Analyse en norme Lp de l'algorithme d'itérations sur les valeurs avec approximations

Rémi Munos

Analyse en norme L_p de l'algorithme d'itérations sur les valeurs avec approximations

Rémi Munos

Projet SequeL, INRIA Futurs
Grappa, Université Lille 3, 59653 Villeneuve d'Ascq Cedex
remi.munos@inria.fr

RÉSUMÉ. *L'algorithme d'*itérations sur les valeurs avec approximations *(IVA) permet de résoudre des problèmes de décision markoviens en grande dimension en approchant la fonction valeur optimale par une séquence de représentations V_n calculées itérativement selon $V_{n+1} = \mathcal{A}\mathcal{T}V_n$ où $\mathcal{T}$ est l'*opérateur de Bellman *et $\mathcal{A}$ un* opérateur d'approximation*, ce dernier pouvant s'implémenter selon un algorithme* d'apprentissage supervisé *(AS). Les résultats usuels établissent des bornes sur la performance de IVA en fonction de la norme L_∞ des erreurs d'approximation induites par l'algorithme d'AS. Cependant, un algorithme d'AS résout généralement un problème de régression en minimisation une norme L_p ($p \geq 1$), rendant les majorations d'erreur en L_∞ inadéquates. Dans cet article, nous étendons ces résultats de majoration à des normes L_p pondérées. Ceci permet d'exprimer les performances de l'algorithme IVA en fonction de la puissance d'approximation de l'algorithme d'AS, ce qui garantit la finesse et l'intérêt applicatif de ces bornes. Nous illustrons numériquement la qualité des majorations obtenues pour un problème de remplacement optimal.*

ABSTRACT. Approximate Value Iteration *(AVI) is a method for solving a large Markov Decision Problem by approximating the optimal value function with a sequence of value representations V_n processed by means of the iterations $V_{n+1} = \mathcal{A}\mathcal{T}V_n$ where $\mathcal{T}$ is the so-called* Bellman operator *and $\mathcal{A}$ an* approximation operator*, which may be implemented by a* Supervised Learning *(SL) algorithm. Previous results relate the asymptotic performance of AVI to the L_∞-norm of the approximation errors induced by the SL algorithm. Unfortunately, the SL algorithm usually perform a minimization problem in L_p-norms ($p \geq 1$), rendering the L_∞ performance bounds inadequate. In this paper, we extend these performance bounds to weighted L_p-norms. This enables to relate the performance of AVI to the approximation power of the SL algorithm, which guarantees the tightness and pratical interest of these bounds. We illustrate the tightness of the bounds on an optimal replacement problem.*

MOTS-CLÉS : *processus de décision markovien, programmation dynamique avec approximation.*

KEYWORDS: *Markov decision process, approximate dynamic programming.*

RSTI - RIA – 21/2007. Prise de décision séquentielle, pages 55 à 76

1. Introduction

Nous considérons la résolution d'un *processus de décision markovien* (PDM) (Puterman, 1994) en grande dimension dans le cas actualisé avec horizon temporel infini.

L'algorithme d'*itérations sur les valeurs* consiste à calculer la fonction valeur optimale V^* par évaluation successive de fonctions V_n selon le schéma d'itération $V_{n+1} = \mathcal{T}V_n$, où $\mathcal{T}$ est l'*opérateur de Bellman*. Grâce à une propriété de contraction -en norme L_∞- de l'opérateur $\mathcal{T}$, les itérés V_n convergent vers V^* lorsque $n \to \infty$. Cependant, cette méthode est inapplicable lorsque le nombre d'états est tel qu'une représentation exacte des valeurs est impossible à mémoriser. Nous devons alors représenter les fonctions à l'aide d'un nombre modéré de coefficients et considérer une résolution approchée de la solution ; ce qui nous mène à définir l'algorithme d'**itérations sur les valeurs avec approximation** (IVA).

IVA est très populaire et est depuis longtemps implémenté de diverses manières en *programmation dynamique* (PD) (Samuel, 1959, Bellman *et al.,* 1959) et plus récemment dans le contexte de l'*apprentissage par renforcement* et la *PD avec approximation* (Bertsekas *et al.,* 1996, Sutton *et al.,* 1998). IVA construit une séquence de représentations V_n calculées itérativement selon

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n, \qquad [1]$$

où $\mathcal{T}$ est l'*opérateur de Bellman* et $\mathcal{A}$ un *opérat*eur *d'approximation*, lequel peut être implémenté par un algorithme d'*apprentissage supervisé* (AS) (voir *e.g.* (Hastie *et al.,* 2001)).

Puisque nous allons utiliser plusieurs normes, nous rappelons leur définition maintenant. Soit $u \in \mathbb{R}^N$. Sa norme L_∞ est $||u||_\infty := \sup_{1\leq i\leq N} |u(i)|$. Soit μ une mesure de probabilité sur l'espace fini $\{1, \ldots, N\}$. La (semi) norme L_p (pour $p \geq 1$) pondérée par le poids μ (celle-ci est notée $L_{p,\mu}$) est définie par $||u||_{p,\mu} := \left[\sum_x \mu(x)|u(x)|^p\right]^{1/p}$. De plus, lorsque μ est uniforme, nous notons $||\cdot||_p$ la norme L_p non pondérée.

Une implémentation typique de IVA est un algorithme combinant des itérations sur les valeurs alternées par des étapes de régression : pour un espace fonctionnel donné $\mathcal{F}$, on définit à chaque étape une nouvelle représentation $V_{n+1} \in \mathcal{F}$ en projetant sur $\mathcal{F}$ l'image par l'opérateur de Bellman de l'estimation courante V_n. A titre d'illustration, une version échantillonnée de cet algorithme serait la suivante : à l'étape n, on choisit K états $(x_k)_{1\leq k\leq K}$ tirés de manière indépendante selon une certaine distribution μ sur l'espace d'états, on calcule les valeur itérées par l'opérateur de Bellman $\{v_k := \mathcal{T}V_n(x_k)\}$, puis on fait appel à un algorithme d'AS avec les données d'apprentissage $\{(x_k, v_k)\}_{1\leq k\leq K}$ (les $\{x_k\}$ étant les entrées et $\{v_k\}$ les sorties désirées). Ce dernier renvoie une fonction V_{n+1} qui minimise une erreur empirique (définie par l'algorithme d'AS considéré), par exemple

$$V_{n+1} = \arg\min_{g\in\mathcal{F}} \frac{1}{K} \sum_{1\leq k\leq K} l(g(x_k) - v_k),$$

où la fonction l est habituellement une fonction quadratique ou valeur absolue (ou des variantes, comme la perte ϵ-insensible utilisée dans les SVMs (Vapnik, 1998)).

Il s'agit d'une version échantillonnée d'un problème de minimisation en norme pondérée (par μ) quadratique ou absolue (notées $L_{p,\mu}$, où $p = 2$ or 1 respectivement), et des majorations sur l'écart entre l'erreur empirique minimale et la norme $L_{p,\mu}$ de l'erreur d'approximation $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$ peuvent s'établir en fonction du nombre d'échantillons K utilisé et la complexité (ou capacité) de l'espace $\mathcal{F}$ (habituellement caractérisée par le *nombre de couverture* ou la *dimension de Vapnik-Chervonenkis* de $\mathcal{F}$, voir (Pollard, 1984, Vapnik, 1998) par exemple).

Il est donc naturel de chercher à établir des majorations sur les performances de IVA en termes de normes L_p ($p \geq 1$) des erreurs d'approximation $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$. Malheureusement, l'analyse usuelle en PD utilise principalement la norme L_∞ (Bertsekas *et al.,* 1996, Gordon, 1999). Par exemple, la performance asymptotique des politiques déduites des représentations calculées par IVA peut être majorée par la norme L_∞ des erreurs d'approximation $||V_{n+1} - \mathcal{T}V_n||_\infty$ (voir la section 2). Cependant, cette majoration n'est pas très informative puisque l'erreur d'approximation L_∞ est difficile à contrôler et peu utile en pratique puisque la plupart des algorithmes d'AS connus résolvent un problème de minimisation en norme L_p. Puisque la plupart des opérateurs d'approximation et algorithmes d'AS fournissent de bonnes régressions en minimisant une norme L_p, il apparaît donc essentiel de pouvoir analyser la performance de l'algorithme IVA en utilisant cette même norme.

L'objectif de cet article est de fournir des majorations en norme L_p de la performance de l'algorithme IVA, permettant ainsi d'évaluer cette performance en fonction de la puissance d'approximation de l'algorithme d'AS utilisé.

