Susan's version of Auer's UCB

To reduce notational burden!

$T = 2$, $|\mathcal{A}| = 2$, $|\mathcal{S}^1| = 2$ $\quad \mathcal{S}^1 = \{0, 1\}$, $\mathcal{A} = \{0, 1\}$

M is number of episodes.

From the prior $m-1$ episodes we will form $\pi^m = \begin{pmatrix} \pi_0^m \\ \pi_1^m \end{pmatrix}$ e.g. π^m is a function of

$$\mathcal{H}_{m-1} = \left\{ S_{i0}\, A_{i0}\, S_{i1}\, A_{i1} \ldots A_{iT-1}, S_{iT} \quad i < m \right\}$$

UCB Algorithm (here $T = 2$)

for $j = 1, \ldots, M$ episodes.

learner sees $S_{j0} = s_0$

for $t = 0, \ldots, T-1$

learner selects A_{jt} using π^j

learner receives S_{jt+1}

$R_{jt} = R(S_{jt}, A_{jt}, S_{jt+1})$

end for

form π^{j+1} from j prior episodes (UCB part)

end for

- Prior to $j = m_0$, episodes are run to ensure that we see at least 1 (state, action) pair for each possible (state, action) $\in \mathcal{S} \times \mathcal{A}$ (maybe select actions uniformly in $\mathcal{A}$)
We ignore this in following.

How do we form π_j from $\mathcal{H}_{j-1}$?

Some Statistics:

For $t = 0, 1$

$$N_{sa}^{t,m-1} = \sum_{j=1}^{m-1} 1\{S_{jt} = s, A_{jt} = a\}$$

where $N_{sa}^{m-1} = N_{sa}^{0,m-1} + N_{sa}^{1,m-1}$

$$\rho_{sa}^{t,m-1} = \sum_{j=1}^{m-1} 1\{S_{jt} = s, A_{jt} = a\} R_{jt}$$

where $\rho_{sa}^{m-1} = \rho_{sa}^{0,m-1} + \rho_{sa}^{1,m-1}$

$$p_{sas'}^{m-1} = \sum_{j=1}^{m-1} \sum_{t=0}^{1} 1\{S_{jt} = s, A_{jt} = a\} 1\{S_{j,t+1} = s'\}$$

For $t = 0, 1$

$$\tilde{r}_a^{t,m-1}(s) = \frac{\rho_{sa}^{t,m-1}}{N_{sa}^{t,m-1}} + \sqrt{\frac{c \log m}{N_{sa}^{t,m-1}}}$$

$$\tilde{r}_a^{m-1} = \frac{\rho_{sa}^{m-1}}{N_{sa}^{m-1}} + \sqrt{\frac{c \log m}{N_{sa}^{m-1}}}$$

(UCB version of average reward)

Form a UCB version of the transition probabilities:

$$\tilde{P}_a^{m-1}(s,s') = \frac{P_{sas'}^{m-1}}{N_{sa}^{m-1}} + \sqrt{\frac{c \log m}{N_{sa}^{m-1}}} \qquad N_{sa}^{m-1} = N_{sa}^{0,m-1} + N_{sa}^{1,m-1}$$

Note that $\tilde{P}_a^{m-1}(s,s')$ may be > 1.

(I will drop the superscript $m-1$ on $\tilde{r}_a(s)$, $\tilde{P}_a(s,s')$:

For any given π_1 we fix this as follows

$$\tilde{P}_a^{\pi_1,m-1}(s,1) = \mathbb{1}\left\{\tilde{r}^1_{\pi_1(1)}(1) \geq \tilde{r}^1_{\pi_1(0)}(0)\right\} \min\left(1, \tilde{P}_a(s,1)\right) + \mathbb{1}\left\{\tilde{r}^1_{\pi_1(1)}(1) < \tilde{r}^1_{\pi_1(0)}(0)\right\}\left(1 - \min\left(1, \tilde{P}_a(s,0)\right)\right)$$

$$\tilde{P}_a^{\pi_1,m-1}(s,0) = 1 - \tilde{P}_a^{\pi_1}(s,1)$$

$$= \mathbb{1}\left\{\tilde{r}^1_{\pi_1(1)}(1) \geq \tilde{r}^1_{\pi_1(0)}(0)\right\}\left(1 - \min\left(1, \tilde{P}_a(s,1)\right)\right) + \mathbb{1}\left\{\tilde{r}^1_{\pi_1(1)}(1) < \tilde{r}^1_{\pi_1(0)}(0)\right\} \min\left(1, \tilde{P}_a(s,0)\right)$$

