Algorithme d'itération sur les politiques avec approximation linéaire

Rémi Munos
Remi.Munos@polytechnique.fr
Centre de Mathématiques Appliquées
École Polytechnique. 91128 Palaiseau, France.

Résumé

Nous utilisons, dans le cadre d'un espace continu, les résultats de majoration en norme L_2 (Munos, 2003) de la performance des politiques calculées par un algorithme d'*itération sur les politiques avec approximation*. Ces majorations sont plus fines que celles usuelles en norme L_∞ (Bertsekas & Tsitsiklis, 1996) et sont pertinentes lorsque l'on considère des approximations de type régression par les moindres carrés (car la régression se fait en minimisant une erreur L_2). Nous présentons deux approches d'approximation basées sur le *résidu de Bellman* et sur les *différences temporelles*, indiquons leur possible implémentation par un algorithme d'*apprentissage par renforcement*, et les expérimentons sur un problème de remplacement optimal.

1 Introduction

Considérons un **Processus de Décision Markovien** (PDM) (Puterman, 1994; Bertsekas & Tsitsiklis, 1996) défini sur un espace d'état continu $X \subset I\!R^n$ et soit $\mathbf{B}$ une tribu borélienne.

Les dynamiques du PDM dépendent de la fonction **probabilité de transition** $P(x, a, B)$ qui donne la probabilité que l'état suivant soit dans $B \in \mathbf{B}$ sachant que l'état courant est x et que l'action choisie est $a \in A$, où A est un ensemble d'actions possibles (A peut être discret ou continu).

Rappelons qu'une probabilité de transition $P(x, a, B)$ est un noyau de transition positif tel que pour tout $x \in X$, $a \in A$, $P(x, a, X) = 1$. De plus, P définit à la fois des opérateurs droit et gauche : pour toute fonction mesurable h sur $X \times A$, Ph représente la fonction $Ph(x, a) := \int_X P(x, a, dy)h(y)$, et pour toute mesure μ sur $\mathbf{B}$, μP représente la mesure $\mu P(a, B) := \int_X \mu(dx)P(x, a, B)$. Le produit P_1P_2 de deux noyaux est défini par $P_1P_2(x, a, B) := \int_X P_1(x, a, dy)$ $P_2(y, a, B)$.

Nous appelons **politique** π une application de X dans A.

Notons $P^\pi(x, B) := P(x, \pi(x), B)$. Soit $r(x, a, y)$ la **récompense** reçue lors de la transition : $(x \in X, a \in A) \to y \in X$. Notons $r^\pi(x, y) := r(x, \pi(x), y)$ et $r^\pi(x) := \int_X P^\pi(x, dy)r^\pi(x, y)$.

Pour toute mesure positive μ telle que $\mu(X) = 1$, nous définissons le produit scalaire $\langle f, h\rangle_\mu := \int_X \mu(dx)f(x)h(x)$ et la semi-norme : $||h||_\mu = \langle h, h\rangle_\mu^{1/2}$.

Nous considérons un problème actualisé à horizon temporel infini. Pour une politique donnée π, la **fonction valeur** V^π en x est l'espérance de la somme des récompenses futures actualisées, partant de x et suivant π :

$$V^\pi(x) = \mathbb{E}_{x,\pi}[\sum_{t=0}^{\infty} \gamma^t r_t],$$

où r_t est la récompense reçue à l'instant t et $\gamma \in [0,1)$ le *coefficient d'actualisation*. La fonction V^π satisfait l'équation de Bellman (Puterman, 1994)

$$V^\pi(x) = r^\pi(x) + \gamma \int_X P^\pi(x, dy) V^\pi(y).$$

Donc V^π est le point fixe de l'opérateur T^π défini, pour tout fonction mesurable h sur X, par $T^\pi h := r^\pi + \gamma P^\pi h$. Donc $(I - \gamma P^\pi)V^\pi = r^\pi$. Puisque P^π est un opérateur stochastique, il possède des valeurs propres de module inférieur à un, donc $(I - \gamma P^\pi)$ est inversible et l'on peut écrire

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi = \sum_{k=0}^{\infty} \gamma^k (P^\pi)^k r^\pi.$$

La **fonction valeur optimale** V^* est le gain lorsqu'on utilise une **politique optimale** π^* : $V^* = V^{\pi^*} = \sup_\pi V^\pi$.

Maintenant, considérons l'espace produit état-action $X \times A$. Par abus de notation, notons à nouveau P^π l'opérateur droit défini, pour toute fonction q mesurable sur $X \times A$, par

$$P^\pi q(x, a) := \int_X P(x, a, dy) q(y, \pi(y)).$$

Soit $r(x, a) := \int_X P(x, a, dy) r(x, a, y)$, et notons T^π l'opérateur défini sur $X \times A$ par : $T^\pi q(x, a) := r(x, a) + \gamma P^\pi q(x, a)$ (par soucis de simplicité nous utilisons les mêmes notations pour les opérateurs P^π et T^π défini sur les espace d'état X et d'état-action $X \times A$).

Il est utile de définir la **fonction-Q** $Q^\pi(x, a)$ définie, pour toute politique π, comme le point fixe de T^π : $Q^\pi(x, a) := T^\pi Q^\pi(x, a)$, et la **fonction-Q optimale** $Q^*(x, a)$, point fixe de T^{π^*}. Ainsi, nous avons $V^\pi(x) = Q^\pi(x, \pi(x))$ et $V^*(x) = Q^*(x, \pi^*(x)) = \max_a Q^*(x, a)$.

