UNIVERSIDAD POLITECNICA SALESIANA

Nombre:Luis Orellana

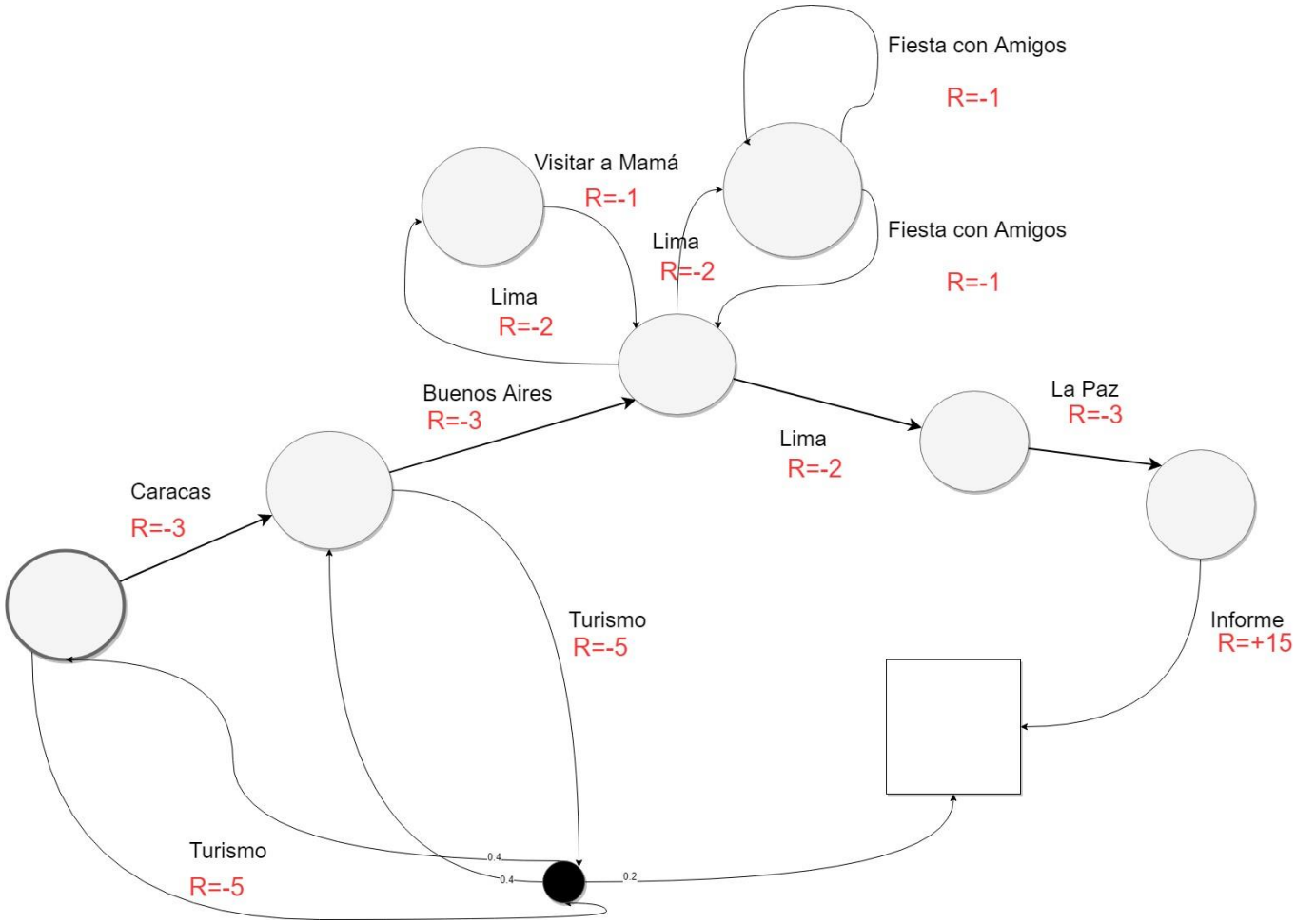
Procesos de Decisión de Markov

Los procesos de decisión de Markov poseen la misma forma que un proceso de recompensa de Markov pero en este caso se le agregan decisiones que van a ser tomadas por nuestros agente, es decir que ya no se moverá libremente en el sistema. Ahora para adentrarnos en la definición agregaremos variables a la tupla ya antes explicada en la primera parte de este post <S, A, P, R, γ>