After $m-1$ episodes we form

$$\pi_1^{m-1}(s) = \arg\max_a \tilde{r}^1_a(s) \qquad \text{(like UCB)}$$

$$= \arg\max_{\pi_1} \tilde{V}^{1,m-1}_{\pi}(s) \qquad \text{where } \tilde{V}^{1,m-1}_{\pi}(s) \triangleq \tilde{r}^1_{\pi_1(s)}(s)$$

$$\pi_0^{m-1}(s) = \arg\max_a \left\{\tilde{r}_a(s) + \sum_{s'} \tilde{P}_a^{\pi_1,m-1}(s,s')\, \tilde{r}^1_{\pi^{m-1}}(s')\right\}$$

$$= \arg\max_{\pi_0} \tilde{V}^{2,m-1}_{\pi_0}(s)$$

where $\tilde{V}^{2,m-1}_{\pi_0}(s) = \tilde{r}_{\pi_0(s)}(s) + \sum_{s'} \tilde{P}^{\pi_1,m-1}_{\pi_0(s)}(s,s')\, \tilde{r}^1_{\pi_1^{m-1}(s')}(s')$

So in the m^{th} episode

$$A_{m0} = \pi_0^{m-1}(s_0) \quad ; \quad A_{m1} = \pi_1^{m-1}(S_{m1}) .$$

(drop superscript of m-1 on $\tilde{V}$, $\tilde{P}^{\hat{\pi}_1}_a$)

Recall the expected regret is

$$R_M(L, D, \pi) = E\left[\sum_{j=1}^{M} V^2_{\pi^*}(s_0) - \sum_{j=1}^{M} V^2_{\hat{\pi}_j}(s_0)\right]$$

$T = 2 \quad |S^1| = 2$
$|A| = 2$

We prove

$$R_M(L, D, \pi) = O\left(\frac{\log M}{\Delta}\right)$$

where $\Delta = \min_{\pi \neq \pi^*} V^2_{\pi^*}(s_0) - V^2_{\pi}(s_0)$

$\pi = \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix}$

regret

$$= \sum_{m=1}^{M} V^2_{\pi^*}(s_0) - V^2_{\pi^{m-1}}(s_0)$$

$$= \sum_{m=1}^{M} \left(V^2_{\pi^*}(s_0) - V^2_{\pi^{m-1}}(s_0) \right) * \mathbb{1}\{A_{m0} = \pi_0^{m-1}(s_0), A_{m1} = \pi_1^{m-1}(s_{m1})\} * \mathbb{1}\{\tilde{V}^{2,m-1}_{\pi^{m-1}}(s_0) - \tilde{V}^{2,m-1}_{\pi^*}(s_0) \geq 0\}$$

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} \mathbb{1}\{A_{m0} = \pi_0(s_0), A_{m1} = \pi_1(s_{m1})\} * \mathbb{1}\{\tilde{V}^{2,m-1}_{\pi}(s_0) - \tilde{V}^{2,m-1}_{\pi^*}(s_0) \geq 0\}$$

where $\Delta_\pi = V^2_{\pi^*}(s_0) - V^2_{\pi}(s_0)$

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} \mathbb{1}\{A_{m0} = \pi_0(s_0), A_{m1} = \pi_1(s_{m1})\} * \mathbb{1}\{\tilde{V}^{2,m-1}_{\pi}(s_0) - \tilde{V}^{2,m-1}_{\pi^*}(s_0) \geq 0\} * \mathbb{1}\{N^{m-1}_{s_0 \pi_0(s_0)} \geq l, N^{1,m-1}_{\pi_1} \geq l_\pi\} + \sum_{\pi \neq \pi^*} \Delta_\pi \cdot l_\pi$$

Since

$$N^{1,m-1}_{\pi_1} = \sum_{j=1}^{m-1} \mathbb{1}\{A_{j1} = \pi_1(s_{j1})\}$$

Note

$$N_{\pi_1}^{1,m-1} = \sum_{j=1}^{m-1} 1\{A_{j1} = \pi_1(1), S_{j1} = 1\} + \sum_{j=1}^{m-1} 1\{A_{j1} = \pi_1(0), S_{j1} = 0\}$$

$$= N_{1\pi_1}^{1,m-1} + N_{0\pi_1}^{1,m-1}$$

(we will select ℓ_π later)

