Optimisation par Simulation et Apprentissage par Renforcement

Frédérick Garcia
fgarcia@toulouse.inra.fr

INRA Unité de Biométrie et Intelligence Artificielle Toulouse

Tutoriel RFIA - Apprentissage par renforcement et Robotique
19/01/2010

Introduction

Objectif du tutorial :

- présenter les bases de l'optimisation par simulation
- expliciter quelques liens établis ou à creuser entre optimisation par simulation et apprentissage par renforcement

Plan :

Introduction

Conception par simulation de systèmes pilotés

- Systèmes dynamiques complexes
- Simulation du fonctionnement du système dans le temps
- Conception par simulation :
  - choix d'une famille de politiques de contrôle $\pi \in \Pi$
  - optimisation de π par simulation du système, pour le critère

$$J = E[L(\xi)]$$

- apprentissage par renforcement / optimisation par simulation

Optimisation par simulation

(Fu et al., 2008)

- hypothèses générales sur J et L, mais généralement modèles stochastiques à évènements discrets
- Choix d'une famille de politiques paramétrées $\pi(\theta)$, $\theta \in \Theta$

$$J(\theta^*) = \max_{\theta \in \Theta} E[L(\xi)]$$

- domaine $\Theta \subseteq \mathbb{R}^p$ simple discret ou continu
- Plusieurs approches itératives :
  - méthodes d'optimisation ordinale
  - méthodes de gradient stochastique
  - méthodes de surfaces de réponse séquentielles
  - méthodes d'optimisation globales
  - méthodes meta-heuristiques

Optimisation par simulation

optimisation ordinale

Domaines discrets de taille moyenne

- Recherche rapide de solution de bonne qualité
- $\Theta = \{\theta_1, \theta_2, \ldots, \theta_p\}$.
- $J(\theta_i)$ estimé par

$$\bar{J}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} L_{\theta_i}(\xi_{ij})$$

- Après $N = N_1 + \cdots N_p$ simulations allouées à la résolution du problème, sélection de θ_b avec

$$b = \underset{i}{\operatorname{argmax}}\, \bar{J}_i.$$

Méthodes optimales d'allocation

- Maximisation de la probabilité de *sélection correcte* du système optimal θ^* :

$$P(SC) = P(\theta_b = \theta^*)$$

- Existence de règles simples de contrôle assurant une convergence exponentiellement rapide en N de $P(SC)$ vers 1
- règle OCBA (optimal computing budget allocation), Chen et al. (2000)

$$\frac{N_b}{N_i} = \sigma_b \sqrt{\sum_{j=1, j\neq b}^{p} \frac{1}{\sigma_j^2} \rho_{ij}^2}, \quad i \neq b,$$

avec

$$\rho_{ij} = \left(\frac{\sigma_j / \Delta_j}{\sigma_i / \Delta_i} \right)^2, \quad \Delta_i = J_b - J_i$$

Optimisation par simulation

gradient stochastique

Domaines continus

Approximation stochastique : version stochastique de la montée de gradient en déterministe

$$\theta_{n+1} = \theta_n + \alpha_n \hat{\nabla} J(\theta_n)$$

avec $\hat{\nabla} J(\theta_n)$ une estimation du gradient

$$\nabla J(\theta_n) = \begin{pmatrix} \cdots \\ \frac{\partial}{\partial \theta_i} E[L(\xi)] \\ \cdots \end{pmatrix}$$

et

$$\alpha_n \to 0.$$

Optimisation par simulation

surfaces de réponse séquentielles

Response Surface methodology (RSM) pour domaines continus (Barton and Meckesheimer, 2006)
Construction itérative autour de θ_n d'un modèle réduit de J :

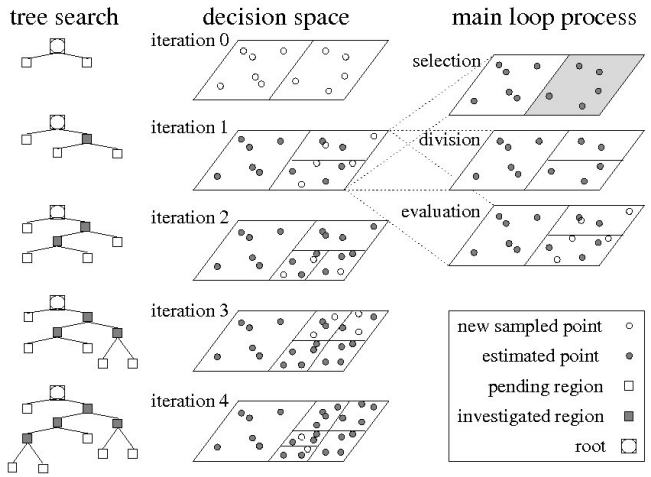
1. construire un plan d'expérience
2. simuler le plan
3. construire un modèle réduit de J
4. déterminer l'optimum de l'approximation
5. faire un test de continuation-arrêt
6. itérer si continuation