Puisque l'espace d'état est continu, nous ne pouvons pas utiliser des méthodes de résolution exacte, telles que les méthodes d'*itération sur les valeurs* ou d'*itération sur les politiques* (Puterman, 1994). Au lieu de cela, nous considérons un algorithme d'**itérations sur les politiques avec approximation** (IPA) (Bertsekas & Tsitsiklis, 1996) défini itérativement par les deux étapes suivantes :

- *Evaluation avec approximation de la politique* : pour une politique π_k, on génère une approximation Q_k de la fonction-Q Q^{π_k} (ou, de manière équivalente, une approximation V_k de la fonction valeur V^{π_k}).
- *Amélioration de la politique* : on génère une nouvelle politique π_{k+1} *déduite de* Q_k, c'est à dire, telle qu'en tout x,

$$\pi_{k+1}(x) \in \arg\max_{a \in A} Q_k(x, a).$$

La partie 2 résume les résultats (Munos, 2003) de majorations sur l'erreur de performance en fonction des erreurs d'approximation ou du résidu de Bellman (Baird, 1995) au cours des étapes d'évaluation de la politique.

La partie 3 détaille le cas de l'*approximation linéaire* (lorsque les fonctions paramétrées sont des combinaisons linéaires d'un dictionnaire donné de fonctions, appelées "features"). Ce cas a souvent été utilisé en combinaison avec des algorithmes de type *différences temporelles* TD(λ) (Tsitsiklis & Van Roy, 1996), ou les méthodes de moindres carrés LSTD(0) (Bradtke & Barto, 1996), LSTD(λ) (Boyan, 1999) et appliqué, avec succès, à les problèmes de contrôle en grande dimension (Lagoudakis & Parr, 2001). Deux approches sont considérées pour l'évaluation de la politique : la **solution du résidu quadratique (RQ)** (le paramètre choisi est celui qui minimise la norme du résidu de Bellman) et la **solution des différences temporelles (DT)** (la fonction paramétrée est le point fixe de l'opérateur de Bellman composé avec l'opérateur de projection sur l'espace des fonctions considérées).

Nous décrivons un algorithme d'*apprentissage par renforcement* (Sutton & Barto, 1998) dans la partie 4 et testons numériquement les deux approches d'approximation précédentes pour un problème de remplacement optimal dans la partie 5.

2 Majoration d'erreur pour l'algorithme IPA

Notons π_k la politique obtenue par l'algorithme IPA à l'itération k. Soit V_k une approximation de la fonction valeur V^{π_k} et π_{k+1} une politique déduite de V_k, i.e. :

$$\pi_{k+1}(x) := \arg\max_{a \in A} \int_X P(x,a,dy)[\gamma V_k(y) + r(x,a,y)].$$

Un résultat de (Munos, 2003) fournit des majorations d'erreur en norme quadratique sur la perte $V^* - V^{\pi_k}$ résultant de l'utilisation de la politique π_k au lieu de la politique optimale, en fonction des **erreurs d'approximation** $V_k - V^{\pi_k}$ ou des **résidus de Bellman** $V_k - T^{\pi_k} V_k$. Nous l'exprimons ici dans le cas espace continu.

Theorème 1 (Munos, 2003) *Définissons les opérateurs stochastiques*

$$\begin{aligned}
S_k &= \frac{(1-\gamma)^2}{2}(I - \gamma P^{\pi^*})^{-1}[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}], \\
\widetilde{S}_k &= \frac{(1-\gamma)^2}{2}(I - \gamma P^{\pi^*})^{-1}[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I + \gamma P^{\pi_k}) + P^{\pi^*}].
\end{aligned}$$

Pour toute mesure positive μ sur $\mathbf{B}$ telle que $\mu(X) = 1$, posons $\mu_k = \mu S_k$ et $\widetilde{\mu}_k = \mu \widetilde{S}_k$. Alors μ_k et $\widetilde{\mu}_k$ sont des mesures positives sur $\mathbf{B}$, telles que $\mu_k(X) = 1$ et $\widetilde{\mu}_k(X) = 1$, et on a

$$\limsup_{k \to \infty} ||V^* - V^{\pi_k}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k \to \infty} ||V_k - T^{\pi_k} V_k||_{\mu_k}, \qquad (1)$$

$$\limsup_{k \to \infty} ||V^* - V^{\pi_k}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k \to \infty} ||V_k - V^{\pi_k}||_{\widetilde{\mu}_k}. \qquad (2)$$

Remarquons que la majoration (2) en norme L_2 est plus fine que la majoration usuelle en norme L_∞ de (Bertsekas & Tsitsiklis, 1996) :

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} ||V_k - V^{\pi_k}||_\infty. \tag{3}$$

3 Approximation linéaire

Dans cette partie nous considérons des approximations linéaires des fonctions-Q, et donnons des hypothèses sur le PDM à partir desquelles nous pouvons déduire des bornes sur les résidus $||V_k - T^{\pi_k} V_k||_{\mu_k}$ et les erreurs d'approximation $||V_k - V^{\pi_k}||_{\widetilde{\mu}_k}$.

3.1 Approximation linéaire à base de *features*

Nous considérons une classe de fonctions $Q_\alpha(x,a) = \sum_{i=1}^K \alpha_i \phi_i(x,a)$ paramétrées linéairement par un paramètre α (vecteur de taille K), où les $(\phi_i)_{1\leq i\leq K}$ sont un ensemble de fonctions, appelées *features*.

