, on reprend la construction de la remarque 1 et on introduit l'espace de probabilité des échantillons $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu}) = (\mathbb{W}, \mathcal{W}, \mu)^{\otimes \mathbb{N}}$.

- Un élément $\tilde{w} \in \widetilde{\mathbb{W}}$ est une suite $\{w^{(k)}\}_{k \in \mathbb{N}^*}$ d'éléments de \mathbb{W} , et itérer la relation (13) revient à associer à une trajectoire des bruits $(w^{(1)}, \dots, w^{(k)}, \dots)$ une trajectoire de commandes $(u^{(1)}, \dots, u^{(k)}, \dots)$. La variable $u^{(k)}$ est de la forme $u^{(k)} = \tilde{U}^{(k)}(w^{(1)}, \dots, w^{(k)}, \dots)$, $\tilde{U}^{(k)}$ étant une variable aléatoire définie sur $(\widetilde{\mathbb{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$.
- La forme de la relation (13) fait que $u^{(k)}$ ne dépend en fait que des bruits $w^{(l)}$, $l \leq k$. On définit alors $\tilde{\mathcal{F}}^{(k)}$ comme étant la sous-tribu de $\widetilde{\mathcal{W}}$ générée par l'*application projection* de l'espace $\mathbb{W}^{\mathbb{N}}$ dans \mathbb{W}^k qui à $(w^{(1)}, \dots, w^{(k)}, \dots)$ associe ses k premières composantes. La suite $\{\tilde{\mathcal{F}}^{(k)}\}_{k \in \mathbb{N}^*}$ est une filtration, et le processus $\{\tilde{U}^{(k)}\}_{k \in \mathbb{N}^*}$ est adapté à cette filtration. La notion d'espérance conditionnelle par rapport à la sous-tribu $\tilde{\mathcal{F}}^{(k)}$ revient dans ce cas à la notion intuitive "d'espérance une fois les k premières composantes $(w^{(1)}, \dots, w^{(k)})$ de la trajectoire des bruits fixées". De même, la notion de convergence presque sûre s'interprète intuitivement comme "pour presque toutes les trajectoires $(w^{(1)}, \dots, w^{(k)}, \dots)$ du bruit, la suite $\{u^{(k)}\}_{k \in \mathbb{N}^*}$ engendrée par l'algorithme converge vers u^\sharp ".

Ce cadre probabiliste à base de trajectoires est bien évidemment équivalent à celui que nous avons présenté sur l'espace primitif Ω , ce dernier ayant comme on l'a déjà dit l'avantage que toutes les variables aléatoires sont définies sur le même espace de probabilité. \square

De nombreux ouvrages présentent les outils probabilistes utilisés dans le cadre de ce cours. On consultera par exemple les ouvrages de référence [2] ou [9]–[8].

2.3 Résultats de convergence

Convergence en moyenne quadratique. Reprenant les travaux de [31] et [14], 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, mais sous des conditions qui ne sont pas forcément réalistes. On donne pour commencer la définition suivante.

Définition 1. On dit qu'une suite de réels positifs $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est une σ -suite si la série qu'elle engendre est divergente, la série de ses carrés étant quant à elle convergente :

$$\sum_{k \in \mathbb{N}} \epsilon^{(k)} = +\infty \quad , \quad \sum_{k \in \mathbb{N}} \epsilon^{(k)2} < +\infty . \quad (15)$$

Puis, \mathbf{W} étant une variable aléatoire à valeurs dans \mathbb{W} et U^{ad} étant une partie convexe fermée d'un espace de Hilbert \mathbb{U} , on fait les hypothèses suivantes pour la convergence.

HC1 La variable aléatoire $j(u, \mathbf{W}) : \Omega \rightarrow \mathbb{R}$ est mesurable et son espérance existe pour tout $u \in U^{\text{ad}}$.

HC2 La fonction $j(\cdot, w) : \mathbb{U} \rightarrow \mathbb{R}$ est convexe, semi-continue inférieurement, propre, différentiable pour tout $w \in \mathbb{W}$.

HC3 Le gradient de j par rapport à u est borné uniformément en u et en w :

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

HC4 Le problème (10) admet un ensemble U^\sharp non vide de solutions, qui vérifie la relation :

$$\forall u \in U^{\text{ad}}, J(u) - J^\sharp \geq c (\text{dist}_{U^\sharp}(u))^2,$$

où J^\sharp est la valeur du minimum de (10) et où $\text{dist}_{U^\sharp}(\cdot)$ est la fonction distance à l'ensemble U^\sharp .

HC5 La suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est une σ -suite décroissante.

