Literature review
Programación Dinámica Estocástica

Procesos de Decisión de Markov Robustos

Diego F. Fonseca. V. *

Resumen. En el presente trabajo pretendemos dar una visión de los enfoques denominados robustos por medio de los cuales se abordan los Procesos de Decisión de Markov MDP, esto surgen cuando se desconoce la matriz de transición del proceso, las recompensas o se desconocen ambas cosas. En ese sentido, presentaremos tres enfoques que son el s, a-rectangular, s-rectangular y el basado en métricas de Wasserstein, intentaremos presentar las condiciones y herramientas que permiten encontrar las políticas óptimas.

Palabras clave: s, a-rectangular, s-rectangular, Wasserstein.

1 Introducción

Los Procesos de Decisión de Markov MDP[1] son una herramienta importante y útil para abordar muchas aplicaciones en contextos como el mercado de opciones financieras, la optimización estocástica y problemas en los cueles se deben tomar decisiones en un entorno aleatorio, incluso, ideas tan populares en la actualidad como lo es el Aprendizaje Reforzado (Reinforcement Learning) es un caso particular de un MDP, esto evidencia la importancia de estudiar a profundidad los Procesos de Decisión de Markov. El estudio de los MDP podría decirse que tiene grandes avances para el caso en el cual se conocen todos los elementos que constituyen el MDP y siempre y cuando la función objetivo posea determinada forma, sin embargo, si se desconocen algunos elementos, especialmente los que imprimen la aleatoriedad al proceso, entonces no es tan claro el panorama, precisamente, abordar el MDP teniendo en cuenta dicho desconocimiento sin asumir condiciones sobre la ley que rige la aleatoriedad del proceso es a lo que llamaremos un **enfoque robusto** del MDP. Al respecto, para determinadas funciones objetivo, en [3], [6] y [7] cada uno propuso un enfoque robusto para el cual idearon herramientas que permiten obtener políticas óptimas, entendiendo una política óptima como una sucesión de decisiones dentro del MDP que permiten obtener la ganancia máxima donde la ganancia esta dada en términos de la función objetivo, no obstante, el calculo de dichas política esta ligado a la forma de un conjunto que se conoce como *conjunto de ambigüedad*, esto se debe a que en los tres enfoques que presentaremos el problema de encontrar un política óptima involucra un problema de óptimización que tiene como región factible el conjunto de ambigüedad, de modo que la complejidad de dicho problema esta ligada a la forma del conjunto de ambigüedad. Por lo tanto, muestro propósito es ilustrar estos enfoques robustos presentando las ideas principales de una manera concisa que permita eventualmente aplicar estos enfoques a diferentes situaciones.

Este trabajo se compone de las siguientes partes. En la Sección 2 presentamos todos los elementos que constituyen un Proceso de Decisión de Markov, además, también se exponen seis de las funciones objetivo que en dichos procesos se suelen maximizar, aunque los enfoques robustos solo se concentran en dos de estas funciones objetivo. Por otro lado, en la Sección 3 abordamos cada uno de los enfoques robustos destinado una subsección para cada uno, específicamente abordaremos el enfoques s, a-rectangular que fue propuesto en 2005 en [3], el enfoque s-rectangular que se propuso en 2013 en [6]

* Universidad de los Andes,Bogotá, Colombia.
[1] Abreviamos como MDP por sus siglas en ingles.

y el enfoque por medio de métricas de Wasserstein que fue propuesto en 2017 en [7], para cada uno de estos enfoque trataremos de exponer las condiciones y herramientas que permiten encontrar una política robusta en el sentido robusto de cada caso.

2 Procesos de Decisión de Markov MDP

Un proceso de decisión de Markov (MDP) es una 5-tupla dada por $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathbb{Q}, r)$ donde $\mathcal{T}$ es el horizonte de tiempo, este puede ser discreto, finito ó infinito, $\mathcal{S}$ es el espacio de estados, $\mathcal{A}$ es el conjunto de acciones, ambos son considerados finitos, y $\mathcal{A}_s \subseteq \mathcal{A}$ es el conjunto de acciones admisibles estando en el estado s. $\mathbb{Q}(s'|s,a)$ para $s' \in \mathcal{S}$ y $(s,a) \in \mathcal{S} \times \mathcal{A}_s$ representa la probabilidad de pasar al estado s' dado que estando en el estado s se tomo la acción a, $\mathbb{Q}$ es llamado *kernel de transición*. El ultimo elemento de la tupla es r el cual se puede ver como $r = \{r_t\}_{t\in\mathcal{T}}$ donde $r_t(a,s)$ representa la recompensa de tomar la acción a estando en el estado s en el tiempo t. En ese sentido, asumimos que $\mathcal{T} = \{0, 1, \ldots, T-1, T\}$ donde T puede ser $T < \infty$ o $T = \infty$. Denotamos por H_t la historia del proceso hasta el tiempo t, es decir, $H_t = \{h_t = (s_0, a_0, s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t) \mid (s_i, a_i) \in \mathcal{S} \times \mathcal{A}_{s_i}\}$, también consideramos $H = \bigcup_{t\in\mathcal{T}} H_t$. Una política π es una aplicación que toma una historia $h_t = (s_0, a_0, s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ y lo envía a una medida de probabilidad sobre $\mathcal{A}_{s_t}$, es decir, $\pi(h_t) \in \mathcal{P}(\mathcal{A}_{s_t})$. Si π solo depende de s_t entonces π es llamada *política Markoviana*, además, si π no depende de t entonces π es llamada *política estacionaria*. Denotamos como Π al conjunto de todas las políticas.