De telles architectures linéaires inclues les méthodes d'agrégation d'état, les décompositions polynômiales, de Fourier, sur des bases d'ondelettes et les méthodes aux éléments finis.

Soit $[\Phi]$ l'espace de ces fonctions paramétrées. A l'itération k, l'étape d'évaluation approchée de la politique sélectionne une “bonne” approximation Q_{α_k} (notée Q_k par simplicité) de la fonction Q^{π_k}, au sens où une certaine (semi-) norme $||Q_k - Q^{\pi_k}||_{\rho_k}$ est minimisée. Ici, ρ_k est une distribution sur l'espace état-action $X \times A$, ce qui défini un produit scalaire $\langle f, h\rangle_\rho := \int_{X\times A} \rho(dx,da) f(x,a) h(x,a)$ (cette notation présuppose que A est un espace continu, mais si ce n'est pas le cas, nous avons simplement $\langle f,h\rangle_\rho := \sum_{a\in A} \int_X \rho(dx,a) f(x,a)h(x,a)$) et la semi-norme $||f||_\rho := \langle f,f\rangle^{1/2}$.

Nous décrivons deux approches possibles pour la résolution de ce problème de minimisation (Schoknecht, 2002) :

- Déterminer le paramètre α qui minimise la norme du résidu :

  $$\min_\alpha ||R_\alpha||_{\rho_k}$$

  où $R_\alpha(x,a) := Q_\alpha(x,a) - T^{\pi_k} Q_\alpha(x,a)$ est le **résidu de Bellman**. Cette solution, appelée **solution du résidu quadratique (RQ)**, est obtenue en résolvant un système linéaire de taille K :

  $$A\alpha = b, \text{ avec } \left\{ \begin{array}{ccc} A_{ij} & = & \langle \phi_i - \gamma P^{\pi_k}\phi_i, \phi_j - \gamma P^{\pi_k}\phi_j\rangle_{\rho_k}, \\ b_i & = & \langle \phi_i - \gamma P^{\pi_k}\phi_i, r^{\pi_k}\rangle_{\rho_k}. \end{array} \right. \tag{4}$$

  Ce problème possède toujours une solution lorsque la famille des ϕ_i est libre (puisqu'alors la matrice A est symétrique, définie positive). Remarquons que ce problème peut être considéré comme un problème de régression linéaire avec un ensemble de fonctions de base $\{\psi_i := \phi_i - \gamma P^{\pi_k}\phi_i\}_{i=1..K}$: déterminer α qui minimise $||\sum_{i=1}^K \alpha_i \psi_i - r^{\pi_k}||_{\rho_k}$.
- Déterminer le point fixe de l'opérateur conjugué $\Pi_{\rho_k} T^{\pi_k}$, i.e. l'opérateur de Bellman suivi de l'opérateur de projection sur $[\Phi]$ (selon la norme $||\cdot||_{\rho_k}$), i.e. Q_k satisfait $Q_k = \Pi_{\rho_k} T^{\pi_k} Q_k$. Cette solution, appelée la **solution**

des différences temporelles (DT), est obtenue en résolvant un système linéaire de taille K :

$$A\alpha = b, \text{ avec } \left\{ \begin{array}{lcl} A_{ij} & = & < \phi_i, \phi_j - \gamma P^{\pi_k}\phi_j >_{\rho_k}, \\ b_i & = & < \phi_i, r^{\pi_k} >_{\rho_k} . \end{array} \right. \tag{5}$$

Ici A n'est pas toujours inversible.

Ces méthodes sont appelées des *méthodes de projection* (Judd, 1998) : on cherche α tel que le résidu est orthogonal à un ensemble de fonctions test (les fonctions de base ϕ_i dans le cas DT, les dérivées $\partial_{\alpha_i} R_\alpha$ dans le cas RQ).

La matrice A et le vecteur b peuvent être estimés (voir la partie 4) à partir de données expérimentales sur les transitions provenant de diverses sources, par exemple de manière incrémentale (Boyan, 1999) ou à partir de l'observation de trajectoires induites par différentes politiques (Lagoudakis & Parr, 2001), ou encore à partir de données provenant de connaissances a priori sur les dynamiques d'état.

Définissons la fonction valeur approchée $V_k(x) = Q_k(x, \pi_k(x))$. Les majorations sur l'erreur d'approximation $||V_k - V^{\pi_k}||_{\widetilde{\mu}_k}$ et sur le résidu de Bellman $||V_k - V^{\pi_k}||_{\mu_k}$ sont dérivées en fonction de la **capacité de représentation** $\varepsilon_{[\Phi]}$ **de l'architecture d'approximation** considérée :

$$\varepsilon_{[\Phi]} := \sup_\pi \inf_\alpha ||Q_\alpha - Q^\pi||_{\rho_\pi}. \tag{6}$$

(la distribution ρ_π pouvant dépendre de la politique π).

Nous énonçons les majorations d'erreur sur la perte $V^* - V^{\pi_k}$ pour les solutions RQ (paragraphe 3.2) et DT (paragraphe 3.3) en fonction de $\varepsilon_{[\Phi]}$, sous une hypothèse de stochasticité uniforme.

3.2 Solution du résidu quadratique

Considérons α_k le paramètre qui minimise le résidu de Bellman en norme quadratique ρ_k (solution du système (4)). Nous faisons l'hypothèse suivante sur le PDM.

