Aprendizaje Reforzado: Guía 1: Multi-armed Bandits

José Saint Germain

25 de septiembre de 2025

1. Ejercicio 1

Demostrar que,si conociéramos exactamente el valor de cada acción, es decir si *greedy* $A_t = \arg\max_a Q_t(a)$ en el sentido de que permite maximizar las recompensas totales.

1.1. Respuesta

En primer lugar debemos definir qué es la recompensa total. Una forma de hacerlo es sumando las recompensas en cada instante de tiempo y acumulándolas.

Entonces definimos a la recompensa total G_T como:

$$G_T = \sum_{t=1}^{T} R_t$$

¿Cuál es la recompensa total en promedio? Para saberlo calculamos el retorno esperado. Ese es el valor esperado G_T dada una secuencia de acciones T:

Retorno esperado: $E[G_T|A_1 = a_1, A_2 = a_2, ..., A_T = a_T]$

Vamos a recurrir a las propiedades del valor esperado, dado que es una función lineal.

$$\text{linealidad del } \mathbb{E}[\cdot] \begin{cases} \mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \\ \mathbb{E}[\alpha \cdot X] = \alpha \cdot \mathbb{E}[X] \end{cases}$$

Aplicando esas propiedades obtenemos que:

$$Retornoesperado = \sum_{t=1}^{T} E[R_t|A_t = a_t]$$

$$Retornoesperado = E[R_1|A_1 = a_1] + E[R_2|A_2 = a_2] + ... + E[R_T|A_T = a_T]$$

La esperanza de R_1 dada la secuencia de acciones depende solo de la acción A_1, porque la recompensa en el primer instante es independiente de la acción que se toma en instantes siguientes. Entonces esto hace que esta sumatoria equivalga a la sumatoria de las recompensas condicionadas solo a la acción tomada en el mismo instante de tiempo.

Entonces, para maximizar la suma debo maximizar cada término. Esos términos se maximizan cuando elijo la acción que da la máxima recompensa en promedio. Eso es exactamente igual a elegir $A_t = \arg\max_a Q_t(a)$, es decir la política Greedy.

2. Ejercicio 2

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

2.1. Respuesta

La probabilidad de seleccionar la acción greedy es 1-ϵ, por lo tanto es de 0.9.

3. Ejercicio 3

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

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

, puede calcularse incrementalmente con la siguiente fórmula:

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

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

3.1. Respuesta

Siguiendo la expresión del enunciado, podemos expresar Q_{n+1} de la siguiente manera:

$$Q_{n+1} = \frac{R_1 + R_2 + \ldots + R_{n-1} + R_n}{n}$$

A su vez, esta expresión podemos separarla en:

$$\frac{R_1 + R_2 + \ldots + R_{n-1}}{n} + \frac{R_n}{n}$$

Adicionalmente, para tener $n-1$ en el denominador del primer sumando, podemos multiplicar y dividir por $n-1$:

$$\frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1} \cdot \frac{n-1}{n} + \frac{R_n}{n}$$

Reescribimos $\frac{R_1+R_2+\ldots+R_{n-1}}{n-1}$ como Q_n:

$$= Q_n \cdot \frac{n-1}{n} + \frac{1}{n} R_n \tag{1}$$

$$= Q_n \cdot \frac{n}{n} - \frac{Q_n}{n} + \frac{1}{n} R_n \tag{2}$$

$$= Q_n - \frac{Q_n}{n} + \frac{1}{n} R_n \tag{3}$$

$$= Q_n + \frac{1}{n}[R_n - Q_n] \tag{4}$$

La ventaja de esta fórmula desde un punto de vista computacional es que no es necesario almacenar todas las recompensas obtenidas en cada paso, sino que solo es necesario almacenar el valor actual de Q_n y el número de veces que se ha seleccionado la acción n. Esto reduce el uso de memoria y hace que el cálculo sea más eficiente.

4. Ejercicio 4

Considere un problema k armedbandit con $k = 4$ acciones. Considere la aplicación de un algoritmo bandit usando selección de acciones ϵ-greedy, estimación incremental de los valores de cada acción y valores iniciales nulos $Q(a) = 0 \forall a$. 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.

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

4.1. Respuesta

En el primer paso de la secuencia, todos los valores iniciales son nulos, por lo tanto se eligó una acción aleatoria $A_1 = 1$ con recompensa $R_1 = 1$.