Pour commencer, mentionnons que, bien entendu, les normes sont équivalentes en dimension finie (ce qui est le cas ici puisque l'on considère un espace d'états fini). Ainsi $||\cdot||_p \leq ||\cdot||_\infty \leq N^{1/p}||\cdot||_p$ (pour des normes non pondérées, avec N étant le nombre d'états), et la borne usuelle L_∞ pour IVA décrite à la section 2 peut être utilisée pour en déduire une borne L_p. Cependant, le terme $N^{1/p}$ (très grand pour des problèmes de grande taille) mène à des majorations très mauvaises.

L'article suit le plan suivant. A la section 2, nous rappelons les résultats de majoration usuels en norme L_∞. La section 3 fait un rapide survol de techniques d'approximation et d'algorithmes d'AS. L'outil principal développé dans cet article est l'obtention de majorations composante par composante pour IVA ; celles-ci sont détaillées à la section 4. Les résultats de majoration sur la performance en norme L_p sont énoncés à la section 5 et le résultat principal de cet article est énoncé par le théorème 1. Un paragraphe fournit une intuition de ces résultats dans le cas particulier où l'algorithme IVA converge, ce qui mène à des bornes exprimées en fonction du résidu de Bellman. L'extension au cas d'un espace continu est considéré à la section 6 et une expérimentation numérique traitant un problème de remplacement optimal est détaillée.

Préliminaires

Nous décrivons maintenant le cadre considéré ici des processus de décision markoviens (PDM) en horizon temporel infini avec récompenses actualisées.

Soit X l'espace d'états, supposé fini, où N désigne le nombre d'états, et A l'espace d'actions, supposé fini aussi. Les résultats énoncés dans cet article s'étendent au cas d'espaces infinis (soit dénombrables, soit continus, comme illustré à la section 6). Notons $p(x, a, y)$ la probabilité de transition vers l'état suivant y lorsque l'état courant est x et l'action choisie a, et $r(x, a, y)$ la récompense reçue lors de la transition correspondante.

Nous appelons une *politique (markovienne* ou *stationnaire)* π une application de X dans A. Notons P^π la matrice carrée de taille N dont les éléments sont $P^\pi(x, y) := p(x, \pi(x), y)$ et r^π le vecteur de taille N de composantes $r^\pi(x) := \sum_y p(x, \pi(x), y) r(x, \pi(x), y)$.

La performance d'une politique donnée π est évaluée par la *fonction valeur* V^π associée (considérée comme un vecteur de taille N), définie par l'espérance de la somme des récompenses actualisées à venir lorsqu'on suit la politique π :

$$V^\pi(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r(x_t, a_t, x_{t+1}) | x_0 = x, a_t = \pi(x_t)\right],$$

où $\gamma \in [0, 1)$ est appelé le *coefficient d'actualisation*. Il est bien connu que V^π est le point fixe de l'opérateur $\mathcal{T}^\pi : \mathbb{R}^N \to \mathbb{R}^N$ défini, pour tout vecteur $W \in \mathbb{R}^N$, par $\mathcal{T}^\pi W := r^\pi + \gamma P^\pi W$.

La *fonction valeur optimale* $V^* := \sup_\pi V^\pi$ est le point fixe de l'opérateur de Bellman $\mathcal{T}$ défini, pour tout $W \in \mathbb{R}^N$, $x \in X$, par

$$\mathcal{T}W(x) := \max_{a \in A} \sum_y p(x, a, y)[r(x, a, y) + \gamma W(y)].$$

On dit qu'une politique π est *déduite* de $W \in \mathbb{R}^N$ si, pour tout $x \in X$,

$$\pi(x) \in \arg\max_{a \in A} \sum_y p(x, a, y)[r(x, a, y) + \gamma W(y)].$$

L'objectif du PDM consiste à déterminer une politique optimale π^*, c'est-à-dire telle qu'en tout $x \in X$, $V^{\pi^*}(x) = \max_\pi V^\pi(x)$. Il est facile de montrer que toute politique déduite de V^* est optimale.

2. Performance de l'algorithme IVA en norme L_∞

Considérons l'**algorithme IVA** défini par l'itération [1] et notons

$$\varepsilon_n := \mathcal{T}V_n - V_{n+1} \in \mathbb{R}^N \qquad [2]$$

l'**erreur d'approximation** à l'étape n. En général, IVA ne converge pas, mais néanmoins son comportement asymptotique peut être analysé. Si les erreurs d'approximation sont uniformément majorées $||\varepsilon_n||_\infty \leq \varepsilon$, alors une borne sur la différence entre la performance asymptotique des politiques π_n déduites des approximations V_n et la performance de la politique optimale est (voir par exemple (Bertsekas *et al.,* 1996)) :

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\varepsilon. \quad [3]$$

Puisque la preuve est très simple, elle est rappelée ici.

Preuve. D'après l'inégalité triangulaire, la propriété de contraction (d'un facteur γ) des opérateurs de Bellman $\mathcal{T}$ et $\mathcal{T}^{\pi_n}$, et du fait que π_n est déduite de V_n (*i.e.* $\mathcal{T}^{\pi_n}V_n = \mathcal{T}V_n$), nous avons :

$$\begin{aligned} ||V^* - V^{\pi_n}||_\infty &\leq ||\mathcal{T}V^* - \mathcal{T}^{\pi_n}V_n||_\infty + ||\mathcal{T}^{\pi_n}V_n - \mathcal{T}^{\pi_n}V^{\pi_n}||_\infty \\ &\leq \gamma||V^* - V_n||_\infty + \gamma(||V_n - V^*||_\infty + ||V^* - V^{\pi_n}||_\infty), \end{aligned}$$

donc :

$$||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{1-\gamma}||V^* - V_n||_\infty. \quad [4]$$

De plus, $||V^* - V_{n+1}||_\infty \leq ||\mathcal{T}V^* - \mathcal{T}V_n||_\infty + ||\mathcal{T}V_n - V_{n+1}||_\infty \leq \gamma||V^* - V_n||_\infty + \varepsilon$. Maintenant, en passant à la limite supérieure, il vient $\limsup_{n\to\infty} ||V^* - V_n||_\infty \leq \varepsilon/(1-\gamma)$, ce qui, combiné à [4] mène à [3].□

Cette borne L_∞ s'exprime en fonction d'une erreur d'approximation uniforme sur tout le domaine, qui est en général difficile à dominer, particulièrement pour des problèmes de grande taille. De plus, elle n'est pas très utile en pratique puisque la plupart des opérateurs d'approximation et algorithmes d'AS résolvent un problème de minimisation en norme L_1 ou L_2 (bien que certains approximateurs de fonction utilisant la norme L_∞, tels les *averagers*, aient été étudiés dans le cadre de la PD (Gordon, 1995, Guestrin *et al.,* 2001)).

3. Opérateurs d'approximation et algorithmes d'apprentissage supervisé

Dans cette section nous faisons un survol du problème de l'approximation de fonction dans le contexte de l'apprentissage statistique (voir par exemple (Hastie *et al.,* 2001)). A titre d'illustration, considérons un algorithme d'AS prenant pour entrée des données $\{(x_k, v_k)\}_{1\leq k\leq K}$, où les états $x_k \in X$ sont tirés de manière indépendante selon une distribution μ sur X et les valeurs v_k sont des réalisations non biaisées d'une variable aléatoire (dépendant de x_k) de moyenne (inconnue) $f(x_k)$, et retournant une

fonction g (dans une classe de fonctions donnée $\mathcal{F}$) minimisant une perte empirique ; par exemple g est solution du problème de minimisation :

$$\inf_{g \in \mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} l(v_k - g(x_k)),$$

où la fonction perte l est une fonction absolue ou quadratique. Si les valeurs v_k ne sont pas bruitées (*i.e.* $v_k = f(x_k)$), il s'agit simplement d'une opération d'approximation qui retourne une représentation compacte $g \in \mathcal{F}$ d'une fonction inconnue f en minimisant une certaine norme L_p empirique (ici $p = 1$ or 2) à partir des données. Il s'agit donc d'une version échantillonnée d'un problème de minimisation en norme pondérée $L_{p,\mu}$, et des majorations sur l'erreur d'approximation $||g - f||_{p,\mu}$ peuvent être déduites en fonction du nombre d'échantillons K et de la puissance d'approximation (ou capacité) de l'espace $\mathcal{F}$, caractérisée par des quantités telles que le *nombre de couverture* ou la *dimension de Vapnik-Chervonenkis (Pollard, 1984, Vapnik, 1998).*

L'approximation linéaire consiste à réaliser une projection sur un espace vectoriel engendré par une famille finie de fonctions, et inclut les décompositions sur des Splines*,* fonctions radiales, bases de Fourier ou ondelettes. Souvent, une meilleure approximation est obtenue lorsque la famille de fonctions est choisie selon f. Cette *approximation non linéaire* est particulièrement efficace quand f possède des régularités locales (par exemple, dans des bases d'ondelettes adaptatives (Mallat, 1997) de telles fonctions sont représentées de manière compacte avec peu de coefficients non nuls). Des algorithmes dits *greedy* (par exemple le *Matching Pursuit* et diverses variantes (Davies *et al.,* 1997)) sélectionnent les meilleures fonctions de base dans un dictionnaire donné. La théorie de l'approximation étudie les erreurs d'approximations en fonction de la régularité de f (DeVore, 1997).