Optimisation par simulation

approches globales

domaines discrets de grande taille ou continus

- stochastic branch and bound
- nested partitions (Shi and Olafsson, 2000)
- P2 (Crespo et al., 2009)
  1. partionner la région la plus prometteuse
  2. échantillonage alétoire des sous-régions
  3. détermination de la nouvelle région la plus prometteuse



Optimisation par simulation

métaheuristiques

- adaptation au cas stochastique des méthodes à J déterministe
- méthodes itératives locales $\theta_{n+1} \in \mathcal{V}(\theta_n)$ ou populationnelles
  - tabu search,
  - recuit simulé,
  - algorithmes génétiques, ...
- peu de considérations statistiques ou probabilistes
- très présents dans les packages commerciaux de simulation

Méthodes de gradient en OS

Approximation stochastique

Approximation stochastique : recherche itérative du zéro d'une fonction bruitée (Robbins and Monro, 1951)

$$\theta_{n+1} = \theta_n + \alpha_n U(\theta_n, \xi)$$

Convergence p.s. de θ_n vers une solution θ^*

$$E[U(\theta^*, \xi)] = 0$$

sous des hypothèses de régularité de U et pour $\sum \alpha_n = \infty$, $\sum {\alpha_n}^2 < \infty$:

$$n^{1/2}(\theta_n - \theta^*) \xrightarrow{dist.} \mathcal{N}(0, \Sigma) \text{ quand } n \to \infty$$

Application à l'optimisation

$$J(\theta^*) = \max_{\theta \in \Theta} E[L(\xi)]$$

pour une montée de gradient stochastique

$$\theta_{n+1} = \theta_n + \alpha_n \hat{\nabla} J(\theta_n)$$

avec

$$\nabla J(\theta) = \begin{pmatrix} \cdots \\ \frac{\partial}{\partial \theta^i} E[L(\xi)] \\ \cdots \end{pmatrix}$$

Problème : estimation du gradient

Différences finies

Méthode naïve de différence finie:

$$\frac{\partial}{\partial \theta^i} E[L(\xi)] \approx \frac{\hat{J}(\theta + \delta^i) - \hat{J}(\theta)}{\delta^i}$$

avec

$$\hat{J}(\theta) = \frac{1}{N} \sum_{j=1}^{N} L_\theta(\xi_j)$$

- coût élevé de l'estimation : $N(p+1)$ simulations par itération
- N élevé $\Rightarrow$ variance $\searrow$
- δ élevé $\Rightarrow$ variance $\searrow$ mais biais $\nearrow$
- différences symétriques $(\theta + \delta^i)$ et $(\theta - \delta^i)$: biais $\searrow$ mais $2Np$ simulations.

Différences finies

Kiefer-Wolfowitz

$N = 1$: différence finie et approximation stochastique (FDSA), méthode de Kiefer-Wolfowitz (1952)

$$\frac{\partial}{\partial \theta^i} E[L(\xi)](\theta_n) \approx \frac{L_{\theta_n + \delta_n e^i}(\xi) - L_{\theta_n - \delta_n e^i}(\xi')}{2\delta_n}$$

avec $e^i = (0, \cdots, 1, 0, \cdots)$ vecteur de base de $\mathbb{R}^p$

- $\delta_n \to 0$
- $\sum \alpha_n = \infty$, $\sum \alpha_n {\delta_n}^2 < \infty$, $\sum (\alpha_n / \delta_n)^2 < \infty$,
- variance importante, qui $\nearrow$ quand $\delta_n \to 0$.

$$n^{1/3}(\theta_n - \theta^*) \xrightarrow{dist.} \mathcal{N}(\mu, \Sigma) \text{ quand } n \to \infty$$

Différences finies

Perturbations simultanées

Méthode des perturbations simultanées (SPSA), Spall (1992)

$$\frac{\partial}{\partial \theta_i} E[L(\xi)](\theta_n) \approx \frac{L_{\theta_n + \delta_n \Delta}(\xi) - L_{\theta_n - \delta_n \Delta}(\xi')}{2\delta_n \Delta_i}$$

avec $\Delta = (\cdots, +/-1, \cdots)$ aléatoire dans $\{-1, 1\}^p$

- même vitesse de convergence que FDSA
- 2 simulations par itération, efficace pour p grand.