En el segundo paso, podemos calcular $Q(1)_2$ y $Q(2)_1$:

$$Q(1)_2 = Q(1)_1 + \frac{1}{1}[R_1 - Q(1)_1] = 0 + \frac{1}{1}[1 - 0] = 1$$

$$Q(2)_1 = 0 \text{ (valor inicial)}$$

Entonces sabemos que $Q(1)_2 > Q(2)_1$, por lo tanto la acción $A_2 = 2$ con recompensa $R_2 = 1$ fue tomada de forma aleatoria.

En el tercer paso, podemos calcular $Q(1)_2$ y $Q(2)_2$:

$$Q(1)_2 = 1 \text{ (calculado en el paso anterior)}$$
$$Q(2)_2 = Q(2)_1 + \frac{1}{1}[R_2 - Q(2)_1] = 0 + \frac{1}{1}[1 - 0] = 1$$

En este caso $Q(1)_2 = Q(2)_2$, por lo tanto la acción $A_3 = 2$ con recompensa $R_3 = -2$ pudo haber sido tomada de forma aleatoria o greedy.

En el cuarto paso, podemos calcular $Q(1)_2$ y $Q(2)_3$:

$$Q(1)_2 = 1 \text{ (calculado en el paso anterior)}$$
$$Q(2)_3 = Q(2)_2 + \frac{1}{2}[R_3 - Q(2)_2] = 1 + \frac{1}{2}[-2 - 1] = -0{,}5$$

En este caso $Q(1)_2 > Q(2)_3$, por lo tanto la acción $A_4 = 2$ con recompensa $R_4 = 2$ fue tomada de forma aleatoria.

En el quinto y último paso, podemos calcular $Q(1)_2$, $Q(2)_4$ y $Q(3)_1$:

$$Q(1)_2 = 1 \text{ (calculado en el paso anterior)}$$
$$Q(2)_4 = Q(2)_3 + \frac{1}{3}[R_4 - Q(2)_3] = -0{,}5 + \frac{1}{3}[2 - (-0{,}5)] = 0{,}33$$
$$Q(3)_1 = 0 \text{ (valor inicial)}$$

En este caso $Q(1)_2 > Q(2)_4 > Q(3)_1$, por lo tanto la acción $A_5 = 3$ con recompensa $R_5 = 0$ fue tomada de forma aleatoria.

5. Ejercicio 5

[Programación] Aplique el algoritmo bandit $\epsilon - greedy$ con $\epsilon = 0$(greedy), $\epsilon = 0{,}01$ y $\epsilon = 0{,}1$ a un problema κ - armed bandit con $\kappa = 10$ 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?

Con sigma = 0 los banded arm bandits van a dar siempre igual a cero, generando que el algoritmo nunca pueda mejorar su política y quede estancado en ese valor.

Con sigma = 1 y sigma = 10, el algoritmo puede explorar y encontrar mejores políticas, pero con sigma = 10 la varianza es mucho mayor, por lo que el algoritmo logra aproximarse a las colas superiores de las distribuciones que son 10 veces más largas.

6. Ejercicio 6

Dada la fórmula adaptiva del valor $Q_(n+1) = Q_n + a[R_n - Q_n]$ con $a \in (0, 1)$, demostrar que

- $Q_{n+1} = (1-a)^n Q_1 + \sum_{i=1}^{n} a(1-a)^{n-i} R_i$
- $(1-a)^n + \sum_{i=1}^{n} a(1-a)^{n-i} = 1$

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

6.1. Respuesta

6.1.1. Demostración 1

Demostraremos que $Q_{n+1} = (1-a)^n Q_1 + \sum_{i=1}^{n} a(1-a)^{n-i} R_i$

$$Q_{n+1} = Q_n + \alpha[R_n - Q_n] \tag{1}$$
$$Q_{n+1} = Q_n + \alpha R_n - \alpha Q_n \tag{2}$$
$$Q_{n+1} = \alpha R_n + (1-\alpha) Q_n \tag{3}$$
$$Q_{n+1} = \alpha R_n + (1-\alpha)[\alpha R_{n-1} + (1-\alpha) Q_{n-1}] \tag{4}$$
$$Q_{n+1} = \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \tag{5}$$
$$Q_{n+1} = \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2[\alpha R_{n-2} + (1-\alpha) Q_{n-2}] \tag{6}$$
$$Q_{n+1} = \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \cdots + (1-\alpha)^{n-1} \alpha R_1 + (1-\alpha)^n Q_1 \tag{7}$$
$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} a(1-a)^{n-i} R_i \tag{8}$$

6.1.2. Demostración 2

Demostraremos que $(1-\alpha)^n Q_1 + \sum_{i=1}^{n} a(1-a)^{n-i} = 1$

Sabemos que $\sum_{i=1}^{n} a(1-a)^{n-i}$ se puede escribir de la siguiente manera: $\alpha[(1-\alpha)^{n-1} + (1-\alpha)^{n-2} + ... + 1]$.