En apprentissage statistique (Hastie *et al.,* 2001), des exemples algorithmes d'AS incluent les *Réseaux de neurones, l'apprentissage localement pondéré, la régression par noyaux* (Atkeson *et al.,* 1997), *les Support-Vectors* et les *méthodes à noyaux* dans les espaces de Hilbert à noyau reproduisant (Vapnik *et al.,* 1997, Vapnik, 1998).

Dans la suite de cet article, nous ne ferons pas la distinction entre apprentissage supervisé et approximation de fonction puisque cela dépend de la méthode particulière utilisée pour résoudre le problème de minimisation. Nous dirons que $\mathcal{A}$ retourne une **approximation** $g = \mathcal{A}f$ à $\epsilon-$près de f (dans une certaine norme $||\cdot||$) si $||f - g|| \leq \varepsilon$.

4. Majorations composante par composante

Dans cette section, nous établissons des bornes composante par composante, à partir desquelles, les majorations L_p seront déduites à la prochaine section. La borne L_∞ précédemment énoncée [3] est aussi une conséquence immédiate de ces bornes par composantes.

4.1. *Borne par composantes sur la performance de l'algorithme IVA*

Une majoration par composantes sur la performance asymptotique des politiques π_n déduites de V_n est établie dans le lemme ci-dessous.

Lemme 1 *Considérons l'algorithme IVA défini par [1] et notons $\varepsilon_n = \mathcal{T}V_n - V_{n+1} \in \mathbb{R}^N$ l'erreur d'approximation à l'étape n. Soit π_n une politique déduite de V_n. Alors :*

$$\limsup_{n\to\infty} V^* - V^{\pi_n} \leq \limsup_{n\to\infty} (I - \gamma P^{\pi_n})^{-1} \qquad [5]$$
$$\Big(\sum_{k=0}^{n-1} \gamma^{n-k} \big[(P^{\pi^*})^{n-k} + P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_{k+2}} P^{\pi_{k+1}} \big] |\varepsilon_k| \Big),$$

où $|\varepsilon_k|$ désigne le vecteur composé des valeurs absolues des composantes de ε_k.

Afin de prouver ce lemme, nous énonçons le résultat préliminaire suivant :

Lemme 2 *Soit A une matrice inversible dont tous les éléments de son inverse sont positifs. Alors, les solutions de l'inégalité $Au \leq b$ sont aussi solutions de $u \leq A^{-1}b$.*

Preuve du Lemme 2. Soit u une solution de $Au \leq b$. Cela signifie qu'il existe un vecteur c de composantes positives tel $Au = b - c$, donc $u = A^{-1}b - A^{-1}c$. Puisque tous les éléments de $A^{-1}c$ sont positifs, nous en déduisons que $u \leq A^{-1}b$. □

Preuve du Lemme 1. A partir de la définition de $\mathcal{T}$ et $\mathcal{T}^\pi$, nous avons l'inégalité composante par composante : $\mathcal{T}V_k \geq \mathcal{T}^{\pi^*}V_k$ et $\mathcal{T}V^* \geq \mathcal{T}^{\pi_k}V^*$, donc

$$\begin{aligned} V^* - V_{k+1} &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_k + \mathcal{T}^{\pi^*}V_k - \mathcal{T}V_k + \varepsilon_k \leq \gamma P^{\pi^*}(V^* - V_k) + \varepsilon_k \\ V^* - V_{k+1} &= \mathcal{T}V^* - \mathcal{T}^{\pi_k}V^* + \mathcal{T}^{\pi_k}V^* - \mathcal{T}V_k + \varepsilon_k \geq \gamma P^{\pi_k}(V^* - V_k) + \varepsilon_k, \end{aligned}$$

où, à la seconde ligne, nous avons utilisé le fait que π_k est déduite de V_k, *i.e.* $\mathcal{T}^{\pi_k}V_k = \mathcal{T}V_k$. Nous en déduisons par récurrence,

$$V^* - V_n \leq \sum_{k=0}^{n-1} \gamma^{n-k-1} (P^{\pi^*})^{n-k-1} \varepsilon_k + \gamma^n (P^{\pi^*})^n (V^* - V_0), \qquad [6]$$

$$\begin{aligned} V^* - V_n \geq & \sum_{k=0}^{n-1} \gamma^{n-k-1} (P^{\pi_{n-1}} P^{\pi_{n-2}} \dots P^{\pi_{k+1}}) \varepsilon_k \\ & + \gamma^n (P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_1})(V^* - V_0). \end{aligned} \qquad [7]$$

Maintenant, en utilisant à nouveau la définition de π_n et le fait que $\mathcal{T}V_n \geq \mathcal{T}^{\pi^*}V_n$, nous avons :

$$\begin{aligned} V^* - V^{\pi_n} &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_n + \mathcal{T}^{\pi^*}V_n - \mathcal{T}V_n + \mathcal{T}V_n - \mathcal{T}^{\pi_n}V^{\pi_n} \\ &\leq \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_n + \mathcal{T}V_n - \mathcal{T}^{\pi_n}V^{\pi_n} \\ &= \gamma P^{\pi^*}(V^* - V_n) + \gamma P^{\pi_n}(V_n - V^{\pi_n}) \\ &= \gamma P^{\pi^*}(V^* - V_n) + \gamma P^{\pi_n}(V_n - V^* + V^* - V^{\pi_n}), \end{aligned}$$

donc $(I - \gamma P^{\pi_n})(V^* - V^{\pi_n}) \leq \gamma(P^{\pi^*} - P^{\pi_n})(V^* - V_n)$. Et puisque $(I - \gamma P^{\pi_n})$ est inversible et son inverse $\sum_{k\geq 0}(\gamma P^{\pi_n})^k$ a tous ses éléments positifs, nous utilisons le Lemme 2 pour en déduire que :

$$V^* - V^{\pi_n} \leq \gamma(I - \gamma P^{\pi_n})^{-1}(P^{\pi^*} - P^{\pi_n})(V^* - V_n).$$

Ceci, combiné à [6] et [7], et après avoir pris la valeur absolue (remarquons que le vecteur $V^* - V^{\pi_n}$ est positif), nous mène à

$$\begin{aligned} V^* - V^{\pi_n} \leq\ & (I - \gamma P^{\pi_n})^{-1} \\ & \Big\{ \sum_{k=0}^{n-1} \gamma^{n-k} \big[(P^{\pi^*})^{n-k} + (P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_{k+1}}) \big] |\varepsilon_k| \qquad [8] \\ & + \gamma^{n+1} \big[(P^{\pi^*})^{n+1} + (P^{\pi_n} P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_1}) \big] |V^* - V_0| \Big\}. \end{aligned}$$

Nous en déduisons [5] en prenant la limite supérieure. □

4.2. *Borne sur la performance en fonction du résidu de Bellman*

Dans cette section, nous établissons une borne composante par composante sur la performance relative (par rapport à l'optimale) d'une politique π déduite d'une fonction $V \in \mathbb{R}^N$ quelconque en fonction du résidu de Bellman de V. Ce résultat étend la majoration usuelle en norme L_∞ (Williams *et al.,* 1993) :

$$||V^* - V^\pi||_\infty \leq \frac{2}{1-\gamma} ||\mathcal{T}V - V||_\infty. \qquad [9]$$

La majoration analogue, par composantes, est énoncée maintenant.

Lemme 3 *Soit $V \in \mathbb{R}^N$ et π une politique déduite de V. Alors*

$$V^* - V^\pi \leq \Big[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^\pi)^{-1} \Big] |\mathcal{T}V - V|. \qquad [10]$$

Nous remarquons immédiatement que [9] se déduit de ce résultat puisque, pour toute matrice stochastique P, $||(I-\gamma P)^{-1}||_\infty = 1/(1-\gamma)$.