Théorème 1. (Convergence en moyenne quadratique)

Sous les hypothèses **HC1–HC2–HC3–HC4–HC5**, la suite $\{U^{(k)}\}_{k \in \mathbb{N}}$ de variables aléatoires générée par l'algorithme 1 du gradient stochastique converge en moyenne quadratique vers l'ensemble U^\sharp :

$$\lim_{k \rightarrow +\infty} \mathbb{E} \left(\text{dist}_{U^\sharp}(U^{(k)})^2 \right) = 0.$$

Preuve. U^\sharp étant un convexe fermée, la projection sur cet ensemble est bien définie. Soit $\{u^{(k)}\}_{k \in \mathbb{N}}$ une réalisation de l'algorithme 1 et soit $\bar{u}^{(k)}$ la projection de $u^{(k)}$ sur U^\sharp :

$$\text{dist}_{U^\sharp}(u^{(k)})^2 = \|u^{(k)} - \bar{u}^{(k)}\|^2.$$

Notant $d^{(k)} = \text{dist}_{U^\sharp}(u^{(k)})^2$, utilisant le fait que la projection est contractante et **HC3**, on a :

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

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

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

Prenant de part et d'autre de cette inégalité l'espérance conditionnelle par rapport à la sous-tribu $\mathcal{F}^{(k)}$, utilisant les propriétés de mesurabilité des variables aléatoires, le fait que $\mathbb{E}(\nabla_u j(U^{(k)}, W^{(k+1)}) | \mathcal{F}^{(k)}) = \nabla J(U^{(k)})$, la convexité de J ainsi que l'hypothèse **HC4**, on obtient :¹¹

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

Prenant ensuite l'espérance, il vient :

$$\mathbb{E}(D^{(k+1)}) \leq (1 - 2\epsilon^{(k)} c) \mathbb{E}(D^{(k)}) + \epsilon^{(k)2} m^2. \quad (16)$$

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

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

Comme la suite de terme général $\prod_{l=0}^k (1 - 2\epsilon^{(l)} c)$ converge vers zéro (voir proposition 2 page 24) et que la suite de

terme général $\sum_{l=0}^k \epsilon^{(l)2}$ converge (hypothèse **HC5**), on en déduit le résultat annoncé. □

¹¹L'hypothèse **HC3** et le fait que $j(u, W(\cdot))$ soit intégrable implique, par un argument de convergence dominée, que le gradient de j par rapport à u est lui aussi intégrable.

On a même un résultat précis concernant la vitesse de décroissance en moyenne de la distance $\mathbf{D}^{(k)}$.

Théorème 2. (Vitesse de convergence en moyenne quadratique)

Sous les mêmes hypothèses et avec les mêmes notations que dans le théorème 1, choisissant la suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ de la forme :

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

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

$$\mathbb{E} \left(\text{dist}_{U^\sharp}(\mathbf{U}^{(k)})^2 \right) \leq \frac{1}{\frac{c^2}{m^2} k + \frac{1}{d^{(0)}}} \quad \forall k \in \mathbb{N}.$$

Preuve. On repart de l'inégalité (16) :

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

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 :

$$(\alpha k + \beta) \mathbb{E} \left(\mathbf{D}^{(k)} \right) \leq 1,$$

est vérifiée avec les choix $\alpha = \frac{c^2}{m^2}$, $\beta = \frac{1}{d^{(0)}}$ et $\gamma = \frac{c}{m^2}$ (voir [14] pour plus de détails). □

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

$$\frac{1}{k} \left(\frac{m}{c} \right)^2. \quad (17)$$

D'après les hypothèses **HC3** et **HC4**, 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.

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

Proposition 1. *L'opération de projection sur U^{ad} est contractante.*

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

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

Les conditions d'optimalité de ce problème, évaluées au point $\text{proj}_{U^{\text{ad}}}(v)$, s'écrivent :

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

Intervertissant les rôles de u et v , on obtient :

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

Additionnant ces deux dernières inégalités, il vient :

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

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

Proposition 2. Soit $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ une suite réelle vérifiant : $\epsilon^{(k)} \searrow 0$ et $\sum \epsilon^{(k)} = +\infty$. Alors, pour tout $\alpha > 0$, la suite de terme général $\{\rho^{(k)}\}_{k \in \mathbb{N}}$ 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 $\epsilon^{(k)} \rightarrow 0$), 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 $\{\rho^{(k)}\}_{k \in \mathbb{N}}$ 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 $\{\rho^{(k)}\}_{k \in \mathbb{N}}$ converge vers zéro. □

Autres résultats de convergence. Les résultats de convergence obtenus au paragraphe précédent se trouvent dans [14]. On a choisi de les présenter car ils sont représentatifs des résultats dont on dispose sur le gradient stochastique (convergence et vitesse) et car leur démonstration est simple. Ils ne sont cependant pas entièrement satisfaisants, pour les raisons suivantes.

- Tout d'abord, l'hypothèse **HC3** 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 [14] un théorème de convergence presque sûre ainsi que l'estimation de vitesse de convergence associée, mais ces résultats sont obtenus sous l'hypothèse **HC3**.¹²

On donnera au §3.3 un théorème de convergence presque sûre très général pour une famille d'algorithmes incluant l'algorithme 1. Dans ce théorème, établi dans le cadre convexe, l'hypothèse **HC3** de gradient de j par rapport à u borné uniformément en w est remplacée par une hypothèse de gradient *linéairement* borné en u uniformément en w :

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

Une telle hypothèse n'est pas surprenante dans une méthode de type gradient.¹³ La démonstration du théorème correspondant fait appel à des outils probabilistes évolués comme la théorie des quasi-martingales et sera détaillée au §3.

