Exercise 2.7

$$\beta_n = \frac{\alpha}{O_n}$$

$O_0 = 0$.

$O_1 = 0 + \alpha.(1-0) = \alpha$.

$O_2 = \alpha + \alpha(1-\alpha)$

$= \alpha + \alpha - \alpha^2 = 2\alpha - \alpha^2$

$= 1 - (1-\alpha)^2$

$O_3 = 1 - (1-\alpha)^2 + \alpha(1 - (1-(1-\alpha)^2))$

$= 1 - (1-\alpha)^2 + \alpha(1-\alpha)^2$

$= 1 - (1-\alpha)^2 [1-\alpha]$

$= 1 - (1-\alpha)^3$

$\therefore O_n = 1 - (1-\alpha)^n$.

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

$$= \beta_n R_n + (1-\beta_n)\left[\beta_{n-1} R_{n-1} + (1-\beta_{n-1}) \cdot Q_{n-1}\right]$$

$$= \beta_n R_n + (1-\beta_n)\beta_{n-1} R_{n-1} + (1-\beta_n)\cdot(1-\beta_{n-1}) Q_{n-1}$$

$$= \beta_n R_n + (1-\beta_n)\cdot\beta_{n-1} R_{n-1} + (1-\beta_n)(1-\beta_{n-1})\left[\beta_{n-2} R_{n-2} + (1-\beta_{n-2}) Q_{n-2}\right]$$

$$= \beta_n R_n + (1-\beta_n)\cdot\beta_{n-1} R_{n-1} + (1-\beta_n)\cdot(1-\beta_{n-1})\beta_{n-2} R_{n-2}$$

$$+ (1-\beta_n)\cdot(1-\beta_{n-1})(1-\beta_{n-2})\cdot Q_{n-2}$$

$\gg \lambda$

$$= \sum_{i=0}^{n-1} \beta_{n-i} \prod_{j=0}^{i} (1-\beta_{n-j})\, R_{a_i}$$

$$+ \prod_{i=1}^{n} (1-\beta_i)\, Q_{a_1}$$

$$\Rightarrow \lambda + \prod_{i=1}^{n} \left(1 - \frac{\alpha}{1-(1-\alpha)^i}\right) Q_1 \cdot \gamma$$

as $n \to \infty$

$$1 - \left(\frac{\alpha}{1-(1-\alpha)}\right)^n .$$

when $n = 1$.

$$\beta_1 = \frac{1}{1-(1-\alpha)^0} \cdot \frac{\alpha}{1-(1-\alpha)^1} = 1 .$$

$$\therefore \text{So the product } \prod_{i=1}^{n} (1-\beta_i) = 0$$

hence

the term γ will be 0, so bias

due to θ_1 disappears.

Ex. 2.6

The early part of the graph shows spike because it is exploring during that time and sometimes it will take the ~~best~~ optimal step and sometime it won't. But once, it gets a stable estimate, the % of the optimal actions grows.

To avoid spikes, we can discourage exploring and encourage the system to do more of exploitation.