A Proof of Theorem 1

Proof. In order to prove Thm. 1, we first need the following proposition about the confidence intervals used in computing the optimistic reward $\widetilde{r}(s,a)$.

Proposition 2. *Let assume $\|\theta_a^*\|_2 \le B$. If $\widehat{\theta}_{t,a}$ is computed as in Eq. 4 and $c_{t,a}$ is defined as in Eq. 6, then*

$$\mathbb{P}\left(r(s,a) \le \widehat{r}(s,a) + c_{t,a}\|x_{s,a}\|_{V_{t,a}^{-1}}\right) \le \frac{t^{-\alpha}}{K}.$$

Proof. By definition of $\rho(s,a)$ we have $0 \le \rho(s,a) \le \sum_{\tau=1}^{w} \frac{1}{\tau} < \log(w) + 1 \doteq \ell_w$. Thus $1 \le \|x_{s,a}\|_2^2 \le \sum_{j=0}^{d} \ell_w^j = \frac{1-\ell_w^{d+1}}{1-\ell_w} = L_w^2$. Using Thm. 2 of [1], we have with probability $1-\delta$,

$$\|\widehat{\theta}_{t,a} - \theta_a^*\|_{V_{t,a}} \le R\sqrt{(d+1)\log\left(\frac{1+T_{t,a}L_w^2/\lambda}{\delta}\right)} + \lambda^{1/2}B.$$

Thus for all $s \in S$ we have,

$$|r(s,a) - \widehat{r}(s,a)| = |x_{s,a}^{\mathsf{T}}\widehat{\theta}_{t,a} - x_{s,a}^{\mathsf{T}}\theta_a^*| \le \|x_{s,a}\|_{V_{t,a}^{-1}}\|\widehat{\theta}_a - \theta_a^*\|_{V_{t,a}}.$$

Using $\delta = \frac{t^{-\alpha}}{K}$ concludes the proof. □

An immediate result of Prop. 2 is that the estimated average reward of $\widetilde{\pi}_k$ in the optimistic MDP $\widetilde{M}_k$ is an upper-confidence bound on the optimal average reward, i.e., for any t (the probability follows by a union bound over actions)

$$\mathbb{P}\big(\eta^* > \eta^{\widetilde{\pi}_k}(\widetilde{M}_k)\big) \le t^{-\alpha}. \tag{7}$$

We are now ready to prove the main result.

Proof of Thm. 1. We follow similar steps as in [9]. We split the regret over episodes as

$$\Delta(\mathcal{A},T) = \sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\big(\eta^* - r(s_t,a_t)\big) = \sum_{k=1}^{m}\Delta_k.$$

Let $\mathcal{T}_{k,a} = \{t_k \le t < t_{k+1} : a_t = a\}$ be the steps when action a is selected during episode k. We upper bound the per-episode regret as

$$\Delta_k = \sum_{a\in[K]}\sum_{t\in\mathcal{T}_{k,a}}\big(\eta^* - r(s_t,a)\big) \le \sum_{t=t_k}^{t_{k+1}-1}\big(\widetilde{\eta}_k - \widetilde{r}_k(s_t,a)\big) + \sum_{a\in[K]}\sum_{t\in\mathcal{T}_{k,a}}\big(\widetilde{r}_k(s_t,a) - r(s_t,a)\big),$$

where the inequality directly follows from the event that $\widetilde{\eta}_k \ge \eta^*$ (Eq. 7) with probability $1 - T^{-\alpha}$. Notice that the low-probability event of failing confidence intervals can be treated as in [9].

We proceed by bounding the first term of Eq. 8. Unlike in the general online learning scenario, in our setting the transition function f is known and thus the regret incurred from bad estimates of the dynamics is reduced to zero. Furthermore, since we are dealing with deterministic MDPs, the optimal policy converges to a loop over states. When starting a new policy, we may start from a state outside its loop. Nonetheless, it is easy to verify that starting from any state s, it is always possible to reach any desired state s' in at most w steps (i.e., the size of the history window). As a result, within each episode k the difference between the cumulative reward ($\sum_t \widetilde{r}_k(s_t,a)$) and the (optimistic) average reward ($(t_{k+1}-t_k)\widetilde{\eta}_k$) in the loop never exceeds w. Furthermore, since episodes terminate when one action doubles its number of samples, using a similar proof as [9], we have that the number of episodes is bounded as $m \le K\log_2(\frac{8T}{K})$. As a result, the contribution of the first term of Eq. 8 to the overall regret is bounded as

$$\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\big(\widetilde{\eta}_k - \widetilde{r}_k(s_t,a)\big) \le Kw\log_2\left(\frac{8T}{K}\right). \tag{8}$$