Preuve du Lemme 3. Nous utilisons le fait que $\mathcal{T}V \geq \mathcal{T}^{\pi^*}V$ et la définition de π (*i.e.* $\mathcal{T}V = \mathcal{T}^{\pi}V$) pour en déduire

$$\begin{aligned} V^* - V^\pi &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V + \mathcal{T}^{\pi^*}V - \mathcal{T}V + \mathcal{T}V - \mathcal{T}^\pi V^\pi \\ &\leq \gamma P^{\pi^*}(V^* - V^\pi + V^\pi - V) + \gamma P^\pi(V - V^\pi), \end{aligned}$$

et donc $(I-\gamma P^{\pi^*})(V^* - V^\pi) \leq \gamma(P^{\pi^*} - P^\pi)(V^\pi - V)$. A nouveau, puisque $(I-\gamma P^{\pi^*})$ est inversible et son inverse a tous ses éléments positifs, nous déduisons, d'après le Lemme 2 que

$$V^* - V^\pi \leq \gamma(I-\gamma P^{\pi^*})^{-1}(P^{\pi^*} - P^\pi)(V^\pi - V).$$

De plus,

$$\begin{aligned} (I-\gamma P^\pi)(V^\pi - V) &= V^\pi - V - \gamma P^\pi V^\pi + \gamma P^\pi V \\ &= r^\pi + \gamma P^\pi V - (r^\pi + \gamma P^\pi V^\pi) + V^\pi - V \\ &= \mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi + V^\pi - V = \mathcal{T}V - V, \end{aligned}$$

donc

$$\begin{aligned} V^* - V^\pi &\leq \gamma(I-\gamma P^{\pi^*})^{-1}(P^{\pi^*} - P^\pi)(I-\gamma P^\pi)^{-1}(\mathcal{T}V - V) \\ &= (I-\gamma P^{\pi^*})^{-1}\Big[(I-\gamma P^\pi) - (I-\gamma P^{\pi^*})\Big](I-\gamma P^\pi)^{-1}(\mathcal{T}V - V) \\ &= \Big[(I-\gamma P^{\pi^*})^{-1} - (I-\gamma P^\pi)^{-1}\Big](\mathcal{T}V - V) \\ &\leq \Big[(I-\gamma P^{\pi^*})^{-1} + (I-\gamma P^\pi)^{-1}\Big]|\mathcal{T}V - V|. \qquad \square \end{aligned}$$

5. Majorations en norme L_p

Nous généralisons ici les résultats en norme L_∞ à des normes L_p. La principale intuition qui rend cette extension possible est simple et réside dans les résultats de majoration par composantes établis précédemment.

Considérons deux vecteurs u et v à composantes positives, tels que l'on ait inégalité vectorielle $u \leq Qv$, avec Q une matrice stochastique. Bien sûr, ceci implique que $||u||_\infty \leq ||v||_\infty$ (puisque $||Q||_\infty = 1$), mais de plus, si ν et μ sont des mesures de probabilité sur X telles que $\nu Q \leq \mu$ (où νQ est le produit matriciel de Q avec la mesure ν définie en tant que vecteur ligne), alors nous déduisons aussi que $||u||_{p,\nu} \leq ||v||_{p,\mu}$.

En effet, nous avons

$$\begin{aligned} ||u||_{p,\nu}^p = \sum_{x\in X} \nu(x)|u(x)|^p &\leq \sum_{x\in X} \nu(x)\big[\sum_{y\in X} Q(x,y)v(y)\big]^p \\ &\leq \sum_{x\in X} \nu(x) \sum_{y\in X} Q(x,y)v(y)^p \\ &\leq \sum_{y\in X} \mu(y)v(y)^p = ||v||_{p,\mu}^p, \end{aligned}$$

en utilisant l'inégalité de Jensen.

On peut directement appliquer cette idée aux bornes définies en fonction du résidu de Bellman. Avec $u = V^* - V^\pi$, $v = \frac{2}{1-\gamma}|\mathcal{T}V - V|$, la borne par composantes [10] permet de déduire la borne L_∞ [9], et permet aussi de déduire des bornes L_p pour des mesures de pondération ν et μ appropriés (voir la section 5.3 ci-dessous). Cette idée s'applique aussi pour l'algorithme d'IVA. La borne par composantes [5] permet de déduire la borne L_∞ [3] et rend aussi possible l'extension à des bornes L_p.

5.1. *Définition des constantes de régularité*

Nous définissons maintenant les constantes de régularité $C(\mu)$, $C_1(\nu,\mu)$, et $C_2(\nu,\mu)$, qui dépendent du PDM, sous lesquelles les mesures ν et μ peuvent être comparées. Soient ν et μ deux mesures de probabilité sur X.

Définition 1 *Nous appelons $C(\mu) \in \mathbb{R}^+ \cup \{+\infty\}$ la **constante de régularité des probabilités de transition**, définie par la plus petite constante C telle que pour tous $x \in X$, $y \in X$, $a \in A$,*

$$p(x,a,y) \leq C\mu(y),$$

(s'il n'existe pas de telle constante, nous posons $C(\mu) = \infty$).

Pour tout entier $m \geq 1$, nous définissons $c(m) \in \mathbb{R}^+ \cup \{+\infty\}$ la plus petite constante C telle que, pour toute séquence de m politiques $\pi_1, \pi_2, \ldots, \pi_m$,

$$\nu P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m} \leq C\mu, \qquad [11]$$

(à nouveau, $c(m) = \infty$ s'il n'existe pas de telle constante) et notons $c(0) := 1$. Remarquons que ces constantes dépendent de ν et μ.

*Nous appelons $C_1(\nu,\mu) \in \mathbb{R}^+ \cup \{+\infty\}$ (resp. $C_2(\nu,\mu)$) la **constante de régularité de la distribution d'états futurs de premier (resp. second) ordre d'actualisation** :*

$$C_1(\nu,\mu) := (1-\gamma)\sum_{m\geq 0} \gamma^m c(m), \qquad [12]$$

$$C_2(\nu,\mu) := (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c(m). \qquad [13]$$

Remarquons que nous sommes intéressés par les situations où les constantes $C(\mu)$, $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$ sont finies, ce qui est garanti dès que la mesure μ est strictement positive.

La constante de régularité des probabilités de transition $C(\mu)$ est introduite dans (Munos, 2003) pour établir des bornes en norme L_2 sur les performances de l'algorithme d'itérations sur les politiques. Alors que $C(\mu)$ fournit une information sur les transitions immédiates du PDM, $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$ décrivent une propriété de régularité relative (par rapport à μ) de la distribution d'états futurs sachant que l'état initial est tiré selon ν. De manière informelle, la distribution d'états futurs est la répartition dans l'espace d'états de la fréquence de visite des états lorsque l'on suit une séquence de politiques. La distribution de premier ordre considère un coefficient d'actualisation de γ^m (où m est l'horizon temporel), celle de second ordre utilise un autre facteur de pondération de $(m+1)\gamma^m$. En d'autres termes, pour n'importe quelle séquence de politiques, $\pi_1,\ldots,\pi_m$, la distribution d'états futurs (de premier et second ordre) partant de ν et utilisant cette séquence de politiques (*i.e.* $\{x_i \sim p(x_{i-1},\pi_i(x_{i-1}),\cdot)\}_{1\leq i\leq m}$) est majorée par ces constantes ($C_1(\nu,\mu)$ et $C_2(\nu,\mu)$) fois μ : pour tous x_0, y de X,

$$(1-\gamma)\sum_{m\geq 0}\gamma^m \Pr\big(x_m = y \big| x_0\sim\nu,\pi_1,\ldots,\pi_m\big) \quad\leq\quad C_1(\nu,\mu)\mu(y),$$

$$(1-\gamma)^2\sum_{m\geq 1} m\gamma^{m-1} \Pr\big(x_m = y \big| x_0\sim\nu,\pi_1,\ldots,\pi_m\big) \quad\leq\quad C_2(\nu,\mu)\mu(y).$$

5.2. *Majorations en norme L_p pour l'algorithme IVA*

Le résultat suivant établit des bornes sur la performance de l'algorithme d'IVA en fonction de la norme $L_{p,\mu}$ des erreurs d'approximation $\varepsilon_n = V_{n+1} - \mathcal{T}V_n$.

Théorème 1 *Soient μ et ν deux mesures de probabilité sur X. Considérons l'algorithme IVA défini par [1]. Notons π_n une politique déduite de V_n et $\varepsilon_n = V_{n+1} - \mathcal{T}V_n \in I\!R^N$ l'erreur d'approximation à chaque étape. Soit $\varepsilon > 0$ et supposons que $\mathcal{A}$ retourne une approximation V_{n+1} à ε-près, en norme $L_{p,\mu}$ (avec $p \geq 1$), de $\mathcal{T}V_n$,* i.e. *$||\varepsilon_n||_{p,\mu} \leq \varepsilon$, pour tout $n \geq 0$. Alors :*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \quad\leq\quad \frac{2\gamma}{(1-\gamma)^2}\big[C(\mu)\big]^{1/p}\varepsilon, \qquad [14]$$

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{p,\nu} \quad\leq\quad \frac{2\gamma}{(1-\gamma)^2}\big[C_2(\nu,\mu)\big]^{1/p}\varepsilon. \qquad [15]$$

Remarquons que les parties droites de ces inégalités utilisent une norme L_p pondérée et que la partie gauche du premier résultat [14] utilise une norme L_∞ alors que la partie gauche du deuxième résultat [15] est en norme L_p. Le premier résultat ne dépend pas de la distribution ν et peut directement être comparé à la borne L_∞ [3].