Différences finies

Common seed

Pour minimiser la variance de

$$\frac{L_{\theta+\delta}(\xi) - L_{\theta-\delta}(\xi')}{2\delta}$$

on utilise des tirages communs pour ξ et ξ' :

$$\frac{L_{\theta+\delta}(\xi) - L_{\theta-\delta}(\xi)}{2\delta}$$

En terme de simulation, initialisation du générateur aléatoire avec la même graine commune en $\theta + \delta$ et $\theta - \delta$.
Accéleration du taux de convergence de FDSA et SPSA de $n^{-1/3}$ à $n^{-1/2}$ (Kleinman, 1999)

Différences finies

2nd ordre

Utilisation de la matrice Hessienne

$$H = \left(\frac{\partial^2}{\partial \theta^i \partial \theta^j} E[L(\xi)] \right)_{ij}$$

comme gain matriciel pour guider et accélerer la convergence des algorithmes stochastiques adaptatifs (Benveniste et al., 1990)

$$\theta_{n+1} = \theta_n + \alpha_n \hat{H}_n^{-1} \hat{\nabla} J(\theta_n)$$

- gain matriciel optimal = $H^{-1}(\theta^*)$ inconnu
- $\hat{H}_n$ estimée adaptativement à partir de $\hat{\nabla} J(\theta)$,
- utilisation possible de la méthode des perturbations simultanées : 3 simulations par itération (Spall, 2000).

Différences finies

Moyennisation

Les méthodes de moyennisation (averaging) permettent d'approcher le gain optimal asymptotique dans le cadre Robbins-Monro (Polyak and Juditsky, 1992)
Utilisation possible pour les méthodes de gradient stochastique :

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_n \hat{\nabla} J(\theta_n) \\ \bar{\theta}_{n+1} &= \tfrac{1}{n+1}\theta_{n+1} + (1 - \tfrac{1}{n+1})\bar{\theta}_n \end{aligned}$$

- comportement asymptotique optimal non garanti
- intérêt pour $H(\theta^*)$ élevée
- autorise des gains α_n plus élevés

Méthodes de dérivation

Les méthodes de différences finies apparaissent comme des méthodes brutales de dérivation, avec hypothèse boite noire pour L.
Pour certains types d'applications :

- modèles de file d'attente
- modèles de gestion de stock avec des politiques (s, S)
- modèles d'options en finance
- MDP

la structure de L est connue et autorise des méthodes plus fines d'estimation de $\hat{\nabla} J(\theta)$ (Fu, 2008)

Méthodes de dérivation

Analyse des perturbations infinitésimales (IPA)

l'IPA suppose que la distribution de ξ ne dépend pas de θ :

$$J(\theta) = E[L(\xi)] = \int_{\Xi} L_\theta(\xi) dP(\xi)$$

Sous certaines hypothèses (fortes) de régularité de L, on a alors

$$\begin{aligned} \frac{\partial}{\partial \theta^i} J(\theta) &= \int_{\Xi} \frac{\partial}{\partial \theta^i} L_\theta(\xi) dP(\xi) \\ &= E[\frac{\partial}{\partial \theta^i} L_\theta(\xi)] \end{aligned}$$

Utile si $\frac{\partial}{\partial \theta^i} L_\theta(\xi)$ peut être simulé exactement

Méthodes de dérivation

Analyse des perturbations infinitésimales (IPA)

- exemples :

$$\begin{array}{ll} L_\theta(\xi) = \theta + Z(\xi) & \frac{dL}{d\theta} = 1 \\ L_\theta(\xi) = \theta Z(\xi) & \frac{dL}{d\theta} = \frac{1}{\theta} L_\theta(\xi) \\ L_\theta(\xi) = \bar{\theta} + \theta Z(\xi) & \frac{dL}{d\theta} = \frac{1}{\theta}(L_\theta(\xi) - \bar{\theta}) \end{array}$$

- dérivation de fonctions composées le long d'un chemin de simulation
- en pratique l'échange ∇ et $E[\ ]$ n'est pas toujours valide directement

