Multi-Arm Bandits.

* Valor real de una acción: valor o recompensa promedio que se obtiene cuando se escoge la acción. lo denotamos por $q(a)$.

* Valor estimado de una acción "a" en un instante "t": $Q_t(a)$.

$$Q_t(a) = \frac{\sum_i^t R_i}{N_t(a)}$$

Donde $N_t(a)$ es el número de veces que se ha seleccionado "a" hasta el instante "t".

Si $N_t(a) = 0$, ~~deno~~ entonces $Q_t(a)$ adopta un valor por defecto, p.ej. $Q_t(a) = 0$.

cuando $N_t(a) \longrightarrow \infty$, entonces $Q_t(a) \longrightarrow q(a)$

$$\lim_{N_t(a) \to \infty} Q_t(a) = q(a)$$ (Por la ley de los grandes números).

¿Cómo usamos $Q_t(a)$ y para qué? Una $Q_t(a)$ nos ayuda a seleccionar acciones en cada paso temporal. Una regla sencilla (la más sencilla) para escoger acciones en cada instante es seleccionar aquella con un valor estimado más alto:

$$A_t^* = \arg\max_a Q_t(a)$$

Si siempre seguimos esta fórmula para escoger A_t, seguimos una selección de acciones "glotona" o "avara". Si permitimos que, con cierta probabilidad baja ε, se seleccione una acción que no es la que maximiza $Q_t(a)$ pero que ~~reduce nuestra~~ permite explorar en busca de otras acciones (más) óptimas, entonces ésta es una regla de selección ε-avara.

$$Q_{k+1} = \frac{1}{k}\sum_{i=1}^{k} R_i = \frac{1}{k}\left(R_k + \sum_{i=1}^{k-1} R_i\right) = \frac{1}{k}\left[R_k + (k-1)\,Q_k\right] =$$

$$= \frac{1}{k}\left[R_k + (k-1)Q_k + Q_k - Q_k\right] = \frac{1}{k}\left[R_k + k\,Q_k - Q_k + Q_k - Q_k\right] =$$

$$= \frac{1}{k}\left[R_k + k\,Q_k - Q_k\right] = Q_k + \frac{1}{k}\left[R_k - Q_k\right]$$

Por tanto, en vez de necesitar guardar un array $R_1, \dots, R_{k-1}$, guardamos

$\to Q_k$

$\to k$, obtenemos R_k y calculamos

Q_{k+1}.

¿Qué pasa si el premio no es estacionario?

→ En ese caso, tiene sentido darle más peso a ~~prem~~ recompensas nuevas que a las antiguas.

¡Ojo! Con el truco anterior, hemos llegado a

$$Q_{k+1} = Q_k + \frac{1}{k}\left[R_k - Q_k\right]$$

Que es de la forma:

Estimación Nueva $\leftarrow$ Estimación anterior $+\ \omega \cdot$ [error]

Donde el error es la diferencia entre la estimación anterior del valor de la acción y la recompensa recibida. Esto es: estamos acercando Q_{k+1} a R_k una $\frac{1}{k}$ parte más que Q_k.

Esto funciona sólo si ~~la recompensa que se obtiene~~ el valor $q(a)$ es estacionario. Sin embargo, si $q(a)$ varía con el tiempo, entonces hay que dar más peso a las recompensas recientes.

En vez de usar $\frac{1}{k}$ con k variable, usemos α y fija: $\alpha \in \,]0,1]$

$$Q_{k+1} = Q_k + \alpha\left[R_k - Q_k\right] = \alpha R_k + Q_k - \alpha Q_k = \alpha R_k + (1-\alpha) Q_k =$$

$$= \alpha R_k + (1-\alpha)\left[\alpha R_{k-1} + (1-\alpha) Q_{k-1}\right] =$$

$$= \alpha R_k + (1-\alpha)\cdot\alpha R_{k-1} + (1-\alpha)^2 Q_{k-1} =$$

$$= \alpha R_k + (1-\alpha)\alpha R_{k-1} + \alpha(1-\alpha)^2 R_{k-2} + \ldots \alpha(1-\alpha)^{k-1} R_1 +$$

$$+ (1-\alpha)^k Q_1$$

Es decir: $Q_{k+1} = (1-\alpha)^k Q_1 + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} R_i$

lo anterior es un promedio pesado porque $(1-\alpha)^k + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} = 1$

El peso $\alpha(1-\alpha)^{k-i}$ que se le da a R_i depende de cuántos pasos temporales separan a R_i del presente $(k-i)$

$(1-\alpha) < 1 \Rightarrow$ a mayor $(k-i)$, menor peso. De hecho, α nos permite controlar el ritmo de decrecimiento de esta función.

