6

Introduction et motivations

6.1 Problématique dans le cas déterministe

On a vu dans la Partie I de cet ouvrage que l'on sait associer, dans le cadre *déterministe*, les idées de l'*optimisation* avec celles issues de la *décomposition-coordination*. Partant du problème d'optimisation (4.1) (écrit dans le cadre statique) ou du problème d'optimisation (2.26) (écrit dans le cadre dynamique), on suppose que la taille et/ou la complexité de ce problème sont telles que l'on peut difficilement calculer sa solution à l'aide des algorithmes d'optimisation classiques (on dit alors que l'on a affaire à un *grand système*). L'idée des techniques de décomposition et coordination consiste à remplacer le problème (4.1) ou (2.26) par une suite de problèmes dits auxiliaires, telle que la suite des solutions des ces problèmes auxiliaires converge vers la solution du problème de départ. Cette transformation se fait à l'aide d'un principe général en optimisation, appelé *Principe du Problème Auxiliaire* (PPA). Ce principe est de type *variationnel*, ce qui signifie que les algorithmes que l'on obtient à l'aide du PPA sont dans leur essence de même nature que l'algorithme du gradient.

Comme on l'a montré dans la Partie I de cet ouvrage, le PPA se prête particulièrement bien à un objectif de décomposition, en ce sens qu'il permet de formuler des problèmes auxiliaires dont la minimisation peut être menée de manière décomposée. Par exemple, dans le cas du problème (4.1), on peut s'arranger pour que chaque problème auxiliaire se scinde en N sous-problèmes auxiliaires ne dépendant chacun que d'un sous-vecteur de la variable u, qui se met donc sous la forme $(u_1, \ldots, u_N)$, de telle sorte que l'on ait à résoudre N « petits » sous-problèmes auxiliaires plutôt qu'un seul problème auxiliaire de grande taille. Dans ce cadre déterministe, le principe de la méthode de décomposition est donc de reconstituer la solution $u^\sharp$ du problème initial par concaténation des sous-vecteurs $u_i^\sharp$, limites des suites des solutions des sous-problèmes auxiliaires.

On notera que, dans le cas déterministe, il n'existe pas de différence fondamentale entre le problème statique (4.1) et le problème dynamique (2.26) :

les méthodes algorithmiques mises en œuvre pour résoudre l'un ou l'autre de ces problèmes sont très semblables puisque le premier consiste à chercher une solution dans l'espace $\mathcal{U}$ alors que la solution du second se trouve dans l'espace-produit $\mathcal{U}^T$. La principale différence vient du fait que l'on peut utiliser plus ou moins habilement le temps dans la formulation et la résolution du problème (2.26), par exemple en calculant les gradients à l'aide de la technique de l'état-adjoint.

6.2 Extensions au cas stochastique

Le but de la deuxième partie de cet ouvrage est de proposer des méthodes permettant d'étendre au cas *stochastique* les techniques de décomposition et de coordination pour des problèmes de type (4.1) ou (2.26).

Dans toute la suite, les variables aléatoires seront notées avec des lettres majuscules grasses, comme par exemple $\boldsymbol{W}$, et les réalisations des variables aléatoires seront notées avec des lettres minuscules normales, comme par exemple w.

6.2.1 Cas statique

Problème

On considère pour commencer le problème (4.1), en oubliant dans un premier temps les contraintes explicites (4.1b). On suppose que la fonction J s'écrit comme l'espérance d'une fonction j dépendant d'une variable u appartenant à un espace $\mathcal{U}$ et d'une variable aléatoire $\boldsymbol{W}$ définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans un espace $\mathcal{W}$ muni d'une tribu $\mathcal{W}$. Ainsi, la même valeur de u s'applique pour toutes les valeurs w que peut prendre la variable aléatoire $\boldsymbol{W}$:

$$J(u) = \mathbb{E}\left(j(u, \boldsymbol{W})\right) \ .$$

Le problème d'optimisation stochastique à résoudre est alors :

$$\min_{u \in U^{\mathrm{ad}}} \ \mathbb{E}\left(j(u, \boldsymbol{W})\right) \ . \tag{6.1}$$

Cette situation est qualifiée de *statique*, au sens où la valeur optimale $u^\sharp$ que l'on cherche ne s'adapte pas aux valeurs que prend $\boldsymbol{W}$ ([1]) : on dit que l'on est en *boucle ouverte* pour signifier que, en tant que fonction de w, la commande optimale $u^\sharp$ est une *constante*.

D'un point de vue théorique, on peut calculer, en tout point u par lequel cheminera l'algorithme d'optimisation choisi pour résoudre le problème, la valeur $J(u)$ de la fonction objectif et son gradient $\nabla J(u)$, si bien que le problème

1. La valeur optimale $u^\sharp$ dépend par contre de la *loi de probabilité* de la variable aléatoire $\boldsymbol{W}$.

d'optimisation stochastique (6.1) se ramène au problème déterministe (4.1) : les méthodes de décomposition s'y appliquent alors sans difficulté particulière. Cependant, d'un point de vue pratique, chaque évaluation de $J(u)$ ou de $\nabla J(u)$ nécessite un calcul d'espérance, ce qui peut s'avérer très consommateur en temps de calcul sur un ordinateur, surtout si l'espace $\mathcal{W}$ dans lequel la variable aléatoire $\boldsymbol{W}$ prend ses valeurs est de grande dimension...

Gradient stochastique

Une approche mieux adaptée que la précédente à ce type de problème est celle du *gradient stochastique* dont le principe est de faire évoluer *simultanément* le calcul de l'espérance *et* la méthode de gradient, en s'appuyant sur la méthode de Monte Carlo pour l'évaluation des espérances. L'idée de la méthode consiste à tirer une suite $(w^1, \ldots, w^k, \ldots)$ de réalisations de la variable aléatoire $\boldsymbol{W}$ suivant sa loi de probabilité, et à faire évoluer la variable u pour chaque nouvelle valeur de l'aléa en effectuant un pas de gradient sur la fonction j en fixant $\boldsymbol{W}$ à cette valeur. L'évolution de u doit être suffisamment lente pour que le phénomène de moyenne lié à l'espérance se fasse au cours des itérations. Plus précisément, la forme générale de l'algorithme de gradient stochastique est la suivante :

- on se donne $u^0 \in U^{\mathrm{ad}}$ pour initialiser l'algorithme,
- à l'itération k, on effectue un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$ suivant sa loi (pour le calcul de l'espérance), et on effectue un pas de gradient en u (pour l'optimisation) :

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

  où ε^k est le terme d'une suite de réels positifs qui converge « lentement » vers zéro[2].

On constate que derrière le gradient stochastique se trouve une idée de nature variationnelle, que l'on devrait donc pouvoir adapter au cadre de la décomposition-coordination.

Cette seconde partie de l'ouvrage va porter sur l'étude du gradient stochastique, sa généralisation, sa convergence et son efficacité. Dans la mesure où le PPA permet de traiter des problèmes d'optimisation avec contraintes explicites, on étendra le gradient stochastique au cas des contraintes Θ *déterministes*.

Interprétation intuitive

On a dit que l'idée du gradient stochastique consiste à profiter des itérations d'optimisation pour effectuer le calcul d'espérance. Pour rendre

2. car si ε^k converge trop vite vers zéro, l'algorithme peut s'arrêter trop tôt...

cette affirmation plus concrète, on considère l'algorithme de gradient stochastique, en supposant que l'ensemble $U^{\rm ad}$ est égal à l'espace $\mathcal{U}$ tout entier et en choisissant le pas ε^k égal à $1/k$ ([3]) :

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

Sommant k fois cette formule à partir d'un indice k_0 donné et supposant :

- d'une part que la fonction : $u \mapsto \nabla_u j(u, w)$ est suffisamment régulière,
- d'autre part que l'indice k_0 est suffisamment « grand » pour que l'on soit proche de la convergence de l'algorithme, et donc pour que les $u^{k_0+\ell}$ soient peu différents les uns des autres pour $0 \leq \ell \leq k-1$,

on obtient la première approximation suivante :

$$\begin{aligned} u^{k_0+k} &= u^{k_0} - \sum_{\ell=0}^{k-1} \frac{1}{k_0+\ell}\nabla_u j(u^{k_0+\ell}, w^{k_0+\ell+1}) \, , \\ &\approx u^{k_0} - \sum_{\ell=0}^{k-1} \frac{1}{k_0+\ell}\nabla_u j(u^{k_0}, w^{k_0+\ell+1}) \, . \end{aligned}$$

Pour simplifier les écritures, on introduit la notation :

$$a^{k_0+\ell} = \nabla_u j(u^{k_0}, w^{k_0+\ell+1}) \, . \tag{6.2}$$

Supposant alors que l'indice k est suffisamment « petit » devant k_0, on remplace les termes $1/(k_0+\ell)$ par leur développement limité au premier ordre, d'où la nouvelle approximation :

$$\begin{aligned} u^{k_0+k} &\approx u^{k_0} - \frac{1}{k_0}\sum_{\ell=0}^{k-1}\Big(1-\frac{\ell}{k_0}\Big)a^{k_0+\ell} \, , \\ &\approx u^{k_0} - \frac{k}{k_0}\Bigg(\underbrace{\frac{1}{k}\sum_{\ell=0}^{k-1} a^{k_0+\ell}}_{T_1} + \underbrace{\frac{1}{k}\sum_{\ell=0}^{k-1}\frac{\ell}{k_0}a^{k_0+\ell}}_{T_2}\Bigg) \, . \end{aligned}$$

- De la définition (6.2) de $a^{k_0+\ell}$, on déduit que le terme T_1 est une approximation de type Monte Carlo du gradient $\nabla J(u^{k_0})$ pourvu que l'indice k soit suffisamment « grand ».
- Supposant que la norme du gradient partiel par rapport à u de la fonction j est uniformément bornée par une constante C, on obtient la majoration $\|T_2\| \leq (k/k_0)C$.
- Le terme k/k_0 en facteur des termes T_1 et T_2 correspond au pas du gradient dans l'expression précédente.

3. On verra au Chapitre 7 que $1/k$ est un choix « canonique » pour le pas de l'algorithme du gradient stochastique.

On fait alors le choix :

$$k = k_0^{\frac{1}{3}} ,$$

qui permet de concilier les hypothèses effectuées sur les indices k et k_0 pourvu que k_0 soit « grand ». Ce choix est tel que le pas de gradient $k/k_0 = k_0^{-2/3}$, qui tend vers zéro avec k_0, est le terme d'une série divergente. De plus, le terme T_2 tend vers zéro lorsque k_0 tend vers l'infini et peut donc être négligé, de telle sorte que l'équation de récurrence s'écrit :

$$u^{k_0+k} \approx u^{k_0} - k_0^{-\frac{2}{3}} \nabla J(u^{k_0}) .$$

Sous cette forme, l'algorithme du gradient stochastique s'interprète comme une méthode de gradient classique, et il apparaît clairement que le gradient stochastique utilise les itérations d'optimisation pour reconstituer l'espérance du *gradient de la fonction objectif* plutôt que l'espérance de la fonction objectif elle-même.

6.2.2 Cas dynamique

Les problèmes d'optimisation stochastique dynamique, c'est à dire ceux pour lesquels les variables de décision dépendent d'observations effectuées sur les variables aléatoires du problème [4], ne seront pas traités dans le cadre de cet ouvrage. Cependant, on présente brièvement quelques caractéristiques de ces problèmes afin de souligner les différences avec le cas statique. On se limite ici au cas des problèmes de type commande optimale, qui sont les plus emblématiques du cas dynamique. On pourra par exemple consulter le Chapitre 1 de Carpentier et collab. (2015) pour une présentation plus détaillée des problèmes d'optimisation stochastique dans le cadre dynamique.

Problème

L'extension du problème (2.26) au cas stochastique est de considérer un système dynamique soumis à chaque pas de temps à une perturbation aléatoire $\boldsymbol{W}_t$ dont les réalisations sont notées w_t. La dynamique décrivant l'évolution du système consiste en une relation définissant la condition initiale (aléatoire) du système :

$$x_0 = f_{-1}(w_0) , \tag{6.3a}$$

et une relation définissant l'évolution du système sur le pas de temps $[t, t+1[$:

$$x_{t+1} = f_t(x_t, u_t, w_{t+1}) , \quad t = 0, \dots, T-1 , \tag{6.3b}$$

dans laquelle le choix des indices temporels des arguments de la fonction f_t est là pour indiquer que l'on choisit au pas de temps $[t, t+1[$ la valeur de

4. alors que dans le cas statique vu au §6.2.1, la même valeur de la variable de décision u s'applique pour toutes les valeurs de la variable aléatoire $\boldsymbol{W}$...

la commande u_t *avant* de connaître la valeur w_{t+1} de la perturbation $\boldsymbol{W}_{t+1}$ affectant ce même pas de temps. On dira alors que la structure d'information du problème est en *Décision-Hasard* (DH), car on décide d'abord de la commande, le hasard n'étant révélé que dans un deuxième temps. Comme les w_t sont des réalisations des variables aléatoires $\boldsymbol{W}_t$, et comme les décisions u_t sont elles aussi des réalisations de variables aléatoires $\boldsymbol{U}_t$ (ce point sera détaillé au paragraphe suivant), on déduit des relations (6.3a) et (6.3b) que les x_t sont elles aussi des réalisations de variables aléatoires $\boldsymbol{X}_t$. Le but de l'optimisation est alors de minimiser une fonction coût de la forme :

$$\mathbb{E}\left(\sum_{t=0}^{T-1} L_t(\boldsymbol{X}_t, \boldsymbol{U}_t, \boldsymbol{W}_{t+1}) + K(\boldsymbol{X}_T)\right) , \tag{6.3c}$$

où le coût instantané L_t dépend de l'état $\boldsymbol{X}_t$, de la commande $\boldsymbol{U}_t$ et est lui aussi perturbé par l'aléa $\boldsymbol{W}_{t+1}$, tandis que la fonction K représente une pénalisation de l'état $\boldsymbol{X}_T$ qu'atteint le système à la fin de la période d'optimisation. L'opération d'espérance dans la relation (6.3c) est effectuée par rapport à l'ensemble des variables aléatoires du problème.

On suppose qu'il n'y a pas de dynamique caché dans l'évolution des aléas au cours du temps, ce qui revient à dire que les bruits $\boldsymbol{W}_t$ et $\boldsymbol{W}_{t'}$ à deux pas de temps différents t et t' sont des variables aléatoires indépendantes.

Structure d'information

Pour donner un sens au problème, il faut préciser la nature des variables par rapport auxquelles on optimise. Dans le cas considéré ci-dessus, il paraît souhaitable que la valeur de la commande u_t à mettre en œuvre sur le pas de temps $[t, t+1[$ dépende des aléas qui ont affectés le système jusqu'au début de ce pas de temps, car la connaissance de ce qui s'est passé sur le système permet en général de mieux l'optimiser. Le calcul de la commande u_t doit donc se baser sur l'ensemble des informations disponibles sur le système au début du pas de temps $[t, t+1[$. Le fait que l'on n'utilise que les informations passées pour calculer u_t correspond à une commande respectant le principe de *causalité*, selon lequel il ne serait pas rationnel de faire dépendre la commande d'informations non disponibles au moment de la décision, comme par exemple les valeurs *futures* des perturbations.

On formalise ces considérations de la manière suivante. On suppose qu'une *observation* z_t est effectuée sur le système au début de chaque pas de temps $[t, t+1[$, cette observation s'écrivant par exemple sous la forme :

$$z_t = h_t(x_t, w_t) , \quad t = 0, \ldots, T-1 . \tag{6.3d}$$

La totalité des *informations disponibles* au début du pas de temps $[t, t+1]$ est elle-même une fonction des observations passées, que l'on note :

$$y_t = C_t(z_0, \ldots, z_t) , \quad t = 0, \ldots, T-1 . \tag{6.3e}$$

Un cas particulier important est celui où il n'y a pas de perte d'information au cours du temps :

$$y_t = (z_0, \dots, z_t) \ .$$

On parle alors de système à *mémoire parfaite*. D'autres cas sont envisageables, comme celui de la *mémoire instantanée* où l'on ne se rappelle que de la dernière observation effectuée :

$$y_t = z_t \ .$$

Comme on l'a déjà noté, il est raisonnable de chercher la commande u_t dans la classe des fonctions ne dépendant que de l'information y_t disponible sur le système au moment où l'on prend la décision. Cette contrainte peut se mettre sous la forme fonctionnelle suivante :

$$u_t = g_t(y_t) \ . \tag{6.3f}$$

Une commande de la forme (6.3f) est dite en *boucle fermée* sur les observations, par opposition au cas de la boucle ouverte rencontré dans le cas statique où la commande est choisie indépendamment des valeurs prises par les perturbations et donc de toute information. Bien sûr, lorsque l'on n'observe rien (c'est-à-dire lorsque l'observation est la même quels que soient les aléas affectant le système), la commande u_t est associée à une fonction constante, et l'on retrouve le cas de la boucle ouverte.

Une fois fixées les fonctions de commande $g_0, \dots, g_{T-1}$, les variables x_t, z_t, y_t et u_t du problème (6.3) deviennent toutes des *variables aléatoires*. L'évaluation de l'espérance du coût (6.3c) est alors possible, et l'optimisation consiste à minimiser cette espérance par rapport aux fonctions $g_0, \dots, g_{T-1}$.

Remarque 6.1. On voit ainsi apparaître une différence fondamentale entre les situations déterministe et stochastique, puisque l'optimisation dans le premier cas se fait par rapport à des constantes, alors qu'elle se fait par rapport à des fonctions dépendant des valeurs w_t des perturbations $\boldsymbol{W}_t$ dans le second cas. Les notions de causalité et de boucle fermée n'ont d'ailleurs aucun sens en déterministe, puisque l'évolution future du système à un pas de temps donné ne dépend que des commandes appliquées et peut donc être anticipée par l'optimisation.

Résolution

On distingue classiquement les trois cas suivants dans l'optimisation des systèmes dynamiques stochastiques.

– *Cas markovien.* On observe sans perturbation l'état x_t du système[5] :

$$z_t = x_t \quad \text{et} \quad y_t = (z_0, \dots, z_t) \ .$$

5. On rappelle que l'on suppose que les bruits à deux pas de temps différents sont des variables aléatoires indépendantes, de telle sorte que les seules équations régissant l'évolution au cours du temps du système sont les équations (6.3b).

On montre alors que les fonctions $g_0^\sharp, \ldots, g_{T-1}^\sharp$ réalisant la solution optimale du problème (6.3) ne dépendent en fait que de l'état instantané du système : la loi de commande optimale sur le pas de temps $[t, t+1[$ est de la forme :

$$u_t = g_t^\sharp(x_t) \ .$$

Pour calculer le coût optimal et les lois de commande associées, on dispose de la méthode de la *programmation dynamique*, qui consiste à résoudre, en tout point de l'espace dans lequel vit l'état x_t du système, une équation récurrente en temps rétrograde comportant un opérateur de minimisation. On notera que, dans le cas de l'observation sans perturbation de l'état, ce résultat reste inchangé si l'information y_t est égale à l'observation z_t seule plutôt qu'à tout le passé de ces observations.

– *Cas de la structure d'information classique.* L'observation z_t que l'on fait sur le système à chaque pas de temps est de la forme générale (6.3d), mais on suppose que l'accumulation des informations au cours du temps se fait sans perte de mémoire (mémoire parfaite) :

$$z_t = h_t(x_t, w_t) \ , \quad \text{et} \ \ y_t = (z_0, \ldots, z_t) \ .$$

La résolution du problème de commande optimale est possible et se fait de la manière suivante :
  – on écrit d'abord l'équation du filtre régissant l'évolution au cours du temps de la loi de probabilité de l'état x_t conditionné par l'observation disponible y_t ;
  – on résout ensuite une équation de programmation dynamique, qui se développe dans l'espace (a priori de dimension infinie) des lois de probabilité conditionnelle de l'état.

– *Cas général.* On ne sait pas écrire les conditions d'optimalité associées au problème.

Pour un exposé détaillé de la commande optimale stochastique dans le cas markovien et dans le cas de l'information classique (aussi appelé cas de l'information incomplète), on se référera à Puterman (2009), ou encore au cours Quadrat et Viot (2000). Pour l'analyse d'un (contre-) exemple dans le cas général, on pourra consulter l'article Witsenhausen (1968). On pourra aussi se référer à Carpentier et collab. (2015) pour une présentation des problèmes d'optimisation stochastique dans le cas dynamique.

Programmation dynamique et décomposition

La mise en œuvre de la méthode de la programmation dynamique n'est en pratique possible que pour les systèmes dont l'état x_t est à valeurs dans un espace de petite dimension (inférieure ou égale à 3 pour fixer les idées). Ce phénomène est connu sous le nom de la *malédiction de la dimension* et empêche en pratique l'utilisation directe de la programmation dynamique sur la plupart des systèmes industriels et financiers en vraie grandeur.

Si l'on cherche alors, pour l'optimisation des grands systèmes dynamiques stochastiques, à marier la technique de programmation dynamique aux méthodes de décomposition-coordination, on se trouve confronté à une nouvelle difficulté. Pour l'illustrer, on considère un système se décomposant en N sous-systèmes interagissant entre eux, l'état x et la commande u du système initial se mettant respectivement sous la forme $(x_1, \ldots, x_N)$ et $(u_1, \ldots, u_N)$. On suppose que l'on est dans le cas markovien : la loi de commande du sous-système i au pas de temps $[t, t+1[$ est donc de la forme :

$$u_{i,t} = g_t(x_{1,t}, \ldots, x_{N,t}) \ ,$$

et dépend donc de l'état de *tous* les sous-systèmes. La résolution de l'équation de programmation dynamique associée au sous-système i doit être développée sur l'état du système tout entier, ce qui ne correspond pas à l'idée de formuler des sous-systèmes correspondant chacun à un sous-vecteur d'état $x_{i,t}$, comme on s'y attendrait dans un processus de décomposition-coordination. Dans ce cas, c'est la classe dans laquelle on cherche les lois de commande qui ne permet pas d'effectuer la décomposition du système.

Cette impossibilité s'explique par le fait que l'on cherche ici à décomposer l'espace de départ des fonctions g_t définissant les commandes (espace dans lequel vit l'état x) alors que la décomposition fonctionne bien lorsque l'on décompose l'espace d'arrivée (espace dans lequel vit la commande u). En effet, pour une application f définie sur un espace $\mathcal{A}$ et à valeurs dans un espace $\mathcal{B}$, une décomposition en produit cartésien de l'espace d'arrivée $\mathcal{B}$ induit le même type de décomposition pour l'application f. Mais ce n'est pas le cas lorsque l'on considère une décomposition de l'espace de départ $\mathcal{A}$, l'application $f(a_1, \ldots, a_N)$ n'étant en général pas représentable par une collection d'applications $f_i(a_i)$.

Remarque 6.2. On pourrait décider arbitrairement de limiter l'optimisation à la classe des lois de commande décentralisées, de la forme :

$$u_{i,t} = g_{i,t}(x_{i,t}) \ .$$

La décomposition et la mise en œuvre de la programmation dynamique par sous-système redeviennent possibles dans certains cas, au prix d'un calcul global d'espérance (voir par exemple Delebecque et Quadrat (1978)). Mais on se contente alors de chercher une solution *sous-optimale* du problème, et il est facile de se rendre compte sur des exemples simples que la classe des lois de commande décentralisée peut conduire à des solutions tellement sous-optimales qu'elles sont inacceptables. Ainsi, dans le cas de deux unités de production indépendantes ayant à satisfaire conjointement une consommation, si l'on suppose que la première unité produit à faible coût tout en étant sujette à des pannes fréquentes alors que la seconde unité produit à coût élevé sans aucune panne, il est clair que la commande optimale de la deuxième unité consiste à se substituer à la première unité en cas de panne de cette dernière, ce qui montre bien qu'ignorer l'état de la première unité dans la commande de la seconde ne peut pas être satisfaisant.

Programmation stochastique

Puisque l'on est dans l'incapacité d'optimiser les systèmes dynamiques stochastiques de grande taille par la programmation dynamique, on se tourne vers d'autres techniques pour résoudre le problème, en particulier le formalisme de la *programmation stochastique*[6]. L'idée générale est de *ne pas transporter les lois de probabilité* dans les espaces où vivent l'état et la commande (comme le fait la programmation dynamique), mais de représenter ces variables d'état et de commande comme des variables aléatoires (définies sur l'espace de probabilité original Ω). La discrétisation du problème est alors possible (par exemple par une approche de type Monte Carlo), mais il faut s'assurer qu'elle respecte la structure d'information associée au problème.

Sans rentrer dans trop de détails, précisons un peu cet autre point de vue.

1. Puisqu'on a vu au paragraphe précédent que le problème vient de la décomposition de l'espace de départ, une façon d'éviter cet écueil est de manipuler des variables aléatoires, qui sont donc des fonctions définies sur l'espace Ω. On ne cherchera pas à décomposer cet espace, et donc on ne considérera que le cas de la décomposition de l'espace d'arrivée des variables aléatoires.
2. Il faut alors renoncer aux lois de feedbacks et trouver une autre façon d'exprimer les dépendances informationnelles entre les différentes variables aléatoires. Ceci est décrit dans Carpentier et collab. (2015), notamment dans le contexte de problèmes du type commande optimale.

Par exemple, le point de vue programmation stochastique utilise la technique des arbres de scénarios pour représenter les contraintes de causalité. L'exploitation de ce point de vue dans un objectif de décomposition est illustré dans Carpentier et collab. (1995).

6.3 Un exemple statique et dynamique

On étudie un exemple mélangeant les caractéristiques des problèmes (6.1) et (6.3), et représentant un problème typique de l'optimisation stochastique dans lequel on cherche à réaliser le meilleur compromis entre des coûts d'investissement et de fonctionnement, par exemple dans un réseau de distribution de services (télécommunication, électricité), dans le transport aérien...

1. **Investissement**. On doit choisir la capacité u d'un outil de production. Ce choix est fait une fois pour toutes, en connaissant la distribution de probabilité des aléas pouvant affecter le système, mais sans connaître la valeur de l'aléa qui se réalisera : la commande u est donc en boucle ouverte, et on note $\alpha(u)$ le coût associé à l'installation de la capacité u.

6. On pourra consulter le site `http://stoprog.org/` qui contient un grand nombre d'informations (présentations, livres, articles, logiciels, exemples...) relatives à la communauté de la programmation stochastique.

2. **Fonctionnement**. Une fois l'investissement u réalisé, le fonctionnement du système est perturbé par une variable aléatoire $\boldsymbol{W}$ que l'on observe. À investissement u et à réalisation w de $\boldsymbol{W}$ connus, on dispose d'une commande v permettant d'agir afin de satisfaire le service rendu par le système, et dont la mise en œuvre induit un coût noté $\beta(u, v, w)$. La commande v dépend de u et w, et est donc en boucle fermée sur l'aléa. En tant que fonction de w, la commande v peut être vue comme la réalisation d'une variable aléatoire $\boldsymbol{V}$ mesurable par rapport à $\boldsymbol{W}$.

Le problème consiste à minimiser par rapport à u et $\boldsymbol{V}$ l'espérance du coût total, soit :

$$\min_{(u,\boldsymbol{V})} \mathbb{E}\big(\alpha(u) + \beta(u, \boldsymbol{V}, \boldsymbol{W})\big) \ . \tag{6.4}$$

On va comparer deux approches pour la résolution de ce problème, l'une basée sur le respect de la structure d'information associée (qui fournira la solution du problème), et l'autre plus directe (qui conduira à un échec).

(A) On suppose que l'on est capable de résoudre le problème d'optimisation lié au seul fonctionnement : pour des valeurs données u^k et w^{k+1} de l'investissement et de la variable aléatoire $\boldsymbol{W}$, on effectue la minimisation :

$$\min_{v} \beta\big(u^k, v, w^{k+1}\big) \ , \tag{6.5}$$

dont la solution optimale, notée v^{k+1}, est en fait une fonction $v^\sharp(u^k, w^{k+1})$ dépendant à la fois des valeurs u^k et w^{k+1} de l'investissement et de la perturbation. Une fois le problème du fonctionnement optimal à investissement donné résolu, on peut appliquer l'algorithme du gradient stochastique à la seule variable d'investissement u, qui est quant à elle en boucle ouverte sur l'aléa.

On est donc conduit à mettre en œuvre l'algorithme suivant :

- on se donne une valeur initiale u^0,
- au pas k de l'algorithme,
  - on tire une valeur w^{k+1} de $\boldsymbol{W}$, indépendant des tirages précédents,
  - on résout complètement [7] le problème de minimisation en v :

$$v^{k+1} = \arg\min_{v} \beta\big(u^k, v, w^{k+1}\big) \ , \tag{6.6a}$$

  - on effectue une étape de gradient de pas ε^k sur u :

$$u^{k+1} = u^k - \varepsilon^k\big(\nabla\alpha(u^k) + \nabla_u\beta(u^k, v^{k+1}, w^{k+1})\big) \ . \tag{6.6b}$$

Cet algorithme fournit la solution $u^\sharp$ du problème (6.4). Une fois l'investissement optimal $u^\sharp$ déterminé, la résolution du problème de fonctionnement seul pour cet investissement et pour une valeur donnée w de la variable aléatoire $\boldsymbol{W}$ permet de calculer $v^\sharp(u^\sharp, w)$, meilleure façon de piloter le système. La commande optimale en boucle fermée $\boldsymbol{V}^\sharp$ est alors : $\boldsymbol{V}^\sharp = v^\sharp(u^\sharp, \boldsymbol{W})$.