Méthodes de dérivation

Rapport de vraisemblance / fonction score

Cas où c'est la loi de ξ qui dépend de θ (densité f_θ) :

$$J(\theta) = E[L(\xi)] = \int_\Xi L(\xi) f_\theta(\xi) d\xi$$

Alors

$$\begin{aligned} \frac{\partial}{\partial \theta^i} J(\theta) &= \int_\Xi L(\xi) \frac{\partial}{\partial \theta^i} f_\theta(\xi) d\xi \\ &= \int_\Xi \left(L(\xi) \frac{\frac{\partial f_\theta}{\partial \theta^i}(\xi)}{f_\theta(\xi)} \right) f_\theta(\xi) d\xi = E[L(\xi) \frac{\frac{\partial f_\theta}{\partial \theta^i}(\xi)}{f_\theta(\xi)}] \end{aligned}$$

Estimateur $L(\xi) \frac{\nabla f_\theta(\xi)}{f_\theta(\xi)}$ dit du rapport de vraisemblance ou fonction score (LR/SF)

Méthodes de dérivation

Rapport de vraisemblance / fonction score

En pratique, $\xi = (\xi_1, \ldots, \xi_t)$ le long d'une trajectoire, selon des distributions $f_{i,\theta}(\xi_i)$
Alors

$$\nabla E[L(\xi)] = E\left[L(\xi) \sum_{i=1}^{t} \nabla \ln f_{i,\theta}(\xi_i)\right]$$

Apprentissage par renforcement

PDM

- J et L reposent sur les processus décisionnels de Markov (PDM)
  - $< S, A, P, R >$
  - $L = r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots + \gamma^t r_t + \cdots$
  - politique π : fonction de $S \rightarrow A$
- Fonction de valeur V
  - $V : S \rightarrow \mathbb{R}$, associée à chaque politique π
  - $V(s) = E[r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots + \gamma^t r_t + \cdots \mid \pi, s_0 = s]$
  - V^* associée à la politique optimale π^*
- Equation d'optimalité de Bellman en V^*

$$V^*(s) = \max_a \sum_{s'} p(s' \mid s, a)\{r(s, a, s') + \gamma V^*(s')\}, \quad \forall s$$

$$\pi^*(s) = \operatorname{argmax}_a \sum_{s'} p(s' \mid s, a)\{r(s, a, s') + \gamma V^*(s')\}, \quad \forall s$$

Apprentissage par renforcement

Méthodes

Méthodes itératives sur paramètres continus :

- Approximation de la fonction de valeur optimale
- Optimisation directe de la politique π

Apprentissage par renforcement

Approximation de la fonction de valeur

- Paramétrisation des fonctions de valeur $V^{\omega}(s)$ (ou $Q^{\omega}(s,a)$)
- Mise à jour itérative en cours de simulation de ω

$$\omega_{k+1} = \omega_k + \alpha_k \Delta(< s_k, a_k, c_k, s_{k+1} >, \omega_k)$$

- Politique optimale approchée :

$$\begin{aligned}
\hat{\pi}^*(s) &= \underset{a}{\operatorname{argmin}}\, Q^{\omega^*}(s,a) \\
&= \underset{a}{\operatorname{argmin}} \sum_{s'} \hat{p}(s' \mid s,a)\{r(s,a,s') + \gamma V^{\omega^*}(s')\} \\
&= \underset{a}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \{r(s,a,s'_i) + \gamma V^{\omega^*}(s'_i)\}
\end{aligned}$$

Apprentissage par renforcement

Optimisation directe de la politique

- Paramétrisation des politiques $\pi_\theta(s)$
- Mise à jour itérative en cours de simulation de θ

$$\theta_{k+1} = \theta_k + \alpha_k \Delta(< s_k, a_k, c_k, s_{k+1} >, \theta_k)$$

- Possibilité de coupler apprentissage de la fonction de valeur et optimisation de la politique : architectures acteur-critique

Approximation stochastique et apprentissage par renforcement

Q-learning

Equation de Bellman :

$$V^*(s) = \max_a \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma V^*(s')\}, \quad \forall s$$

De manière équivalente :

$$\begin{aligned} V^*(s) &= \max_a Q^*(s,a), \quad \forall s \\ Q^*(s,a) &= \sum_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma \max_{a'} Q^*(s',a')\}, \quad \forall s,a \\ Q^*(s,a) &= E[r(s,a,s') + \gamma \max_{a'} Q^*(s',a')], \quad \forall s,a \end{aligned}$$