Como $(1-\alpha) < 1$, podemos usar la fórmula de la serie geométrica:

$$\alpha \frac{1-(1-\alpha)^n}{1-(1-\alpha)} = \alpha \frac{1-(1-\alpha)^n}{\alpha} = 1-(1-\alpha)^n$$

Remplazándolo en la ecuación inicial, obtenemos:

$$(1-\alpha)^n + 1 - (1-\alpha)^n = 1$$

7. Ejercicio 7

Demostrar que la 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 la hipótesis de convergencia y por lo tanto no está garantizada su convergencia.

7.1. Respuesta

Analicemos cada condición por separado cuando $\alpha_n = \alpha$ es constante para todo n.

Condición 1: Verificación de $\sum_{n=1}^{\infty} \alpha_n = \infty$

$$\sum_{n=1}^{\infty} \alpha_n = \sum_{n=1}^{\infty} \alpha \tag{1}$$

$$= \alpha \sum_{n=1}^{\infty} 1 \tag{2}$$

$$= \alpha \cdot \infty = \infty \tag{3}$$

Dado que $\alpha > 0$, esta condición **SÍ se satisface**.

Condición 2: Verificación de $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$

$$\sum_{n=1}^{\infty} \alpha_n^2 = \sum_{n=1}^{\infty} \alpha^2 \tag{4}$$

$$= \alpha^2 \sum_{n=1}^{\infty} 1 \tag{5}$$

$$= \alpha^2 \cdot \infty = \infty \tag{6}$$

Dado que $\alpha^2 > 0$ es constante, esta condición **NO se satisface**.

Puesto que no se verifica la condición 2, la fórmula adaptativa $Q_{n+1} = Q_n + \alpha[R_n - Q_n]$ con $\alpha \in (0, 1]$ constante **NO garantiza su convergencia**.

8. Ejercicio 8

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

8.1. Respuesta

La inicialización optimista hace que todas las acciones tengan un valor inicial alto, lo que incentiva la exploración de todas las acciones. Una vez que la acción óptima es seleccionada, su valor se actualiza con la recompensa real obtenida, que es menor que el valor inicial optimista. Esto provoca que el valor de esa acción disminuya, creando el *spike* observado en la figura.

9. Ejercicio 9

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

9.1. Respuesta

Para que una función defina una distribución de probabilidades discreta válida, debe cumplir dos condiciones:

1. Cada probabilidad debe ser mayor o igual a 0.
2. La suma de todas las probabilidades debe ser igual a 1.

9.1.1. Demostración 1: $0 \leq p_i(a) \leq 1$

Para cualquier acción a:

- $e^{H_i(a)} > 0$ para todo valor de $H_i(a)$, ya que la función exponencial es siempre positiva
- $\sum_{a'} e^{H_i(a')} > 0$, ya que es suma de términos estrictamente positivos

Por lo tanto:

$$p(a) = \frac{e^{H_i(a)}}{\sum_{a'} e^{H_i(a')}} \geq 0$$

A su vez, observemos que $e^{H_i(a)}$ es uno de los términos en la suma $\sum_{a'} e^{H_i(a')}$.

