Recherche locale de politique dans un espace convexe

Bruno Scherrer, Matthieu Geist

Recherche locale de politique dans un espace convexe

Bruno Scherrer * Matthieu Geist †

Résumé

En apprentissage par renforcement, la recherche locale de politique est une approche classique permettant de prendre en compte de grands espaces d'état. Formellement, elle consiste à chercher localement dans un espace de politiques paramétrées la solution qui va maximiser la fonction de valeur associée, moyennée selon une loi prédéfinie sur les états. La première contribution de cet article montre que si l'espace de politiques est convexe, *tout optimum local* (approché) présente une *garantie globale de performance*. Malheureusement, supposer la convexité de l'espace de recherche est une hypothèse forte : elle n'est pas satisfaite par les représentations usuelles des politiques et définir une paramétrisation non triviale qui satisfasse cette propriété est difficile. Une solution naturelle pour palier ce problème est d'optimiser la fonction objectif associée grâce à une montée de gradient fonctionnel, la recherche étant contrainte à l'enveloppe convexe de l'espace de politiques. Il s'avère que l'algorithme résultant est une légère généralisation du schéma d'itération conservative de la politique. Ainsi, notre seconde contribution consiste à souligner cette connexion originale entre recherche locale de politique et programmation dynamique approchée.

1 Introduction

Cet article s'intéresse au problème de l'apprentissage par renforcement (Sutton & Barto, 1998), formalisé grâce aux processus décisionnels de Markov (PDM) (Puterman, 1994), dans le cas où la taille de l'espace d'état nécessite des approximations. D'un côté, la programmation dynamique approchée (PDA) est une approche standard pour gérer les grands espaces d'état. Elle consiste à mimer, dans une forme approchée, des algorithmes standard conçus pour optimiser globalement la politique (c'est-à-dire maximiser la fonction de valeur associée en chaque état) lorsque le modèle est connu. D'un autre côté, la recherche locale de politique (RLP) consiste à paramétrer la politique (parfois appelée "acteur") et à

*Equipe BIGS, INRIA Nancy Grand-Est, Nancy, France, bruno.scherrer@inria.fr

†UMI 2958, Georgia Tech - CNRS, CentraleSupélec, Université Paris-Saclay, Metz, France, matthieu.geist@centralesupelec.fr

maximiser localement l'espérance (selon une loi sur les états prédéfinie) de la valeur associée. Cela peut être fait grâce à une montée de gradient naturel (Baxter & Bartlett, 2001 ; Kakade, 2001), possiblement combinée à l'utilisation d'un critique (Sutton et al., 1999 ; Peters & Schaal, 2008), grâce à une approche de type espérance-maximisation (EM) (Kober & Peters, 2011), voire directement en utilisant une approche de type optimisation en boîte noire (Heidrich-Meisner & Igel, 2008 ; Fix & Geist, 2012). Les approches de type RLP fonctionnent particulièrement bien en pratique : l'essentiel des références mentionnées ci-avant décrivent des problèmes de référence et des applications telles que la robotique, où la RLP est compétitive avec la PDA. De façon surprenante, les approches basées sur des gradients ou sur EM, qui sont notoirement enclines à fournir des optima locaux, ne semblent pas pénalisées dans leur application à l'apprentissage par renforcement. De façon encore plus surprenante, Kakade (2001) a montré qu'une montée de gradient naturel dans l'espace des politiques fournissait de meilleurs résultats que la PDA sur le jeu de Tetris.

En s'appuyant sur le travail fondateur de Bertsekas et Tsitsiklis (1996), il a été montré que les algorithmes de PDA bénéficient de garanties globales de performance, bornant l'erreur faite en utilisant la politique estimée plutôt que la politique optimale par une fonction des erreurs d'approximations faites à chaque itération : Munos (2003) traite de l'itération de la politique approchée (IPA), Munos (2007) de l'itération de la valeur approchée (IVA) et plus généralement Scherrer et al. (2012) traitent d'une forme approchée de l'itération modifiée de la politique. Autant que nous le sachions, de telles garanties n'existent pas pour la RLP. En termes généraux, le mieux que l'on puisse espérer de la RLP est l'obtention d'un optimum local de la fonction objectif (l'espérance de la valeur associée) maximisée, et l'importante question de la qualité de cette solution (en termes de valeur) reste ouverte. C'est la perspective principale mentionnée par Bhatnagar et al. (2007), où la convergence d'une famille d'algorithmes acteur-critique est démontrée : *il est important de caractériser la qualité des solutions vers lesquelles on converge*. La motivation principale de cet article est d'approfondir la compréhension de la RLP.

Notre principale contribution (le théorème 3 de la section 3) montre que si l'espace de politiques où l'on effectue la RLP est un sous-ensemble convexe de l'espace des politiques stochastique (autrement dit, l'espace de recherche est stable par mélange stochastique), alors *n'importe quel optimum local (approché) de la fonction de valeur moyenne bénéficie d'une garantie globale de performance*, similaire à celle offerte par la PDA (les bornes associées sont même plus fines, voir la section 5). Après avoir expliqué pourquoi proposer des politiques paramétrées qui satisfassent l'hypothèse de convexité semble particulièrement difficile, nous proposons dans la section 4 une solution algorithmique basée sur le *boosting* (vu comme une montée de gradient fonctionnel), qui permet d'effectuer la RLP dans l'enveloppe convexe d'un espace de politiques déterministes. Il s'avère que l'algorithme obtenu est une légère généralisation du schéma d'itération conservative de la politique (Kakade & Langford, 2002), originellement introduit dans le cadre de la PDA. Ainsi, une autre contribution de cet article est cette connexion originale entre PDA et RLP. La section 5 discute les résul-

tats de l'analyse proposée. Notamment, une comparaison est faite aux bornes de la PDA et les conséquences pratiques des résultats démontrés sont discutées. La section 6 ouvre quelques perspectives. Dans la prochaine section, nous commençons par présenter le contexte et par énoncer formellement le problème de la recherche locale de politique.

2 Contexte et notations

Soit Δ_X l'ensemble des mesures de probabilité sur l'ensemble fini X (muni de sa tribu discrète) et Y^X l'ensemble des applications de X vers l'ensemble fini Y. Par convention, tous les vecteurs sont des vecteurs colonne, sauf les mesures de probabilité qui sont des vecteurs ligne (multiplication à gauche). Nous considérons un PDM $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$ (Puterman, 1994 ; Bertsekas, 1995), où $\mathcal{S}$ est un espace d'état fini [1], $\mathcal{A}$ un espace d'action fini, $P \in (\Delta_{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}}$ la dynamique markovienne ($P(s'|s,a)$ est la probabilité de transiter vers l'état s' quand l'action a a été appliquée dans l'état s), $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ est la fonction de récompense bornée et $\gamma \in [0,1[$ est le facteur d'actualisation.

