5

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

$$\min_{u \in U^{\mathrm{ad}}} J(u) \qquad \text{sous la contrainte} \qquad \Theta(u) \in -C \,. \tag{5.1}$$

Comme au §4, on suppose que la fonction J est l'espérance d'une fonction j définie sur $\mathbb{U} \times \mathbb{W}$ à valeurs dans $\overline{\mathbb{R}}$, supposée intégrable pour tout $u \in U^{\mathrm{ad}}$:

$$J(u) = \mathbb{E}\left(j(u, \boldsymbol{W})\right) \,. \tag{5.2a}$$

De même, on suppose que la fonction Θ représente l'espérance d'une fonction θ définie sur $\mathbb{U} \times \mathbb{W}$ à valeurs dans $\mathbb{V}$, intégrable pour tout $u \in U^{\mathrm{ad}}$:

$$\Theta(u) = \mathbb{E}\left(\theta(u, \boldsymbol{W})\right) \,. \tag{5.2b}$$

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 (5.2b) et utiliser dans un algorithme de type Arrow-Hurwicz stochastique des évaluations de la fonction θ plutôt que de son espérance Θ.

5.1 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 (4.9), 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 remplace alors la résolution du problème (5.1) par la résolution de la suite de problèmes auxiliaires :

$$u^{(k+1)} \in \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} g^{(k)} - \nabla K(u^{(k)}) \, , u \right\rangle + \epsilon^{(k)} \left\langle p^{(k)} \, , \vartheta^{(k)} \cdot u \right\rangle , \tag{5.3a}$$

$$p^{(k+1)} = \mathrm{proj}_{C^\star} \left(p^{(k)} + \rho^{(k)} \, \theta(u^{(k+1)}, w^{(k+1)}) \right) , \tag{5.3b}$$

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 $\vartheta^{(k)}$.

L'algorithme qui en découle est le suivant.

Algorithme 5.1. (PPA stochastique et contraintes en espérance)

1. Choisir $(u^{(0)}, p^{(0)}) \in U^{\mathrm{ad}} \times C^\star$, et deux suites $\left\{\epsilon^{(k)}\right\}_{k\in\mathbb{N}}$ 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 (5.3a).
4. Calculer $p^{(k+1)}$ par la formule de mise à jour (5.3b).
5. Incrémenter l'indice k de 1 et retourner à l'étape 2.

Remarque 5.2. Avec le choix de noyau $K(u) = \|u\|^2/2$, le problème auxiliaire (5.3) se met sous la forme :

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

Pour étudier la convergence de l'algorithme 5.1, on le formule en terme de variables aléatoires 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 prend alors la forme :

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \epsilon^{(k)} \boldsymbol{G}^{(k)} - \nabla K(\boldsymbol{U}^{(k)}) \, , u \right\rangle + \epsilon^{(k)} \left\langle \boldsymbol{P}^{(k)} \, , \boldsymbol{\vartheta}^{(k)} \cdot u \right\rangle , \tag{5.4a}$$

$$\boldsymbol{P}^{(k+1)} = \mathrm{proj}_{C^\star} \left(\boldsymbol{P}^{(k)} + \rho^{(k)} \, \theta(\boldsymbol{U}^{(k+1)}, \boldsymbol{W}^{(k+1)}) \right) . \tag{5.4b}$$

La minimisation dans (5.4a) et la projection dans (5.4b) sont effectuées ω par ω. Le résultat de la minimisation (5.4a) dépend de ω et est noté $\boldsymbol{U}^{(k+1)}$.

Théorème 5.3.
On suppose que les hypothèses suivantes sont vérifiées.

