La complexité d'échantillonnage des processus de décision markovien robuste est inférieure à celle des processus de décision markovien classique.

Pierre Clavier [1,2] & Laixi Shie [3] & Erwan Le Pennec [1]

[1] *CMAP, Ecole Polytechnique, pierre.clavier@poltechnique.edu ; erwan.le-pennec@polytechnique.edu ;*
[2] *INRIA Paris, HeKA*
[3] *California Institute of Technology, laixis@caltech.edu*

Résumé. Ce travail étudie la complexité de l'échantillonnage des processus de décision de Markov robustes (RMDP) L'objectif est d'optimiser les performances dans le pire des cas lorsque l'environnement se situe dans un ensemble d'incertitudes défini entourant un certain processus de Markov décisionnels (MDP) dit nominal. Malgré des efforts récents, la complexité d'échantillonnage des processus décisionnels de Markov robustes reste indéterminée. Bien que cette question ait été étudiée dans certains cas spécifiques, la généralisation des résultats existants reste incertaine, en particulier en comparaison avec les MDPs standards. En supposant l'accès à un modèle génératif qui échantillonne à partir du MDP nominal, nous examinons la complexité d'échantillonnage des RMDPs en utilisant une norme arbitraire comme fonction de "distance" pour l'ensemble d'incertitude, sous deux conditions couramment adoptées *sa*-rectangulaire et *s*-rectangulaire. Nous fournissons une borne supérieure quasi-optimale et une borne inférieure minimax correspondante pour les scénarios *sa*- rectangulaires. Pour les scénarios *s*-rectangulaires, nous améliorons la borne supérieure de pointe et dérivons une borne inférieure pour la norme L_∞ et L_1 . Les résultats impliquent que les RMDPs peuvent être plus efficaces en termes d'échantillonnage que les MDP standards sous des normes générales dans les cas *sa*- et *s*-rectangulaires.

Mots-clés. Processus de Markov Décisionnels Robustes, complexité d'échantillonnage,

L'apprentissage par renforcement (RL) (Sutton, 1988) est un paradigme clé de l'apprentissage automatique, particulièrement remarquable pour son succès dans les applications pratiques. Le cadre de l'apprentissage par renforcement est souvent modélisé dans le contexte d'un processus de décision de Markov (MDP) et se concentre sur l'apprentissage de stratégies de prise de décision efficaces fondées sur des interactions avec un environnement. Cependant, les travaux de Mannor et al. (2004) ont mis en évidence une vulnérabilité du RL, révélant sa sensibilité aux erreurs d'estimation dans les probabilités de récompense et de transition. Un exemple typique des problèmes rencontrés par le RL est lorsque qu'en raison d'un écart entre les simulations et les applications réelles (dit sim-to-real), les politiques apprises dans des environnements idéalisés et simulés peuvent échouer de manière catastrophique lorsqu'elles sont déployées dans des environnements avec de légers changements ou des perturbations adverses (Klopp et al., 2017; Mahmood et al., 2018).

Pour résoudre ce problème, les MDP robustes (RMDP), proposés par Iyengar (2005) et Nilim and El Ghaoui (2005), ont fait l'objet d'une attention considérable. Les RMDP sont

formulés comme des problèmes max-min, recherchant des politiques qui résistent aux erreurs d'estimation du modèle dans un ensemble d'incertitudes spécifié. Malgré les avantages de la robustesse, la résolution des RMDP est NP-hard pour les ensembles d'incertitude généraux (Nilim and El Ghaoui, 2005). Pour surmonter cette difficulté, l'hypothèse de rectangularité des ensemble d'incertitude est souvent adoptée, les ensembles d'incertitude sont ainsi structurés comme des produits de sous-ensembles indépendants pour chaque état ou paire état-action, désignés par les hypothèses s-rectangulaire ou sa-rectangulaire (voir les définitions (5) et (7)). Ces deux hypothèses facilitent l'utilisation de méthodes telles que l'itération robuste de la valeur et l'itération robuste de la politique, en préservant de nombreuses propriétés structurelles des MDP (Ho et al., 2021). Les ensembles s-rectangulaires, bien que moins restrictifs, posent de plus grands défis, tandis que les ensembles sa-rectangulaires permettent des politiques déterministes apparentées aux MDP non robustes (Wiesemann et al., 2013). Enfin, il est important de noter que, si l'incertitude de la récompense peut facilement être gérée, il est plus difficile de gérer l'incertitude du noyau de transition, (Kumar et al., 2022; Derman et al., 2021).