En este proceso s_t y a_t pueden verse como realizaciones de las variables aleatorias X_t y Y_t, en ese sentido, la transición del proceso es dada por

$$X_{t+1} = f(X_t, Y_t, \xi_t)$$

Donde ξ_t es una variable aleatoria que por lo general tiene influencia en $\mathbb{Q}$, es decir, $\mathbb{Q}$ se puede ver como

$$\mathbb{Q}(s'|s,a) = \mathbb{P}\left[f(s,a,\xi_t) = s'\right].$$

El objetivo es encontrar una política π que maximice el valor esperado de una función J que esta en términos de X, Y y la recompensa, en ese sentido, tenemos que Y_t depende de π de modo que X_{t+1} depende de $\mathbb{Q}$ y π. Dependiendo de las necesidades de cada problema se tienen diferentes funciones objetivo J, en realidad esta ultima se puede ver como una variable aleatoria, si estamos trabajando en un espacio medible $(\Omega, \mathcal{F})$ tenemos que $J : (\Omega, \mathcal{F}) \to \mathbb{R}$, entonces, llamando $R(s,\pi,\mathbb{Q}) := \mathbb{E}_{\pi,\mathbb{Q}}[J|X_0 = s]$ se tiene que el objetivo es encontrar π de tal manera que alcance el máximo

$$\max_{\pi} R(s,\pi,\mathbb{Q}). \tag{1}$$

Dependiendo del valor de T tenemos las siguientes opciones de J y R:

$\boxed{T < \infty :}$ En este caso $J := \sum_{t=0}^{T-1} r_t(X_t, Y_t) + r_T(X_T)$, entonces

$$R(s,\pi,\mathbb{Q}) := \mathbb{E}_{\pi,\mathbb{Q}}\left[\sum_{t=0}^{T-1} r_t(X_t, Y_t) + r_T(X_T) \,\middle|\, X_0 = s\right]. \tag{2}$$

$\boxed{T = \infty :}$ Para este caso existen varias funciones objetivo, nuestro interés se enfoca en las siguientes funciones:

Total expected cost: En este caso tenemos $J := \lim_{N\to\infty} \sum_{t=0}^{N-1} r_t(X_t, Y_t)$. Entonces

$$R(s,\pi,\mathbb{Q}) := \mathbb{E}_{\pi,\mathbb{Q}}\left[\lim_{N\to\infty}\sum_{t=0}^{N-1} r_t(X_t,Y_t)\middle| X_0 = s\right].$$

Bajo ciertas condiciones se puede demostrar que

$$R(s,\pi,\mathbb{Q}) := \lim_{N\to\infty}\mathbb{E}_{\pi,\mathbb{Q}}\left[\sum_{t=0}^{N-1} r_t(X_t,Y_t)\middle| X_0 = s\right]. \tag{3}$$

Average expected cost: En este caso tenemos $J := \lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} r_t(X_t, Y_t)$. Entonces

$$R(s,\pi,\mathbb{Q}) := \mathbb{E}_{\pi,\mathbb{Q}}\left[\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1} r_t(X_t,Y_t)\middle| X_0 = s\right].$$

Bajo ciertas condiciones se puede demostrar que

$$R(s,\pi,\mathbb{Q}) := \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\pi,\mathbb{Q}}\left[\sum_{t=0}^{N-1} r_t(X_t,Y_t)\middle| X_0 = s\right]. \tag{4}$$

Discount expected cost: En este caso tenemos $J := \lim_{N\to\infty} \sum_{t=0}^{N-1} \alpha^t r_t(X_t, Y_t)$ donde $\alpha \in (0,1)$, entonces

$$\begin{aligned} R(s,\pi,\mathbb{Q}) &:= \mathbb{E}_{\pi,\mathbb{Q}}\left[\lim_{N\to\infty}\sum_{t=0}^{N-1}\alpha^t r_t(X_t,Y_t)\middle| X_0 = s\right] \\ &= \lim_{N\to\infty}\mathbb{E}_{\pi,\mathbb{Q}}\left[\sum_{t=0}^{N-1}\alpha^t r_t(X_t,Y_t)\middle| X_0 = s\right]. \leftarrow \text{bajo ciertas condiciones.} \end{aligned} \tag{5}$$

Risk-sensitive expected cost: En este caso no se define J, considerando $\lambda \neq 0$ se define R directamente como

$$R(s,\pi,\mathbb{Q}) := \lim_{N\to\infty}\frac{1}{N\lambda}\log\left(\mathbb{E}_{\pi,\mathbb{Q}}\left[\exp\left(\lambda\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right)\middle| X_0 = s\right]\right). \tag{6}$$

Si se conoce $\mathbb{Q}$ y r entonces para el caso (2), (4) y (5) el problema (1) se puede formular como un linear programming, otra forma es emplear algoritmos que aproximan la solución eficientemente conocidos como *value iteration* and *policy iteration*, estos algoritmos basicamente son formas de solucionar ecuaciones de óptimalidad que surgen de cada problema de optimización. El problema (1) para R en (6) no es tan fácil, este es abordado con más detalle en [1] y [5], aunque dicho estudio se hace para casos específicos de MDP, no son resultados que abarca un universo amplios de MDPs como si ocurre para R de la forma (2), (4) y (5).