Preuve du Théorème 1. Remarquons que la constante $C(\mu)$ est toujours plus grande que $C_2(\nu, \mu)$ pour toute distribution ν. En effet, pour tout $m \geq 1$, $c(m) \leq C(\mu)$. Donc $C_2(\nu, \mu) \leq (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} C(\mu) = C(\mu)$. Donc, si [15] a lieu pour tout ν, choisir comme ν un Dirac respectivement en chaque état implique [14]. Ainsi, nous n'avons qu'à montrer [15]. Nous pouvons réécrire [8] selon

$$V^* - V^{\pi_n} \leq \frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2} \Big[\sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k| + \alpha_n A_n |V^* - V_0| \Big],$$

avec les coefficients positifs $\{\alpha_k\}_{0\leq k\leq n}$

$$\begin{aligned} \alpha_k &:= \frac{(1-\gamma)\gamma^{n-k-1}}{1-\gamma^{n+1}}, \text{ pour } 0 \leq k < n \\ \text{et } \alpha_n &:= \frac{(1-\gamma)\gamma^n}{1-\gamma^{n+1}}, \end{aligned}$$

(remarquons que $\sum_{k=0}^{n} \alpha_k = 1$), et les matrices stochastiques $\{A_k\}_{0\leq k\leq n}$:

$$\begin{aligned} A_k &:= \frac{1-\gamma}{2} (I - \gamma P^{\pi_n})^{-1} \big[(P^{\pi^*})^{n-k} + (P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_{k+1}}) \big], \text{ pour } 0 \leq k < n \\ A_n &:= \frac{1-\gamma}{2} (I - \gamma P^{\pi_n})^{-1} \big[(P^{\pi^*})^{n+1} + (P^{\pi_n} P^{\pi_n} \dots P^{\pi_1}) \big]. \end{aligned}$$

Puisque les vecteurs des deux côtés de cette majoration ont des composantes positives, nous pouvons prendre leur norme $L_{p,\nu}$ tout en conservant l'inégalité :

$$\begin{aligned} & ||V^* - V^{\pi_n}||_{p,\nu}^p \\ & \quad \leq \Big[\frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2} \Big]^p \sum_{x\in X} \nu \Big[\sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k| + \alpha_n A_n |V^* - V_0| \Big]^p \\ & \quad \leq \Big[\frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2} \Big]^p \sum_{x\in X} \nu \Big[\sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k|^p + \alpha_n A_n |V^* - V_0|^p \Big], \quad [16] \end{aligned}$$

en utilisant deux fois l'inégalité de Jensen (puisque les coefficients $\{\alpha_k\}_{0\leq k\leq n}$ sont de somme 1 et que les matrices A_k sont stochastiques) (*i.e.* convexité of $x \to |x|^p$). Le second terme dans les crochets disparaît quand on passe à la limite supérieure. Et

d'après la définition des coefficients $c(m)$, $\nu A_k \leq (1-\gamma)\sum_{m\geq 0}\gamma^m c(m+n-k)\mu$, ainsi, le premier terme de [16] vérifie

$$\begin{aligned}\sum_x \nu \sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k|^p &\leq \sum_{k=0}^{n-1} \alpha_k (1-\gamma) \sum_{m\geq 0} \gamma^m c(m+n-k) ||\varepsilon_k||_{p,\mu}^p \\ &\leq \frac{(1-\gamma)^2}{1-\gamma^{n+1}} \sum_{m\geq 0}\sum_{k=0}^{n-1} \gamma^{m+n-k-1} c(m+n-k)\varepsilon^p \\ &\leq \frac{1}{1-\gamma^{n+1}} C_2(\nu,\mu)\varepsilon^p,\end{aligned}$$

où l'on a remplacé les α_k par leur valeur et utilisé le fait que $||\varepsilon_k||_{p,\mu} \leq \varepsilon$. En prenant la limite supérieure dans [16], nous en déduisons [15]. □

Que se passe-t'il lorsque IVA converge ?

D'un point de vue théorique, il n'y a pas de raison pour que l'algorithme d'IVA converge. Cependant, dans certaines applications, on observe expérimentalement que l'algorithme converge. Il est intéressant de remarquer que dans ces cas, nous pouvons déduire des majorations plus fines (pour $\gamma > 1/2$). En effet, la convergence d'IVA signifie qu'il existe $V \in \mathbb{R}^N$ tel que $\lim_{n\to\infty} V_n = V$. Donc, en passant à la limite dans [1], on déduit que V est un point fixe de l'opérateur $\mathcal{AT}$, *i.e.* $V = \mathcal{AT}V$, et que l'erreur d'approximation [2] tend vers le résidu $\mathcal{T}V - V$ de V.

Ainsi, la performance asymptotique de IVA est la performance de la politique π déduite de V, et peut donc être majorée en fonction du résidu $\mathcal{T}V - V$. Il se trouve que les bornes établies en fonction du résidu de Bellman (la borne en norme L_∞ [9] ou la borne par composantes [10]), utilisent un coefficient $2/(1-\gamma)$ au lieu de $2\gamma/(1-\gamma)^2$ (pour les bornes IVA), fournissant une majoration plus fine dès que $\gamma > 1/2$. Le prochain paragraphe étend les majorations basées sur le résidu de Bellman aux normes L_p.

5.3. *Majorations en norme L_p en fonction du résidu de Bellman*

Ici, nous établissons une majoration de la performance d'une politique π déduite de V (où $V \in \mathbb{R}^N$) en fonction de la normes $L_{p,\mu}$ de son résidu $\mathcal{T}V - V$.

Théorème 2 *Soit V un vecteur de taille N et π une politique déduite de V. Soient μ et ν deux distributions de probabilités sur X. Alors :*

$$||V^* - V^\pi||_\infty \leq \frac{2}{(1-\gamma)} \left[C(\mu)\right]^{1/p} ||\mathcal{T}V - V||_{p,\mu}, \quad [17]$$

$$||V^* - V^\pi||_{p,\nu} \leq \frac{2}{(1-\gamma)} \left[C_1(\nu,\mu)\right]^{1/p} ||\mathcal{T}V - V||_{p,\mu}. \quad [18]$$

A nouveau, le premier résultat [17] fournit une majoration L_∞ sur la performance, qui peut être directement comparé à la borne L_∞ [9], alors que le second résultat [18] utilise la norme L_p.

Preuve du Théorème 2. Nous pouvons réécrire [10] sous la forme :

$$V^* - V^\pi \leq \frac{2}{1-\gamma} A|\mathcal{T}V - V|,$$

où A est la matrice stochastique

$$A = \frac{1-\gamma}{2}\Big[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^\pi)^{-1}\Big].$$

En utilisant l'idée décrite dans l'introduction de ce chapitre, nous déduisons :

$$\begin{aligned} ||V^* - V^\pi||_{p,\nu}^p &\leq \Big[\frac{2}{1-\gamma}\Big]^p \sum_{x\in X} \nu(x)\Big[A|\mathcal{T}V - V|\Big]^p(x) \\ &\leq \Big[\frac{2}{1-\gamma}\Big]^p \sum_{x\in X} \nu(x)\Big[A|\mathcal{T}V - V|^p\Big](x), \end{aligned} \quad [19]$$

d'après l'inégalité de Jensen. Maintenant, utilisant la définition des coefficients $c(m)$, $\nu A \leq (1-\gamma)\sum_{m\geq 0}\gamma^m c(m)\mu = C_1(\nu,\mu)\mu$, donc :

$$\begin{aligned} ||V^* - V^\pi||_{p,\nu}^p &\leq \Big[\frac{2}{1-\gamma}\Big]^p C_1(\nu,\mu)\mu|\mathcal{T}V - V|^p \\ &= \Big[\frac{2}{1-\gamma}\Big]^p C_1(\nu,\mu)||\mathcal{T}V - V||_{p,\mu}^p, \end{aligned}$$

qui démontre [18]. Maintenant, puisque $C(\mu) \geq C_1(\nu,\mu)$ pour tout ν, choisir pour ν un Dirac respectivement en chaque état permet de déduire [17]. □

De manière intuitive, la composante $A(x,y)$ de la matrice A indique une majoration sur la contribution du résidu de Bellman en y à l'erreur de performance en l'état x. En effet,

$$V^*(x) - V^\pi(x) \leq \frac{2}{1-\gamma}\sum_{y\in X} A(x,y)|\mathcal{T}V - V|(y).$$

Il est clair d'après [19] que si nous choisissons $\mu = \nu A$, alors la majoration L_p devient :

$$||V^* - V^\pi||_{p,\nu} \leq \frac{2}{(1-\gamma)}||\mathcal{T}V - V||_{p,\mu}. \quad [20]$$