S_0

regret

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} 1\{\tilde{V}_{\pi}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0\} * 1\{N_{s_0 \pi_0(s_0)}^{m-1} \geq \ell_\pi, N_{\pi_1}^{1,m-1} \geq \ell_\pi\}$$

$$+ \sum_{\pi \neq \pi^*} \Delta_\pi \ell_\pi$$

$$1\{\tilde{V}_{\pi}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0\} =$$

$$1\left\{\left(\tilde{V}_{\pi}^{2,1}(s_0) - V_{\pi}^{2}(s_0)\right) - \left(\tilde{V}_{\pi^*}^{2,1}(s_0) - V_{\pi^*}^{2}(s_0)\right) + \left(V_{\pi}^{2}(s_0) - V_{\pi^*}^{2}(s_0)\right) \geq 0\right\}$$

Recall $\tilde{V}^{2,m-1}_{\pi}(s_0) = \tilde{r}^{(s_0)}_{\pi_0(s_0)} + \sum_{s'} \tilde{p}^{\pi_1}_{\pi_0(s_0)}(s_0,s') \hat{r}^{1}_{\pi_1(s')}(s')$

$V^{2}_{\pi}(s_0) = r^{(s_0)}_{\pi_0(s_0)} + \sum_{s'} p^{(s_0,s')}_{\pi_0(s_0)} r^{(s')}_{\pi_1(s')}$

(I am often leaving off the superscript of $m-1$ as in $\tilde{r}^{m-1}_{\pi_1(s')}(s')$ or $\tilde{p}^{\pi_1,m-1}_{\pi_0(s_0)}(s_0,s')$)

So $\tilde{V}^{2,m-1}_{\pi}(s_0) - V^{2}_{\pi}(s_0) = \tilde{r}^{(s_0)}_{\pi_0(s_0)} - r^{(s_0)}_{\pi_0(s_0)}$
$+ \sum_{s'} p_{\pi_0(s_0)}(s_0,s') \left(\hat{r}^{1}_{\pi_1(s')}(s') - r^{(s')}_{\pi_1(s')} \right)$
$+ \sum_{s} \left(\tilde{p}^{\pi_1}_{\pi_0(s_0)}(s_0,s') - p^{(s_0,s')}_{\pi_0(s_0)} \right) \hat{r}^{1}_{\pi_1(s')}(s')$

Because we only have 2 states

$\sum_{s'} \left(\tilde{p}_a^{\pi_1}(s_0,s') - p_a(s_0,s') \right) \hat{r}^{1}_{\pi_1(s')}(s')$

$= \left(\tilde{p}_a^{\pi_1}(s_0,1) - p_a(s_0,1) \right) \left(\hat{r}^{1}_{\pi_1(1)}(1) - \hat{r}^{1}_{\pi_1(0)}(0) \right)$

$= \left(\min\left(1, \tilde{p}_a(s_0,1)\right) - p_a(s_0,1) \right) \left(\hat{r}^{1}_{\pi_1(1)}(1) - \hat{r}^{1}_{\pi_1(0)}(0) \right)^{+}$
$+ \left(\min\left(1, \tilde{p}_a(s_0,0)\right) - p_a(s_0,0) \right) \left(\hat{r}^{1}_{\pi_1(0)}(0) - \hat{r}^{1}_{\pi_1(1)}(1) \right)^{+}$

(Recall that $\tilde{r}^{1}_{a}(s) = \frac{\rho^{1,m-1}_{sa}}{N^{1,m-1}_{sa}} + \sqrt{\frac{c \log(tm)}{N^{1,m-1}_{sa}}}$)

On the set $\tilde{p}_{\pi_0(s_0)}(s_0,1) \le 1 \cap \tilde{p}_{\pi_0(s_0)}(s_0,0) \le 1$

