1.

a)

we know that softmax can be denoted as:

$$k = \frac{e^{Qt(a)/\mathrm{t}}}{\sum_{i=1}^{n} e^{Qt(i)/t}}$$

we can get the

$$\ln\mathrm{K} = \frac{\mathrm{Qt(a)}}{\mathrm{t}} - ln^{\left(e^{\frac{Qt(a)}{t}}\right)*(1+\sum\frac{e^{Qt(i)/t}}{e^{Qt(a)/t}})}$$

when t => 0, $e^{Qt(i)/t} \ll e^{\frac{Qt(a)}{t}}$. So $\frac{e^{Qt(i)/t}}{e^{Qt(a)/t}} => 0, \sum\frac{e^{Qt(i)/t}}{e^{Qt(a)/t}} => 0$

So $\ln\mathrm{K}$ => $\frac{\mathrm{Qt(a)}}{\mathrm{t}}$ $-ln^{\left(e^{\frac{Qt(a)}{t}}\right)} =$ $\frac{\mathrm{Qt(a)}}{\mathrm{t}} - \frac{\mathrm{Qt(a)}}{\mathrm{t}}$ = 0 ;

So k => 1. Therefore, when t -> 0, it becomes the same as greedy action selection, always choose the biggest one.

b)

we know that softmax can be denoted as:

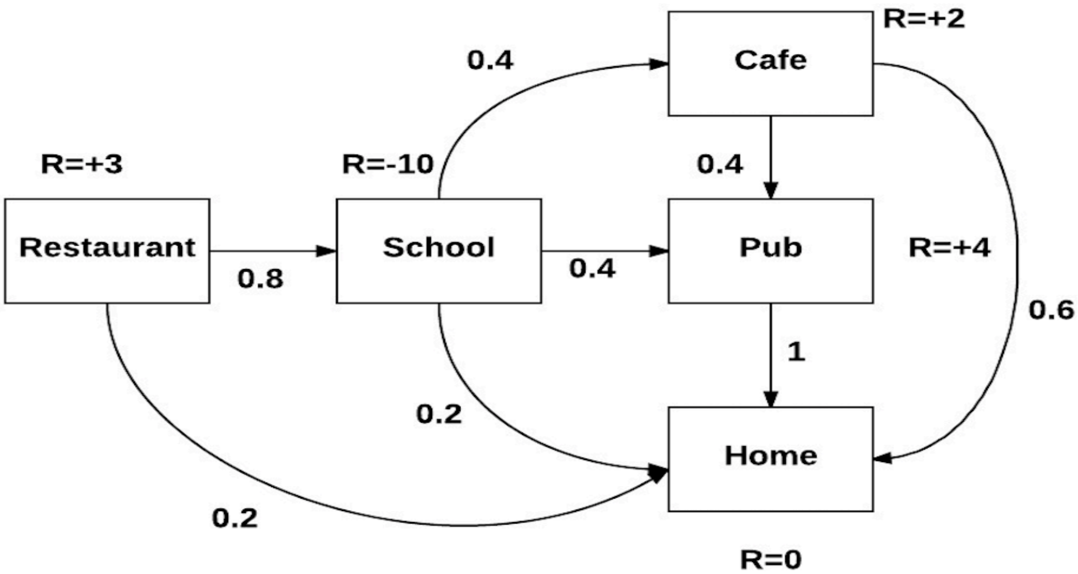
$$k = \frac{e^{Qt(a)/\mathrm{t}}}{\sum_{i=1}^{n} e^{Qt(i)/t}}$$

So for two actions, the probability of either one can be denoted as:

$$\mathrm{Pi} = \frac{e^{\frac{Qi}{t}}}{e^{\frac{Q1}{t}}+e^{\frac{Q2}{t}}} = \frac{1}{1+e^{-(Q1-Q2)/t}}$$

So it becomes a Logistic function.