7. en supposant que ce problème admet une unique solution

Remarque 6.3. On a utilisé un résultat « classique » en optimisation (voir (Cohen, 2000, Exercice 4.69)) qui dit que, si une fonction ϕ est le résultat de la minimisation d'une fonction J par rapport à l'un de ses arguments :

$$\phi(u) = \min_v J(u, v) \ ,$$

alors, en notant $\widehat{v}(u)$ une solution de ce problème, et sous des hypothèses « raisonnables » de convexité, de continuité et de différentiabilité de la fonction J, la fonction ϕ est elle aussi différentiable, et que son gradient est égal au gradient partiel de J par rapport à u évalué au point $\widehat{v}(u)$ qui réalise le minimum :

$$\nabla\phi(u) = \nabla_u J(u, \widehat{v}(u)) \ .$$

(B) Un inconvénient pratique de l'algorithme (6.6) est que l'on est amené à effectuer à chaque itération une minimisation complète en la variable v. Dans (6.6b), les pas ε^k ont pour rôle de permettre un effet de moyenne sur les réalisations w^{k+1} afin d'approximer l'espérance, tandis que les itérées v^{k+1} fournies par (6.6a) « suivent » les tirages w^{k+1}. On peut alors penser à un algorithme alternatif « à deux vitesses » dans lequel la variable v est mise à jour par un algorithme de gradient « à grands pas » permettant de suivre « à vitesse suffisante » les sauts successifs des tirages de $\boldsymbol{W}$. On propose donc l'algorithme suivant :

- on se donne des valeurs initiales u^0 et v^0,
- au pas k de l'algorithme,
  - on tire une valeur w^{k+1} de $\boldsymbol{W}$,
  - on effectue une étape de gradient de pas ρ sur v :

$$v^{k+1} = v^k - \rho\nabla_v\beta(u^k, v^k, w^{k+1}) \ ,$$

  - on effectue une étape de gradient de pas ε^k sur u :

$$u^{k+1} = u^k - \varepsilon^k\big(\nabla\alpha(u^k) + \nabla_u\beta(u^k, v^{k+1}, w^{k+1})\big) \ .$$

Cependant, cet algorithme ne fournit pas la solution en u du problème (6.4), car on peut montrer que la limite de la suite $\{u^k\}$ qu'il engendre dépend du paramètre ρ que l'on a choisi pour faire évoluer l'algorithme en v. Intuitivement, ceci est dû au fait que la récurrence sur les variables v^k ne converge pas assez vite par rapport à l'évolution « lente » des variables u^k, et ce malgré le « grand pas » ρ. On se référera à (Culioli, 1987, pp. 114–117) pour une discussion plus approfondie de ce point.

En conclusion, on doit donc être très méfiant quant à la possibilité de dérouler des méthodes de type gradient stochastique traitant au même niveau des variables en boucle ouverte et des variables en boucle fermée.

7

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

7.1 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 $(\mathcal{W}, \mathcal{W})$. On note $\mu = \mathbb{P} \circ \boldsymbol{W}^{-1}$ la loi de probabilité sur $\mathcal{W}$ résultant du transport de la loi $\mathbb{P}$ par $\boldsymbol{W}$. On se donne un espace de Hilbert $\mathcal{U}$ (dont le produit scalaire et la norme sont notés $\langle \cdot\,, \cdot \rangle$ et $\|\cdot\|$), une partie convexe fermée non vide U^{ad} de $\mathcal{U}$ et une fonction j définie sur $\mathcal{U} \times \mathcal{W}$ à valeurs dans $\overline{\mathbb{R}}$. On note $J(u)$ l'espérance de la fonction $j(u, \boldsymbol{W})$ (supposée intégrable pour tout $u \in U^{\mathrm{ad}}$) :

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

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 :

$$\nabla J(u) = \mathbb{E}\left(\nabla_u j(u, \boldsymbol{W})\right) \ , \tag{7.1}$$

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

$$\min_{u \in U^{\mathrm{ad}}} J(u) \ , \tag{7.2}$$

déjà considéré au §3.1.1. 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 (7.2) un algorithme de type gradient (gradient conjugué, quasi-Newton, Newton, ...). Le plus simple de ces algorithmes est celui du gradient projeté (voir §3.3.2), dont une itération s'écrit :

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

où ε est un scalaire positif représentant la longueur du pas de gradient. Dans cette approche, on s'attaque en fait au problème *déterministe* (7.2) et l'aspect stochastique est caché dans le calcul de $\nabla J(u^k)$ comme une espérance (voir (7.1)). 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 $\mathcal{W}$ dont la dimension peut être grande.

On considère alors le problème (7.2), dans lequel on remplace $J(u)$ par son expression en fonction de j :

$$\min_{u \in U^{\mathrm{ad}}} \mathbb{E}\left(j(u, \boldsymbol{W})\right) \ . \tag{7.3}$$

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 (7.3) par l'approximation suivante :

$$\min_{u \in U^{\mathrm{ad}}} \frac{1}{k} \sum_{l=1}^{k} j(u, w^l) \ , \tag{7.4}$$

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 (7.4) 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 (7.3).

7.2 Algorithme du gradient stochastique

On présente ici l'algorithme du gradient stochastique et le cadre probabiliste associé.

7.2.1 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 appelle k-échantillon de la variable aléatoire $\boldsymbol{W}$ 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 7.1 (Algorithme du gradient stochastique).

1. Choisir un $u^0 \in U^{\mathrm{ad}}$ initial, et une suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ 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 :

$$u^{k+1} = \mathrm{proj}_{U^{\mathrm{ad}}}\left(u^k - \varepsilon^k \nabla_u j(u^k, w^{k+1})\right) \ . \tag{7.5}$$

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 §7.6.

Remarque 7.2. L'Algorithme 7.1 correspond à la mise en œuvre numérique de la méthode du gradient stochastique. Il est nécessaire, lorsque l'on souhaite étudier les propriétés de convergence 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 7.1 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 7.1 peut être interprétée comme une relation de récurrence sur des variables aléatoires $\boldsymbol{U}^k$ à valeurs dans l'espace $\mathcal{U}$:

$$\boldsymbol{U}^{k+1} = \mathrm{proj}_{U^{\mathrm{ad}}}\left(\boldsymbol{U}^k - \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1})\right) \ , \tag{7.6}$$

l'opérateur de projection dans l'équation (7.6) devant être interprété comme une projection « ω par ω » dans l'espace $\mathcal{U}$, et les valeurs u^k obtenues par application de (7.5) correspondant à une réalisation des variables aléatoires $\boldsymbol{U}^k$ données par (7.6) :

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

Dans la formulation (7.6), l'Algorithme 7.1 engendre une suite de variables aléatoires $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ dont on peut étudier la convergence au sens probabiliste. Dans la suite de cet ouvrage, on passera sans précaution particulière d'une écriture de type (7.5) portant sur des « valeurs » à une écriture de type (7.6) portant sur des « variables aléatoires », ce changement ne posant comme on vient de le voir aucune difficulté.

Remarque 7.3. Pour que la description en termes de variables aléatoires de l'algorithme de gradient stochastique soit valide, il faut être capable de construire une suite $\{\boldsymbol{W}^k\}_{k\in\mathbb{N}^\star}$ (où $\mathbb{N}^\star$ est l'ensemble des entiers naturels non nuls) 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 $\widetilde{\mathcal{W}} = \mathcal{W}^{\mathbb{N}}$ 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 $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$ comme étant les *applications coordonnées*[2] :

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

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 $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$ 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 $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{W}}, \widetilde{\mu})$. Dans toute la suite, pour simplifier les notations, cet espace produit sera noté $(\Omega, \mathcal{A}, \mathbb{P})$. On notera que cet espace est suffisamment « gros » pour contenir un échantillon de taille infinie de $\boldsymbol{W}$, ce qui est une condition indispensable pour pouvoir mettre en œuvre la relation de récurrence (7.6) de l'algorithme du gradient stochastique.

7.2.2 Exemple

On va illustrer l'algorithme 7.1 dans le cadre de l'estimation statistique, et plus précisément comme une 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 :

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

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

$$\boldsymbol{U}^k = \frac{1}{k} \sum_{l=1}^{k} \boldsymbol{W}^l \ . \tag{7.7}$$

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 (7.7), on a :

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

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

Posant $\varepsilon^k = 1/(k+1)$ 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 :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k - \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) \,. \tag{7.8}$$

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 (7.8) s'interprète comme l'Algorithme 7.1 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 ε^k tend vers zéro lorsque k tend vers l'infini, alors que le pas d'un algorithme de gradient classique est constant ; cependant, ε^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.

7.2.3 Cadre probabiliste

Une itération de la méthode du gradient stochastique (7.6) peut se mettre sous la forme générale suivante :

$$\boldsymbol{U}^{k+1} = \mathcal{R}^k\left(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}\right) \,. \tag{7.9}$$

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.

3. Voir l'Annexe 2.3, dans laquelle est discuté le choix d'une série divergente (plutôt que convergente) dans le cas d'un algorithme de sous-gradient.

– L'utilisation récursive de la relation (7.9) 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 :

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

– Définissant la fonction φ^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 165, la notion de convergence adaptée à l'étude de la suite engendrée par la relation (7.9) est celle de la convergence presque sûre :

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

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).

7.3 Premiers résultats

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

$$\min_{u \in U^{\mathrm{ad}}} \mathbb{E}\left(j(u, \boldsymbol{W})\right) \ , \tag{7.10}$$

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

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

$$\sum_{k\in\mathbb{N}} \varepsilon^k = +\infty \quad , \quad \sum_{k\in\mathbb{N}} \left(\varepsilon^k\right)^2 < +\infty \ . \tag{7.11}$$

On fait alors les hypothèses suivantes.

Hypothèses 7.5.

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 $j(\cdot, w) : \mathcal{U} \to \mathbb{R}$ est propre[4], convexe, s.c.i. (semi-continue inférieurement), différentiable pour tout $w \in \mathcal{W}$.
3. Le gradient partiel de j par rapport à u est borné uniformément en u et en w([5]) :
$$\exists m > 0,\ \forall u \in U^{\mathrm{ad}},\ \forall w \in \mathcal{W},\ \left\|\nabla_u j(u, w)\right\| \leq m \ .$$
4. Le problème (7.10) 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 (7.10) et où $\mathrm{dist}_{U^\sharp}(\cdot)$ est la fonction distance à l'ensemble $U^\sharp$.
5. La suite $\left\{\varepsilon^k\right\}_{k\in\mathbb{N}}$ est une σ-suite décroissante.

7.3.1 Convergence

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

4. On dit qu'une fonction à valeurs réelles est propre si elle n'est jamais égale à $-\infty$ et si elle n'est pas identiquement égale à $+\infty$.
5. Cette hypothèse très forte sera discutée au §7.3.4.

Théorème 7.6. *Sous les hypothèses 7.5, la suite* $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ *constituée par les variables aléatoires engendrée par l'Algorithme 7.1 converge en moyenne quadratique vers l'ensemble* $U^\sharp$ *:*

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

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 7.1 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 un opérateur non expansif[6] et l'hypothèse 7.5-3, on a :

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

Comme on l'a vu dans la Remarque 7.2, cette dernière inégalité peut se réécrire en termes de variables aléatoires :

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

Prenant de part et d'autre de l'inégalité l'espérance conditionnelle par rapport à la sous-tribu $\mathcal{F}^k = \sigma(\boldsymbol{W}^1, \ldots, \boldsymbol{W}^k)$, utilisant les propriétés de mesurabilité des variables aléatoires et 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 7.5-4, on obtient[7] :

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

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

$$\mathbb{E}\big(\boldsymbol{D}^{k+1}\big) \le \big(1 - 2\varepsilon^k c\big)\mathbb{E}\big(\boldsymbol{D}^k\big) + \big(\varepsilon^k\big)^2 m^2\ . \tag{7.12}$$

Par récurrence, on montre que, pour k_0 donné et pour tout $n \in \mathbb{N}$, on a :

6. Voir l'Exercice 4.8 pour cette propriété.

7. L'hypothèse 7.5-3 et le fait que $j(u, \boldsymbol{W}(\cdot))$ soit intégrable impliquent, par un argument de convergence dominée, que le gradient de j par rapport à u est lui aussi intégrable.

$$\mathbb{E}\big(\boldsymbol{D}^{k_0+n+1}\big) \leq \left(\prod_{l=0}^{n}\big(1-2\varepsilon^{k_0+l}c\big)\right)\mathbb{E}\big(\boldsymbol{D}^{k_0}\big) + \left(\sum_{l=0}^{n}\big(\varepsilon^{k_0+l}\big)^2\right)m^2 \,.$$

Comme la suite de terme général $\displaystyle\prod_{l=0}^{k}\big(1-2\varepsilon^{l}c\big)$ converge vers zéro (voir la Proposition 7.8) et que la suite de terme général $\displaystyle\sum_{l=0}^{k}\big(\varepsilon^{l}\big)^2$ converge (hypothèse 7.5-5), on en déduit que la suite de terme général $\mathbb{E}\big(\boldsymbol{D}^{k}\big)$ converge vers zéro, d'où le résultat annoncé. □

7.3.2 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 7.7. *Sous les hypothèses du Théorème 7.6, et en choisissant une suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ de la forme :*

$$\varepsilon^k = \frac{1}{c\;k+\frac{m^2}{c\;d^0}} \,,$$

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

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

Preuve. On repart de l'inégalité (7.12) :

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

et on choisit une suite ε^k de la forme :

$$\varepsilon^k = \frac{\gamma}{\alpha k+\beta} \,,$$

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). □

7.3.3 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 :

$$\frac{1}{k}\left(\frac{m}{c}\right)^2 . \tag{7.13}$$

Par l'hypothèse 7.5-3, on a que la constante m est une borne supérieure de l'écart type de la norme du gradient de la fonction j([8]). De plus, utilisant l'hypothèse 7.5-4 et les inégalités (3.42) et (3.4), on montre que la constante c est une borne inférieure de la constante de forte convexité de la fonction J. Cette interprétation sera utilisée §7.5 pour la comparaison de différentes versions de l'algorithme du gradient stochastique.

7.3.4 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 7.5-3 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 §7.2 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 7.5-3.

On donnera au §8.2 un théorème de convergence presque sûre très général pour une famille d'algorithmes incluant l'Algorithme 7.1. Dans ce théorème, établi dans le cadre convexe, l'hypothèse 7.5-3 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 :

$$\exists c_1 > 0,\ \exists c_2 > 0,\ \forall w \in \mathcal{W},\ \forall u \in U^{\mathrm{ad}}\ ,\quad \|\nabla_u j(u,w)\| \leq c_1\,\|u\| + c_2\ . \tag{7.14}$$

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 (voir la définition (3.44)). La démonstration du théorème correspondant fait appel à des outils probabilistes évolués comme la théorie des quasi-martingales et sera détaillée au Chapitre 8.

8. Soit $\boldsymbol{X} = \|\nabla_u j(u, \boldsymbol{W})\|$. Alors, $\mathrm{Var}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{X}^2) - \mathbb{E}(\boldsymbol{X})^2 \leq \mathbb{E}(\boldsymbol{X}^2) \leq m^2$.

7.3.5 Lemme technique

On a utilisé dans la preuve du théorème de convergence la propriétés suivante.

Proposition 7.8. *Soit* $\{\varepsilon^k\}_{k\in\mathbb{N}}$ *une suite décroissante de réels positifs telle que* $\varepsilon^k \to 0$ *et* $\sum \varepsilon^k = +\infty$. *Alors, pour tout* $\alpha > 0$, *la suite de terme général* $\{\rho^k\}_{k\in\mathbb{N}}$ *avec :*

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

converge vers zéro.

Preuve. Notant k_0 le premier indice tel que l'on ait $0 \leq 1 - \alpha\varepsilon^l \leq 1$ pour tout $l \geq k_0$ (cet indice existe car $\varepsilon^k \to 0$), on se ramène, à une constante multiplicative près, au cas où le produit définissant le terme ρ^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 :

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

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

7.4 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 $\mathcal{U}$ l'espace de Hilbert $\mathbb{R}^n$ et on considère une fonction $h : \mathcal{U} \to \mathcal{U}$, 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 :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k + \varepsilon^k \big(h(\boldsymbol{U}^k) + \boldsymbol{\xi}^{k+1}\big) \ . \tag{7.15}$$

Cet algorithme est très fortement lié à celui du gradient stochastique. En effet, dans le cas où l'ensemble U^{ad} est l'espace $\mathcal{U}$ tout entier, la projection sur U^{ad} correspond à l'identité et l'algorithme du gradient stochastique 7.1 s'écrit alors :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k - \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) \ . \tag{7.16}$$

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

$$h(u) = -\nabla J(u) \ , \tag{7.17a}$$

$$\boldsymbol{\xi}^{k+1} = \nabla J(\boldsymbol{U}^k) - \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) \ , \tag{7.17b}$$

la formule de mise à jour du gradient stochastique (7.16) est identique à celle de l'approximation stochastique (7.15). On notera que trouver un point $u^\sharp \in \mathcal{U}$ 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 (7.2). Le changement de signe dans la relation (7.17a) entre la fonction h et le gradient ∇J est dû à la différence historique de point de vue entre la communauté de l'approximation stochastique et celle de l'optimisation.

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 (7.15). 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 $\{\mathcal{F}^k\}_{k\in\mathbb{N}}$.

Dans tout le §7.4, l'espace de Hilbert $\mathcal{U}$ sera l'espace $\mathbb{R}^n$ et l'ensemble admissible U^{ad} sera égal à l'espace $\mathcal{U}$ tout entier. Si cette dernière restriction peut être levée sans aucune difficulté dans le cadre du Théorème 7.10 traitant de la convergence (en utilisant le fait que l'opérateur de projection est non expansif), il n'en est pas de même pour le Théorème 7.13 concernant la vitesse asymptotique de convergence, qui ne peut être établi que dans le cas $U^{\mathrm{ad}} = \mathcal{U}$.

7.4.1 Théorème de Robbins-Monro

On s'intéresse d'abord à la convergence de la suite de variables aléatoires $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ engendrée par (7.15). Pour cela, on fait les hypothèses suivantes.

Hypothèses 7.9.

1. La variable aléatoire $\boldsymbol{U}^0$ est $\mathcal{F}^0$-mesurable.
2. La fonction $h : \mathcal{U} \to \mathcal{U}$ est continue et vérifie les propriétés suivantes :
   – $\exists\, u^\sharp \in \mathcal{U},\ h(u^\sharp) = 0$ et $\langle h(u)\,, u - u^\sharp\rangle < 0,\ \forall u \neq u^\sharp$;
   – $\exists\, a > 0,\ \forall u \in \mathcal{U},\ \|h(u)\|^2 \leq a\big(1 + \|u\|^2\big)$.
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 $\{\varepsilon^k\}_{k\in\mathbb{N}}$ est une σ-suite.

On remarquera que l'hypothèse 7.9-2, dont la première propriété est de même nature que l'inéquation variationnelle (3.4), implique que $u^\sharp$ est l'unique zéro de la fonction h.

Théorème 7.10. *Sous les hypothèses 7.9, la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ de variables aléatoires engendrées par (7.15) converge presque sûrement vers $u^\sharp$.*

Voir (Duflo, 1997, §1.4) pour la preuve.

On peut faire le lien entre les hypothèses du Théorème 7.10 et celles que l'on pourrait formuler pour obtenir la solution du problème (7.3) 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 (7.17) 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 7.9-2 en découle (existence et unicité de la solution du problème (7.3)). De même, la première partie de l'hypothèse 7.9-3 est une conséquence immédiate de (7.17). Pour ce qui concerne la seconde partie des hypothèses 7.9-2 et 7.9-3, elles sont reliées à l'hypothèse de gradient linéairement borné (7.14) qui, par la propriété $(a+b)^2 \leq 2(a^2+b^2)$, implique que

$$\exists\, c_3 > 0,\ c_4 > 0,\ \forall u \in \mathbb{R}^n,\ \forall w \in \mathcal{W},\ \left\|\nabla_u j(u,w)\right\|^2 \leq c_3 \left\|u\right\|^2 + c_4 \ ,$$

ce qui entraîne

$$\left\|\nabla J(u)\right\|^2 \leq c_3 \left\|u\right\|^2 + c_4 \ .$$

On notera que les hypothèses faites sur la fonction j sont naturelles dans le cadre de l'optimisation convexe. On donnera un résultat de convergence plus général de l'algorithme de gradient stochastique au Chapitre 8.

7.4.2 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 (7.15). 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 préciser la notion de σ-suite qui a été donnée par la Définition 7.4.

Définition 7.11. *Une suite de réels positifs* $\{\varepsilon^k\}_{k\in\mathbb{N}}$ *est une* $\sigma(\alpha,\beta,\gamma)$-suite *si elle est telle que :*

$$\varepsilon^k = \frac{\alpha}{k^\gamma + \beta} \ ,$$

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 7.9 déjà faites pour l'étude de la convergence les nouvelles hypothèses suivantes.

Hypothèses 7.12.

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[9].

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 $\{\varepsilon^k\}_{k\in\mathbb{N}}$ est une $\sigma(\alpha,\beta,\gamma)$-suite.
5. La matrice $H - \lambda I$ est définie positive, le coefficient λ étant défini par :

$$\lambda = \begin{cases} 0 & \text{si } \gamma < 1 \,, \\ \frac{1}{2\alpha} & \text{si } \gamma = 1 \,. \end{cases} \tag{7.18}$$

On notera que, dans le cadre du problème d'optimisation initial (7.3) 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 7.12-2 correspond quant 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, dit « de la limite centrale », précisant la vitesse à laquelle les itérées $\boldsymbol{U}^k$ engendrées par (7.15) convergent vers $u^\sharp$.

Théorème 7.13. *Sous les hypothèses 7.9 et 7.12, la suite* $\{(\varepsilon^k)^{-\frac{1}{2}}(\boldsymbol{U}^k - u^\sharp)\}_{k\in\mathbb{N}}$ *converge en distribution vers la loi normale centrée de matrice de covariance* Σ *:*

$$\frac{1}{\sqrt{\varepsilon^k}}\left(\boldsymbol{U}^k - u^\sharp\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Sigma) \,, \tag{7.19}$$

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

$$(H - \lambda I)\Sigma + \Sigma(H - \lambda I) = \Gamma \,. \tag{7.20}$$

Voir (Duflo, 1996, Chapitre 4) pour la preuve.

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

9. Le symbole O correspond à la notation « Grand O » : $f(x) = \mathrm{O}(g(x))$ 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 7.14. *Soit A 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 :*

$$A\Sigma + \Sigma A^\top = \Gamma \,, \tag{7.21}$$

et cette solution a pour expression :

$$\Sigma = \int_0^{+\infty} e^{-tA}\Gamma e^{-tA^\top} \mathrm{d}t \,. \tag{7.22}$$

Remarque 7.15. Ce résultat reste vrai si la matrice Γ est symétrique semi-définie positive : dans ce cas, la matrice Σ donnée par (7.22) est elle aussi semi-définie positive, et est solution de l'équation (7.21).

Utilisant explicitement le fait que, dans le Théorème 7.13, les pas ε^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 7.13 se réécrit sous la forme :
$$k^{\frac{\gamma}{2}}\left(\boldsymbol{U}^k - u^\sharp\right) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0,\alpha\Sigma\big) \,. \tag{7.23}$$
On constate que le coefficient β n'a aucune influence sur le comportement asymptotique de l'algorithme [10].
2. De la relation (7.23), 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 (7.23) ne tient pas. En effet, la solution Σ de l'équation de Lyapunov (7.20) dépend de λ, et donc de α (voir (7.18)), 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 ($\mathcal{U} = \mathbb{R}$), H et Γ sont des réels et la solution de l'équation de Lyapunov (7.20) 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 7.12-5.

10. Ceci était prévisible car β devient rapidement négligeable au dénominateur devant le terme en k. Le 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 (7.23). 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.

7.5 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 ε^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.

7.5.1 Algorithme de Newton stochastique

Pour appliquer cette idée à l'algorithme du gradient stochastique, on se donne une matrice A carrée de dimension n symétrique définie positive. On garde dans la forme des pas ε^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 (7.16) 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 :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k + \frac{A}{k+\beta}\big(h(\boldsymbol{U}^k) + \boldsymbol{\xi}^{k+1}\big) \,. \tag{7.24}$$

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 7.12-5 devient :

Hypothèses 7.16.
La matrice $AH - \dfrac{I}{2}$ est définie positive.

Le Théorème 7.13 s'applique alors avec Ah comme champ de vecteurs, $A\boldsymbol{\xi}^k$ comme bruits et avec $\lambda = 1/2$, et implique pour la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ engendrée par l'algorithme à gain matriciel (7.24) le résultat de convergence suivant :

$$\sqrt{k}(\boldsymbol{U}^k - u^\sharp) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_A) \,, \tag{7.25}$$

la matrice de covariance asymptotique Σ_A étant donnée par :

$$\big(AH - \frac{I}{2}\big)\Sigma_A + \Sigma_A\big(HA - \frac{I}{2}\big) = A\Gamma A \,. \tag{7.26}$$

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 (7.24), que l'on appelle alors « algorithme de Newton stochastique ».

Théorème 7.17. *Le choix $A^\sharp = H^{-1}$ comme gain dans la relation (7.24) minimise la variance asymptotique Σ_A définie par (7.26) 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 7.13 correspondant à l'algorithme (7.24) peut toujours se mettre sous la forme :

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

Reportant cette expression dans (7.26), 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 7.14 et la Remarque 7.15, 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 7.18. 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 (7.24) 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 qui reste de toute façon 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 7.7 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 7.19. *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 qui soit Newton-efficace ?* La difficulté vient du fait que l'algorithme à gain matriciel optimal ne peut pas être directement mis en œuvre car le gain 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.

7.5.2 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 $\mathcal{U}$ tout entier, la phase de mise à jour classique :

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

par le calcul en deux étapes suivant :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k - \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) \,, \tag{7.27a}$$

$$\boldsymbol{U}_{\mathrm{M}}^{k+1} = \frac{1}{k+1}\sum_{l=1}^{k+1} \boldsymbol{U}^l \,, \tag{7.27b}$$

dans lequel la première étape (7.27a) est identique à celle du gradient stochastique classique, la deuxième étape (7.27b) 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 (7.27b) :

$$\boldsymbol{U}_{\mathrm{M}}^{k+1} = \boldsymbol{U}_{\mathrm{M}}^{k} + \frac{1}{k+1}\big(\boldsymbol{U}^{k+1} - \boldsymbol{U}_{\mathrm{M}}^{k}\big) \,. \tag{7.27c}$$

On remarquera que, par le théorème de Césaro (Hardy (2000)), 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 7.10, et en particulier avec des pas ε^k de la forme :

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

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 (7.27), appelé « algorithme du gradient stochastique moyenné », tient à ses propriétés asymptotiques. Les hypothèses que l'on fait alors sont semblables à celles ayant permis d'établir le Théorème 7.13 de la limite centrale, mais on restreint l'hypothèse 7.12-4 au cas où le coefficient γ est *strictement* inférieur à 1.

Hypothèses 7.20.
La suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ 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 7.13 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 7.21. *Sous les hypothèses 7.9 et 7.12, dans laquelle on remplace 7.12-4 par l'hypothèse 7.20, l'algorithme* (7.27) *du gradient stochastique moyenné est Newton-efficace :*

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

Voir (Duflo, 1996, Chapitre 4) pour la preuve.

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}.

7.6 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.

7.6.1 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 $\|u^{k+1} - u^k\|$ car cette différence converge mécaniquement vers zéro par le biais des hypothèses faites sur les pas ε^k. Le gradient $\nabla_u j(u^k, w^{k+1})$ n'a lui 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 :

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

on doit être capable de construire un test d'arrêt raisonnable, au moins dans le cas sans contrainte ($U^{\mathrm{ad}} = \mathcal{U}$). 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.

7.6.2 Réglage de l'algorithme standard

La deuxième question porte sur la forme de la suite de pas $\left\{\varepsilon^k\right\}_{k\in\mathbb{N}}$. On a vu lors de l'études de la vitesse de convergence qu'il est raisonnable de prendre des pas ε^k de la forme $\frac{1}{k^\gamma}$, avec $\frac{1}{2} < \gamma \leq 1$. D'après le Théorème 7.13, 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 $\alpha\Sigma$ varie 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 ε^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 ε^k. On citera :

1. la méthode de projection de 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 (1958), dont l'idée 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 à valeurs entières $\boldsymbol{N}^k$ 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 :

$$\boldsymbol{\varepsilon}^k = \frac{\alpha}{\boldsymbol{N}_k^\gamma + \beta}\,;$$

3. une règle multiplicative d'adaptation du pas (voir 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 ε^k mal choisie. Une bonne règle de conduite consiste à ne diminuer le pas ε^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)).

7.6.3 Réglage de l'algorithme moyenné

On a montré l'algorithme du gradient stochastique moyenné était Newton-efficace à la condition de choisir des pas ε^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 ε^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 (7.27a) s'est approché de la zone de convergence, ce qui permet de ne pas tenir compte dans la moyenne des premières itérées de l'algorithme.