Si $\alpha = 1 \rightarrow$ sólo importa R_k (se asume por convención que $0^0 = 1$).

De alguna forma, α mide la "memoria" del algoritmo.

Podemos tomar α variable ($\alpha = \alpha_k(a)$). Para que ~~sea una buena función~~ se asegure convergencia, debe cumplir:

$$\sum_{k=1}^{\infty} \alpha_k(a) = \infty \quad y \quad \sum_{k=1}^{\infty} \alpha_k^2(a) < \infty$$

El caso $\alpha_k(a) = \frac{1}{k}$ garantiza ambas ~~cos~~ condiciones.

La ecuación quedaría en ese caso:

$$Q_{k+1} = \left(1 - \frac{1}{k}\right)^k Q_1 + \sum_{i=1}^{k} \frac{1}{k}\left(1-\frac{1}{k}\right)^{k-i} R_i \quad (*)$$

OJO! Esto no es lo mismo que la ec. 2.3 de Sutton y Barto!

~~Las ecuaciones anteriores~~

En ese caso, tenemos:

$$Q_{k+1} = Q_k + \alpha_k [R_k - Q_k] = \alpha_k R_k + \overset{(1-\alpha_k)}{\vee} Q_k =$$

$$= \cancel{\frac{1}{k} R_k + \left(1-\frac{1}{k}\right)\left[\frac{1}{k-1}\left(R_{k-1} - Q_{k-1}\right)\right] = \frac{R_k}{k} + \frac{1}{k-1}\left(1-\frac{1}{k}\right) R_{k-1}}$$

$$= \alpha_k R_k + (1-\alpha_k)\left[\alpha_{k-1} R_{k-1} + (1-\alpha_{k-1}) Q_{k-1}\right] =$$

$$= \alpha_k R_k + \alpha_{k-1}\left[1-\alpha_k\right] R_{k-1} + (1-\alpha_k)(1-\alpha_{k-1}) Q_{k-1} = \dots$$

... (Que no ~~es~~ lleva a la misma ecuación que $*$)

2.5 Valores iniciales optimistas.

Todos los métodos anteriores se pueden escribir en términos de Q_1 (están sesgados por el valor inicial).

Lo bueno de ésto es que podemos usarlo para introducir algún tipo de conocimiento a priori de qué valores de las recompensas se han de esperar.

Valores ~~a~~ de Q_1 altos empujan al algoritmo a esforzarse más durante la exploración.

2.6. Upper-Confidence-Bound Action Selection. (Selección de acciones por el límite superior de confianza).

~~Es~~

El método ε-greedy fuerza a la exploración ~~en~~ un ε% de las veces, pero lo hace indiscriminadamente. Sería interesante hacerlo con criterio. Una criterio de selección más informado ~~es~~ es:

$$A_t = \arg\max_a \left[Q_t(a) + c \cdot \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Medida de la ~~varianza~~ incertidumbre asociada a una acción

→ c determina el intervalo de confianza. Controla el grado de exploración.

→ el cociente ~~del~~ es una "estimación de la varianza": si una acción no se ha probado a menudo, $N_t(a)$ será pequeño y el cociente será grande (mucha incertidumbre). Por otra parte, cada vez que una acción que no es "a" se escoge, la incertidumbre de "a" aumenta (lentamente).

2.7. "Bandas de Gradiente".

Podemos también seleccionar acciones en función de una preferencia que definimos sobre ellas: $H_t(a)$.

La preferencia no tiene interpretación en términos de recompensa.

Lo que importa en este caso es el orden ~~entre~~ ~~defin~~ entre acciones definido por las preferencias, no el valor de las preferencias per se.