3 Procesos de Decisión de Markov Robustos RMDP

En realidad el termino "robusto" no hace referencia a el Proceso de Decisión de Markov MDP sino a la forma de abordar el problema de encontrar la política que maximice la recompensa esperada, esta forma de abordar este problema surge cuando se desconocen algunas de las componentes que definen el MDP, específicamente, cuando no se conoce $\mathbb{Q}$ ó r_t ó no se conocen ambas, el caso mas

común es asumir que no se conoce $\mathbb{Q}$ y que r_t si se conoce, asumiendo esto último, definimos $\mathbb{Q}_{s,a} := \{\mathbb{Q}(s'|s,a)\}_{s'\in\mathcal{S}}$ y[2] $\mathbb{Q}_s := \{\mathbb{Q}(s'|s,a)\}_{(s',s)\in\mathcal{S}\times\mathcal{A}_s}$, note que $\mathbb{Q}_{s,a}$ se puede ver como un elemento de $\mathcal{P}(\mathcal{S})$, y $\mathbb{Q}_s$ como un elemento de $\mathcal{P}(\mathcal{S})^{|\mathcal{S}||\mathcal{A}|}$. Además, considerando $\mathcal{S} = \left\{s^{(0)}, s^{(1)}, \ldots, s^{(l)}\right\}$ y $\mathcal{A} = \left\{a^{(0)}, a^{(1)}, \ldots, a^{(m)}\right\}$, entonces $\mathbb{Q}$ se puede expresar como

$$\begin{aligned}\mathbb{Q} &= \left(\mathbb{Q}_{s^{(0)},a^{(0)}}, \ldots, \mathbb{Q}_{s^{(l)},a^{(0)}}, \mathbb{Q}_{s^{(0)},a^{(1)}}, \ldots, \mathbb{Q}_{s^{(l)},a^{(1)}}, \ \cdots \ , \mathbb{Q}_{s^{(0)},a^{(m)}}, \ldots, \mathbb{Q}_{s^{(l)},a^{(m)}}\right) \\ &= \left(\mathbb{Q}_{s^{(0)}}, \mathbb{Q}_{s^{(1)}}, \ldots, \mathbb{Q}_{s^{(l)}}\right)\end{aligned}$$

Donde esto último se puede entender como una concatenación de vectores.

El enfoque robusto de un MDP consiste en considerar un conjunto $\mathcal{D}$ llamado *conjunto de ambigüedad* el cual satisface que $\mathbb{Q} \in \mathcal{D}$ con alta probabilidad, entonces, en lugar de abordar (1) se aborda el problema

$$\max_{\pi} \min_{\overline{\mathbb{Q}}\in\mathcal{D}} R(s,\pi,\overline{\mathbb{Q}}). \tag{7}$$

La dificultad de este enfoque radica en obtener ecuaciones de optimalidad y que estas puedan ser solucionadas de manera eficiente, esto dependerá de la forma que tenga el conjunto $\mathcal{D}$, en ese sentido, a continuación presentaremos las formas de definir $\mathcal{D}$ mas importantes.

3.1 $\mathcal{D}$ usando conjuntos s,a-rectangular

Este enfoque fue inicialmente propuesto en [2] y [3] simultáneamente, este consiste en asumir que existe un conjunto $\mathcal{D}_{s,a} \subset \mathcal{P}(\mathcal{S})$ para el cual se puede garantizar que $\mathbb{Q}_{s,a} \in \mathcal{D}_{s,a}$ con alta probabilidad, entonces se define el *conjunto de ambigüedad* como

$$\mathcal{D} := \bigtimes_{(s,a)\in\mathcal{S}\times\mathcal{A}_s} \mathcal{D}_{s,a}.$$

Lo que sigue es presentar las ecuaciones de óptimalidad, primero abordaremos el caso de horizonte finito, es decir, $T < \infty$ y consideramos la función de recompensa (2). Note que en nuestro MDP $\mathbb{Q}$ es estacionaria, es decir, no depende del t, sin embargo, según [3] este hecho no facilita el análisis (7), en ese sentido, es pertinente no limitarnos a las probabilidades de transición estacionaria, de modo que consideramos un nuevo conjunto de ambigüedad

$$\mathscr{D} := \mathcal{D}^{T+1}$$

Ya que sabemos que las probabilidades de transición son estacionarias, entonces el conjunto donde $\mathbb{Q}$ puede estar con mayor probabilidad es en el conjunto de portabilidades de transición estacionarias en $\mathscr{D}$ que es dado por

$$\mathscr{D}^{st} := \left\{\{\mathbb{Q}_t\}_{t\in\mathcal{T}} \in \mathscr{D} \mid \mathbb{Q}_t = \mathbb{Q}_j \ \forall\, t \neq j,\ t,j \in \mathcal{T}\right\}.$$

Note que $\mathscr{D}^{st}$ esta en biyección con $\mathcal{D}$, en ese sentido, denotando por

$$\phi_T(s,\Pi,\mathscr{P}) := \max_{\pi\in\Pi} \min_{\overline{\mathbb{Q}}\in\mathscr{P}} R(s,\pi,\overline{\mathbb{Q}}) \tag{8}$$

para cualquier conjunto de ambigüedad $\mathscr{P}$, tenemos que

$$\phi_T(s,\Pi,\mathscr{D}) \leq \phi_T(s,\Pi,\mathscr{D}^{st}) = \max_{\pi} \min_{\overline{\mathbb{Q}}\in\mathcal{D}} R(s,\pi,\overline{\mathbb{Q}}).$$

[2] Otra forma de notar $\mathbb{Q}_{s,a}$ es como una función, esto es, $\mathbb{Q}_{s,a} := \mathbb{Q}(\cdot|s,a)$.