1. *U^{ad} est une partie convexe fermée non vide d'un l'espace de Hilbert $\mathbb{U}$, et C est un cône convexe fermé saillant d'un autre espace de Hilbert $\mathbb{V}$.*
2. *La fonction $j : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ 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 $w \in \mathbb{W}$.*
4. *La fonction $j(\cdot, w)$ est à gradient linéairement borné uniformément en w :*
$$\exists c_1 > 0,\ \exists c_2 > 0,\ \forall w \in \mathbb{W},\ \forall u \in U^{\mathrm{ad}}\ ,\quad \|\nabla_u j(u, w)\| \leq c_1 \|u\| + c_2\ .$$
5. *La fonction J est convexe, Lipschitzienne, coercive sur l'ensemble U^{ad}.*
6. *La fonction $\theta : \mathbb{U} \times \mathbb{W} \to \mathbb{V}$ est telle que l'application $(u, w) \mapsto \langle p, \theta(u, w)\rangle$ est une intégrande normale pour tout $p \in C^\star$, 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 :*
$$\exists \lambda > 0,\ \exists \mu > 0,\ \forall w \in \mathbb{W},\ \forall u, v \in U^{\mathrm{ad}}\ ,\quad \|\theta(u, w) - \theta(v, w)\| \leq \lambda \|u - v\| + \mu\ .$$
8. *Pour tout $w \in \mathbb{W}$, la fonction θ est différentiable, et sa différentielle par rapport à u est bornée par une constante ϱ, uniformément en w.*
9. *La variance associée à la fonction de contrainte θ est bornée par une fonction quadratique :*
$$\exists \gamma > 0,\ \exists \delta > 0,\ \forall u \in U^{\mathrm{ad}}\ ,\quad \mathbb{E}\big(\big\|\theta(u, \boldsymbol{W}) - \Theta(u)\big\|^2\big) \leq \gamma \|u\|^2 + \delta\ .$$
10. *La fonction Θ est C-convexe, Lipschitzienne de rapport L_Θ.*
11. *Les contraintes sont qualifiées et le Lagrangien L est stable.*
12. *La fonction K est propre, fortement convexe de module b, semi-continue inférieurement, et elle est différentiable sur un sous-ensemble ouvert contenant U^{ad}.*
13. *Les deux suites $\big\{\epsilon^{(k)}\big\}_{k\in\mathbb{N}}$ et $\big\{\rho^{(k)}\big\}_{k\in\mathbb{N}}$ sont des σ-suites.*
14. *La suite quotient $\big\{\epsilon^{(k)}/\rho^{(k)}\big\}_{k\in\mathbb{N}}$ est décroissante.*

On a alors les conclusions suivantes.

1. *Le problème (5.1) admet un ensemble de points selle $U^\sharp \times P^\sharp$ non vide.*
2. *Le problème (5.4a) admet une solution $\boldsymbol{U}^{(k+1)}$ unique.*
3. *Pour tout $p^\sharp \in P^\sharp$, la suite de variables aléatoires $\big\{L(\boldsymbol{U}^{(k)}, p^\sharp)\big\}_{k\in\mathbb{N}}$ 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 5.1 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* $U^\sharp$*, ensemble des solutions du problème.*

Remarque 5.4. L'hypothèse 13 implique que la série de terme général $\epsilon^{(k)}\rho^{(k)}$ est convergente[1]. L'hypothèse 14 implique l'existence d'un réel positif α tel que $\epsilon^{(k)} \leq \alpha\rho^{(k)}$ pour tout $k \in \mathbb{N}$.

Remarque 5.5. Comme on l'a déjà noté dans la remarque 3.4 suivant le théorème de convergence du gradient stochastique généralisé sans contrainte explicite, 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 (5.3a) dans lequel on remplacerait le terme linéarisé $\left\langle p^{(k)}\,, \vartheta^{(k)} \cdot u\right\rangle$ par le terme $\left\langle p^{(k)}\,, \theta(u, w^{(k+1)})\right\rangle$, car on ne saurait alors rien dire sur la convexité du problème auxiliaire de minimisation en résultant.

Pour simplifier les écritures dans la preuve de ce théorème, on introduit la définition et la notation suivantes.

Définition 5.6. *Soit* $x = \{x^{(k)}\}_{k\in\mathbb{N}}$ *et* $y = \{y^{(k)}\}_{k\in\mathbb{N}}$ *deux suites de réels positifs. On dit que la suite* y *est bornée de manière affine par la suite* x*, relation que l'on note :*

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

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

$$y^{(k)} \leq a\, x^{(k)} + b\,.$$

De cette définition, on déduit les propriétés élémentaires suivantes.