Approximation stochastique et apprentissage par renforcement

Q-learning

On pose $\xi = (s, a, r, s')$, et $H(Q, \xi)_{s,a} = (r + \gamma \max_{a'} Q(s', a') - Q(s, a))$
Alors

$$E[H(Q^*, \xi)] = 0$$

La méthode de Robbins-Monro donne directement

$$Q_{n+1} = Q_n + \alpha_n H(Q_n, \xi_n)$$

soit

$$Q_{n+1}(s_n, a_n) = Q_n(s_n, a_n) + \alpha_n (r_n + \gamma \max_{a'} Q_n(s_{n+1}, a') - Q_n(s_n, a_n))$$

(Williams, 1992 ; Baird and Moore, 1999 ; Sutton et al., 2000 ; Baxter and Bartlett, 2001)
Politique π_θ aléatoire

$$\pi_\theta(s) = a \text{ avec la proba } q(a \mid s; \theta) \quad \forall s$$

Exemples :

$$q(a \mid s; \theta) = \frac{\theta_{s,a}}{\sum_b \theta_{s,b}}, \quad , \theta_{s,a} \geq 0$$

$$q(a \mid s; \theta) = \frac{\exp \theta.\phi(s,a)}{\sum_b \exp \theta.\phi(s,b)}$$

avec

$$\phi(s,a) = (\phi_1(s,a), \ldots, \phi_p(s,a))$$

MDP à un coup ($T = 1$), état initial tiré selon $\mu(s)$

$$J(\theta^*) = \max_\theta E[r(s,a,s')]$$

Méthode de la fonction score

$$\begin{aligned}
\nabla E[r(s,a,s')] &= \nabla \left[\sum_{s,a,s'} \mu(s) q(a|s;\theta) p(s'|s,a) r(s,a,s') \right] \\
&= \sum_{s,a,s'} \left[\frac{\nabla q(a|s;\theta)}{q(a|s;\theta)} r(s,a,s') \right] \mu(s) q(a|s;\vec{\theta}) p(s'|s,a) \\
&= E\left[\frac{\nabla q(a|s;\theta)}{q(a|s;\theta)} r(s,a,s') \right].
\end{aligned}$$

Avec $q(a \mid s;\theta) = \frac{\exp \theta.\phi(s,a)}{\sum_b \exp \theta.\phi(s,b)}$:

$$\frac{1}{q(a|s;\theta)} \frac{\partial q(a|s;\theta)}{\partial \theta_i} = \phi_i(s,a) - \sum_b \phi_i(s,b) q(b|s;\theta),$$

soit

$$\frac{\nabla q(a|s;\theta)}{q(a|s;\theta)} = \phi(s,a) - \sum_b q(b|s;\theta)\phi(s,b)$$

Horizon $T \geq 1$

$$J(\theta^*) = \max_{\theta} E[r(s_0, a_0, s_1) + \gamma r(s_1, a_1, s_2) + \cdots + \gamma^{T-1} r(s_{T-1}, a_{T-1}, s_T)]$$

Une démarche similaire conduit à

$$\nabla J(\theta) = \nabla E[\sum_{t=0}^{T-1} \gamma^t r_t] = E\left[\sum_{t=0}^{T-1} \gamma^t r_t \sum_{t'=0}^{t-1} \frac{\nabla q(a_{t'}|s_{t'};\theta)}{q(a_{t'}|s_{t'};\theta)}\right]$$

Méthodes adaptatives en horizon infini (Baxter and Bartlett, 2001)

Horizon infini, équation de bellman

$$V_\theta(s) = \sum_a q(a|s;\theta) \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V_\theta(s'))$$

On montre que

$$\nabla V_\theta(s) = E\left[\sum_{t=0}^{\infty} \gamma^t \frac{\nabla q(a_t|s_t;\theta)}{q(a_t|s_t;\theta)} Q_\theta(s,a)|s_0 = s\right]$$

d'où

$$\begin{aligned} \nabla J(\theta) &= \sum_s \mu(s) \nabla V_\theta(s) \\ &= \sum_s \mu_\theta^\gamma(s) \sum_a \nabla q(a|s;\theta) Q_\theta(s,a) \end{aligned}$$