- S es una lista de estados a los cuales puede pertenecer.
- A es una lista de acciones.
- P es una matriz de transición de estado.
- R es la función de recompensa.
- γ es el valor de descuento.

En este caso se agregará una lista de acciones que se tomará, debido a que no será un problema en el cual el agente se moverá aleatoriamente en la cadena de Markov si no más bien será una serie de decisiones que nos llevarán a hallar la mayor recompensa.

Ahora que ya hemos definido lo que es un proceso de decisión de Markov actualizaremos nuestro ejemplo en base a estados y decisiones.



En la parte inferior de la cadena de Markov se ve representado una decisión que puede transportarte aleatoriamente (según una probabilidad) a uno de los estados con los cuales está relacionado.

Políticas

Las políticas se pueden describir como una distribución π de acciones dado estados.

$$\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$$

La política depende del estado actual y no de los estados pasados. La política es independiente del tiempo, es decir son estacionarias.

Función de Valor

La función de valor se divide en dos conceptos separados, la función estado-valor, que representa que tan bueno es un estado (o que tanto valor tiene) y la función acción-valor que representa que tan positivo es tomar cierta acción y que tanto valor nos va a devolver.

La función de estado-valor básicamente representa el retorno esperado desde un estado s siguiendo la política π

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$

E representa la expectativa cuando se prueba todas las acciones dada la política π. La función de acciónvalor iniciando desde el estado s, como la anterior pero tomando una acción a y luego siguiendo la política π para ver cuanto retorno nos trae.

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s,\ A_t = a]$$

Ecuación de expectativa de Bellman

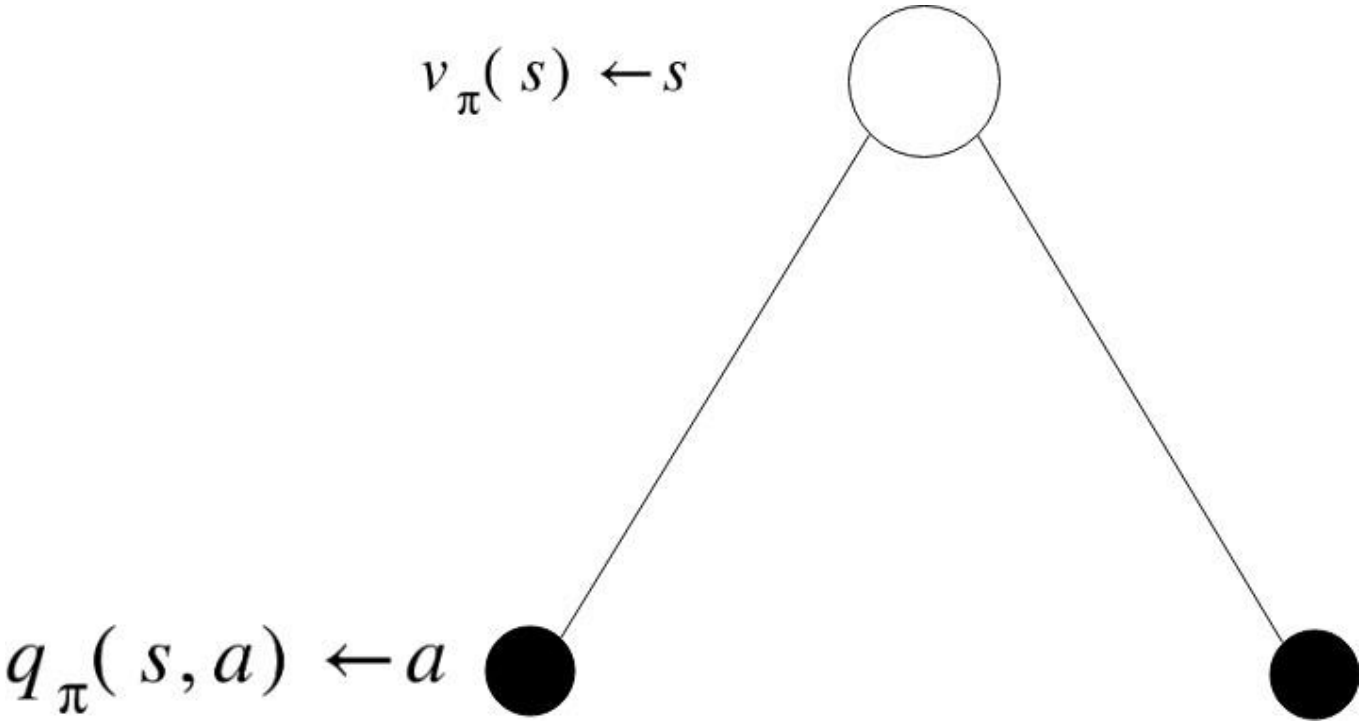
Esta vez volvemos a ver la ecuación de Bellman desde otro punto de vista: Para la función de estado-valor puede ser descompuesta en la recompensa inmediata mas el valor descontado del siguiente estado siguiendo la política π.

