

練 2.1

$$1 - \varepsilon + \frac{\varepsilon}{2} = 0.75 \quad //$$

練 2.2

0 step 後 ... $Q_1(a) = 0 \quad (\forall a)$

$A_1 = 1$ は greedy は選ばれず or ランダムに選択された.

1 step 後 ... $Q_2(a) = \begin{cases} 1 & (a=1) \\ 0 & (a \neq 1) \end{cases}$

$A_2 = 2$ は ランダムに選択された.

2 step 後 ... $Q_3(a) = \begin{cases} 1 & (a=1, 2) \\ 0 & (\text{otherwise}) \end{cases}$

$A_3 = 2$ は greedy は選ばれず or ランダムに選択された

3 step 後 ... $Q_4(a) = \begin{cases} 1 & (a=1) \\ \frac{3}{2} & (a=2) \\ 0 & (\text{otherwise}) \end{cases}$

$A_f = 2$ は greedy に選ばれて or ランダムに選択された。

4 step 後 ... $Q_5(a) = \begin{cases} 1 & (a=1) \\ \frac{5}{3} & (a=2) \\ 0 & (\text{otherwise}) \end{cases}$

$A_5 = 3$ は ランダムに選ばれた。

練2.3

$\varepsilon = 0, 0.01, 0.1$ の比較

- 長期的には $\varepsilon = 0.01$ や 0.1 で最適方策が得られる。得られた後には最適行動を選択する確率は $\varepsilon = 0.01$ の方が下ままで。

累積報酬では $\varepsilon = 0.01$ の方が大きくなる。

練2.4

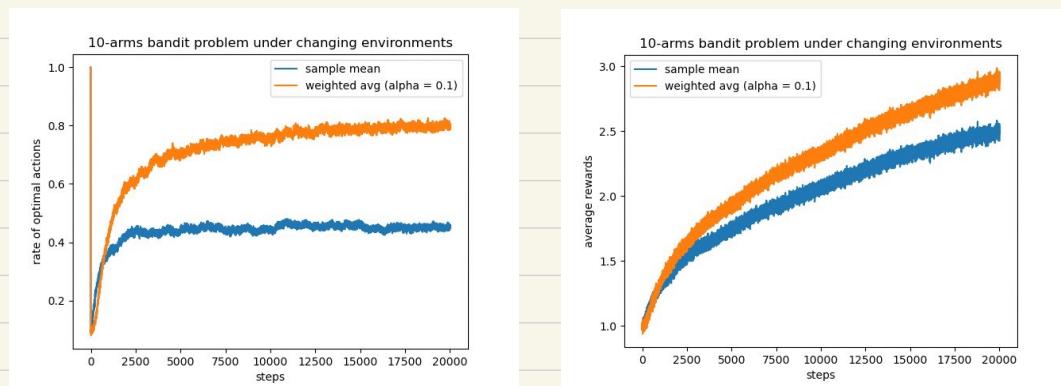
$$\begin{aligned} Q_{n+1} &= Q_n + \alpha_n (R_n - Q_n) \\ &= (1 - \alpha_n) \left\{ Q_{n-1} + \alpha_{n-1} (R_{n-1} - Q_{n-1}) \right\} + \alpha_n R_n \end{aligned}$$

:

$$= \left\{ \prod_{k=1}^n (1 - \alpha_k) \right\} Q_1 + \sum_{i=1}^n \left\{ \alpha_i R_i \prod_{j=i+1}^n (1 - \alpha_j) \right\}$$

$$R_i \text{ の重み } \prod_{j=i+1}^n (1 - \alpha_j) \quad //$$

練2.5 2000回の実験(平均)



練2.6

- 最初の何回かで (≥ 10 step), optimal actionを学習。このgreedyは遂行率で一時的に最適行動の割合が上昇する。しかし, $Q_t(a^*) + \gamma q^*(a^*)$ に収束するにつれて, 他の行動, $Q_t(a)$ の割合 $Q_t(a^*)$ より大きくなるべからず。すると, 最適行動の割合が再び低下する。

練2.7

$$\bar{O}_n := \bar{O}_{n-1} + \alpha (1 - \bar{O}_{n-1}) \quad (\bar{O}_0 := 0)$$

$$\beta_n = \frac{\alpha}{\bar{O}_n}$$

$$\text{つまり}, \beta_1 = 1, \beta_2 = \frac{1}{2-\alpha}, \beta_3 = \frac{1}{(1-\alpha)(2-\alpha)}, \dots$$

$$\text{つまり}, Q_2 = Q_1 + \beta_1 (R_1 - Q_1) = R_1$$

$$Q_3 = Q_2 + \beta_2 (R_2 - Q_2) = \frac{1-\alpha}{2-\alpha} R_1 + \frac{1}{2-\alpha} R_2$$

Q₁を固定

(証)

さて, \bar{O}_n の一般項を求める.

$$\bar{O}_n = (1-\alpha) \bar{O}_{n-1} + \alpha$$

$$\bar{O}_n - 1 = (1-\alpha) (\bar{O}_{n-1} - 1)$$

$$\bar{O}_n = 1 - (1-\alpha)^n$$

$\beta_1 = 1, \beta_n = \alpha$ の 単調減少性.