On trouvera enfin au §15.4 le théorème de convergence presque sûre donné dans [12] concernant les algorithmes stochastiques généraux, dans lequel il n'est pas fait d'hypothèse explicite de convexité sur la fonction j . Ce résultat provient plus de la théorie de l'approximation stochastique que de celle de l'optimisation. Il est obtenu sous des hypothèses assez peu restrictives, mais a le défaut de n'être applicable que lorsque l'ensemble U^{ad} est l'espace \mathcal{U} tout entier.¹⁴

¹²De plus, leur démonstration est moins élémentaire car faisant appel à la théorie des martingales.

¹³Elle constitue une extension au cadre stochastique de l'hypothèse classique "critère à gradient lipschitzien".

¹⁴On pourra cependant, pour des extensions au cas avec projection, se reporter aux références [22] et [23].

Considérations pratiques. La mise en œuvre de l’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.

- 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 $\|u^{(k+1)} - u^{(k)}\|$ car cette différence converge mécaniquement vers zéro par le biais des pas $\epsilon^{(k)}$ (hypothèse **HC5**). 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 $\nabla_{u,j}(U^{(k)}, W^{(k+1)})$ a vocation à converger vers $\nabla J(u^\#)$ et peut donc servir pour effectuer un test de convergence.¹⁵ 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 est 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.

- La deuxième question porte sur la forme de la σ -suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$. L’hypothèse **HC5** suggère de prendre des pas $\epsilon^{(k)}$ de la forme $\frac{1}{k^\gamma}$, avec $\frac{1}{2} < \gamma \leq 1$. En pratique, on constate que le meilleur comportement de l’algorithme est obtenu avec le choix $\gamma = 1$, et la théorie confirme ce choix (voir §2.4).

Mais le réglage du coefficient γ ne prend en compte qu’une partie du comportement asymptotique de l’algorithme. C’est pourquoi on choisit plutôt des pas de la forme :

$$\epsilon^{(k)} = \frac{\alpha}{k^\gamma + \beta}. \quad (18)$$

Indépendamment du coefficient γ déjà évoqué, les coefficients α et β remplissent les rôles suivants :

- le coefficient α a une influence sur la vitesse asymptotique de l’algorithme : son effet *multiplicatif* sur k peut plus ou moins accélérer la convergence ;
- 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 constant, sensiblement 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.

Il est de plus souhaitable de disposer d’une estimation de la longueur du pas de gradient pour l’algorithme déterministe sous-jacent : cette estimation constitue en général une bonne initialisation pour le réglage des pas au début de l’algorithme stochastique (quotient $\frac{\alpha}{\beta}$).

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 [3] qui, en plus d’être un outil théorique permettant d’affaiblir les hypothèses nécessaires à la convergence des approximations stochastiques, (voir §15.4 pour plus de détails sur ce point), 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} ;

¹⁵La condition d’optimalité est $\langle \nabla J(u^\#), u - u^\# \rangle \geq 0 \forall u \in U^{\text{ad}}$; elle s’écrit $\nabla J(u^\#) = 0$ dans le cas $u^\# \in \text{int}(U^{\text{ad}})$.

2. l'algorithme de Kesten [20], précisément décrit dans [13], 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, à partir de l'indice k des itérations de l'algorithme, on définit la suite d'entiers n_k par la relation :

$$n_{k+1} = n_k + \mathbf{1}_{\{\langle \nabla_u j(u^{(k-1)}, \omega^{(k)}) , \nabla_u j(u^{(k)}, \omega^{(k+1)}) \rangle < 0\}} ,$$

dans laquelle le dernier terme prend la valeur 1 si le produit scalaire des deux gradients successifs est négatif (ces gradients forment donc un angle obtus) et 0 sinon ; le pas de l'algorithme est alors défini par :

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

3. une règle multiplicative d'adaptation des pas [30], 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" (et non de manière mécanique comme dans l'utilisation directe de la relation (18)). Signalons enfin qu'il existe toute une littérature concernant les algorithmes stochastiques à *pas constant* (voir par exemple [1] ou [39]).