Gradient stochastique et apprentissage par renforcement

Méthode acteur-critique

Utilisation d'une fonction de valeur approchée $Q^\omega(s,a)$ estimée au cours de la simulation
Critère de compatibilité entre les paramétrages ω et θ

$$\nabla Q^\omega(s,a) = \frac{\nabla q(a|s;\theta)}{q(a|s;\theta)}$$

Exemple pour $q(a \mid s;\theta) = \frac{\exp \theta.\phi(s,a)}{\sum_b \exp \theta.\phi(s,b)}$:

$$Q^\omega(s,a) = \omega.(\phi(s,a) - \sum_b q(b|s;\theta)\phi(s,b))$$

Q^ω linéaire en ϕ_i

Voir aussi le gradient naturel et Natural Actor-Critic (Peters and Schaal, 2008)

Conclusions

- Un réel transfert aujourd'hui du savoir-faire en optimisation par approximation stochastique et dérivation vers la communauté de l'apprentissage par renforcement
- liens entre optimisation ordinale et exploration ou encore optimisation en ligne
- liens entre approximation de fonction de valeur et metamodélisation en optimisation par simulation
- tirages communs et apprentissage par renforcement ?
- optimisation par simulation / apprentissage par renforcement et parallélisme
- des liens à faire avec les résultats sur les SMDP en optimisation par simulation
- apprentissage par renforcement et modèles à évènements discrets (Rachelson et al., 2008) ?

Bibliographie

L. C. Baird and A.W. Moore, Gradient Descent for General Reinforcement Learning, NIPS'98, MIT Press, 1999.

R. R. Barton and M. Meckesheimer, Metamodel-based simulation optimisation, Handbooks in Research and Management Science: Simulation, Elsevier, 2006.

J. Baxter and P. Bartlett, Infinite-Horizon Policy-Gradient Estimation, Journal of Artificial Intelligence Research, 15, 2001.

A. Benveniste, M. Metivier and P. Priouret, Adaptive Algorithms and Stochastic Approximation, Springer-Verlag, 1990

O. Buffet, Méthodes de gradient pour la recherche de politiques paramétrées. Processus décisionnels de Markov en intelligence artificielle, volume 2. Hermes Science Publishing, 2008

C. H. Chen, J. Lin, E. Yucesan and S. E. Chick, Simulation budget allocation for further enhancing the efficiency of ordinal optimization, Discrete Event Dynamic Systems, 2000

O. Crespo, J.-E Bergez, F. Garcia, P2 hierarchical decomposition procedure: application to irrigation strategies design, Operational Research, 2009

M. C. Fu, What you should know about simulation and derivates, Naval Research Logistics, (55)8, 2008

M. C. Fu, C.-H. Chen and L. Shi, Some topics for simulation optimization, Winter Simulation Conference, 2008

J. Kiefer and J. Wolfowitz, Stochastic estimation of a regression function, Annals of mathematical statistics, 23, 1952

N. Kleinman, J. C. Spall and D. Q. Naiman, Simulation-based optimization with stochastic approximation using common random numbers, Management Science, 45, 1999

J. Peters and S. Schaal, Natural Actor-Critic, Neurocomputing 71(7-9), 2008

B. T. Polyak and A. B. Juditsky, Acceleration of Stochastic Approximation by Averaging, SIAM Journal on Control and Optimization, 30, 1992

E. Rachelson, Q. Quesnel, F. Garcia, P. Fabiani, A Simulation-based Approach for Solving Generalized Semi-Markov Decision Processes, ECAI, 2008

Bibliographie

H. Robbins and S. Monro, A Stochastic Approximation Method, Annals of mathematical statistics, 22, 1951

L. Shi and S. Olafsson, Nested partitions methods for global optimization, Operations Research, 48, 2000

O. Sigaud et F. Garcia, Apprentissage par renforcement. Processus décisionnels de Markov en intelligence artificielle, volume 2. Hermes Science Publishing, 2008

J. C. Spall, Multivariate stochastic approximation using a simultaneous pertubation gradient approximation, IEEE Trans. on Automatic Control 37(3), 1992

J. C. Spall, Adaptive stochastic approximation using a simultaneous pertubation gradient approximation, IEEE Trans. on Automatic Control 45, 2000

R. Sutton, D. McAllester, S. Singh and Y. Mansour, Policy Gradient Methods for Reinforcement Learning with Function Approximation, NIPS'99, MIT Press, 2000.

R. J. Williams, Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning, Machine Learning, (8)3, 1992.