# A PROOFS

## A.1 Proof of Theorem 3.1

*Proof.* To prove the theorem, we need to show that Algorithm 1 will take the same action as the index policy for each state $s \in \mathcal{S}$ that:

$$\arg\max_{a_i \in \mathcal{A}} q_t(s, a_i) = \arg\max_{a_i \in \mathcal{A}} \hat{z}_t(s^i, a_i) \, .$$

By subtracting a constant from the left-hand side, we get the following equivalent statement:

$$\arg\max_{a_i \in \mathcal{A}} \Big( q_t(s, a_i) - \hat{v}_{t+1}(s) \Big) = \arg\max_{a_i \in \mathcal{A}} \hat{z}_t(s^i, a_i) \, .$$

Informally, the term $q_t(s, a_i) - \hat{v}_{t+1}(s)$ measures the advantage of pulling an arm $a_i$ in comparison with a virtual action of "pulling no arm" instead. Next we show that $q_t(s, a_i) - \hat{v}_{t+1}(s) = \hat{z}_t(s_i, a_i)$.

Fix an action $a_i$, let $s_\tau$ be some state, and let $S_{\tau+1}$ to be the random variable representing the state that follows $s^i_\tau$ after taking action $a_i$. Than:

$$\begin{aligned} q_t(s_\tau, a_i) - \hat{v}_{t+1}(s_\tau) &= \mathbb{E}\left[\, \hat{v}_{t+1}(S_{\tau+1}) + r(s_\tau, a_i, S_{\tau+1}) \,\right] - \hat{v}_{t+1}(s_\tau) = \\ &= \mathbb{E}\left[\, r(s_\tau, a_i, S_{\tau+1}) \,\right] + \mathbb{E}\left[\, \hat{v}_{t+1}(S_{\tau+1}) \,\right] - \hat{v}_{t+1}(s_\tau) = \\ &= r(s^i_\tau, a_i) + \mathbb{E}\left[\, \hat{v}_{t+1}(S_{\tau+1}) \,\right] - \hat{v}_{t+1}(s_\tau) \end{aligned}$$

Using the assumed linearly separable form of $\hat{v}_t(s_\tau) = \sum_{a_i \in \mathcal{A}} \hat{v}^i_t(s^i_\tau)$, we get:

$$\begin{aligned} r(s^i_\tau, a_i) + \mathbb{E}\left[\, \hat{v}_{t+1}(S_{\tau+1}) \,\right] - \hat{v}_{t+1}(s_\tau) &= r(s^i_\tau, a_i) + \mathbb{E}\left[ \sum_{a_j \in \mathcal{A}} \hat{v}^j_{t+1}(S^j_{\tau+1}) \right] - \sum_{a_j \in \mathcal{A}} \hat{v}^j_{t+1}(s^j_\tau) = \\ &= r(s^i_\tau, a_i) + \sum_{a_j \in \mathcal{A}} \Big( \mathbb{E}\left[ \, \hat{v}^{a'}_{t+1}(S^j_{\tau+1}) \, \right] - \hat{v}^j_{t+1}(s^j_\tau) \Big) \end{aligned}$$

Notice next that $S^j_{\tau+1} = s^j_\tau$ whenever $a_j \neq a_i$, since the state of an arm does not change unless the arm is pulled. Thus we can further simplify the sum as follows:

$$\begin{aligned} & r(s^i_\tau, a_i) + \sum_{a_j \in \mathcal{A}} \Big( \mathbb{E}\left[ \, \hat{v}^j_{t+1}(S^j_{\tau+1}) \, \right] - \hat{v}^j_{t+1}(s^j_\tau) \Big) = \\ = r(s^i_\tau, a_i) + & \sum_{a_j \in \mathcal{A} \setminus \{a_i\}} \Big( \mathbb{E}\left[ \, \hat{v}^j_{t+1}(S^j_{\tau+1}) \, \right] - \hat{v}^j_{t+1}(s^j_\tau) \Big) + \mathbb{E}\left[\, \hat{v}^i_{t+1}(S^i_{\tau+1}) \,\right] - \hat{v}^i_{t+1}(s^i_\tau) = \\ = r(s^i_\tau, a_i) + & \sum_{a_j \in \mathcal{A} \setminus \{a_i\}} \Big( \mathbb{E}\left[ \, \hat{v}^j_{t+1}(s^j_\tau) \, \right] - \hat{v}^j_{t+1}(s^j_\tau) \Big) + \mathbb{E}\left[\, \hat{v}^i_{t+1}(S^i_{\tau+1}) \,\right] - \hat{v}^i_{t+1}(s^i_\tau) = \\ & = r(s^i_\tau, a_i) + \mathbb{E}\left[\, \hat{v}^i_{t+1}(S^i_{\tau+1}) \,\right] - \hat{v}^i_{t+1}(s^i_\tau) \, . \end{aligned}$$