Hypothèse 1 (Stochasticité uniforme) *Soit $\overline{\mu}$ une mesure positive sur X (avec $\overline{\mu}(X) = 1$), par exemple la mesure de Lebesgue. Nous supposons qu'il existe une constante C telle que pour tous $x \in X$, $a \in A$, $B \in \mathbf{B}$,*

$$P(x, a, B) \leq C\overline{\mu}(B) \tag{7}$$

Cette hypothèse permet de prévenir la masse de la distribution de présence des états après une itération (pour tout contrôle possible) de s'accumuler sur quelques états particuliers. En particulier, les PDM déterministes sont proscrits. Cette hypothèse est vraie lorsque les probabilités de transition admettent une représentation $P^\pi(x, B) = \int_B p^\pi(x, y)dy$ avec une densité bornée $p^\pi(x, \cdot) < \infty$. Une autre hypothèse moins contraignante (et permettant d'établir des majorations similaires) qui couvre les PDMs déterministes est décrite dans (Munos, 2004).

Considérons la distribution $\rho_\pi^\lambda = \overline{\mu}(I - \lambda P^\pi)^{-1}(1 - \lambda)$ (pour $0 \leq \lambda < 1$), qui est construite à partir d'une distribution initiale $\overline{\mu}$ à partir de laquelle des transitions selon π ont lieu pendant un temps qui est une variable aléatoire exponentielle de loi $\lambda^t(1-\lambda)$. Donc ρ_π^λ correspond à la distribution d'une chaîne de Markov qui part d'un état initial tiré selon $\overline{\mu}$ et qui, à chaque instant suit la politique π avec une probabilité λ, ou repart d'un nouvel état (tiré selon $\overline{\mu}$) avec la probabilité $1-\lambda$. Remarquons que lorsque λ tend vers 0 alors ρ_π^λ tend vers la distribution constante $\overline{\mu}$, et lorsque λ tend vers 1, ρ_π^λ tend vers la distribution stationnaire associée à la politique π. Nous avons le résultat suivant.

Theorème 2 (Munos, 2003) *Supposons que l'hypothèse 1 soit satisfaite avec une distribution $\overline{\mu}$ et une constant C, et que la distribution dans (6) est $\rho_\pi = \overline{\mu}(I-\lambda P^\pi)^{-1}(1-\lambda)$ (avec $0 \le \lambda < 1$), alors*

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \le \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{C}{1-\lambda}}\left(1+\gamma\sqrt{\min(\frac{C}{1-\lambda},\frac{1}{\lambda})}\right)\varepsilon_{[\Phi]}. \quad (8)$$

3.3 Solution des différences temporelles

Le système (5) peut être instable puisque, selon la distribution ρ_k (qui défini la norme $||\cdot||_{\rho_k}$ utilisée pour l'opérateur de projection Π_{ρ_k}), les valeurs propres de la matrice A peuvent être proches de 0, auquel cas, l'approximation Q_k de Q^{π_k} risque d'être très mauvaises. Il a été énoncé, à partir de résultats expérimentaux, que les solutions DT fournissaient de meilleure politiques que les solutions RQ bien que les fonctions valeurs n'étaient pas aussi précisément approchées. La raison invoquée était que la solution DT "*preserved the shape of the value function to some extent rather than trying to fit the absolute values*", ainsi "*the improved policy from the approximate value function is "closer" to the improved policy from the corresponding exact value function*" (Lagoudakis & Parr, 2001).

Cependant, dans (Munos, 2003), nous avons fourni un argument montrant que ceci n'est pas vrai en général, et qu'il fallait maîtriser les valeurs propres de l'opérateur $\Pi_{\rho_k}P^{\pi_k}$ pour déduire des majorations sur l'erreur d'approximation $Q_k - Q^{\pi_k}$ (ainsi que sur les formes relatives des fonctions-Q, i.e. les différences entre les erreurs d'approximation pour différentes actions : $(Q_k - Q^{\pi_k})[(\cdot,a) - (\cdot,b)]$). Nous énoncions qu'en général, la solution DT est moins stable que la solution RQ.

Un cas particulier pour lequel nous sommes capable de contrôler les valeurs propres de $\Pi_{\rho_k}P^{\pi_k}$ est lorsque nous utilisons la norme définie par la distribution stationnaire $\overline{\rho}_k$ de la politique π_k, lorsque celle-ci existe (par exemple si la chaîne de Markov induite par π_k est irréductible et apériodique (Puterman, 1994)). Dans ce cas,

$$||V_k - V^{\pi_k}||_{\overline{\rho}_k} \le ||(I-\gamma\Pi_{\overline{\rho}_k}P^{\pi_k})^{-1}||_{\overline{\rho}_k}\varepsilon_{[\Phi]} \le \frac{1}{1-\gamma||P^{\pi_k}||_{\overline{\rho}_k}}\varepsilon_{[\Phi]} \le \frac{\varepsilon_{[\Phi]}}{1-\gamma}$$

(puisque $||P^{\pi_k}||_{\overline{\rho}_k} = 1$ (Tsitsiklis & Van Roy, 1996)). Si $\overline{\rho}_k$ est minorée, nous déduisons la majoration d'erreur suivante sur la perte $V^* - V^{\pi_k}$:

Theorème 3 (Munos, 2003) *Supposons que l'hypothèse 1 soit satisfaite pour une distribution $\overline{\mu}$ et une constante C, et considérons dans (6) la distribution stationnaire $\overline{\rho}_k$ de π_k avec $\overline{\rho}_k$ minorée par $\frac{1}{\kappa}\overline{\mu}$ (pour une constante $\kappa > 0$), alors*

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \le \frac{2\gamma}{(1-\gamma)^3}\sqrt{\kappa C}\,\varepsilon_{[\Phi]}. \quad (9)$$

Il est intéressant de constater que finalement, les inégalités obtenues (8) et (9) expriment une majoration de la perte $V^* - V^{\pi_k}$ en norme L_∞ en fonction d'une norme L_2 de l'erreur d'approximation $\varepsilon_{[\Phi]}$, et que ces résultats ne se déduisent pas de la majoration (3).