El Teorema 4 en [3] demuestra que la brecha entre $\phi_T(s, \Pi, \mathscr{D})$ y $\phi_T(s, \Pi, \mathscr{D}^{st})$ tiende a cero a medida que T tiende a infinito, esto establece una limitante en el caso de horizonte finito pero es una ventaja en el caso infinito con función de recompensa dada por (5). Por lo tanto, la atención se concentra en $\phi_T(s, \Pi, \mathscr{D})$, para ilustrar la forma como se soluciona este problema expondremos las ideas que se sigue en [3], una de los hechos que se demuestran en [3] es que

$$\phi_T(s, \Pi, \mathscr{D}) = \min_{\overline{\mathbb{Q}} \in \mathscr{D}} \max_{\pi \in \Pi} R(s, \pi, \overline{\mathbb{Q}}) =: \psi_T(s, \Pi, \mathscr{D})$$

Es decir, se puede intercambiar min y max, esto es consecuencia de como se definió $\mathscr{D}$ a partir $\mathcal{D}$ que es s-rectangular, esto centra la atención en $\psi_N(s, \Pi, \mathscr{D})$. Luego, de acuerdo a los expuesto en [4] se sabe que dada $\overline{\mathbb{Q}} \in \mathscr{D}$ el problema $\max_{\pi \in \Pi} R(s, \pi, \overline{\mathbb{Q}})$ se puede reformular como un problema lineal, entonces $\psi_N(s, \Pi, \mathscr{D})$ se puede escribir como el valor óptimo de

$$\psi_N(s, \Pi, \mathscr{D}) = \begin{cases} \min\limits_{\overline{\mathbb{Q}} \in \mathscr{D},\, v_0, v_1, \ldots, v_{T-1} \in \mathbb{R}^{|\mathcal{S}|}} & \sum\limits_{s' \in \mathcal{S}} c(s') v_0(s') \\ \text{sujeto a} & v_t(\overline{s}) - \sum\limits_{s' \in \mathcal{S}} \overline{\mathbb{Q}}_t(s' | \overline{s}, a) v_{t+1}(s') \geq r_t(\overline{s}, a), \quad \forall\, a \in \mathcal{A},\, \overline{s} \in \mathcal{S},\, t \in \mathcal{T} \end{cases}$$

donde c es tomado de tal manera que $c(s') = 1$ si $s' = s$ y 0 en otro caso. Por lo tanto, todo se limita a solucionar el problema de optimización anterior, de acuerdo a [3] este se soluciona recursivamente mediante la relación

$$v_t(\overline{s}) = \max_{a \in \mathcal{A}} \left\{ r_t(\overline{s}, a) + \inf_{\overline{Q}_{\overline{s},a} \in \mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v_{t+1} \right\} \right\}$$

donde $\overline{Q}_{\overline{s},a} v_{t+1}$ es el producto pnto de dos vectores. La política óptima $\pi^* = (d_0^*, \ldots, d_{T-1}^*)$ también se obtiene recursivamente mediante

$$d_t^*(\overline{s}) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r_t(\overline{s}, a) + \inf_{\overline{Q}_{\overline{s},a} \in \mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v_{t+1} \right\} \right\}.$$

La dificultad de este enfoque robusto radica en calcular $\inf_{\overline{Q}_{\overline{s},a} \in \mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v_{t+1} \right\}$, el éxito en esta tarea dependerá de la forma de $\mathcal{D}_{s,a}$. En resumen, no se soluciona (7) directamente, lo que se hace es aproximarlo mediante $\phi_T(s, \Pi, \mathscr{D})$ que a su vez es igual a $\psi_N(s, \Pi, \mathscr{D})$, esta sera una buena aproximación siempre y cuando el espació de estados $\mathcal{S}$ sea grande.

En [3] también se aborda el caso de horizonte infinito en el caso descontado con recompensa estacionaria, es decir, $T = \infty$ con R dada por (5) con factor de descuento $\alpha \in (0, 1)$ y $r_t = r$, para este caso, en lugar de notar ϕ_T notamos ϕ_∞, con esta convención, se demuestra $\phi_\infty(s, \Pi, \mathscr{D}) = \phi_\infty(s, \Pi, \mathscr{D}^{st})$, es decir, no existe una brecha como si ocurría en el caso finito, además, también se tiene que $\phi_\infty(s, \Pi, \mathscr{D})$ es igual a $v(s)$ donde v es solución de la ecuaciones

$$v(\overline{s}) = \max_{a \in \mathcal{A}} \left\{ r(\overline{s}, a) + \alpha \inf_{\overline{Q}_{\overline{s},a} \in \mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v \right\} \right\} \quad \forall\, \overline{s} \in \mathcal{S}. \tag{9}$$

donde $\overline{Q}_{\overline{s},a} v$ es el producto punto de dos vectores, estas ecuaciones son conocidas como *ecuaciones de óptimalidad*. Para solucionar (9) se recurre a un método iterativo, este parte de hecho que si v soluciona (9) entonces es el limite de la sucesión dada por

$$v_{k+1}(\overline{s}) = \max_{a \in \mathcal{A}} \left\{ r(\overline{s}, a) + \alpha \inf_{\overline{Q}_{\overline{s},a} \in \mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v_k \right\} \right\} \quad \begin{array}{l} \forall\, \overline{s} \in \mathcal{S} \\ k = 1, 2, \ldots \end{array}$$

Una vez solucionada la ecuación (9) entonces la política óptima $\pi^* = (d^*, d^*, \ldots)$ es estacionaria y es dada por

$$d^*(\overline{s}) = \underset{a\in\mathcal{A}}{\operatorname{argmax}} \left\{ r(\overline{s}, a) + \alpha \inf_{\overline{Q}_{\overline{s},a}\in\mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v \right\} \right\} \quad \forall\, \overline{s} \in \mathcal{S}.$$

Las ideas de la demostración de todo lo inmediatamente anterior son similares a las empleadas para el caso de horizonte-finito. De nuevo, la dificultad de este enfoque robusto radica en calcular $\inf_{\overline{Q}_{\overline{s},a}\in\mathcal{D}_{\overline{s},a}} \left\{ \overline{Q}_{\overline{s},a} v \right\}$ y esto dependerá de la forma de $\mathcal{D}_{\overline{s},a}$.