Finally, substituting (5) into the equation above gives us:

$$q_t(s_\tau, a_i) - \hat{v}_{t+1}(s_\tau) = \hat{z}_t(s_\tau, a_i) \, ,$$

which proves the theorem. □

## A.2 Proof of Lemma 4.1

*Proof.* First, we will use the following decomposition of the regret for a given parameter $\mu$:

$$\begin{aligned}
\mathrm{Regret}(\pi, T, \mu) &= \sum_{t=1}^{T} \mathbb{E}\left[\, \mu_{i_u^\star} - \mu_{\pi(S_t)} \,\right] = \\
&\overset{\text{(a)}}{=} \sum_{t=1}^{T} \left(\mathbb{E}\left[\, \mu_{i_u^\star} - \big(q_t(S_t, \pi(S_t)) - v_{t+1}(S_t)\big) \,\right] + \mathbb{E}\left[\, \big(q_t(S_t, \pi(S_t)) - v_{t+1}(S_t)\big) - \mu_{\pi(S_t)} \,\right]\right) = \\
&\overset{\text{(b)}}{\leq} \sum_{t=1}^{T} \left(\mathbb{E}\left[\, \mu_{i_u^\star} - \big(q_t(S_t, a_{i_u^\star}) - v_{t+1}(S_t)\big) \,\right] + \mathbb{E}\left[\, \big(q_t(S_t, \pi(S_t)) - v_{t+1}(S_t)\big) - \mu_{\pi(S_t)} \,\right]\right) = \\
&= \sum_{t=1}^{T} \mathbb{E}\left[ \underbrace{\big(\mu_{i_u^\star} - q_t(S_t, a_{i_u^\star}) + v_{t+1}(S_t)\big)}_{-\varphi_t(\mu, S_t, a_{i_u^\star})} \right] + \sum_{t=1}^{T} \mathbb{E}\left[ \underbrace{\big(q_t(S_t, \pi(S_t)) - v_{t+1}(S_t) - \mu_{\pi(S_t)}\big)}_{\varphi_t(\mu, S_t, \pi(S_t))} \right] = \\
&= \sum_{t=1}^{T} \mathbb{E}_{S_t}\left[\, \varphi_t(\mu, S_t, \pi(S_t)) \,\right] - \sum_{t=1}^{T} \mathbb{E}_{S_t}\left[\, \varphi_t(\mu, S_t, a_{i_u^\star}) \,\right]
\end{aligned}$$

The equality (a) follows by simply adding $0 = v_{t+1}(S_t) - v_{t+1}(S_t)$, and the inequality (b) follows from the optimality of $\pi(S_t)$ with respect to $v_t$:

$$q_t(S_t, \pi(S_t)) \geq q_t(S_t, a),$$

for any action $a \in \mathcal{S}$. □

## A.3 Proof of Theorem 4.1

The theorem follows directly from the following two lemmas.

**Lemma A.1.** *For $\alpha > 1$ and $\varphi_t$ computed for $v^{UCB}$ the following lower bound holds true:*

$$\sum_{t=1}^{T} \mathbb{E}_{S_t}\left[\, \varphi_t(\mu, S_t, a_{i_u^\star}) \,\right] \geq -\sum_{t=1}^{T} \frac{1}{t^{2\alpha-1}} \geq O(1)\,.$$

*Proof.*

$$\begin{aligned}
\mathbb{E}_{S_t}\left[\, \varphi_t(\mu, S_t, a_{i_u^\star}) \,\right] &= \mathbb{E}_{S_t}\left[\, q_t(S_t, a_{i_u^\star}) - v_{t+1}(S_t) - \mu_{i_u^\star} \,\right] = \mathbb{E}\left[\, z^{\text{UCB}}(S_t, a_{i_u^\star}) - \mu_{i_u^\star} \,\right] = \\
&= \mathbb{E}\left[ r(S_t, a_{i_u^\star}) + \sqrt{\frac{\alpha \log t}{T_{i_u^\star}(S_t^{i_u^\star}, a_{i_u^\star})}} - \mu_{i_u^\star} \right] \geq \\
&\overset{\text{(a)}}{\geq} -\mathbb{P}\left[ r(S_t, a_{i_u^\star}) - \mu_{i_u^\star} \leq \sqrt{\frac{\alpha \log t}{T_{i_u^\star}(S_t^{i_u^\star}, a_{i_u^\star})}} \right] \geq \\
&\overset{\text{(b)}}{\geq} -t \exp\left( -2 T_{i_u^\star}(S_t^{i_u^\star}, a_{i_u^\star}) \frac{\alpha \log t}{T_{i_u^\star}(S_t^{i_u^\star}, a_{i_u^\star})} \right) = \\
&= -t \exp(-2\alpha \log t) = -\frac{1}{t^{2\alpha-1}}
\end{aligned}$$

where (a) follows by rewriting expectation and the fact that the rewards are bounded between 0 and 1, and (b) follows from Hoeffding's inequality and the union bound over possible values of $T_{i_u^\star}$. Summing the value above for $t = 1 \ldots T$ proves the lemma. □

**Lemma A.2.** *Assume $\alpha > 1$ and fix an action $a_i \in \mathcal{A}$. Then for a policy $\pi$ greedy with respect to $v^{UCB}$ for following inequality holds true:*

$$\sum_{t=1}^{T} \mathbb{E}_{S_t} \left[ \varphi_t(\mu, S_t, \pi(S_t)) \right] \leq O(\sqrt{T \log T})$$

*Proof.* Consider the bound for a single action:

$$\begin{aligned}
\mathbb{E}_{S_t} \left[ \varphi_t(\mu, S_t, a_i) \right] &= \mathbb{E}_{S_t} \left[ q_t(S_t, a_i) - v_{t+1}(S_t) - \mu_i \right] = \mathbb{E} \left[ z^{\text{UCB}}(S_t, a_i) - \mu_{i_u^\star} \right] \\
&= \mathbb{E} \left[ r(S_t, a_i) + \sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} - \mu_i \right] = \\
&= \mathbb{E} \left[ r(S_t, a_i) - \sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} + 2\sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} - \mu_i \right] = \\
&= \mathbb{E} \left[ r(S_t, a_i) - \sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} - \mu_i \right] + \mathbb{E} \left[ 2\sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} \right] = \\
&\overset{\text{(a)}}{\leq} \mathbb{P} \left[ r(S_t, a_i) - \mu_i \leq \sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} \right] + \mathbb{E} \left[ 2\sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} \right] \\
&\overset{\text{(b)}}{\leq} \frac{1}{t^{2\alpha - 1}} + \mathbb{E} \left[ 2\sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} \right]
\end{aligned}$$

The equalities above follow readily using algebra; (a) and (b) follow by an argument that is identical to the proof of Lemma A.1.

Considering just the state in which arm $a_i$ is pulled, it remains to upper bound the following expression:

$$\begin{aligned}
\sum_{t=1}^{T} 2\,\mathbb{E} \left[ \mathbf{1}_{a_i = \pi(S_t)} \sqrt{\frac{\alpha \log t}{T_i(S_t^i, a_i)}} \right] &\leq \sum_{t=1}^{T} 2\,\mathbb{E} \left[ \mathbf{1}_{a_i = \pi(S_t)} \sqrt{\frac{\alpha \log T}{T_i(S_t^i, a_i)}} \right] \leq \\
&\overset{\text{(d)}}{\leq} \sum_{t=1}^{T} 2\,\mathbb{E} \left[ \sqrt{\frac{\alpha \log T}{t}} \right] \overset{\text{(e)}}{\leq} O(\sqrt{T \log T})\,.
\end{aligned}$$

The inequality (d) follows by upper bounding the error by assuming that the arm was pulled in every step, and (e) follows by upper bounding the sum by an integral. See, for example, the proof of Proposition 2 in Russo and Van Roy [2014]. □