Une politique stochastique $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}}$ associe à chaque état s une mesure de probabilité $\pi(.|s)$ sur l'espace $\mathcal{A}$ des actions. Un espace de politiques Π, sous-ensemble de $(\Delta_{\mathcal{A}})^{\mathcal{S}}$, est dit convexe (ou de façon équivalente stable par mélange stochastique) si il satisfait

$$\forall \pi, \pi' \in \Pi, \quad \forall \alpha \in]0,1[, \quad (1-\alpha)\pi + \alpha\pi' \in \Pi.$$

Pour une politique donnée π, nous définissons $r_\pi \in \mathbb{R}^{\mathcal{S}}$ par

$$r_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) = \mathbb{E}_{a\sim\pi(.|s)}[r(s,a)]$$

et $P_\pi \in (\Delta_{\mathcal{S}})^{\mathcal{S}}$ par

$$P_\pi(s'|s) = \sum_{a\in\mathcal{A}} \pi(a|s) P(s'|s,a) = \mathbb{E}_{a\sim\pi(.|s)}[P(s'|s,a)].$$

La fonction de valeur v_π quantifie la qualité d'une politique π pour chaque état s en mesurant le cumul espéré et décompté de récompenses obtenues en démarrant en cet état puis en suivant la politique :

$$v_\pi(s) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r_\pi(s_t) | s_0 = s, s_{t+1} \sim P_\pi(.|s_t)\right].$$

L'opérateur de Bellman T_π d'une politique π associe à chaque fonction $v \in \mathbb{R}^{\mathcal{S}}$ la fonction définie par

$$[T_\pi v](s) = \mathbb{E}\left[r_\pi(s) + \gamma v(s') | s' \sim P_\pi(.|s)\right]$$

1. Les résultats présentés peuvent facilement être étendus à des espaces dénombrables voire continus et compacts, voir par exemple le formalisme utilisé par Scherrer et al. (2015). La restriction au cas fini est faite pour la clarté d'exposition.

ou, de façon plus compacte, $T_\pi v = r_\pi + \gamma P_\pi v$. La fonction de valeur v_π est l'unique point fixe de T_π.

Il existe au moins une politique π_*, optimale dans le sens où elle satisfait $v_{\pi_*}(s) \geq v_\pi(s)$ pour tout état s et toute politique π. La fonction de valeur v_* est l'unique point fixe de l'opérateur de Bellman non-linéaire défini par

$$v_* = Tv_* \text{ avec } Tv = \max_{\pi \in \mathcal{A}^\mathcal{S}} T_\pi v$$

où l'opérateur max se lit ici par composante. Etant donné une fonction $v \in \mathbb{R}^\mathcal{S}$, une politique π' est dite gloutonne par rapport à v si $T_{\pi'} v = Tv$ et nous notons $\mathcal{G}(\pi)$ l'ensemble des politiques gloutonnes par rapport à la valeur v_π d'une politique π. Les notions de fonction de valeur optimale et de politique gloutonne sont fondamentales dans le cadre du contrôle optimal : toute politique π_* gloutonne par rapport à la valeur optimale est une politique optimale et sa valeur satisfait $v_{\pi_*} = v_*$. Ainsi, une caractérisation équivalente de l'optimalité d'une politique π est qu'elle soit gloutonne par rapport à sa propre valeur :

$$\pi \in \mathcal{G}(\pi). \tag{1}$$

Pour une distribution μ, nous définissons la mesure d'occupation décomptée [2] induite par la politique π quand l'état initial est échantillonné selon μ par $d_{\mu,\pi} = (1-\gamma)\mu(I - \gamma P_\pi)^{-1}$ (rappelons que μ est un vecteur ligne par convention), où $(I - \gamma P_\pi)^{-1} = \sum_{t \geq 0} (\gamma P_\pi)^t$. Il est aisé de vérifier que $\mu v_\pi = \frac{1}{1-\gamma} d_{\mu,\pi} r_\pi$. Pour deux mesures μ et ν, nous écrivons $\left\| \frac{\mu}{\nu} \right\|_\infty$ la plus petite constante C satisfaisant $\mu(s) \leq C\nu(s)$ pour tout état $s \in \mathcal{S}$. Cette constante est la norme infini du ratio par composante des mesures, d'où la notation. C'est un coefficient de concentrabilité qui mesure le désaccord entre les mesures.

D'un point de vue algorithmique, les méthodes de programmation dynamique calculent la paire (v_*, π_*) de façon itérative. Quand le problème est grand et ne peut être résolu exactement, la PDA réfère à des implémentations bruitées de ces méthodes exactes, où le bruit est causé par les approximations faites à chaque itération. Par exemple, l'IVA et l'IPA correspondent respectivement aux schémas suivants :

$$v_{k+1} = Tv_k + \epsilon_k \quad \text{et} \quad \begin{cases} v_k = v_{\pi_k} + \epsilon_k \\ \pi_{k+1} \in \mathcal{G}(v_k) \end{cases}.$$

Dans le cadre de la RLP qui nous intéresse ici, notons Π l'espace où est cherchée une solution. Pour une mesure (de probabilité, arbitraire) d'intérêt ν définie *a priori*, le problème de la RLP peut être exprimé comme suit :

trouver $\pi \in \Pi$ telle que π est un maximum local de $J_\nu(\pi) = \mathbb{E}_{s \sim \nu}[v_\pi(s)]$.

Supposons être capable de trouver une politique $\pi \in \Pi$ qui est un optimum local (approché). Une question naturelle est de savoir ce que l'on peut dire

2. Quand elle existe, cette mesure tend vers la distribution stationnaire de la chaîne de Markov P_π lorsque le facteur d'actualisation tend vers 1.

de la distance de la valeur de cette politique, v_π, à la valeur de la politique optimale, $v_* = v_{\pi_*}$. De façon assez surprenante, et contrairement à la plupart des problèmes d'optimisation, nous donnons une condition sur l'espace de recherche Π qui permet d'obtenir une garantie non triviale de performance. C'est l'objet de la section suivante.

3 Principal résultat

Pour énoncer notre résultat principal, nous devons définir une relaxation de l'ensemble des politiques gloutonnes par rapport à une politique donnée.

Définition 1 (politiques ϵ-gloutonnes en μ-moyenne)**.** *Nous notons $\mathcal{G}_\Pi(\pi, \mu, \epsilon)$ l'ensemble des politiques qui sont ϵ-gloutonnes par rapport à π en μ-moyenne, formellement défini par :*

$$\mathcal{G}_\Pi(\pi, \mu, \epsilon) = \{\pi' \in \Pi \text{ tel que } \forall \pi'' \in \Pi,\ \mu T_{\pi'} v_\pi + \epsilon \geq \mu T_{\pi''} v_\pi\}\,.$$

C'est en effet une relaxation de $\mathcal{G}$, car pour toutes politiques π et π' nous avons :

$$\begin{aligned} \pi' \in \mathcal{G}(\pi) \quad &\Leftrightarrow \quad \forall \mu \in \Delta_\mathcal{S},\ \pi' \in \mathcal{G}_\Pi(\pi, \mu, 0) \\ &\Leftrightarrow \quad \exists \mu \in \Delta_\mathcal{S},\ \mu > 0,\ \pi' \in \mathcal{G}_\Pi(\pi, \mu, 0). \end{aligned}$$

Nous pouvons maintenant énoncer le premier résultat important.

Théorème 1. *Soit π une politique de Π et ν la mesure d'intérêt définissant le problème de la RLP. Les deux assertions suivantes sont équivalentes :*

$$\forall \pi' \in \Pi, \quad \lim_{\alpha \to 0} \frac{\nu v_{(1-\alpha)\pi + \alpha\pi'} - \nu v_\pi}{\alpha} \leq \epsilon. \tag{2}$$

$$\pi \in \mathcal{G}_\Pi(\pi, d_{\nu,\pi}, (1-\gamma)\epsilon). \tag{3}$$

L'équation (3) stipule que la politique π est approximativement gloutonne par rapport à elle-même, ce qui peut être vu comme une version relâchée de la condition d'optimalité exprimée équation (1). Comme nous le montrons plus tard, c'est cela qui nous permet de proposer une garantie globale de performance pour la politique π. L'équation (2) stipule que la politique π est un optimum local approché de $\pi \mapsto J_\nu(\pi)$, si l'espace de recherche est convexe. En effet, quelle que soit la direction dans laquelle on regarde dans cet espace, l'amélioration locale autour de π est bornée par ϵ. Le théorème 1 a donc le corollaire suivant.

Corollaire 1. *Supposons l'espace Π convexe. Toute politique π qui est un ϵ-optimum local de $\pi \mapsto J_\nu(\pi)$ (dans le sens de l'équation (2)) satisfait l'équation de Bellman relaxée (3).*

Nous allons maintenant démontrer le théorème 1. Pour cela, le lemme technique suivant est utile.