we have

$\tilde{V}^{2,m-1}_{\pi}(s_0) - V^{2}_{\pi}(s_0) = \frac{\rho^{m-1}_{s_0,\pi_0(s_0)}}{N_{s_0\pi_0}} - r^{(s_0)}_{\pi_0(s_0)} - \sqrt{\frac{c \log m}{N_{s_0\pi_0}}}$
$+ \sum_{s} p^{(s_0,s)}_{\pi_0(s_0)} \left(\frac{\rho^{m-1}_{s_1,\pi_1(s_1)}}{N_{s_1\pi_1(s_1)}} - r^{(s)}_{\pi_1(s)} \right)$
$+ \left(\frac{\tilde{p}_{s_0\pi_0(s_0)1}}{N_{s_0\pi_0}} - p^{(s_0,1)}_{\pi_0(s_0)} - \sqrt{\frac{c \log m}{N_{s_0\pi_0}}} \right) * \left(\hat{r}^{1}_{\pi_1(1)}(1) - \hat{r}^{1}_{\pi_1(0)}(0) \right)^{+}$
$+ \left(\frac{\tilde{p}_{s_0\pi_0(s_0)0}}{N_{s_0\pi_0}} - p^{(s_0,0)}_{\pi_0(s_0)} - \sqrt{\frac{c \log m}{N_{s_0\pi_0}}} \right) * \left(\hat{r}^{1}_{\pi_1(0)}(0) - \hat{r}^{1}_{\pi_1(1)}(1) \right)^{+}$
$+ 2\sqrt{\frac{c \log m}{N_{s_0\pi_0}}} + 2\sqrt{\frac{c \log m}{N_{s_0\pi_0}}} \left| \hat{r}^{1}_{\pi_1(0)}(0) - \hat{r}^{1}_{\pi_1(1)}(1) \right|$

$$\geq \frac{\rho^{m-1}_{s_0,\pi_0(s_0)}}{N_{s_0\pi_0}} - r_{\pi_0(s_0)}(s_0) - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}}$$

$$+ \sum_s p^{(s_0,s)}_{\pi_0(s_0)} \left(\frac{\rho^{m-1}_{s_1,\pi_1(s_1)}}{N_{s_1\pi_1(s_1)}} - r^{(s)}_{\pi_1(s)} - \sqrt{\frac{c\log m}{N_{s,\pi_1}}} \right)^+$$

$$+ \left(\frac{\widetilde{P}_{s_0\pi_0(s_0)1}}{N_{s_0\pi_0}} - p^{(s_0,1)}_{\pi_0(s_0)} - \sqrt{\frac{c\log m}{N_{s\pi_0}}} \right) * \left(\hat{r}^{(1)}_{\pi_1(1)} - \hat{r}^{(0)}_{\pi_1(0)} \right)^+$$

$$+ \left(\frac{\widetilde{P}_{s_0\pi_0(s_0)0}}{N_{s_0\pi_0}} - p^{(s_0,0)}_{\pi_0(s_0)} - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \right) * \left(\hat{r}^{(0)}_{\pi_1(0)} - \hat{r}^{(1)}_{\pi_1(1)} \right)^+$$

$$+ 2\sqrt{\frac{c\log m}{N_{s_0\pi_0}}} + 2\sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \left| \hat{r}'^{(0)}_{\pi_1(0)} - \hat{r}'^{(1)}_{\pi_1(1)} \right|$$

$$+ \sum_s p^{(s_0,s)}_{\pi_0(s_0)} \sqrt{\frac{c\log m}{N_{s\pi_1}}} \mathbb{1}\left\{ \frac{\rho^{m-1}_{s,\pi_1(s)}}{N_{s\pi_1}} - r_{\pi_1}(s) > \sqrt{\frac{c\log m}{N_{s\pi_1}}} \right\}$$

In the following we ignore that fact that the above inequality was only shown to hold under $\widetilde{P}_{\pi_0(s_0)}(s_0,1) \leq 1 \wedge \widetilde{P}_{\pi_0(s_0)}(s_0,0) \leq 1$ and we will assume that $|\widetilde{r}'_a(s)| \leq 1$ as well. This can be fixed albeit with more technical, drawn out derivations!

Recall

$$\mathbb{1}\left\{ \widetilde{V}^{z_1,m-1}_{\pi}(s_0) - \widetilde{V}^{x_1,m-1}_{\pi^*}(s_0) \geq 0 \right\} =$$

$$\mathbb{1}\left\{ \left(\widetilde{V}^{z_1}_{\pi}(s_0) - V^2_{\pi}(s_0) \right) - \left(\widetilde{V}^{z_1}_{\pi^*}(s_0) - V^2_{\pi^*}(s_0) \right) + \left(V^2_{\pi}(s_0) - V^2_{\pi^*}(s_0) \right) \geq 0 \right\}$$