La question de l'efficacité de l'échantillonnage est centrale dans les problèmes de RL allant de la pratique à la théorie. Bien que les bornes minimax soient atteintes dans les travaux de Azar et al. (2013); Li et al. (2023) dans le contexte des MDP classiques, cet objectif n'est pas encore atteint en général, dans le contexte des RMDP. Plus précisément, il existe des travaux antérieurs sur la complexité de l'échantillon de RL robuste pour quelques divergences spécifiques telles que la variation totale TV, L_p, χ^2, KL, et Wasserstein (Yang et al., 2022a; Zhou et al., 2021; Panaganti and Kalathil, 2022), alors que de tels résultats restent incertains pour des classes plus générales de divergences. À ce jour, à notre connaissance, les résultats de la complexité de l'échantillon qui atteignent l'optimalité minimax pour tout rayon d'incertitude sont limités à un seul cas à savoir la distance TV dans le cas sa-rectangulaire. (Shi et al., 2023).

Dans ce travail, nous nous concentrons sur la question de la complexité d'échantillonnage des RMDP avec une norme arbitraire. Cette généralisation est intéressante à la fois en pratique et en théorie. En pratique, de nombreuses applications reposent des approches qui impliquent des normes arbitraires et ad-hoc, différentes de celles qui ont déjà été étudiées théoriquement. Par exemple, le contrôle robuste peut être spécifique à une tâche, utilisant une distance de Mahalanobis (Jiang and Zhang, 2018) pour construire les ensembles d'incertitude. D'un point de vue théorique, il est intéressant d'étudier le coût statistique de la robustesse en RL dans des scénarios plus généraux, ce qui conduit à deux questions ouvertes auxquelles nous essayerons de répondre. L'une d'entre elles concerne sur la complexité d'échantillonnage pour la résolution du RL robuste par rapport à la résolution de RL classique. En particulier, pour le cas spécifique de la distance TV, Shi et al. (2023) a montré que la complexité d'échantillonnage pour résoudre le RL robuste est au plus la même et parfois (lorsque le niveau d'incertitude est relativement grand) plus petite que celle de du RL standard. Cela motive la question ouverte suivante :

Le RL robuste est-il plus efficace en termes d'échantillons que le RL standard pour des normes générales ?

Une seconde question concerne les comparaisons entre la complexité d'échantillonnage de

la résolution des RMDPs s-rectangulaires et celle de la résolution de RMDPs avec l'hypothèse de sa-rectangularité. On peut souligner que les RMDPs s-rectangulaires ont des formulations d'optimisation plus compliquées avec des variables supplémentaires (niveaux d'incertitude pour chaque action) à optimiser. Cela conduit à une classe plus riche de politique optimale, à savoir des politiques stochastiques dans les cas s-rectangulaires, contrairement à la classe des politiques déterministes pour les cas sa-rectangulaires. En outre, la limite supérieure de la complexité d'échantillonnage existante pour la résolution des RMDP s-rectangulaires est plus grande que celle de la résolution des RMDP sa-rectangulaires (Yang et al., 2022a) pour les cas étudiés. Cela motive la questions suivante:

La résolution de s-rectangulaires RMDPs nécessite-t-elle, en effet, plus d'échantillons que la résolution de sa-rectangulaires RMDPs avec des normes générales ?