Proposition 5.7. *Soit* $\{x^{(k)}\}_{k\in\mathbb{N}}$*,* $\{y^{(k)}\}_{k\in\mathbb{N}}$ *et* $\{z^{(k)}\}_{k\in\mathbb{N}}$ *trois suites de réels positifs. Alors,*

1. $y^{(k)} \leq \mathcal{L}(x^{(k)}) \;\Rightarrow\; y^{(k)} \leq \mathcal{L}(x^{(k)}) + \alpha$ *pour toute constante* $\alpha \geq 0$*,*
2. $y^{(k)} \leq \mathcal{L}(x^{(k)}) \;\Rightarrow\; (y^{(k)})^2 \leq \mathcal{L}((x^{(k)})^2)$*,*
3. $y^{(k)} \leq \mathcal{L}(x^{(k)})$ *et* $z^{(k)} \leq \mathcal{L}(x^{(k)}) \;\Rightarrow\; y^{(k)}z^{(k)} \leq \mathcal{L}((x^{(k)})^2)$*.*

Preuve. La preuve de la propriété 1 est évidente. La preuve des propriétés 2 et 3 s'appuie respectivement sur le fait que $(ax+b)^2 \leq 2a^2x^2 + 2b^2$ et que $(a_1x + b_1)(a_2x + b_2) \leq \big(\max\{a_1, a_2\}x + \max\{b_1, b_2\}\big)^2$.

Ces propriétés seront utilisées pour la démonstration du théorème 5.3, qui est donnée maintenant.

1. car on a $\epsilon^{(k)}\rho^{(k)} \leq \big((\epsilon^{(k)})^2 + (\rho^{(k)})^2\big)/2$

Preuve. La démonstration des 2 premières conclusions du théorème découle des théorèmes généraux relatifs à l'optimisation convexe sous contraintes. Le fait que la solution $\boldsymbol{U}^{(k+1)}$ du problème (5.4a) 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 3.3 pour plus de détails). La démonstration des 2 dernières conclusions se fait en suivant le schéma « habituel » de preuve.

Le fait que $u^{(k+1)}$ soit solution du problème (5.3a) est caractérisé pr la condition d'optimalité suivante :

$$\forall u \in U^{\mathrm{ad}},\ \big\langle \nabla K(u^{(k+1)}) - \nabla K(u^{(k)}) + \epsilon^{(k)}\big(g^{(k)} + (\vartheta^{(k)})^{\top}\cdot p^{(k)}\big)\,, u - u^{(k+1)}\big\rangle \geq 0\,. \quad (5.5)$$

1. **Choix de la fonction de Lyapunov**. Soit $(u^\sharp, p^\sharp) \in U^\sharp \times P^\sharp$ un point selle du problème (5.1). On choisit la fonction de Lyapunov de telle sorte que :

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

Par la définition de $\psi^{(k)}$ et la forte convexité du noyau K, on obtient les inégalités :

$$\big\|u^{(k)} - u^\sharp\big\|^2 \leq \mathcal{L}\big(\psi^{(k)}\big)\,, \quad (5.6a)$$

$$\big\|p^{(k)} - p^\sharp\big\|^2 \leq \frac{\rho^{(k)}}{\epsilon^{(k)}}\mathcal{L}\big(\psi^{(k)}\big)\,. \quad (5.6b)$$

On déduit alors de (5.6a) et de l'hypothèse GLB que :

$$\big\|g^{(k)}\big\|^2 \leq \mathcal{L}\big(\psi^{(k)}\big)\,. \quad (5.6c)$$

2. **Majorations**.
   a. On majore pour commencer la variation $\big\|u^{(k+1)} - u^{(k)}\big\|$. Évaluant la condition d'optimalité (5.5) au point $u = u^{(k)}$ et utilisant la forte convexité de K, on obtient :

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

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

$$\big\|u^{(k+1)} - u^{(k)}\big\| \leq \frac{\epsilon^{(k)}}{b}\big\|g^{(k)} + (\vartheta^{(k)})^{\top}\cdot p^{(k)}\big\|\,. \quad (5.7)$$

Par l'hypothèse 8 et la relation (5.6b), on obtient :