7.6.4 Illustration numérique

On considère l'exemple quadratique suivant :

$$\min_{u\in\mathbb{R}^{10}} \mathbb{E}\Big(\frac{1}{2}u^\top Au + \boldsymbol{W}^\top u\Big)\,, \tag{7.28}$$

où A est une matrice symétrique définie positive et où $\boldsymbol{W}$ est un vecteur aléatoire gaussien à valeurs dans $\mathbb{R}^{10}$, de moyenne μ et de matrice de covariance Γ. La solution de ce problème est : $u^\sharp = -A^{-1}\mu$, que l'on peut approximer par l'estimateur de Monte Carlo $\widehat{\boldsymbol{U}}^k$:

$$\widehat{\boldsymbol{U}}^k = -\frac{1}{k}\sum_{l=1}^{k} A^{-1}\boldsymbol{W}^l\,. \tag{7.29}$$

Cet estimateur de $u^\sharp$ est *efficace*, ce qui veut dire que sa variance atteint la borne de Cramer-Rao (voir Kay (1993) pour plus de détails sur l'estimation statistique) :

$$k\mathrm{Var}\big(\widehat{\boldsymbol{U}}^k\big) = A^{-1}\Gamma A^{-1}\,. \tag{7.30}$$

On a utilisé dans cet exemple les données numériques suivantes :

$$A = \mathrm{diag}(1,2,3,4,5,6,7,8,9,10) \quad , \quad \mu = (0\,0\,0\,0\,0\,0\,0\,0\,0\,0)^\top\,,$$

et

$$\Gamma = \begin{pmatrix}
11.13 & 1.40 & 1.19 & 1.92 & 1.49 & 1.34 & 1.52 & 1.28 & 1.04 & 1.27\\
1.40 & 12.29 & 1.20 & 1.92 & 2.09 & 2.42 & 0.84 & 1.70 & 1.31 & 1.63\\
1.19 & 1.20 & 12.56 & 2.06 & 2.19 & 1.23 & 1.40 & 0.41 & 1.08 & 1.46\\
1.92 & 1.92 & 2.06 & 12.04 & 1.89 & 1.05 & 1.96 & 1.53 & 1.53 & 1.86\\
1.49 & 2.09 & 2.19 & 1.89 & 12.85 & 0.68 & 1.26 & 0.72 & 0.64 & 0.83\\
1.34 & 2.42 & 1.23 & 1.05 & 0.68 & 13.21 & 1.49 & 1.78 & 1.05 & 0.80\\
1.52 & 0.84 & 1.40 & 1.96 & 1.26 & 1.49 & 12.82 & 1.58 & 1.18 & 2.08\\
1.28 & 1.70 & 0.41 & 1.53 & 0.72 & 1.78 & 1.58 & 11.59 & 1.21 & 1.09\\
1.04 & 1.31 & 1.08 & 1.53 & 0.64 & 1.05 & 1.18 & 1.21 & 11.60 & 1.69\\
1.27 & 1.63 & 1.46 & 1.86 & 0.83 & 0.80 & 2.08 & 1.09 & 1.69 & 12.20
\end{pmatrix}.$$

Gradient stochastique standard

Utilisant des pas de longueur $\varepsilon^k = \alpha/(k+\beta)$, l'application de l'algorithme du gradient stochastique standard sur l'exemple (7.28) conduit à :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k - \frac{\alpha}{k+\beta}\big(A\boldsymbol{U}^k + \boldsymbol{W}^{k+1}\big)\ . \tag{7.31}$$

La Figure 7.1 montre quatre exécutions de l'algorithme avec différentes valeurs du coefficient α, le quotient α/β restant constant et égal à $0,1$. Pour chaque valeur de α, on a représenté l'évolution au cours des itérations de la norme de l'estimateur de Monte Carlo (courbe noire) et de la norme de la variable $\boldsymbol{U}^k$ fournie par l'algorithme du gradient stochastique (courbe gris clair). De manière évidente, choisir une valeur « faible » $\alpha = 0,3$ (graphique en haut à gauche) fait que l'algorithme ne converge pas [11], alors que choisir une valeur « grande » $\alpha = 5,0$ ou $\alpha = 10,0$ (graphiques en bas) conduit à un comportement oscillatoire excessif. Pour cet exemple, le choix $\alpha = 1$ (graphique en haut à droite) peut être considéré comme optimal.

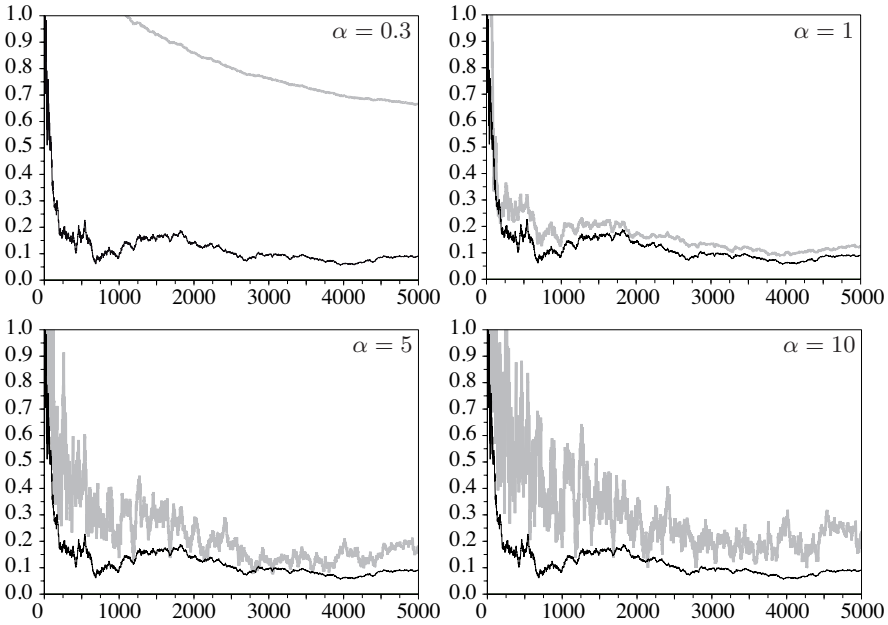

Fig. 7.1. Gradient stochastique standard pour différentes valeurs de α

Afin de préciser le comportement asymptotique de l'algorithme du gradient stochastique, on calcule la matrice de covariance des itérées $\boldsymbol{U}^k$. De l'équation (7.31), on déduit :

$$\begin{aligned}\mathrm{Var}\big(\boldsymbol{U}^{k+1}\big) &= \mathrm{Var}\Big(\big(I - \varepsilon^k A\big)\boldsymbol{U}^k - \varepsilon^k \boldsymbol{W}^{k+1}\Big)\\ &= \big(I - \varepsilon^k A\big)\mathrm{Var}\big(\boldsymbol{U}^k\big)\big(I - \varepsilon^k A\big) + \big(\varepsilon^k\big)^2 \Gamma\ ,\end{aligned}$$

11. du moins en un temps raisonnable...

où I représente la matrice identité de taille 10. La limite de la suite de ces matrices de covariance normalisées $k\mathrm{Var}\left(\boldsymbol{U}^k\right)$ peut alors être comparée à la borne de Cramer-Rao (7.30). Les valeurs propres minimale $\lambda_{\min}$ et maximale $\lambda_{\max}$ de ces matrices limite sont données dans le Tableau 7.1 pour différentes valeurs du couple (α, β). On remarque que la plus grande valeur propre de la borne

Algorithme du gradient stochastique standard	$\lambda_{\min}$	$\lambda_{\max}$
$\alpha = 0.3$ — $\beta = 3.0$	0.192	6170.542
$\boldsymbol{\alpha = 1.0}$ — $\boldsymbol{\beta = 10.0}$	0.556	**11.286**
$\alpha = 5.0$ — $\beta = 50.0$	2.664	32.056
$\alpha = 10.0$ — $\beta = 100.0$	5.299	60.936
Borne de Cramer-Rao	**0.108**	**11.258**

Tableau 7.1. Valeurs propres extrêmes de la matrice de covariance asymptotique de l'algorithme du gradient stochastique standard

de Cramer-Rao coïncide avec celle de la « meilleure » matrice de covariance (obtenue avec le choix $\alpha = 1$). Cette dernière remarque illustre le résultat donné dans Dodu et collab. (1981), que l'on a rappelé au Théorème 7.7, et qui dit que la plus grande valeur propre de la matrice de covariance normalisée « optimale » est de l'ordre de $(m/c)^2$, où c (resp. m) est la constante de forte convexité (resp. la borne supérieure de la norme du gradient) de la fonction objectif.

Gradient stochastique moyenné

On applique ensuite à l'exemple (7.28) l'algorithme du gradient stochastique moyenné, ce qui conduit à :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k - \frac{\alpha}{k^\gamma + \beta}\left(A\boldsymbol{U}^k + \boldsymbol{W}^{k+1}\right),$$
$$\boldsymbol{U}_{\mathrm{M}}^{k+1} = \frac{1}{k+1}\sum_{l=1}^{k+1}\boldsymbol{U}^l .$$

On utilise les mêmes valeurs de α et β que celles prises pour l'algorithme du gradient stochastique standard, mais on choisit alors un coefficient γ strictement inférieur à 1, égal à 2/3. Les quatre exécutions de l'algorithme moyenné sont représentées sur la Figure 7.2. Pour chaque exécution, on a représenté l'évolution au cours des itérations de la norme de l'estimateur de Monte Carlo (7.29) (courbe noire), de la norme de la variable $\boldsymbol{U}^k$ fournie par l'algorithme du gradient stochastique avant moyennisation (courbe gris clair) et de la norme de la variable $\boldsymbol{U}_{\mathrm{M}}^k$ obtenue après moyenne (courbe gris foncé). On constate comme auparavant qu'augmenter le paramètre α (de $0,3$ à $10,0$) amplifie les oscillations dans l'algorithme du gradient stochastique avant moyennisation. Par contre, le comportement de l'algorithme après

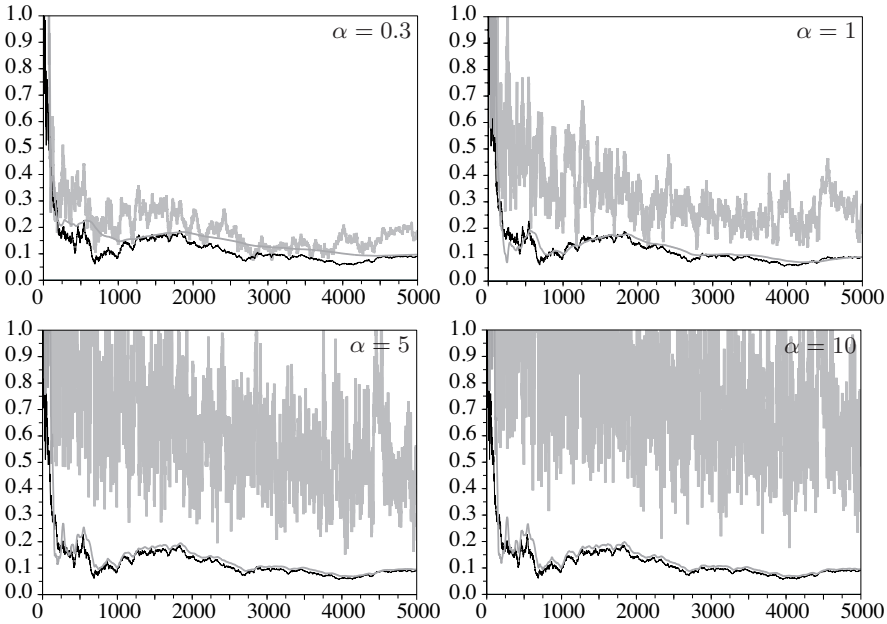

Fig. 7.2. Gradient stochastique moyenné pour différentes valeurs de α

moyenne est remarquablement stable et ne dépend pratiquement pas de la valeur du coefficient α, d'où le qualificatif de « robuste » donné à l'algorithme moyenné.

Comme dans le cas de l'algorithme standard, il est possible de calculer la matrice de covariance des itérées $\boldsymbol{U}_{\mathrm{M}}^{k}$ et d'obtenir les valeurs propres minimale et maximale de la limite de cette suite de matrices pour différents choix du paramètre α. Ces valeurs propres extrêmes sont données dans le Tableau 7.2. On observe alors que le spectre complet associé à la borne de Cramer-Rao est

Algorithme du gradient stochastique moyenné	$\lambda_{\min}$	$\lambda_{\max}$
$\alpha = 0.3$ — $\beta = 3.0$	0.108	11.360
$\alpha = 1.0$ — $\beta = 10.0$	0.108	11.288
$\alpha = 5.0$ — $\beta = 50.0$	0.108	11.264
$\alpha = 10.0$ — $\beta = 100.0$	0.108	11.262
Borne de Cramer-Rao	**0.108**	**11.258**

Tableau 7.2. Valeurs propres extrêmes de la matrice de covariance asymptotique de l'algorithme du gradient stochastique moyenné

obtenu, et cela quelle que soit la valeur du paramètre α. Cette constatation expérimentale est en accord avec le Théorème 7.13 caractérisant la vitesse de convergence du gradient stochastique moyenné.

7.7 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 (Chapitre 8 et Chapitre 9), on va étudier en détail la convergence du gradient stochastique, d'abord dans le cas sans contraintes, puis dans le cas des contraintes déterministes. Cette étude se fera dans le cadre du Principe du Problème Auxiliaire.

Puis, au dernier chapitre (Chapitre 10), 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.

Il faut aussi noter que les méthodes de type gradient stochastique sont depuis quelques années très populaires dans la communauté « Machine Learning ». On pourra consulter les articles Bousquet et Bottou (2008) et Bach et Moulines (2013). On pourra enfin consulter avec profit l'article Nemirovski et collab. (2009) dans lequel les auteurs mélangent l'idée du gradient stochastique avec celle de la « mirror descent method » afin d'obtenir un algorithme performant. Anticipant les résultats des deux chapitres à suivre, on pourra noter la grande proximité d'idée entre la mirror descent method et le Principe du Problème Auxiliaire.

8

Le Principe du Problème Auxiliaire en optimisation stochastique sur un ensemble admissible

Dans le cadre de l'optimisation convexe déterministe, on a présenté au Chapitre 3 le Principe du Problème Auxiliaire (PPA), qui consiste à remplacer le problème :

$$\min_{u \in U^{\mathrm{ad}}} J(u) \ , \tag{8.1}$$

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

$$\min_{u \in U^{\mathrm{ad}}} \Big(K(u) + \big\langle \varepsilon \nabla J(u^k) - \nabla K(u^k) \, , u \big\rangle \Big) \ , \tag{8.2}$$

et sa solution u^{k+1} permet de formuler le problème auxiliaire d'indice $k+1$. La fonction K est *choisie* par l'utilisateur, et on l'appelle la « fonction auxiliaire ». On rappelle les principales propriétés de ce principe.

- Le PPA permet de *retrouver* les algorithmes classiques d'optimisation : ainsi, avec un choix de fonction auxiliaire $K(u) = \|u\|^2/2$, le problème (8.2) prend la forme :

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

  dont la solution s'obtient de manière analytique :

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

  ce qui correspond à un algorithme de gradient avec un pas fixe ε.
- Étant donnée une décomposition de u sous la forme $u = (u_1, \ldots, u_N)$, le PPA permet de *décomposer* chaque problème auxiliaire (8.2) en N sous-problèmes indépendants, pourvu que la contrainte $u \in U^{\mathrm{ad}}$ se mette sous la forme de N contraintes $u_i \in U_i^{\mathrm{ad}}$ et pourvu que l'on choisisse une fonction auxiliaire K additive par rapport à la décomposition de u.

On va maintenant étendre ce principe au cas stochastique en le mélangeant à l'idée du gradient stochastique. En effet, l'algorithme du PPA étant par nature itératif, on peut profiter de ces itérations pour accumuler simultanément les tirages indépendants du bruit aux fins d'approximation des espérances dans l'esprit du gradient stochastique décrit au chapitre précédent.

8.1 Algorithme du Principe du Problème Auxiliaire stochastique

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

On considère le problème de minimisation (7.3) :

$$\min_{u \in U^{\mathrm{ad}}} J(u)\,, \tag{8.3}$$

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

$$\min_{u \in U^{\mathrm{ad}}} \left(K(u) + \left\langle \varepsilon \nabla J(u^k) - \nabla K(u^k)\,, u \right\rangle \right).$$

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

$$\min_{u \in U^{\mathrm{ad}}} \left(K(u) + \left\langle \varepsilon^k \nabla_u j(u^k, w^{k+1}) - \nabla K(u^k)\,, u \right\rangle \right), \tag{8.4}$$

où w^{k+1} est un tirage de la variable aléatoire $\boldsymbol{W}$. On en déduit l'algorithme suivant, dit « du PPA stochastique ».

Algorithme 8.1.

1. Choisir un point initial $u^0 \in U^{\mathrm{ad}}$, et une suite $\{\varepsilon^k\}_{k \in \mathbb{N}}$ de réels positifs.
2. À l'itération k, effectuer un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$.
3. Résoudre le problème auxiliaire (8.4), dont la solution est notée u^{k+1}.
4. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Remarque 8.2. Avec la fonction auxiliaire $K(u) = \|u\|^2/2$, le problème auxiliaire (8.4) se met sous la forme :

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

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

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

ce qui est précisément l'Algorithme 7.1 du gradient stochastique.

Remarque 8.3. Dans le Chapitre 3, on s'est placé dans le cas où la fonction objectif J s'écrit comme la somme de deux fonctions $\boldsymbol{\mathcal{J}}$ et $\mathfrak{J}$, la partie $\boldsymbol{\mathcal{J}}$ étant linéarisée lors de l'utilisation du PPA et la partie $\mathfrak{J}$ étant gardée en l'état. On rappelle que cette distinction prend tout son intérêt dans le cadre de la décomposition, la partie $\mathfrak{J}$ étant alors supposée additive par rapport à la décomposition de l'espace $\mathcal{U}$ et la partie $\boldsymbol{\mathcal{J}}$ étant remplacée par son approximation linéaire au premier ordre. Dans le cadre du PPA stochastique, on se contente d'étudier le cas où $\mathfrak{J}$ est identiquement nulle, ce qui permet d'alléger (un peu) les écritures. Il n'y a cependant pas de difficulté supplémentaire à considérer le cas général. On se reportera à (Culioli, 1987, §2.5.1) pour plus de détails.

De même que dans le cas de l'Algorithme 7.1, une hypothèse fondamentale pour la convergence de cet algorithme du PPA stochastique est que les tirages $(w^1, \dots, w^k)$ 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 $(\boldsymbol{W}^1, \dots, \boldsymbol{W}^k)$ *indépendantes* de même loi que $\boldsymbol{W}$. Comme on l'a noté dans la Remarque 7.2, les propriétés de convergence de l'Algorithme 8.1 s'étudient en termes de variables aléatoires : on s'intéresse donc au problème de minimisation :

$$\min_{u \in U^{\mathrm{ad}}} \left(K(u) + \left\langle \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) - \nabla K(\boldsymbol{U}^k)\,, u \right\rangle \right), \tag{8.5}$$

dans lequel il est sous-entendu que la minimisation est faite $\ll \omega$ par $\omega \gg$, de telle sorte que le résultat de la minimisation dépend de ω. Les résultats qui suivent ont été obtenus par J. C. Culioli dans le cadre de sa thèse Culioli (1987) et publiés dans Culioli et Cohen (1990).

8.2 Théorème de convergence

On considère la multi-application définie sur Ω à valeurs dans $\mathcal{U}$ qui à chaque ω associe l'arg min obtenu par résolution du problème (8.5), et on suppose que cette multi-application admet une sélection mesurable [1] notée $\boldsymbol{U}^{k+1}$.

Remarque 8.4. On fera dans le Théorème 8.5 suffisamment d'hypothèses pour que la solution du problème (8.5) soit unique, de telle sorte que la multi-application qui à ω associe l'arg min soit en fait une application. Cela ne change pas le fait qu'il faut se préoccuper de sa mesurabilité afin de pouvoir pouvoir parler de la variable aléatoire $\boldsymbol{U}^{k+1}$.

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

La question de la convergence de l'Algorithme 8.1 et celle du lien de cet algorithme avec le problème de départ (8.3) sont réglées par le théorème suivant.

Théorème 8.5. *On fait les hypothèses suivantes.*

1. U^{ad} *est une partie convexe fermée non vide de l'espace de Hilbert* $\mathcal{U}$.
2. *La fonction* $j : \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ *est une intégrande normale*[2], *et l'espérance de* $j(u, \boldsymbol{W})$ *existe pour tout* $u \in U^{\mathrm{ad}}$.
3. *Pour tout* $w \in \mathcal{W}$, *la fonction* $j(\cdot, w) : \mathcal{U} \to \mathbb{R}$ *est propre*[3], *convexe, semi-continue inférieurement, différentiable sur un sous-ensemble ouvert contenant* U^{ad}.
4. *La fonction* $j(\cdot, w)$ *est à gradient linéairement borné en* u *uniformément en* w :

$$\exists c_1 > 0,\ \exists c_2 > 0,\ \forall w \in \mathcal{W},\ \forall u \in U^{\mathrm{ad}},\ \|\nabla_u j(u, w)\| \leq c_1 \|u\| + c_2 \ . \quad (8.6)$$

5. *La fonction* J *est Lipschitzienne, coercive sur l'ensemble* U^{ad}.
6. *La fonction* K *est propre, fortement convexe de constante* b, *semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant* U^{ad}.
7. *La suite* $\{\varepsilon^k\}_{k\in\mathbb{N}}$ *est une* σ*-suite*[4].

On a alors les conclusions suivantes.

1. *Le problème* (8.3) *admet un ensemble de solutions* $U^\sharp$ *non vide.*
2. *Le problème* (8.5) *admet une solution* $\boldsymbol{U}^{k+1}$ *unique.*
3. *La suite de variables aléatoires* $\{J(\boldsymbol{U}^k)\}_{k\in\mathbb{N}}$ *converge vers* $\min_{u\in U^{\mathrm{ad}}} J(u)$, *presque sûrement.*
4. *La suite des solutions* $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ *engendrée par l'Algorithme 8.1 est bornée presque sûrement, et tout point d'accumulation d'une réalisation de cette suite appartient à l'ensemble optimal* $U^\sharp$.

Si l'on fait l'hypothèse supplémentaire que la fonction J *est fortement convexe, alors la suite* $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ *engendrée par l'Algorithme 8.1 converge presque sûrement vers l'unique solution* $u^\sharp$ *du problème* (8.3).

Remarque 8.6. Il n'y a pas de difficulté supplémentaire à faire l'analyse de convergence dans le cas où la fonction $j(\cdot, w) : \mathcal{U} \to \mathbb{R}$ est seulement sous-différentiable. Il faut alors remplacer l'hypothèse (8.6) par une hypothèse de sous-gradient linéairement borné :

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

On consultera Culioli et Cohen (1990) pour plus de détails.

2. Voir §8.5 pour cette notion.
3. Voir la note en bas de page 168.
4. Voir la Définition 7.4.

Preuve. La démonstration des deux premières conclusions de ce théorème découle des théorèmes généraux relatifs à l'optimisation convexe. Le fait que $\boldsymbol{U}^{k+1}$, unique solution du problème (8.5) soit une variable aléatoire, et donc une application mesurable, provient de ce que l'on a supposé que j est une intégrande normale (voir le Théorème 8.16 de l'Annexe 8.5). Plus précisément, supposons que $\boldsymbol{U}^k$ soit mesurable. On considère le problème (8.5) et on note :

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

La propriété de semi-continuité inférieure de la fonction auxiliaire K fait que l'intégrande $(u, w) \mapsto K(u)$ est une intégrande autonome, et donc une intégrande normale (Rockafellar et Wets, 1998, Example 14.30). Comme on a supposé que j est une intégrande normale, on en déduit par (Rockafellar et Wets, 1998, Theorem 14.56) que $\nabla_u j$ et ∇K sont aussi des intégrandes normales. Par composition avec les applications mesurables $\boldsymbol{U}^k$ et $\boldsymbol{W}^{k+1}$, on en déduit que l'application φ^k est mesurable. L'intégrande $(\omega, u) \mapsto \big\langle \varphi^k(\omega)\ , u\big\rangle$, étant mesurable en ω et continue en u, est une intégrande de Carathéodory (Rockafellar et Wets, 1998, Example 14.29). En tant que somme de deux intégrandes normales, Φ^k est elle aussi une intégrande normale. Par (Rockafellar et Wets, 1998, Theorem 14.37), on en déduit que l'application $\boldsymbol{U}^{k+1}$, unique solution du problème (8.5) est une application mesurable[5].

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

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

Pour alléger les écritures, on utilisera la notation :

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

5. Les résultats sur les intégrandes utilisés dans cette preuve se trouvent dans Rockafellar et Wets (1998), où ils sont énoncés dans le cas où l'espace $\mathcal{U}$ est de dimension finie (voir annexe, §8.5). Ceci devrait donc limiter le champ d'application du théorème aux espaces de Hilbert de dimension finie. Des résultats équivalents concernant les intégrandes sont donnés dans Hess (1995) pour le cas où $\mathcal{U}$ est un espace de Banach séparable, mais ces résultats sont d'un abord plus difficile et ne sont pas présentés dans le cadre de cet ouvrage. On considérera néanmoins que ces derniers résultats s'appliquent et donc que le Théorème 8.5 est valide pour les espaces de Hilbert de dimension infinie.

Le fait que u^{k+1} soit solution du problème (8.4) est caractérisé par la condition d'optimalité suivante :

$$\forall u \in U^{\mathrm{ad}},\ \left\langle \nabla K(u^{k+1}) - \nabla K(u^k) + \varepsilon^k g^k\,, u - u^{k+1} \right\rangle \geq 0\,. \tag{8.7}$$

1. **Choix de la fonction de Lyapunov.** Soit $u^\sharp \in U^\sharp$ une solution du problème d'optimisation (8.3). À l'instar de (3.20), on adopte la fonction de Lyapunov ℓ suivante :

$$\ell(u) = K(u^\sharp) - K(u) - \left\langle \nabla K(u)\,, u^\sharp - u \right\rangle .$$

La forte convexité de K implique la majoration (voir (3.42)) :

$$\frac{b}{2}\left\|u - u^\sharp\right\|^2 \leq \ell(u)\,, \tag{8.8}$$

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

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

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

où $\left\{u^k\right\}_{k\in\mathbb{N}}$ est la suite des solutions engendrée par l'Algorithme 8.1 pour une réalisation $(w^1, \ldots, w^k, \ldots)$ d'un échantillon de taille infinie de la variable $\boldsymbol{W}$.

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

– Par convexité de K, on a :

$$T_1 \leq 0\,.$$

– L'écriture de la condition d'optimalité (8.7) au point $u = u^\sharp$ implique :

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

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

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

– L'écriture de la condition (8.7) au point $u = u^k$ et la propriété de forte monotonie de ∇K impliquent :

$$b\left\|u^{k+1} - u^k\right\|^2 \leq \varepsilon^k \left\langle g^k\,, u^k - u^{k+1} \right\rangle .$$

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

$$\left\|u^{k+1}-u^k\right\| \leq \frac{\varepsilon^k}{b}\left\|g^k\right\| , \tag{8.9}$$

d'où, en appliquant une fois encore l'inégalité de Schwarz au terme T_4 :

$$T_4 \leq \frac{\varepsilon^k}{b}\left\|g^k\right\|^2 .$$

Une conséquence immédiate de (8.6) est qu'il existe des constantes c_3 et c_4 telles que l'on ait : $\left\|g^k\right\| \leq c_3\left\|u^k-u^\sharp\right\|+c_4$. Élevant cette inégalité au carré, utilisant $(a+b)^2 \leq 2(a^2+b^2)$ ainsi que la relation (8.8), on obtient :

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

d'où la nouvelle majoration :

$$T_4 \leq \frac{\varepsilon^k}{b}\big(\alpha\ell(u^k)+\beta\big) .$$

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

$$\ell(u^{k+1})-\ell(u^k) \leq \varepsilon^k\big(j(u^\sharp,w^{k+1})-j(u^k,w^{k+1})\big)+\frac{\big(\varepsilon^k\big)^2}{b}\big(\alpha\ell(u^k)+\beta\big) .$$

On écrit alors cette dernière inégalité en termes de variables aléatoires (voir la Remarque 7.2), et on en prend de part et d'autre l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^k$ engendrée par les k variables aléatoires $(\boldsymbol{W}^1,\ldots,\boldsymbol{W}^k)$. Comme la variable aléatoire $\boldsymbol{U}^k$ est par construction mesurable par rapport à la tribu $\mathcal{F}^k$, on en déduit que $\mathbb{E}\big(\ell(\boldsymbol{U}^k)\ \big|\ \mathcal{F}^k\big)=\ell(\boldsymbol{U}^k)$. Comme la variable aléatoire $\boldsymbol{W}^{k+1}$ est indépendante des $\boldsymbol{W}^l$ précédentes et donc de $\boldsymbol{U}^k$, on en déduit que $\mathbb{E}\big(j(\boldsymbol{U}^k,\boldsymbol{W}^{k+1})\ \big|\ \mathcal{F}^k\big)=J(\boldsymbol{U}^k)$. On obtient finalement :

$$\mathbb{E}\big(\ell(\boldsymbol{U}^{k+1})\ \big|\ \mathcal{F}^k\big)-\ell(\boldsymbol{U}^k) \leq \alpha^k\ell(\boldsymbol{U}^k)+\beta^k + \varepsilon^k\big(J(u^\sharp)-J(\boldsymbol{U}^k)\big) , \tag{8.10}$$

où $\alpha^k=\big(\alpha/b\big)\big(\varepsilon^k\big)^2$ et $\beta^k=\big(\beta/b\big)\big(\varepsilon^k\big)^2$ sont les termes de deux séries convergentes. On rappelle que le terme $J(u^\sharp)-J(\boldsymbol{U}^k)$ est presque sûrement négatif ou nul étant donné que $u^\sharp$ est une solution du problème (8.3).

3. **Analyse de convergence.** Une application directe du Théorème 8.8 de Robbins-Siegmund permet de montrer que la suite de variables aléatoires $\big\{\ell(\boldsymbol{U}^k)\big\}_{k\in\mathbb{N}}$ converge presque sûrement vers une variable aléatoire $\boldsymbol{\ell}^\infty$ bornée presque sûrement [6], et que l'on a :

6. i.e. l'ensemble $\big\{\omega\in\Omega\ \big|\ \lim_{k\to+\infty}\ell(\boldsymbol{U}^k)(\omega)=+\infty\big\}$ est de mesure nulle

$$\sum_{k=0}^{+\infty} \varepsilon^k \big(J(\boldsymbol{U}^k) - J(u^\sharp)\big) < +\infty \ , \quad \mathbb{P}\text{-p.s.} \ . \tag{8.11}$$

4. **Limites des suites.** Comme la suite $\big\{\ell(\boldsymbol{U}^k)\big\}_{k\in\mathbb{N}}$ est bornée presque sûrement, on déduit de la relation (8.8) que la suite $\big\{\boldsymbol{U}^k\big\}_{k\in\mathbb{N}}$ est elle aussi bornée presque sûrement. L'hypothèse (8.6) implique alors que la suite $\big\{\nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1})\big\}_{k\in\mathbb{N}}$ est elle aussi bornée presque sûrement. Enfin, on déduit de la relation (8.9) que la même conclusion est vraie pour la suite $\big\{\|\boldsymbol{U}^{k+1} - \boldsymbol{U}^k\|/\varepsilon^k\big\}_{k\in\mathbb{N}}$. Cette dernière propriété, associée à la relation (8.11) et au fait que la fonction J soit Lipschitzienne, permet d'utiliser le Lemme 8.10 pour conclure que la suite $\big\{J(\boldsymbol{U}^k)\big\}_{k\in\mathbb{N}}$ converge presque sûrement vers $J(u^\sharp)$, valeur optimale du problème (8.3).