Por definición de suma:

$$e^{H_i(a)} \leq \sum_{a'} e^{H_i(a')}$$

Dividiendo ambos lados por $\sum_{a'} e^{H_i(a')} > 0$:

$$\frac{e^{H_i(a)}}{\sum_{a'} e^{H_i(a')}} \leq 1$$

Entonces, combinando ambas desigualdades:

$$0 \leq p_i(a) = \frac{e^{H_i(a)}}{\sum_{a'} e^{H_i(a')}} \leq 1$$

9.2. Demostración 2: $\sum_a p_i(a) = 1$

Calculemos la suma sobre todas las acciones:

$$\sum_a p_i(a) = \sum_a \frac{e^{H_i(a)}}{\sum_{a'} e^{H_i(a')}}$$

Factorizando el denominador común:

$$= \frac{1}{\sum_{a'} e^{H_i(a')}} \cdot \sum_a e^{H_i(a)}$$

Como estamos sumando sobre todas las acciones posibles, ambas sumas recorren el mismo conjunto:

$$\sum_a e^{H_i(a)} = \sum_{a'} e^{H_i(a')}$$

Sustituyendo:

$$\sum_a p_i(a) = \frac{1}{\sum_{a'} e^{H_i(a')}} \cdot \sum_{a'} e^{H_i(a')} = 1$$

Por lo tanto, la suma de todas las probabilidades es igual a 1.

10. Ejercicio 10

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

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

10.1. Respuesta

Para simplificar la notación, definamos:

- $S = \sum_{k=1}^{K} e^{H(k)}$ (denominador común)
- $p(x) = \frac{e^{H(x)}}{S}$

10.1.1. Caso 1: $\alpha = x$

Aplicamos la derivada del numerador.

$$\frac{\partial}{\partial H(x)}[e^{H(x)}] = e^{H(x)}$$

Aplicamos la derivada del denominador. Recordemos que $S = \sum_{k=1}^{K} e^{H(k)} = e^{H(1)} + e^{H(2)} + ... + e^{H(K)}$

$$\frac{\partial S}{\partial H(x)} = \frac{\partial}{\partial H(x)} \left[e^{H(1)} + e^{H(2)} + ... + e^{H(K)}\right]$$

Aplicando la derivada término por término:

- $\frac{\partial}{\partial H(x)}[e^{H(1)}] = 0$ si $1 \neq x$, pero $= e^{H(1)}$ si $1 = x$
- $\frac{\partial}{\partial H(x)}[e^{H(2)}] = 0$ si $2 \neq x$, pero $= e^{H(2)}$ si $2 = x$
- ...
- $\frac{\partial}{\partial H(x)}[e^{H(x)}] = e^{H(x)}$ (este término siempre contribuye)
- ...

Solo el término $e^{H(x)}$ tiene derivada no-cero, por lo tanto:

$$\frac{\partial S}{\partial H(x)} = e^{H(x)}$$

Aplicamos la regla del cociente:

$$\frac{\partial p(x)}{\partial H(x)} = \frac{e^{H(x)} \cdot S - e^{H(x)} \cdot e^{H(x)}}{S^2}$$

$$= \frac{e^{H(x)}(S - e^{H(x)})}{S^2}$$

$$= \frac{e^{H(x)}}{S} \cdot \frac{S - e^{H(x)}}{S}$$

$$= p(x) \cdot \left(1 - \frac{e^{H(x)}}{S}\right)$$

$$= p(x)(1 - p(x))$$

10.1.2. Caso 2: $\alpha \neq x$

Cuando $\alpha \neq x$, calculamos $\frac{\partial p(x)}{\partial H(\alpha)}$.

Aplicamos derivada del numerador:

$$\frac{\partial}{\partial H(\alpha)}[e^{H(x)}] = 0$$

(ya que $e^{H(x)}$ no depende de $H(\alpha)$ cuando $\alpha \neq x$)

Aplicamos derivada del denominador: Recordemos que $S = \sum_{k=1}^{K} e^{H(k)} = e^{H(1)} + e^{H(2)} + ... + e^{H(K)}$

$$\frac{\partial S}{\partial H(\alpha)} = \frac{\partial}{\partial H(\alpha)} \left[e^{H(1)} + e^{H(2)} + ... + e^{H(K)}\right]$$

Aplicando la derivada término por término:

- $\frac{\partial}{\partial H(\alpha)}[e^{H(1)}] = 0$ si $1 \neq \alpha$, pero $= e^{H(1)}$ si $1 = \alpha$
- $\frac{\partial}{\partial H(\alpha)}[e^{H(2)}] = 0$ si $2 \neq \alpha$, pero $= e^{H(2)}$ si $2 = \alpha$
- ...
- $\frac{\partial}{\partial H(\alpha)}[e^{H(\alpha)}] = e^{H(\alpha)}$ (este término siempre contribuye)
- ...

Solo el término $e^{H(\alpha)}$ tiene derivada no-cero, por lo tanto:

$$\frac{\partial S}{\partial H(\alpha)} = e^{H(\alpha)}$$

Aplicamos regla del cociente:

$$\frac{\partial p(x)}{\partial H(\alpha)} = \frac{0 \cdot S - e^{H(x)} \cdot e^{H(\alpha)}}{S^2}$$

$$= \frac{-e^{H(x)} \cdot e^{H(\alpha)}}{S^2}$$

$$= -\frac{e^{H(x)}}{S} \cdot \frac{e^{H(\alpha)}}{S}$$

$$= -p(x) \cdot p(\alpha)$$

11. Ejercicio 11

Demostrar que la regla de actualización por gradiente ascendiente estocástico:

$$H_{t+1}(\alpha) = H_t(\alpha) + \alpha \frac{\partial E[R_t]}{\partial H_t(\alpha)'}$$

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

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

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

11.1. Respuesta

11.1.1. Definición del Problema

Tenemos la regla de actualización por gradiente ascendente estocástico:

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

donde $E[R_t] = \sum_{x=1}^{K} p_t(x) q_*(x)$ es el valor esperado de la recompensa.

11.1.2. Objetivo

Demostrar que esta regla puede escribirse como:

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

donde:

- C es una constante cualquiera (baseline)
- A_t es la acción seleccionada en el tiempo t
- R_t es la recompensa observada
- α representa la tasa de aprendizaje (renombrada de c)

11.1.3. Paso 1: Calcular el Gradiente del Valor Esperado

El valor esperado de la recompensa es:

$$E[R_t] = \sum_{x=1}^{K} p_t(x) q_*(x)$$

Calculamos $\frac{\partial E[R_t]}{\partial H_t(\alpha)}$:

$$\frac{\partial E[R_t]}{\partial H_t(\alpha)} = \frac{\partial}{\partial H_t(\alpha)} \left[\sum_{x=1}^{K} p_t(x) q_*(x) \right]$$

$$= \sum_{x=1}^{K} q_*(x) \frac{\partial p_t(x)}{\partial H_t(\alpha)}$$

12. Paso 2: Usar las Derivadas de SOFTMAX

De la demostración anterior, sabemos que:

$$\frac{\partial p_t(x)}{\partial H_t(\alpha)} = \begin{cases} p_t(x)(1 - p_t(x)) & \text{si } \alpha = x \\ -p_t(x) p_t(\alpha) & \text{si } \alpha \neq x \end{cases}$$

Sustituyendo:

$$\frac{\partial E[R_t]}{\partial H_t(\alpha)} = \sum_{x=1}^{K} q_*(x) \frac{\partial p_t(x)}{\partial H_t(\alpha)}$$

Separando los casos:

$$= q_*(\alpha) \cdot p_t(\alpha)(1 - p_t(\alpha)) + \sum_{x \neq \alpha} q_*(x) \cdot (-p_t(x) p_t(\alpha))$$

$$= q_*(\alpha) p_t(\alpha)(1 - p_t(\alpha)) - p_t(\alpha) \sum_{x \neq \alpha} q_*(x) p_t(x)$$

13. Paso 3: Simplificar la Expresión

Podemos reescribir:

$$\sum_{x \neq \alpha} q_*(x) p_t(x) = \sum_{x=1}^{K} q_*(x) p_t(x) - q_*(\alpha) p_t(\alpha) = E[R_t] - q_*(\alpha) p_t(\alpha)$$

Sustituyendo:

$$\frac{\partial E[R_t]}{\partial H_t(\alpha)} = q_*(\alpha)p_t(\alpha)(1 - p_t(\alpha)) - p_t(\alpha)(E[R_t] - q_*(\alpha)p_t(\alpha))$$

$$= q_*(\alpha)p_t(\alpha) - q_*(\alpha)p_t^2(\alpha) - p_t(\alpha)E[R_t] + q_*(\alpha)p_t^2(\alpha)$$

$$= q_*(\alpha)p_t(\alpha) - p_t(\alpha)E[R_t]$$

$$= p_t(\alpha)(q_*(\alpha) - E[R_t])$$

13.0.1. Paso 4: El Problema del Gradiente Ascendente Estocástico

En la práctica, **no conocemos** $q_*(\alpha)$ para todas las acciones. Solo observamos:

- La acción seleccionada A_t
- La recompensa obtenida R_t

Por tanto, necesitamos **estimar** el gradiente usando solo esta información limitada.

