2. An Example

We make two assumptions before we start with the example. For simplicity we consider the random environment as a *stationary random environment* and we are using the *linear reward-penalty scheme*.

Let's say, the robot roams through a room and shall learn how to avoid obstacles, then a *stationary random environment* simply means, that the probabilities are everywhere in the room the same, that the robot will hit an obstacle. We will discuss later in detail how to gain such penalty probabilities as a function of the position in the room.

Let us assume, the robot can choose from the set $\underline{\alpha} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ of actions. We could define these actions for instance as follows: α_1: *drive forwards*, α_2: *drive backwards*, α_3: *turn right* and α_4: *turn left.*

$i = 1, 2, ..., r$

$r = 4$

$p_i(1) = \frac{1}{r} = \frac{1}{4}$

Let $a = b = \frac{1}{2}$

Let's assume now, the initial action α_1 (which has been selected randomly) has led to an input $\beta = 0$ (reward) at the time $t = n$. The new probabilities are then calculated as follows:

$$\beta = 0: \quad p_j(n+1) = \begin{cases} p_j(n) + a \cdot (1 - p_j(n)) & j = i \\ p_j(n) \cdot (1 - a) & \forall\, j \neq i \end{cases}$$

$p_j(n+1) = p_j(n) + a \cdot (1 - p_j(n))$ for α_1

$$p_j(n+1) = \frac{1}{4} + \frac{1}{2} \cdot \left(1 - \frac{1}{4}\right) = \frac{5}{8}$$

$p_j(n+1) = p_j(n) \cdot (1 - a)$ for $\alpha_2, \alpha_3, \alpha_4$

$$p_j(n+1) = \frac{1}{4} \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{8}$$

As it is requested that $\forall\, n: \sum_{i=1}^{r} p_i(n) = 1$

$$\sum_{j=1}^{r} p_j(n+1) = \frac{5}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 1$$

I.e. after the input from the environment $\beta(n) = 0$, the probability that action α_1 will be chosen as action $\alpha(n+1)$ has been increased to $\frac{5}{8}$ while the probabilities that one of the actions α_2, α_3 or α_4 will be chosen has been decreased to $\frac{1}{8}$.

The same we compute now if the initial action α_1 has led to an input $\beta = 1$ (penalty) at the time $t = n$.

$$\beta = 1: \quad p_j(n+1) = \begin{cases} p_j(n) \cdot (1-b) & j = i \\ \dfrac{b}{r-1} + p_j(n) \cdot (1-b) & \forall j \neq i \end{cases}$$

$p_j(n+1) = p_j(n) \cdot (1-b)$ for α_1

$$p_j(n+1) = \frac{1}{4} \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{8}$$

$p_j(n+1) = \dfrac{b}{r-1} + p_j(n) \cdot (1-b)$ for $\alpha_2, \alpha_3, \alpha_4$

$$p_j(n+1) = \frac{\frac{1}{2}}{4-1} + \frac{1}{4} \cdot \left(1 - \frac{1}{2}\right) = \frac{7}{24}$$

$$\sum_{j=1}^{r} p_j(n+1) = \frac{1}{8} + \frac{7}{24} + \frac{7}{24} + \frac{7}{24} = 1$$

I.e. after the input from the environment $\beta(n) = 1$, the probability that action α_1 will be chosen as action $\alpha(n+1)$ has been decreased to $\frac{1}{8}$ while the probabilities that one of the actions α_2, α_3 or α_4 will be chosen has been increased to $\frac{7}{24}$.

From this example it can be also seen immediately that the limits of a probability p_i for $n \to \infty$ are either 0 or 1. Therefore the robot learns to choose the optimal action asymptotically. It shall be noted, that it converges not always to the correct action; but the probability that it converges to the wrong one can be made arbitrarily small by making the learning parameter a small.