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

De cette dernière majoration et de la relation (5.6c), par l'inégalité triangulaire[2], on déduit :

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

la dernière inégalité provenant du second point de la remarque 5.4. On obtient finalement la majoration :

$$\left\|g^{(k)} + (\vartheta^{(k)})^\top \cdot p^{(k)}\right\|^2 \leq \frac{\rho^{(k)}}{\epsilon^{(k)}} \mathcal{L}\big(\psi^{(k)}\big) \; , \tag{5.8}$$

et donc, par (5.7) :

$$\left\|u^{(k+1)} - u^{(k)}\right\|^2 \leq \big(\epsilon^{(k)}\rho^{(k)}\big)\mathcal{L}\big(\psi^{(k)}\big) \; . \tag{5.9}$$

b. On majore ensuite la variation $\left\|p^{(k+1)} - p^\sharp\right\|^2$. Utilisant la relation $p^\sharp = \mathrm{proj}_{C^\star}\big(p^\sharp + \rho^{(k)}\Theta(u^\sharp)\big)$, vraie pour tout $\rho^{(k)} > 0$, ainsi que la relation (5.3b) définissant $p^{(k+1)}$, et comme l'opérateur de projection sur $C^\star$ est contractant, 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 :

$$\begin{aligned}\left\|p^{(k+1)} - p^\sharp\right\|^2 \leq \; & \left\|p^{(k)} - p^\sharp\right\|^2 \\ & + 2\rho^{(k)}\big\langle p^{(k)} - p^\sharp \, , \theta(u^{(k+1)}, w^{(k+1)}) - \Theta(u^\sharp)\big\rangle \\ & + \underbrace{\big(\rho^{(k)}\big)^2 \left\|\theta(u^{(k+1)}, w^{(k+1)}) - \Theta(u^\sharp)\right\|^2}_{T_1} \; .\end{aligned} \tag{5.10}$$

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

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

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

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

$$T_1 \leq 3\big(\rho^{(k)}\big)^2\Big(\underbrace{\big\|\theta(u^{(k+1)},w^{(k+1)})-\theta(u^{(k)},w^{(k+1)})\big\|^2}_{T1,1} + \underbrace{\big\|\theta(u^{(k)},w^{(k+1)})-\Theta(u^{(k)})\big\|^2}_{T1,2} + \underbrace{\big\|\Theta(u^{(k)})-\Theta(u^\sharp)\big\|^2}_{T1,3}\Big) .$$

– De l'hypothèse 7 et de la majoration (5.9), on obtient :

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

– De l'hypothèse 10 et de la majoration (5.6a), 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 (5.10), et multipliant de part et d'autre de l'inégalité par $\epsilon^{(k)}/2\rho^{(k)}$, on obtient :

$$\begin{aligned}\frac{\epsilon^{(k)}}{2\rho^{(k)}}\big\|p^{(k+1)}-p^\sharp\big\|^2 \leq{}& \frac{\epsilon^{(k)}}{2\rho^{(k)}}\big\|p^{(k)}-p^\sharp\big\|^2 \\ &+ \epsilon^{(k)}\big\langle p^{(k)}-p^\sharp\,,\theta(u^{(k+1)},w^{(k+1)})-\Theta(u^\sharp)\big\rangle \\ &+ \frac{3}{2}\big(\epsilon^{(k)}\rho^{(k)}\big)\big\|\theta(u^{(k)},w^{(k+1)})-\Theta(u^{(k)})\big\|^2 \\ &+ \big(\epsilon^{(k)}\rho^{(k)}\big)\mathcal{L}\big(\psi^{(k)}\big) . \end{aligned} \tag{5.11}$$

c. On majore pour finir la variation $\psi^{(k+1)}-\psi^{(k)}$. On a :

$$\begin{aligned}\psi^{(k+1)}-\psi^{(k)} ={}& \underbrace{K(u^{(k)})-K(u^{(k+1)})-\big\langle\nabla K(u^{(k)})\,,u^{(k)}-u^{(k+1)}\big\rangle}_{T2,1} \\ &+ \underbrace{\big\langle\nabla K(u^{(k)})-\nabla K(u^{(k+1)})\,,u^\sharp-u^{(k+1)}\big\rangle}_{T2,2} \\ &+ \frac{\epsilon^{(k+1)}}{2\rho^{(k+1)}}\big\|p^{(k+1)}-p^\sharp\big\|^2 - \frac{\epsilon^{(k)}}{2\rho^{(k)}}\big\|p^{(k)}-p^\sharp\big\|^2 .\end{aligned}$$