Using $\mathbb{1}\{A+B+C \geq 0\} \leq \mathbb{1}\{A \geq 0\} + \mathbb{1}\{B \geq 0\} + \mathbb{1}\{C \geq 0\}$:

$$\mathbb{1}\left\{ \tilde{V}^{2,m-1}_{\pi,1}(s_0) - \tilde{V}^{2,m-1}_{\pi^*}(s_0) \geq 0 \right\}$$

$$\leq \mathbb{1}\Bigg\{ \frac{\rho^{m-1}_{s_0,\pi_0(s_0)}}{N_{s_0\pi_0}} - r_{\pi_0(s_0)}(s_0) - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} + \sum_s p_{\pi_0(s_0)}(s_0,s) \left(\frac{\rho^{m-1}_{s,\pi_1(s_1)}}{N_{s_1\pi_1(s_1)}} - r_{\pi_1(s)}(s) - \sqrt{\frac{c\log m}{N_{s,\pi_1}}} \right)^+$$

$$+ \left(\frac{\tilde{P}_{s_0\pi_0(s_0)1}}{N_{s_0\pi_0}} - p_{\pi_0(s_0)}(s_0,1) - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \right) * \left(\hat{r}^{1}_{\pi_1(1)}(1) - \hat{r}^{1}_{\pi_1(0)}(0) \right)^+$$

$$+ \left(\frac{\tilde{P}_{s_0\pi_0(s_0)0}}{N_{s_0\pi_0}} - p_{\pi_0(s_0)}(s_0,0) - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \right) * \left(\hat{r}^{1}_{\pi_1(0)}(0) - \hat{r}^{1}_{\pi_1(1)}(1) \right)^+ \geq 0 \Bigg\}$$

$$+ \mathbb{1}\left\{ -\left(\tilde{V}^{2,1}_{\pi^*}(s_0) - V^{2}_{\pi^*}(s_0) \right) \geq 0 \right\}$$

$$+ \mathbb{1}\Bigg\{ V^{2}_{\pi}(s_0) - V^{2}_{\pi^*}(s_0) + 2\sqrt{\frac{c\log m}{N_{s_0\pi_0}}} + 2\sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \left| \hat{r}^{1}_{\pi_1(0)}(0) - \hat{r}^{1}_{\pi_1(1)}(1) \right| + \Bigg\}$$

$$\left\{ \sum_s p_{\pi_0(s_0)}^{(s_0,s)} \sqrt{\frac{c\log m}{N_{s\pi_1}}} \mathbb{1}\left\{ \frac{\rho_{s,\pi_1(s)}^{m-1}}{N_{s\pi_1}} - r_{\pi_1}(s) > \sqrt{\frac{c\log m}{N_{s\pi_1}}} \right\} \right\} \geq 0$$

Recall the regret is

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} \mathbb{1}\left\{ \tilde{V}_{\pi}^{2,m-1}(s_0) - \tilde{V}_{\pi^*}^{2,m-1}(s_0) \geq 0 \right\} * \mathbb{1}\left\{ N_{s_0 \pi_0(s_0)}^{m-1} \geq \ell_\pi, N_{\pi_1}^{m-1} \geq \ell_\pi \right\}$$

$$+ \sum_{\pi \neq \pi^*} \Delta_\pi \ell_\pi$$

Consider the contribution of the * term, pg 17 to the regret.

$$\leq \mathbb{1}\left\{ 0 \leq -\Delta_\pi + 6\sqrt{\frac{c\log m}{N_{s\pi_0}}} + \sum_s p_{\pi_0(s_0)}^{(s_0,s)} \sqrt{\frac{c\log m}{N_{s\pi_1}}} \mathbb{1}\left\{ \frac{\rho_{s\pi_1}^{m-1}}{N_{s\pi_1}} - r_{\pi_1}(s) > \sqrt{\frac{c\log m}{N_{s\pi_1}}} \right\} \right\}$$
$$* \mathbb{1}\left\{ N_{s_0\pi_0}^{m-1} \geq \ell_{\pi_1}, N_{1\pi_1}^{m-1} + N_{0\pi_0}^{m-1} \geq \ell_\pi \right\}$$

note $N_{1\pi_1}^{m-1} + N_{0\pi_1}^{m-1} \geq \ell_\pi \Rightarrow N_{1\pi_1}^{m-1} \geq \frac{1}{2}\ell_\pi \cup N_{0\pi_1}^{m-1} \geq \ell_{\frac{1}{2}\pi}$.