Lemme 1. *Pour toutes politiques π et π', nous avons*

$$v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_\pi - v_\pi).$$

Démonstration. La preuve utilise le fait que l'équation d'évaluation de Bellman, $v_\pi = r_\pi + \gamma P_\pi v_\pi$, implique que $v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$. Alors,

$$\begin{aligned} v_{\pi'} - v_\pi &= (I - \gamma P_{\pi'})^{-1} r_{\pi'} - v_\pi \\ &= (I - \gamma P_{\pi'})^{-1}(r_{\pi'} + \gamma P_{\pi'} v_\pi - v_\pi) \\ &= (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_\pi - v_\pi). \end{aligned}$$

□

Preuve du théorème 1. Pour tout $\alpha \in]0, 1[$ et pour toute politique $\pi' \in \Pi$, notons $\pi_\alpha = (1 - \alpha)\pi + \alpha\pi'$. Grâce au lemme 1, nous avons :

$$\nu(v_{\pi_\alpha} - v_\pi) = \nu(I - \gamma P_{\pi_\alpha})^{-1}(T_{\pi_\alpha} v_\pi - v_\pi).$$

En remarquant que $r_{\pi_\alpha} = (1 - \alpha) r_\pi + \alpha r_{\pi'}$ et que $P_{\pi_\alpha} = (1 - \alpha) P_\pi + \alpha P_{\pi'}$, on a directement que $T_{\pi_\alpha} v_\pi = (1 - \alpha) T_\pi v_\pi + \alpha T_{\pi'} v_\pi$. Ainsi, en utilisant le fait que $v_\pi = T_\pi v_\pi$, nous obtenons :

$$\begin{aligned} T_{\pi_\alpha} v_\pi - v_\pi &= (1 - \alpha) T_\pi v_\pi + \alpha T_{\pi'} v_\pi - v_\pi \\ &= \alpha(T_{\pi'} v_\pi - v_\pi). \end{aligned}$$

Parallèlement à cela, nous avons (pour α suffisamment petit)

$$\begin{aligned} (I - \gamma P_{\pi_\alpha})^{-1} &= (I - \gamma P_\pi + \alpha\gamma(P_\pi - P_{\pi'}))^{-1} \\ &= (I - \gamma P_\pi)^{-1} \left(I - \alpha\gamma(P_{\pi'} - P_\pi)(I - \gamma P_\pi)^{-1}\right)^{-1} \\ &= (I - \gamma P_\pi)^{-1}(I + \alpha M), \end{aligned}$$

où M est bornée (la forme exacte de cette matrice n'est pas importante). En assemblant ces résultats, nous obtenons

$$\nu(v_{\pi_\alpha} - v_\pi) = \alpha\nu(I - \gamma P_\pi)^{-1}(T_{\pi'} v_\pi - v_\pi) + O(\alpha^2).$$

En passant à la limite, nous avons

$$\begin{aligned} \lim_{\alpha \to 0} \frac{\nu(v_{\pi_\alpha} - v_\pi)}{\alpha} &= \nu(I - \gamma P_\pi)^{-1}(T_{\pi'} v_\pi - v_\pi) \\ &= \frac{1}{1 - \gamma} d_{\nu,\pi}(T_{\pi'} v_\pi - v_\pi), \end{aligned}$$

et le résultat annoncé suit. □

La deuxième étape de notre analyse consiste à montrer que la condition d'optimalité relâchée de l'équation (3) implique une garantie globale de performance. Pour exprimer ce résultat, nous devons d'abord définir la notion de "complexité ν-gloutonne" relative à l'espace de politiques, qui mesure à quel point Π permet d'approcher l'opérateur de gloutonnerie, pour une distribution ν sur les états.

Définition 2 (Complexité ν-gloutonne)**.** *Nous définissons $\mathcal{E}_\nu(\Pi)$ la complexité ν-gloutonne d'un espace de politiques π par*

$$\mathcal{E}_\nu(\Pi) = \max_{\pi\in\Pi} \min_{\pi'\in\Pi} \left(d_{\nu,\pi}\left(Tv_\pi - T_{\pi'}v_\pi\right)\right).$$

Comme $Tv_\pi - T_\pi v_\pi = Tv_\pi - v_\pi \geq 0$, nous avons que $\mathcal{E}_\nu(\Pi) \geq 0$, pour tout espace de politiques Π. Dans le cas limite où Π contient toutes les politiques déterministes, nous avons $\mathcal{E}_\nu(\Pi) = 0$.

Grâce à cette définition, nous pouvons énoncer le résultat suivant.

Théorème 2. *Si $\pi \in \mathcal{G}_\Pi(\pi, d_{\nu,\pi}, \epsilon)$, alors pour toute politique π' et pour toute mesure $\mu \in \Delta_\mathcal{S}$, nous avons*

$$\mu v_{\pi'} \leq \mu v_\pi + \frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi'}}{\nu}\right\|_\infty (\mathcal{E}_\nu(\Pi) + \epsilon).$$

Notons que ce théorème est en fait une légère généralisation [3] du théorème 6.2 de Kakade et Langford (2002).

Démonstration. En utilisant à nouveau le lemme 1 et le fait que $Tv_\pi \geq T_{\pi'}v_\pi$, nous avons

$$\begin{aligned}\mu(v_{\pi'} - v_\pi) &= \mu(I - \gamma P_{\pi'})^{-1}(T_{\pi'}v_\pi - v_\pi) \\ &= \frac{1}{1-\gamma} d_{\mu,\pi'}(T_{\pi'}v_\pi - v_\pi) \leq \frac{1}{1-\gamma} d_{\mu,\pi'}(Tv_\pi - v_\pi).\end{aligned}$$

Comme $Tv_\pi - v_\pi \geq 0$ et $d_{\nu,\pi} \geq (1-\gamma)\nu$, nous obtenons

$$\begin{aligned}\mu(v_{\pi'} - v_\pi) &\leq \frac{1}{1-\gamma}\left\|\frac{d_{\mu,\pi'}}{\nu}\right\|_\infty \nu(Tv_\pi - v_\pi) \\ &\leq \frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi'}}{\nu}\right\|_\infty d_{\nu,\pi}(Tv_\pi - v_\pi).\end{aligned}$$

En utilisant le fait que $d_{\nu,\pi}(Tv_\pi - v_\pi) = (d_{\nu,\pi}Tv_\pi - d_{\nu,\pi}v_\pi)$, nous obtenons finalement

$$\begin{aligned}&\mu(v_{\pi'} - v_\pi) \\ &\leq \frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi'}}{\nu}\right\|_\infty \left(d_{\nu,\pi}Tv_\pi - \max_{\pi'\in\Pi} d_{\nu,\pi}T_{\pi'}v_\pi + \max_{\pi'\in\Pi} d_{\nu,\pi}T_{\pi'}v_\pi - d_{\nu,\pi}v_\pi\right) \\ &\leq \frac{1}{(1-\gamma)^2}\left\|\frac{d_{\mu,\pi'}}{\nu}\right\|_\infty (\mathcal{E}_\nu(\Pi) + \epsilon).\end{aligned}$$

□

3. Le théorème 2 est vrai pour toute politique π', pas seulement pour la politique optimale, et le terme d'erreur est séparé (ce qui est nécessaire pour fournir un résultat plus général, autrement dit pour séparer biais et variance).

Notre résultat principal découle d'une combinaison évidente du corollaire 1 et du théorème 2.