Cette majoration peut nous inspirer un algorithme pour approcher V^* dans un certain espace fonctionnel donné $\mathcal{F}$ en résolvant directement le problème de minimisation du résidu de Bellman dans $\mathcal{F}$:

$$\min_{V \in \mathcal{F}} ||\mathcal{T}V - V||^p_{p,\mu},$$

où la distribution μ dépend de V à travers la politique π déduite de V, *i.e.* $\mu = \nu A = \frac{1-\gamma}{2}\nu\Big[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^{\pi})^{-1}\Big]$. Notons $\mu = (\mu^\pi + \mu^*)/2$ avec $\mu^\pi = (1-\gamma)\nu(I - \gamma P^\pi)^{-1}$ étant la distribution actualisée d'états futurs partant de ν et suivant la politique π, et $\mu^* = (1-\gamma)\nu(I - \gamma P^{\pi^*})^{-1}$ la distribution analogue mais suivant la politique optimale π^*. La norme $L_{p,\mu}$ du résidu à minimiser possède deux contributions :

$$||\mathcal{T}V - V||^p_{p,\mu} = \frac{1}{2}\Big(||\mathcal{T}V - V||^p_{p,\mu^\pi} + ||\mathcal{T}V - V||^p_{p,\mu^*}\Big). \qquad [21]$$

Nous pouvons considérer une méthode d'optimisation itérative, telle qu'une méthode de gradient, où, à chaque itération, un résidu empirique est calculé et minimisé. Minimiser le premier terme dans [21] peut se réaliser facilement en échantillonnant des états selon μ^π (ce qui peut s'implémenter en choisissant un état initial $x \sim \nu$ puis en effectuant des transitions selon la politique courante π pendant un horizon temporel défini par une variable aléatoire exponentielle de coefficient γ). Le second terme est plus difficile à implémenter car il n'est pas facile d'échantillonner des états selon μ^* puisque π^* est inconnue. A la place de μ^*, on peut considérer une distribution proche d'une uniforme, ou utiliser une distribution d'états futurs obtenue en suivant une politique exploratrice, où chaque action est choisie avec une probabilité non nulle.

5.4. *Intuition à propos des constantes* $C(\mu)$, $C_1(\nu,\mu)$, *et* $C_2(\nu,\mu)$

Donnons quelque intuition concernant ces constantes lorsque l'on considère une distribution uniforme $\mu = (\frac{1}{N} \dots \frac{1}{N})$. Dans ce cas, d'après sa définition, $C(\mu)$ est toujours plus petite que le nombre d'états N. $C(\mu)$ est égale à N s'il existe au moins une transition déterministe (*i.e.* il existe $x, y \in X$, $a \in A$, tels que $p(x,a,y) = 1$). Dans ce cas, la borne L_p (disons pour $p = 1$) [14] n'est pas meilleure que la majoration L_∞ [3] combinée avec le simple résultat de comparaison des normes $||\cdot||_\infty \leq N||\cdot||_1$.

Ainsi, la borne L_p [14] (respectivement [17]) est plus fine que celle en norme L_∞ [3] (resp. [9]) dès que la constante de régularité $C(\mu)$ est plus petite que le nombre d'états. Un cas intéressant pour lequel ceci a lieu est le cas d'un espace d'états continu, lorsque le noyau de transition admet une densité par rapport à μ, auquel cas $C(\mu)$ est la borne supérieure de cette densité. Le cas continu sera brièvement abordé à la section 6 et illustré sur un problème de remplacement optimal.

Maintenant, considérons les constantes $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$.

– La plus grande valeur de ces constantes est obtenue dans un PDM où pour une certaine politique π, tous les états transitent vers un état particulier – disons l'état 1–

avec probabilité 1. Ainsi, pour tout ν et tout m, on a $\nu(P^\pi)^m = (1\,0\,\ldots\,0) \leq c(m)\mu$ pour $c(m) = N$ (avec égalité pour l'état 1), donc $C_1(\nu,\mu) = C_2(\nu,\mu) = N$. Il s'agit du pire cas car la masse de distribution des états futurs s'accumule sur un seul état. Dans ce cas, la borne L_p [15] (respectivement [18]) peut en fait se déduire de celle L_∞ [3] (resp. [9]) puisque $||\cdot||_p \leq ||\cdot||_\infty$ et $||\cdot||_\infty \leq N^{1/p}||\cdot||_p$.

– La plus petite valeur possible de ces constantes est obtenue dans un PDM où les probabilités de transition sont uniformes $p(x,a,y) = 1/N$, pour tous $x, y \in X$ et $a \in A$. Quand ν et μ sont toutes les deux uniformes, alors $c(m) = 1$ et $C_1(\nu,\mu) = C_2(\nu,\mu) = 1$ (Il s'agit bien de la plus petite valeur possible puisque pour ν uniforme et pour toute matrice stochastique P, nous avons $\max_y \sum_x \nu(x)P(x,y) \geq 1/N$).

Remarquons cependant qu'un PDM déterministe ne définit pas nécessairement des grandes valeurs des constantes $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$ (contrairement au cas de la constante $C(\mu)$). En effet, dans un PDM n'ayant que des politiques consistant à réaliser des permutations d'états (pour lesquelles chaque état dispose d'un unique successeur et d'un unique prédécesseur), on a $C(\mu) = N$ puisque les transitions sont déterministes (voir précédemment), mais on a $C_1(\nu,\mu) = C_2(\nu,\mu) = 1$ pour des distributions ν et μ uniformes (puisque pour tout $m \geq 0$, $c(m) = 1$). Un autre exemple pour lequel les constantes de régularité des distributions d'états futurs sont faibles (et indépendantes du nombre d'états N) est le PDM « chaînon » décrit au prochain paragraphe.

Ainsi, les constantes de régularité $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$ expriment l'accumulation (relativement à μ) maximale (*i.e.* pour tout politique) de la distribution d'états futurs (avec prise en compte d'un facteur d'actualisation du premier ou deuxième ordre) sachant que l'état initial est tiré selon ν. Une valeur faible de ces constantes signifie que la masse de la distribution d'états futurs (partant de ν) ne s'accumule pas sur des états pour lesquels μ est faible. Ainsi, en l'absence de connaissance de la répartition d'états futurs pour toute politique, afin d'obtenir des valeurs raisonnables de ces constantes, il est désirable de choisir μ proche d'une distribution uniforme (cette condition a déjà été mentionnée dans des travaux précédents (Koller *et al.,* 2000, Kakade *et al.,* 2002, Munos, 2003) afin de garantir une certaine stabilité des étapes d'amélioration de la politique dans l'algorithme d'itérations sur les politiques avec approximations).

5.5. *Illustration sur le PDM* chaînon

Nous illustrons le fait que la borne en norme L_p [15] établie au Théorème 1 est plus fine que celle en norme L_∞ [3] (combinée avec l'inégalité $||\cdot||_\infty \leq N^{1/p}||\cdot||_p$) sur l'exemple du PDM *chaînon* défini dans (Lagoudakis *et al.,* 2003) (voir la figure 1). Il s'agit d'un exemple où la constante $C(\mu)$ est élevée (sa valeur est N, le nombre d'états) mais où les constantes $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$ sont petites (et indépendantes de N).

La chaîne est constituée de N états avec deux états terminaux : les états 1 et N. Dans chaque état intérieur $2 \leq x \leq N-1$ il y a deux actions possibles : droite

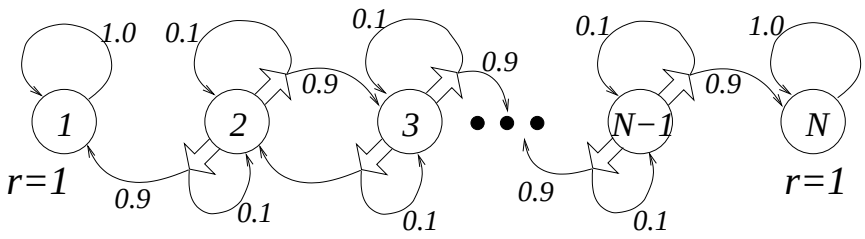

Figure 1. *Le PDM chaînon*

et gauche, qui ont pour effet de déplacer l'état dans la direction correspondante avec une probabilité 0.9, ou de rester sur place avec une probabilité 0.1. La récompense dépend uniquement de l'état courant et vaut 1 aux états terminaux et 0 ailleurs : $r = (1\,0\ldots0\,1)'$.

Nous considérons une approximation de la fonction valeur dans l'espace $\mathcal{F} := \{V(x) = \alpha + \beta x\}_{\alpha\in\mathbb{R},\beta\in\mathbb{R}}$ où $x \in \{1,\ldots,N\}$ est l'indice de l'état. L'approximation initiale est zéro : $V_0 = (0\ldots0)'$. Ainsi $\mathcal{T}V_0 = (1\,0\ldots0\,1)'$. La meilleure approximation de $\mathcal{T}V_0$ dans $\mathcal{F}$ au sens de la minimisation de la norme L_∞ est la fonction constante $V_1 = (\frac{1}{2}\ldots\frac{1}{2})'$ qui fournit l'erreur d'approximation $||V_1 - \mathcal{T}V_0||_\infty = \frac{1}{2}$.