The second term in Eq. 8 refers to the (cumulative) reward estimation error and it can be decomposed as

$$|\widetilde{r}_k(s_t,a) - r(s_t,a)| \leq |\widetilde{r}_k(s_t,a) - \widehat{r}_k(s_t,a)| + |\widehat{r}_k(s_t,a) - r(s_t,a)|.$$

We can bound the cumulative sum of the second term as (similar for the first, since $\widetilde{r}_k$ belongs to the confidence interval of $\widehat{r}_k$ by construction)

$$\begin{aligned}\sum_{k=1}^{m}\sum_{a\in[K]}\sum_{t\in\mathcal{T}_{k,a}} |\widehat{r}_k(s_t,a) - r(s_t,a)| &\leq \sum_{k=1}^{m}\sum_{a\in[K]}\sum_{t\in\mathcal{T}_{k,a}} c_{t,a}\|x_{s_t,a}\|_{V_{a,t}^{-1}} \\ &\leq c_{\max}\sum_{a\in[K]}\sqrt{\sum_{k=1}^{m}\sum_{t\in\mathcal{T}_{k,a}}\|x_{s_t,a}\|^2_{V_{a,t}^{-1}}}\sqrt{T_a},\end{aligned}$$

where the first inequality follows from Prop. 2 with probability $1-T^{-\alpha}$, and T_a is the total number of times a has been selected at step T. Let $\mathcal{T}_a = \cup_k \mathcal{T}_{k,a}$, then using Lemma 11 of [1], we have

$$\sum_{t\in\mathcal{T}_a}\|x_{s_t,a}\|^2_{V_{t,a}^{-1}} \leq 2\log\frac{\det(V_{T,a})}{\det(\lambda I)},$$

and by Lem. 10 of [1], we have

$$\det(V_{t,a}) \leq (\lambda + tL_w^2/(d+1))^{d+1},$$

which leads to

$$\begin{aligned}\sum_{k=1}^{m}\sum_{a\in[K]}\sum_{t\in\mathcal{T}_{k,a}} |\widehat{r}_k(s_t,a) - r(s_t,a)| &\leq c_{\max}\sum_{a\in[K]}\sqrt{T_a}\sqrt{2(d+1)\log\Big(\frac{\lambda+tL_w^2}{\lambda(d+1)}\Big)} \\ &\leq c_{\max}\sqrt{2KT(d+1)\log\Big(\frac{\lambda+tL_w^2}{\lambda(d+1)}\Big)}.\end{aligned}$$

Bringing all the terms together gives the regret bound. □

B Experiments Details

Genre	$\theta^*_{a,0}$	$\theta^*_{a,1}$	$\theta^*_{a,2}$	$\theta^*_{a,3}$	$\theta^*_{a,4}$	$\theta^*_{a,5}$
Action	3.1	0.54	-1.08	0.78	-0.22	0.02
Comedy	3.34	0.54	-1.08	0.78	-0.22	0.02
Adventure	3.51	0.86	-2.7	3.06	-1.46	0.24
Thriller	3.4	1.26	-2.9	2.76	-1.14	0.16
Drama	2.75	1.0	0.94	-1.86	0.94	-0.16
Children	3.52	0.1	0.0	-0.3	0.2	-0.04
Crime	3.37	0.32	1.12	-3.0	2.26	-0.54
Horror	3.54	-0.68	1.84	-2.04	0.82	-0.12
SciFi	3.3	0.64	-1.32	1.1	-0.38	0.02
Animation	3.4	1.38	-3.44	3.62	-1.62	0.24

Table 3: Reward parameters of each genre for the *movielens* experiment.

The parameters used in the MovieLens experiment are reported in Table 3.