   On note alors Ω_0 le sous-ensemble (de mesure nulle) de Ω sur lequel la suite $\big\{\ell(\boldsymbol{U}^k)\big\}_{k\in\mathbb{N}}$ n'est pas bornée, et Ω_1 le sous-ensemble (de mesure nulle lui aussi) de Ω sur lequel la relation (8.11) n'est pas vérifiée. On a donc : $\mathbb{P}\big(\Omega_0 \cup \Omega_1\big) = 0$.

   Soit $\omega \notin \Omega_0 \cup \Omega_1$. La suite des réalisations $\big\{u^k\big\}_{k\in\mathbb{N}}$ associée à cet élément ω est bornée et chaque u^k appartient à $U^{\rm ad}$, partie fermée de $\mathcal{U}$. Par un argument de compacité[7], on conclut que l'on peut extraire de la suite $\big\{u^k\big\}_{k\in\mathbb{N}}$ une sous-suite convergente $\big\{u^{\Phi(k)}\big\}_{k\in\mathbb{N}}$ (on remarquera que la sous-suite $\big\{\Phi(k)\big\}_{k\in\mathbb{N}}$ *dépend* de l'aléa ω). Soit $\bar{u}$ la limite de la suite $\big\{u^{\Phi(k)}\big\}_{k\in\mathbb{N}}$. Par la semi-continuité inférieure de la fonction J, on a :

$$J(\bar{u}) \le \liminf_{k\to+\infty} J(u^{\Phi(k)}) = J(u^\sharp) \ .$$

   On en déduit que $\bar{u} \in U^\sharp$.

Pour conclure, on considère le cas où la fonction J est fortement convexe de constante a. Alors, le problème (8.3) a une unique solution $u^\sharp$, caractérisée par l'inéquation variationnelle :

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

La condition de forte convexité sur J conduit à :

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

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

La convergence presque sûre de $J(\boldsymbol{U}^k)$ vers $J(u^\sharp)$ entraîne la convergence presque sûre de $\left\|\boldsymbol{U}^k - u^\sharp\right\|$ vers zéro, ce qui achève la démonstration. □

Remarque 8.7. Il existe de nombreuses situations dans lesquelles la fonction $j(\cdot, w)$ n'est pas convexe, alors que son espérance J l'est. On notera que, dans le théorème précédent, l'hypothèse de convexité faite sur la fonction j n'est pas nécessaire et peut être remplacée par une hypothèse de convexité sur son espérance J. Il suffit dans la preuve de garder le terme $T_3 = \left\langle g^k \, , u^\sharp - u^k \right\rangle$, d'en prendre l'espérance conditionnelle par rapport à $\mathcal{F}^k$, ce qui produit $\left\langle \nabla J(\boldsymbol{U}^k) \, , u^\sharp - \boldsymbol{U}^k \right\rangle$, et d'utiliser l'hypothèse de convexité sur J pour majorer ce terme par $J(u^\sharp) - J(\boldsymbol{U}^k)$.

8.3 Conclusions

On a énoncé et démontré un théorème très général de convergence pour l'algorithme issu du principe du problème auxiliaire dans le cas stochastique. Ce théorème inclut bien évidemment le cas du gradient stochastique standard (avec une fonction auxiliaire de la forme $K(u) = \|u\|^2/2$) ; il inclut aussi l'algorithme de Newton stochastique (avec pour fonction auxiliaire $K(u) = \langle u \, , Au \rangle/2$).

D'un point de vue théorique, les hypothèses sous lesquelles le Théorème 8.5 a été énoncé sont « raisonnables » dans le cadre de l'optimisation convexe. On notera en particulier que l'on ne fait pas d'hypothèse de forte convexité sur la fonction objectif du problème d'optimisation. On a pris le parti dans ce chapitre de présenter les résultats dans le cadre différentiable. On consultera Culioli (1987) et Culioli et Cohen (1990) pour des résultats analogues dans le cadre plus général des fonctions sous-différentiables.

Pour ce qui concerne la décomposition, si l'utilisation du principe du problème auxiliaire ouvre la voie à la décomposition des problèmes d'optimisation stochastique de type (8.3), cette possibilité a en pratique moins d'impact que dans le cadre déterministe : en effet, la vitesse de l'algorithme du gradient stochastique généralisé est conditionnée par les pas ε^k, et n'est donc pas très différente de la vitesse de la méthode du gradient stochastique standard, qui, comme tout algorithme de gradient, est elle-même une méthode de décomposition ! Ceci étant, le choix d'une fonction auxiliaire « proche » de la fonction objectif originale permet, comme dans le cas déterministe, de décomposer les calculs et d'améliorer le conditionnement des sous-problèmes à résoudre à chaque itération de l'algorithme du PPA stochastique.

8.4 Annexe : théorème de Robbins-Siegmund et lemme technique

Le théorème suivant, connu sous le nom de « théorème de Robbins-Siegmung », est l'un des résultats-clés de la théorie de l'approximation stochastique.

Théorème 8.8. *On considère quatre suites de variables aléatoires à valeurs réelles positives* $\{\boldsymbol{\Lambda}^k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\alpha}^k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\beta}^k\}_{k\in\mathbb{N}}$ *et* $\{\boldsymbol{\eta}^k\}_{k\in\mathbb{N}}$, *que l'on suppose toutes les quatre adaptées à une filtration* $\{\mathcal{F}^k\}_{k\in\mathbb{N}}$ *donnée. On suppose de plus que la relation suivante est vérifiée :*

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

et que l'on a :

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

Alors, la suite de variables aléatoires $\{\boldsymbol{\Lambda}^k\}_{k\in\mathbb{N}}$ *converge presque sûrement vers* $\boldsymbol{\Lambda}^\infty$, *variable aléatoire presque sûrement bornée*[8], *et on a de plus que :*

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

Voir (Duflo, 1997, Theorem 1.3.12) pour la preuve.

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

Corollaire 8.9. *On considère cinq suites de variables aléatoires à valeurs réelles positives* $\{\boldsymbol{\Lambda}^k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\alpha}^k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\beta}^k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\gamma}^k\}_{k\in\mathbb{N}}$, *et* $\{\boldsymbol{\eta}^k\}_{k\in\mathbb{N}}$, *que l'on suppose adaptées à une filtration* $\{\mathcal{F}^k\}_{k\in\mathbb{N}}$. *On suppose de plus que*

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

et que l'on a :

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

Alors, $\{\boldsymbol{\Lambda}^k\}_{k\in\mathbb{N}}$ *converge presque sûrement vers une variable aléatoire* $\boldsymbol{\Lambda}^\infty$ *bornée presque sûrement, et on a de plus :*

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

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

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

$$1+\widetilde{\alpha}^k = \frac{1+\alpha^k}{1-\gamma^k}\ ,\ \ \widetilde{\beta}^k = \frac{\beta^k}{1-\gamma^k}\ ,\ \ \widetilde{\eta}^k = \frac{\eta^k}{1-\gamma^k}\ .$$

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

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

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

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

Lemme 8.10. *Soit J une fonction définie sur un espace de Hilbert $\mathcal{U}$ à valeurs dans $\mathbb{R}$, Lipschitzienne de constante L. Soit $\{u^k\}_{k\in\mathbb{N}}$ une suite d'éléments de $\mathcal{U}$ et soit $\{\varepsilon^k\}_{k\in\mathbb{N}}$ une suite de nombres réels positifs vérifiant :*

(a) $\displaystyle\sum_{k\in\mathbb{N}} \varepsilon^k = +\infty$,

(b) $\exists \mu \in \mathbb{R},\ \displaystyle\sum_{k\in\mathbb{N}} \varepsilon^k \left|J(u^k)-\mu\right| < +\infty$,

(c) $\exists \delta > 0,\ \forall k \in \mathbb{N},\ \left\|u^{k+1}-u^k\right\| \leq \delta\varepsilon^k$.

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

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

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

(i) D'après l'hypothèse *(b)*, on a :

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

d'où l'on déduit que :

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

(ii) Les hypothèses *(a)* et *(b)* impliquent que le cardinal de l'ensemble N_α n'est pas fini.

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

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

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

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

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

d'où le résultat. □

8.5 Annexe : intégrande normale et sélection mesurable

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

On se donne un espace Ω muni de sa tribu $\mathcal{A}$, ainsi qu'un espace de Hilbert $\mathcal{U}$ de dimension finie [9].

Définition 8.11. *On dit que la multi-application $S : \Omega \rightrightarrows \mathcal{U}$ est mesurable si, pour tout ouvert $\mathcal{O} \subset \mathcal{U}$, l'ensemble $S^{-1}(\mathcal{O}) = \left\{\omega \in \Omega \mid S(\omega) \cap \mathcal{O} \neq \emptyset\right\}$ est mesurable, c'est-à-dire si $S^{-1}(\mathcal{O}) \in \mathcal{A}$.*

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

Théorème 8.12. *Une multi-application $S : \Omega \rightrightarrows \mathcal{U}$ mesurable et à valeurs fermées admet une sélection mesurable : il existe une application mesurable $s : \Omega \to \mathcal{U}$ telle que $s(\omega) \in S(\omega)$ pour tout $\omega \in \Omega$.*

9. Comme on l'a déjà mentionné dans la note en bas de page 193, on pourra trouver dans Hess (1995) des résultats concernant les intégrandes normales dans le cas où $\mathcal{U}$ est un espace de Banach séparable.

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

Définition 8.13. *Une application $f : \Omega \times \mathcal{U} \to \overline{\mathbb{R}}$ qui est mesurable par rapport à ω pour tout $u \in \mathcal{U}$ est appelée une integrande.*

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

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

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

Définition 8.14. *Une intégrande normale f est une intégrande telle que la multi-application épigraphique S_f associée soit mesurable et à valeurs fermées.*

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

Théorème 8.15. *Soit $f : \Omega \times \mathcal{U} \to \overline{\mathbb{R}}$ une intégrande normale. Alors, l'application $u \mapsto f(\omega, u)$ est semi-continue inférieurement pour tout $\omega \in \Omega$, l'application $\omega \mapsto f(\omega, u)$ est mesurable pour tout $u \in \mathcal{U}$, et l'application $\omega \mapsto f\big(\omega, \boldsymbol{U}(\omega)\big)$ est mesurable pour toute application $\boldsymbol{U}$ mesurable.*

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

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

Théorème 8.16. *Soit une intégrande normale $f : \Omega \times \mathcal{U} \to \overline{\mathbb{R}}$, à laquelle on associe l'application p définie par :*

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

et la multi-application P définie par :

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

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

9

Le Principe du Problème Auxiliaire en optimisation stochastique sous contraintes explicites

On considère dans ce chapitre les problèmes sous contraintes explicites de la forme (4.1), qui ont été étudiés dans le cas déterministe au Chapitre 4, et on cherche à étendre l'étude au cas stochastique.

Comme au Chapitre 8, on se donne un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et une variable aléatoire $\boldsymbol{W}$ à valeurs dans l'espace $\mathcal{W}$ muni de sa tribu $\mathcal{W}$. On se donne un espace de Hilbert $\mathcal{U}$ ainsi qu'une partie non vide $U^{\rm ad}$ de $\mathcal{U}$, et une fonction j définie sur $\mathcal{U} \times \mathcal{W}$ à valeurs dans $\overline{\mathbb{R}}$. On note J l'espérance de j (supposée intégrable pour tout $u \in U^{\rm ad}$) :

$$J(u) = \mathbb{E}\left(j(u, \boldsymbol{W})\right) \ .$$

Pour exprimer les contraintes, on se donne un nouvel espace de Hilbert $\mathcal{C}$, un cône C inclus dans cet espace et une application Θ définie sur $\mathcal{U}$ à valeurs dans $\mathcal{C}$. On s'intéresse alors au problème suivant :

$$\min_{u \in U^{\rm ad}} J(u) \qquad \text{sous la contrainte} \qquad \Theta(u) \in -C \ . \tag{9.1}$$

On traitera dans ce chapitre le cas où la fonction objectif J s'exprime comme l'espérance d'une fonction aléatoire, alors que la contrainte Θ est prise sous forme déterministe uniquement. On présentera au Chapitre 10 une extension du gradient stochastique au cas où la fonction Θ est l'espérance d'une fonction aléatoire θ définie sur $\mathcal{U} \times \mathcal{W}$ à valeurs dans $\mathcal{C}$.

9.1 Rappels du cas déterministe

Le Lagrangien L du problème (9.1) est donné par [1] :

$$L(u, p) = J(u) + \langle p \, , \Theta(u) \rangle \ .$$

Sous les conditions classiques de convexité, de continuité et de qualification des contraintes, la résolution de (9.1) se fait en maxi-minimisant le Lagrangien.

1. On se reportera au §4.4 pour toutes les notions relatives à la théorie de la dualité.

9.1.1 Algorithmes d'Uzawa et d'Arrow-Hurwicz

La manière la plus classique d'effectuer la maxi-minimisation est de mettre en œuvre l'algorithme d'*Uzawa*, qui, à chaque itération, consiste à alterner une étape de *minimisation complète* du Lagrangien en u et une étape de type *pas de gradient* pour la maximisation en p. L'itération k de l'algorithme d'Uzawa consiste donc, connaissant le couple (u^k, p^k), à calculer :

$$u^{k+1} \in \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} J(u) + \left\langle p^k \,, \Theta(u) \right\rangle , \tag{9.2a}$$

$$p^{k+1} = \mathrm{proj}_{C^*}\left(p^k + \rho\, \Theta(u^{k+1})\right) . \tag{9.2b}$$

Une autre façon de procéder, connue sous le nom d'algorithme d'*Arrow-Hurwicz*, consiste à chaque itération à alterner un pas de gradient pour la minimisation en u *et* un pas de gradient pour la maximisation en p. Une itération de l'algorithme d'Arrow-Hurwicz revient, connaissant le couple (u^k, p^k), à calculer :

$$u^{k+1} = \mathrm{proj}_{U^{\mathrm{ad}}}\Big(u^k - \varepsilon\big(\nabla J(u^k) + \big(\Theta'(u^k)\big)^\top \cdot p^k\big)\Big) , \tag{9.3a}$$

$$p^{k+1} = \mathrm{proj}_{C^*}\left(p^k + \rho\, \Theta(u^{k+1})\right) . \tag{9.3b}$$

9.1.2 Principe du Problème Auxiliaire

La généralisation de la décomposition par les prix par le Principe du Problème Auxiliaire a été présentée au Chapitre 4. À travers une démarche initiée au §4.1, elle peut finalement se résumer à choisir une fonction K définie sur l'espace $\mathcal{U}$ à valeurs dans $\mathbb{R}$, et à résoudre le problème (9.1) par le biais de la résolution d'une suite de problèmes auxiliaires indexés par k. Le k-ième problème auxiliaire comporte deux étapes, l'une de minimisation en u et l'autre de maximisation en p. La résolution de ce problème auxiliaire constitue la k-ième itération de l'algorithme du PPA qui s'écrit :

$$u^{k+1} \in \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon \nabla J(u^k) - \nabla K(u^k) \,, u \right\rangle + \varepsilon \left\langle p^k \,, \Theta'(u^k) \cdot u \right\rangle , \tag{9.4a}$$

$$p^{k+1} = \mathrm{proj}_{C^*}\left(p^k + \rho\, \Theta(u^{k+1})\right) . \tag{9.4b}$$

Dans la phase de minimisation en u de cet algorithme, une variante consiste à remplacer l'approximation du premier ordre $\left\langle p^k \,, \Theta'(u^k) \cdot u \right\rangle$ par le terme $\left\langle p^k \,, \Theta(u) \right\rangle$, ce qui conduit à une étape de minimisation en u de la forme :

$$u^{k+1} \in \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon \nabla J(u^k) - \nabla K(u^k) \,, u \right\rangle + \varepsilon \left\langle p^k \,, \Theta(u) \right\rangle . \tag{9.4c}$$

On a vu au Chapitre 4 que l'application du PPA dans ce cadre permettait, d'une part de retrouver les algorithmes d'Uzawa (étape (9.4c) avec les choix $K(u) = J(u)$ et $\varepsilon = 1$), et d'Arrow-Hurwicz (étape (9.4a) avec le choix $K(u) = \|u\|^2/2$), et d'autre part de décomposer l'étape de minimisation en u de chaque problème auxiliaire en choisissant une fonction auxiliaire K additif par rapport à la décomposition de l'espace $\mathcal{U}$, sous réserve, pour le choix (9.4c), que Θ ait elle aussi une forme additive.

9.2 Extension stochastique de l'algorithme d'Uzawa ?

Une première tentative pour appliquer l'idée du gradient stochastique au problème (9.1) est, dans le cadre de l'algorithme d'Uzawa, de remplacer dans (9.2) la valeur $J(u)$ par la valeur $j(u, w^{k+1})$, où w^{k+1} est un tirage de la variable aléatoire $\boldsymbol{W}$. Cette manière de faire conduirait à l'Algorithme 9.1 ci-dessous. On montre alors à l'aide d'un contre-exemple que cet algorithme ne peut pas fonctionner.

9.2.1 Tentative d'algorithme d'Uzawa stochastique

On propose l'algorithme suivant, que l'on pourrait qualifier de « algorithme d'Uzawa stochastique », alternant des étapes de minimisation en la variable primale u et des remises à jour de type gradient en la variable duale p.

Algorithme 9.1.

1. Choisir un couple de points initiaux $(u^0, p^0) \in U^{\mathrm{ad}} \times C^*$, ainsi qu'une suite $\{\rho^k\}_{k\in\mathbb{N}}$ de réels positifs.
2. À l'itération k, effectuer un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$.
3. Calculer $u^{k+1} \in \underset{u\in U^{\mathrm{ad}}}{\arg\min}\, j(u, w^{k+1}) + \langle p^k\,, \Theta(u)\rangle$.
4. Calculer $p^{k+1} = \mathrm{proj}_{C^*}\big(p^k + \rho^k\, \Theta(u^{k+1})\big)$.
5. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Plutôt que de mener l'étude théorique de convergence de cet algorithme, on va montrer sur un exemple simple que l'algorithme proposé ne permet pas d'obtenir la solution du problème de départ. On s'appuiera pour cela sur le résultat de (Kushner et Clark, 1978, Theorem 2.3.1) qui, dans le cadre de l'approximation stochastique étudié au §7.4, établit un lien entre la convergence de l'algorithme :

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k + \varepsilon^k \Big(h(\boldsymbol{U}^k) + \boldsymbol{\xi}^{k+1} \Big)\,, \tag{9.5}$$

et le comportement asymptotique de l'équation différentielle ordinaire :

$$\dot{u} = h(u)\,. \tag{9.6}$$

Dans l'équation (9.6), la notation $\dot{u}$ représente la dérivée de u par rapport à un « temps continu » t qui vient remplacer les itérations discrètes indexées par k. Si l'on écrit l'équation (9.5) sous la forme :

$$\frac{\boldsymbol{U}^{k+1} - \boldsymbol{U}^k}{\varepsilon^k} = h(\boldsymbol{U}^k) + \boldsymbol{\xi}^{k+1} ,$$

on peut interpréter le membre de gauche de cette dernière égalité comme la discrétisation par différence finie d'une dérivée dans laquelle le coefficient ε^k joue le rôle d'un incrément de temps. On notera que le temps t est « moralement » la somme des ε^k, qui tend bien vers $+\infty$ lorsque $\{\varepsilon^k\}_{k\in\mathbb{N}}$ est une σ-suite.

Plus précisément, on utilisera le théorème suivant (Benveniste et collab., 1990, Chapter 2, Theorem 7).

Théorème 9.2. *Soit $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ la suite engendrée par l'Algorithme (9.5). On suppose qu'il existe un point $u^\sharp \in \mathcal{U}$ tel que :*

$$\mathbb{P}\big(\lim_{k\to+\infty} \boldsymbol{U}^k = u^\sharp\big) > 0 .$$

Alors, le point $u^\sharp$ est un point d'équilibre stable de l'équation différentielle ordinaire (9.6).

En d'autres termes, ce résultat indique qu'un algorithme de type (9.5) ne peut converger (s'il converge…) que vers un point d'équilibre stable de l'équation différentielle ordinaire associée.

9.2.2 Contre-exemple

On considère le problème défini de la manière suivante.
- $\mathcal{U} = \mathbb{R}^2$ et $U^{\rm ad} = \mathcal{U}$.
- $\mathcal{C} = \mathbb{R}$ et $C = \{0\}$ (contrainte égalité).
- $\mathcal{W} = \mathbb{R}^4$ et $w = (a_1, a_2, b_1, b_2)$.
- $j(u, w) = \frac{1}{2}\big(a_1 {u_1}^2 + a_2 {u_2}^2\big) + \big(b_1 u_1 + b_2 u_2\big)$.
- $\Theta(u) = \theta_1 u_1 + \theta_2 u_2$.

On suppose que les variables aléatoires $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{B}_1$ et $\boldsymbol{B}_2$ sont intégrables, que $\boldsymbol{A}_1$ et $\boldsymbol{A}_2$ sont strictement positives, et que θ_1 et θ_1 sont des constantes réelles non nulles. Le problème à résoudre est :

$$\min_{u\in\mathbb{R}^2} \mathbb{E}\,\big(j(u, \boldsymbol{W})\big) \qquad \text{sous} \qquad \Theta(u) = 0 .$$

Résolution du problème déterministe

Il n'y a aucune difficulté sur ce problème à calculer les espérances et résoudre le problème déterministe associé. Notant $(u_1^\sharp, u_2^\sharp)$ la solution primale et $p^\sharp$ le multiplicateur optimal associé à la contrainte, les conditions d'optimalité du problème s'écrivent :

$$\begin{aligned}
&\mathbb{E}\left(\boldsymbol{A}_1\right) u_1^\sharp + \mathbb{E}\left(\boldsymbol{B}_1\right) + \theta_1 p^\sharp = 0\,,\\
&\mathbb{E}\left(\boldsymbol{A}_2\right) u_2^\sharp + \mathbb{E}\left(\boldsymbol{B}_2\right) + \theta_2 p^\sharp = 0\,,\\
&\theta_1 u_1^\sharp + \theta_2 u_2^\sharp = 0\,.
\end{aligned}$$

On en déduit que la valeur du multiplicateur optimale est :

$$p^\sharp = -\frac{\theta_1\Big(\mathbb{E}\left(\boldsymbol{B}_1\right)/\mathbb{E}\left(\boldsymbol{A}_1\right)\Big) + \theta_2\Big(\mathbb{E}\left(\boldsymbol{B}_2\right)/\mathbb{E}\left(\boldsymbol{A}_2\right)\Big)}{\theta_1\Big(\theta_1/\mathbb{E}\left(\boldsymbol{A}_1\right)\Big) + \theta_2\Big(\theta_2/\mathbb{E}\left(\boldsymbol{A}_2\right)\Big)}\,. \tag{9.7}$$

Mise en œuvre de l'algorithme d'Uzawa stochastique

La k-ième itération de l'Algorithme 9.1 revient à résoudre le système :

$$\begin{aligned}
&a_1^{k+1} u_1^{k+1} + b_1^{k+1} + \theta_1 p^k = 0\,,\\
&a_2^{k+1} u_2^{k+1} + b_2^{k+1} + \theta_2 p^k = 0\,,
\end{aligned}$$

et à mettre à jour la valeur p^{k+1} du multiplicateur par la relation :

$$p^{k+1} = p^k + \rho^k\Big(\theta_1 u_1^{k+1} + \theta_2 u_2^{k+1}\Big)\,.$$

Les deux premières relations permettent de calculer u_1^{k+1} et u_2^{k+1} en fonction de p^k. Reportant ces valeurs dans la troisième relation, on en déduit l'équation de récurrence stochastique décrivant l'évolution du multiplicateur au cours des itérations :

$$\boldsymbol{P}^{k+1} = \boldsymbol{P}^k - \rho^k\left(\Big(\frac{\theta_1^2}{\boldsymbol{A}_1^{k+1}} + \frac{\theta_2^2}{\boldsymbol{A}_2^{k+1}}\Big)\boldsymbol{P}^k - \Big(\theta_1\frac{\boldsymbol{B}_1^{k+1}}{\boldsymbol{A}_1^{k+1}} + \theta_2\frac{\boldsymbol{B}_2^{k+1}}{\boldsymbol{A}_2^{k+1}}\Big)\right).$$

Utilisant le Théorème 9.2, on sait que cette équation de récurrence stochastique ne peut converger (si elle converge…) que vers l'unique point $p^\star$ d'équilibre stable de l'équation différentielle ordinaire associée, à savoir :

$$p^\star = -\frac{\theta_1\Big(\mathbb{E}\left(\boldsymbol{B}_1/\boldsymbol{A}_1\right)\Big) + \theta_2\Big(\mathbb{E}\left(\boldsymbol{B}_2/\boldsymbol{A}_2\right)\Big)}{\theta_1\Big(\theta_1/\mathbb{E}\left(\boldsymbol{A}_1\right)\Big) + \theta_2\Big(\theta_2/\mathbb{E}\left(\boldsymbol{A}_2\right)\Big)}\,. \tag{9.8}$$

Si l'on compare alors les relations (9.7) et (9.8), on constate que, dès que les variables aléatoires $\boldsymbol{A}_1$ et $\boldsymbol{B}_1$ ne sont pas *indépendantes*, on a d'une manière générale $\mathbb{E}\left(\boldsymbol{B}_1/\boldsymbol{A}_1\right) \neq \mathbb{E}\left(\boldsymbol{B}_1\right)/\mathbb{E}\left(\boldsymbol{A}_1\right)$, et donc $p^\sharp \neq p^\star$.

9.2.3 Conclusion

Cet exemple montre que, même si l'Algorithme 9.1 converge (ce que l'on n'a pas cherché à prouver), il ne fournit de toute façon pas toujours la solution

du problème initial. L'exemple choisi (linéaire-quadratique) est suffisamment simple et générique pour pouvoir exclure l'algorithme de la panoplie permettant de résoudre des problèmes de type (9.1), et donc pour tirer la conclusion que l'algorithme d'Uzawa ne peut pas être étendu au cas stochastique.

Si l'on compare l'algorithme de gradient stochastique standard à cette tentative d'extension de l'algorithme d'Uzawa au cadre stochastique, on constate que, alors que l'itérée u^{k+1} de l'algorithme de gradient stochastique diffère de u^k par une formule faisant intervenir un pas ε^k tendant vers zéro, l'itérée u^{k+1} de cette extension de l'algorithme d'Uzawa résulte d'une minimisation dépendant essentiellement du tirage w^{k+1} ; ces itérées u^k « sautent » donc au gré de ces réalisations indépendantes w^k, ce qui ne permet pas de reconstituer l'effet de moyenne typique des itérations du gradient stochastique qui, comme on l'a expliqué au Chapitre 6, repose de façon essentielle sur une variation « lente » des itérées u^k.

9.3 Extension stochastique de l'algorithme issu du PPA

L'extension stochastique de l'algorithme issu du principe du problème auxiliaire consiste à remplacer dans les relations (9.4) le gradient $\nabla J(u^k)$ (correspondant à une espérance portant sur la variable aléatoire $\boldsymbol{W}$) par le terme $\nabla_u j(u^k, w^{k+1})$ (correspondant à un tirage de la variable $\boldsymbol{W}$) ; on aboutit donc, pour la résolution du problème (9.1), à la résolution de la suite de problèmes auxiliaires :

$$u^{k+1} \in \arg\min_{u\in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon^k \nabla_u j(u^k, w^{k+1}) - \nabla K(u^k)\,, u \right\rangle + \varepsilon^k \left\langle p^k\,, \Theta'(u^k)\cdot u \right\rangle\,, \tag{9.9a}$$

$$p^{k+1} = \mathrm{proj}_{C^*}\left(p^k + \varepsilon^k\, \Theta(u^{k+1})\right)\,. \tag{9.9b}$$

Comme on l'a vu au §9.1.2 dans le cas déterministe, il est possible de remplacer dans la phase de minimisation en u l'approximation du premier ordre $\left\langle p^k\,, \Theta'(u^k)\cdot u\right\rangle$ par le terme $\left\langle p^k\,, \Theta(u)\right\rangle$, ce qui conduit à une étape de minimisation en u de la forme :

$$u^{k+1} \in \arg\min_{u\in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon^k \nabla_u j(u^k, w^{k+1}) - \nabla K(u^k)\,, u \right\rangle + \varepsilon^k \left\langle p^k\,, \Theta(u)\right\rangle\,. \tag{9.9c}$$

Le choix entre les formulations (9.9a) et (9.9c) se fera dans les applications en fonction des possibilités de décomposition offertes par la fonction Θ ; on notera que, dans la formulation (9.9a), le terme dépendant de Θ est *linéaire* en u et est donc additif par rapport à n'importe quelle décomposition de l'espace $\mathcal{U}$.

On propose alors l'algorithme suivant comme extension de l'algorithme issu du PPA au cadre stochastique sous contrainte déterministe.

Algorithme 9.3.

1. Choisir $(u^0, p^0) \in U^{\mathrm{ad}} \times C^*$, et une suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ de réels positifs.
2. À l'itération k, effectuer un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$.
3. Calculer u^{k+1} solution du problème auxiliaire (9.9c).
4. Calculer p^{k+1} par la formule de mise à jour (9.9b).
5. Incrémenter l'indice k de 1 et retourner à l'étape 2.

On fait sur cet algorithme les commentaires suivants.

– Avec le choix de fonction auxiliaire $K(u) = \|u\|^2/2$, et utilisant la formulation (9.9a) plutôt que (9.9c), la minimisation en u dans le problème auxiliaire (9.9) se met sous la forme :

$$\min_{u\in U^{\mathrm{ad}}} \frac{1}{2}\|u\|^2 + \left\langle \varepsilon^k \nabla_u j(u^k, w^{k+1}) - u^k + \varepsilon^k \big(\Theta'(u^k)\big)^\top \cdot p^k \, , u \right\rangle ,$$

dont la solution u^{k+1} se calcule explicitement :

$$u^{k+1} = \mathrm{proj}_{U^{\mathrm{ad}}}\big(u^k - \varepsilon^k \nabla_u j(u^k, w^{k+1}) - \varepsilon^k \big(\Theta'(u^k)\big)^\top \cdot p^k\big) ,$$

et correspond exactement à un pas de l'extension « naturelle » au cas stochastique de l'algorithme d'Arrow-Hurwicz déterministe (comparer avec les relations (9.3)).