3.2 $\mathcal{D}$ usando conjuntos s-rectangular

Este enfoque es inicialmente propuesto en [6], este asume que existe un conjunto $\mathcal{D}_s \subset \mathcal{P}(\mathcal{S})^{|\mathcal{S}||\mathcal{A}_s|}$ para el cual se puede garantizar que $\mathbb{Q}_s \in \mathcal{D}_s \subseteq \mathcal{P}(\mathcal{S})^{|\mathcal{S}||\mathcal{A}|}$, entonces para este caso se define el *conjunto de ambigüedad* como

$$\mathcal{D} := \underset{s\in\mathcal{S}}{\times} \mathcal{D}_s.$$

A diferencia del enfoque s-rectangular, en este caso no es fácil obtener ecuaciones de óptimalidad para el problema (7), es más, en [6], que es el primer trabajo que presenta este enfoque, se trabaja con una forma especifica de $\mathcal{D}_s$ para así poder obtener un método que permita solucionar (7). En esta subsección vamos a abordar estas ideas. En primer lugar, este enfoque es ideado para tratar MDPs de horizonte infinito descontado con recompensa estacionaria, es decir, $T = \infty$ con R dada por (5) con factor de descuento $\alpha \in (0, 1)$ y $r_t = r$. Por lo tanto, se considera

$$\mathcal{D} := \left\{ \overline{\mathbb{Q}} \in \mathcal{P}(\mathcal{S})^{|\mathcal{S}||\mathcal{A}|} \;\middle|\; \exists \xi \in \Xi \text{ tal que } \overline{\mathbb{Q}}_{s,a} = \mathbb{Q}^{\xi}(\cdot|s, a) := \kappa_{s,a} + K_{s,a}\xi,\ \forall\, (s, a) \in \mathcal{S} \times \mathcal{A} \right\}$$

donde $\Xi \subseteq \mathbb{R}^q$ para algún q, $\kappa_{s,a} \in \mathbb{R}^{|\mathcal{S}|}$ y $K_{s,a} \in \mathbb{R}^{|\mathcal{S}|\times q}$. Considerando $\mathcal{A} = \left\{a^{(0)}, \ldots, a^{(m)}\right\}$, se dice que $\mathcal{D}$ es **s-rectangular** si este se puede expresar de la forma

$$\mathcal{D} = \underset{s\in\mathcal{S}}{\times} \mathcal{D}_s \quad \text{donde} \quad \mathcal{D}_s = \left\{ (\overline{\mathbb{Q}}_{s,a^{(0)}}, \ldots, \overline{\mathbb{Q}}_{s,a^{(m)}}) \mid \overline{\mathbb{Q}} \in \mathcal{D} \right\}.$$

Note que no todo $\mathcal{D}$ es s-recatangular, poseer esta propiedad depende de la forma como se considere Ξ, y $K_{s,a}$. En adelante, asumimos que $\mathcal{D}$ es s-rectangular, también adoptamos la misma notación usada en el caso s, a-rectangular, es decir, $\phi_\infty(s, \Pi, \mathcal{D})$ tiene el mismo significado que se expuso en (8). Dada la dependencia de $\mathcal{D}$ respecto a Ξ, entonces podemos escribir

$$\phi_\infty(s, \Pi, \mathcal{D}) = \max_{\pi} \min_{\xi\in\Xi} R(s, \pi, \mathbb{Q}^{\xi}) =: \phi_\infty(s, \Pi, \Xi).$$

Dada una política π y $\xi \in \Xi$ y teniendo en cuenta que R es dada por (5), entonces del Teorema 2.2 de [4] se sigue que para todo $\overline{s} \in \mathcal{S}$

$$w(\overline{s}) \leq \sum_{a\in\mathcal{A}} \pi(a|\overline{s}) \sum_{s'\in\mathcal{S}} \mathbb{Q}^{\xi}(s'|\overline{s}, a) r(\overline{s}, a) + \alpha \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} \pi(a|\overline{s}) \mathbb{Q}^{\xi}(s'|\overline{s}, a) w(s') \;\Rightarrow\; w(\overline{s}) \leq R(\overline{s}, \pi, \mathbb{Q}^{\xi})$$

es más,

$$w(\overline{s}) = \sum_{a\in\mathcal{A}} \pi(a|\overline{s}) \sum_{s'\in\mathcal{S}} \mathbb{Q}^{\xi}(s'|\overline{s}, a) r(\overline{s}, a) + \alpha \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} \pi(a|\overline{s}) \mathbb{Q}^{\xi}(s'|\overline{s}, a) w(s') \;\Leftrightarrow\; w(\overline{s}) = R(\overline{s}, \pi, \mathbb{Q}^{\xi}).$$

Por lo tanto, considerando $c \in \mathbb{R}^{|\mathcal{S}|}$ tal que $c(\overline{s}) = 1$ si $\overline{s} = s$ y 0 en otro caso, se obtiene que

$$R(s, \pi, \mathbb{Q}^{\xi}) = \left\{ \begin{array}{ll} \sup\limits_{w\in\mathbb{R}^{|\mathcal{S}|}} & c \cdot w \\ \text{sujeto a} & \\ & w(\overline{s}) \leq \sum\limits_{a\in\mathcal{A}} \pi(a|\overline{s}) \sum\limits_{s'\in\mathcal{S}} \mathbb{Q}^{\xi}(s'|\overline{s}, a) r(\overline{s}, a) + \alpha \sum\limits_{s'\in\mathcal{S}} \sum\limits_{a\in\mathcal{A}} \pi(a|\overline{s}) \mathbb{Q}^{\xi}(s'|\overline{s}, a) w(s') \\ & \forall\, \overline{s} \in \mathcal{S} \end{array} \right.$$