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s]$$

Para la función de acción-valor puede tambien ser descompuesta en la expectativa por la recompensa actual y el valor de la acción que se eligió desde ese punto para adelante.

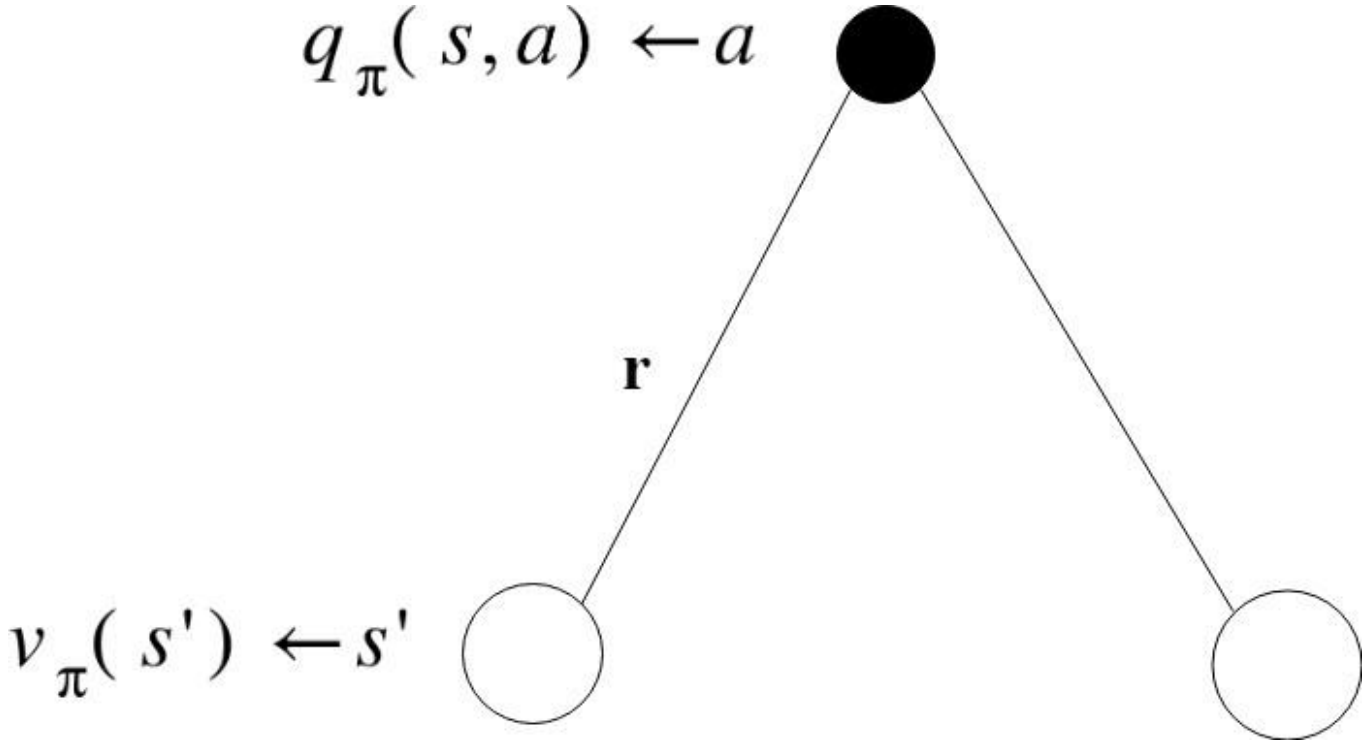
$$q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma q_\pi\left(S_{t+1}, A_{t+1}\right) \middle| S_t = s, A_t = a\right]$$