Choisissons des distributions uniformes : $\nu = \mu = (\frac{1}{N}\ldots\frac{1}{N})$. En norme L_1, la meilleure approximation est $V_1 = (0\ldots0)'$ (pour $N > 4$) qui fournit l'erreur $||V_1 - \mathcal{T}V_0||_1 = \frac{2}{N}$. En norme L_2, il s'agit de $V_1 = (\frac{2}{N}\ldots\frac{2}{N})'$ avec l'erreur $||V_1 - \mathcal{T}V_0||_2 = \frac{\sqrt{2N-4}}{N}$.

Dans ces trois cas, nous pouvons montrer par récurrence que les approximations successives V_n sont constantes, donc $\mathcal{T}V_n = r + \gamma V_n$ et les erreurs d'approximation successives sont les mêmes qu'à la première itération : pour tout $n \geq 0$, $||V_{n+1} - \mathcal{T}V_n||_\infty = \frac{1}{2}$, $||V_{n+1} - \mathcal{T}V_n||_1 = \frac{2}{N}$, et $||V_{n+1} - \mathcal{T}V_n||_2 = \frac{\sqrt{2N-4}}{N}$.

Puisque V_n est constante, toute politique π_n est déduite de V_n. Donc, pour $\pi_n = \pi^*$ les parties droites de [3] et [15] sont égales à zéro. Maintenant, afin de comparer leur parties gauches, nous désirons calculer les constantes $C(\mu)$, $C_1(\nu,\mu)$ et $C_2(\nu,\mu)$.

Puisque l'état 1 transite vers lui-même de manière déterministe, nous n'avons pas de meilleure valeur que $C(\mu) = N$.

Maintenant, le pire cas dans [11] est obtenu lorsque la masse de la distribution d'états futurs se concentre sur un état particulier – par exemple l'état 1 – ce qui correspond à une politique π_{Left} qui choisit l'action gauche en tout état. Nous observons alors que pour $\nu = \mu$,

$$\nu(P^{\pi_{\mathrm{Left}}})^m(x) \leq \nu(P^{\pi_{\mathrm{Left}}})^m(1) \leq (1 + 0.9m)\mu(x),$$

pour tout $x \geq 0$, donc $c(m) \leq 1 + 0.9m$. Nous en déduisons que les constantes $C_1(\nu,\mu) \leq (1-\gamma)\sum_{m\geq0}\gamma^m(1+0.9m)$ et $C_2(\nu,\mu) \leq (1-\gamma)^2\sum_{m\geq1} m\gamma^{m-1}(1+$

$0.9m$) sont majorées par des quantités qui sont **indépendantes du nombre d'états** N.

Donc, si nous considérons les performances de IVA en norme L_1, la borne [15] (pour $p = 1$) fournit une approximation d'ordre $O(N^{-1})$, alors que la borne L_1 qui pourrait se déduire du résultat en L_∞ [3] combiné à la comparaison de normes $||\cdot||_\infty \leq N||\cdot||_1$ fournirait une approximation d'ordre $O(1)$ seulement.

De manière similaire, la borne en norme L_2 établit une majoration d'ordre $O(N^{-1/2})$, alors que la borne L_∞ [3] combinée avec $||\cdot||_\infty \leq N^{1/2}||\cdot||_2$ ne donne qu'une approximation en $O(1)$.

Ainsi, lorsque notre algorithme d'apprentissage supervisé retourne des fonctions qui minimisent une erreur d'approximation en norme L_p (ce qui est habituellement le cas), **la borne [15] est arbitrairement plus informative que [3] pour des grandes valeurs de N.**

6. Application numérique en domaine continu

Tous les résultats précédents se généralisent au cas d'un espace d'états continu. Pour illustration, nous commençons par redéfinir brièvement les constantes de régularité dans ce contexte et appliquons l'algorithme IVA à un problème de remplacement optimal, pour lequel la constante $C(\mu)$ est explicitement calculée.

Notons $P(x, a, B)$ le noyau de transition du PDM, où B est un sous-ensemble mesurable de X. Pour une politique stationnaire, $\pi : X \to A$, nous notons $P^\pi(x, B) = P(x, \pi(x), B)$, qui définit un opérateur linéaire à droite (défini sur l'espace des fonctions V bornées mesurables sur X) : $P^\pi V(x) = \int_X V(y)P^\pi(x, dy)$, et un opérateur linéaire à gauche (défini sur l'espace des mesures de probabilité μ sur X) : $\mu P^\pi(B) = \int_X P^\pi(x, B)\mu(dx)$. Le produit de deux noyaux P^{π_1} et P^{π_2} étant défini par $P^{\pi_1}P^{\pi_2}(x, B) = \int_X P^{\pi_1}(x, dy)P^{\pi_2}(y, B)$.

6.1. *Constantes de régularité*

Les constantes de régularité sont définies ainsi : soient ν et μ deux mesures de probabilité sur X. La constante de régularité du noyau de transition $C(\mu)$ est la plus petite constante C telle que pour tout $x \in X$, tout sous-ensemble mesurable B de X, tout $a \in A$, $p(x, a, B) \leq C\mu(B)$. Si le noyau de probabilité $P(x, a, B)$ admet une densité par rapport à la mesure μ, alors la constante $C(\mu)$ est la borne supérieure de cette densité. Ce cas est illustré dans l'exemple numérique ci-dessous.

Les constantes de régularité de la distribution actualisée des états futurs $C_1(\nu, \mu)$ et $C_2(\nu, \mu)$ sont définies similairement à partir de [12] et [13].

6.2. *Problème de remplacement optimal*

Cette simulation numérique illustre les finesses respectives des bornes L_∞, L_1, et L_2 obtenues lors de l'application d'un algorithme d'IVA pour résoudre un problème de contrôle optimal en domaine continu, extrait de (Rust, 1996).

Une variable monodimensionnelle $x_t \in [0, x_{\max}]$ mesure l'utilisation accumulée d'un certain produit (par exemple le compteur kilométrique d'une voiture). $x_t = 0$ désigne un produit tout neuf. A chaque instant discret t, il y a deux décisions possibles : soit conserver ($a_t = \mathrm{K}$), soit remplacer ($a_t = \mathrm{R}$) le produit, auquel cas, un coût supplémentaire $C_{replace}$ (de vente du produit courant suivi du rachat d'un nouveau bien) est perçu. Les transitions suivent une loi exponentielle de paramètre β avec une queue tronquée : si l'état suivant y est plus grand qu'une valeur maximale fixée $x_{\max}$ (par exemple un état critique d'usure de la voiture) alors un nouvel état est immédiatement tiré et une pénalité $C_{dead} > C_{replace}$ est reçue. Les densités de transition sont donc définies ainsi : en notant $q(x) := \beta e^{-\beta x}/(1 - e^{-\beta x_{\max}})$,

$$p(x, a = \mathrm{R}, y) = \begin{cases} q(y) & \text{si } y \in [0, x_{\max}] \\ 0 & \text{sinon.} \end{cases}$$

$$p(x, a = \mathrm{K}, y) = \begin{cases} q(y - x) & \text{si } y \in [x, x_{\max}] \\ q(y - x + x_{\max}) & \text{si } y \in [0, x) \\ 0 & \text{sinon.} \end{cases}$$

Le coût immédiat (opposé d'une récompense) $c(x)$ est la somme d'une fonction monotone lentement croissante (qui correspond par exemple à des coûts de maintenance) et d'une fonction coût discontinue ponctuelle (*e.g.* les coûts de révision ou d'assurance).

Le coût immédiat et la fonction valeur (calculée à l'aide d'une discrétisation fine du domaine) sont représentés sur la figure 2.