Por consiguiente, se obtiene que

$$\inf_{\xi\in\Xi} R(s,\pi,\mathbb{Q}^\xi) = \inf_{\xi\in\Xi}\begin{cases}\sup\limits_{w\in\mathbb{R}^{|\mathcal{I}|}} & c\cdot w\\ \text{sujeto a} & \\ & w(\overline{s}) \le \sum\limits_{a\in\mathcal{A}}\pi(a|\overline{s})\sum\limits_{s'\in\mathcal{S}}\mathbb{Q}^\xi(s'|\overline{s},a)r(\overline{s},a) + \alpha\sum\limits_{s'\in\mathcal{S}}\sum\limits_{a\in\mathcal{A}}\pi(a|\overline{s})\mathbb{Q}^\xi(s'|\overline{s},a)w(s')\\ & \forall\,\overline{s}\in\mathcal{S}\end{cases}$$

$$= \begin{cases}\sup\limits_{v:\Xi\xrightarrow{\text{cont}}\mathbb{R}^{|\mathcal{S}|}} & \max\limits_{\xi\in\Xi} c\cdot v_\xi\\ \text{sujeto a} & \\ & v_\xi(\overline{s}) \le \sum\limits_{a\in\mathcal{A}}\pi(a|\overline{s})\sum\limits_{s'\in\mathcal{S}}\mathbb{Q}^\xi(s'|\overline{s},a)r(\overline{s},a) + \alpha\sum\limits_{s'\in\mathcal{S}}\sum\limits_{a\in\mathcal{A}}\pi(a|\overline{s})\mathbb{Q}^\xi(s'|\overline{s},a)v_\xi(s')\\ & \forall\,\overline{s}\in\mathcal{S}\end{cases}$$

donde $\xrightarrow{\text{cont}}$ indica que la función es *continua*. Los detalles que justifican la última igualdad son presentados en [6]. Optimizar sobre las funciones continuas es intratable, de modo que se aproxima el valor óptimo del problema en cuestión mediante otro problema que difiere del anterior en el hecho de que se cambia $\xrightarrow{\text{cont}}$ por $\xrightarrow{\text{aff}}$ que significa función afín. Así pues, el problema que nos interesa abordar se convierte en

$$\max_{\pi\in\Pi}\inf_{\xi\in\Xi} R(s,\pi,\mathbb{Q}^\xi) = \begin{cases}\sup\limits_{\pi\in\Pi}\sup\limits_{v:\Xi\xrightarrow{\text{cont}}\mathbb{R}^{|\mathcal{S}|}} & \max\limits_{\xi\in\Xi} c\cdot v_\xi\\ \text{sujeto a} & \\ & v_\xi(\overline{s}) \le \sum\limits_{a\in\mathcal{A}}\pi(a|\overline{s})\sum\limits_{s'\in\mathcal{S}}\mathbb{Q}^\xi(s'|\overline{s},a)r(\overline{s},a) + \alpha\sum\limits_{s'\in\mathcal{S}}\sum\limits_{a\in\mathcal{A}}\pi(a|\overline{s})\mathbb{Q}^\xi(s'|\overline{s},a)v_\xi(s')\\ & \forall\,\overline{s}\in\mathcal{S}\end{cases} \tag{10}$$

Este problema se soluciona de la siguiente manera, se considera la función $\Psi_s : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ dada por

$$\Psi_s(w) := \max_{\pi\in\Pi}\inf_{\xi\in\Xi}\left\{\sum_{a\in\mathcal{A}}\pi(a|s)\sum_{s'\in\mathcal{S}}\mathbb{Q}^\xi(s'|\overline{s},a)r(s,a) + \alpha\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\pi(a|s)\mathbb{Q}^\xi(s'|s,a)w(s')\right\}.$$

Esta función es contractiva, de modo que esta tiene un punto fijo w^*, entonces, la función v^* que es solución óptima de (10) es la función constante $v^*(\xi) = w^*$ y la política óptima π^* es dada por

$$\pi^* \in \operatorname*{argmax}_{\pi\in\Pi}\inf_{\xi\in\Xi}\left\{\sum_{a\in\mathcal{A}}\pi(a|s)\sum_{s'\in\mathcal{S}}\mathbb{Q}^\xi(s'|\overline{s},a)r(s,a) + \alpha\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\pi(a|s)\mathbb{Q}^\xi(s'|s,a)w^*(s')\right\}.$$

Note que del hecho que Ψ_s sea contractiva se puede inferir un especie de método de *value iteration*. De nuevo, como ocurrió en el caso s,a-rectangulat, en este enfoque existen varios factores que pueden hacer fácil o difícil abordar este MDP robusto, en este caso esos factores recaen en la forma de Ξ, $\kappa_{s,a}$ y $K_{s,a}$. Otro aspecto que genera inquietud es el hecho de que los resultados expuestos estén ligados a la forma como se definió $\mathcal{D}$.

