Assumptions:

- If the agent exicutes ϵ-greedy with constant ϵ, the environment does not need to be continious
- If the agent exicutes a policy that converges to ϵ-greedy with constant ϵ in the limit $t \to \infty$ then the environment needs to be continious around $p_i = \epsilon$ and $p_i = 1 - \epsilon$

If the agent behaves like ϵ-greedy, for at fixed $\epsilon > 0$, in the limit $t \to \infty$ then the set of explored action probabilities will eventually be dominated by values p_i such that $p_i \in \{\epsilon, 1-\epsilon\}$.

For some q, we have that

$$\langle p \rangle = q(1-\epsilon) + (1-q)\epsilon = q - 2q\epsilon + \epsilon \tag{1}$$

$$\begin{aligned} \langle p^2 \rangle &= q(1-\epsilon)^2 + (1-q)\epsilon^2 = q - 2q\epsilon + \epsilon^2 & (2) \\ &= \langle p \rangle - \epsilon(1-\epsilon) & (3) \end{aligned}$$

This allowes us to simplify the expected learned utility difference, Δ (see table) as a function of only $\langle p \rangle$ and the constants of the problem. Note that because $\langle p \rangle$ is dominated by $p_i \in \{\epsilon, 1-\epsilon\}$, we must get $\epsilon \le \langle p \rangle \le (1-\epsilon)$

Becasue of law ot large numbers (?), Δ will eventual eventually describe the agents belif arbitary well. Let $\Delta = Q(a_1) - Q(a_2)$ and $p_i = P(a_1|t=i)$. If currently $\Delta > 0$, the agent will exicute $p_i = 1 - \epsilon$ and drive $\langle p \rangle$ towards this value. If currently $\Delta < 0$, the agent will exicute $p_i = \epsilon$ and drive $\langle p \rangle$ towards that value. Therefore the only posible stable points are

$$\langle p \rangle = \epsilon \quad \& \quad \Delta \le 0 \tag{4}$$

$$\langle p \rangle = 1 - \epsilon \quad \& \quad \Delta \ge 0 \tag{5}$$

$$\Delta = 0 \quad \& \quad \frac{d\Delta}{d\langle p \rangle} \le 0 \tag{6}$$

(4) and (5) represent constant p_i at ether extreme, ϵ or $1-\epsilon$. (6) represent fluctuating p_i but with a stable $\langle p \rangle$ in between the two extremes.

Concider an environment with concists of the agent playing a game agianst it self with an arbitary payof matrix. The copy will always use the same action probabilites as the agent, for any given round.

		copy	
		a_1	a_2
self	a_1	M_{11}	M_{12}
	a_2	M_{12}	M_{12}

This generalisation covers, for example, *Prisoners' dilemma agains copy* and *Death in Damaskus* but not *Absent minded driver* and *Evidential blackmail.*