Le terme $T_{2,1}$ est négatif ou nul par convexité du noyau K et peut donc être négligé. Écrivant la condition d'optimalité (5.5) au point $u=u^\sharp$, on majore le terme $T_{2,2}$ par $\epsilon^{(k)}\big\langle g^{(k)}+(\vartheta^{(k)})^\top\cdot p^{(k)}\,,u^\sharp-u^{(k+1)}\big\rangle$. Utilisant l'hypothèse de décroissance 14 et la majoration (5.11), il vient :

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

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

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

– Appliquant l'inégalité de Schwartz au terme $T_{4,3}$ et utilisant les relations (5.8) et (5.9), il vient :

$$T_{4,3} \leq \big(\epsilon^{(k)}\rho^{(k)}\big)\mathcal{L}\big(\psi^{(k)}\big)\,.$$

– Appliquant l'inégalité de Schwartz au terme $T_{4,4}$ et utilisant la relation (5.6b), l'hypothèse 7 et la majoration (5.9), il vient :

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

On obtient donc la majoration :

$$\begin{aligned}\psi^{(k+1)} - \psi^{(k)} \leq & \underbrace{\epsilon^{(k)}\big\langle g^{(k)} + (\vartheta^{(k)})^\top \cdot p^{(k)}\,, u^\sharp - u^{(k)}\big\rangle}_{T_{5,1}} \\ & + \underbrace{\epsilon^{(k)}\big\langle p^{(k)} - p^\sharp\,, \theta(u^{(k)}, w^{(k+1)}) - \Theta(u^\sharp)\big\rangle}_{T_{5,2}} \\ & + \frac{3}{2}\underbrace{\big(\epsilon^{(k)}\rho^{(k)}\big)\big\|\theta(u^{(k)}, w^{(k+1)}) - \Theta(u^{(k)})\big\|^2}_{T_{5,3}} \\ & + \big(\epsilon^{(k)}\rho^{(k)}\big)\mathcal{L}\big(\psi^{(k)}\big)\,. \qquad (5.12)\end{aligned}$$

Cette inégalité, que l'on a écrite sur les réalisations des différentes variables aléatoires qui y sont impliquées, peut aussi s'écrire sur les variables aléatoires elles-même. 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 $(\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(k)})$. On rappelle que la variable aléatoire $\boldsymbol{W}^{(k+1)}$ est indépendante des $\boldsymbol{W}^{(l)}$ précédentes. Les variables aléatoires $\boldsymbol{U}^{(k)}$, $\boldsymbol{P}^{(k)}$ et donc $\boldsymbol{\Psi}^{(k)}$ sont par construction mesurables par rapport à la tribu $\mathcal{F}^{(k)}$, alors que les variables aléatoires $\boldsymbol{G}^{(k)}$ et $(\boldsymbol{\vartheta}^{(k)})^\top$ dépendent de $\boldsymbol{W}^{(k+1)}$.

– On considère d'abord le terme $\boldsymbol{T}_{5,1}$:

$$\boldsymbol{T}_{5,1} = \epsilon^{(k)} \big\langle \boldsymbol{G}^{(k)} + (\boldsymbol{\vartheta}^{(k)})^\top \cdot \boldsymbol{P}^{(k)} \, , u^\sharp - \boldsymbol{U}^{(k)} \big\rangle \, .$$

Prenant l'espérance conditionnelle par rapport à $\mathcal{F}^{(k)}$, on obtient :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{T}_{5,1} \;\big|\; \mathcal{F}^{(k)}\big) &= \epsilon^{(k)} \big\langle \nabla J(\boldsymbol{U}^{(k)}) + \big(\Theta'(\boldsymbol{U}^{(k)})\big)^\top \cdot \boldsymbol{P}^{(k)} \, , u^\sharp - \boldsymbol{U}^{(k)} \big\rangle \\ &\leq \epsilon^{(k)} \Big(J(u^\sharp) - J(\boldsymbol{U}^{(k)}) \\ &\qquad + \big\langle \boldsymbol{P}^{(k)} \, , \Theta(u^\sharp) - \Theta(\boldsymbol{U}^{(k)}) \big\rangle \Big) \, , \end{aligned}$$

la dernière inégalité provenant, d'une aprt de la convexité de J, et d'autre part de la C-convexité de Θ.