Ahora a partir de los siguientes gráficos veremos como v y q se relacionan unos a otros y entraremos en mayor profundidad en la generalización de la ecuación de expectativa de Bellman. En el siguiente gráfico muestra el estado “s” y las posibles acciones que puede tomar (siendo representado por los puntos negros). Si queremos saber cual es el valor de “s” necesitamos promediar las acciones que se puedan tomar desde ese estado, en otras palabras lo que se debe hacer para saber el valor del estado “s” es observar cual es el valor de las siguientes acciones que se pueden tomar dada la política π desde el estado “s” y promediarlas.



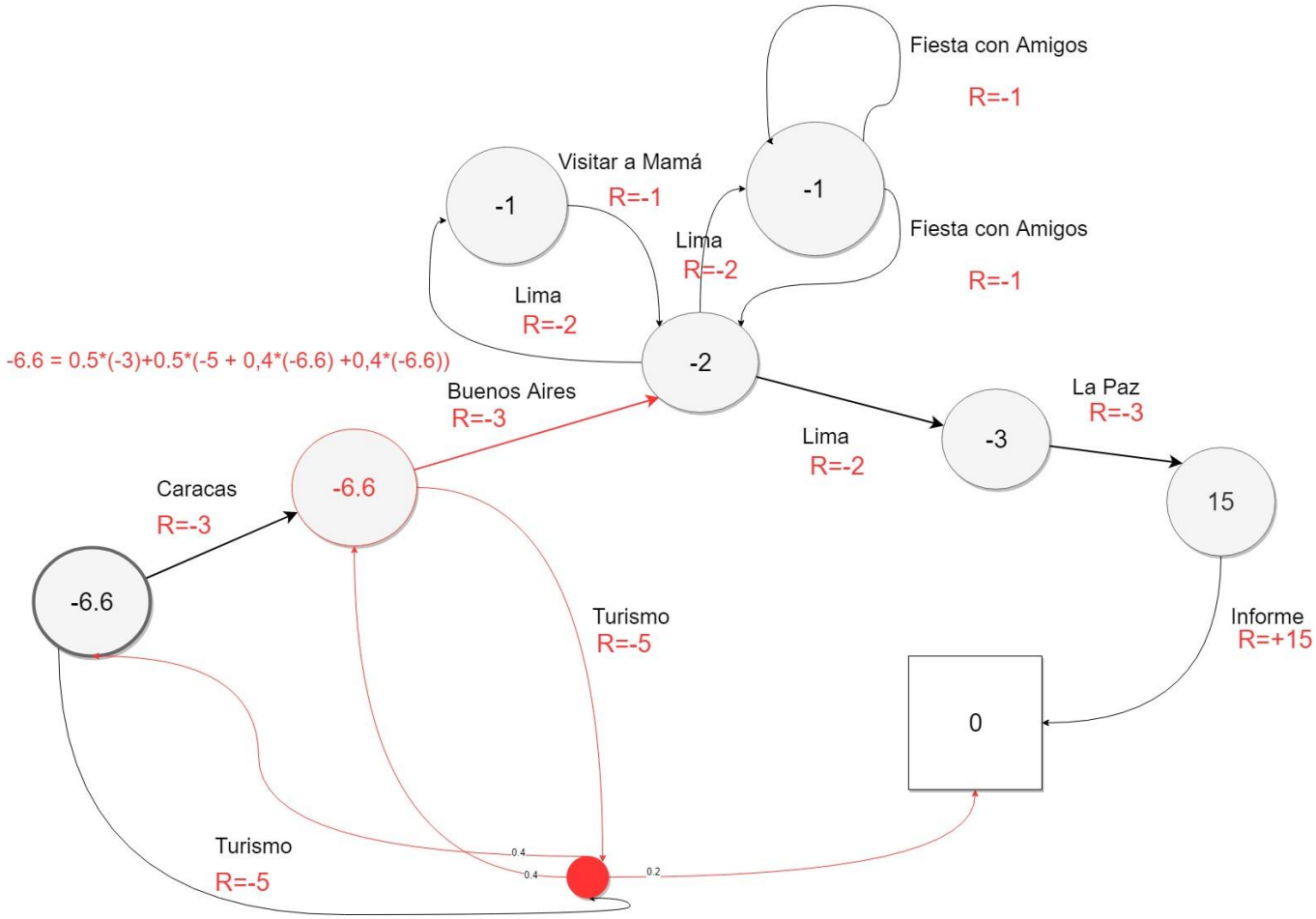
$$v_\pi(s) = \sum_{a \in A} \pi(a|s)\, q_\pi(s, a)$$

Ahora vamos a ver lo contrario, comenzaremos seleccionando una acción, dado el estado “s”, seleccionamos la acción “a” la cual dada la dinámica de nuestro Proceso de Decisión de Markov puede llevarlo hacia el estado de la derecha o el de la izquierda, entonces para hallar el valor de esta acción tendríamos que promediar el valor de los estados donde nos puede llevar esa acción junto a las probabilidades de la dinámica de nuestro Proceso de Decisión de Markov (la cual se denota en la formula por P), en otras palabras la probabilidad que nos pueda llevar a el estado de la derecha y al de la izquierda.