4 Apprentissage par Renforcement

Ici nous supposons que les probabilités de transition $P(x, a, B)$ sont inconnues de l'agent : celui-ci dispose seulement d'un *modèle génératif* qui permet de tirer des échantillons selon $P(x, a, \cdot)$ (Sutton & Barto, 1998). La matrice A et le vecteur b dans les systèmes (4) et (5) peuvent être approchés à partir de données expérimentales, comme cela est expliqué dans (Lagoudakis & Parr, 2001) : une base de données D_k est construite par observation de bouts de trajectoires. Les données (x, a, y, r) résultent des transitions : (état x, action a) vers état $y \sim P(x, a, \cdot)$ avec une récompense r. Pour une politique donnée π_k, nous sélectionnons dans D_k, un ensemble de données $\{(x_n, a_n, y_n, r_n)\}_{n=1..N}$ où $a_n = \pi_k(x_n)$. Les états-actions (x_n, a_n) forment une distribution empirique ρ_k.

4.1 Estimateur pour la solution DT

Un estimateur non-biaisé de A et b pour le système (5) est

$$\begin{aligned}\widehat{A}_{ij} &= \frac{1}{N}\sum_{n=1}^{N} \phi_i(x_n, a_n)[\phi_j(x_n, a_n) - \gamma\phi_j(y_n, \pi_k(y_n))],\\ \widehat{b}_i &= \frac{1}{N}\sum_{n=1}^{N} \phi_i(x_n, a_n) r_n.\end{aligned}$$

Quand le nombre de données n tend vers l'infini, la solution $\widehat{\alpha}$ du système approché $\widehat{A}\widehat{\alpha} = \widehat{b}$ tend vers α, la solution du système DT (5) (si celle-ci existe).

4.2 Estimateur pour la solution RQ

Il faut faire attention ici. Nous pourrions penser qu'un estimateur similaire pour le système RQ (4) est :

$$\widehat{A}_{ij} = \frac{1}{N}\sum_{n=1}^{N} [\phi_i(x_n, a_n) - \gamma\phi_i(y_n, \pi_k(y_n))][\phi_j(x_n, a_n) - \gamma\phi_j(y_n, \pi_k(y_n))].$$

Or, cela est faux (et peut peut-être expliquer pourquoi les simulations numériques utilisant cet estimateur ne donnent pas d'aussi bons résultats que ceux de l'estimateur DT, comme indiqué dans (Lagoudakis & Parr, 2001)). En effet, un tel estimateur $\widehat{A}_{ij}$ minimiserait

$$\int_{X\times A} \rho_k(dx, da) \int_X P(x, a, dy) \left(Q_\alpha(x, a) - [\gamma Q_\alpha(y, \pi_k(y)) + r(x, a, y)]\right)^2,$$

au lieu de minimiser

$$\int_{X\times A} \rho_k(dx, da) \left(Q_\alpha(x, a) - \int_X P(x, a, dy)[\gamma Q_\alpha(y, \pi_k(y)) + r(x, a, y)]\right)^2.$$

Commettre cette erreur mènerait à une paramétrisation qui tendrait à réduire la variance des valeurs-Q itérées à un pas. Le problème vient du fait que pour un couple (x,a) donné, le produit $P^\pi\phi_i(x,a)P^\pi\phi_j(x,a) = \mathbb{E}[X_i]\mathbb{E}[X_j]$ (avec X_l (pour $l = i$ ou j) la variable aléatoire $\phi_l(y,\pi_k(y))$, où $y \sim P(x,a,\cdot)$) est différent de $\mathbb{E}[X_iX_j]$ dès que X_i et X_j sont corrélées. Donc, pour estimer $\langle P^\pi\phi_i, P^\pi\phi_j\rangle_{\rho_k}$, nous devons décorréler X_i et X_j. Deux possibilités sont les suivantes :

- Si nous possédons un modèle génératif, pour chaque état-action (x_n,a_n), nous tirons deux échantillons indépendants y_n et y'_n à partir de $P(x_n,a_n,\cdot)$.
- Si nous ne disposons que d'un seul échantillon y_n (tiré selon $P(x_n,a_n,\cdot)$) par état-action (x_n,a_n), nous pouvons considérer le plus proche voisin x'_n de x_n tel que (x'_n,a_n,y'_n,r'_n) est dans la base de données D_k. Cette approximation est valable lorsque $P(\cdot,a,B)$ est régulière et si D_k contient suffisamment de données.