$$\text{つまり}, \beta_n - \text{一般項}, \beta_n = \frac{\alpha}{1 - (1-\alpha)^n} \quad (n \geq 1)$$

また,

$$Q_{n+1} = Q_n + \beta_n (R_n - Q_n)$$

$$\begin{aligned}
 &= (1 - \beta_n) (Q_{n-1} + \beta_{n-1} (R_{n-1} - Q_{n-1})) + \beta_n R_n \\
 &\vdots \\
 &= \underbrace{\left\{ \prod_{i=1}^n (1 - \beta_i) \right\}}_0 Q_1 + \sum_{i=1}^n \left\{ R_i \beta_i \prod_{j=i+1}^n (1 - \beta_j) \right\} \\
 &= \sum_{i=1}^n \left\{ R_i \beta_i \prod_{j=i+1}^n (1 - \beta_j) \right\}
 \end{aligned}$$

Q_1 为随机变量。且有：

$$\frac{\beta_{i+1} \prod_{j=i+2}^n (1 - \beta_j)}{\beta_i \prod_{j=i+1}^n (1 - \beta_j)} = \frac{\beta_{i+1}}{\beta_i (1 - \beta_{i+1})} = \frac{\frac{\alpha}{1 - (1-\alpha)^{i+1}}}{\frac{\alpha}{1 - (1-\alpha)^i} \cdot \frac{1 - (1-\alpha)^{i+1} - \alpha}{1 - (1-\alpha)^{i+1}}}$$

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

$$= \frac{1}{1 - \alpha}$$

即，重对数随机变量的极限分布为 $(\alpha < 1, \epsilon \neq 0) \quad (\alpha = 1, \epsilon \neq 0, \beta_n = \alpha)$

$$\text{最短} \Leftrightarrow \sum_{i=1}^n \beta_i \prod_{j=i+1}^n (1-\beta_j) = 1 \text{ を示す}.$$

(i) $n=1$ の場合も自明

(ii) $n=k$ の場合も成立する。

$$\sum_{i=1}^{k+1} \beta_i \prod_{j=i+1}^{k+1} (1-\beta_j) = \beta_{k+1} + (1-\beta_{k+1}) \underbrace{\sum_{i=1}^k \beta_i \prod_{j=i+1}^k (1-\beta_j)}_1 = 1$$

練2.8

最初の10回で、全ての行動が1回ずつ選択される。11回目では、最初の10回で報酬率が最大のもとが選ばれることで、平均報酬率が大きくなる。12回目以降は、11回目で選んで行動した $N_t(a)$ が大きくなるので、その行動が選ばれる確率。
(特にこの下限は)

練2.9

$$\begin{aligned} \text{Tr}(a_1) &= \frac{e^{H_t(a_1)}}{e^{H_t(a_1)} + e^{H_t(a_2)}} \\ &= \frac{1}{1 + e^{H_t(a_2) - H_t(a_1)}} \\ &= \frac{1}{1 + e^{-(H_t(a_1) - H_t(a_2))}} \end{aligned}$$

$H_t(a_1) - H_t(a_2)$ が大きいほど、 a_1 を選ぶ。 //

練2.10

各ステップでの状態にかかる確率を

行動1を選ぶ確率を $\pi(a_1)$ とすると、

$$\begin{aligned} E[R] &= 0.5 \times \left(0.1 \times \pi(a_1) + 0.2 \times (1 - \pi(a_1)) \right) \\ &\quad + 0.5 \times \left(0.9 \times \pi(a_1) + 0.8 \times (1 - \pi(a_1)) \right) \\ &= 0.5 \times (1) \\ &= 0.5 \end{aligned}$$

つまり、 $\pi(a_1)$ をどのように設定しても、報酬の期待値は0.5。

• 各ステップでの状態がわかること。

H-SAで行動1を選択する確率を $\pi_A(a_1)$ ($H\rightarrow B$ も同様) とする。

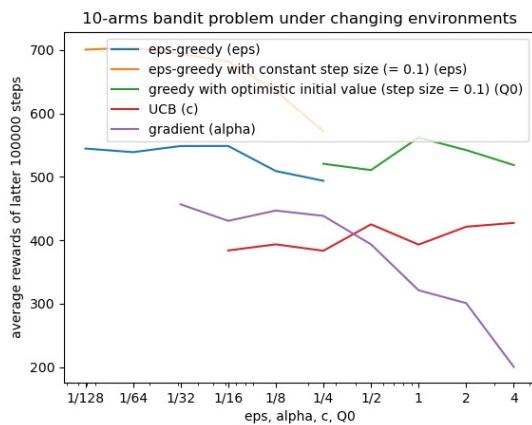
$$E[R] = 0.5 \times (0.1 \times \pi_A(a_1) + 0.2 \times (1 - \pi_A(a_1)))$$

$$+ 0.5 \times (0.9 \times \pi_B(a_1) + 0.8 \times (1 - \pi_B(a_1)))$$

$$\leq 0.55 \quad (\text{等号成立は } \pi_A(a_1) = 0, \pi_B(a_1) = 1)$$

//

練2.11 100回実験して平均をとる。



• 非定常環境下で、step size一定のeps-greedyが最も良い。