13.0.2. Paso 5: Aproximación Estocástica del Gradiente

Del Paso 3 obtuvimos:

$$\frac{\partial E[R_t]}{\partial H_t(\alpha)} = p_t(\alpha)(q_*(\alpha) - E[R_t])$$

Idea clave: En lugar de usar todos los $q_*(x)$, usamos solo la información de la acción seleccionada.

Para la acción seleccionada A_t, tenemos la observación $R_t \approx q_*(A_t)$.

13.0.3. Paso 6: Construcción del Estimador

Construimos un estimador no sesgado del gradiente:

Si $\alpha = A_t$ **(acción seleccionada)**:

- Sabemos que R_t es una estimación de $q_*(A_t)$
- Estimamos: $\frac{\partial E[R_t]}{\partial H_t(\alpha)} \approx p_t(\alpha)(R_t - E[R_t])$

Si $\alpha \neq A_t$ (acción no seleccionada):

- No observamos $q_*(\alpha)$ directamente
- Pero por simetría del problema, el efecto debe ser proporcional a $p_t(\alpha)$
- Estimamos: $\frac{\partial E[R_t]}{\partial H_t(\alpha)} \approx -p_t(\alpha)(R_t - E[R_t])$

13.0.4. Paso 7: Introducir el Baseline

Como $E[R_t]$ también es desconocido, lo reemplazamos por un baseline C:

- C puede ser cualquier constante (típicamente el promedio de recompensas pasadas)
- Esto no introduce sesgo porque $E[C] = C$ es constante

13.0.5. Paso 8: La Regla de Actualización Final

Aplicando $H_{t+1}(\alpha) = H_t(\alpha) + c\frac{\partial E[R_t]}{\partial H_t(\alpha)}$ con $c = \alpha$:

Si $\alpha = A_t$ (acción seleccionada):

$$H_{t+1}(\alpha) = H_t(\alpha) + \alpha \cdot p_t(\alpha)(R_t - C)$$

Pero recordemos que de las derivadas de SOFTMAX sabemos que para $\alpha = A_t$:

$$\frac{\partial p_t(A_t)}{\partial H_t(\alpha)} = p_t(\alpha)(1 - p_t(\alpha))$$

El factor correcto es $(1 - p_t(\alpha))$, por tanto:

$$H_{t+1}(\alpha) = H_t(\alpha) + \alpha(R_t - C)(1 - p_t(\alpha))$$

Si $\alpha \neq A_t$ (acción no seleccionada):

$$\begin{aligned} H_{t+1}(\alpha) &= H_t(\alpha) + \alpha \cdot (-p_t(\alpha))(R_t - C) \\ &= H_t(\alpha) - \alpha(R_t - C)p_t(\alpha) \end{aligned}$$

13.0.6. Verificación de Consistencia

Podemos verificar que este estimador es no sesgado:

$$E\left[\sum_{\alpha=1}^{K}\frac{\partial E[R_t]}{\partial H_t(\alpha)}\right] = E\left[(1-p_t(A_t))(R_t-C) - \sum_{\alpha\neq A_t} p_t(\alpha)(R_t-C)\right]$$

$$= E\left[(R_t-C)\left(1-p_t(A_t)-\sum_{\alpha\neq A_t} p_t(\alpha)\right)\right]$$

$$= E\left[(R_t-C)(1-\sum_{\alpha=1}^{K} p_t(\alpha))\right] = E[(R_t-C)\cdot 0] = 0$$

Esto confirma que nuestro estimador preserva la propiedad de que la suma de todos los gradientes es cero.

Por lo tanto:

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

13.0.7. Interpretación

1. **Acción seleccionada** ($\alpha = A_t$):
   - Si $R_t > C$: aumenta $H_t(\alpha) \rightarrow$ aumenta $p_t(\alpha)$
   - Si $R_t < C$: disminuye $H_t(\alpha) \rightarrow$ disminuye $p_t(\alpha)$
2. **Acciones no seleccionadas** ($\alpha \neq A_t$):
   - Si $R_t > C$: disminuye $H_t(\alpha) \rightarrow$ disminuye $p_t(\alpha)$
   - Si $R_t < C$: aumenta $H_t(\alpha) \rightarrow$ aumenta $p_t(\alpha)$
3. **El baseline** C: No afecta el valor esperado del gradiente, pero reduce la varianza