2.4 Théorème de la limite centrale

On considère le même problème que précédemment, en supposant dans tout ce qui suit que l'espace \mathbb{U} est égal à \mathbb{R}^d et que l'ensemble U^{ad} est l'espace \mathbb{U} tout entier. La projection sur U^{ad} dans l'algorithme 1 correspond alors à l'identité, et la k -ème itération de l'algorithme de gradient stochastique s'écrit en terme de variables aléatoires :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \epsilon^{(k)} \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) . \quad (19a)$$

On écrit cette itération sous la forme canonique suivante :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} + \epsilon^{(k)} \left(h(\mathbf{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)} \right) , \quad (19b)$$

où la fonction h et la variable aléatoire $\boldsymbol{\xi}^{(k+1)}$ sont définies par :

$$h(u) = -\nabla J(u) \quad , \quad \boldsymbol{\xi}^{(k+1)} = \nabla J(\mathbf{U}^{(k)}) - \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) . \quad (19c)$$

On remarquera d'une part que la condition d'optimalité au point $u^\# \in \mathbb{U}$ s'écrit :

$$\nabla J(u^\#) = 0 = -h(u^\#) ,$$

et d'autre part que $\boldsymbol{\xi}^{(k+1)}$ est un *incrément de martingale* :

$$\mathbb{E} \left(\boldsymbol{\xi}^{(k+1)} \mid \mathcal{F}^{(k)} \right) = 0 .$$

Normalité asymptotique. On va donner un résultat de type “théorème de la limite centrale” (TCL) précisant la normalité asymptotique des itérées $\mathbf{U}^{(k)}$ de l’algorithme du gradient stochastique. Ce résultat sera utilisé pour comparer la vitesse de convergence de différentes mises en œuvre de l’algorithme. Le théorème suivant est un cas particulier de celui présenté dans [15]. Sa démonstration, assez technique, n’est pas donnée ici. On pourra aussi consulter le §15.4 pour un résultat de même nature. On fait les hypothèses suivantes pour le comportement asymptotique.

HV1 La suite de variables aléatoires $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ converge presque sûrement vers u^\sharp .

HV2 La fonction J est deux fois continûment différentiable, et sa matrice hessienne H au point u^\sharp ($H = J''(u^\sharp)$) est définie positive.¹⁶ On notera $c > 0$ la *plus petite valeur propre* de H .

HV3 La suite des matrices de covariance conditionnelle $\{\mathbb{E}(\boldsymbol{\xi}^{(k+1)}(\boldsymbol{\xi}^{(k+1)})^\top \mid \mathcal{F}^{(k)})\}_{k \in \mathbb{N}}$ converge presque sûrement vers Γ , matrice symétrique définie positive déterministe égale à la matrice de covariance du gradient partiel de j par rapport à u au point u^\sharp :

$$\Gamma = \mathbb{E} \left(\nabla_u j(u^\sharp, \mathbf{W}) (\nabla_u j(u^\sharp, \mathbf{W}))^\top \right) .$$

HV4 Il existe $\delta > 0$, tel que l’on ait : $\sup_{k \in \mathbb{N}} \mathbb{E} (\|\boldsymbol{\xi}^{(k+1)}\|^{2+\delta} \mid \mathcal{F}^{(k)}) < +\infty$.

HV5 La suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est de la forme : $\epsilon^{(k)} = \frac{\alpha}{k^\gamma + \beta}$, avec $\alpha > 0$, $\beta > 0$ et $\frac{1}{2} < \gamma \leq 1$.

HV6 La matrice $H - \lambda I$ est définie positive,¹⁷ le coefficient λ étant défini par :

$$\lambda = \begin{cases} 0 & \text{si } \gamma < 1 \\ \frac{1}{2\alpha} & \text{si } \gamma = 1 \end{cases} \quad (20)$$

On a alors le théorème suivant précisant la vitesse à laquelle les itérées $\mathbf{U}^{(k)}$ de l’algorithme du gradient stochastique convergent vers u^\sharp .

Théorème 3. (Théorème de la limite centrale – TCL)

Sous les hypothèses **HV1**, **HV2**, **HV3**, **HV4**, **HV5** et **HV6**, la suite normalisée de variables aléatoires $\left\{ (\epsilon^{(k)})^{-\frac{1}{2}} (\mathbf{U}^{(k)} - u^\sharp) \right\}_{k \in \mathbb{N}}$ converge en loi vers une loi normale centrée de matrice de covariance Σ :

$$\frac{1}{\sqrt{\epsilon^{(k)}}} (\mathbf{U}^{(k)} - u^\sharp) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma) , \quad (21)$$

la matrice de covariance Σ étant solution de l’équation de Lyapunov :

$$(H - \lambda I)\Sigma + \Sigma(H - \lambda I) = \Gamma . \quad (22)$$

Preuve. Voir [15, Ch. 4]. □

Remarque 3. On rappelle le résultat classique caractérisant la solution d’une équation de Lyapunov.

Proposition 3. Soit K une matrice définie positive et C une matrice symétrique définie positive de même dimension. Alors, il existe une matrice S symétrique définie positive, solution unique de l’équation de Lyapunov :

$$KS + SK^\top = C ,$$

et cette solution est :

$$S = \int_0^{+\infty} \exp(-tK)C \exp(-tK^\top) dt . \quad \square$$

¹⁶Au voisinage de u^\sharp , on a donc : $h(u) = -H(u - u^\sharp) + O(\|u - u^\sharp\|^2)$, avec $-H$ matrice *attractive*.

¹⁷Dans le cas $\gamma = 1$, une condition équivalente à cette propriété de positivité est : $2\alpha c > 1$.

Vitesse de convergence. Du théorème précédent, on tire les premières conclusions suivantes.

1. Utilisant la forme des pas imposée par l'hypothèse **HV5**, et constatant que le coefficient β n'a aucune influence asymptotique, on déduit du théorème 3 la propriété de convergence suivante :

$$k^{\frac{\gamma}{2}} \left(\mathbf{U}^{(k)} - u^\# \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \alpha \Sigma) . \quad (23)$$

Le choix optimal, c'est-à-dire celui qui conduit à la vitesse de convergence la plus élevée, est donc $\gamma = 1$. On retrouve ainsi la vitesse en $\frac{1}{\sqrt{k}}$ du théorème de la limite centrale classique.

2. Dans le cas optimal $\gamma = 1$, le coefficient α doit être choisi de telle sorte que l'hypothèse **HV6** soit satisfaite, d'où la condition :

$$\alpha > \frac{1}{2c} .$$

Le raisonnement simpliste consistant à prendre un α aussi petit que possible pour diminuer la covariance asymptotique dans la relation (23) ne tient pas. En effet, la solution Σ de l'équation de Lyapunov (22) dépend de λ , et donc de α , de telle sorte que la matrice de covariance $\alpha \Sigma$ ne varie pas *linéairement* avec α . Ainsi, dans le cas scalaire ($d = 1$), H et Γ sont des réels, on a : $H = c$, et la solution de l'équation de Lyapunov (22) est :

$$\Sigma = \frac{\alpha \Gamma}{2\alpha c - 1} .$$

La variance $\alpha \Sigma$ peut être minimisée par rapport à α , le minimum étant atteint pour la valeur $\alpha^\# = \frac{1}{c}$ (qui vérifie bien la condition : $2\alpha^\# c - 1 > 0$).

On se place maintenant dans le cas optimal $\gamma = 1$. Il reste alors à rendre aussi petite que possible (au sens des matrices définies positives) la matrice de covariance dans la relation (23). 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

Algorithme de Newton stochastique. En optimisation déterministe, il est bien connu qu'une façon d'améliorer sensiblement le comportement des algorithmes à direction de descente est de pré-multiplier 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.

Pour appliquer cette idée à l'algorithme du gradient stochastique, on se donne une matrice A carrée de dimension d symétrique et inversible. 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 (19) l'itération courante de gradient stochastique par la nouvelle relation :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \frac{A}{k + \beta} \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) = \mathbf{U}^{(k)} + \frac{A}{k + \beta} \left(h(\mathbf{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)} \right) . \quad (24)$$

Tout se passe alors comme si l'on mettait en œuvre l'algorithme du gradient stochastique avec :

- un champ $\tilde{h} = Ah$,
- des bruits $\tilde{\boldsymbol{\xi}}^{(k)} = A\boldsymbol{\xi}^{(k)}$,
- des pas $\epsilon^{(k)} = \frac{1}{k + \beta}$.

Moyennant la satisfaction de l'hypothèse **HV6**, qui prend ici la forme :

HV6' La matrice $AH - \frac{I}{2}$ est définie positive,

le théorème 3 s'applique et fournit la convergence en loi pour l'algorithme du gradient stochastique à gain matriciel (24) :

$$\sqrt{k} \left(\mathbf{U}^{(k)} - u^\# \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_A), \quad (25)$$

la matrice de covariance asymptotique Σ_A vérifiant la relation :

$$\left(AH - \frac{I}{2} \right) \Sigma_A + \Sigma_A \left(HA - \frac{I}{2} \right) = A \Gamma A. \quad (26)$$

On a alors le théorème suivant, qui caractérise le choix optimal du gain matriciel dans l'algorithme de gradient stochastique.

Théorème 4. (Algorithme de Newton stochastique)

Dans l'algorithme de gradient stochastique à gain matriciel (24), le choix :

$$A^\# = H^{-1},$$

rend minimale la matrice de covariance asymptotique, qui a pour expression :

$$\Sigma^\# = H^{-1} \Gamma H^{-1}.$$

Preuve. Écrivant la matrice de covariance asymptotique Σ_A donnée par le théorème 3 pour l'algorithme de gradient stochastique à gain matriciel (24) sous la forme :

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

et reportant cette expression dans la relation (26), on obtient :

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

La matrice Δ_A vérifie donc une équation de Lyapunov et est, par la proposition 3, définie positive dès que $A \neq H^{-1}$. On en déduit que l'inégalité (pour l'ordre de la semi-définie positivité) $\Sigma_A \geq H^{-1} \Gamma H^{-1}$ est vérifiée pour toute matrice A symétrique définie positive, l'égalité étant obtenue pour la valeur $A^\# = H^{-1}$ du gain matriciel. \square

Remarque 4. Le gain H^{-1} correspond à la matrice hessienne de la fonction J évaluée au point $u^\#$, d'où le nom "algorithme de Newton stochastique" donné à l'algorithme (24) avec ce choix particulier de gain. Bien sûr, les pas utilisés dans l'algorithme stochastique doivent être de longueur $1/k$ (et non de longueur 1 comme dans l'algorithme de Newton déterministe). \square

Remarque 5. Dans le cas déterministe, l'utilisation d'une méthode de type Newton plutôt qu'une méthode de type gradient conduit à une vitesse de convergence q-quadratique (de type a^{2k} , $|a| < 1$) plutôt que q-linéaire (de type a^k). Dans le cas stochastique, les méthodes à gain scalaire et à gain matriciel ont toutes les deux une vitesse de convergence de type $\frac{a}{\sqrt{k}}$. L'amélioration apportée par le gain matriciel est due à la constante multiplicative a intervenant dans cette vitesse : notant 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 de covariance optimale $H^{-1} \Gamma H^{-1}$ est de l'ordre du quotient $(\frac{m}{c})^2$, qui correspond à la borne sur la vitesse obtenue pour l'algorithme du gradient stochastique à gain scalaire (voir la relation (17)). On en déduit que la vitesse de convergence des deux méthodes est la même dans la direction où la vitesse de convergence est la plus faible. Ce résultat se vérifie d'ailleurs très bien de manière expérimentale (voir la boîte à outils "Gradient stochastique" [38]). \square

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

Définition 2. On dit qu'un algorithme de gradient stochastique est Newton-efficace s'il a le même comportement en loi que l'algorithme de Newton stochastique.

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. On peut alors se poser les deux questions suivantes.

1. *Comment utiliser en pratique l'algorithme de Newton stochastique ?*

La difficulté vient du fait que cet algorithme ne peut pas être directement mis en œuvre car le gain optimal H^{-1} dépend du point $u^\#$ que l'on cherche ! Plutôt que de proposer des algorithmes approximant la matrice H^{-1} au cours des itérations, on verra au paragraphe suivant une technique simple permettant d'obtenir un algorithme Newton-efficace.

2. *L'algorithme de Newton stochastique est-il efficace au sens de l'estimation ?*¹⁸

Dans la théorie de l'estimation, il est possible de comparer la qualité de différents estimateurs d'une même quantité déterministe. On montre en particulier qu'il existe une borne, dite de *Cramer-Rao*, en dessous de laquelle la variance d'un estimateur ne peut descendre (voir §12 pour un rappel de ce résultat). On peut alors se poser la question du lien entre la borne fournie par l'algorithme de Newton stochastique et celle de Cramer-Rao. Ce point fait l'objet du dernier paragraphe.

Moyennisation. Comme on l'a déjà souligné, l'algorithme de Newton stochastique ne peut pas être directement mis en œuvre à cause de la matrice de gain H^{-1} qui dépend du point $u^\#$ que l'on cherche à déterminer. Afin de contourner cette difficulté, B. T. Polyak a proposé dans [32] de modifier l'algorithme de gradient stochastique 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 :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \epsilon^{(k)} \nabla_{u,j}(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) .$$

par le calcul en deux étapes suivant :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \epsilon^{(k)} \nabla_{u,j}(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) , \quad (27a)$$

$$\widehat{\mathbf{U}}^{(k+1)} = \frac{1}{k+1} \sum_{l=1}^{k+1} \mathbf{U}^{(l)} , \quad (27b)$$

dans lequel la première étape (27a) est identique à celle du gradient stochastique classique, la deuxième étape (27b) 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 (27b) :

$$\widehat{\mathbf{U}}^{(k+1)} = \widehat{\mathbf{U}}^{(k)} + \frac{1}{k+1} (\mathbf{U}^{(k+1)} - \widehat{\mathbf{U}}^{(k)}) . \quad (27c)$$

On remarquera que, par le théorème de Césaro, la convergence presque sûre de la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ implique la convergence de la suite moyennée $\{\widehat{\mathbf{U}}^{(k)}\}_{k \in \mathbb{N}}$. Sous les conditions du §2.3, et en particulier avec des pas $\epsilon^{(k)}$ de la forme :

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

¹⁸Il y a sur ce point une subtilité de terminologie qu'il faut bien noter : la **Newton-efficacité d'un algorithme** de gradient stochastique correspond au fait d'atteindre la matrice de covariance de l'algorithme de Newton stochastique, alors que l'**efficacité d'un estimateur** correspond au fait d'atteindre la borne de Cramer-Rao.

on sait que la suite $\{\widehat{\mathbf{U}}^{(k)}\}_{k \in \mathbb{N}}$ converge vers la solution $u^\#$ du problème.

L'intérêt essentiel de l'algorithme moyenné (27) tient à ses propriétés asymptotiques. Les hypothèses que l'on fait alors sont semblables à celles ayant permis d'établir le théorème 3 de la limite centrale, mais on restreint l'hypothèse **HV5** au cas où le coefficient γ est *strictement* inférieur à 1.

HV5' La suite $\{\epsilon^{(k)}\}_{k \in \mathbb{N}}$ est de la forme : $\epsilon^{(k)} = \frac{\alpha}{k^\gamma + \beta}$, avec $\alpha > 0$, $\beta > 0$ et $\frac{1}{2} < \gamma < 1$.

Remarque 6. C'est bien sur avec la suite $\{\widehat{\mathbf{U}}^{(k)}\}_{k \in \mathbb{N}}$ obtenue *après moyennisation* que l'on a des propriétés de convergence intéressantes. Pour la suite $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ provenant de l'étape (27a), la vitesse de convergence est inférieure strictement à $\frac{1}{\sqrt{k}}$ (d'après le théorème 3 et avec la condition $\gamma < 1$), et donc moins bonne que celle de l'algorithme de gradient stochastique bien réglé. \square

Théorème 5. (Optimalité asymptotique du gradient stochastique moyenné)

Sous les hypothèses HV1, HV2, HV3, HV4, HV5' et HV6, l'algorithme du gradient stochastique moyenné est Newton-efficace :

$$\sqrt{k}(\widehat{\mathbf{U}}^{(k)} - u^\#) \xrightarrow{\mathcal{L}} \mathcal{N}(0, H^{-1}\Gamma H^{-1}).$$

Preuve. Voir [15, Ch. 4]. \square

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} . Pour sa mise en œuvre, on peut faire les remarques 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 [12] 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 considérablement l'algorithme durant sa phase transitoire, il est préférable de ne commencer le processus de moyennisation que lorsque le gradient stochastique (27a) s'est approché de la zone de convergence.

Lien avec la borne de Cramer-Rao (facultatif). On va montrer que l'algorithme de Newton stochastique n'engendre pas forcément une suite d'estimateurs $\{\mathbf{U}^{(k)}\}_{k \in \mathbb{N}}$ qui soient asymptotiquement efficaces au sens de la théorie de l'estimation.

(a) Tout d'abord, on repart des résultats sur le maximum de vraisemblance donnés au §12.3. Soit \mathbf{W} une variable aléatoire dont la densité $p(w, m)$ dépend d'un paramètre m à déterminer, la loi associée étant notée \mathbb{P} . Alors, $(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k+1)})$ étant des variables aléatoires indépendantes de même loi que \mathbf{W} , l'estimateur de m obtenu par la méthode du maximum de vraisemblance :

$$\widehat{\mathbf{U}}^{(k+1)} = \arg \max_u \frac{1}{k} \sum_{l=1}^{k+1} \ln(p(\mathbf{W}^{(l)}, u)), \tag{28}$$

est asymptotiquement sans biais et efficace :

$$\sqrt{k}(\widehat{\mathbf{U}}^{(k+1)} - m) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, (\mathcal{I}_{\mathbf{W}}(m))^{-1}\right).$$

$\mathcal{I}_{\mathbf{W}}(m)$ étant la matrice d'information de Fisher. La fonction que l'on maximise dans (28) est une approximation de type Monte-Carlo de :

$$\mathbb{E} (\ln (p(\mathbf{W}, u))) .$$

Notant j la fonction définie par : $j(u, w) = \ln (p(w, u))$, cette dernière expression suggère d'utiliser pour résoudre le problème de maximisation l'algorithme de Newton stochastique :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} + \frac{A^\sharp}{k} \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) , \quad (29)$$

où A^\sharp est le gain matriciel optimal. Utilisant comme précédemment les notations :

$$h(u) = \mathbb{E} (\nabla_u j(u, \mathbf{W})) \quad , \quad \boldsymbol{\xi}^{(k+1)} = \nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) - h(\mathbf{U}^{(k)}) ,$$

et supposant que l'algorithme converge vers le point $u^\sharp = m$, on montre facilement que l'on a :

$$A^\sharp = - (\nabla h(u^\sharp))^{-1} = \left(\mathcal{I}_{\mathbf{W}}(m) \right)^{-1} , \quad \lim_{k \rightarrow +\infty} \mathbb{E} (\boldsymbol{\xi}^{(k+1)} (\boldsymbol{\xi}^{(k+1)})^\top | \mathcal{F}^{(k)}) = \mathcal{I}_{\mathbf{W}}(m) .$$

La matrice de covariance optimale dans l'algorithme de Newton est alors égale à $[\mathcal{I}_{\mathbf{W}}(m)]^{-1}$, inverse de la matrice d'information de Fisher, qui est aussi la borne de Cramer-Rao pour l'estimation de m .

En conclusion, l'itéré $\mathbf{U}^{(k)}$ de l'algorithme (29) appliqué à la maximisation de la vraisemblance est un estimateur asymptotiquement efficace du paramètre m , de matrice de covariance $[\mathcal{I}_{\mathbf{W}}(m)]^{-1}$.