$$\leq \mathbb{1}\left\{ 0 \leq -\Delta_\pi + 6\sqrt{\frac{c\log m}{\ell_\pi}} + \sum_s p_{\pi_0}^{(s_0,s)} \sqrt{\frac{c\log m}{N_{s\pi_1}}} \mathbb{1}\left\{ \frac{\rho_{s\pi_1}^{m-1}}{N_{s\pi_1}} - r_{\pi_1}(s) > \sqrt{\frac{c\log m}{N_{s\pi_1}}} \right\} \right\}$$
$$* \mathbb{1}\left\{ N_{1\pi_1}^{m-1} \geq \frac{1}{2}\ell_{\pi_1} \right\}$$

$$+ 1\Big\{ 0 \le -\Delta_\pi + 6\sqrt{\frac{C\log m}{\ell}} + \sum_s p_{\pi_0}(s_0,s)\sqrt{\frac{C\log m}{N_{s\pi_1}}}\, 1\Big\{\frac{\rho_{s\pi_1}^{m-1}}{N_{s\pi_1}} - r_{\pi_1}(s) > \sqrt{\frac{C\log m}{N_{s\pi_1}}}\Big\}\Big\}$$

$$* \; 1\{N_{0\pi_1}^{m-1} \ge \tfrac{1}{2}\ell_\pi\}$$

(Consideration of the two terms is the same – so consider only the 1st term)

$$1\Big\{ 0 \le -\Delta_\pi + 6\sqrt{\frac{C\log m}{\ell_\pi}} + \sum_s p_{\pi_0}(s_0,s)\sqrt{\frac{C\log m}{N_{s\pi_1}}}\, 1\Big\{\frac{\rho_{s\pi_1}^{m-1}}{N_{s\pi_1}} - r_{\pi_1}(s) > \sqrt{\frac{C\log m}{N_{s\pi_1}}}\Big\}\Big\}$$

$$* \; 1\{N_{1\pi_1}^{m-1} \ge \tfrac{1}{2}\ell_\pi\}$$

$$\le 1\Big\{ 0 \le -\Delta_\pi + 6\sqrt{\frac{C\log m}{\ell_\pi}} + p_{\pi_0}(s_0,1)\sqrt{\frac{C\log m}{\frac{1}{2}\ell_\pi}} + p_{\pi_0}(s_0,0)\sqrt{\frac{C\log m}{N_{0,\pi_1}}}\, 1\Big\{\frac{\rho_{0\pi_1}^{m-1}}{N_{0\pi_1}} - r_{\pi_1}(0) > \sqrt{\frac{C\log m}{N_{0\pi_1}}}\Big\}\Big\}$$

$$\# \mathbb{1}\left\{ N_{1\pi_1}^{m-1} \geq \tfrac{1}{2} l_\pi \right\}$$

$$= \mathbb{1}\left\{ 0 \leq -\Delta_\pi + 6\sqrt{\frac{c\log m}{l_\pi}} + \rho_{\pi_0(s_0,1)} \sqrt{\frac{c\log m}{l_\pi}} \right\}$$

$$+ \mathbb{1}\left\{ \frac{\rho_{0\pi_1}^{m-1}}{N_{0\pi_1}} - r_{\pi_1}(0) > \sqrt{\frac{c\log m}{N_{0\pi_1}}} \right\}$$

we will choose l_π so that 1st indicator is zero $\forall m$! e.g. $l_\pi = \frac{8^2 c \log M}{\Delta_\pi^2}$

So the regret is bounded above by

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} \mathbb{1}\left\{ \tilde{V}_{\pi,1}^{2^m-1}(s_0) - \tilde{V}_{\pi^*}^{2^m-1}(s_0) \geq 0 \right\} * \mathbb{1}\left\{ N_{s_0\pi_0(s_0)}^{m-1} \geq l_\pi, N_{\pi_1}^{m-1} \geq l_\pi \right\}$$

$$+ \sum_{\pi \neq \pi^*} \Delta_\pi \, l_\pi$$

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \, l_\pi +$$

$$\sum_{\pi \neq \pi^*} \Delta_\pi \cdot \sum_{m=1}^{M} \mathbb{1}\left\{ \frac{\rho_{s_0,\pi_0(s_0)}^{m-1}}{N_{s_0\pi_0}} - r_{\pi_0(s_0)}^{(s_0)} - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \geq 0 \right\}$$

$$+ \mathbb{1}\left\{ \frac{\rho_{s_1,\pi_1(s_1)}^{m-1}}{N_{s_1\pi_1(s_1)}} - r_{\pi_1(s_1)}^{(s_1)} - \sqrt{\frac{c\log m}{N_{s_1\pi_1}}} \geq 0 \right\}$$

$$+ 1\left\{\frac{\widetilde{P}^{m-1}_{s_0\pi_0(s_0),1}}{N_{s_0\pi_0}} - P^{(s_0,1)}_{\pi_0(s_0)} - \sqrt{\frac{c\log m}{N_{s\pi}}} \geq 0\right\}$$

$$+ 1\left\{\frac{\widetilde{P}^{m-1}_{s_0\pi_0(s_0),0}}{N_{s_0\pi_0}} - P^{(s_0,0)}_{\pi_0(s_0)} - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \geq 0\right\}$$

$$+ 1\left\{-\left(\widetilde{V}^{2,1}_{\pi^*}(s_0) - V^{2}_{\pi^*}(s_0)\right) \geq 0\right\}$$

$$+ 1\left\{\frac{P^{m-1}_{0\pi_1}}{N_{0\pi_1}} - r_{\pi_1}(0) - \sqrt{\frac{c\log m}{N_{0\pi_1}}} \geq 0\right\}$$

$$+ 1\left\{\frac{P^{m-1}_{1\pi_1}}{N_{1\pi_1}} - r_{\pi_1}(1) - \sqrt{\frac{c\log m}{N_{1\pi_1}}} \geq 0\right\}$$

The expectation of each term is treated the same: for example, consider

$$P\left\{\frac{P^{m-1}_{s_0,\pi_0(s_0),s}}{N_{s_0\pi_0}} - P^{(s_0,s)}_{\pi_0(s_0)} - \sqrt{\frac{c\log m}{N_{s_0\pi_0}}} \geq 0\right\}$$

$$\leq P\left[\exists\, n,\ 1 \leq n \leq m-1 \text{ for which } \frac{P^{n-1}_{s_0,\pi_0(s_0),s}}{n} - P^{(s_0,s)}_{\pi_0(s_0)} - \sqrt{\frac{c\log m}{n}} \geq 0\right]$$

$$\leq$$

$$\sum_{n=1}^{m} P\left[\frac{P^{n-1}_{s_0 \pi_0(s_0) s}}{n} - P_{\pi_0(s_0)}(s_0, s) - \sqrt{\frac{c \log m}{n}} > 0 \right]$$

$$\leq \sum_{n=1}^{m} P\left[\frac{1}{n} \sum_{j=1}^{n} 1\{A_{j0} = \pi_0(s_0), S_{j0} = s_0\} \left(1\{S_{j1} = s\} - P_{\pi_0(s_0)}(s_0, s) \right) > \sqrt{\frac{c \log m}{n}} \right]$$

now apply Hoeffding - Azuma (for non-independent terms)

note that

$$E\left[1\{A_{j0} = \pi_0(s_0), S_{j0} = s_0\} \left(1\{S_{j1} = s\} - P_{\pi_0(s_0)}(s_0, s) \right) \middle| \mathcal{H} \right] = 0$$

to see that

$$\leq \exp\left\{ -2n \frac{c \log m}{n} \right\} = \exp\{-2c \log m\}$$

$$= m^{-2c}$$

$$\sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} \sum_{n=1}^{m} m^{-2c}$$

$$= \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} m^{-2(c-1)}$$

sums finitely for $c \geq 1$.

This the expected regret is bounded above by

$$\Delta_\pi = V^2_{\pi^*}(s_0) - V^2_\pi(s_0)$$

$$E\left[\sum_{m=1}^{M} V^2_{\pi^*}(s_0) - V^2_{\pi^{m-1}}(s_0)\right]$$

$$\leq \sum_{\pi \neq \pi^*} \Delta_\pi \frac{8^2 c \log M}{\Delta_\pi^2}$$

$$+ \text{some finite integer} * \sum_{\pi \neq \pi^*} \Delta_\pi \sum_{m=1}^{M} m^{-2(c-1)}$$

$$= O\left(\frac{\log M}{\min_{\pi \neq \pi^*} \Delta_\pi} |A|^{|S|}\right)$$