3.3 $\mathcal{D}$ usando conjuntos definidos por medio de la Métrica de Wasserstein

Este enfoque tiene marcadas diferencias con los anteriores, este es presentado por primera vez en [7] y es formulado para el caso $T < \infty$, para este enfoque se consideran la matriz de transición y la recompensa no estacionarias y desconocidas, en ese sentido, notamos dichos elementos desconocidos $\mathbb{Q}^t$ y r_t. Note que a diferencia de los enfoques anteriores en este caso r_t también es desconocida. Para este enfoque se asume que el par $(\mathbb{Q}^t_s, r_{t,s})$ es un vector aleatorio, donde $\mathbb{Q}^t_s$ se interpreta como se

definió al inicio de la Sección 3 y $r_{t,s}$ se interpreta como $r_{t,s} = (r_t(a^{(0)}, s), r_t(a^{(1)}, s), \ldots, r_t(a^{(m)}, s))$, en ese sentido, tenemos que $(\mathbb{Q}_s^t, r_{t,s}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}_s|+|\mathcal{A}_s|}$. Asumir dicha aleatoriedad en parte generaliza la noción de MDP que definimos en la Sección 2, además, en la practica cuando se desconoce $\mathbb{Q}^t$ y r_t una acción sensata es asumir que son aleatorios si no se cuenta con evidencia para inferir que son deterministas. Se asume también que r_T y $\mathbb{Q}^T$ son conocidas, para evitar confusiones se nota $r_N = q$. En consecuencia, estas consideraciones obliga a modificar algunos de los elementos que definían el MDP, denotando por μ_t la distribución conjunta del par $(\mathbb{Q}_{s_t}^t, r_{t,s_t})$ donde s_t es el estado donde esta el proceso en el tiempo t, definimos la historia de este proceso hasta el tiempo t como

$$H_t = \left\{ h_t = (s_0, a_0, s_1, a_1, \mu_1, \ldots, s_{t-1}, a_{t-1}, \mu_{t-1}, s_t) \;\middle|\; \begin{array}{l} (s_i, a_i) \in \mathcal{S} \times \mathcal{A}_{s_i} \\ (\mathbb{Q}_{s_t}^t, r_{t,s_t}) \sim \mu_t \end{array} \right\}$$

también se define la *historia extendida* como

$$H_t^e = \left\{ h_t = (s_0, a_0, s_1, a_1, \mu_1, \ldots, s_{t-1}, a_{t-1}, \mu_{t-1}, s_t, a_t) \;\middle|\; \begin{array}{l} (s_i, a_i) \in \mathcal{S} \times \mathcal{A}_{s_i} \\ (\mathbb{Q}_{s_t}^t, r_{t,s_t}) \sim \mu_t \end{array} \right\}.$$

La idea es considerar un conjunto $\mathcal{D}_{s_t} \subseteq \mathcal{P}(\mathbb{R}^{|\mathcal{S}||\mathcal{A}_{s_t}|+|\mathcal{A}_{s_t}|})$ tal que $\mu_t \in \mathcal{D}_{s_t}$, el hecho de considerar $\mathcal{P}(\mathbb{R}^{|\mathcal{S}||\mathcal{A}_{s_t}|+|\mathcal{A}_{s_t}|})$ es la gran diferencia con el enfoque s-rectangular, además, estos dos enfoque están pensados para MDPs diferentes, empezando porque en s-rectangular la matriz de transición y la recompensa son estacionarias. En el contexto de [7] a $\mathcal{D}_{s_t}$ se le llama *conjunto de ambigüedad*, sin embargo, en el contexto en el que hemos venido trabajando el conjunto de ambigüedad es el que en [7] se codifica como

$$\Gamma := \{\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_T) \mid \gamma_t : H_t^e \to \mathcal{D}_{s_t},\ \forall\, t \in \mathcal{T} \setminus \{T\}\}$$

Esta ultima definición puede llegar a no ser ta clara, pero lo que debe importar es que dada una política π y $\gamma \in \Gamma$, el par (π, γ) inducen una medida de probabilidad que distribuye a todo lo que tiene aleatoriedad en la expresión $\sum_{t=0}^{T-1} r_t(X_t, Y_t) + q(X_T)$. Bajo estas consideraciones todo se resume a solucionar el problema de optimización

$$\phi_T(s, \Pi, \Gamma) := \max_{\pi \in \Pi} \inf_{\gamma \in \Gamma} \mathbb{E}_{\pi,\gamma} \left[\sum_{t=0}^{T-1} r_t(X_t, Y_t) + q(X_T) \middle| X_0 = s \right]. \tag{11}$$

De nuevo, como en todos los enfoques robustos, la dificultad de (11) radica en la forma de Γ que a su vez radica en la forma de $\mathcal{D}_s$, sin embargo, considerando

$$v_t(s) := \max_{\pi \in \Pi} \inf_{\gamma \in \Gamma} \mathbb{E}_{\pi,\gamma} \left[\sum_{\tau=t}^{T-1} r_\tau(X_\tau, Y_\tau) + q(X_T) \middle| X_t = s \right].$$

y siguiendo la estrategia similar a como en [4] se obtiene las ecuaciones de Bellman sumado a otros resultados de otros autores citados en [7] se obtiene un análogo de la *ecuaciones de Belleman*, para este caso estas son dadas por

$$\begin{aligned} v_t(s) &= \sup_{d \in \triangle(\mathcal{A}_s)} \inf_{\mu \in \mathcal{D}_s} \mathbb{E}_\mu \left[\sum_{a \in \mathcal{A}_s} d(s) \left(r_t(s,a) + \sum_{s' \in \mathcal{S}} \mathbb{Q}^t(s'|s,a) v_{t+1}(s') \right) \right] \\ &= \sup_{d \in \triangle(\mathcal{A}_s)} \inf_{\mu \in \mathcal{D}_s} \int_{\mathcal{X}_s} \sum_{a \in \mathcal{A}_s} d(s) \left(r_t(s,a) + \sum_{s' \in \mathcal{S}} \mathbb{Q}^t(s'|s,a) v_{t+1}(s') \right) d\mu(\mathbb{Q}_s^t, r_{t,s}) \end{aligned} \tag{12}$$

donde $\mathcal{X}_s$ es el soporte de μ, $v_T(s) = q(s)$, además, la política óptima π^* se escribe en términos de d, es decir, si d alcanza el supremo en (12) entonces $\pi^*(\cdot|s) = d$. El conjunto $\triangle(\mathcal{A}_s)$ es el conjunto

de medidas de probabilidad en $\mathcal{A}_s$ que se escribe con un $\triangle$ pues este se puede ver como un simplex. De acuerdo a estas ecuaciones de optimalidad se infiere que $\phi_T(s, \Pi, \Gamma) = v_0(s)$.