Distribución de las preferencias:

$$Pr\{A_t = a\} = \frac{e^{H_t(a)}}{\sum_{b=1}^{n} e^{H_t(b)}} = \pi_t(a)$$

$\Rightarrow$ A mayor preferencia de "a", mayor probabilidad de "a".

$$H_{t+1}(A_t) = H_t(A_t) + \alpha \cdot (R_t - \bar{R}_t) \cdot [1 - \pi_t(A_t)]$$

$$H_{t+1}(a) = H_t(a) - \alpha (R_t - \bar{R}_t) \pi_t(a) \quad \forall a \neq A_t, \text{ con } \alpha > 0.$$

Podemos entender este algoritmo mejor si lo interpretamos como una aproximación estocástica a un ascenso de gradiente:

La preferencia de cada acción se incrementa de manera proporcional al incremento producido por esa acción en el rendimiento:

$$H_{t+1}(a) \doteq H_t(a) + \alpha \cdot \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$

Donde $\mathbb{E}[R_t] = \sum_b \pi_t(b) \cdot q(b)$

Es decir: selecciono una acción "a" según $\pi_t(a)$.

Sabemos que ~~el~~ ~~pr~~ la recompensa recibida en total (promedio) se calcula como:

$$E[R_t] = \sum_b \pi_t(b) q(b)$$

Y entonces podemos obtener el peso o la medida en que $E[R_t]$ ha variado debido a $H_t(a)$, ya que $\pi_t(a) = \dfrac{e^{H_t(a)}}{\sum_b e^{H_t(b)}}$:

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial \sum_b \pi_t(b) q(b)}{\partial H_t(a)} = \sum_b q(b) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

Veamos este diferencial:

Si $a = b$:

$$\frac{\partial \pi_t(a)}{\partial H_t(a)} = \frac{\partial \left[\dfrac{e^{H_t(a)}}{\sum_b e^{H_t(b)}} \right]}{\partial H_t(a)} = \frac{e^{H_t(a)} \cdot \sum_b e^{H_t(b)} - e^{H_t(a)} \cdot e^{H_t(a)}}{\left[\sum_b e^{H_t(b)} \right]^2} =$$

$$= \frac{e^{H_t(a)} \cdot \left[\sum_b e^{H_t(b)} - e^{H_t(a)} \right]}{\left[\sum_b e^{H_t(b)} \right]^2} = \frac{e^{H_t(a)}}{\sum_b e^{H_t(b)}} \cdot \frac{\sum_b e^{H_t(b)} - e^{H_t(a)}}{\sum_b e^{H_t(b)}} =$$

$$= \pi_t(a) \left[1 - \pi_t(a) \right]$$

Si $a \neq b$:

$$\frac{\partial \pi_t(b)}{\partial H_t(a)} = \frac{- e^{H_t(b)} \cdot e^{H_t(a)}}{\left[\sum_b e^{H_t(b)} \right]^2} = - \pi_t(b) \cdot \pi_t(a)$$

Por tanto:

$$\frac{\partial \Pi_t(b)}{\partial H_t(a)} = \Pi_t(a)\,(\mathbb{1}_{a=b} - \Pi_t(b)) \;\; \forall a, b$$

Y entonces $\dfrac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_b q(b)\, \Pi_t(a)\,[\mathbb{1}_{a=b} - \Pi_t(b)] =$

$$= \sum_b \left(q(b) - \underline{X}_t\right) \Pi_t(a)\,[\mathbb{1}_{a=b} - \Pi_t(b)] \quad (*)$$

- - - -

Nótese que $\sum_b \Pi_t(a)\,[\mathbb{1}_{b=a} - \Pi_t(b)] =$

$$= \sum_b \Pi(a) \cdot \mathbb{1}_{b=a} - \sum_a \Pi_t(a)\Pi_t(b) = \Pi(b) - \Pi(b) = 0.$$

Y por tanto $X_t \sum_b \Pi_t(a)\,[\mathbb{1}_{a=b} - \Pi_t(b)] = 0$ y se puede incluir sin peligro!

- - - -

Después de llegar a $(*)$, multiplicamos y ~~sumam~~ dividimos por $\Pi_t(b)$:

$$\cancel{\sum_b \Pi_t(b)\,(q(b) - \underline{X}_t)\,\frac{1}{\Pi_t(b)}}\;\frac{\partial \Pi_t(b)}{\partial H_t(a)}$$

$$\sum_b \left[(q(b) - \overline{X}_t)\, \Pi_t(b)\; \overset{A_t}{\Pi_t(b)}\,(\mathbb{1}_{a=A_t} - \Pi_t(a)) \cdot \frac{1}{\Pi_t(A_t)} \right] =$$

$$= \mathbb{E}\left[(R_t - \overline{R}_t)\, \Pi_t(A_t)\,(\mathbb{1}_{a=A_t} - \Pi_t(a)) \cdot \frac{1}{\Pi_t(A_t)} \right]$$