– Comme on l'a déjà discuté à la section précédente, cet algorithme général ne peut pas se spécialiser en celui d'Uzawa car, comme on le verra par la suite, les pas ε^k doivent tendre vers 0 de telle sorte que prendre un pas constant (égal à 1) comme dans l'algorithme d'Uzawa n'est pas envisageable.

On considère maintenant l'application à la décomposition de cet algorithme. Utilisant la forme linéarisée (9.9a) en Θ dans (9.9), supposant que l'espace $\mathcal{U}$ et l'ensemble admissible U^{ad} se mettent sous forme de produits cartésiens $\mathcal{U}_1 \times \ldots \times \mathcal{U}_N$ et $U_1^{\mathrm{ad}} \times \ldots \times U_N^{\mathrm{ad}}$, avec pour tout i l'inclusion $U_i^{\mathrm{ad}} \subset \mathcal{U}_i$, et en choisissant la fonction auxiliaire K sous forme additive par rapport à cette décomposition :

$$K(u) = \sum_{i=1}^{N} K_i(u_i) \, , \quad \text{avec} \quad u_i \in \mathcal{U}_i \, ,$$

l'étape de minimisation en u s'écrit :

$$\min_{(u_1,\ldots,u_N)\in U^{\mathrm{ad}}} \sum_{i=1}^{N} \Big(K_i(u_i) + \left\langle \varepsilon^k \nabla_{u_i} j(u^k, w^{k+1}) - \nabla K_i(u_i^k) \, , u_i \right\rangle + \varepsilon^k \left\langle \big(\Theta'_{u_i}(u^k)\big)^\top \cdot p^k \, , u_i \right\rangle \Big) ,$$

qui se décompose en N sous-problèmes indépendants dont la i-ième instance s'écrit :

$$\min_{u_i \in U_i^{\mathrm{ad}}} K_i(u_i) + \left\langle \varepsilon^k \nabla_{u_i} j(u^k, w^{k+1}) - \nabla K_i(u_i^k)\,, u_i \right\rangle + \varepsilon^k \left\langle \left(\Theta'_{u_i}(u^k)\right)^\top \cdot p^k\,, u_i \right\rangle .$$

Lorsque la fonction Θ est additive par rapport à la décomposition de l'espace :

$$\Theta(u) = \sum_{i=1}^{N} \Theta_i(u_i)\ , \quad \text{avec} \quad u_i \in \mathcal{U}_i\ ,$$

le produit scalaire $\left\langle \left(\Theta'_{u_i}(u^k)\right)^\top \cdot p^k\,, u_i \right\rangle$ apparaissant dans le i-ième sous-problème ci-dessus peut être remplacé par le terme $\left\langle p^k\,, \Theta_i(u_i) \right\rangle$. On obtient ainsi une version décomposée de l'algorithme (9.9) dans lequel l'étape de minimisation est (9.9c) plutôt que (9.9a).

Remarque 9.4. Ces considérations sont de même nature que celles faites au Chapitre 4 à propos de la fonction Θ s'écrivant sous la forme $\boldsymbol{\Theta} + \mathfrak{T}$, la partie $\mathfrak{T}$ étant supposée être de structure additive par rapport à une décomposition de l'espace $\mathcal{U}$. On se reportera à la Remarque 8.3 pour des considérations analogues à propos de la fonction objectif J.

On va donner au §9.4 un résultat de convergence pour l'Algorithme 9.3 dans le cas où le Lagrangien du problème déterministe associé est *stable en* u([2]). Puis, on donnera au §9.5 un résultat adapté au cas où la fonction J est *fortement convexe*, ce qui permettra d'utiliser des « grands pas » dans l'étape de remise à jour du multiplicateur. Enfin, on s'intéressera au §9.6 au cas où la fonction J est *simplement convexe*, cas pour lequel on donnera un algorithme basé sur l'utilisation du Lagrangien augmenté.

9.4 Théorème de convergence sous condition de stabilité du Lagrangien en u

Comme cela a été dit dans la Remarque 7.2, la convergence de l'Algorithme 9.3 s'étudie en termes de variables aléatoires, l'étape de minimisation en u s'écrivant[3] :

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) - \nabla K(\boldsymbol{U}^k)\,, u \right\rangle + \varepsilon^k \left\langle \boldsymbol{P}^k\,, \Theta(u) \right\rangle , \qquad (9.10a)$$

2. Voir le commentaire 2.5c page 20 pour la notion de stabilité du Lagrangien.
3. On se limite à la formulation (9.9c) pour l'étape de minimisation en u. En fait, l'analyse de convergence serait quasiment identique pour la forme linéarisée (9.9a).

et la remise à jour des variables duales étant :

$$\boldsymbol{P}^{k+1} = \mathrm{proj}_{C^*}\left(\boldsymbol{P}^k + \varepsilon^k\, \Theta(\boldsymbol{U}^{k+1})\right)\ . \tag{9.10b}$$

On se place dans le cas où le Lagrangien du problème (9.1) est stable en u. La question de la convergence de l'Algorithme 9.3 est réglée par le résultat suivant.

Théorème 9.5. *On fait les hypothèses suivantes.*

1. *U^{ad} est une partie convexe fermée non vide d'un l'espace de Hilbert $\mathcal{U}$, et C est un cône convexe fermé saillant*[4] *d'un autre espace de Hilbert $\mathcal{C}$.*
2. *La fonction $j : \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ est une intégrande normale, et l'espérance de $j(u, \boldsymbol{W})$ existe pour tout $u \in U^{\mathrm{ad}}$.*
3. *Pour tout $w \in \mathcal{W}$, la fonction $j(\cdot, w) : \mathcal{U} \to \mathbb{R}$ est propre, convexe, semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant U^{ad} et son gradient partiel par rapport à u est noté $\nabla_u j(u, w)$.*
4. *La fonction $j(\cdot, w)$ vérifie l'hypothèse (8.6) de gradient linéairement borné en u uniformément en w.*
5. *La fonction J est Lipschitzienne, coercive sur l'ensemble U^{ad}.*
6. *La fonction Θ est C-convexe, Lipschitzienne de constante L_Θ.*
7. *Les contraintes sont qualifiées et le Lagrangien L est stable.*
8. *La fonction K est propre, fortement convexe de constante b, semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant U^{ad}.*
9. *La suite $\left\{\varepsilon^k\right\}_{k\in\mathbb{N}}$ est une σ-suite.*

On a alors les conclusions suivantes.

1. *Le problème (9.1) admet un ensemble de points selle $U^\sharp \times P^\sharp$ non vide.*
2. *Le problème (9.10a) admet une solution $\boldsymbol{U}^{k+1}$ unique.*
3. *Pour tout $p^\sharp \in P^\sharp$, la suite de variables aléatoires $\left\{L(\boldsymbol{U}^k, p^\sharp)\right\}_{k\in\mathbb{N}}$ converge presque sûrement vers $L(u^\sharp, p^\sharp)$, avec $u^\sharp \in U^\sharp$.*
4. *Les suites $\left\{\boldsymbol{U}^k\right\}_{k\in\mathbb{N}}$ et $\left\{\boldsymbol{P}^k\right\}_{k\in\mathbb{N}}$ engendrées par l'Algorithme 9.3 sont bornées presque sûrement, et tout point d'accumulation d'une réalisation de la suite $\left\{\boldsymbol{U}^k\right\}_{k\in\mathbb{N}}$ appartient à l'ensemble $U^\sharp$ des solutions du problème.*

Remarque 9.6. L'hypothèse de stabilité en u du Lagrangien n'est en pratique pas aisée à vérifier. On peut la remplacer par la condition (plus forte) de stricte convexité de la fonction J. Il faut noter que sans l'hypothèse de stabilité, il y a peu de chances de pouvoir dire quoi que ce soit des limites de la suite de variables aléatoires $\left\{\boldsymbol{U}^k\right\}_{k\in\mathbb{N}}$.

4. Voir (4.35) pour cette notion.

Preuve. La démonstration des deux premières conclusions de ce théorème découle des théorèmes généraux relatifs à l'optimisation convexe en présence de contraintes. Le fait que la solution $\boldsymbol{U}^{k+1}$ du problème (9.10a) soit une variable aléatoire, et donc une fonction mesurable, provient de ce que l'on a supposé que j était une intégrande normale (voir le Théorème 8.5 et notamment le raisonnement fait au début de la preuve de ce théorème). La démonstration des deux dernières conclusions se fait en suivant le déroulement « classique » en quatre étapes de la preuve de convergence d'un algorithme d'optimisation stochastique, à savoir :

1. choix d'une fonction de Lyapunov opérant sur (u^k, p^k),
2. majoration de sa variation d'une itération de l'algorithme sur l'autre,
3. convergence de l'algorithme, à l'aide d'un argument de type martingale,
4. analyse des limites des suites permettant de caractériser la solution.

Pour alléger les écritures, on utilisera la notation :

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

Le fait que u^{k+1} soit solution du problème (9.9c) est caractérisé par la condition d'optimalité suivante :

$$\forall u \in U^{\mathrm{ad}}, \; \big\langle \nabla K(u^{k+1}) - \nabla K(u^k) + \varepsilon^k g^k \, , u - u^{k+1} \big\rangle \\ + \varepsilon^k \big\langle p^k \, , \Theta(u) - \Theta(u^{k+1}) \big\rangle \geq 0 \; . \quad (9.11)$$

1. **Choix de la fonction de Lyapunov**. Soit $(u^\sharp, p^\sharp) \in U^\sharp \times P^\sharp$ un point selle du Lagrangien du problème (9.1). On choisit la fonction de Lyapunov ℓ de la forme :

$$\ell(u,p) = K(u^\sharp) - K(u) - \big\langle \nabla K(u) \, , u^\sharp - u \big\rangle + \frac{1}{2} \big\| p - p^\sharp \big\|^2 \; ,$$

et l'on note :

$$\psi^k = \ell(u^k, p^k) \; .$$

De la forte convexité de K et de la définition de ℓ, on déduit que ℓ est bornée inférieurement et coercive :

$$\big\| u^k - u^\sharp \big\|^2 \leq \frac{2}{b} \, \psi^k \quad \text{et} \quad \big\| p^k - p^\sharp \big\|^2 \leq 2 \, \psi^k \; . \quad (9.12)$$

2. **Majorations**.
   a. On majore pour commencer la quantité $\big\| u^{k+1} - u^k \big\|$. Évaluant la condition d'optimalité (9.11) au point $u = u^k$:

$$\big\langle \nabla K(u^k) - \nabla K(u^{k+1}) \, , u^k - u^{k+1} \big\rangle \leq \big\langle \varepsilon^k g^k \, , u^k - u^{k+1} \big\rangle \\ + \varepsilon^k \big\langle p^k \, , \Theta(u^k) - \Theta(u^{k+1}) \big\rangle \; .$$

Utilisant la forte convexité de K, l'inégalité de Schwarz et le caractère Lipschitzien de Θ, on obtient :

$$b\|u^{k+1}-u^k\|^2 \le \varepsilon^k\|g^k\|\|u^k-u^{k+1}\| + \varepsilon^k\|p^k\|L_\Theta\|u^k-u^{k+1}\| ,$$

et donc :

$$b\|u^{k+1}-u^k\| \le \varepsilon^k\Big(\|g^k\|+L_\Theta\|p^k\|\Big) .$$

L'hypothèse (8.6) et la majoration (9.12) impliquent l'existence de constantes c_3 et c_4 telles que :

$$\|u^{k+1}-u^k\| \le \varepsilon^k\Big(c_3\sqrt{\psi^k}+c_4\Big) . \tag{9.13}$$

b. On majore ensuite la quantité $\|p^{k+1}-p^\sharp\|$. À l'aide de la relation [5] :

$$p^\sharp = \mathrm{proj}_{C^*}\left(p^\sharp+\varepsilon^k\Theta(u^\sharp)\right) , \tag{9.14}$$

de la relation (9.9b) et comme l'opérateur de projection sur C^* est non expansif, on obtient :

$$\|p^{k+1}-p^\sharp\| \le \|p^k-p^\sharp+\varepsilon^k\big(\Theta(u^{k+1})-\Theta(u^\sharp)\big)\| .$$

Élevant cette inégalité au carré, développant le terme de droite et utilisant le fait que Θ est Lipschitzienne, il vient :

$$\|p^{k+1}-p^\sharp\|^2 \le \|p^k-p^\sharp\|^2+\big(\varepsilon^k L_\Theta\big)^2\|u^{k+1}-u^\sharp\|^2 + 2\varepsilon^k\langle p^k-p^\sharp\,,\Theta(u^{k+1})-\Theta(u^\sharp)\rangle . \tag{9.15}$$

c. On majore pour finir l'écart $\psi^{k+1}-\psi^k$. Formant cette différence, utilisant le fait que la fonction K est convexe et que donc $K(u^k)-K(u^{k+1})+\langle\nabla K(u^k)\,,u^{k+1}-u^k\rangle\le 0$, ainsi que la relation (9.15), on obtient :

$$\psi^{k+1}-\psi^k \le \langle\nabla K(u^k)-\nabla K(u^{k+1})\,,u^\sharp-u^{k+1}\rangle + \varepsilon^k\langle p^k-p^\sharp\,,\Theta(u^{k+1})-\Theta(u^\sharp)\rangle + \frac{\big(\varepsilon^k L_\Theta\big)^2}{2}\|u^{k+1}-u^\sharp\|^2 .$$

Par la condition d'optimalité (9.11) évaluée au point $u=u^\sharp$, on obtient :

$$\psi^{k+1}-\psi^k \le +\varepsilon^k\langle g^k\,,u^\sharp-u^{k+1}\rangle + \varepsilon^k\langle p^\sharp\,,\Theta(u^\sharp)-\Theta(u^{k+1})\rangle + \frac{\big(\varepsilon^k L_\Theta\big)^2}{2}\|u^{k+1}-u^\sharp\|^2 .$$

5. Cette relation, vraie pour toute valeur $\varepsilon^k>0$, est équivalente à l'inégalité de gauche du point selle du Lagrangien (voir Exercice 4.7).

Remplaçant $u^\sharp - u^{k+1}$ par $u^\sharp - u^k + u^k - u^{k+1}$ et $\Theta(u^\sharp) - \Theta(u^{k+1})$ par $\Theta(u^\sharp) - \Theta(u^k) + \Theta(u^k) - \Theta(u^{k+1})$ dans les deux premiers termes du membre de droite de cette inégalité, appliquant deux fois l'inégalité de Schwarz et utilisant l'hypothèse (8.6), il vient :

$$\begin{aligned}\psi^{k+1} - \psi^k \leq\ & \varepsilon^k \langle g^k\,, u^\sharp - u^k\rangle + \varepsilon^k \langle p^\sharp\,, \Theta(u^\sharp) - \Theta(u^k)\rangle \\ & + \varepsilon^k \|u^{k+1} - u^k\| \big(\|g^k\| + L_\Theta \|p^\sharp\|\big) \\ & + \frac{\big(\varepsilon^k L_\Theta\big)^2}{2} \|u^{k+1} - u^\sharp\|^2\,.\end{aligned}$$

Utilisant les majorations (9.12), (9.13) ainsi que l'hypothèse (8.6) dans l'inégalité ci-dessus, et passant à son interprétation en termes de variables aléatoires (voir la Remarque 7.2), on montre par un calcul simple l'existence de constantes c_5, c_6 et c_7 telles que :

$$\begin{aligned}\boldsymbol{\Psi}^{k+1} - \boldsymbol{\Psi}^k \leq \varepsilon^k \langle \boldsymbol{G}^k\,, u^\sharp - \boldsymbol{U}^k\rangle + \varepsilon^k \langle p^\sharp\,, \Theta(u^\sharp) - \Theta(\boldsymbol{U}^k)\rangle \\ + (\varepsilon^k)^2 \big(c_5 \boldsymbol{\Psi}^k + c_6 + c_7 \boldsymbol{\Psi}^{k+1}\big)\,, \quad (9.16)\end{aligned}$$

On rappelle que $\boldsymbol{W}^{k+1}$ est indépendante des $\boldsymbol{W}^1$ pour $l \leq k$, et que $\boldsymbol{U}^k$ et $\boldsymbol{\Psi}^k$ sont par construction des variables aléatoires mesurables par rapport à la tribu $\mathcal{F}^k$. On en déduit que :

$$\mathbb{E}\big(\langle \boldsymbol{G}^k\,, u^\sharp - \boldsymbol{U}^k\rangle \mid \mathcal{F}^k\big) = \langle \nabla J(\boldsymbol{U}^k)\,, u^\sharp - \boldsymbol{U}^k\rangle \leq J(u^\sharp) - J(\boldsymbol{U}^k)\,,$$

la dernière inégalité provenant de la convexité de J. Prenant de part et d'autre de l'inégalité (9.16) l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^k$ engendrée par les k variables aléatoires $(\boldsymbol{W}^1, \ldots, \boldsymbol{W}^k)$, on obtient :

$$\begin{aligned}\mathbb{E}\big(\boldsymbol{\Psi}^{k+1} \mid \mathcal{F}^k\big) - \boldsymbol{\Psi}^k \leq \alpha^k \boldsymbol{\Psi}^k + \beta^k + \gamma^k \mathbb{E}\big(\boldsymbol{\Psi}^{k+1} \mid \mathcal{F}^k\big) \\ + \varepsilon^k \big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)\big)\,, \quad (9.17)\end{aligned}$$

où α^k, β^k et γ^k sont les termes de trois séries convergentes. Le terme $L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)$ est toujours négatif ou nul (inégalité de droite caractérisant le point selle $(u^\sharp, p^\sharp)$ du Lagrangien L).

3. **Analyse de convergence.** Une application directe du Corollaire 8.9 du théorème de Robbins-Siegmund permet de montrer que la suite de variables aléatoires $\{\boldsymbol{\Psi}^k\}_{k\in\mathbb{N}}$ converge presque sûrement vers une variable aléatoire $\boldsymbol{\Psi}^\infty$ bornée presque sûrement, et que l'on a :

$$\sum_{k=0}^{+\infty} \varepsilon^k \big(L(\boldsymbol{U}^k, p^\sharp) - L(u^\sharp, p^\sharp)\big) < +\infty\,, \quad \mathbb{P}\text{-p.s.}\,. \qquad (9.18)$$

4. **Limites des suites.** Du fait que la suite $\{\boldsymbol{\Psi}^k\}_{k\in\mathbb{N}}$ est bornée presque sûrement, on déduit de (9.12) que les suites $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ et $\{\boldsymbol{P}^k\}_{k\in\mathbb{N}}$ sont bornées presque sûrement. Par la relation (9.13), il en est de même pour la suite $\{\|\boldsymbol{U}^{k+1}-\boldsymbol{U}^k\|/\varepsilon^k\}_{k\in\mathbb{N}}$. Cette dernière propriété, associée à la relation (9.18) et au fait que les fonctions J et Θ soient Lipschitziennes, permet d'utiliser le Lemme 8.10 : on en déduit alors que la suite $\{L(\boldsymbol{U}^k,p^\sharp)\}_{k\in\mathbb{N}}$ converge presque sûrement vers $L(u^\sharp,p^\sharp)$.
   On note alors Ω_0 le sous-ensemble (de mesure nulle) de Ω sur lequel la suite $\{\boldsymbol{\Psi}^k\}_{k\in\mathbb{N}}$ n'est pas bornée, et Ω_1 le sous-ensemble (de mesure nulle lui aussi) de Ω sur lequel la relation (9.18) n'est pas vérifiée.
   Soit $\omega \notin \Omega_0\cup\Omega_1$. La suite des réalisations $\{u^k\}_{k\in\mathbb{N}}$ associée à cet élément ω est bornée et chaque u^k appartient à $U^{\rm ad}$, partie fermée de $\mathcal{U}$. Par un argument de compacité, on conclut que l'on peut extraire de la suite $\{u^k\}_{k\in\mathbb{N}}$ une sous-suite convergente $\{u^{\Phi(k)}\}_{k\in\mathbb{N}}$. Soit $\bar{u}$ la limite de la suite $\{u^{\Phi(k)}\}_{k\in\mathbb{N}}$. La semi-continuité inférieure du Lagrangien L implique que l'on a :

$$L(\bar{u},p^\sharp)\le\liminf_{k\to+\infty}L(u^{\Phi(k)},p^\sharp)=L(u^\sharp,p^\sharp)\ .$$

   On en déduit que, presque sûrement, $\overline{u}$ est solution du problème de minimisation sur $U^{\rm ad}$ du Lagrangien L pour $p=p^\sharp$ fixé. La stabilité en u du Lagrangien implique alors que $\overline{u}\in U^\sharp$.

Pour conclure, on notera que dans le cas où la fonction J est strictement convexe, l'ensemble des solutions du problème (9.1) est un singleton :

$$U^\sharp=\{u^\sharp\}\ .$$

Le Lagrangien L est alors stable, et toute réalisation de la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ engendrée par l'Algorithme 9.3 a un *unique* point d'accumulation et converge donc toute entière vers $u^\sharp$. □

9.5 Cas fortement convexe : utilisation de grands pas

Lorsque la fonction J est *fortement convexe*, on dispose d'une variante intéressante de l'Algorithme 9.3 car il est alors possible de remettre à jour les multiplicateurs p^k avec un pas ρ *constant* (au lieu de pas ε^k décroissants vers zéro), pourvu que l'on contraigne les itérées p^k à rester dans une boule $\mathcal{B}(0,R)$ (de centre en 0 et de rayon R) et que cette boule soit assez grande pour contenir au moins un multiplicateur optimal $p^\sharp$. Ce procédé, (illustré par la Figure 4.1) a déjà été évoqué pour proposer une variante à l'Algorithme 4.1 (voir le commentaire 4.9c).

L'intérêt de cette variante est que l'utilisation de « grands pas » ρ permet une convergence plus rapide des variables duales p^k. L'évolution des variables

primales u^k se fait toujours quant à elle avec des « petits pas » ε^k afin de permettre l'effet de moyenne nécessaire pour obtenir la solution du problème initial. L'algorithme associé à cette variante est le suivant.

Algorithme 9.7.

1. Choisir $u^0 \in U^{\mathrm{ad}}$ et $p^0 \in C^* \cap \mathcal{B}(0,R)$, choisir un réel $\rho > 0$ et une suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ de réels positifs.
2. À l'itération k, effectuer un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$.
3. Calculer u^{k+1} solution du problème auxiliaire (9.9c).
4. Calculer $p^{k+1} = \mathrm{proj}_{C^*\cap\mathcal{B}(0,R)}\big(p^k + \rho\,\Theta(u^{k+1})\big)$.
5. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Les conditions de convergence de cet algorithme sont spécifiées par le théorème suivant.

Théorème 9.8. *En plus des hypothèses du Théorème 9.5, on suppose que :*
– *la fonction J est fortement convexe*[6] *de constante a,*
– *la σ-suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ est décroissante,*
– *le coefficient ρ est tel que $0 < \rho \leq a/L_\Theta^2$.*

Alors, toute réalisation de la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ engendrée par l'Algorithme 9.7 converge presque sûrement vers $u^\sharp$, unique solution du problème (9.1).

Preuve. La démonstration suivant un démarche très similaire à celle du Théorème 9.5, on reprend les éléments de cette preuve en indiquant les différences qui apparaissent dans le cadre de cette variante.

1. **Choix de la fonction de Lyapunov**. Soit $(u^\sharp, p^\sharp) \in U^\sharp \times P^\sharp$ un point selle du problème (9.1), où on suppose que $p^\sharp$ appartient à la boule $\mathcal{B}(0,R)$. On définit :

$$\psi^k = K(u^\sharp) - K(u^k) - \left\langle \nabla K(u^k)\,, u^\sharp - u^k \right\rangle + \frac{\varepsilon^k}{2\rho}\left\|p^k - p^\sharp\right\|^2 .$$

On notera qu'il n'existe alors pas de fonction de Lyapunov ℓ qui soit telle que $\psi^k = \ell(u^k, p^k)$. On est ici obligé de considérer une suite de fonctions ℓ^k variant avec k : on n'est donc plus tout-à-fait dans même schéma de preuve que dans le cas du Théorème 9.5.
Comme dans le Théorème 9.5, on a la majoration :

$$\left\|u^k - u^\sharp\right\|^2 \leq \frac{2}{b}\,\psi^k . \tag{9.19}$$

Par contre, la majoration $\left\|p^k - p^\sharp\right\|^2 \leq 2\psi^k$ n'est plus valide ; c'est pourquoi il faut s'assurer dans l'Algorithme 9.7 que $\left\|p^k\right\|$ reste majorée par R.

6. condition qui implique que J est coercive et que L est stable en u

2. **Majorations**.
   a. Afin de majorer la quantité $\|u^{k+1}-u^k\|$, on procède comme dans la preuve du Théorème 9.5, et on obtient :

$$b\|u^{k+1}-u^k\| \leq \varepsilon^k\big(\|g^k\| + L_\Theta\|p^k\|\big)\ .$$

   L'hypothèse (8.6), la majoration (9.19) et le fait que la norme du multiplicateur p^k soit majorée par R impliquent l'existence de constantes positives c_3 et c_4 telles que :

$$\|u^{k+1}-u^k\| \leq \varepsilon^k\big(c_3\sqrt{\psi^k}+c_4\big)\ . \tag{9.20}$$

   b. Pour la majoration de la quantité $\|p^{k+1}-p^\sharp\|$, on utilise les mêmes arguments que dans le Théorème 9.5. En effet, l'équation (9.14) peut s'écrire avec ρ à la place de ε^k et avec la projection sur $C^*\cap\mathcal{B}(0,R)$ grâce à l'hypothèse que $p^\sharp\in\mathcal{B}(0,R)$. En utilisant le caractère non expansif de cette projection, on obtient donc :

$$\begin{aligned}\|p^{k+1}-p^\sharp\|^2 \leq \|p^k-p^\sharp\|^2 + \big(\rho\, L_\Theta\big)^2\|u^{k+1}-u^\sharp\|^2\\ + 2\rho\,\langle p^k-p^\sharp\,,\Theta(u^{k+1})-\Theta(u^\sharp)\rangle\ .\end{aligned}$$

   Multipliant de part et d'autre de cette inégalité par ε^k/ρ, et utilisant le fait que la suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ est décroissante, on obtient :

$$\begin{aligned}\frac{\varepsilon^{k+1}}{\rho}\|p^{k+1}-p^\sharp\|^2 \leq \frac{\varepsilon^k}{\rho}\|p^k-p^\sharp\|^2 + \varepsilon^k\rho L_\Theta^2\|u^{k+1}-u^\sharp\|^2\\ + 2\varepsilon^k\langle p^k-p^\sharp\,,\Theta(u^{k+1})-\Theta(u^\sharp)\rangle\ .\end{aligned}$$

   c. Pour l'écart $\psi^{k+1}-\psi^k$, comme dans la preuve du Théorème 9.5, on a :

$$\begin{aligned}\psi^{k+1}-\psi^k \leq \varepsilon^k\langle g^k\,,u^\sharp-u^k\rangle + \varepsilon^k\langle p^\sharp\,,\Theta(u^\sharp)-\Theta(u^k)\rangle\\ + \varepsilon^k\|u^{k+1}-u^k\|\big(\|g^k\|+L_\Theta\|p^\sharp\|\big)\\ + \frac{\varepsilon^k\rho L_\Theta^2}{2}\|u^{k+1}-u^\sharp\|^2\ .\end{aligned}$$

   Utilisant l'hypothèse (8.6) et la majoration (9.19), et écrivant l'inégalité précédente en termes de variables aléatoires, on en déduit l'existence de constantes positives c_5 et c_6 telles que :

$$\begin{aligned}\boldsymbol{\Psi}^{k+1}-\boldsymbol{\Psi}^k \leq \varepsilon^k\langle \boldsymbol{G}^k\,,u^\sharp-\boldsymbol{U}^k\rangle + \varepsilon^k\langle p^\sharp\,,\Theta(u^\sharp)-\Theta(\boldsymbol{U}^k)\rangle\\ + (\varepsilon^k)^2\big(c_5\boldsymbol{\Psi}^k+c_6\big) + \frac{\varepsilon^k\rho\, L_\Theta^2}{2}\|\boldsymbol{U}^{k+1}-u^\sharp\|^2\ .\end{aligned}$$

   Prenant de part et d'autre de cette inégalité l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^k$ engendrée par les k variables aléatoires

$(\boldsymbol{W}^1,\dots,\boldsymbol{W}^k)$, et utilisant la forte convexité de la fonction J([7]), on obtient :

$$\begin{aligned}\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)-\boldsymbol{\Psi}^k &\le \varepsilon^k\big(L(u^\sharp,p^\sharp)-L(\boldsymbol{U}^k,p^\sharp)\big)\\ &\quad+\frac{\varepsilon^k}{2}\Big(\rho\,L_\Theta^2\big\|\boldsymbol{U}^{k+1}-u^\sharp\big\|^2-a\big\|\boldsymbol{U}^k-u^\sharp\big\|^2\Big)\\ &\quad+(\varepsilon^k)^2\Big(c_5\boldsymbol{\Psi}^k+c_6\Big)\,. \qquad (9.21)\end{aligned}$$

Grâce à la condition $\rho\,L_\Theta^2\le a$, on obtient :

$$\begin{aligned}\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)-\boldsymbol{\Psi}^k &\le \varepsilon^k\big(L(u^\sharp,p^\sharp)-L(\boldsymbol{U}^k,p^\sharp)\big)\\ &\quad+\frac{\varepsilon^k a}{2}\Big(\big\|\boldsymbol{U}^{k+1}-u^\sharp\big\|^2-\big\|\boldsymbol{U}^k-u^\sharp\big\|^2\Big)\\ &\quad+(\varepsilon^k)^2\Big(c_5\boldsymbol{\Psi}^k+c_6\Big)\,. \qquad (9.22)\end{aligned}$$

Avec les notations $X=\boldsymbol{U}^{k+1}-u^\sharp$ et $Y=\boldsymbol{U}^k-u^\sharp$, les majorations suivantes :

$$\begin{aligned}\varepsilon^k\,a\big(\,\|X\|^2-\|Y\|^2\,\big)&=\varepsilon^k\,a\,\langle X+Y\,,X-Y\rangle\\ &\le\frac{1}{2}\big((\varepsilon^k)^2\,\|X+Y\|^2+a^2\,\|X-Y\|^2\,\big)\\ &\le(\varepsilon^k)^2\big(\,\|X\|^2+\|Y\|^2\,\big)+\frac{a^2}{2}\,\|X-Y\|^2\\ &\le(\varepsilon^k)^2\big(\,\|X\|^2+\|Y\|^2\,\big)+a^2\,\|X-Y\|^2\;,\end{aligned}$$

permettent de transformer l'inégalité (9.22) en :

$$\begin{aligned}\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)-\boldsymbol{\Psi}^k &\le \varepsilon^k\big(L(u^\sharp,p^\sharp)-L(\boldsymbol{U}^k,p^\sharp)\big)\\ &\quad+\frac{(\varepsilon^k)^2}{2}\Big(\big\|\boldsymbol{U}^{k+1}-u^\sharp\big\|^2+\big\|(\boldsymbol{U}^k-u^\sharp\big\|^2\Big)\\ &\quad+\frac{a^2}{2}\big\|\boldsymbol{U}^{k+1}-\boldsymbol{U}^k\big\|^2\\ &\quad+(\varepsilon^k)^2\Big(c_5\boldsymbol{\Psi}^k+c_6\Big)\,.\end{aligned}$$

Utilisant (9.19) et (9.20), on en déduit l'existence de constantes positives c_7, c_8 et c_9 telles que :

$$\begin{aligned}\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)-\boldsymbol{\Psi}^k &\le \varepsilon^k\big(L(u^\sharp,p^\sharp)-L(\boldsymbol{U}^k,p^\sharp)\big)\\ &\quad+(\varepsilon^k)^2\Big(c_7\boldsymbol{\Psi}^k+c_8+c_9\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)\Big)\,.\end{aligned}$$

7. à savoir $J(u^\sharp)\ge J(\boldsymbol{U}^k)+\langle\boldsymbol{G}^k\,,u^\sharp-\boldsymbol{U}^k\rangle+(a/2)\big\|u^\sharp-\boldsymbol{U}^k\big\|^2$

3. **Analyse de convergence**. La suite de la démonstration est identique à celle du Théorème 9.5 : l'application directe du Corollaire 8.9 montre la suite de variables aléatoires $\{\boldsymbol{\Psi}^k\}_{k\in\mathbb{N}}$ converge presque sûrement vers une variable aléatoire $\boldsymbol{\Psi}^\infty$ bornée presque sûrement, et on a :

$$\sum_{k=0}^{+\infty} \varepsilon^k \big(L(\boldsymbol{U}^k, p^\sharp) - L(u^\sharp, p^\sharp)\big) < +\infty \ , \quad \mathbb{P}\text{-p.s.} \ .$$

4. **Limites des suites**. Comme dans le Théorème 9.5, avec de plus la forte convexité de J et donc l'unicité de la solution $u^\sharp$, on conclut que la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ engendrée par l'Algorithme 9.7 est bornée presque sûrement et converge vers $u^\sharp$. □

Remarque 9.9. Il est clair que la forte convexité de J est nécessaire pour pouvoir prendre un pas ρ constant dans l'étape de mise à jour des multiplicateurs, car sans cette condition, même en supposant J strictement convexe, la fonction duale, tout en restant différentiable, ne serait pas forcément à gradient Lipschitzien si bien qu'il faudrait utiliser des « petits pas » pour les mises à jour de p^k (voir le Lemme 4.4 et la discussion qui le précède).