La complejidad de la ecuaciones (12) dependen de la forma de $\mathcal{D}_s$, en ese enfoque robusto se considera $\mathcal{D}_s$ como una bola respecto a una métrica de Wasserstein centrada en una distribución empírica y radio fijo que debe ser ajustado, en ese sentido se asume que tenemos acceso a realizaciones del vector aleatorio $(\mathbb{Q}_s^t, r_{t,s})$, es decir, existe una muestra $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N$ de $(\mathbb{Q}_s^t, r_{t,s})$, donde $\widehat{\xi}_i = (\mathbb{Q}_{s,i}^t, r_{t,s,i})$, entonces se define la distribución empírica

$$\widehat{\mu}_s := \frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$$

donde $\delta_{\widehat{\xi}_i}$ es el delta de Dirac concentrado en $\widehat{\xi}_i$. Entonces, dada cualquier $\mu \in \mathcal{P}(\mathbb{R}^{|\mathcal{S}||\mathcal{A}_{s_t}|+|\mathcal{A}_{s_t}|})$ se define la **distancia de Wasserstein** entre

$$W_p(\mu, \widehat{\mu}_s) := \inf_{\eta \in \mathcal{P}(\mathcal{X}_s \times \mathcal{X}_s)} \left\{ \left(\int_{\mathcal{X}_s \times \mathcal{X}_s} d(x,y)^p \mathbf{d}\eta(x,y) \right)^{\frac{1}{p}} \;\middle|\; \eta \text{ tiene marginales en } \mu \text{ y } \widehat{\mu}_s \right\}$$

donde $\mathbf{d}$ es una métrica en $\mathcal{X}_s \times \mathcal{X}_s$. Entonces, se considera $\mathcal{D}_s$ como

$$\mathcal{D}_s = \mathcal{B}_\varepsilon(\widehat{\mu}_s) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^{|\mathcal{S}||\mathcal{A}_{s_t}|+|\mathcal{A}_{s_t}|}) \;\middle|\; W_p(\mu, \widehat{\mu}_s) \leq \varepsilon^p \right\}.$$

Existen varios resultados que permiten escribir problemas de optimización estocástica robusta como problemas de optimización semi-infinita y en algunos caso de optimización convexa finito, en este caso, el problema de minimización que se encuentra al interior de (12) es de optimización estocástica robusta, entonces, (12) se puede reformular como

$$v_t(s) = \begin{cases} \sup\limits_{d \in \triangle(\mathcal{A}_s),\, \lambda \geq 0,\, s_i \in \mathbb{R}} & -\lambda \varepsilon^p + \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} k \\ \text{sujeto a} & \\ & \inf\limits_{\xi := (\mathbb{Q}_s^{(\xi)}, r_s^{(s)}) \in \mathcal{X}_s} \left\{ \lambda \mathbf{d}(\xi, \widehat{\xi}_i)^p + \displaystyle\sum_{a \in \mathcal{A}_s} d(s) \left(r^\xi(s,a) + \sum_{s' \in \mathcal{S}} \mathbb{Q}^\xi(s'|s,a) v_{t+1}(s') \right) \right\} \geq k_i, \\ & \forall\, i = 1, 2, \ldots, N \end{cases} \tag{13}$$

En conclusión, (12) se pude expresar como un problema de optimización finito (13) siempre y cuando $\mathcal{X}_s$ sea finito, poder solucionar este problema de optimización permite que de (12) se puedan plantear algoritmos análogos a *value iteration* ó *policy iteration* para este caso.

4 Conclusiones

Existen dificultades en cada uno de los enfoques robustos, en algunos casos el universo de MDPs para los cuales es procedente cada enfoque puede llegar a ser limitado, queda como tarea para un proyecto a futuro evaluar el desempeño de cada uno de estos perspectivas robustas ya que los autores no realizan dicha evaluación, podemos intuir que la razón de esta omisión es que para ejemplos interesantes con un numero considerable de estados y acciones la implementación no es fácil, esto quizá se debe a que en la mayoría de los casos se esta llegando a un método iterativo donde cada iteración implica solucionar un problema de optimización, esto puede llegar a ser costoso computacionalmente.

Referencias

1. C. R Cavazos and H. D. Hernández. Vanishing discount approximations in controlled Markov chains with risk-sensitive average criterion. *Advances in Applied Probability*, 50(1):204–230, 2017.
2. G. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30:257–280, 05 2005.
3. A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53:780–798, 10 2005.
4. M. L. Puterman. *Markov Decision Processes*. Wiley, 1994.
5. Q. Wei and X. Chen. Risk-sensitive average continuous-time Markov decision processes with unbounded rates. *Optimization*, 68(4):773–800, 2019.
6. W. Wiesemann, D. Kuhn, and B. Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.
7. I. Yang. A convex optimization approach to distributionally robust markov decision processes with wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.