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, with convergent $\epsilon < 1/2$, 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

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

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

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 $\epsilon \leq \langle p \rangle \leq (1 - \epsilon)$

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

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

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

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

(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 agent playing a game with an arbitary payof matrix, agianst a copy of it self. 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_{21}	M_{22}

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

If $p_i \in \{\epsilon, 1-\epsilon\}$, then

$$\Delta = M_{11} - M_{22} + \left(\frac{M_{12} - M_{11}}{\langle p \rangle} - \frac{M_{21} - M_{22}}{1 - \langle p \rangle} \right) \epsilon(1-\epsilon) \tag{7}$$

$$\frac{d\Delta}{d\langle p \rangle} = \left(\frac{M_{11} - M_{12}}{\langle p \rangle^2} - \frac{M_{21} - M_{22}}{(1 - \langle p \rangle)^2} \right) \epsilon(1-\epsilon) \tag{8}$$

The roots of this expression are

$$\Delta = 0 \implies \begin{cases} \langle p \rangle = \dfrac{M_{11} - M_{12}}{M_{11} - M_{22}} \epsilon + \mathcal{O}(\epsilon^2) \\ \quad \textbf{or} \\ 1 - \langle p \rangle = \dfrac{M_{21} - M_{22}}{M_{11} - M_{22}} \epsilon + \mathcal{O}(\epsilon^2) \end{cases} \tag{9}$$

The derivatives at those points are

$$\left. \frac{d\Delta}{d\langle p \rangle} \right|_{\langle p \rangle = \frac{M_{11} - M_{12}}{M_{11} - M_{22}} \epsilon} = \frac{(M_{11} - M_{22})^2}{M_{11} - M_{12}} + \mathcal{O}(\epsilon) \tag{10}$$

$$\left. \frac{d\Delta}{d\langle p \rangle} \right|_{1 - \langle p \rangle = \frac{M_{21} - M_{22}}{M_{11} - M_{22}} \epsilon} = \frac{(M_{11} - M_{22})^2}{M_{22} - M_{21}} + \mathcal{O}(\epsilon) \tag{11}$$

If $\epsilon \to 0$ when $t \to \infty$, then

$$\langle p \rangle \text{ can converge to } 0 \quad \textbf{iff} \quad (M_{11} < M_{22}) \textbf{or} (M_{12} < M_{22}) \tag{12}$$

$$\langle p \rangle \text{ can converge to } 1 \quad \textbf{iff} \quad (M_{22} < M_{11}) \textbf{or} (M_{21} < M_{11}) \tag{13}$$