(b) On suppose maintenant que le paramètre m intervenant dans la loi de probabilité de la variable \mathbf{W} correspond à l'espérance de cette variable. On a vu que le calcul de cette espérance par la méthode de Monte-Carlo :

$$\mathbf{U}^{(k+1)} = \frac{1}{k+1} \sum_{l=1}^{k+1} \mathbf{W}^{(l)} . ,$$

peut s'interpréter comme un algorithme de gradient stochastique :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \frac{1}{k+1} (\mathbf{U}^{(k)} - \mathbf{W}^{(k+1)}) . \quad (30)$$

Cet algorithme est en fait l'algorithme de Newton stochastique car la matrice hessienne de la fonction J de ce problème est égale à la matrice identité ; notant Γ la matrice de covariance de la variable \mathbf{W} , on a que la covariance optimale de cet algorithme de Newton stochastique est égale à Γ .

En conclusion, l'algorithme de Newton stochastique (30) appliqué au calcul de la moyenne m par la méthode de Monte-Carlo fournit un estimateur du paramètre m , de matrice de covariance Γ .

Reprenant les conclusions des paragraphes **(a)** et **(b)**, on obtient que, pour une loi de probabilité paramétrée par sa moyenne m et de matrice de covariance Γ , l'inégalité ¹⁹:

$$\Gamma \geq \left(\mathcal{I}_{\mathbf{W}}(m) \right)^{-1} ,$$

est toujours vraie. Dans le cas gaussien, on sait que cette inégalité est en fait une égalité : la méthode de Monte-Carlo fournit un estimateur efficace de l'espérance d'une loi normale. Mais cette inégalité peut aussi être *stricte*.²⁰ L'algorithme stochastique (29) est alors asymptotiquement plus efficace que l'algorithme de Newton stochastique (30). **Il est donc possible de construire des algorithmes stochastiques plus efficaces que l'algorithme de Newton.**

¹⁹au sens des matrices semi-définies positives

²⁰On considérera par exemple le cas de la première loi de Laplace, de densité $\frac{1}{2\sigma} \exp\left(-\frac{|w-m|}{\sigma}\right)$, dont la variance est égale à $2\sigma^2$, l'inverse de la quantité d'information de Fisher $\mathcal{I}_{\mathbf{W}}(m)$ étant pour sa part égale à σ^2 .

Remarque 7. Ce résultat est mentionné dans [32] dans le cas plus général où le gradient de la fonction j est de la forme :

$$\nabla_u j(\mathbf{U}^{(k)}, \mathbf{W}^{(k+1)}) = \nabla J(\mathbf{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)},$$

les bruits $\boldsymbol{\xi}^{(k+1)}$ étant indépendants et identiquement distribués, de densité p et de moyenne nulle. Alors, supposant que la matrice d'information de Fisher \mathcal{I}_ξ existe et est finie :

$$\mathcal{I}_\xi = \int \frac{\nabla p(\xi)(\nabla p(\xi))^\top}{p(\xi)} d\xi,$$

notant H la matrice hessienne de J au point u^\sharp et $\varphi(x) = \ln(p(x))$, on montre que l'algorithme stochastique :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} + \frac{H^{-1}\mathcal{I}_\xi^{-1}}{k} \nabla\varphi(\nabla J(\mathbf{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)}), \quad (31)$$

qui fait intervenir une *transformation non linéaire* du gradient, est tel que $\mathbf{U}^{(k)}$ est un estimateur asymptotiquement sans biais et efficace de u^\sharp , la borne de Cramer-Rao correspondante étant :

$$H^{-1}\mathcal{I}_\xi^{-1}H^{-1}.$$

L'algorithme de Newton stochastique appliqué au problème de la minimisation de J s'écrit :

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} - \frac{H^{-1}}{k} (\nabla J(\mathbf{U}^{(k)}) + \boldsymbol{\xi}^{(k+1)}), \quad (32)$$

et sa covariance asymptotique est :

$$H^{-1}\Gamma H^{-1},$$

où Γ est la matrice de covariance du bruit $\boldsymbol{\xi}$. Comme dans le cas des paragraphes (a) et (b), l'algorithme (31), qui correspond la maximisation de la vraisemblance pour la famille de densité $\pi(\xi, u) = p(\nabla J(u) + \xi)$, est asymptotiquement plus efficace que l'algorithme de Newton (32). \square

2.6 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,

en insistant un peu sur les aspects "mise en œuvre" de l'algorithme. Il y a d'autres points intéressants concernant cette méthode, comme par exemple la *loi quadratique forte des grands nombres* permettant d'obtenir en fonction des itérées de l'algorithme une estimation de la matrice de covariance asymptotique, ou la *procédure de Kiefer-Wolfowitz* dans laquelle l'estimation du gradient de j est calculée par différence finie entre deux estimations de la fonction j . Plus intéressant encore, des travaux récents ont permis d'étendre le champ d'application du gradient stochastique au *problèmes d'optimisation stochastique en boucle fermée* ! Ces extensions, ainsi que quelques autres, seront décrites au §6.

La méthode du gradient stochastique est un cas particulier de la théorie plus générale des *algorithmes stochastiques*, algorithmes pour lesquels la convexité ne joue pas un rôle central (du moins explicitement). C'est en particulier de cette théorie que vient le théorème de la limite centrale et ses extensions. Cette théorie plus générale sera présentée au §15.

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