Contributions. Dans ce travail, nous abordons chacune des deux questions discutées ci-dessus. En particulier, nous fournissons la première analyse de complexité d'échantillon pour les RMDP avec des normes générales sous les conditions de s- et sa-rectangularité. Par commodité, nous présentons une comparaison détaillée de l'état de l'art existant et nos résultats dans le tableau 1 et discutons des contributions et de leurs implications ci-dessous. En ce qui concerne la première question, nous illustrons nos résultats dans les cas sa- et s-rectangulaire dans la Figure 1. Dans le cas de la sa-rectangularité, nous dérivons une borne supérieure de complexité d'échantillon pour les RMDP en utilisant des normes générales (Théorème 2.1) de l'ordre de :

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right). \tag{1}$$

De plus, nous fournissons une borne inférieure minimax correspondante (Théorème 2.2) qui confirme la quasi-optimalité de la borne supérieure pour la quasi-totalité de la plage du niveau d'incertitude. Cela correspond à la complexité d'échantillonnage quasi-optimale dérivée dans Shi et al. (2023) pour le cas spécifique de la distance TV et sa rectangulaire, tout en étant valable pour n'importe quelle norme arbitraire. Dans le cas d'une s-rectangularité, nous fournissons une borne supérieure de complexité pour la résolution des RMDP avec n'importe quelle norme générale de l'ordre de :

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right).$$

Ce résultat améliore l'état de l'art antérieur $\widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ dans Clavier et al. (2023) pour le cas spécifique de L_p lorsque $\tilde{\sigma} \lesssim 1-\gamma$ par au moins un facteur de $1/(1-\gamma)$. De plus, nous présentons une borne inférieure pour un cas représentatif avec la norme L_∞, qui atteint la borne supérieure. À notre connaissance, il s'agit de la première borne inférieure pour la résolution de RMDPs avec s-rectangularité. Enfin, nous sommes en mesure de jeter un nouvel éclairage sur la seconde question grâce à nos nouveaux résultats. En particulier, comme l'illustre la figure 1, nos résultats mettent en évidence le fait que le RL robuste est au moins aussi efficace que le RL standard pour les normes générales, et qu'il peut parfois l'être davantage. Ce résultat a une importance pratique considérable et constitue une motivation

Résultat	Reference	Distance	*sa*-rectangulaire		*s*-rectangulaire	
			$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma < \sigma_{\max}$	$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma < \sigma_{\max}$
Borne supérieure	Yang et al. (2022b)	TV	$\frac{S^2A(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A^2(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A^2(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$
	Panaganti and Kalathil (2022)	TV	$\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$	×	×
	Shi et al. (2023)	TV	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$	×	×
	Clavier et al. (2023)	L_p	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^4\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^4\varepsilon^2}$
	Ce travail	$\|\cdot\|$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$
Borne inférieure	Yang et al. (2022b)	TV	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA(1-\gamma)}{\sigma^4\varepsilon^2}$	×	×
	Shi et al. (2023)	TV	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	×	×
	Ce travail	L_∞	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$

Table 1: Comparaisons avec des résultats antérieurs (aux termes logarithmiques près) de la complexité nécessaire pour obtenir la recherche une politique optimale à ε près pour un processus de décision de Markov robuste, où σ est le rayon de l'ensemble d'incertitude et $\sigma_{\max}$ défini dans 2.1.

essentielle pour l'utilisation et l'étude de la robustesse. Plus précisément, le RL robuste ne réduit pas seulement la vulnérabilité du RL aux erreurs d'estimation et aux écarts entre les simulations et le réel, mais conduit également à une meilleure efficacité en termes de complexité d'échantillonnage. En termes de comparaison des implications statistiques de la *sa*- et de la *s*- rectangularité, nos résultats montrent que la résolution de RMDPs *s*-rectangulaires n'est pas plus difficile que la résolution de RMDPs *sa*-rectangulaires en termes d'exigence d'échantillon (voir le théorème 2.3)