– On considère ensuite le terme $\boldsymbol{T}_{5,2}$:

$$\boldsymbol{T}_{5,2} = \epsilon^{(k)} \big\langle \boldsymbol{P}^{(k)} - p^\sharp \, , \theta(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) - \Theta(u^\sharp) \big\rangle \, .$$

Prenant l'espérance conditionnelle par rapport à $\mathcal{F}^{(k)}$, on obtient :

$$\mathbb{E}\big(\boldsymbol{T}_{5,2} \;\big|\; \mathcal{F}^{(k)}\big) = \epsilon^{(k)} \big\langle \boldsymbol{P}^{(k)} - p^\sharp \, , \Theta(\boldsymbol{U}^{(k)}) - \Theta(u^\sharp) \big\rangle \, .$$

– On considère enfin le terme $\boldsymbol{T}_{5,3}$:

$$\boldsymbol{T}_{5,3} = \big(\epsilon^{(k)} \rho^{(k)}\big) \big\| \theta(\boldsymbol{U}^{(k)}, \boldsymbol{W}^{(k+1)}) - \Theta(\boldsymbol{U}^{(k)}) \big\|^2 \, .$$

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 (5.6a), on obtient :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{T}_{5,3} \;\big|\; \mathcal{F}^{(k)}\big) &\leq \big(\epsilon^{(k)} \rho^{(k)}\big) \big(\gamma \big\| \boldsymbol{U}^{(k)} \big\|^2 + \delta\big) \\ &\leq \big(\epsilon^{(k)} \rho^{(k)}\big) \mathcal{L}\big(\boldsymbol{\Psi}^{(k)}\big) \, . \end{aligned}$$

Prenant l'espérance conditionnelle par rapport à la tribu $\mathcal{F}^{(k)}$ dans la relation (5.12) et y reportant les majorations des termes $\boldsymbol{T}_{5,1}$, $\boldsymbol{T}_{5,2}$ et $\boldsymbol{T}_{5,3}$, on obtient :

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{\Psi}^{(k+1)} \;\big|\; \mathcal{F}^{(k)}\big) - \boldsymbol{\Psi}^{(k)} &\leq \big(\epsilon^{(k)} \rho^{(k)}\big) \mathcal{L}\big(\boldsymbol{\Psi}^{(k)}\big) \\ &\quad + \epsilon^{(k)} \big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^{(k)}, p^\sharp)\big) \, , \end{aligned}$$

où L est le Lagrangien associé au problème (5.1). On en déduit l'existence de constantes c_1 et c_2 telles que :

$$\mathbb{E}\big(\boldsymbol{\Psi}^{(k+1)} \mid \mathcal{F}^{(k)}\big) - \boldsymbol{\Psi}^{(k)} \leq \big(\epsilon^{(k)}\rho^{(k)}\big)\big(c_1\boldsymbol{\Psi}^{(k)} + c_2\big) + \epsilon^{(k)}\big(L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^{(k)}, p^\sharp)\big) \ . \quad (5.13)$$

Comme noté à la remarque 5.4, la série de terme général $\big(\epsilon^{(k)}\rho^{(k)}\big)$ est convergente. Le terme $L(u^\sharp, p^\sharp) - L(\boldsymbol{U}^{(k)}, p^\sharp)$ est quant à lui négatif.