Remarque 9.10. Une légère modification de la preuve permet de choisir le « grand pas » ρ tel que :

$$0 < \rho < \frac{2a}{L_\Theta^2} \ .$$

En effet, de la condition $\rho < 2a/L_\Theta^2$, on déduit l'existence d'un $\delta > 0$, que l'on peut supposer inférieur à $1/2$, tel que :

$$\rho L_\Theta^2 \leq 2a(1-\delta) \ . \tag{9.23}$$

Utilisant le fait la fonction $L(\cdot, p^\sharp)$ est fortement convexe de constante a, et que son unique minimum est atteint au point $u^\sharp$, on obtient que :

$$L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp) \leq -\frac{a}{2}\big\|\boldsymbol{U}^k - u^\sharp\big\|^2 \ .$$

Pour tout $\delta \in]0, 1/2[$, on dispose donc de la majoration :

$$\begin{aligned} L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp) \leq -(1-2\delta)\frac{a}{2}\big\|\boldsymbol{U}^k - u^\sharp\big\|^2 \\ + 2\delta\big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)\big) \ . \end{aligned} \tag{9.24}$$

L'utilisation de (9.23) et (9.24) dans l'inégalité (9.21) conduisent alors à :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{\Psi}^{k+1} \mid \mathcal{F}^k\big) - \boldsymbol{\Psi}^k \leq 2\delta\varepsilon^k\big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)\big) \\ + \varepsilon^k a(1-\delta)\Big(\big\|\boldsymbol{U}^{k+1} - u^\sharp\big\|^2 - \big\|\boldsymbol{U}^k - u^\sharp\big\|^2\Big) \\ + (\varepsilon^k)^2\Big(c_5\boldsymbol{\Psi}^k + c_6\Big) \ . \end{aligned}$$

La suite de la preuve est inchangée.

9.6 Cas simplement convexe et Lagrangien augmenté

On se place maintenant dans le cas où la fonction J est *simplement convexe*. On dispose encore d'un algorithme de résolution du problème (9.1) et du théorème de convergence associé pourvu que l'on fasse appel au Lagrangien augmenté plutôt qu'au Lagrangien ordinaire. Le Lagrangien augmenté a été présenté au Chapitre 5. Pour le problème (9.1), il s'écrit :

$$L_c(u,p) = J(u) + \zeta_c\big(p, \Theta(u)\big)\ , \tag{9.25a}$$

l'expression de la fonction ζ_c étant donnée par :

$$\zeta_c(p,\theta) = \frac{1}{2c}\big(\big\|\mathrm{proj}_{C^*}\left(p+c\theta\right)\big\|^2 - \|p\|^2\big)\ . \tag{9.25b}$$

On rappelle que la fonction ζ_c est concave en p, convexe en θ et qu'elle est différentiable, l'expression de ses gradients partiels étant :

$$\nabla_p\zeta_c(p,\theta) = \frac{1}{c}\big(\mathrm{proj}_{C^*}\left(p+c\theta\right) - p\big)\ , \tag{9.25c}$$

$$\nabla_\theta\zeta_c(p,\theta) = \mathrm{proj}_{C^*}\left(p+c\theta\right)\ . \tag{9.25d}$$

On a vu au Chapitre 5 dans le cas *déterministe* un algorithme utilisant à la fois le principe du problème auxiliaire et le Lagrangien augmenté. La k-ième itération de cet algorithme s'écrit :

$$\begin{aligned} u^{k+1} &= \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}}\ K(u) + \big\langle \varepsilon\nabla J(u^k) - \nabla K(u^k)\,, u\big\rangle \\ &\qquad\qquad + \varepsilon\big\langle \nabla_\theta\zeta_c\big(p^k, \Theta(u^k)\big)\,, \Theta(u)\big\rangle\ , \\ p^{k+1} &= p^k + \rho\nabla_p\zeta_c\big(p^k, \Theta(u^{k+1})\big)\ , \end{aligned}$$

et son principal avantage est que l'on peut prouver sa convergence sans qu'il soit nécessaire de faire une hypothèse de forte convexité sur la fonction J.

L'extension au cas stochastique de cet algorithme consiste à remplacer le gradient $\nabla J(u^k)$ par $\nabla_u j(u^k, w^{k+1})$; la résolution du problème (9.1) est alors remplacée par la résolution de la suite de problèmes auxiliaires :

$$\begin{aligned} u^{k+1} &\in \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}} K(u) + \big\langle \varepsilon^k\nabla_u j(u^k, w^{k+1}) - \nabla K(u^k)\,, u\big\rangle \\ &\qquad\qquad + \varepsilon^k\big\langle \nabla_\theta\zeta_c\big(p^k, \Theta(u^k)\big)\,, \Theta(u)\big\rangle\ . \qquad (9.26a) \\ p^{k+1} &= p^k + \varepsilon^k\nabla_p\zeta_c\big(p^k, \Theta(u^{k+1})\big)\ . \qquad (9.26b) \end{aligned}$$

Algorithme 9.11.

1. Choisir $(u^0, p^0) \in U^{\mathrm{ad}} \times C^*$, et une suite $\big\{\varepsilon^k\big\}_{k\in\mathbb{N}}$ de réels positifs.

2. À l'itération k, effectuer un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$.
3. Calculer u^{k+1} solution du problème auxiliaire (9.26a).
4. Calculer p^{k+1} par la formule de mise à jour (9.26b).
5. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Remarque 9.12. Du point de vue algorithmique, on notera les différences suivantes entre l'utilisation du Lagrangien augmenté et celle d'un Lagrangien ordinaire.

– Dans l'étape (9.26a) de minimisation en u, on utilise dans le dernier produit scalaire le terme $\nabla_\theta \zeta_c\big(p^k, \Theta(u^k)\big) = \mathrm{proj}_{C^*}\big(p^k + c\Theta(u^k)\big)$ plutôt que le terme p^k : ceci correspond à une sorte d'anticipation dans l'utilisation du multiplicateur pour le calcul de u^{k+1}.
– L'expression de $\nabla_p \zeta_c\big(p^k, \Theta(u^{k+1})\big)$ permet d'interpréter la remise à jour de p^{k+1} comme la succession des deux opérations suivantes :

$$\text{projection :}\quad p^{k+1/2} = \mathrm{proj}_{C^*}\big(p^k + c\Theta(u^{k+1})\big)\,, \tag{9.27a}$$

$$\text{relaxation :}\quad p^{k+1} = \big(1 - (\varepsilon^k/c)\big)p^k + \big(\varepsilon^k/c\big)p^{k+1/2}\,. \tag{9.27b}$$

On en déduit qu'initialiser l'algorithme avec un $p^0 \in C^*$ conduit à des multiplicateurs $p^k \in C^*$ pourvu que le cône C soit convexe (et que les pas ε^k restent plus petits que le coefficient c, afin que (9.27b) puisse s'interpréter comme une combinaison convexe de p^k et $p^{k+1/2}$).

On considère alors un échantillon $\{\boldsymbol{W}^k\}_{k\in\mathbb{N}}$ de taille infinie de la variable aléatoire $\boldsymbol{W}$, et les étapes 3 et 4 de l'algorithme 9.11 prennent la forme :

$$\min_{u\in U^{\mathrm{ad}}} K(u) + \big\langle \varepsilon^k \nabla_u j(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) - \nabla K(\boldsymbol{U}^k)\,, u\big\rangle + \varepsilon^k \big\langle \mathrm{proj}_{C^*}\big(\boldsymbol{P}^k + c\Theta(\boldsymbol{U}^k)\big)\,, \Theta(u)\big\rangle\,, \tag{9.28a}$$

dont la solution est notée $\boldsymbol{U}^{k+1}$, et

$$\boldsymbol{P}^{k+1} = \boldsymbol{P}^k + \varepsilon^k \nabla_p \zeta_c\big(\boldsymbol{P}^k, \Theta(\boldsymbol{U}^{k+1})\big)\,. \tag{9.28b}$$

On rappelle que les opérations de minimisation et de projection dans (9.28a) sont effectuées ω par ω. Le théorème suivant précise les conditions de convergence de l'Algorithme 9.11.

Théorème 9.13. *On suppose que les hypothèses suivantes sont vérifiées.*

1. *U^{ad} est une partie convexe fermée non vide d'un l'espace de Hilbert $\mathcal{U}$, et C est un cône convexe fermé saillant d'un autre espace de Hilbert $\mathcal{C}$.*
2. *La fonction $j : \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ est une intégrande normale, et l'espérance de $j(u, \boldsymbol{W})$ existe pour tout $u \in U^{\mathrm{ad}}$.*

3. *Pour tout $w \in \mathcal{W}$, la fonction $j(\cdot, w) : \mathcal{U} \to \mathbb{R}$ est propre, convexe, semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant U^{ad}, et son gradient partiel par rapport à u est noté $\nabla_u j(u, w)$.*
4. *La fonction $j(\cdot, w)$ vérifie l'hypothèse (8.6) de gradient linéairement borné en u uniformément en w.*
5. *La fonction J est Lipschitzienne, coercive sur l'ensemble U^{ad}.*
6. *La fonction Θ est C-convexe, Lipschitzienne de constante L_Θ.*
7. *Les contraintes sont qualifiées.*
8. *La fonction K est propre, fortement convexe de constante b, semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant U^{ad}.*
9. *La suite $\{\varepsilon^k\}_{k\in\mathbb{N}}$ est une σ-suite.*

On a alors les conclusions suivantes.

1. *Le problème (9.1) admet un ensemble de points selle $U^\sharp \times P^\sharp$ non vide.*
2. *Le problème (9.28a) admet une solution $\boldsymbol{U}^{k+1}$ unique.*
3. *Les suites $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ et $\{\boldsymbol{P}^k\}_{k\in\mathbb{N}}$ engendrées par l'Algorithme 9.11 sont bornées presque sûrement, et tout point d'accumulation d'une réalisation de la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ appartient à $U^\sharp$, ensemble des solutions du problème.*

Preuve. Pour alléger les écritures, on utilise les notations :

$$\begin{aligned} g^k &= \nabla_u j\big(u^k, w^{k+1}\big)\,, \\ q^k &= \mathrm{proj}_{C^*}\big(p^k + c\Theta(u^k)\big)\,, \\ p^{k+1/2} &= \mathrm{proj}_{C^*}\big(p^k + c\Theta(u^{k+1})\big)\,. \end{aligned}$$

La démonstration suivant le même schéma que celle des Théorèmes 9.5 et 9.8 et contenant des points très semblables à ceux des démonstrations de ces théorèmes, on détaillera surtout les différences qui apparaissent dans le cadre de cette variante. Le fait que u^{k+1} soit solution du problème (9.26a) est caractérisé par le fait que, pour tout $u \in U^{\mathrm{ad}}$,

$$\begin{aligned} \big\langle \nabla K(u^{k+1}) - \nabla K(u^k) + \varepsilon^k g^k\,, u - u^{k+1}\big\rangle \\ + \varepsilon^k \big\langle q^k\,, \Theta(u) - \Theta(u^{k+1})\big\rangle \geq 0\,. \end{aligned} \tag{9.29}$$

1. **Choix de la fonction de Lyapunov**. Soit $(u^\sharp, p^\sharp) \in U^\sharp \times P^\sharp$ un point selle du problème (9.1). On choisit la fonction de Lyapunov ℓ de la forme :

$$\ell(u, p) = K(u^\sharp) - K(u) - \big\langle \nabla K(u)\,, u^\sharp - u\big\rangle + \frac{1}{2}\big\|p - p^\sharp\big\|^2\,,$$

et l'on note :

$$\psi^k = \ell(u^k, p^k)\,.$$

On déduit de la forte convexité de K et de la définition de ℓ les deux relations :

$$\left\|u^k - u^\sharp\right\|^2 \leq \frac{2}{b}\,\psi^k\,, \tag{9.30a}$$

$$\left\|p^k - p^\sharp\right\|^2 \leq 2\,\psi^k\,, \tag{9.30b}$$

ce qui prouve que ℓ est bornée inférieurement et coercive.

2. **Majorations**.
Le fait que la projection soit non expansive, le caractère Lipschitz de Θ et les inégalités (9.30) permettent d'écrire :

$$\begin{aligned}\left\|q^k\right\| &= \left\|\mathrm{proj}_{C^*}\left(p^k + c\Theta(u^k)\right)\right\| \\ &\leq \left\|p^k + c\Theta(u^k) - \left(p^\sharp + c\Theta(u^\sharp)\right)\right\| + \left\|p^\sharp + c\Theta(u^\sharp)\right\| \\ &\leq \left\|p^k - p^\sharp\right\| + cL_\Theta\left\|u^k - u^\sharp\right\| + \left\|p^\sharp + c\Theta(u^\sharp)\right\|\,,\end{aligned}$$

d'où l'existence de coefficients a_1 et a_2 tels que

$$\left\|q^k\right\| \leq a_1 + a_2\sqrt{\psi^k}\,. \tag{9.31}$$

a. Pour la quantité $\left\|u^{k+1} - u^k\right\|$, un raisonnement identique à celui fait dans la démonstration du Théorème 9.5 associé à la majoration (9.31) montre que :

$$\left\|u^{k+1} - u^k\right\| \leq \varepsilon^k\left(a_3\sqrt{\psi^k} + a_4\right)\,. \tag{9.32}$$

b. Dans la mesure où $p^0 \in C^*$ et donc $p^k \in C^*$ pour tout indice k, comme la projection est non expansive et Θ Lipschitzienne, on a :

$$\begin{aligned}\left\|p^{k+1/2} - p^k\right\| &= \left\|\mathrm{proj}_{C^*}\left(p^k + c\Theta(u^{k+1})\right) - p^k\right\| \\ &\leq cL_\Theta\left\|u^{k+1} - u^\sharp\right\| + c\left\|\Theta(u^\sharp)\right\|\,.\end{aligned}$$

De (9.30a), on conclut à l'existence de coefficients b_1 et b_2 tels que [8]

$$\left\|p^{k+1/2} - p^k\right\| \leq b_1 + b_2\sqrt{\psi^{k+1}}\,. \tag{9.33}$$

La relation (9.27) définissant p^{k+1} s'écrit sous la forme :

$$p^{k+1} - p^k = \frac{\varepsilon^k}{c}\left(p^{k+1/2} - p^k\right)\,. \tag{9.34}$$

De la majoration (9.33), on déduit l'existence de coefficients b_3 et b_4 tels que :

$$\left\|p^{k+1} - p^k\right\| \leq \varepsilon^k\left(b_3 + b_4\sqrt{\psi^{k+1}}\right)\,. \tag{9.35}$$

8. Un calcul analogue permettrait, dans le cas où $p^k \notin C^*$, d'obtenir l'existence de coefficients b_1, b_2 et b_3 tels que $\left\|p^{k+1/2} - p^k\right\| \leq b_1 + b_2\sqrt{\psi^k} + b_3\sqrt{\psi^{k+1}}$.

Enfin, utilisant que la projection est contractante et que l'application Θ est Lipschitzienne, on a :

$$\left\|p^{k+1/2} - q^k\right\| \leq cL_\Theta \left\|u^{k+1} - u^k\right\| ,$$

et donc, par (9.32), l'existence de coefficients b_5 et b_6 tels que

$$\left\|p^{k+1/2} - q^k\right\| \leq \varepsilon^k \left(b_5 + b_6\sqrt{\psi^k}\right) . \tag{9.36}$$

c. L'écart $\psi^{k+1} - \psi^k$ s'écrit :

$$\begin{aligned}\psi^{k+1} - \psi^k = &\underbrace{K(u^k) - K(u^{k+1}) - \left\langle \nabla K(u^k) \,, u^k - u^{k+1}\right\rangle}_{T_1} \\ &+ \underbrace{\left\langle \nabla K(u^k) - \nabla K(u^{k+1}) \,, u^\sharp - u^{k+1}\right\rangle}_{T_2} \\ &+ \underbrace{\frac{1}{2}\left\|p^{k+1} - p^\sharp\right\|^2 - \frac{1}{2}\left\|p^k - p^\sharp\right\|^2}_{T_3} .\end{aligned}$$

- La fonction auxiliaire K étant convexe, le terme T_1 est négatif ou nul.
- Par la condition d'optimalité (9.29) évalué en $u = u^\sharp$, on obtient :

$$T_2 \leq \varepsilon^k \left\langle g^k \,, u^\sharp - u^{k+1}\right\rangle + \varepsilon^k \left\langle q^k \,, \Theta(u^\sharp) - \Theta(u^{k+1})\right\rangle .$$

Comme ζ_c est convexe en θ, et que $q^k = \nabla_\theta \zeta_c\left(p^k, \Theta(u^k)\right)$, on a :

$$\left\langle q^k \,, \Theta(u^\sharp) - \Theta(u^k)\right\rangle \leq \zeta_c\left(p^k, \Theta(u^\sharp)\right) - \zeta_c\left(p^k, \Theta(u^k)\right) .$$

Par des calculs élémentaires, on montre alors que :

$$\begin{aligned}T_2 \leq\ &\varepsilon^k \left\langle g^k \,, u^\sharp - u^k\right\rangle \\ &+ \varepsilon^k \left(\zeta_c\left(p^k, \Theta(u^\sharp)\right) - \zeta_c\left(p^k, \Theta(u^k)\right)\right) \\ &+ \underbrace{\varepsilon^k \left\|u^{k+1} - u^k\right\| \left(\left\|g^k\right\| + L_\Theta \left\|q^k\right\|\right)}_{T_{2,1}} .\end{aligned}$$

Utilisant (9.32), l'hypothèse (8.6) et les relations (9.30a) et (9.31), on obtient l'existence de constantes c_1 et c_2 telles que :

$$T_{2,1} \leq \left(\varepsilon^k\right)^2 \left(c_1 + c_2\psi^k\right) .$$

- Le terme T_3 s'écrit :

$$\begin{aligned}T_3 &= \frac{1}{2}\left\langle p^{k+1} - p^k \,, p^{k+1} + p^k - 2p^\sharp\right\rangle \\ &= \frac{1}{2}\left\|p^{k+1} - p^k\right\|^2 + \left\langle p^{k+1} - p^k \,, p^k - p^\sharp\right\rangle .\end{aligned}$$

Utilisant la relation (9.34) et par des calculs élémentaires, on déduit :

$$T_3 = \frac{1}{2c^2}\big(\varepsilon^k\big)^2\big\|p^{k+1/2}-p^k\big\|^2 + \frac{\varepsilon^k}{c}\big\langle q^k-p^k\,,p^k-p^\sharp\big\rangle + \frac{\varepsilon^k}{c}\big\langle p^{k+1/2}-q^k\,,p^k-p^\sharp\big\rangle\,.$$

Le gradient partiel $\nabla_p\zeta_c\big(p^k,\Theta(u^k)\big)$ étant égal à $\big(q^k-p^k\big)/c$, on déduit de la concavité en p de ζ_c que :

$$\frac{1}{c}\big\langle q^k-p^k\,,p^k-p^\sharp\big\rangle \le \zeta_c\big(p^k,\Theta(u^k)\big)-\zeta_c\big(p^\sharp,\Theta(u^k)\big)\,.$$

On obtient finalement :

$$T_3 \le \varepsilon^k\big(\zeta_c\big(p^k,\Theta(u^k)\big)-\zeta_c\big(p^\sharp,\Theta(u^k)\big)\big) + \underbrace{\frac{1}{2c^2}\big(\varepsilon^k\big)^2\big\|p^{k+1/2}-p^k\big\|^2}_{T_{3,1}} + \underbrace{\frac{\varepsilon^k}{c}\big\|p^{k+1/2}-q^k\big\|\big\|p^k-p^\sharp\big\|}_{T_{3,2}}\,.$$

Utilisant les relations (9.33), (9.36) et (9.30b) ainsi que l'inégalité générale $2ab \le a^2+b^2$, on en déduit l'existence de coefficients c_3, c_4 et c_5 tels que :

$$T_{3,1}+T_{3,2} \le \big(\varepsilon^k\big)^2\big(c_3+c_4\psi^k+c_5\psi^{k+1}\big)\,.$$

Regroupant les majorations des termes T_1, T_2 et T_3 et des sous-termes $T_{2,1}$ et $T_{3,1}+T_{3,2}$, on obtient :

$$\psi^{k+1}-\psi^k \le \varepsilon^k\big(\big\langle g^k\,,u^\sharp-u^k\big\rangle + \zeta_c\big(\Theta(u^\sharp),p^k\big)-\zeta_c\big(\Theta(u^k),p^\sharp\big)\big) + \big(\varepsilon^k\big)^2\big(c_6+c_7\psi^k+c_8\psi^{k+1}\big)\,.$$

Écrivant cette dernière inégalité en termes de variables aléatoires et en en prenant l'espérance conditionnelle par rapport à la tribu engendrée par $(\boldsymbol{W}^1,\dots,\boldsymbol{W}^k)$, puis utilisant la convexité de la fonction J et la définition du Lagrangien augmenté, on obtient :

$$\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)-\boldsymbol{\Psi}^k \le \varepsilon^k\big(L_c(u^\sharp,\boldsymbol{P}^k)-L_c(\boldsymbol{U}^k,p^\sharp)\big) + (\varepsilon^k)^2\big(c_6+c_7\boldsymbol{\Psi}^k+c_8\mathbb{E}\big(\boldsymbol{\Psi}^{k+1}\;\big|\;\mathcal{F}^k\big)\big)\,. \quad (9.37)$$

3. **Analyse de convergence**. Par le Corollaire 8.9 du théorème de Robbins-Siegmund, on en déduit que la suite $\{\boldsymbol{\Psi}^k\}_{k\in\mathbb{N}}$ converge presque sûrement vers une variable aléatoire bornée, et que la série de terme général $\varepsilon^k\big(L_c(u^\sharp,\boldsymbol{P}^k)-L_c(\boldsymbol{U}^k,p^\sharp)\big)$ est convergente. De la double inégalité du point selle :
$$L_c(u^\sharp,\boldsymbol{P}^k)\le L_c(u^\sharp,p^\sharp)\le L_c(\boldsymbol{U}^k,p^\sharp)\ ,$$
on en déduit qu'il en est de même des deux séries de terme général $\varepsilon^k\big(L_c(u^\sharp,p^\sharp)-L_c(u^\sharp,\boldsymbol{P}^k)\big)$ et $\varepsilon^k\big(L_c(\boldsymbol{U}^k,p^\sharp)-L_c(u^\sharp,p^\sharp)\big)$.
4. **Limites des suites**. Les majorations (9.32) et (9.35) permettent l'utilisation du Lemme 8.10 sur les deux séries dont on a montré la convergence au point 3 ci-dessus. Les suites $\{L_c(\boldsymbol{U}^k,p^\sharp)\}_{k\in\mathbb{N}}$ et $\{L_c(u^\sharp,\boldsymbol{P}^k)\}_{k\in\mathbb{N}}$ convergent donc toutes les deux vers $L_c(u^\sharp,p^\sharp)$. Des majorations (9.30), on déduit que les suites $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ et $\{\boldsymbol{P}^k\}_{k\in\mathbb{N}}$ sont bornées presque sûrement. Grâce aux propriétés de continuité du Lagrangien L_c, on obtient que tout point d'accumulation d'une réalisation de la suite $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ est solution du problème (9.1), et que tout point d'accumulation d'une réalisation de la suite $\{\boldsymbol{P}^k\}_{k\in\mathbb{N}}$ est un multiplicateur optimal du Lagrangien. □

Remarque 9.14. Comme on l'a vu au §9.5, le fait de supposer la fonction objectif J fortement convexe implique que la fonction duale ψ (définie par (2.5)) est différentiable de dérivée Lipschitzienne (par le Lemme 4.4), ce qui a permis d'utiliser des « grands pas » pour la remise à jour des variables duales dans l'Algorithme 9.7. Mais faire appel, comme on l'a fait ici, au Lagrangien augmenté a pour conséquence d'utiliser la transformée de Moreau-Yosida ψ_c de la fonction duale ψ, et on sait par le Théorème 5.14 que ψ_c est différentiable de dérivée Lipschitzienne. On peut donc espérer disposer d'une version de l'Algorithme 9.11 dans laquelle les variables duales seraient remises à jour avec des « grands pas ». À notre connaissance, la convergence d'un tel algorithme n'a pas été étudiée.

9.7 Conclusions

On a montré dans ce chapitre comment généraliser la méthode du PPA stochastique au cas des problèmes d'optimisation soumis à des contraintes déterministes, et on a vu que cette généralisation était possible dans le cadre de l'algorithme de Arrow-Hurwicz et non dans celui d'Uzawa. Cependant, le fait de pouvoir traiter des problèmes d'optimisation stochastique sous contraintes déterministes ne couvre pas l'ensemble des cas que l'on peut rencontrer en pratique.

En fait, on peut distinguer au moins les trois types de contraintes suivantes dans le cadre de l'optimisation stochastique :

1. les contraintes presque sûres :
$$\theta(u,\boldsymbol{W})\in -C\quad \mathbb{P}\text{-p.s.}\ ,$$

2. les contraintes en probabilité :

$$\mathbb{P}\big(\theta(u, \boldsymbol{W}) \in -C\big) \geq \pi \, .$$

3. les contraintes en espérance :

$$\mathbb{E}\big(\theta(u, \boldsymbol{W})\big) \in -C \, .$$

Les contraintes presque sûres sont en général trop restrictives pour pouvoir être prises en compte dans le cadre de la boucle ouverte, car il s'agit en fait de satisfaire un ensemble de contraintes indexées par ω, pour presque tout $\omega \in \Omega$. Quant aux contraintes en probabilité, leur traitement direct entraîne de grandes difficultés mathématiques (perte de convexité, voir de connexité de l'ensemble admissible induit par les contraintes). Enfin, le cas des contraintes en espérance n'est pas fréquent en pratique, car il est souvent difficile de donner un sens à la satisfaction « en moyenne » d'un besoin.

Cependant, l'extension des algorithmes vus dans ce chapitre au cas des contraintes en espérance paraît relativement abordable : dans l'esprit du gradient stochastique, il semblerait raisonnable de profiter des itérations de gradient pour reconstituer l'espérance du gradient de la fonction objectif *et* l'espérance de la contrainte. De plus, le fait de pouvoir écrire une contrainte en probabilité comme une contrainte en espérance :

$$\mathbb{P}\big(\theta(u, \boldsymbol{W}) \in -C\big) = \mathbb{E}\big(\mathbf{1}_{\{\theta(u,\boldsymbol{W})\in -C\}}\big) \, , \tag{9.38}$$

où $\mathbf{1}$ dénote la fonction indicatrice d'ensemble[9], fait que l'on peut espérer s'attaquer numériquement aux problèmes sous contraintes en probabilité par le biais des contraintes en espérance. Il faudra alors, pour surmonter les difficultés liées à la non convexité engendrée par la fonction indicatrice[10] présente dans (9.38), utiliser un Lagrangien augmenté sur une contrainte en espérance, et donc savoir appliquer le gradient stochastique à des fonctions non linéaires de l'espérance. Ce point sera abordé dans le Chapitre 10.

9. La fonction indicatrice d'un sous-ensemble mesurable A de Ω est définie par :

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{sinon} \end{cases} \, ,$$

et donc $\mathbb{P}(A) = \mathbb{E}(\mathbf{1}_A)$.