Alors, un estimateur non-biaisé de A et b pour le système RQ (4) est

$$
\begin{aligned}
\widehat{A}_{ij} &= \frac{1}{N}\sum_{n=1}^{N}[\phi_i(x_n,a_n)-\gamma\phi_i(y_n,\pi_k(y_n))][\phi_j(x_n,a_n)-\gamma\phi_j(y'_n,\pi_k(y'_n))],\\
\widehat{b}_i &= \frac{1}{N}\sum_{n=1}^{N}[\phi_i(x_n,a_n)-\gamma\phi_i(y_n,\pi_k(y_n))]\, r_n.
\end{aligned}
$$

5 Problème de remplacement optimal

5.1 Description

Ce problème, décrit dans (Rust, 1996), considère une variable à une dimension $x_t \in \mathbb{R}_+$ qui mesure l'état d'usure d'un bien (tel que le compteur kilométrique d'une voiture par exemple). $x_t = 0$ signifie un bien neuf. A chaque instant discret t, il y a deux décisions possibles : soit *garder* ($a_t = 0$) soit *remplacer* ($a_t = 1$) le bien (ce qui entraîne un coût supplémentaire C de vente du bien actuel et du rachat d'un nouveau bien). Nous considérons la densité de transition exponentielle suivante

$$
\begin{aligned}
p(x,a=0,y) &= \begin{cases} \beta e^{-\beta(y-x)} & \text{si } y \geq x \\ 0 & \text{si } y < x \end{cases}\\
p(x,a=1,y) &= \begin{cases} \beta e^{-\beta y} & \text{si } y \geq 0 \\ 0 & \text{si } y < 0 \end{cases}
\end{aligned}
$$

et la fonction récompense (opposée d'un coût) : $r(x,a=0,y) = -c(x)$ (où nous supposons que $c(x)$ est une fonction croissante) et $r(x,a=1,y) = -C-c(0)$. La fonction valeur optimale satisfait l'équation de la programmation dynamique :

$$V^*(x) = \max\{-c(x)+\gamma\int_x^\infty p(x,0,y)V^*(y)dy, -C-c(0)+\gamma\int_0^\infty p(x,1,y)V^*(y)dy\}.$$

Ici, il est possible d'écrire explicitement sa valeur :

$$V^*(x) = \begin{cases} \int_x^{\overline{x}} \frac{c'(y)}{1-\gamma}(1-\gamma e^{-\beta(1-\gamma)(y-x)})dy - \frac{c(\overline{x})}{1-\gamma} & \text{si} \quad x \leq \overline{x} \\ \frac{-c(\overline{x})}{1-\gamma} & \text{si} \quad x > \overline{x} \end{cases}$$

où $\overline{x}$ est l'unique solution de

$$C = \int_0^{\overline{x}} \frac{c'(y)}{1-\gamma}(1-\gamma e^{-\beta(1-\gamma)y})dy.$$

La politique optimale est $\pi^*(x) = 0$ si $x \in [0, \overline{x}]$, et $\pi^*(x) = 1$ si $x > \overline{x}$.

5.2 Résultats numériques

Nous avons choisi les valeurs numériques suivantes $\gamma = 0.6$, $\beta = 0.5$, $C = 30$, $c(x) = 4x$. Donc ici $\overline{x} \simeq 4.8665$ et la fonction valeur optimale est

$$V^*(x) = \begin{cases} -10x + 30(e^{0.2(x-\overline{x})} - 1) & \text{pour} \quad x \le \overline{x} \\ -10\overline{x} & \text{pour} \quad x > \overline{x} \end{cases}$$

(représentée sur la figure 1).

Considérons les 8 fonctions de base suivantes (polynômes de degré ≤ 3) : $\phi_i(x,0) = x^{i-1}$, $\phi_i(x,1) = 0$, pour $i \in \{1, \cdots, 4\}$, et $\phi_i(x,0) = 0$, $\phi_i(x,1) = x^{i-5}$, pour $i \in \{5, \cdots, 8\}$.

Pour chaque politique π_k, nous construisons une base de données $D_k = \{(x_n, a_n, y_n, r_n)\}_{n=1..N}$ en utilisant la distribution $\rho_k^{\lambda,\epsilon}$ sur $X \times A$ (pour $0 \le \lambda < 1$ et $0 \le \epsilon \le 1$), définie selon le protocole suivant :

- x_0 est tiré selon une distribution uniforme $\overline{\mu}$ (sur $[0, 10]$).
- Pour chaque x_n, l'action a_n est choisie selon la politique stochastique suivante : l'action $\pi_k(x_n)$ est choisie avec une probabilité $1-\epsilon$, et une action (uniformément) aléatoire avec une probabilité ϵ. L'état suivant y_n est tiré selon $P(x_n, a_n, \cdot)$.
- L'état suivant x_{n+1} est y_n avec une probabilité $1-\lambda$, ou réinitialisé (tiré selon $\overline{\mu}$) avec une probabilité λ.

Nous pouvons écrire $\rho_\pi^{\lambda,\epsilon} = \overline{\mu}[I - \lambda((1-\epsilon)P^\pi + \epsilon\frac{1}{|A|}\sum_{a\in A} P^a)]^{-1}(1-\lambda)$.

Nous remarquons que l'hypothèse 1 est satisfaite : la densité de transition $p(x,a,y)$ étant majorée. Donc, le théorème 2 s'applique à la solution RQ.

Dans ces simulations, nous portons notre attention sur le domaine $[0, 10]$. La figure 1 montre la fonction valeur optimale et les fonctions-Q approchées pour les deux actions possibles lors des premières itérations de l'algorithme, pour les solutions RQ et DT, pour des distributions $\rho_\pi^{\lambda,\epsilon}$ avec $\lambda = 0.5$ et $\epsilon = 0.1$ échantillonnées avec $N = 1000$ points.