Les valeurs numériques sont $\gamma = 0.6$, $\beta = 0.6$, $C_{replace} = 50$, $C_{dead} = 70$, et $x_{\max} = 10$. Nous considérons une distribution uniforme μ sur le domaine $[0, x_{\max}]$. Nous choisissons K points (avec $K = 200$ ou 2000) uniformément répartis sur le domaine $\{x_k := kx_{\max}/K\}_{0 \le k < K}$ et réalisons, à chaque itération, un problème de régression quadratique :

$$V_{n+1} = \arg\min_{f \in \mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} [f(x_k) - \mathcal{T}V_n(x_k)]^2,$$

où $\mathcal{F}$ est l'espace engendré par une famille de cosinus (avec $M = 20$ ou $M = 40$ fonctions de base) :

$$\mathcal{F} := \left\{ f(x) = \sum_{m=1}^{M} \alpha_m \cos(m\pi \frac{x}{x_{\max}}) \right\}_{\alpha \in \mathbb{R}^M}.$$

Nous commençons avec une fonction valeur initiale $V_0 = 0$. La figure 3 représente la première itération (pour la grille à $K = 200$ points) : les valeurs itérées par

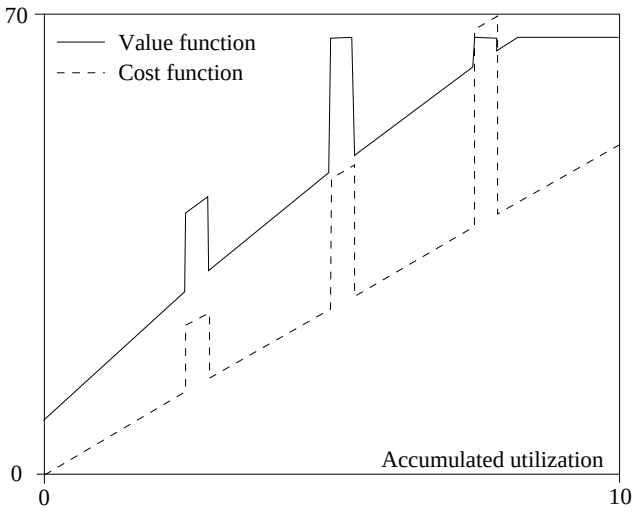

Figure 2. *Fonctions coût et valeur*

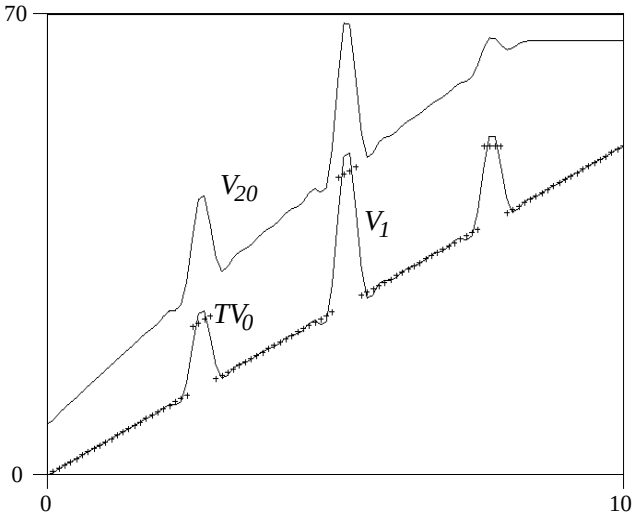

Figure 3. $\mathcal{T}V_0$ *(les croix),* V_1 *et* V_{20}

l'opérateur de Bellman $\mathcal{T}V_0$ (indiquées par les croix), la régression correspondante V_1 (meilleure approximation de $\mathcal{T}V_0$ dans l'espace $\mathcal{F}$). La fonction valeur approchée

après 20 itérations (lorsque l'on n'observe plus d'amélioration significative) est aussi représentée.

La constante de régularité $C(\mu)$ est définie par la valeur maximale de la densité de transition (puisque μ est uniforme), ainsi : $C(\mu) = q(0)x_{\max} = \beta x_{\max}/(1 - e^{-\beta x_{\max}}) \simeq 6$.

	$\|\|\varepsilon_n\|\|_\infty$	$C(\mu)\|\|\varepsilon_n\|\|_1$	$\sqrt{C(\mu)}\|\|\varepsilon_n\|\|_2$
$K = 200,\ M = 20$	12.4	0.367	1.16
$K = 2000,\ M = 40$	12.4	0.0552	0.897

Tableau 1. *Comparaison des parties droites des bornes en norme L_∞, L_1 et L_2*

Le tableau 1 donne les valeurs des parties droites (en négligeant la constante $2\gamma/(1-\gamma)^2$ qui est commune à toutes les bornes) des équations [3] et [14] pour $p = 1$ et 2, leur partie gauche étant la même puisqu'elles utilisent la même norme L_∞. Nous remarquons que les valeurs obtenues pour les bornes L_1 et L_2 [14] sont plus petites que celles en L_∞ [3]. De plus, nous observons que les erreurs d'approximation L_1 et L_2 tendent vers 0 lorsque K, le nombre d'échantillons, et M, le nombre de fonctions de base, tendent vers l'infini, alors que ce n'est pas le cas pour la borne L_∞. La raison est que puisque la fonction coût est discontinue, l'erreur d'approximation L_∞ (en utilisant un espace $\mathcal{F}$ de fonctions continues, comme cela est le cas avec les cosinus utilisés ici) ne pourra jamais être inférieure à la moitié de la plus grande discontinuité, même pour des grandes valeurs de K et M. Cet exemple illustre le fait que la borne L_p [14] peut être arbitrairement plus fine que la borne L_∞ [3].

7. Conclusion

Le théorème 1 fournit un outil intéressant pour majorer les performances de l'algorithme d'IVA en fonction de la norme L_p des erreurs d'approximation, donc en termes de la richesse d'approximation de l'algorithme d'apprentissage supervisé (lorsque celui-ci réalise un problème de minimisation en norme L_p, comme cela est le cas généralement). Le fait de pouvoir analyser la performance de l'algorithme d'IVA en utilisant la même norme que celle utilisée par l'opérateur de régression garantit la finesse et l'utilité applicative de ces majorations. Mentionnons tout de même que pour que ces bornes soient utiles, il convient de savoir estimer une majoration sur les constantes de régularité $C(\mu)$, $C_1(\nu, \mu)$ et $C_2(\nu, \mu)$, ce qui peut s'avérer difficile dans certains cas. Nous avons illustré le cas de petites valeurs de $C_1(\nu, \mu)$ et $C_2(\nu, \mu)$ pour le PDM chaînon, et de petite valeur de $C(\mu)$ pour le problème de remplacement optimal.

Des extensions immédiates sont l'utilisation d'autres fonctions perte l, telles que la *perte ϵ-insensible* (utilisée dans les SVM) ou la *perte de Huber* (utilisée pour la régression robuste) (Vapnik, 1998) ainsi que l'application de ces résultats à des problèmes de jeux markoviens.

8. Bibliographie

Atkeson C. G., Moore A. W., Schaal S. A., « Locally Weighted Learning », *AI Review*, 1997.

Bellman R., Dreyfus S., « Functional Approximation and Dynamic Programming », *Math. Tables and other Aids Comp.*, vol. 13, p. 247-251, 1959.

Bertsekas D. P., Tsitsiklis J., *Neuro-Dynamic Programming*, Athena Scientific, 1996.

Davies G., Mallat S., Avellaneda M., « Adaptive Greedy Approximations », *J. of Constr. Approx.*, vol. 13, p. 57-98, 1997.

DeVore R., *Nonlinear Approximation*, Acta Numerica, 1997.

Gordon G., « Stable function approximation in dynamic programming », *Proceedings of the International Conference on Machine Learning*, 1995.

Gordon G. J., Approximate solutions to Markov Decision Processes, PhD thesis, CS department, Carnegie Mellon University, Pittsburgh, PA, 1999.

Guestrin C., Koller D., Parr R., « Max-norm Projections for Factored MDPs », *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer Series in Statistics, 2001.

Kakade S., Langford J., « Approximately Optimal Approximate Reinforcement Learning », *Proceedings of the 19th International Conference on Machine Learning*, 2002.

Koller D., Parr R., « Policy Iteration for Factored MDPs », *Proceedings of the 16th conference on Uncertainty in Artificial Intelligence*, 2000.

Lagoudakis M., Parr R., « Least-Squares Policy Iteration », *Journal of Machine Learning Research*, vol. 4, p. 1107-1149, 2003.

Mallat S., *A Wavelet Tour of Signal Processing*, Academic Press, 1997.

Munos R., « Error Bounds for Approximate Policy Iteration », *19th International Conference on Machine Learning*, 2003.

Pollard D., *Convergence of Stochastic Processes*, Springer Verlag, New York, 1984.

Puterman M. L., *Markov Decision Processes, Discrete Stochastic Dynamic Programming*, A Wiley-Interscience Publication, 1994.

Rust J., *Numerical Dynamic Programming in Economics*, In Handbook of Computational Economics, Elsevier, North Holland, 1996.

Samuel A., « Some studies in machine learning using the game of checkers », *IBM Journal on Research and Development*p. 210-229, 1959. Reprinted in *Computers and Thought*, E.A. Feigenbaum and J. Feldman, editors, McGraw-Hill, New York, 1963.

Sutton R. S., Barto A. G., « Reinforcement Learning : An Introduction », *Bradford Book*, 1998.

Vapnik V., *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

Vapnik V., Golowich S. E., Smola A., « Support Vector Method for Function Approximation, Regression Estimation and Signal Processing », *In Advances in Neural Information Processing Systems*p. 281-287, 1997.

Williams R., Baird L., « Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions », *Technical Report NU-CCS-93-14. Northeastern University*, 1993.