Théorème 3. *Supposons que l'espace* Π *est convexe. Alors, tout politique* π *qui est un* ϵ*-optimum local de* $\pi \mapsto J_\nu(\pi)$ *(dans le sens de l'équation* (2)*) bénéficie de la garantie globale de performance suivante :*

$$\mathbb{E}_{s\sim\mu}[v_*(s) - v_\pi(s)] \leq \frac{1}{1-\gamma} \left\| \frac{d_{\mu,\pi_*}}{\nu} \right\|_\infty \left(\frac{\mathcal{E}_\nu(\Pi)}{1-\gamma} + \epsilon \right).$$

4 A propos de l'hypothèse convexe

Le résultat remarquable de la section précédente—une connexion entre optimalité *locale* et garantie *globale*—repose sur l'hypothèse que l'espace Π est convexe (ou, de façon équivalente, stable par mélange stochastique). Bien que cette hypothèse puisse sembler faible dans un premier temps, nous allons montrer qu'elle est en fait très forte, et non satisfaite en générale. Nous proposons alors une approche algorithmique naturelle pour effectuer la RLP dans l'enveloppe convexe d'un espace de politiques Π (qui n'est, lui, pas nécessairement convexe).

4.1 Une hypothèse forte

Une approche standard (dans le cas d'actions continues) consiste à paramétrer une fonction u_θ associant des actions à des états, puis à la considérer comme l'espérance d'une distribution gaussienne, c'est-à-dire

$$\pi_\theta(a|s) \propto \exp\left(-\frac{1}{2} \left\| a - u_\theta(s) \right\|^2_{\Sigma^{-1}} \right),$$

où Σ est une matrice de variance-covariance prédéfinie. Evidemment, l'espace de ces politiques n'est pas convexe, dans la mesure où un mélange de distributions gaussiennes n'est en général pas une distribution gaussienne. Une autre approche (dans le cas d'actions discrètes) consiste à adopter une distribution de Gibbs paramétrée, c'est-à-dire

$$\pi_\theta(a|s) \propto \exp\left(\theta^\top \psi(s,a) \right),$$

où $\theta^\top \psi(s,a)$ peut être vue comme une fonction de score (ou une fonction de valeur sur les couples état-action). Ici encore, l'espace de politiques résultant n'est pas convexe en général.

En fait, nous considérons comme un problème ouvert le fait de proposer une paramétrisation non triviale qui définisse un espace de politiques convexe (par non triviale, nous entendons un espace qui n'est ni la combinaison convexe d'un petit nombre de politiques, ni l'enveloppe convexe de tout l'espace des politiques déterministes $\mathcal{A}^\mathcal{S}$). Même s'il n'y a qu'un unique état, le problème n'est pas évident : il s'agit de trouver une mesure de probabilité qui soit stable par

mélange stochastique, et nous n'avons pas de réponse satisfaisante à ce jour. Une alternative, que nous développons dans la section suivante, consiste à considérer pour Π l'enveloppe convexe d'un ensemble de politiques (non-convexe).

4.2 *Boosting*

Soit $\mathcal{P}$ un espace de politiques et $\Pi = \text{co}(\mathcal{P})$ son enveloppe convexe. Nous proposons d'utiliser du *boosting* pour trouver un maximum local de $J_\nu(\pi)$ dans Π. Plus précisément, nous proposons d'appliquer l'algorithme AnyBoost.L1 de Mason et al. (1999), qui exprime le *boosting* comme une montée de gradient dans l'espace fonctionnel (des politiques) et contraint la recherche à l'enveloppe convexe des politiques de $\mathcal{P}$. Soit $\nabla J_\nu(\pi)$ le gradient fonctionnel (par rapport à π) de la fonction objectif de la RLP. Appliqué à notre problème, AnyBoost.L1 fonctionne comme suit. A chaque iteration k, nous avons la politique courante π_{k-1} et nous effectuons les étapes suivantes :

1. calcul de $h_k \in \text{argmax}_{h\in\mathcal{P}}\langle \nabla J_\nu(\pi_{k-1}), h\rangle$;
2. mise à jour de la politique : $\pi_k = (1-\alpha_k)\pi_{k-1} + \alpha_k h_k$, avec $\alpha_k \in]0,1[$ le taux d'apprentissage.

L'idée de base est de faire une montée de gradient fonctionnel sur $J_\nu(\pi)$. Toutefois, le gradient $\nabla J_\nu(\pi_{k-1})$ n'appartient généralement pas à $\mathcal{P}$, donc nous cherchons une politique h la plus colinéaire possible avec $\nabla J_\nu(\pi_{k-1})$, ce qui est la première étape. La seconde étape met à jour la politique, en faisant un mélange stochastique entre l'ancienne et la politique h_k, le coefficient de mélange α_k étant le taux d'apprentissage de la montée de gradient. Pour obtenir un algorithme pratique, il est nécessaire d'étudier le problème d'optimisation de la première étape.

Proposition 1. *Nous avons que*

$$\underset{h\in\mathcal{P}}{\text{argmax}}\langle \nabla J(\pi), h\rangle = \underset{h\in\mathcal{P}}{\text{argmin}}\, d_{\nu,\pi}(Tv_\pi - T_h v_\pi).$$

En particulier, supposons que $\mathcal{P}$ est un espace de politiques déterministes et définissons $q_\pi(s,a) = [T_a v_\pi](s)$, la fonction de valeur sur les couples état-action de la politique π (en notant avec un léger abus de notation T_a l'opérateur de Bellman pour la politique qui associe l'action a à chaque état), alors

$$\underset{h\in\mathcal{P}}{\text{argmax}}\langle \nabla J(\pi), h\rangle = \underset{h\in\mathcal{P}}{\text{argmin}} \sum_{s\in S} d_{\nu,\pi}(s)\left(\max_{a\in\mathcal{A}} q_\pi(s,a) - q_\pi(s,h(s))\right).$$

Cela peut être vu comme une version approchée de l'étape gloutonne de l'algorithme d'itération de la politique, qui peut être implémentée comme un problème de classification multi-classe à coût sensitif, ou comme une régression ℓ_p de la fonction q_π.

Preuve de la proposition 1. Le gradient fonctionnel [4] de J_ν est

$$\nabla J_\nu(\pi) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d_{\nu,\pi}(s) \sum_{a \in \mathcal{A}} \nabla \pi(a|s) q_\pi(s,a).$$

C'est une extension directe du résultat classique de Sutton et al. (1999). Ensuite, il s'agit de calculer son produit scalaire avec une fonction h de $\mathcal{P}$:

$$\begin{aligned}
\langle \nabla J(\pi), h \rangle &= \frac{1}{1-\gamma} \langle \sum_s d_{\nu,\pi}(s) \sum_a \nabla \pi(a|s) q_\pi(s,a), h \rangle \\
&= \frac{1}{1-\gamma} \sum_s d_{\nu,\pi}(s) \sum_a \langle \nabla \pi(a|s), h \rangle q_\pi(s,a) \\
&= \frac{1}{1-\gamma} \sum_s d_{\nu,\pi}(s) \sum_a h(a|s) q_\pi(s,a) \\
&= \frac{1}{1-\gamma} d_{\nu,\pi}(T_h v_\pi).
\end{aligned}$$

Ceci permet de conclure :

$$\begin{aligned}
\operatorname*{argmax}_{h \in \mathcal{H}} \langle \nabla J, h \rangle &= \operatorname*{argmax}_{h \in \mathcal{H}} d_{\nu,\pi}(T_h v_\pi) \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} d_{\nu,\pi}(T v_\pi - T_h v_\pi).
\end{aligned}$$

□

4.3 Connexion à l'itération conservative de la politique

Ainsi, le *boosting* appliqué à la RLP consiste à calculer un mélange stochastique de politiques, chaque nouvelles composante du mélange étant la solution d'une approximation de l'opérateur de gloutonnerie par rapport à la fonction de valeur du mélange précédent. Il s'avère que l'itération conservative de la politique (ICP) (Kakade & Langford, 2002) est un cas particulier de cet algorithme général, la seule différence étant que l'ICP choisit des valeurs spécifiques du taux d'apprentissage (de façon à garantir une amélioration).