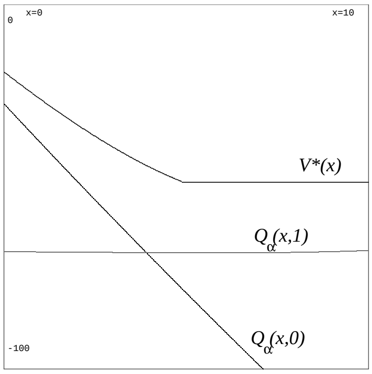
Une phase stationnaire est atteinte au bout de 4 itérations (après cela, il n'y a plus de décroissance de l'erreur $||V_k - V^{\pi_k}||_{\rho^{\lambda,\epsilon}_{\pi_k}}$). Les erreur quadratiques pour les solutions RQ au cours des premières itérations sont : $||V_1 - V^{\pi_1}||_{\rho_1} = 18.272$, $||V_2 - V^{\pi_2}||_{\rho_2} = 0.275$, $||V_3 - V^{\pi_3}||_{\rho_3} = 0.164$, $||V_4 - V^{\pi_4}||_{\rho_4} = 0.285$.

Celles obtenues pour les solutions DT sont : $||V_1 - V^{\pi_1}||_{\rho_1} = 18.980$, $||V_2 - V^{\pi_2}||_{\rho_2} = 0.399$, $||V_3 - V^{\pi_3}||_{\rho_3} = 0.252$, $||V_4 - V^{\pi_4}||_{\rho_4} = 0.159$.

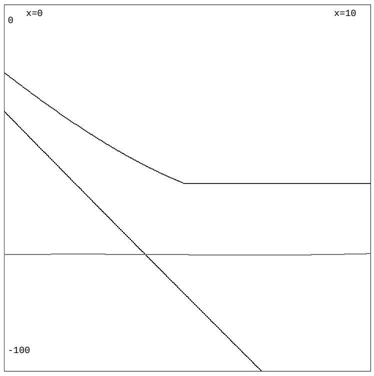
Ainsi, les erreurs d'approximation obtenues par les solutions DT et RQ sont comparables et relativement proches de 0 : dans les deux cas, la fonction valeur est bien approchée.

Maintenant si nous considérons les mêmes simulations mais avec une distribution $\rho_\pi^{\lambda,\epsilon}$ avec $\lambda = 0.5$ et $\epsilon = 0$ (i.e. il n'y a pas d'exploration : la politique déterministe π_k est toujours utilisée), alors, à la seconde itération, la matrice DT a des valeurs propres très proches de 0, ce qui produit une divergence des

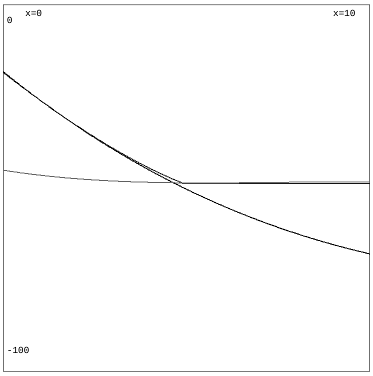
FIG. 1 – Premières itérations pour les systèmes RQ et DT. Ici $\epsilon = 0.1$. Ordre de lecture : (a), (c), (e), puis (b), (d), (f).