$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

Ambas propiedades se pueden mezclar de diferentes maneras para unificar ambas formulas en formas mas complejas. A continuación veremos como sería aplicada la Ecuación de Expectativa de Bellman en nuestro proceso de decisión de Markov de ejemplo.



En este proceso hallaremos la función de estado-valor según la política que estamos utilizando (escoger cada uno de los caminos con igual probabilidad) iniciaremos en el segundo estado con igual probabilidad de escoger la acción de “ir a buenos aires” que nos traería una recompensa de -3 o “hacer turismo” con recompensa de -5, el hecho de haber escogido la acción de hacer turismo nos puede transportar, por dinámicas del proceso de decisión de Markov, hacia el primer estado, al segundo o al último (con

probabilidades de 0.4, 0.4 y 0.2 respectivamente), entonces dado este caso al sumar todos los posibles resultados obtenemos cuanto sería el valor esperado de este estado.

Ahora, en este post hemos aprendido como hallar el valor para cada uno de los estados, utilizando la política de tener la misma probabilidad de caer en cualquier estado, pero todavía no tenemos claro como elegir el mejor camino que podría existir en nuestro proceso de decisión de Markov, es decir en cual obtendríamos la mayor cantidad de recompensa, aquí donde necesitamos la función de valor óptimo.