. 2. MRP



let’s take Restaurant :0, School :1, Café: 2, Pub: 3, Home:4 .
Transition Probability matrix is P, the Reward function is R.

$$P = \begin{bmatrix} 0.8 \quad\quad 0.2 \\ 0.40.40.2 \\ 0.40.6 \\ 1 \\ 0 \end{bmatrix} \qquad\qquad R = \begin{bmatrix} 3 \\ -10 \\ 2 \\ 4 \\ 0 \end{bmatrix}$$

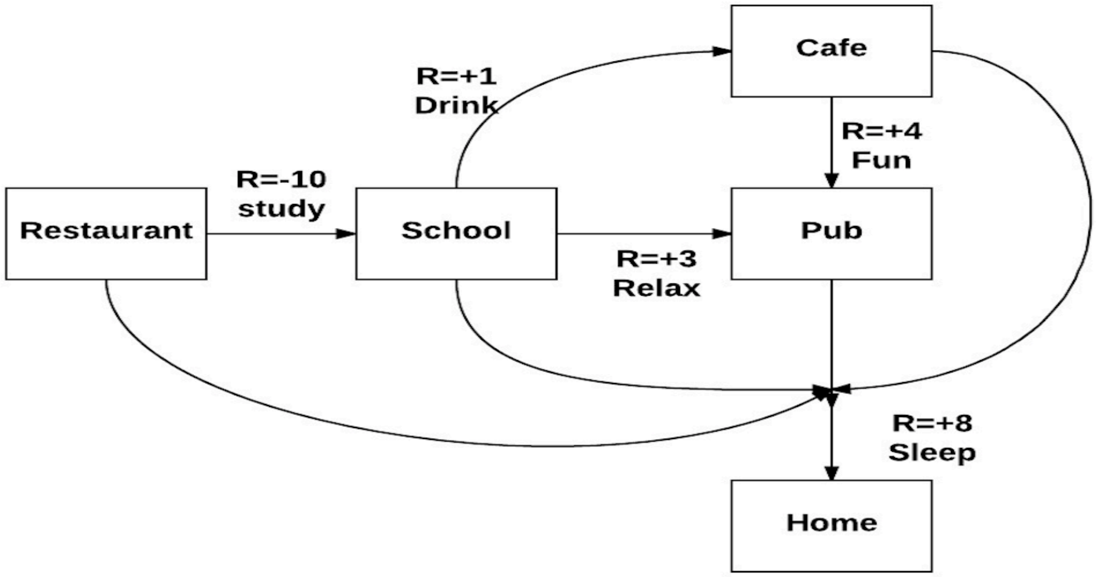
According to Bellman Equation:

$$V = R + rPV$$

$$V = (I - rP)^{-1}R$$

$$(I\text{-}rP)^{-1} = \begin{bmatrix} \multicolumn{5}{l}{10.80.320.4481} \\ 0 & 1 & 0.4 & 0.56 & 1 \\ 0 & 0 & 1 & 0.4 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad\qquad V = \begin{bmatrix} 3 \\ -10 \\ 2 \\ 4 \\ 0 \end{bmatrix}$$

3.MDP



For our MDP, the r = 1, and has a stochastic policy. π(a|s) is the same.
Transition Probability matrix is P, the Reward function is R.
let's take Restaurant :0, School :1, Café: 2, Pub: 3, Home:4 .
Follow the rule:

$$\mathcal{R}_s^\pi = \sum_{a \in \mathcal{A}} \pi(a|s)\mathcal{R}_s^a$$

$$P^\pi = \begin{bmatrix} 0.5 \quad 0.5 \\ 0.330.330.33 \\ 0.5 \;\; 0.5 \\ 1 \\ 0 \end{bmatrix} \qquad R^\pi = \begin{bmatrix} -1 \\ 4 \\ 6 \\ 8 \\ 0 \end{bmatrix}$$

According to Bellman Equation:

$$v_\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi$$

$$(\mathrm{I}\text{-}\mathrm{r}P^\pi)^{-1} = \begin{bmatrix} 10.50.16650.249750.995 \\ 0 & 1 & 0.333 & 0.4995 & 0.999 \\ 0 & 0 & 1 & 0.5 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad V^\pi = \begin{bmatrix} 3.997 \\ 9.994 \\ 10 \\ 8 \\ 0 \end{bmatrix}$$

4. Derive the weighted-average update rule (5.5) from (5.4). Follow the pattern of the derivation of the unweighted rule (2.3).

Let $C_{n+1} = C_n + W_{n+1}$, where $C_0 = 0$;

$$V_{n+1} = \frac{\sum_{k=1}^{n} Wk*Gk}{\sum_{k=1}^{n} Wk} = \frac{Wn*Gn+Vn*(Cn-Wn)}{Cn} = Vn + \frac{Wn*Gn-Vn*Wn}{Cn}$$

$$= Vn + \frac{Wn}{Cn}(Gn - Vn)$$