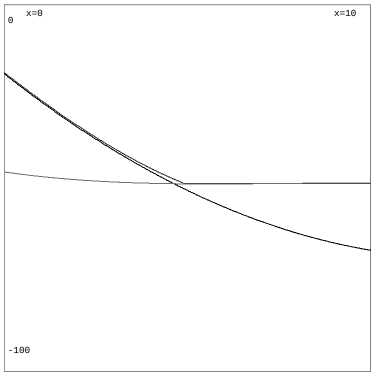
(a) RQ itération 1



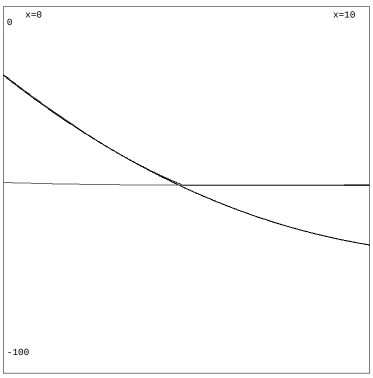
(b) DT itération 1



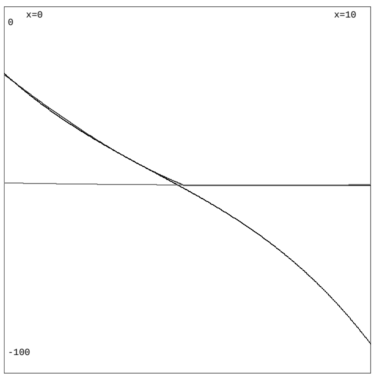
(c) RQ itération 2



(d) DT itération 2

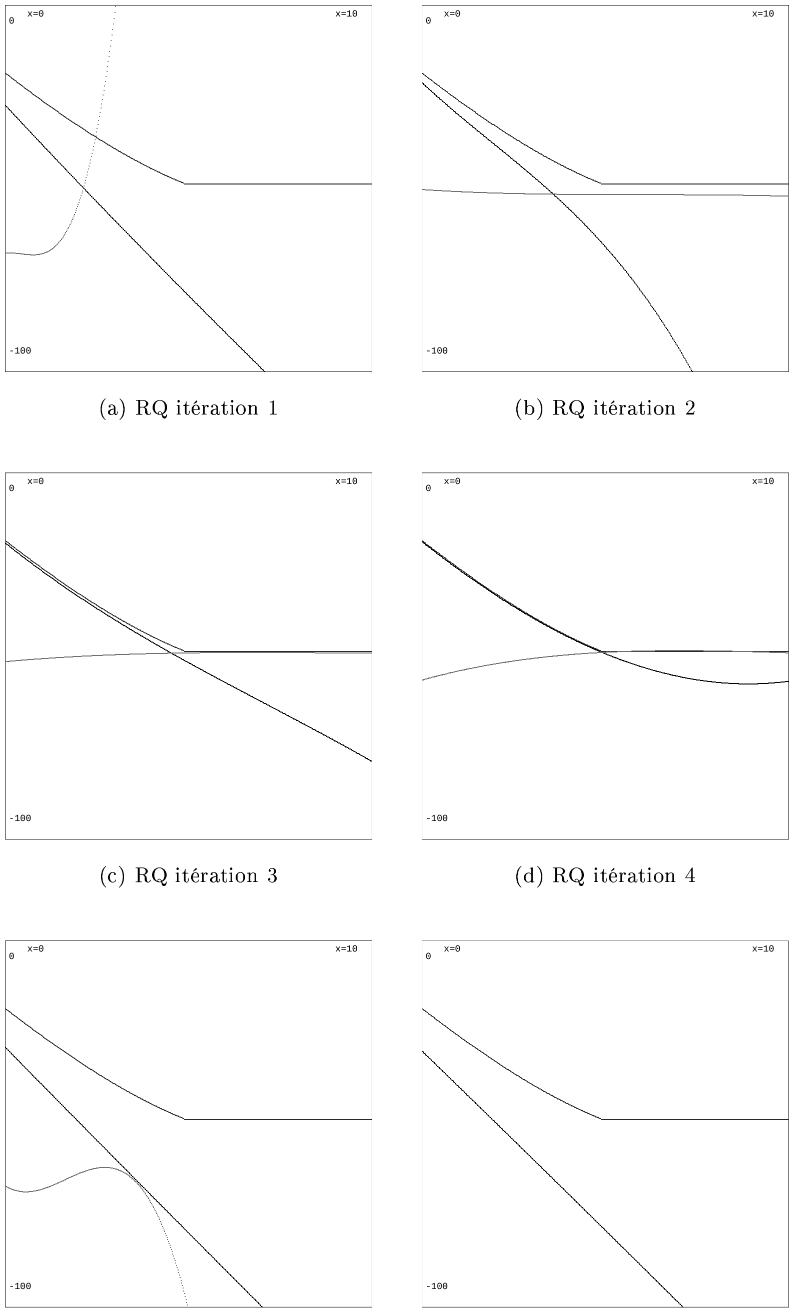


(e) RQ itération 3



(f) DT itération 3

FIG. 2 – Premières itérations pour les systèmes RQ et DT. Ici $\epsilon = 0$. Ordre de lecture : (a), (b), (c), (d), puis (e), (f).



(a) RQ itération 1

(b) RQ itération 2

(c) RQ itération 3

(d) RQ itération 4

(e) DT itération 1

(f) DT itération 2

paramètres successifs α_k. Le système RQ n'est pas affecté pas ce problème. Ces simulations sont illustrées sur la figure 2.

Cette instabilité est fréquemment observée lorsqu'on utilise la solution DT avec une distribution qui n'est proche ni de la distribution stationnaire ni d'une distribution uniforme.

6 Conclusion

Au cours de ces simulations numériques simples, nous avons observé que l'algorithme d'itération sur les politiques avec approximation linéaire utilisant les solutions RQ et DT fournissent des résultats très similaires lorsque les deux méthodes s'appliquent. Construire un estimateur non-biaisé pour le système RQ est plus difficile que pour le système DT d'un point de vue computationnel puisqu'il nécessite 2 échantillons indépendants pour chaque donnée (si nous disposons d'un modèle génératif). Cependant il apparaît que la solution RQ est plus robuste et s'applique à davantage de situations que la solution DT, qui se montre instable pour certaines distributions.

References

Baird, L. C. (1995). Residual algorithms : Reinforcement learning with function approximation. *Machine Learning : proceedings of the Twelfth International Conference.*

Bertsekas, D. P., & Tsitsiklis, J. (1996). *Neuro-dynamic programming.* Athena Scientific.

Boyan, J. (1999). Least-squares temporal difference learning. *Proceedings of the 16th International Conference on Machine Learning*, 49–56.

Bradtke, S., & Barto, A. (1996). Linear least-squares algorithms for temporal difference learning. *Journal of Machine Learning*, *22*, 33–57.

Judd, K. (1998). *Numerical methods in economics.* MIT Press.

Lagoudakis, M., & Parr, R. (2001). Model free least-squares policy iteration. *Technical Report CS-2001-05, Department of Computer Science, Duke University.*

Munos, R. (2003). Error bounds for approximate policy iteration. *19th International Conference on Machine Learning.*

Munos, R. (2004). Error bounds for approximate value iteration. *Technical report N° 527, CMAP, Ecole Polytechnique.*

Puterman, M. L. (1994). *Markov decision processes, discrete stochastic dynamic programming.* A Wiley-Interscience Publication.

Rust, J. (1996). *Numerical dynamic programming in economics.* In Handbook of Computational Economics. Elsevier, North Holland.

Schoknecht, R. (2002). Optimality of reinforcement learning algorithms with linear function approximation. *Proceedings of the 15th Neural Information Processing Systems conference.*

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning : An introduction. *Bradford Book.*

Tsitsiklis, J., & Van Roy, B. (1996). An analysis of temporal difference learning with function approximation. *Technical report LIDS-P-2322, MIT.*