Función de Valor Óptimo

Como era en los casos anteriores, esta función de valor tambien se divide en valor estado y acción, cuyas definiciones pasaremos a revisar.

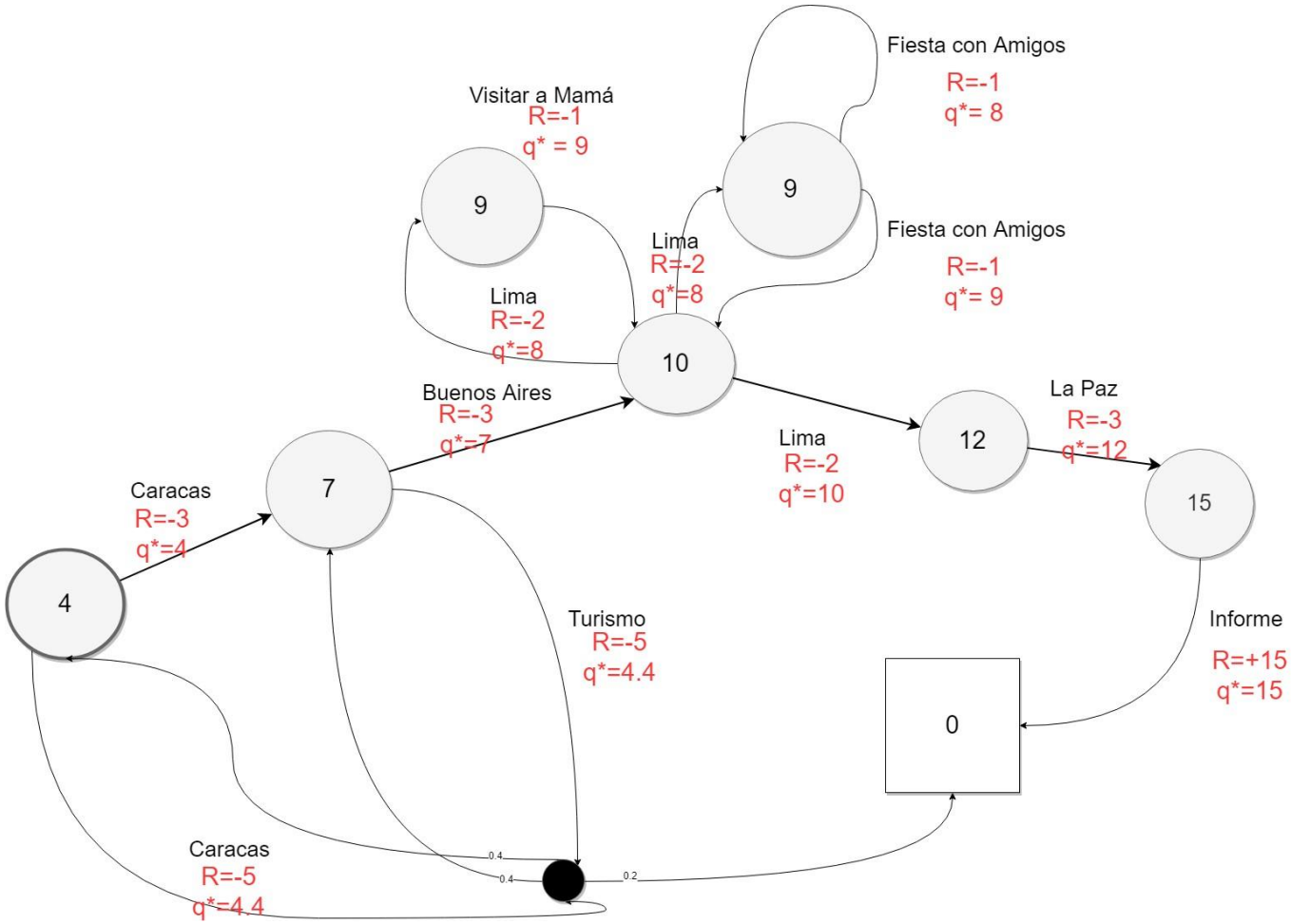
La función de estado-valor optimo es la que posee el mayor valor sobre todas las políticas, en otras palabras se puede representar como el camino óptimo que traería mayor valor dentro de todo el sistema, y todas las políticas posibles.