Si l'algorithme obtenu par *boosting* n'est pas vraiment nouveau, il apporte des clarifications sur les liens qui existent entre la RLP, la PDA et l'ICP. Premièrement, il montre que l'ICP peut être obtenue comme étant une méthode de RLP, alors qu'elle a été originellement dérivée comme une méthode de PDA, la motivation étant de résoudre le problème de dégradation des politiques de l'IAP (Kakade & Langford, 2002). Cela tisse un lien entre la RLP et la PDA, qui n'a pas été documenté dans la littérature à ce jour, et qui souligne le fait que

4. Par gradient fonctionnel nous entendons dérivée au sens de Fréchet, pour l'espace de Hilbert adéquat, c'est-à-dire l'ensemble des classes d'équivalence des fonctions $g \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ telles que $\sum_s \nu(s) \sum_a g(a,s)^2$ soit fini, muni du produit scalaire $\langle g_1, g_2 \rangle = \sum_s \nu(s) \sum_a g_1(s,a) g_2(s,a)$. Voir par exemple (Geist, 2015) pour plus de détails concernant ce type d'espace.

l'ICP est à la frontière de ces deux approches générales. Cela fournit quelques pistes d'amélioration pour l'ICP (qui a des garanties fortes mais est lente en générale). Il est envisageable de choisir le taux d'apprentissage en se basant sur la littérature concernant le *boosting* et les problèmes d'optimisation associés, ou sur des heuristiques du domaine (voire d'optimiser le taux d'apprentissage avec du *line search*). Enfin, AnyBoost.L1 est une approche naturelle pour chercher un optimum local de J_ν dans un espace convexe de politiques. Etudier la possibilité d'utiliser des algorithmes alternatifs est une perspective de recherche intéressante.

5 Discussion

Dans cette section, nous discutons le lien entre notre analyse et des travaux antérieurs, nous comparons les garanties obtenues avec celles de la PDA (en nous focalisant plus particulièrement sur l'IPA) et nous discutons certaines conséquences pratiques et théoriques de nos résultats.

5.1 Une analyse proche

Une garantie de performance très similaire à celle du théorème 2 a d'abord été obtenue pour l'ICP par Kakade et Langford (2002). Ce résultat a certainement été considéré comme propre à l'ICP, qui n'a semble-t-il pas été beaucoup utilisé en pratique, vraisemblablement en raison d'une implémentation relativement complexe. En revanche, nous avons montré dans cet article que cette garantie de performance est valide pour n'importe quelle approche qui trouve une politique satisfaisant l'identité de Bellman relâchée donnée équation (3), dont l'ICP découle naturellement, comme nous l'avons montré section 4.

Bien que le résultat principal de cet article soit le théorème 3, et comme le théorème 2 apparaît sous une forme très proche dans l'article de Kakade et Langford (2002), notre contribution technique principale est le théorème 1, qui souligne la connexion qui existe entre optimalité locale et la caractérisation relâchée de l'optimalité. Un résultat, à première vue similaire, est proposé par Kakade (2001) pour la RLP par montée de gradient naturel : le théorème 3 de cet article montre que les mises à jour par le gradient naturel déplacent la politique dans la direction de la solution que fournirait une mise à jour de programmation dynamique. L'auteur écrit : "*Le gradient naturel peu être efficace loin du maximum, dans le sens où il déplace la politique vers le choix d'actions gloutonnes optimales*". Bien qu'il y ait une connexion évidente avec notre travail, ce résultat est limité car (de façon similaire à ce que nous venons également d'expliquer pour l'ICA) *(i)* il semble spécifique au gradient naturel (alors que notre résultat est général) et *(ii)* il n'est pas exploité pour fournir une garantie globale de performance.

5.2 Relations aux bornes de la programmation dynamique approchée

La garantie de performance de n'importe que algorithme de programmation dynamique approchée implique *(i)* une dépendance quadratique en l'horizon moyen $\frac{1}{1-\gamma}$, *(ii)* un coefficient de concentrabilité (qui quantifie le désaccord entre la pire—au sens de la politique suivie—mesure d'occupation décomptée en commençant selon une mesure d'intérêt et la mesure utilisée pour contrôler les erreurs d'estimation) et *(iii)* un terme d'erreur lié à l'erreur d'estimation faite à chaque itération (qui peut être causée par l'estimation des valeurs et/ou des politiques). Selon la quantité estimée, une comparaison de ces erreurs d'estimation peut être difficile. Pour faciliter la comparaison, la discussion qui suit se focalise sur l'IAP. Notons toutefois que plusieurs aspects de cette comparaison sont plus généraux et s'appliquent globalement à la PDA. L'IAP génère une séquence de politiques : à chaque itération, la nouvelle politique est approximativement gloutonne par rapport à la valeur de la politique précédente. Cela peut être effectué grâce une régression ℓ_p de la fonction de valeur sur les couples état-action (Bertsekas, 1995 ; Munos, 2003 ; Lagoudakis & Parr, 2003a) ou grâce à une classification multi-classe à coût sensitif (Lagoudakis & Parr, 2003b ; Fern et al., 2006 ; Lazaric et al., 2010). Quelle que soit l'approche, la séquence de politiques appartient (implicitement pour la régression ou explicitement pour la classification) à un espace $\mathcal{P}$, qui est typiquement un ensemble de *politiques déterministes*. Pour une politique initiale π_0 et et une mesure donnée ν, l'IAP s'itère comme suit :

$$\begin{gathered}\text{choisir } \pi_{k+1} \in \mathcal{P} \\ \text{de façon à (approximativement) minimiser } \nu(Tv_{\pi_k} - T_{\pi_{k+1}} v_{\pi_k}).\end{gathered}$$

Cela est similaire à l'ICP/la RLP boostée, au faits près que *(i)* l'IAP utilise ν plutôt que $d_{\nu,\pi}$ pour approcher la politique gloutonne et *(ii)* l'IAP est optimiste (dans le sens où $\alpha_k = 1$). Pour énoncer la borne de l'IAP, nous avons besoin d'un autre coefficient de concentrabilité et d'une nouvelle erreur permettant de caractériser la qualité de l'espace $\mathcal{P}$. Soit $C_{\mu,\nu}$ le coefficient de concentrabilité définit par

$$C_{\mu,\nu} = (1-\gamma)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \gamma^{i+j} \sup_{\pi \in \mathcal{A}^{\mathcal{S}}} \left\| \frac{\mu (P_{\pi_*})^i (P_\pi)^j}{\nu} \right\|_\infty .$$

La complexité de l'espace $\mathcal{P}$ est quantifié par (de façon similaire à $\mathcal{E}_\nu$) :

$$\mathcal{E}'_\nu(\mathcal{P}) = \max_{\pi \in \mathcal{P}} \min_{\pi' \in \mathcal{P}} (\nu(Tv_\pi - T_{\pi'} v_\pi)).$$

Soit aussi e l'erreur d'estimation qui tend vers zéro lorsque le nombre d'échantillons tend vers l'infini (à une vitesse dépendant de l'estimateur choisi). La garantie de performance de l'IAP (Munos, 2003 ; Antos et al., 2008 ; Lazaric et al., 2011, 2010 ; Ghavamzadeh & Lazaric, 2012) peut s'exprimer comme suit :

$$\limsup_{k \to \infty} \mu(v_* - v_{\pi_k}) \leq \frac{C_{\mu,\nu}}{(1-\gamma)^2} (\mathcal{E}'_\nu(\mathcal{P}) + e).$$

Tableau 1 – Comparaison des garanties de performance pour la RLP et l'IAP.

	terme borné	horizon	concentrabilité	erreur
RLP	$\mu(v_* - v_\pi)$	$\frac{1}{(1-\gamma)^2}$	$\left\|\frac{d_{\mu,\pi_*}}{\nu}\right\|_\infty$	$\mathcal{E}_\nu(\Pi) + \epsilon(1-\gamma)$
IAP	$\limsup_{k\to\infty} \mu(v_* - v_{\pi_k})$	$\frac{1}{(1-\gamma)^2}$	$C_{\mu,\nu}$	$\mathcal{E}'_\nu(\mathcal{P}) + e$

Cette borne est comparable à celle du théorème 3, en considérant les trois termes impliqués : l'horizon moyen, le coefficient de concentrabilité et l'erreur de gloutonnerie. Chaque terme est discuté ci-après, un bref résumé étant fourni dans le tableau 1. Comme expliqué section 5.1, la borne de la RLP est similaire à celle de l'ICP, et la borne de l'ICP a été comparée à des instantiations spécifiques de l'IAP par Ghavamzadeh et Lazaric (2012). La discussion que nous proposons peut être vue comme étant complémentaire : nous considérons l'IAP plus généralement, nous proposons de nouveaux éléments de comparaison, et nous comparons les méthodes empiriquement.