10. Plus grave encore, la fonction indicatrice est *discontinue*, et on ne peut donc parler de son gradient.

10

Extensions de la méthode du gradient stochastique

On présente dans ce chapitre un certain nombre de résultats concernant des extensions et des variations autour du gradient stochastique.

On se place dans le même cadre qu'au Chapitre 9 : on considère un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et une variable aléatoire $\boldsymbol{W}$ à valeurs dans l'espace $\mathcal{W}$ muni de sa tribu $\mathcal{W}$. On considère un espace de Hilbert $\mathcal{U}$ ainsi qu'une partie non vide U^{ad} de $\mathcal{U}$, et une fonction J définie sur $\mathcal{U}$ à valeurs dans $\overline{\mathbb{R}}$. Pour exprimer les contraintes, on introduit un autre espace de Hilbert $\mathcal{C}$, un cône C inclus dans cet espace et une application Θ définie sur $\mathcal{U}$ à valeurs dans $\mathcal{C}$. On s'intéresse alors au problème suivant :

$$\min_{u \in U^{\mathrm{ad}}} J(u) \qquad \text{sous la contrainte} \qquad \Theta(u) \in -C \,. \tag{10.1}$$

Comme au Chapitre 9, on suppose que la fonction J est l'espérance d'une fonction j définie sur $\mathcal{U} \times \mathcal{W}$ à valeurs dans $\overline{\mathbb{R}}$, supposée intégrable pour tout $u \in U^{\mathrm{ad}}$:

$$J(u) = \mathbb{E}\left(j(u, \boldsymbol{W})\right) \,. \tag{10.2a}$$

On se place dans le cadre plus général qu'au Chapitre 9 en supposant que la fonction Θ représente elle aussi l'espérance d'une fonction θ définie sur $\mathcal{U} \times \mathcal{W}$ à valeurs dans $\mathcal{C}$, intégrable pour tout $u \in U^{\mathrm{ad}}$:

$$\Theta(u) = \mathbb{E}\left(\theta(u, \boldsymbol{W})\right) \,. \tag{10.2b}$$

Tout comme un algorithme de type gradient stochastique utilise le gradient de la fonction j évalué en des réalisations de la variable aléatoire $\boldsymbol{W}$ plutôt que le gradient de la fonction J, on va montrer comme tirer parti de la forme particulière (10.2b) et utiliser dans un algorithme de type Arrow-Hurwicz stochastique des évaluations de la fonction θ plutôt que de son espérance Θ.

10.1 Contrainte en espérance et Lagrangien

L'extension « naturelle » de l'algorithme du gradient stochastique issu du principe du problème auxilaire au cas des contraintes en espérance consiste,

à partir des relations (9.9), à remplacer les évaluations de Θ et de sa dérivée par rapport à u par des évaluations de θ et de sa dérivée par rapport à u. On aborde alors la résolution du problème (10.1) par la résolution de la suite de problèmes auxiliaires :

$$u^{k+1} \in \arg\min_{u\in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon^k g^k - \nabla K(u^k)\,, u\right\rangle + \varepsilon^k \left\langle p^k\,, \vartheta^k \cdot u\right\rangle\,, \tag{10.3a}$$

$$p^{k+1} = \mathrm{proj}_{C^*}\left(p^k + \rho^k\, \theta(u^{k+1}, w^{k+1})\right)\,, \tag{10.3b}$$

relations dans lesquelles on a utilisé les notations :

$$g^k = \nabla_u j(u^k, w^{k+1})\,,$$
$$\vartheta^k = \theta'_u(u^k, w^{k+1})\,,$$

pour représenter respectivement le gradient partiel par rapport à u de la fonction j et la dérivée partielle par rapport à u de la fonction θ. La notation $(\vartheta^k)^\top$ représente l'adjoint de l'opérateur linéaire ϑ^k.

L'algorithme qui en découle est le suivant.

Algorithme 10.1.

1. Choisir $(u^0, p^0) \in U^{\mathrm{ad}} \times C^*$, et deux suites $\left\{\varepsilon^k\right\}_{k\in\mathbb{N}}$ et $\left\{\rho^k\right\}_{k\in\mathbb{N}}$ de réels positifs.
2. À l'itération k, effectuer un tirage w^{k+1} de la variable aléatoire $\boldsymbol{W}$.
3. Calculer u^{k+1} solution du problème auxiliaire (10.3a).
4. Calculer p^{k+1} par la formule de mise à jour (10.3b).
5. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Remarque 10.2. Avec le choix de fonction auxiliaire $K(u) = \|u\|^2/2$, le problème auxiliaire (10.3) se met sous la forme :

$$u^{k+1} = \mathrm{proj}_{U^{\mathrm{ad}}}\left(u^k - \varepsilon^k\left(g^k + (\vartheta^k)^\top \cdot p^k\right)\right)\,,$$
$$p^{k+1} = \mathrm{proj}_{C^*}\left(p^k + \rho^k\, \theta(u^{k+1}, w^{k+1})\right)\,.$$

L'étude de la convergence de l'Algorithme 10.1, s'effectue en considérant un échantillon $\{\boldsymbol{W}^k\}_{k\in\mathbb{N}}$ de taille infinie de la variable aléatoire $\boldsymbol{W}$, le problème auxiliaire à l'étape k prenant alors la forme :

$$\min_{u\in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon^k \boldsymbol{G}^k - \nabla K(\boldsymbol{U}^k)\,, u\right\rangle + \varepsilon^k \left\langle \boldsymbol{P}^k\,, \boldsymbol{\vartheta}^k \cdot u\right\rangle\,, \tag{10.4a}$$

dont la solution est notée $\boldsymbol{U}^{k+1}$, et la mise à jour des multiplicateurs étant :

$$\boldsymbol{P}^{k+1} = \mathrm{proj}_{C^*}\left(\boldsymbol{P}^k + \rho^k\, \theta(\boldsymbol{U}^{k+1}, \boldsymbol{W}^{k+1})\right)\,. \tag{10.4b}$$

On rappelle que les opérations de minimisation dans (10.4a) et de projection dans (10.4b) sont effectuées « ω par ω ».

Théorème 10.3. *On suppose que les hypothèses suivantes sont vérifiées.*

1. U^{ad} *est une partie convexe fermée non vide d'un l'espace de Hilbert* $\mathcal{U}$*, et* C *est un cône convexe fermé saillant d'un autre espace de Hilbert* $\mathcal{C}$*.*
2. *La fonction* $j : \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ *est une intégrande normale, et l'espérance de* $j(u, \boldsymbol{W})$ *existe pour tout* $u \in U^{\mathrm{ad}}$*.*
3. *La fonction* $j(\cdot, w)$ *est propre, semi-continue inférieurement, et est différentiable sur un sous-ensemble ouvert contenant* U^{ad}*, pour tout* $w \in \mathcal{W}$*.*
4. *La fonction* $j(\cdot, w)$ *vérifie l'hypothèse* (8.6) *de gradient linéairement borné en* u *uniformément en* w*.*
5. *La fonction* J *est convexe, Lipschitzienne, coercive sur l'ensemble* U^{ad}*.*
6. *La fonction* $\theta : \mathcal{U} \times \mathcal{W} \to \mathcal{C}$ *est telle que l'application* $(u, w) \mapsto \langle p\,, \theta(u, w)\rangle$ *est une intégrande normale pour tout* $p \in C^*$*, et l'espérance de* $\theta(u, \boldsymbol{W})$ *existe pour tout* $u \in U^{\mathrm{ad}}$*.*
7. *La fonction* θ *est sous-Lipschitzienne uniformément en* w *:*

$$\exists \lambda > 0,\ \exists \mu > 0,\ \forall w \in \mathcal{W},\ \forall u, v \in U^{\mathrm{ad}}\ ,\quad \|\theta(u, w) - \theta(v, w)\| \leq \lambda \|u - v\| + \mu\ .$$

8. *Pour tout* $w \in \mathcal{W}$*, la fonction* θ *est différentiable, et sa différentielle par rapport à* u *est bornée par une constante* ς*, uniformément en* w*.*
9. *La variance associée à la fonction de contrainte* θ *est bornée par une fonction quadratique :*

$$\exists \gamma > 0,\ \exists \delta > 0,\ \forall u \in U^{\mathrm{ad}}\ ,\ \ \mathbb{E}\big(\big\|\theta(u, \boldsymbol{W}) - \Theta(u)\big\|^2\big) \leq \gamma \|u\|^2 + \delta\ .$$

10. *La fonction* Θ *est* C*-convexe, Lipschitzienne de constante* L_Θ*.*
11. *Les contraintes sont qualifiées et le Lagrangien* L *est stable en* u*.*
12. *La fonction* K *est propre, fortement convexe de constante* $\mathfrak{b}$*, semi-continue inférieurement, et elle est différentiable sur un sous-ensemble ouvert contenant* U^{ad}*.*
13. *Les deux suites* $\{\varepsilon^k\}_{k\in\mathbb{N}}$ *et* $\{\rho^k\}_{k\in\mathbb{N}}$ *sont des* σ*-suites.*
14. *La suite quotient* $\{\varepsilon^k/\rho^k\}_{k\in\mathbb{N}}$ *est décroissante.*

On a alors les conclusions suivantes.

1. *Le problème* (10.1) *admet un ensemble de points selle* $U^\sharp \times P^\sharp$ *non vide.*
2. *Le problème* (10.4a) *admet une solution* $\boldsymbol{U}^{k+1}$ *unique.*
3. *Pour tout* $p^\sharp \in P^\sharp$*, la suite de variables aléatoires* $\{L(\boldsymbol{U}^k, p^\sharp)\}_{k\in\mathbb{N}}$ *converge presque sûrement vers* $L(u^\sharp, p^\sharp)$*, avec* $u^\sharp \in U^\sharp$*.*
4. *Les suites* $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ *et* $\{\boldsymbol{P}^k\}_{k\in\mathbb{N}}$ *engendrées par l'Algorithme 10.1 sont bornées presque sûrement, et tout point d'accumulation d'une réalisation de la suite* $\{\boldsymbol{U}^k\}_{k\in\mathbb{N}}$ *appartient à* $U^\sharp$*, ensemble des solutions du problème.*

Remarque 10.4. L'hypothèse 13 implique que la série de terme général $\varepsilon^k \rho^k$ est convergente[1]. L'hypothèse 14 implique l'existence d'un réel positif α tel que $\varepsilon^k \leq \alpha \rho^k$ pour tout $k \in \mathbb{N}$.

Remarque 10.5. Comme on l'a déjà noté dans la Remarque 8.7, on ne fait pas ici d'hypothèse de convexité sur la fonction j, mais plutôt une hypothèse (moins restrictive) de convexité sur la fonction J. De même, on ne fait pas d'hypothèse de C-convexité sur la fonction $u \mapsto \theta(u, w)$, que l'on remplace par une hypothèse de C-convexité sur la fonction Θ. Mais l'absence d'une telle hypothèse sur θ interdit de considérer la variante du problème auxiliaire (10.3a) dans lequel on remplacerait le terme linéarisé $\langle p^k \, , \vartheta^k \cdot u \rangle$ par le terme $\langle p^k \, , \theta(u, w^{k+1}) \rangle$, car on ne saurait alors rien dire sur la convexité du problème auxiliaire de minimisation en résultant.

On introduit le « raccourci » suivant qui permettra de simplifier les écritures dans la preuve du Théorème (10.3). Étant donné deux suites $\{x^k\}_{k\in\mathbb{N}}$ et $\{y^k\}_{k\in\mathbb{N}}$ de réels positifs, la notation :

$$y^k \leq \mathcal{L}\big(x^k\big) \, , \tag{10.5}$$

signifie qu'il existe deux constantes réelles a et b positives telles que, pour tout $k \in \mathbb{N}$, on ait :

$$y^k \leq a \, x^k + b \, .$$

Dans cette notation $\mathcal{L}$, on ne précise pas les constantes a et b pourvu qu'elles ne dépendent pas de l'indice k. C'est pourquoi (10.5) implique, pour tout $\alpha \geq 0$, la même relation pour les suites $\{x^k + \alpha\}_{k\in\mathbb{N}}$ et $\{y^k\}_{k\in\mathbb{N}}$:

$$y^k \leq \mathcal{L}\big(x^k + \alpha\big) \, .$$

Par contre, dans le cas de la suite $\{\epsilon^k x^k\}_{k\in\mathbb{N}}$ (avec $\epsilon^k \geq 0$), on prendra garde à ce que (10.5) conduit à écrire :

$$\epsilon^k y^k \leq \mathcal{L}(\epsilon^k x^k) \, .$$

On dispose alors des « règles de calcul » suivantes[2] :

$$y^k \leq \mathcal{L}(x^k) \;\Rightarrow\; (y^k)^2 \leq \mathcal{L}((x^k)^2) \, , \tag{10.6a}$$

$$(y^k)^2 \leq \mathcal{L}((x^k)^2) \;\Rightarrow\; y^k \leq \mathcal{L}(x^k) \, , \tag{10.6b}$$

$$y^k \leq \mathcal{L}(x^k) \text{ et } z^k \leq \mathcal{L}(x^k) \;\Rightarrow\; y^k z^k \leq \mathcal{L}((x^k)^2) \, , \tag{10.6c}$$

$$(y^k)^2 \leq \mathcal{L}(x^k) \text{ et } (z^k)^2 \leq \mathcal{L}(x^k) \;\Rightarrow\; y^k z^k \leq \mathcal{L}(x^k) \, . \tag{10.6d}$$

La propriété (10.6a) provient de ce que $(ax + b)^2 \leq 2a^2x^2 + 2b^2$, et la propriété (10.6b) de $ax^2 + b \leq (\sqrt{a}x + \sqrt{b})^2$. La propriété (10.6c) est une

1. car on a $\varepsilon^k \rho^k \leq \big((\varepsilon^k)^2 + (\rho^k)^2\big)/2$
2. On rappelle que toutes les variables et coefficients utilisés ici sont positifs.

conséquence de $(a_1x+b_1)(a_2x+b_2) \leq \big(\max\{a_1,a_2\}x+\max\{b_1,b_2\}\big)^2$, et la propriété (10.6d) résulte de la combinaison des propriétés (10.6b) et (10.6c). Ces propriétés seront utilisées durant la démonstration du Théorème 10.3, qui est donnée maintenant.

Preuve. La démonstration des deux premières conclusions de ce théorème découle des théorèmes généraux relatifs à l'optimisation convexe en présence de contraintes. Le fait que la solution $\boldsymbol{U}^{k+1}$ du problème (10.4a) soit une variable aléatoire, et donc une fonction mesurable, provient de ce que l'on a supposé que $j(\cdot,\cdot)$ et $\langle p\,,\theta(\cdot,\cdot)\rangle$ étaient des intégrandes normales (voir la preuve du Théorème 8.5 pour plus de détails). La démonstration des deux dernières conclusions se fait en suivant le schéma « habituel » de preuve.

Le fait que u^{k+1} soit solution du problème (10.3a) est caractérisé pr la condition d'optimalité suivante :

$$\forall u \in U^{\mathrm{ad}},\ \big\langle \nabla K(u^{k+1}) - \nabla K(u^k) + \varepsilon^k\big(g^k + (\vartheta^k)^\top \cdot p^k\big)\,, u-u^{k+1}\big\rangle \geq 0\,. \quad (10.7)$$

1. **Choix de la fonction de Lyapunov**. Soit $(u^\sharp,p^\sharp) \in U^\sharp \times P^\sharp$ un point selle du problème (10.1). On définit :

$$\psi^k = K(u^\sharp) - K(u^k) - \big\langle \nabla K(u^k)\,, u^\sharp - u^k\big\rangle + \frac{\varepsilon^k}{2\rho^k}\big\|p^k - p^\sharp\big\|^2\,.$$

Comme dans le cas du Théorème 9.8, on notera qu'il n'existe pas de fonction de Lyapunov ℓ telle que $\psi^k = \ell(u^k,p^k)$.
Par la définition de ψ^k et la forte convexité de la fonction auxiliaire K, on obtient les inégalités :

$$\big\|u^k - u^\sharp\big\|^2 \leq \mathcal{L}\big(\psi^k\big)\,, \quad (10.8a)$$

$$\big\|p^k - p^\sharp\big\|^2 \leq \frac{\rho^k}{\varepsilon^k}\mathcal{L}\big(\psi^k\big)\,. \quad (10.8b)$$

On déduit alors de (10.8a) et de l'hypothèse (8.6) que :

$$\big\|g^k\big\|^2 \leq \mathcal{L}\big(\psi^k\big)\,. \quad (10.8c)$$

2. **Majorations**.
   a. On majore pour commencer la quantité $\big\|u^{k+1}-u^k\big\|$. Évaluant la condition d'optimalité (10.7) au point $u = u^k$ et utilisant la forte convexité de K, on obtient :

$$\begin{aligned}\varepsilon^k\big\langle g^k + (\vartheta^k)^\top \cdot p^k\,, u^k - u^{k+1}\big\rangle &\geq \big\langle \nabla K(u^k) - \nabla K(u^{k+1})\,, u^k - u^{k+1}\big\rangle \\ &\geq b\big\|u^k - u^{k+1}\big\|^2\,.\end{aligned}$$

On déduit de l'inégalité de Schwarz que l'on a :

$$\left\|u^{k+1}-u^k\right\| \leq \frac{\varepsilon^k}{b}\left\|g^k+(\vartheta^k)^\top\cdot p^k\right\| . \tag{10.9}$$

Par l'hypothèse 8 et la relation (10.8b), on obtient :

$$\left\|(\vartheta^k)^\top\cdot p^k\right\|^2 \leq \varsigma^2\left\|p^k\right\|^2 \leq \frac{\rho^k}{\varepsilon^k}\mathcal{L}\big(\psi^k\big) .$$

De cette dernière majoration et de la relation (10.8c), par l'inégalité triangulaire[3], on déduit :

$$\begin{aligned}\Big(\frac{\varepsilon^k}{b}\Big)^2\left\|g^k+(\vartheta^k)^\top\cdot p^k\right\|^2 \leq\ & \big(\varepsilon^k\big)^2\mathcal{L}\big(\psi^k\big)+\big(\varepsilon^k\rho^k\big)\mathcal{L}\big(\psi^k\big)\\ \leq\ & \big(\varepsilon^k\rho^k\big)\mathcal{L}\big(\psi^k\big) ,\end{aligned}$$

la dernière inégalité provenant du second point de la Remarque 10.4. On obtient finalement la majoration :

$$\left\|g^k+(\vartheta^k)^\top\cdot p^k\right\|^2 \leq \frac{\rho^k}{\varepsilon^k}\mathcal{L}\big(\psi^k\big) , \tag{10.10}$$

et donc, par (10.9) :

$$\left\|u^{k+1}-u^k\right\|^2 \leq \big(\varepsilon^k\rho^k\big)\mathcal{L}\big(\psi^k\big) . \tag{10.11}$$

b. On majore ensuite la quantité $\left\|p^{k+1}-p^\sharp\right\|^2$. Utilisant la relation $p^\sharp=\mathrm{proj}_{C^*}\big(p^\sharp+\rho^k\Theta(u^\sharp)\big)$, vraie pour tout $\rho^k>0$, ainsi que la relation (10.3b) définissant p^{k+1}, et comme l'opérateur de projection sur C^* est non expansif, on obtient :

$$\left\|p^{k+1}-p^\sharp\right\|^2 \leq \left\|p^k-p^\sharp+\rho^k\big(\theta(u^{k+1},w^{k+1})-\Theta(u^\sharp)\big)\right\|^2 .$$

Développant le carré, il vient :

$$\begin{aligned}\left\|p^{k+1}-p^\sharp\right\|^2 \leq\ & \left\|p^k-p^\sharp\right\|^2\\ & +2\rho^k\big\langle p^k-p^\sharp\,,\theta(u^{k+1},w^{k+1})-\Theta(u^\sharp)\big\rangle\\ & +\underbrace{\big(\rho^k\big)^2\left\|\theta(u^{k+1},w^{k+1})-\Theta(u^\sharp)\right\|^2}_{T_1} .\end{aligned} \tag{10.12}$$

Écrivant la différence $\theta(u^{k+1},w^{k+1})-\Theta(u^\sharp)$ sous la forme :

$$\begin{aligned}&\theta(u^{k+1},w^{k+1})-\theta(u^k,w^{k+1})\\ &\qquad+\theta(u^k,w^{k+1})-\Theta(u^k)\\ &\qquad\qquad+\Theta(u^k)-\Theta(u^\sharp) ,\end{aligned}$$

3. en fait, la relation $\|a+b\|^2\leq 2\big(\|a\|^2+\|b\|^2\big)\ldots$

et utilisant l'inégalité $\|x+y+z\|^2 \leq 3\big(\|x\|^2+\|y\|^2+\|z\|^2\big)$, il vient :

$$T_1 \leq 3\big(\rho^k\big)^2\Big(\underbrace{\big\|\theta(u^{k+1},w^{k+1})-\theta(u^k,w^{k+1})\big\|^2}_{T_{1,1}} + \underbrace{\big\|\theta(u^k,w^{k+1})-\Theta(u^k)\big\|^2}_{T_{1,2}} + \underbrace{\big\|\Theta(u^k)-\Theta(u^\sharp)\big\|^2}_{T_{1,3}}\Big) .$$

– Par l'hypothèse 7, la majoration (10.11) et utilisant la Remarque 10.4, on obtient :
$$T_{1,1} \leq \big(\varepsilon^k\rho^k\big)\mathcal{L}\big(\psi^k\big) \leq \mathcal{L}\big(\psi^k\big) .$$
– De l'hypothèse 10 et de la majoration (10.8a), on déduit :
$$T_{1,3} \leq \mathcal{L}\big(\psi^k\big) .$$

On obtient donc la majoration suivante :

$$T_1 \leq 3\big(\rho^k\big)^2\big\|\theta(u^k,w^{k+1})-\Theta(u^k)\big\|^2 + \big(\rho^k\big)^2\mathcal{L}\big(\psi^k\big) .$$

Reportant cette majoration dans (10.12), et multipliant de part et d'autre de l'inégalité par $\varepsilon^k/2\rho^k$, on obtient :

$$\begin{aligned}\frac{\varepsilon^k}{2\rho^k}\big\|p^{k+1}-p^\sharp\big\|^2 \leq{}& \frac{\varepsilon^k}{2\rho^k}\big\|p^k-p^\sharp\big\|^2 \\ &+ \varepsilon^k\big\langle p^k-p^\sharp\,,\theta(u^{k+1},w^{k+1})-\Theta(u^\sharp)\big\rangle \\ &+ \frac{3}{2}\big(\varepsilon^k\rho^k\big)\big\|\theta(u^k,w^{k+1})-\Theta(u^k)\big\|^2 \\ &+ \big(\varepsilon^k\rho^k\big)\mathcal{L}\big(\psi^k\big) . \end{aligned} \tag{10.13}$$

c. On majore pour finir l'écart $\psi^{k+1}-\psi^k$. On a :

$$\begin{aligned}\psi^{k+1}-\psi^k ={}& \underbrace{K(u^k)-K(u^{k+1})-\big\langle\nabla K(u^k)\,,u^k-u^{k+1}\big\rangle}_{T_{2,1}} \\ &+ \underbrace{\big\langle\nabla K(u^k)-\nabla K(u^{k+1})\,,u^\sharp-u^{k+1}\big\rangle}_{T_{2,2}} \\ &+ \frac{\varepsilon^{k+1}}{2\rho^{k+1}}\big\|p^{k+1}-p^\sharp\big\|^2-\frac{\varepsilon^k}{2\rho^k}\big\|p^k-p^\sharp\big\|^2 .\end{aligned}$$

Le terme $T_{2,1}$ est négatif ou nul par convexité de la fonction auxiliaire K et peut donc être négligé. Écrivant la condition d'optimalité (10.7) au

point $u = u^\sharp$, on majore le terme $T_{2,2}$ par $\varepsilon^k \langle g^k + (\vartheta^k)^\top \cdot p^k \, , u^\sharp - u^{k+1} \rangle$. Utilisant l'hypothèse de décroissance 14 et la majoration (10.13), il vient :

$$\begin{aligned}\psi^{k+1} - \psi^k \leq & \underbrace{\varepsilon^k \langle g^k + (\vartheta^k)^\top \cdot p^k \, , u^\sharp - u^{k+1} \rangle}_{T_{3,1}} \\ & + \underbrace{\varepsilon^k \langle p^k - p^\sharp \, , \theta(u^{k+1}, w^{k+1}) - \Theta(u^\sharp) \rangle}_{T_{3,2}} \\ & + \frac{3}{2} (\varepsilon^k \rho^k) \left\| \theta(u^k, w^{k+1}) - \Theta(u^k) \right\|^2 \\ & + (\varepsilon^k \rho^k) \mathcal{L}(\psi^k) \, .\end{aligned}$$

On écrit le terme $T_{3,1} + T_{3,2}$ sous la forme :

$$\begin{aligned}T_{3,1} + T_{3,2} = & \varepsilon^k \langle g^k + (\vartheta^k)^\top \cdot p^k \, , u^\sharp - u^k \rangle \\ & + \varepsilon^k \langle p^k - p^\sharp \, , \theta(u^k, w^{k+1}) - \Theta(u^\sharp) \rangle \\ & + \underbrace{\varepsilon^k \langle g^k + (\vartheta^k)^\top \cdot p^k \, , u^k - u^{k+1} \rangle}_{T_{4,1}} \\ & + \underbrace{\varepsilon^k \langle p^k - p^\sharp \, , \theta(u^{k+1}, w^{k+1}) - \theta(u^k, w^{k+1}) \rangle}_{T_{4,2}} \, .\end{aligned}$$

– Appliquant l'inégalité de Schwarz au terme $T_{4,1}$ et utilisant les relations (10.10) et (10.11) ainsi que la propriété (10.6d), il vient :

$$T_{4,1} \leq (\varepsilon^k \rho^k) \mathcal{L}(\psi^k) \, .$$

– Appliquant l'inégalité de Schwarz au terme $T_{4,2}$ et utilisant la relation (10.8b), l'hypothèse 7, la majoration (10.11) et pour finir la propriété (10.6d), il vient :

$$T_{4,2} \leq (\varepsilon^k \rho^k) \mathcal{L}(\psi^k) \, .$$

Regroupant ces résultats, on aboutit à la majoration :

$$\begin{aligned}\psi^{k+1} - \psi^k \leq & \underbrace{\varepsilon^k \langle g^k + (\vartheta^k)^\top \cdot p^k \, , u^\sharp - u^k \rangle}_{T_{5,1}} \\ & + \underbrace{\varepsilon^k \langle p^k - p^\sharp \, , \theta(u^k, w^{k+1}) - \Theta(u^\sharp) \rangle}_{T_{5,2}} \\ & + \frac{3}{2} \underbrace{(\varepsilon^k \rho^k) \left\| \theta(u^k, w^{k+1}) - \Theta(u^k) \right\|^2}_{T_{5,3}} \\ & + (\varepsilon^k \rho^k) \mathcal{L}(\psi^k) \, .\end{aligned} \tag{10.14}$$

Cette dernière inégalité s'écrit en termes de variables aléatoires (voir la Remarque 7.2). On prend alors, de part et d'autre de l'inégalité, l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^k$ engendrée par les k variables aléatoires $(\boldsymbol{W}^1, \dots, \boldsymbol{W}^k)$([4]).

– On considère d'abord le terme $\boldsymbol{T}_{5,1}$:

$$\boldsymbol{T}_{5,1} = \varepsilon^k \big\langle \boldsymbol{G}^k + (\boldsymbol{\vartheta}^k)^\top \cdot \boldsymbol{P}^k \, , u^\sharp - \boldsymbol{U}^k \big\rangle \, .$$

Prenant l'espérance conditionnelle par rapport à $\mathcal{F}^k$, on obtient :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{T}_{5,1} \;\big|\; \mathcal{F}^k\big) &= \varepsilon^k \big\langle \nabla J(\boldsymbol{U}^k) + \big(\Theta'(\boldsymbol{U}^k)\big)^\top \cdot \boldsymbol{P}^k \, , u^\sharp - \boldsymbol{U}^k \big\rangle \\ &\leq \varepsilon^k \Big(J(u^\sharp) - J(\boldsymbol{U}^k) + \big\langle \boldsymbol{P}^k \, , \Theta(u^\sharp) - \Theta(\boldsymbol{U}^k) \big\rangle \Big) \, , \end{aligned}$$

la dernière inégalité provenant, d'une part de la convexité de J, et d'autre part de la C-convexité de Θ.

– On considère ensuite le terme $\boldsymbol{T}_{5,2}$:

$$\boldsymbol{T}_{5,2} = \varepsilon^k \big\langle \boldsymbol{P}^k - p^\sharp \, , \theta(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) - \Theta(u^\sharp) \big\rangle \, .$$

Prenant l'espérance conditionnelle par rapport à $\mathcal{F}^k$, on obtient :

$$\mathbb{E}\big(\boldsymbol{T}_{5,2} \;\big|\; \mathcal{F}^k\big) = \varepsilon^k \big\langle \boldsymbol{P}^k - p^\sharp \, , \Theta(\boldsymbol{U}^k) - \Theta(u^\sharp) \big\rangle \, .$$

– On considère enfin le terme $\boldsymbol{T}_{5,3}$:

$$\boldsymbol{T}_{5,3} = \big(\varepsilon^k \rho^k\big) \big\| \theta(\boldsymbol{U}^k, \boldsymbol{W}^{k+1}) - \Theta(\boldsymbol{U}^k) \big\|^2 \, .$$

L'espérance conditionnelle de $\boldsymbol{T}_{5,3}$ par rapport à $\mathcal{F}^k$ se réduit en fait à une simple espérance par rapport à $\boldsymbol{W}^{k+1}$. Utilisant l'hypothèse 9 et la relation (10.8a), on obtient :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{T}_{5,3} \;\big|\; \mathcal{F}^k\big) &\leq \big(\varepsilon^k \rho^k\big) \big(\gamma \big\| \boldsymbol{U}^k \big\|^2 + \delta\big) \\ &\leq \big(\varepsilon^k \rho^k\big) \mathcal{L}\big(\boldsymbol{\Psi}^k\big) \, . \end{aligned}$$

Prenant l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^k$ dans la relation (10.14) et y reportant les majorations des termes $\boldsymbol{T}_{5,1}$, $\boldsymbol{T}_{5,2}$ et $\boldsymbol{T}_{5,3}$, on obtient :

$$\mathbb{E}\big(\boldsymbol{\Psi}^{k+1} \;\big|\; \mathcal{F}^k\big) - \boldsymbol{\Psi}^k \leq \big(\varepsilon^k \rho^k\big) \mathcal{L}\big(\boldsymbol{\Psi}^k\big) + \varepsilon^k \big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)\big) \, ,$$

4. On rappelle que les variables aléatoires $\boldsymbol{U}^k$, $\boldsymbol{P}^k$ et donc $\boldsymbol{\Psi}^k$ sont par construction mesurables par rapport à la tribu $\mathcal{F}^k$, et que les variables aléatoires $\boldsymbol{G}^k$ et $(\boldsymbol{\vartheta}^k)^\top$ dépendent de $\boldsymbol{W}^{k+1}$. De plus, la variable aléatoire $\boldsymbol{W}^{k+1}$ est indépendante des $\boldsymbol{W}^l$ précédentes.

où L est le Lagrangien associé au problème (10.1). On en déduit l'existence de constantes c_1 et c_2 telles que :

$$\mathbb{E}\big(\boldsymbol{\Psi}^{k+1} \;\big|\; \mathcal{F}^k\big) - \boldsymbol{\Psi}^k \leq \big(\varepsilon^k \rho^k\big)\big(c_1 \boldsymbol{\Psi}^k + c_2\big) + \varepsilon^k \big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)\big) \; . \quad (10.15)$$

Comme noté à la Remarque 10.4, la série de terme général $\big(\varepsilon^k \rho^k\big)$ est convergente. Le terme $L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^k, p^\sharp)$ est quant à lui négatif ou nul.

3. **Analyse de convergence.** Par le Théorème 8.9, on obtient que la suite de variables aléatoires $\big\{\boldsymbol{\Psi}^k\big\}_{k\in\mathbb{N}}$ converge presque sûrement vers une variable aléatoire bornée presque sûrement, et que l'on a :

$$\sum_{k=0}^{+\infty} \varepsilon^k \big(L(\boldsymbol{U}^k, p^\sharp) - L(u^\sharp, p^\sharp)\big) < +\infty \; , \quad \mathbb{P}\text{-p.s.} \; . \quad (10.16)$$

4. **Limites des suites.** Du fait que la suite $\big\{\boldsymbol{\Psi}^k\big\}_{k\in\mathbb{N}}$ est bornée presque sûrement, on déduit de (10.8) que les suites $\big\{\boldsymbol{U}^k\big\}_{k\in\mathbb{N}}$ et $\big\{\boldsymbol{P}^k\big\}_{k\in\mathbb{N}}$ sont elles aussi bornées presque sûrement. Les hypothèses du Lemme 8.10 étant satisfaites, on déduit de (10.16) que la suite $\big\{L(\boldsymbol{U}^k, p^\sharp)\big\}_{k\in\mathbb{N}}$ converge presque sûrement vers $L(u^\sharp, p^\sharp)$.
On note alors Ω_0 le sous-ensemble (de mesure nulle) de Ω sur lequel la suite $\big\{\boldsymbol{\Psi}^k\big\}_{k\in\mathbb{N}}$ n'est pas bornée, et Ω_1 le sous-ensemble (de mesure nulle lui aussi) de Ω sur lequel la relation (10.16) n'est pas vérifiée.
Soit $\omega \notin \Omega_0 \cup \Omega_1$. La suite des réalisations $\big\{u^k\big\}_{k\in\mathbb{N}}$ associée à cet élément ω est bornée et chaque u^k appartient à U^{ad}, partie fermée de $\mathcal{U}$. Par un argument de compacité, on conclut que l'on peut extraire de la suite $\big\{u^k\big\}_{k\in\mathbb{N}}$ une sous-suite convergente $\big\{u^{\Phi(k)}\big\}_{k\in\mathbb{N}}$. Soit $\bar{u}$ la limite de la suite $\big\{u^{\Phi(k)}\big\}_{k\in\mathbb{N}}$. La semi-continuité inférieure du Lagrangien L implique :

$$L(\bar{u}, p^\sharp) \leq \liminf_{k\to+\infty} L(u^{\Phi(k)}, p^\sharp) = L(u^\sharp, p^\sharp) \; .$$

On en déduit que, presque sûrement, $\overline{u}$ est solution du problème de minimisation sur U^{ad} du Lagrangien L pour $p = p^\sharp$ fixé. La stabilité en u du Lagrangien implique alors que $\overline{u} \in U^\sharp$. □

Remarque 10.6. Si on fait l'hypothèse supplémentaire de C-convexité sur la fonction $u \mapsto \theta(u, w)$, on peut remplacer dans le problème auxilaire (10.3a) le terme linéaire $\langle p^k \, , \vartheta^k \cdot u \rangle$ par le terme $\langle p^k \, , \theta(u, w^{k+1}) \rangle$, ce qui conduit à la phase de minimisation en u suivante :

$$u^{k+1} \in \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}} K(u) + \big\langle \varepsilon^k g^k - \nabla K(u^k) \, , u \big\rangle + \varepsilon^k \big\langle p^k \, , \theta(u, w^{k+1}) \big\rangle \; ,$$

ce nouveau problème étant fortement convexe et ayant donc une solution unique. La preuve de convergence précédente s'adapte alors facilement à cette variante.

10.2 Perspectives

On a donc proposé un algorithme permettant de traiter « à la Monte Carlo » les problèmes d'optimisation en boucle ouverte sous contraintes en espérance, et on a prouvé sa convergence. Comme on l'a déjà remarqué, bien que de telles contraintes ne soient pas naturelles dans le contexte de l'optimisation, elles permettent de prendre en compte les *contraintes en probabilité* par la transformation :

$$\mathbb{P}\big(\theta(u,\boldsymbol{W}) \in -C\big) = \mathbb{E}\big(\mathbf{1}_{\{\theta(u,\boldsymbol{W})\in -C\}}\big)\ . \tag{10.17}$$

Pour rendre opérationnelle cette remarque, il faut encore se pencher sur les deux points suivants.

1. Il ne suffit pas de faire des hypothèses de convexité sur la fonction θ pour que la contrainte en probabilité $\mathbb{P}\big(\theta(u,\boldsymbol{W}) \in -C\big)$ induise un ensemble convexe dans l'espace $\mathcal{U}$. Les propriétés de connexité, convexité et de différentiabilité des contraintes en probabilité ont été étudiées par de nombreux auteurs (voir par exemple Prekopa (1995), Henrion (2002), Henrion et Strugarek (2008) et (Shapiro et collab., 2009, Chapter 4) pour plus de détails). On se donne plus de chances d'assurer l'existence d'un point selle, au moins localement, en utilisant un Lagrangien augmenté (voir le Chapitre 5).
2. Une autre difficulté vient de ce que la transformation dans (10.17) fait intervenir sous l'espérance la fonction indicatrice d'un ensemble, et que cette fonction n'a pas de bonnes propriétés de continuité ni de différentiabilité. Une manière de passer cette difficulté consiste à transformer la fonction indicatrice en la convolant avec une fonction régulière (c'est la méthode des « mollifiers » proposée dans Ermoliev et collab. (1995)) et de récupérer ainsi la (sous-)différentiabilité nécessaire à un algorithme de type gradient.

Le premier point a été abordé dans (Strugarek, 2006, Chapitre VI). La difficulté pratique vient de ce qu'il est alors nécessaire de disposer d'un algorithme de type gradient stochastique pouvant s'accommoder d'une *fonction non linéaire* de l'espérance de la contrainte, alors que les méthodes de type Monte Carlo cherchent à reconstituer l'espérance proprement dite. Partant d'une contrainte $\mathbb{E}\big(\theta(u,\boldsymbol{W})\big) \in -C$, et notant f la fonction non linéaire de l'espérance apparaissant dans le Lagrangien augmenté associé[5], la contrainte introduit dans la fonction objectif un terme de la forme $f\big(\mathbb{E}\big(\theta(u,\boldsymbol{W})\big)\big)$, dont le gradient est donné par l'expression :

$$\mathbb{E}\big(\theta'_u(u,\boldsymbol{W})\big)^\top \cdot \nabla f\big(\mathbb{E}\big(\theta(u,\boldsymbol{W})\big)\big)\ .$$

5. Voir les relations (9.25) définissant le Lagrangien augmenté, la fonction ζ_c et ses gradients, avec dans le cas qui nous intéresse $\Theta(u) = \mathbb{E}\big(\theta(u,\boldsymbol{W})\big)$.

Ce gradient n'est pas une espérance, mais en comporte deux. On ne peut donc pas appliquer directement un algorithme de gradient stochastique. Cependant, dans le cas de contraintes égalité[6] :

$$\mathbb{E}\big(\theta(u, \boldsymbol{W})\big) = 0 \,,$$

le Lagrangien augmenté prend la forme simple suivante :

$$L_c(u, p) = J(u) + \big\langle p \,, \mathbb{E}\big(\theta(u, \boldsymbol{W})\big)\big\rangle + \frac{c}{2}\big\|\mathbb{E}\big(\theta(u, \boldsymbol{W})\big)\big\|^2 \,.$$

La fonction f considérée ci-dessus est alors la fonction $v \mapsto \|v\|^2/2$, et le gradient du terme quadratique provenant de f est :

$$\mathbb{E}\big(\theta'_u(u, \boldsymbol{W})\big)^\top \cdot \mathbb{E}\big(\theta(u, \boldsymbol{W})\big) \,.$$

Ce produit d'espérance peut toujours s'écrire comme l'espérance d'un produit, à savoir :

$$\mathbb{E}\Big(\big(\theta'_u(u, \boldsymbol{W}_1)\big)^\top \cdot \theta(u, \boldsymbol{W}_2)\Big) \,,$$

où $\boldsymbol{W}_1$ et $\boldsymbol{W}_2$ sont deux variables aléatoires indépendantes et identiquement distribuées, de même loi que $\boldsymbol{W}$. On peut alors proposer l'algorithme de gradient stochastique suivant, basé sur le Lagrangien augmenté :

$$\begin{aligned} u^{k+1} \in \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} & K(u) + \big\langle \varepsilon^k \nabla_u j(u^k, w^{k+1}) - \nabla K(u^k) \,, u \big\rangle \\ & + \varepsilon^k \big\langle p^k + c\theta(u^k, w_2^{k+1}) \,, \theta'_u(u^k, w_1^{k+1}) \cdot u \big\rangle \,, \\ p^{k+1} = \mathrm{proj}_{C^*} & \left(p^k + \rho^k \, \theta(u^{k+1}, w_2^{k+1})\right) \,. \end{aligned}$$

La présence de deux tirages w_1^{k+1} et w_2^{k+1} dans l'algorithme (alors qu'il n'y en a qu'un seul dans l'algorithme basé sur le Lagrangien simple) peut provoquer une plus grande variance asymptotique et donc ralentir la convergence de l'algorithme.

Le second point mentionné plus haut et relatif à la discontinuité de la fonction indicatrice a été traité dans Andrieu et collab. (2011). On choisit de le présenter ici le cas d'une contrainte en probabilité scalaire :

$$\mathbb{P}\big(\theta(u, \boldsymbol{W}) \leq \alpha\big) \geq \pi \,,$$

qui, par la transformation (10.17), conduit à considérer la fonction :

$$\Theta(u) = \mathbb{E}\big(\mathbf{1}_{\mathbb{R}^+}\big(\alpha - \theta(u, \boldsymbol{W})\big)\big) \,.$$

La méthode des mollifiers consiste à choisir une fonction $h : \mathbb{R} \to \mathbb{R}$ régulière, positive ($h(u) \geq 0$), symétrique ($h(u) = h(-u)$), ayant un maximum unique en $u = 0$ et telle que :

6. Pour le cas des contraintes générales $\mathbb{E}\big(\theta(u, \boldsymbol{W})\big) \in -C$, on consultera (Strugarek, 2006, Chapitre VI).

$$\int_{-\infty}^{+\infty} h(u)\,\mathrm{d}u = 1\ .$$

Étant donné une fonction réelle $\phi : \mathbb{R} \to \mathbb{R}$, on se donne un paramètre réel r positif et on considère le produit de convolution :

$$\phi_r(u) = \frac{1}{r}\int_{-\infty}^{+\infty} \phi(v)\, h\Big(\frac{u-v}{r}\Big)\mathrm{d}v\ .$$

La fonction ϕ_r peut être vue comme une approximation de la fonction ϕ, car la la fonction $h(\cdot/r)/r$ converge (au sens des distributions) vers le Dirac quand r tend vers zéro. On applique alors cette méthode à la fonction $\mathbf{1}_{\mathbb{R}^+}$, ce qui conduit à la fonction contrainte « mollifiée »

$$\begin{aligned}\Theta_r(u) &= \frac{1}{r}\mathbb{E}\Big(\int_{-\infty}^{+\infty} \mathbf{1}_{\mathbb{R}^+}(v)\, h\Big(\frac{\alpha-\theta(u,\boldsymbol{W})-v}{r}\Big)\mathrm{d}v\Big)\\ &= \frac{1}{r}\mathbb{E}\Big(\int_{0}^{+\infty} h\Big(\frac{v-\alpha+\theta(u,\boldsymbol{W})}{r}\Big)\mathrm{d}v\Big)\ .\end{aligned}$$

Notant $\mathbf{I}_r(u,w)$ la fonction définie par :

$$\mathbf{I}_r(u,w) = \frac{1}{r}\int_{0}^{+\infty} h\Big(\frac{v-\alpha+\theta(u,w)}{r}\Big)\mathrm{d}v\ ,$$

on a donc :

$$\Theta_r(u) = \mathbb{E}\big(\mathbf{I}_r(u,\boldsymbol{W})\big) \quad , \quad \nabla\Theta_r(u) = \mathbb{E}\big(\nabla_u\mathbf{I}_r(u,\boldsymbol{W})\big)\ ,$$

et un calcul simple montre que l'on a :

$$\nabla_u\mathbf{I}_r(u,w) = \frac{1}{r}h\Big(\frac{\theta(u,w)-\alpha}{r}\Big)\nabla_u\theta(u,w)\ ,$$

cette dernière expression ne faisant plus intervenir de calcul d'intégrale. On a donc montré que $\nabla_u\mathbf{I}_r(u,\boldsymbol{W})$ était un estimateur sans biais de $\nabla\Theta_r(u)$, qui est quant à lui un estimateur *biaisé* de $\nabla\Theta(u)$. Cependant, ce biais disparaît lorsque le paramètre r tend vers zéro. Tout est donc en place pour utiliser un algorithme de type gradient stochastique en utilisant à l'itération k de l'algorithme l'expression $\nabla_u\mathbf{I}_r(u^k,w^{k+1})$ comme approximation du gradient de la contrainte sous l'espérance. Il reste encore à définir comment il convient de faire décroître le paramètre r au cours des itérations pour que l'algorithme fournisse la solution du problème initial. On montre qu'il est optimal de choisir une suite $\{r^k\}_{k\in\mathbb{N}}$ de paramètres de la forme :

$$r^k = \frac{a}{k^{1/5}}\ .$$

On consultera Andrieu et collab. (2011) pour plus de détails sur la méthode.

Littérature

Andrieu, L., G. Cohen et F. Vazquez-Abad. 2011, «Gradient-based simulation optimization under probability constraints», *European Journal of Operational Research*, vol. 212, n° 2, p. 345–351.

Arrow, K. et L. Hurwicz. 1960, «Decentralization and computation in resource allocation», dans *Essays in Economics and Econometrics*, édité par R. Pfouts, University of North Carolina Press, p. 34–104.

Bach, F. et E. Moulines. 2013, «Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n)», dans *Advances in Neural Information Processing Systems*, p. 773–781.

Bellman, R. E. 1957, *Dynamic programming*, Princeton University Press.

Bensoussan, A., J.-L. Lions et R. Temam. 1974, «Sur les méthodes de décomposition, de décentralisation et de coordination et applications», dans *Sur les méthodes numériques en sciences physiques et économiques*, édité par J.-L. Lions et G. Marchouk, Dunod, Paris.

Benveniste, A., M. Métivier et P. Priouret. 1990, *Adaptative Algorithms and Stochastic Approximation*, Springer Verlag.

Bertsekas, D. P. 1976, «Multiplier methods : A survey», *Automatica*, vol. 12, p. 133–145.

Bertsekas, D. P. 1996, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific.

Bonnans, F., J. Gilbert, C. Lemaréchal et C. Sagastizábal. 2006, *Numerical optimization : theoretical and practical aspects*, Springer-Verlag, Berlin. Second revised ed. of translation of 1997 French ed.

Bouleau, N. 1986, *Probabilités de l'ingénieur*, Hermann.

Bousquet, O. et L. Bottou. 2008, «The tradeoffs of large scale learning», dans *Advances in neural information processing systems*, p. 161–168.

Brosilow, C., L. Lasdon et J. Pearson. 1965, «Feasible optimization methods for interconnected systems», dans *Proceedings Joint Automatic Control Conference*, Troy, New-York.

Bryson, A. E. et Y. Ho. 1975, *Applied Optimal Control*, Taylor and Francis.

Carpentier, P., J.-P. Chancelier, G. Cohen et M. de Lara. 2015, *Stochastic Optimization : at the crossroads between discrete time stochastic control and stochastic programming*, Springer.

Carpentier, P., G. Cohen et J.-C. Culioli. 1995, ≪Stochastic optimal control and decomposition-coordination methods≫, dans *Recent Developments in Optimization*, vol. Lecture Notes in Economics and Mathematical Systems No. 429, édité par R. Durier et C. Michelot, Springer-Verlag.

Chen, H. F., L. Guo et A. J. Gao. 1988, ≪Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds≫, *Stoch. Proc. Appl.*, vol. 27, n° 2, p. 217–231.

Cohen, G. 1978, ≪Optimization by decomposition and coordination : a unified approach≫, *IEEE Transactions on Automatic Control*, vol. AC–23, p. 222–232.

Cohen, G. 1980, ≪Auxiliary problem principle and decomposition of optimization problems≫, *Journal of Optimization Theory and Applications*, vol. 32, n° 3, p. 277–305.

Cohen, G. 1984, *Décomposition et coordination en optimisation déterministe différentiable et non différentiable*, thèse de doctorat, Université de Paris IX-Dauphine. Thèse de Doctorat ès-Sciences Mathématiques.

Cohen, G. 1988, ≪Auxiliary problem principle extended to variational inequalities≫, *Journal of Optimization Theory and Applications*, vol. 59, p. 325–333.

Cohen, G. 2000, ≪Convexité et optimisation≫, Unpublished lecture notes, ENPC. URL http://cermics.enpc.fr/~cohen-g/documents/Ponts-cours-A4-NB.pdf.

Cohen, G. et B. Miara. 1990, ≪Optimization with an auxiliary constraint and decomposition≫, *SIAM Journal on Control and Optimization*, vol. 28, p. 137–157.

Cohen, G. et D. Zhu. 1984, ≪Decomposition coordination methods in large scale optimization problems≫, dans *Advances in large scale systems theory and applications*, vol. 1, édité par J. Cruz, JAI Press, Greeenwich, Connecticut, p. 203–266.

Culioli, J.-C. 1987, *Algorithmes de décomposition/coordination en optimisation stochastique*, thèse de doctorat, École Nationale Supérieure des Mines de Paris.

Culioli, J.-C. et G. Cohen. 1990, ≪Decomposition-coordination algorithms in stochastic optimization≫, *SIAM Journal on Control and Optimization*, vol. 28, n° 6, p. 1372–1403.

Dacunha-Castelle, D. et M. Duflo. 1994, *Probabilités et statistiques. Tome 1 : problèmes à temps fixe*, Masson.

Dantzig, G. et P. Wolfe. 1961, ≪The decomposition algorithm for linear program≫, *Econometrica*, vol. 29, p. 767–778.

Delebecque, F. et J.-P. Quadrat. 1978, ≪Contribution of stochastic control singular perturbations averaging and team≫, *IEEE Transaction on Automatic Control*.

Delyon, B. 2000, ≪Stochastic approximation with decreasing gain : convergence and asymptotic theory≫, *Unpublished lecture notes, Université de Rennes*.

Dodu, J.-C., M. Goursat, A. Hertz, J.-P. Quadrat et M. Viot. 1981, ≪Méthodes de gradient stochastique pour l'optimisation des investissements≫, *Bulletin de la Direction des Études et Recherches EDF*, vol. C, n° 2.

Duflo, M. 1996, *Algorithmes stochastiques*, Springer Verlag.

Duflo, M. 1997, *Random Iterative Models*, Springer Verlag.

Dunn, J. 1976, ≪Convexity, monotonicity, and gradient process in hilbert space≫, *Journal of Mathematical Analysis and Applications*, vol. 53, p. 145–158.

Ekeland, I. et R. Temam. 1999, *Convex analysis and variational problems*, SIAM, Philadelphia.

El Farouq, N. 1993, *Algorithmes de résolution d'inéquations variationnelles*, thèse de doctorat, École Nationale Supérieure des Mines de Paris.

El Farouq, N. et G. Cohen. 1998, ≪Progressive regularization of variational inequalities and decomposition algorithms≫, *Journal of Optimization Theory and Applications*, vol. 97, n° 2, p. 407–433.

Ermoliev, Y., V. Norkin et R.-B. Wets. 1995, ≪The minimization of semicontinuous functions : Mollifier subgradients≫, *SIAM Journal on Control and Optimization*, vol. 33, n° 1, p. 149–167.

Findeisen, W., F. Bailey, N. Brdys, K. Malinowski et P. T. A. Wozniak. 1980, *Control and Coordination in Hierarchical Systems*, Wiley, New-York.

Hardy, G. H. 2000, *Divergent series*, vol. 334, American Mathematical Society.

Henrion, R. 2002, ≪On the connectedness of probabilistic constraint sets≫, *Journal of Optimization Theory and Applications*, vol. 112, n° 3, p. 657–663.

Henrion, R. et C. Strugarek. 2008, ≪Convexity of chance constraints with independent random variables≫, *Computational Optimization and Applications*, vol. 41, n° 2, p. 263–276.

Hess, C. 1995, ≪On the measurability of the conjugate and the subdifferential of a normal integrand≫, *Journal of Convex Analysis*, vol. 2, p. 153–165.

Hestenes, M. 1969, ≪Multiplier and gradient methods≫, *Journal of Optimization Theory and Applications*, vol. 4, p. 303320.

Kay, S. M. 1993, *Fundamentals of Statistical Signal Processing : Estimation Theory*, Prenctice Hall.

Kesten, H. 1958, ≪Accelerated stochastic approximation≫, *Annals Math. Statist.*, vol. 29, p. 41–59.

Khalil, H. K. 2002, *Nonlinear Systems*, 3e éd., Prentice-Hall.

Kiefer, J. et J. Wolfowitz. 1952, ≪Stochastic estimation of the maximum of a regression function≫, *Annals Math. Statist.*, vol. 23, p. 462–466.

Kushner, H. J. et D. S. Clark. 1978, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer Verlag.

Kushner, H. J. et G. G. Yin. 2003, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer Verlag.

Lai, T. L. 2003, ≪Stochastic approximation≫, *The Annals of Statistics*, vol. 31, n° 2, p. 391–406.

Lasdon, L. 1970, *Optimization Theory for Large Systems*, Mac Millan, Toronto.

Lasdon, L. et J. Schoeffler. 1965, ≪A multilevel technique for optimization≫, dans *Proceedings Joint Automatic Control Conference*, Troy, New-York.

Marschak, J. et R. Radner. 1971, *The Economic Theory of Teams*, Yale University Press, New Haven, Connecticut.

Mataoui, M. A. 1990, *Contribution à la décomposition et à l'agrégation des problèmes variationnels*, thèse de doctorat, École Nationale Supérieure des Mines de Paris.

Mesarovic, M., D. Macko et Y. Takahara. 1970, *Theory of Hierarchical Multilevel Systems*, Academic Press, New-York.

Moreau, J.-J. 1962, ≪Fonctions convexes duales et points proximaux dans un espace hilbertien≫, *Compte Rendus de l'Académie des Sciences, Paris*, vol. Série A, 255, p. 28972899.

Moreau, J.-J. 1965, ≪Proximité et dualité dans un espace hilbertien≫, *Bulletin de la Société Mathématique de France*, vol. 93, n° 2, p. 273299.

Nemirovski, A., A. Juditsky, G. Lan et A. Shapiro. 2009, ≪Robust stochastic approximation approach to stochastic programming≫, *SIAM Journal on Optimization*, vol. 19, n° 4, p. 1574–1609.

Parikh, N. et S. Boyd. 2013, ≪Proximal algorithms≫, *Foundations and Trends in Optimization*, vol. 1, n° 3, p. 123231.

Plakhov, A. et P. Cruz. 2005, ≪A stochastic approximation algorithm with multiplicative step size adaptation≫, *arXiv, math.ST/0503434*.

Polyak, B. T. 1976, ≪Convergence and convergence rate of iterative stochastic algorithms≫, *Automation and Remote Control*, vol. 37, n° 12, p. 1858–1868. Translation : Avtomatica i Telemekhanika, No. 12, pp. 83-94, 1976.

Polyak, B. T. 1990, ≪New method of stochastic approximation type≫, *Automation and Remote Control*, vol. 51, n° 7, p. 937–946. Translation : Avtomatica i Telemekhanika, No. 7, pp. 98-107, 1990.

Polyak, B. T. et A. B. Juditsky. 1992, ≪Acceleration of stochastic approximation by averaging≫, *SIAM Journal on Control and Optimization*, vol. 30, n° 4, p. 838–855.

Polyak, B. T. et Y. Z. Tsypkin. 1979, ≪Adaptative estimation algorithms (convergence, optimality, stability)≫, *Automation and Remote Control*, vol. 40, n° 3, p. 378–389. Translation : Avtomatica i Telemekhanika, No. 3, pp. 71-84, 1979.

Powell, M. 1969, ≪A method for nonlinear constraints in minimization problems≫, dans *Optimization*, édité par R. Fletcher, Academic Press, New York, NY, p. 283298.

Prekopa, A. 1995, *Stochastic Programming*, Kluwer, Dortrecht.

Puterman, M. L. 2009, *Markov decision processes : discrete stochastic dynamic programming*, vol. 414, John Wiley & Sons.

Quadrat, J.-P. et M. Viot. 2000, ≪Introduction à la commande stochastique - Cours du DEA MMME de l'Université Paris I≫, http://www-rocq.inria.fr/metalau/quadrat/ComSto0.9.pdf.

Renaud, A. 1993, *Algorithmes de régularisation et de décomposition pour les problèmes variationnels monotones*, thèse de doctorat, École Nationale Supérieure des Mines de Paris.

Robbins, H. et S. Monro. 1951, ≪A stochastic approximation method≫, *Annals Math. Statist.*, vol. 22, p. 400–407.

Rockafellar, R. 1970, *Convex Analysis*, Princeton University Press, Princeton.

Rockafellar, R. 1973, ≪The multiplier method of Hestenes and Powell applied to convex programming≫, *Journal of Optimization Theory and Applications*, vol. 12, p. 555–562.

Rockafellar, R. et R.-B. Wets. 1998, *Variational Analysis*, Springer Verlag, Berlin.

Schwartz, L. 1993, *Analyse IV. Applications à la théorie de la mesure*, Hermann, Paris.

Shapiro, A., D. Dentcheva et A. Ruszczyński. 2009, *Lectures on Stochastic Programming : Modeling and Theory*, SIAM, Philadelphia.

Strugarek, C. 2006, *Approches variationnelles et autres contributions en optimisation stochastique*, Thèse de doctorat http://cermics.enpc.fr/theses/, École Nationale des Ponts et Chaussées.

Vazquez-Abad, F. 2006, ≪Stochastic optimisation≫, Unpublished lecture notes, University of Melbourne.

von Neumann, J. et O. Morgenstern. 1953, *Theory of Games and Economic Behavior*, Princeton University Press. 3rd edition.

Wismer, D., éd.. 1971, *Optimization Methods for Large Scale Systems with Applications*, Mac Graw-Hill.

Witsenhausen, H. S. 1968, ≪A counterexample in stochastic optimum control≫, *SIAM Journal on Control*, vol. 6, p. 131–147.

Yosida, K. 1964, *Functional analysis*, Springer Verlag, Berlin.

Zhu, D. 1982, *Optimisation sous-différentiable et méthodes de décomposition*, thèse de doctorat, École Nationale Supérieure des Mines de Paris.

Index

algorithme
  d'Arrow-Hurwicz
    stochastique, 209
  d'Uzawa
    stochastique, 205
  de gradient
    stochastique, *voir* gradient stochastique
  de Newton
    stochastique, 178
  Newton-efficace, 179
approximation stochastique, 161, 172

borne
  Cramer-Rao, 183

contraintes
  en espérance, 227, 229
  en probabilité, 227, 239
  presque sûres, 226
convergence
  en moyenne quadratique, 168
  presque sûre, 167
Cramer Rao, *voir* borne

échantillon, 163

fonction
  indicatrice, 227

GLB, *voir* gradient linéairement borné
gradient
  linéairement borné, 192
gradient stochastique, 149, 161

  algorithme, 163
    moyenné, 180
    PPA, 190
  approche robuste, 186

intégrande, 201
  normale, 192, 201

Lagrangien
  augmenté
    cas stochastique, 220
    contrainte en espérance, 239
Limite centrale (théorème de la), 175
Lyapunov (équation de), 175

méthode
  de l'équation différentielle ordinaire, 205
  des mollifiers, 239
  Monte Carlo, 162
mesurable
  multi-application, 200
  sélection, 192, 200
mollifiers, *voir* méthode
Monte Carlo, *voir* méthode
moyennisation, 179

Newton-eficace, *voir* algorithme

optimisation stochastique
  boucle fermée, 153
  boucle ouverte, 148

Robbins-Monro (théorème de), 173

Robbins-Siegmund (théorème de), 198

sélection mesurable, *voir* mesurable
SA, *voir* approximation stochastique
SAA, *voir* Sample Average Approximation
Sample Average Approximation, 163
σ-suite, 168
$\sigma(\alpha, \beta, \gamma)$-suite, 174