3. **Analyse de convergence.** Par le théorème 3.6, 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 :

$$\sum_{k=0}^{+\infty} \epsilon^{(k)}\big(L(\boldsymbol{U}^{(k)}, p^\sharp) - L(u^\sharp, p^\sharp)\big) < +\infty \ , \quad \mathbb{P}\text{-p.s.} \ . \quad (5.14)$$

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 (5.6) 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 3.7 étant satisfaites, on déduit de (5.14) 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 (5.14) 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 $\mathbb{U}$. 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 fait que l'on a :

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

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 du Lagrangien implique alors que $\overline{u} \in U^\sharp$.

Remarque 5.8. Si on fait l'hypothèse supplémentaire de C-convexité sur la fonction $u \mapsto \theta(u, w)$, on peut remplacer dans le problème auxilaire (5.3a) 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 :

$$u^{(k+1)} \in \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}} K(u) + \big\langle \epsilon^{(k)} g^{(k)} - \nabla K(u^{(k)})\,, u\big\rangle + \epsilon^{(k)}\big\langle p^{(k)}\,, \theta(u, w^{(k+1)})\big\rangle \ ,$$

ce nouveau problème étant fortement convexe et ayant donc une solution unique. La preuve de convergence précédente s'adapte alors facilement à cette variante.

5.2 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 :

$$\mathbb{P}\big(\theta(u, \boldsymbol{W}) \in -C\big) = \mathbb{E}\big(\mathbf{1}_{\{\theta(u,\boldsymbol{W})\in -C\}}\big) \,. \tag{5.15}$$

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 $\mathbb{U}$. Les propriétés de connexité, convexité et de différentiabilité des contraintes en probabilité ont été étudié 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). Une manière de surmonter les potentielles non convexités associées à ce type de contrainte est d'utiliser un *Lagrangien augmenté* plutôt qu'un Lagrangien ordinaire.
2. Une autre difficulté vient de ce que la transformation dans (5.15) 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 (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 [3], la contrainte

3. Voir les relations (4.21) définissant le Lagrangien augmenté, la fonction ζ_c et ses gradients, avec dans le cas qui nous intéresse $\Theta(u) = \mathbb{E}\big(\theta(u, \boldsymbol{W})\big)$.

introduit dans le critère un terme de la forme $f\big(\mathbb{E}\big(\theta(u,\boldsymbol{W})\big)\big)$, dont le gradient est donné par l'expression :

$$\mathbb{E}\big(\theta'_u(u,\boldsymbol{W})\big)^\top \cdot \nabla f\big(\mathbb{E}\big(\theta(u,\boldsymbol{W})\big)\big) \ .$$

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[4] :

$$\mathbb{E}\big(\theta(u,\boldsymbol{W})\big) = 0 \ ,$$

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 :

$$\begin{aligned} u^{(k+1)} \in \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}} & K(u) + \big\langle \epsilon^{(k)} \nabla_u j(u^{(k)}, w^{(k+1)}) - \nabla K(u^{(k)}) \, , u\big\rangle \\ & + \epsilon^{(k)} \big\langle p^{(k)} + c\theta(u^{(k)}, w_2^{(k+1)}) \, , \theta'_u(u^{(k)}, w_1^{(k+1)}) \cdot u\big\rangle \ , \\ p^{(k+1)} = \mathrm{proj}_{C^\star} & \Big(p^{(k)} + \rho^{(k)} \, \theta(u^{(k+1)}, w_2^{(k+1)})\Big) \ . \end{aligned}$$

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 a été traité dans Andrieu et collab. (2011). On choisit de le présenter ici le cas d'une contrainte en probabilité scalaire :

$$\mathbb{P}\big(\theta(u,\boldsymbol{W}) \le \alpha\big) \ge \pi \ ,$$

qui, par la transformation (5.15), conduit à considérer la fonction :

4. Pour le cas des contraintes générales $\mathbb{E}\big(\theta(u,\boldsymbol{W})\big) \in -C$, on consultera (Strugarek, 2006, Chapitre VI).

$$\Theta(u) = \mathbb{E}\big(\mathbf{1}_{\mathbb{R}^+}\big(\alpha - \theta(u, \boldsymbol{W})\big)\big) \ .$$

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 :

$$\int_{-\infty}^{+\infty} h(u)\, \mathrm{d}u = 1 \ .$$

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 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 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.