$$v_*(s) = \max_\pi v_\pi(s)$$

La función optima de acción-valor es el máximo valor que se puede alcanzar luego de haber tomado una acción “a”, sobre todas las políticas. Este valor terminaría siendo el más importante, debido a que si se mapean todos los valores para cada opción ya se tendría resuelto el proceso de decisión de Markov, debido a que ya sabrías que si tomas cierto camino cual es el valor de este, entonces teniendo los valores para cada acción seleccionarías los que te trajeran mayor valor, siempre.

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

En el siguiente ejemplo mostraremos como quedaría nuestro ejemplo si aplicáramos la función de valor óptimo.



Los valores se hallan restando la recompensa por cada acción por ejemplo el penúltimo estado tiene un valor de 15 por que el la recompensa que puede ganar en el futuro, así se van hallando los diferentes valores de los estados basados en las acciones y las recompensas de estas, eso nos muestra que tan bueno es estar en un estado y cuanto valor aporta, pero para saber cual es el mejor camino para tomar debemos basarnos en el valor de q, básicamente el valor de q nos demuestra que acciones resultan con un mayor valor, y por tal motivo que acción debemos tomar y podemos ver q visualmente por medio de los arcos que conectan a los estados en nuestro proceso de decisión de Markov.

Política Óptima

Para encontrar el mejor camino dentro de nuestro proceso de decisión de Markov es necesario encontrar la mejor política para este, a continuación veremos como definirla. Para definir cuando una política es mejor

que otra se toma en cuenta que el valor de cada política sea mayor o igual que otra para cada uno de los estados, por lo tanto una política π no puede ser mejor que π' si uno de sus estados tiene menor valor que π'.

Se puede concluir que:

- Siempre va a haber por lo menos una política óptima para cada proceso de decisión de Markov.
- Si existe mas de una política óptima, todas llegan a la misma función de valor.
- Todas las políticas óptimas logran la óptima función de acción-valor.

Ecuación de Optimidad de Bellman

Por último veremos la aplicación de la ecuación de Bellman para encontrar las acciones mas óptimas a tomar dentro del proceso de decisión de Markov. Como se había visto previamente en la ecuación de expectativa de Bellman, iniciando desde un estado s, donde se tenían dos posibles acciones a tomar y se terminaba promediando el valor, a diferencia de ese caso, en la ecuación de optimidad de Bellman se busca el mayor valor, por lo tanto, sólo nos quedaríamos con la acción que tuviera el mayor valor de todas las posibles acciones.

$$v_*(s) = \max_a q_*(s,a)$$

En el caso de el mayor valor para nuestras acciones que son guiadas por las dinámicas de nuestro proceso de decisión de Markov no podríamos simplemente escoger la de mayor valor, por que en este caso es el sistema el que nos mueve de un estado al otro, luego de tomar una acción, así que la formula se mantendría exactamente igual como la vista en la ecuación de la expectativa de Bellman.

Referencias

https://medium.com/aprendizaje-por-refuerzo-introducci%C3%B3n-al-mundo-del/aprendizaje-por-refuerzoprocesos-de-decisi%C3%B3n-de-markov-parte-2-d219358ecd76 (https://medium.com/aprendizaje-porrefuerzo-introducci%C3%B3n-al-mundo-del/aprendizaje-por-refuerzo-procesos-de-decisi%C3%B3n-demarkov-parte-2-d219358ecd76)

In [ ]: