Guía 1: *Multi-armed bandits*

Ejercicio 1

Demostrar que, si conociéramos exactamente el valor de cada acción, es decir si $Q_t(a) = E[R_t | A_t = a]$, entonces la acción *greedy* $A_t = argmax_a Q_t(a)$ es la acción óptima en el sentido de que permite maximizar las recompensas totales.

Ejercicio 2

En una selección de acciones tipo $\epsilon -$*greedy* con dos acciones posibles y $\epsilon = 0.1$, ¿Cuál es la probabilidad de seleccionar la acción *greedy*?

Ejercicio 3

Demostrar que el valor de una acción después de haber sido seleccionada $n-1$ veces, definido como

$$Q_n = \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1},$$

puede calcularse incrementalmente con la siguiente fórmula:

$$Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n].$$

Describa la ventaja de esta fórmula desde un punto de vista computacional.

Ejercicio 4

Considere un problema $k -$*armed bandit* con $k = 4$ acciones. Considere la aplicación de un algoritmo *bandit* usando selección de acciones $\epsilon -$*greedy*, estimación incremental de los valores de cada acción y valores iniciales nulos $Q_1(a) = 0 \ \forall a$. Suponga la siguiente secuencia de acciones y recompensas: $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. En algunos de estos pasos se ha tomado una decisión aleatoria.

a. ¿En qué pasos definitivamente se tomaron decisiones aleatorias?
b. ¿En qué pasos es posible que la decisión haya sido aleatoria?

Ejercicio 5

[Programación] Aplique el algoritmo *bandit* $\epsilon -$*greedy* con $\epsilon = 0$ (*greedy*), $\epsilon = 0.01$ y $\epsilon = 0.1$ a un problema $k -$*armed bandit* con $k = 10$ acciones. Considere recompensas con medias aleatorias y desvío estándar constante σ. Analice experimentalmente el efecto de del desvío estándar σ evaluando tres casos: $\sigma = 0$ (determinístico), $\sigma = 1$ y $\sigma = 10$. ¿Qué conclusiones puede sacar,

Ejercicio 6

Dada la fórmula adaptativa del valor $Q_{n+1} = Q_n + \alpha[R_n - Q_n]$ con $\alpha \in (0,1]$, demostrar que

a. $Q_{n+1} = (1-\alpha)^n Q_n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$,
b. $(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$,

es decir, Q_{n+1} es un promedio pesado de $Q_n, R_1, R_2, \ \ldots, R_n$.

Ejercicio 7

Demostrar que fórmula adaptativa para calcular el valor $Q_{n+1} = Q_n + \alpha[R_n - Q_n]$ con step-size $\alpha \in (0,1]$ constante no verifica las hipótesis del teorema de convergencia y por lo tanto no está garantizada su convergencia.

Ejercicio 8

En la Figura 2.3 del libro Sutton&Barto (2018), se observa un *spike* en el paso número 11 cuando se utiliza inicialización optimista. De una explicación de este fenómeno.

Ejercicio 9

Demuestre que la función SOFTMAX: $p(a) = \frac{e^{H(a)}}{\sum_{a'=1}^{K} e^{H(a')}}$, define una distribución de probabilidades discreta válida.

Ejercicio 10

Demostrar que las derivadas de la función SOFTMAX $p(x)$ respecto de sus parámetros $H(a)$, $a = 1,2,\cdots,K$, son iguales a:

$$\frac{\partial p(x)}{\partial H(a)} = \begin{cases} p(x)(1-p(x)) & si\ \ x = a \\ -p(x)p(a) & si\ \ x \neq a \end{cases}$$

Ejercicio 11

Demostrar que la regla de actualización por gradiente ascendente estocástico:

$$H_{t+1}(a) \ = \ H_t(a) \ + \ \alpha \frac{\partial E[R_t]}{\partial H_t(a)},$$

con $E[R_t] = \sum_{x=1}^{K} p_t(x) q_*(x)$, puede escribirse de la siguiente manera:

$$H_{t+1}(a) \ = \begin{cases} H_t(a) \ + \ \alpha(R_t - C)(1 - p_t(a)) & si\ \ a = A_t \\ H_t(a) - \ \alpha(R_t - C)p_t(a) & si\ \ a \neq A_t \end{cases}$$

donde C es una constante cualquiera (usualmente se usa $C = \bar{R}_t$, el promedio de recompensas anteriores).