Termes d'horizon. Les deux bornes ont une dépendance quadratique en l'horizon moyen $\frac{1}{1-\gamma}$. Pour la programmation dynamique approchée, Scherrer et Lesner (2012) ont montré que cette dépendance est fine, la seule solution connue pour l'améliorer étant d'introduire des politiques non-stationnaires. L'amélioration possible de la borne (par rapport à l'horizon) pour la RLP est une question ouverte. Nous suggérons section 5.3 une piste pour améliorer cette dépendance.

Coefficients de concentrabilité. Les deux bornes impliquent un coefficient de concentrabilité. Ils peuvent être comparés comme suit.

Théorème 4. *Dans tous les cas, nous avons :* $\left\|\frac{d_{\mu,\pi_*}}{\nu}\right\|_\infty \leq \frac{1}{1-\gamma} C_{\mu,\nu}$. *De plus, il existe toujours une mesure* ν *telle que* $\left\|\frac{d_{\mu,\pi_*}}{\nu}\right\|_\infty < \infty$ *(en choisissant* $\nu = d_{\mu,\pi_*}$*). Toutefois, il peut ne pas exister de mesure* ν *telle que* $C_{\mu,\nu} < \infty$.

Démonstration. Considérons l'inégalité de la première partie. Par définition de d_{μ,π_*} et en utilisant le fait que $d_{\mu,\pi_*} \geq (1-\gamma)\nu$, nous avons

$$\begin{aligned}
C_{\mu,\nu} &= (1-\gamma)^2 \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \gamma^{i+j} \sup_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{S}}} \left\|\frac{\mu(P_{\pi_*})^i (P_\pi)^j}{\nu}\right\|_\infty \\
&\geq (1-\gamma)^2 \left\|\sum_{i,j=0}^{\infty} \gamma^{i+j} \frac{\mu(P_{\pi_*})^{i+j}}{\nu}\right\|_\infty = (1-\gamma)\left\|\sum_{i=0}^{\infty} \gamma^{i} \frac{d_{\mu,\pi_*}(P_{\pi_*})^{i}}{\nu}\right\|_\infty \\
&\geq (1-\gamma)^2 \left\|\sum_{i=0}^{\infty} \gamma^{i} \frac{\mu(P_{\pi_*})^{i}}{\nu}\right\|_\infty = (1-\gamma)\left\|\frac{d_{\mu,\pi_*}}{\nu}\right\|_\infty .
\end{aligned}$$

Traitons à présent le second résultat. Considérons un PDM avec N états et N actions, avec $\mu = \delta_1$ un Dirac sur le premier état, et tel que de cet état

chaque action $a \in [1; N]$ mène dans l'état a de façon déterministe. Ecrivons $c = \sup_{\pi \in \mathcal{A}^{\mathcal{S}}} \| \frac{\mu P_\pi}{\nu} \|_\infty$ le premier terme définissant $C_{\mu,\nu}$. Pour chaque politique π, nous avons $\mu P_\pi \leq c\nu$. Ainsi, pour chaque action a nous avons $\delta_a \leq c\nu \Rightarrow 1 \leq c\nu(a)$. Finalement, $1 = \sum_{i=1}^N \nu(i) \geq \frac{1}{c} \sum_{i=1}^N 1 \Leftrightarrow c \geq N$. Comme cela est vrai pour un $N \in \mathbb{N}$ arbitraire, nous avons $c = \infty$ et donc $C_{\mu,\nu} = \infty$. □

La seconde partie de ce résultat stipule que nous pouvons avoir $\left\| \frac{d_{\mu,\pi_*}}{\nu} \right\|_\infty \ll C_{\mu,\nu}$, ce qui est clairement en faveur de la RLP (et de l'ICP, qui implique le même terme de concentrabilité).

Termes d'erreur. Les deux bornes impliquent un terme d'erreur. Les termes ϵ (RLP) et e (IAP) peuvent être rendus arbitrairement petits en augmentant l'effort computationnel (le temps et la mémoire alloués à l'algorithme ainsi que la quantité d'échantillons utilisés), bien que rien de plus ne puisse être dit en général, sans étudier d'instance spécifique des algorithmes (par exemple, choix du classifieur). Les termes définissant la "complexité gloutonne" des espaces de politiques peuvent être partiellement comparés. Comme ils sont basés sur des distributions qui peuvent être comparées ($d_{\nu,\pi} \geq (1-\gamma)\nu$), nous avons que pour tout espace de politiques Π (Ghavamzadeh & Lazaric, 2012),

$$\mathcal{E}'_\nu(\Pi) \leq \frac{\mathcal{E}_\nu(\Pi)}{1-\gamma}.$$

Cependant, ce résultat ne prend pas en compte le fait que la RLP (ou l'ICP pour la discussion de Ghavamzadeh et Lazaric (2012)) utilise des politiques stochastiques alors que l'IAP utilise des politiques déterministes. Cela rend ces termes non-comparables en général.

Expériences. Pour avoir une idée plus précise des performances relatives de l'IAP et de la RLP, nous appliquons les deux algorithmes à des PDM générés aléatoirement. Pour pouvoir estimer leur qualité, nous considérons des problèmes finis où la fonction de valeur exacte peut être calculée. Plus précisément, nous considérons les "Garnets", introduits par Archibald et al. (1995), qui sont une classe de PDM finis construits aléatoirement. Il ne correspondent à aucune application particulière et sont totalement abstraits, tout en restant représentatifs du type de PDM que l'on peut rencontrer en pratique. Dans nos expériences, un Garnet est paramétré par 4 variables et s'écrit $G(n_S, n_A, b, p)$: n_S est le nombre d'états, n_A est le nombre d'actions, b est le facteur de branchement spécifiant combien d'états suivants sont possibles pour chaque couple état-action (b états sont choisis aléatoirement et uniformément et les probabilités de transitions sont déterminées en générant, toujours aléatoirement et uniformément, $b-1$ points de coupure entre 0 et 1) et p est le nombre de fonctions de base (pour une approximation linéaire de la fonction de valeur). La récompense dépend de l'état : pour un Garnet généré aléatoirement, la récompense associée pour chaque état est tirée uniformément entre 0 et 1. La facteur d'actualisation est choisi égal à $0,99$ pour toutes les expériences.

Les algorithmes RLP et IAP nécessitent de calculer de façon répétée $\mathcal{G}_\Pi$. Autrement dit, ils doivent pouvoir faire appel à un opérateur glouton approché

appliqué à la valeur v_π d'une certaine politique π pour les distributions ν ou $d_{\nu,\pi}$, respectivement. Pour implémenter cet opérateur, nous calculons une estimation bruitée de la valeur v_π, le bruit $u(\iota)$ étant un blanc et uniforme d'amplitude ι, que nous projetons sur $\mathcal{H}$, l'espace engendré par les p fonctions de base choisies, par rapport à la norme quadratique pondérée par μ (projection que nous notons $\Pi_{\mathcal{H},\mu}$). Enfin, nous appliquons l'opérateur glouton (exact) à cette projection bruitée, ce qui définit donc notre espace de politiques. Pour résumer, un appel à l'opérateur glouton approché $\mathcal{G}_\Pi(\pi, \mu, \epsilon)$ revient à calculer $\mathcal{G}(\Pi_{\mathcal{H},\mu}(v_\pi + u(\iota)))$, avec $\mu = \nu$ (IAP) ou $\mu = d_{\nu,\pi}$ (RLP).

Dans nos expériences, nous considérons des Garnets avec $n_s \in \{50, 100, 200\}$ états, $n_a \in \{2, 5\}$ actions, et des facteurs de branchements $b \in \{1, 2, 10\}$. Pour chacune des 2×3^2 combinaisons résultantes possibles, nous générons 30 PDM aléatoires de façon indépendante, $(M_i)_{1 \leq i \leq 30}$. Pour chacun de ces PDM M_i, nous faisons 30 apprentissages indépendants de *(i)* l'IAP et *(ii)* la RLP, avec un taux d'apprentissage de 0.1. Pour chaque apprentissage et chaque algorithme, nous calculons la distance entre la valeur de la politique estimée et la valeur de la politique optimale, $(\Delta_j)_{1 \leq j \leq 30}$. La figure 1 montre les courbes d'apprentissage avec les statistiques des variables aléatoires associées. Dans ce large ensemble de problèmes, la RLP surpasse significativement l'IAP, tant en termes de performance moyenne qu'en termes de variabilité (entre différents apprentissages et entre différents problèmes). Notons toutefois que l'IAP nécessite moins d'itérations (l'erreur moyenne décroît lors des quelques premières itérations, ce qui n'apparaît pas sur la figure 1 en raison de l'échelle). Cela confirme l'importance du meilleur coefficient de concentrabilité de la RLP, dans la mesure où c'est théoriquement son meilleur avantage par rapport à l'IAP.

5.3 Conséquences pratiques et théoriques de l'analyse

Finalement, cette section présente quelques importantes conséquences de notre analyse, en particulier du théorème 3.

Espace de politiques riche et équivalence entre optimalité locale et globale. Si l'espace de politiques est très riche, il est facile de montrer que tout optimum local est en fait un optimum global (ce résultat étant un corollaire direct du théorème 3).

Théorème 5. *Soit $\nu > 0$ une distribution. Supposons que l'espace de politiques est riche, dans le sens où $\mathcal{E}_\nu(\Pi) = 0$, et que π est un optimum local (exact) de J_ν ($\epsilon = 0$). Alors, nous avons $v_\pi = v_*$.*

Si ce résultat est connu dans le cas de politiques tabulaires, il est à notre connaissance nouveau dans ce cadre général (en reconnaissant toutefois que $\mathcal{E}_\nu(\Pi) = 0$ est une hypothèse très forte).

Choix de la distribution d'échantillonnage. Etant donné le résultat du théorème 3, et comme cela a déjà été mentionné à propos de l'ICP par Kakade et Langford (2002) (la borne étant similaire), si l'on veut optimiser la politique par rapport à une distribution μ (c'est-à-dire, de façon à avoir $\mu(v_* - v_\pi)$ petit), alors il faut optimiser la fonction objectif J_ν avec la distribution $\nu \simeq d_{\mu,\pi_*}$

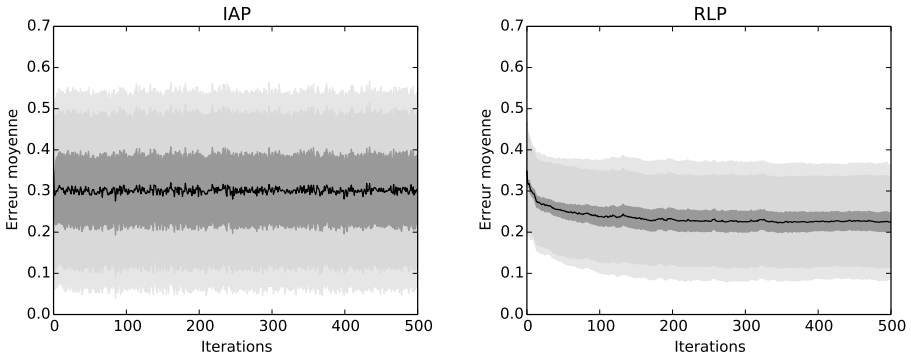

FIGURE 1 – **Courbes d'apprentissage pour l'IAP et la RLP.** Les PDM sont indépendants et identiquement distribués (i.i.d.) selon la loi de M_1. Conditionné à un PDM M_i, les mesures d'erreur sont i.i.d., de loi Δ_1. La ligne centrale est une estimation de l'erreur globale moyenne $E[\Delta_1]$. Les trois régions grises (de la plus foncée à la plus claire) sont des estimations de la variabilité (entre PDM) de l'erreur moyenne $Ect[E[\Delta_1|M_1]]$, la moyenne (entre PRM) de l'écart type de l'erreur $E[Ect[\Delta_1|M_1]]$, et la variabilité (entre PDM) de l'écart type des erreurs $Ect[Ect[\Delta_1|M_1]]$.

(de façon à minimiser le coefficient de concentrabilité $\left\| \frac{d_{\mu,\pi_*}}{\nu} \right\|_\infty$). Idéalement, il faudrait donc échantillonner les états selon des trajectoires obtenues en suivant la politique π_* à partir d'états initiaux distribués selon μ. C'est bien sûr irréaliste, dans la mesure où π_* n'est pas connue. Toutefois, des solutions pratiques peuvent être envisagées.

Premièrement, cela signifie qu'il faut échantillonner les états dans les parties "intéressantes" de l'espace d'état, c'est-à-dire où l'on pense que la politique optimale va nous conduire en partant d'états échantillonnés selon μ. C'est le type d'information qu'un expert du domaine d'intérêt devrait naturellement pouvoir fournir. De plus, bien que nous laissions l'étude précise de cette idée en perspective, une approche naturelle et pratique pour choisir cette distribution ν serait de calculer une séquence de politiques $\pi_1, \pi_2, \ldots$ telles que pour tout i, π_i est un optimum local de $\pi \mapsto J_{d_{\nu,\pi_{i-1}}}(\pi)$, c'est-à-dire le critère pondéré par les régions visitées par la politique précédente π_{i-1}. Il serait particulièrement intéressant d'étudier si la convergence d'un tel processus itératif (si convergence il y a) mène à des garanties intéressantes.

On peut également remarquer que le théorème 3 peut être facilement réécrit plus généralement pour n'importe quelle politique. Si π est un ϵ-optimum local de J_ν sur Π, alors pour toute politique stochastique π' nous avons

$$\mu v_{\pi'} \leq \mu v_\pi + \frac{1}{1-\gamma} \left\| \frac{d_{\mu,\pi'}}{\nu} \right\|_\infty \left(\frac{\mathcal{E}_\nu(\Pi)}{1-\gamma} + \epsilon \right).$$

Ainsi, il est possible d'échantillonner des trajectoires en utilisant un contrôleur

acceptable (et connu) π', de façon à obtenir des échantillons d'états pour optimiser $J_{d_{\nu,\pi'}}$. Plus généralement, si l'on sait où une bonne politique π' mène le système à partir d'une distribution initiale μ, on peut apprendre une politique π qui a la garantie d'être approximativement aussi bonne (et potentiellement meilleure).

Un meilleur problème d'apprentissage ? Avec le résultat du théorème 3, nous avons une dépendance quadratique de la borne en l'horizon moyen $\frac{1}{1-\gamma}$. Pour la programmation dynamique approchée, il est bien connu que cette dépendance est fine (Bertsekas, 1995 ; Scherrer & Lesner, 2012). Pour l'instant, c'est une question ouverte pour la recherche locale de politique. Cependant, il est possible d'améliorer la borne. Nous avons montré que l'ϵ-optimalité locale d'une politique π impliquait qu'une version relâchée de l'optimalité de Bellman était satisfaite, $\pi \in \mathcal{G}_\Pi(\pi, d_{\nu,\pi}, \epsilon)$, ce qui implique le théorème 3. Le résultat suivant, qui implique une version légèrement plus simple de cette caractérisation relâchée de l'optimalité, peut être prouvé de façon similaire au théorème 2 :

$$\text{Si } \pi \in \mathcal{G}_\Pi(\pi, \nu, \epsilon) \text{ alors } \mu v_{\pi'} \leq \mu v_\pi + \frac{1}{1-\gamma} \left\| \frac{d_{\mu,\pi'}}{\nu} \right\|_\infty (\mathcal{E}_\nu(\Pi) + \epsilon).$$

Une politique satisfaisant cette condition aurait donc une dépendance améliorée à l'horizon ($\frac{1}{1-\gamma}$ au lieu de $\frac{1}{(1-\gamma)^2}$). Pour l'instant, nous ne savons pas s'il existe un algorithme efficace pour calculer une politique satisfaisant $\pi \in \mathcal{G}_\Pi(\pi, \nu, \epsilon)$. La garantie précédente suggère que résoudre ce problème pourrait fournir de meilleures politiques que ne le font les approches usuelles (PDA et RLP).

6 Conclusion

Ces dernières années, les algorithmes de recherche locale de politique se sont avérés être des alternatives pratiques viables aux algorithmes plus traditionnels de programmation dynamique approchée. L'obtention de garanties globales de performance pour la RLP, probablement considérée comme un cas désespéré, n'a à notre connaissance jamais été considérée dans la littérature. Dans cet article, nous avons démontré un résultat surprenant : *tout algorithme de recherche locale de politique*, dans la mesure où il est capable de fournir un *optimum local approché* de $J_\nu(\pi)$, bénéficie de *garanties globales de performance*, similaires à celles de la programmation dynamique approchée. Toutefois, cela repose sur une forte hypothèse de convexité de l'espace de recherche, non satisfaite par les algorithmes standard. Affaiblir cette hypothèse est une perspective intéressante (mais difficile, dans la mesure où cette convexité est au cœur de notre analyse).

Pour palier ce problème, nous avons proposé d'appliquer AnyBoost.L1 au problème de recherche local de politique. Si c'est finalement une légère généralisation de l'itération conservative de la politique, et non un nouvel algorithme, notre travail fournit une connexion originale entre la recherche locale de politique, le *boosting* et la programmation dynamique approchée. De plus, cela suggère quelques pistes de recherche. Tout d'abord, AnyBoost.L1 (et donc l'ICP) est une approche naturelle pour travailler dans un espace convexe de

politiques. Une alternative intéressante serait d'étudier la paramétrisation d'un espace convexe. En étant capable de fournir une paramétrisation non triviale, il serait possible d'utiliser beaucoup des algorithmes de RLP de la littérature (par exemple, les algorithmes acteur-critique). Notre analyse suggère également qu'il serait plus intéressant de concevoir des algorithmes qui cherchent une politique π satisfaisant $\pi \in \mathcal{G}_\Pi(\pi, \nu, \epsilon)$, plutôt que de chercher un maximum local de J_ν, comme cela mène à une meilleure borne (dépendance linéaire en l'horizon moyen). Travailler dans cette direction est une perspective de recherche intéressante. Enfin, nos experiences sur les Garnets montrent que la RLP fournit de meilleurs résultats que l'IAP. Approfondir la comparaison de ces approches sur de plus grands problèmes est une perspective naturelle.

Références

Antos A., Szepesvari C., & Munos R. (2008). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning Journal*, *71*, 89-129.

Archibald T., McKinnon K., & Thomas L. (1995). On the Generation of Markov Decision Processes. *Journal of the Operational Research Society*, *46*, 354-361.

Baxter J., & Bartlett P. L. (2001). Infinite-horizon gradient-based policy search. *Journal of Artificial Intelligence Research (JAIR)*, *15*, 319–350.

Bertsekas D. (1995). *Dynamic Programming and Optimal Control.* Athena Scientific.

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

Bhatnagar S., Sutton R. S., Ghavamzadeh M., & Lee M. (2007). Incremental natural actor-critic algorithms. In *Advances in neural information processing systems (nips).*

Fern A., Yoon S., & Givan R. (2006). Approximate Policy Iteration with a Policy Language Bias : Solving Relational Markov Decision Processes. *Journal of Artificial Intelligence Research (JAIR)*, *25*, 75-118.

Fix J., & Geist M. (2012). Monte-Carlo Swarm Policy Search. In *Symposium on Swarm Intelligence and Differential Evolution.* Springer.

Geist M. (2015). Soft-max boosting. *Machine Learning*, *100*(2), 305-332.

Ghavamzadeh M., & Lazaric A. (2012). Conservative and Greedy Approaches to Classification-based Policy Iteration. In *Conference on artificial intelligence (aaai).*

Heidrich-Meisner V., & Igel C. (2008). Evolution strategies for direct policy search. In *International conference on parallel problem solving from nature (ppsn x)*, pp. 428–437.

Kakade S. (2001). A Natural Policy Gradient. In *Advances in neural information processing systems (nips).*

Kakade S., & Langford J. (2002). Approximately optimal approximate reinforcement learning. In *International conference on machine learning (icml).*

Kober J., & Peters J. (2011). Policy Search for Motor Primitives in Robotics. *Machine Learning*, 171-203.

Lagoudakis M., & Parr R. (2003a). Least-squares policy iteration. *Journal of Machine Learning Research (JMLR)*, *4*, 1107–1149.

Lagoudakis M., & Parr R. (2003b). Reinforcement learning as classification : Leveraging modern classifiers. In *International conference on machine learning (icml).*

Lazaric A., Ghavamzadeh M., & Munos R. (2010). Analysis of a classification-based policy iteration algorithm. In *International conference on machine learning (icml).*

Lazaric A., Ghavamzadeh M., & Munos R. (2011). Finite-sample analysis of least-squares policy iteration. *Journal of Machine learning Research*, *13*, 3041-3074.

Mason L., Baxter J., Bartlett P., & Frean M. (1999). *Boosting algorithms as gradient descent in function space* (Rapport technique). Australian National University.

Munos R. (2003). Error bounds for approximate policy iteration. In *International conference on machine learning (icml).*

Munos R. (2007). Performance bounds in Lp norm for approximate value iteration. *SIAM Journal on Control and Optimization.*

Peters J., & Schaal S. (2008). Natural Actor-Critic. *Neurocomputing*, *71*, 1180-1190.

Puterman M. L. (1994). *Markov decision processes : Discrete stochastic dynamic programming.* Wiley-Interscience.

Scherrer B., Gabillon V., Ghavamzadeh M., & Geist M. (2012). Approximate Modified Policy Iteration. In *International Conference on Machine Learning (ICML).*

Scherrer B., Ghavamzadeh M., Gabillon V., Lesner B., & Geist M. (2015). Approximate Modified Policy Iteration and its Application to the Game of Tetris. *Journal of Machine Learning Research (JMLR)*, *16*, 1629-1676.

Scherrer B., & Lesner B. (2012). On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes. In *Advances in neural information processing systems (nips).*

Sutton R., & Barto A. (1998). *Reinforcement Learning, An introduction.* The MIT Press.

Sutton R., McAllester D., Singh S., & Mansour Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in neural information processing systems (nips).*