$$V_\pi(s) = E_\pi[r_t \mid s_t = s] + \gamma \sum_a \pi(a|s) \sum_{s' \in S} p(s'|s,a) V_\pi(s')$$

במקרה שלנו: $S = \{H, O\}$, $A = \{stay, switch\}$

$\pi(a|s) = \frac{1}{2} \quad \forall a \in A, s \in S$ (כלומר, הסוכן בוחר לפעולות בהסתברות שווה)

$$V^\pi(H) = E_\pi(r \mid s=H) + \gamma\Big[\pi(stay|H)\big[P(H|H,stay)V^\pi(H) + P(O|H,stay)V^\pi(O)\big] + \pi(switch|H)\big[P(H|H,switch)V^\pi(H) + P(O|H,switch)V_\pi(O)\big]\Big]$$

נבנה נגזר:

$$\frac{0+1}{2} + \frac{1}{2}\cdot\Big[\frac{1}{2}\big[1\cdot V^\pi(H) + 0\cdot V^\pi(O)\big] +$$

$$\frac{1}{2}\big[0.2\cdot V^\pi(H) + 0.8 V^\pi(O)\big] =$$

$$\frac{1}{2} + \frac{1}{4}\big(1.2 V^\pi(H) + 0.8 V^\pi(O)\big) = \frac{1}{2} + 0.3 V^\pi(H) + 0.2 V^\pi(O) =$$

$$0.7 V^\pi(H) = 0.2 V^\pi(O) + \frac{1}{2}$$

$$\Downarrow$$

$$V^\pi(H) = \frac{2}{7} V^\pi(O) + \frac{5}{7}$$

$$V^\pi(O) = \frac{1}{3}\left(\frac{2}{7} V^\pi(O) + \frac{5}{7}\right) + \frac{4}{3} = 3V^\pi(O) = \frac{2}{7} V^\pi(O) + \frac{5}{7} + 4$$

$$\Downarrow$$

$$\left(2 + \frac{5}{7}\right) V^\pi(O) = \frac{5}{7} + 4 \Rightarrow \underline{V^\pi(O) = 1.736}$$

$$\Downarrow$$

$$V^\pi(H) = \frac{2}{7}\cdot 1.736 + \frac{5}{7} = 1.21$$

$$\Downarrow$$

$$V^\pi(S) = \begin{cases} 1.21 & S = H \\ 1.736 & S = O \end{cases}$$

(2) נתן דוגמא לפי הנחות כי המדיניות האופטימלית היא
למקסם תמיד כאשר נמצאים ב- H ולהשאר תמיד כשנמצאים
ב- O בגלל שאין הסתברות במעקטמות את המצב.

$$\pi^*(a|s) = \begin{cases} 1 & (s=H,\ a=switch) \text{ or } (s=O,\ a=stay) \\ 0 & \text{else} \end{cases}$$

(3) נבדוק האם המדיניות π שהגדרנו למעלה מקיימת את משוואת
האופטימליות.

$$V^{\pi^*}(s) = \max_{a' \in A} \left\{ E[r|s,a'] + \gamma \sum_{s' \in S} P(s'|s,a') V^{\pi^*}(s') \right\}$$

$$V^{\pi}(H) = \max \{ r(H,stay) + \gamma [P(H|H,stay) V^{\pi}(H) + P(O|H,stay) V^{\pi}(O)],\ r(H,switch) + \gamma [P(H|H,switch) V^{\pi}(H) + P(O|H,switch) V^{\pi}(O)] \}$$

$$V^{\pi}(O) = \max \{ r(O,stay) + \gamma (P(H|O,stay) V^{\pi}(H) + P(O|O,stay) \cdot V^{\pi}(O)],\ r(O,switch) + \gamma [P(H|O,switch) V^{\pi}(H) + P(O|O,switch) \cdot V^{\pi}(O)] \}$$

כדאי לשים לב בכל האופן לא איטרציות כאלה, [redacted]
נבדוק איך $V^{\pi}(s)$ מגיע לא שינוי.

$$V^{\pi}(H) = 1 + \frac{1}{2}\left[0\cdot\left[P(H|H,stay)V^{\pi}(H) + P(O|H,stay)V^{\pi}(O)\right] + 1\cdot\left[0.2V^{\pi}(H) + 0.8V^{\pi}(O)\right]\right] = 1 + \frac{1}{2}(0.2V^{\pi}(H) + 0.8V^{\pi}(O))$$

$$V^{\pi}(O) = 2 + \frac{1}{2}\cdot\left[1\left[0V^{\pi}(H) + 1\cdot V^{\pi}(O)\right] + 0\cdot\left[P(H|O,switch)\cdot V^{\pi}(H) + P(O|O,switch)V^{\pi}(O)\right]\right] = 2 + \frac{1}{2}(V^{\pi}(O))$$

$$\Downarrow$$

$$1.3V^{\pi}(O) = 2 \Rightarrow V^{\pi}(O) = \frac{4}{3}.$$

$$V^{\pi}(H) = 1 + \frac{1}{2}(0.2V^{\pi}(H) + 0.8\cdot\frac{4}{3}) = 1 + 0.1V^{\pi}H + 1.066$$

$$\Downarrow$$

$$0.9V^{\pi}(H) = 2.066$$

$$\Downarrow$$

$$V^{\pi}(H) = 2.96$$

כדי נגיב בשני [redacted] שאלות אשמלינן:

$$V^{\pi}(H) = \max\left\{0 + \frac{1}{2}(1\cdot 2.96),\ 1 + \frac{1}{2}(0.2\cdot 2.96 + 0.8\cdot\frac{4}{3})\right\} =$$

$$\max\{1.48, 1.83\} = 1.83 \Rightarrow \arg\max_{a} = switch$$

$$V^{\pi}(O) = \max\left\{2 + \frac{1}{2}(0 + 1\cdot\frac{4}{3}),\ 0 + \frac{1}{2}(1\cdot 2.96) + 0\right\} =$$

$$\max\{2.66, 1.48\} = 2.66 \Rightarrow \arg\max_{a} = stay$$