1 Formulation du problème : Processus de décision de Markov robustes

Processus de décision de Markov (MDP) standard. Un MDP actualisé à horizon infini est représenté par $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, r)$, où $\mathcal{S} = \{1, \cdots, S\}$ et $\mathcal{A} = \{1, \cdots, A\}$ sont les espaces d'état et d'action finis, respectivement, $\gamma \in [0, 1)$ est le facteur d'actualisation, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ représente le noyau de transition des probabilités, et $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ est la fonction de récompense immédiate, qui est supposée être déterministe. De plus, nous supposons que la fonction de récompense est bornée en $(0, 1)$ sans perte de généralité. La politique que nous recherchons est définie par $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, qui spécifie la probabilité de sélection d'action sur l'espace d'action pour tous les états. Enfin, pour caractériser la récompense cumulative, la fonction de valeur $V^{\pi,P}$ pour toute politique π sous le noyau de transition P est définie par $\forall s \in \mathcal{S}$

$$V^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\middle|\, s_0 = s\right]. \tag{2}$$

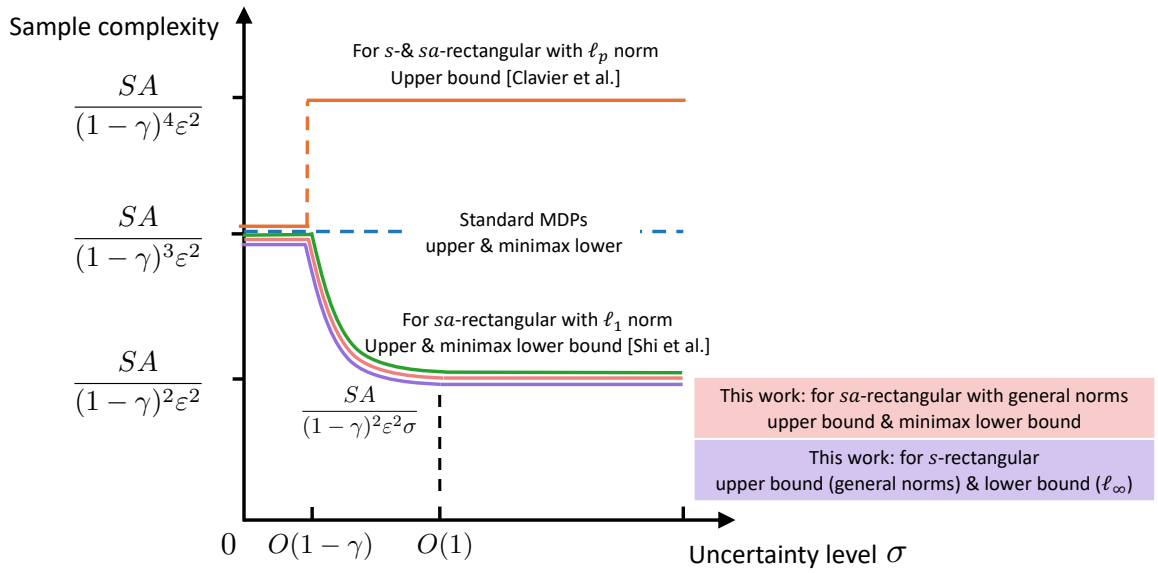

Figure 1: Les résultats de la complexité de l'échantillon pour les RMDP avec sa- et s-rectangularité avec des normes générales et des comparaisons avec les arts antérieurs (Shi et al., 2023) (pour la norme ℓ_1, ou appelée distance de variation totale) et (Clavier et al., 2023) (pour la norme L_p avec $1 \leq p \leq \infty$).

L'espérance est prise sur le caractère aléatoire de la trajectoire $\{s_t, a_t\}_{t=0}^{\infty}$ générée par l'exécution de la politique π sous le noyau de transition P, de sorte que $a_t \sim \pi(\cdot \,|\, s_t)$ et $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ pour tout $t \geq 0$. De la même manière, la Q-fonction $Q^{\pi,P}$ associée à toute politique π sous le noyau de transition P comme : $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$: est définie comme suit

$$Q^{\pi,P}(s,a) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\middle|\, s_0, a_0 = s, a\right], \tag{3}$$

avec une espérance prise sur le caractère aléatoire de la trajectoire sous la politique π.

RMDPs robustes du point de vue de la distribution Nous considérons des MDP robustes sur le plan de la distribution (RMDP) dans le cadre d'un horizon infini actualisé, dénotés par $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}_{\|.\|}(P^0), r\}$, où $\mathcal{S}, \mathcal{A}, \gamma, r$ sont les mêmes ensembles et paramètres que dans les PDM standard. La principale différence par rapport aux MDP standard est qu'au lieu d'utiliser un noyau de transition fixe P, il permet au noyau de transition d'être choisi arbitrairement dans un ensemble d'incertitude prescrit $\mathcal{U}^{\sigma}_{\|.\|}(P^0)$ centré autour d'un noyau *nominal* $P^0 : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, où l'ensemble d'incertitude est spécifié en utilisant une métrique de distance $\|.\|$ de rayon $\sigma > 0$. Cette définition est générale et inclut TV, $L_p, p > 1$ et toute norme, telle que la distance de Mahalanobis, par exemple. Cependant, elle n'inclut pas les divergences telles que KL et χ^2. En particulier, étant donné le noyau de transition nominal P^0 et un certain niveau d'incertitude σ, l'ensemble d'incertitude–avec une norme arbitraire $\| \, \| : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^+$ ou de $\mathbb{R}^{\mathcal{SA}}$ dans le cas s-rectangulaire, est spécifié par

$\mathcal{U}_{\|.\|^\sigma(P^0):=\times_{s,a}\,\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a})}$

$$\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \left\|P_{s,a} - P^0_{s,a}\right\| \leq \sigma\right\}, \tag{4}$$

où nous désignons un vecteur du noyau de transition P ou P^0 au couple état-action (s,a) respectivement par

$$P_{s,a} \coloneqq P(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}, \quad P^0_{s,a} \coloneqq P^0(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}. \tag{5}$$

En d'autres termes, l'incertitude est imposée de manière découplée pour chaque paire état-action, obéissant à la soi-disant *sa*-rectangularité (Zhou et al., 2021; Wiesemann et al., 2013). Dans ce travail, nous considérerons toute norme arbitraire définie comme $\|.\|$. Plus généralement, nous définissons les RMPD s-rectangulaires comme $\mathcal{U}^{\sigma}_{\|.\|}(P) = \otimes_s \mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P_s)$, pour la norme arbitraire $\|.\|$, à l'aide de la définition suivante

$$P_s \coloneqq P(\cdot,\cdot \,|\, s) \in \mathbb{R}^{1\times SA}, \quad P^0_s \coloneqq P^0(\cdot,\cdot \,|\, s) \in \mathbb{R}^{1\times SA}. \tag{6}$$

L'incertitude est imposée de manière découplée pour chaque paire d'états, et un budget fixe donné à un état pour toutes les actions est défini. Pour obtenir une signification similaire pour le rayon de la boule entre les hypothèses *sa*-rectangulaire et s-rectangulaire, nous devons renormaliser le rayon en fonction de la norme comme dans Yang et al. (2022a). L'ensemble d'incertitude s est alors défini en utilisant le rayon renormalisé $\tilde{\sigma}$ comme suit

$$\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P_s) \coloneqq \left\{P'_s \in \Delta(\mathcal{S})^{\mathcal{A}} : \|P'_s - P_s\| \leq \tilde{\sigma} = \sigma \left\|1\right\| \right\}, \tag{7}$$

où 1 représente le vecteur unitaire. Dans le cas spécifique des normes L_1, L_p et L_∞, $\tilde{\sigma}$ est égal à $|\mathcal{A}|, |\mathcal{A}|^{1/p}$ et 1. Notez que cette échelle permet une comparaison équitable entre les PDM *sa*- et s-rectangulaires. Dans les RMDP, nous nous intéressons à la performance la plus défavorable d'une politique π sur tous les noyaux de transition possibles dans l'ensemble d'incertitude. Cette performance est mesurée par la fonction de valeur robuste $V^{\pi,\sigma}$ et la fonction Q robuste $Q^{\pi,\sigma}$ dans $\mathcal{M}_{\mathsf{rob}}$, définies respectivement comme $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$

$$V^{\pi,\sigma}(s) \coloneqq \inf_{P\in\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0)} V^{\pi,P}(s), \tag{8}$$

$$Q^{\pi,\sigma}(s,a) \coloneqq \inf_{P\in\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0)} Q^{\pi,P}(s,a). \tag{9}$$

De la même manière, nous définissons la fonction de valeur de s-rectangularité.

$$V^{\pi,\sigma}_s(s) \coloneqq \inf_{P\in\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P^0)} V^{\pi,P}(s), \tag{10}$$

Noyau nominal empirique. Le noyau empirique de transition nominale $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$ peut être construit sur la base de la fréquence empirique des transitions d'état, c'est-à-dire ,$\forall (s,a) \in \mathcal{S}\times\mathcal{A}$

$$\widehat{P}^0(s' \,|\, s,a) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left\{s_{i,s,a} = s'\right\}, \tag{11}$$

qui mène au RMDP empirique $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}_{\|.\|}(\widehat{P}^0), r\}$. Toutes les quantités définies avec opérateur $\hat{.}$ sont définies comme précédemment, mais dans le MRPD empirique comme les Q fonction ou fonction de valeur empirique $\hat{Q}$ et $\hat{V}$.

2 Garanties théoriques

2.1 Ensemble d'incertitude *sa*-rectangulaire avec normes générales

Pour commencer, nous considérons les RMDPs avec *sa*-rectangularité avec des normes arbitraires. Nous commençons par fournir la limite supérieure de complexité d'échantillon.

Theorem 2.1 (Borne supérieure pour l'hypothèse *sa*-rectangulaire.) *Nous considérons l'ensemble d'incertitude $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\cdot)$ associé à la norme $\|\cdot\|$ et notons $\sigma_{\max} \coloneqq \max_{p_1,q_1\in\Delta(\mathcal{S})} \|p_1 - p_2\|$ le rayon maximal. Pour un niveau de confiance $\delta \in (0,1)$, un facteur d'actualisation $\gamma \in \left[\frac{1}{4}, 1\right)$, et un rayon $\sigma \in (0, \sigma_{\max}]$ nous définissons la politique oracle dans le MDP empirique $\widehat{\pi}$ qui est le résultat du problème d'optimisation dans le RMPDS empirique avec une erreur $\varepsilon_{\mathsf{opt}}$ tel que $\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \le \varepsilon_{\mathsf{opt}}$. Avec probabilité au moins $1-\delta$, on a*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \le \varepsilon + \frac{7\varepsilon_{\mathsf{opt}}}{1-\gamma} \tag{12}$$

pour tout $\varepsilon \in (0, \sqrt{1/\max\{1-\gamma,\sigma\}}]$, si le nombre d'échantillons obéit

$$NSA \gtrsim \frac{C_1 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}$$

avec C_1 une constante universelle positive.

Nous introduisons la borne inférieure minimax-optimale suivante pour les normes générales afin de vérifier de l'intérêt de la borne supérieure ci-dessus.

Theorem 2.2 (Borne inférieure pour l'hypothèse *sa*-rectangulaire) *En considérant l'ensemble d'incertitude $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\cdot)$ associé à une norme arbitraire $\|\cdot\|$ et le rayon maximal $\sigma_{\max} \coloneqq \max_{p_1,q_1\in\Delta(\mathcal{S})} \|p_1 - p_2\|$ nous définissons le tuple $(S, A, \gamma, \sigma, \varepsilon)$, avec $\gamma \in \left[\frac{1}{2}, 1\right)$, $\sigma \in (0, \sigma_{\max}(1-c_0)]$ et $0 < c_0 \le \frac{1}{8}$ une constante positive et $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. Nous pouvons contruire deux RMPDs $\mathcal{M}_0, \mathcal{M}_1$ tel que ayant donné un jeux de données avec N échantillons indépendants pour chaque couple état-action échantillonné du MDP nominal (pour chaque $\mathcal{M}_0$ ou $\mathcal{M}_1$ respectivement), on a*

$$\inf_{\widehat{\pi}} \max_{\mathcal{M}\in\{\mathcal{M}_0,\mathcal{M}_1\}} \left\{ \mathbb{P}_{\mathcal{M}}\Big(\max_{s\in\mathcal{S}} \left[V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s)\right] > \varepsilon \Big) \right\} \ge \frac{1}{8},$$

tant que

$$NSA \le \frac{C_2 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}.$$

Ici C_2 est une constante universelle positive, l'infimum est pris sous tous les estimateurs $\widehat{\pi}$, et $\mathbb{P}_0$ (respectivement. $\mathbb{P}_1$) dénote la probabilité lorsque le RMDP est $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

2.2 Ensemble d'incertitude s-rectangulaire avec normes générales

Pour continuer, nous passons au cas où l'ensemble d'incertitude est construit sous s-rectangularité. Le théorème suivant présente la limite supérieure de la complexité de l'échantillon pour l'apprentissage d'une politique optimale *varepsilon* pour les RMDP avec s-rectangularité.

Theorem 2.3 (Borne supérieure pour l'hypothèse s-rectangulaire.) *Nous considérons ici l'ensemble d'incertitude $\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|\cdot\|}(\cdot)$ sous l'hypothèse s-rectangulaire. Nous définissons également le facteur d'actualisation $\gamma \in \left[\frac{1}{4}, 1\right)$, le rayon d'incertitude $\tilde{\sigma} = \sigma \|1\|$ et le niveau de confiance $\delta \in (0,1)$. Nous définissons la politique oracle dans le MDPs empirique $\widehat{\pi}$ qui est le résultat du problème d'optimisation dans le RMPDS empirique avec une erreur $\varepsilon_{\mathsf{opt}}$ tel que $\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \leq \varepsilon_{\mathsf{opt}}$. Avec probabilité au moins $1-\delta$, on a*

$$\forall s \in \mathcal{S}: \quad V^{\star,\tilde{\sigma}}(s) - V^{\widehat{\pi},\tilde{\sigma}}(s) \leq \varepsilon + \frac{7\varepsilon_{\mathsf{opt}}}{1-\gamma}$$

tant que le nombre d'échantillons obéit

$$NSA \gtrsim \frac{C_3 SA}{(1-\gamma)^2 \varepsilon^2} \min\left\{ \frac{1}{\max\{1-\gamma, \sigma\}}, \frac{1}{\sigma \min_{s\in\mathcal{S}} \left\{ \|\pi_s^*\|_* \|1\|, \|\hat{\pi}_s\|_* \|1\| \right\}} \right\}, \tag{13}$$

avec C_3 une constante positive universelle.

où $\hat{\pi}_s \in \Delta_A$ dénote la politique des RMPD empiriques à l'état s, $\pi_s^* \in \Delta_A$ la politique optimale à l'état s et $\|.\|_*$ la norme duale. De plus, nous fournissons les bornes inférieures pour les normes L_∞ et L_1 dans le théorème suivant :

Theorem 2.4 (Borne inférieure pour l'hypothèse s-rectangulaire.) *Ici, le théorème est valable pour la norme infinie L_∞ et la norme L_1. En considérant l'ensemble d'incertitude $\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|\cdot\|}(\cdot)$ associé à une norme arbitraire $\|\cdot\|$ et le rayon maximal $\sigma_{\max} \coloneqq \max_{p_1,q_1\in\Delta(\mathcal{S})} \|p_1 - p_2\|$ nous définissons le tuple $(S, A, \gamma, \sigma, \varepsilon)$, avec $\gamma \in \left[\frac{1}{2}, 1\right)$, $\sigma \in (0, \sigma_{\max}(1-c_0)]$ et $0 < c_0 \leq \frac{1}{8}$ une constante positive et $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. Ici*

Nous pouvons construire deux RMPDs $\mathcal{M}_0, \mathcal{M}_1$ tel que étant donné un jeu de données avec N échantillons indépendants pour chaque couple état-action échantillonné du MDP nominal (pour chaque $\mathcal{M}_0$ ou $\mathcal{M}_1$ respectivement), on a

$$\inf_{\widehat{\pi}} \max_{\mathcal{M}\in\{\mathcal{M}_0,\mathcal{M}_1\}} \left\{ \mathbb{P}_{\mathcal{M}}\Big(\max_{s\in\mathcal{S}} \left[V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \right] > \varepsilon \Big) \right\} \geq \frac{1}{8},$$

tant que

$$NSA \leq \frac{C_2 SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\} \varepsilon^2}.$$

Ici C_2 est une constante universelle positive, l'infimum est prit sous tous les estimateurs $\widehat{\pi}$, et $\mathbb{P}_0$ (respectivement. $\mathbb{P}_1$) dénote la probabilité lorsque le RMDP est $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

3 Conclusion

Ce travail a fait progresser le domaine en affinant les limites de la complexité de l'échantillon pour apprendre des processus décisionnels de Markov robustes lorsque l'ensemble d'incertitude est caractérisé par une norme arbitaire en supposant la présence d'un modèle génératif. Nos résultats renforcent non seulement le corpus de connaissances existant en améliorant les limites supérieures et inférieures, mais soulignent également que l'apprentissage des MDP s-rectangulaires est moins difficile en termes de complexité d'échantillonnage que les MDP classiques sa-rectangulaires. Ce travail représente un effort considérable pour fournir des résultats avec une borne minimax, car les résultats précédents concernant les cas s-rectangulaires n'étaient pas minimax optimaux. En outre, nous avons établi la complexité d'échantillonnage minimax pour les RMDP en utilisant une norme arbitraire, en démontrant qu'elle n'est jamais plus grande que celle requise pour l'apprentissage des MDP standard. Notre recherche fournit des pistes potentielles pour des travaux futurs, comme l'exploration de la caractérisation de la complexité d'échantillonnage pour les RMDP dans une famille plus large d'ensembles d'incertitude, tels que ceux définis par la f-divergence. Il serait souhaitable de disposer d'une base théorique plus unifiée, puisque la distance entre les mesures de probabilité est plus naturelle à définir à l'aide de divergences. De plus, il serait intéressant de se concentrer également sur la question de l'horizon fini et le cadre linéaire. Une telle extension contribuerait à une compréhension plus complète des cas tabulaires.

References

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Clavier, P., Pennec, E. L., and Geist, M. (2023). Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*.

Derman, E., Geist, M., and Mannor, S. (2021). Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34.

Ho, C. P., Petrik, M., and Wiesemann, W. (2021). Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.

Jiang, C. and Zhang, S.-B. (2018). A novel adaptively-robust strategy based on the mahalanobis distance for gps/ins integrated navigation systems. *Sensors*, 18(3):695.

Klopp, O., Lounici, K., and Tsybakov, A. B. (2017). Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. (2022). Certifying model accuracy under distribution shifts. *arXiv preprint arXiv:2201.12440*.

Li, G., Yan, Y., Chen, Y., and Fan, J. (2023). Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*.

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., and Bergstra, J. (2018). Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2004). Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, page 72.

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.

Yang, W., Zhang, L., and Zhang, Z. (2022a). Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248.

Yang, W., Zhang, L., and Zhang, Z. (2022b). Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR.