A Proof of Proposition 1

As shorthand, we let $d(a,a^*) = D(\mathbb{P}(Y_{t,a} \in \cdot|\mathcal{F}_t, A^* = a^*)||\mathbb{P}(Y_{t,a} \in \cdot|\mathcal{F}_t))$ and $x(a) = \sqrt{d(a,a)}$. By the definition of the instantaneous regret, we have that

$$\begin{aligned}
\boldsymbol{\Delta}_t^T \boldsymbol{\alpha}_t &= \sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)\mathbb{E}[Y_{t,A^*} - Y_{t,a}|\mathcal{F}_t] && (13)\\
&\overset{(a)}{=} \sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)\left(\mathbb{E}[Y_{t,A^*}|\mathcal{F}_t] - \mathbb{E}[Y_{t,a}|\mathcal{F}_t]\right) && (14)\\
&= \mathbb{E}[Y_{t,A^*}|\mathcal{F}_t] - \sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)\mathbb{E}[Y_{t,a}|\mathcal{F}_t] && (15)\\
&\overset{(b)}{=} \sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)\left(\mathbb{E}[Y_{t,a}|\mathcal{F}_t, A^* = a] - \mathbb{E}[Y_{t,a}|\mathcal{F}_t]\right)\\
&\overset{(c)}{\leq} \sqrt{\frac{1}{2}} \sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)\sqrt{d(a,a)} && (16)\\
&= \sqrt{\frac{1}{2}} \sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)x(a), && (17)
\end{aligned}$$

where (a) follows from the linearity of expectation, (b) uses the law of total probability, (c) follows from the Pinsker's inequality.

By the definition of the information gain of observing an action, we have that

$$\begin{aligned}
\boldsymbol{g}_t^T \boldsymbol{\alpha}_t &\overset{(d)}{\geq} (\boldsymbol{G}_t \boldsymbol{h}_t)^T \boldsymbol{\alpha}_t\\
&= \sum_{a\in\mathcal{K}} \left(\sum_{a':a' \xrightarrow{t} a} \boldsymbol{\alpha}_t(a') \right) I_t(A^*; Y_{t,a}) && (18)\\
&\overset{(e)}{=} \sum_{a\in\mathcal{K}} \left(\sum_{a':a' \xrightarrow{t} a} \boldsymbol{\alpha}_t(a') \right) \left(\sum_{a^*\in\mathcal{K}} \boldsymbol{\alpha}_t(a^*) d(a,a^*) \right)\\
&\overset{(f)}{\geq} \sum_{a\in\mathcal{K}} \left(\sum_{a':a' \xrightarrow{t} a} \boldsymbol{\alpha}_t(a') \right) \boldsymbol{\alpha}_t(a) d(a,a) && (19)\\
&= \sum_{a\in\mathcal{K}} \left(\sum_{a':a' \xrightarrow{t} a} \boldsymbol{\alpha}_t(a') \right) \boldsymbol{\alpha}_t(a) \left(x(a)\right)^2, && (20)
\end{aligned}$$

where (d) follows from Proposition 1 of Liu *et al.* [2018], (e) follows from the KL divergence form of mutual information and (f) follows by dropping some nonnegative terms.

As shorthand, we let $z(a) = \frac{\sum_{a':a' \xrightarrow{t} a} \boldsymbol{\alpha}_t(a')}{\boldsymbol{\alpha}_t(a)}$. Now, we are ready to bound the information ration.

$$\begin{aligned}
(\boldsymbol{\Delta}_t^T \boldsymbol{\alpha}_t)^2 &\overset{(g)}{\leq} \frac{1}{2} \left(\sum_{a\in\mathcal{K}} \boldsymbol{\alpha}_t(a)x(a) \right)^2 && (21)\\
&= \frac{1}{2} \left(\sum_{a\in\mathcal{K}} \frac{1}{\sqrt{z(a)}} \sqrt{z(a)} \boldsymbol{\alpha}_t(a)x(a) \right)^2\\
&\overset{(h)}{\leq} \frac{1}{2} \left(\sum_{a\in\mathcal{K}} \frac{1}{z(a)} \right) \left(\sum_{a\in\mathcal{K}} z(a) \left(\boldsymbol{\alpha}_t(a)x(a)\right)^2 \right)\\
&\overset{(i)}{\leq} \frac{1}{2} \left(\sum_{a\in\mathcal{K}} \frac{1}{z(a)} \right) \boldsymbol{g}_t^T \boldsymbol{\alpha}_t && (22)
\end{aligned}$$

where (g) follows from equation (17), (h) follows from Cauchy-Schwartz inequality and (i) follows from equation (20).

B Proof of Theorem 1

First observe that the entropy bounds the expected cumulative information gain.

$$\begin{aligned}
&\mathbb{E} \sum_{t=1}^T \boldsymbol{g}_t^T \boldsymbol{\pi}_t && (23)\\
&= \mathbb{E} \sum_{t=1}^T \left(\sum_{i\in\mathcal{K}} \boldsymbol{\pi}_t(i) \mathbb{E}[H(\boldsymbol{\alpha}_t) - H(\boldsymbol{\alpha}_{t+1})|\mathcal{F}_t, A_t = i] \right)\\
&= \mathbb{E} \sum_{t=1}^T \mathbb{E}[H(\boldsymbol{\alpha}_t) - H(\boldsymbol{\alpha}_{t+1})|\mathcal{F}_t] && (24)\\
&= \mathbb{E} \sum_{t=1}^T \left(H(\boldsymbol{\alpha}_t) - H(\boldsymbol{\alpha}_{t+1})\right) && (25)\\
&\leq H(\boldsymbol{\alpha}_1), && (26)
\end{aligned}$$

Then, we bound the regret of TS-N.

$$\begin{aligned}
\mathbb{E}[R(T,\boldsymbol{\pi})] &= \mathbb{E} \sum_{t=1}^T \boldsymbol{\Delta}_t^T \boldsymbol{\pi}_t = \mathbb{E} \sum_{t=1}^T \frac{\boldsymbol{\Delta}_t^T \boldsymbol{\pi}_t}{\sqrt{\boldsymbol{g}_t^T \boldsymbol{\pi}_t}} \sqrt{\boldsymbol{g}_t^T \boldsymbol{\pi}_t}\\
&\overset{(a)}{\leq} \sqrt{\mathbb{E} \sum_{t=1}^T \frac{(\boldsymbol{\Delta}_t^T \boldsymbol{\pi}_t)^2}{\boldsymbol{g}_t^T \boldsymbol{\pi}_t}} \sqrt{\mathbb{E} \sum_{t=1}^T \boldsymbol{g}_t^T \boldsymbol{\pi}_t}\\
&\overset{(b)}{\leq} \sqrt{\frac{1}{2} \sum_{t=1}^T \mathbb{E}[Q_t(\boldsymbol{\alpha}_t)] H(\boldsymbol{\alpha}_1)} && (27)
\end{aligned}$$

where (a) follows from Holder's inequality and (b) follows from Proposition 1 and equation (26).

C Proof of Theorem 2

As shorthand, we let $d(a,a^*) = D(\mathbb{P}(Y_{t,a} \in \cdot|\mathcal{F}_t, A^* = a^*)||\mathbb{P}(Y_{t,a} \in \cdot|\mathcal{F}_t))$ and $x(a) = \sqrt{d(a,a)}$. By the definition of the instantaneous regret, we have that

$$\boldsymbol{\Delta}_t^T\boldsymbol{\pi}_t = \sum_{a\in\mathcal{K}}\boldsymbol{\pi}_t(a)\mathbb{E}[Y_{t,A^*}-Y_{t,a}|\mathcal{F}_t] \tag{28}$$

$$\overset{(a)}{=}\sum_{a\in\mathcal{K}}\boldsymbol{\pi}_t(a)\left(\mathbb{E}[Y_{t,A^*}|\mathcal{F}_t]-\mathbb{E}[Y_{t,a}|\mathcal{F}_t]\right) \tag{29}$$

$$=\mathbb{E}[Y_{t,A^*}|\mathcal{F}_t]-\sum_{a\in\mathcal{K}}\boldsymbol{\pi}_t(a)\mathbb{E}[Y_{t,a}|\mathcal{F}_t] \tag{30}$$

$$\overset{(b)}{=}\sum_{a\in\mathcal{K}}\boldsymbol{\alpha}_t(a)\mathbb{E}[Y_{t,a}|\mathcal{F}_t,A^*=a]-\sum_{a\in\mathcal{K}}\boldsymbol{\pi}_t(a)\mathbb{E}[Y_{t,a}|\mathcal{F}_t]$$

$$\overset{(c)}{\le}(1-\epsilon)\sum_{a\in\mathcal{K}}\boldsymbol{\alpha}_t(a)\left(\mathbb{E}[Y_{t,a}|\mathcal{F}_t,A^*=a]-\mathbb{E}[Y_{t,a}|\mathcal{F}_t]\right)+\epsilon$$

$$\overset{(d)}{\le}(1-\epsilon)\sqrt{\frac{1}{2}}\sum_{a\in\mathcal{K}}\left(\boldsymbol{\alpha}_t(a)\sqrt{d(a,a)}\right)+\epsilon \tag{31}$$

$$=(1-\epsilon)\sqrt{\frac{1}{2}}\sum_{a\in\mathcal{K}}\left(\boldsymbol{\alpha}_t(a)x(a)\right)+\epsilon, \tag{32}$$

where (a) follows from the linearity of expectation, (b) uses the law of total probability, (c) follows from the fact that $\boldsymbol{\pi}_t=(1-\epsilon)\boldsymbol{\alpha}_t+\epsilon/K$ and the rewards are bounded by 1, (d) follows from the Pinsker's inequality. Step (c) allows us to decompose the regret into the regret from uniform sampling and the regret from Thompson Sampling. Thus, we can further relate the latter regret term to the expected information gain.

By the definition of the information gain of observing an action, we have that

$$\boldsymbol{g}_t^T\boldsymbol{\pi}_t\overset{(e)}{\ge}(\boldsymbol{G}_t\boldsymbol{h}_t)^T\boldsymbol{\pi}_t$$

$$=\sum_{a\in\mathcal{K}}\left(\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')\right)I_t(A^*;Y_{t,a}) \tag{33}$$

$$\overset{(f)}{=}\sum_{a\in\mathcal{K}}\left(\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')\right)\left(\sum_{a^*\in\mathcal{K}}\boldsymbol{\alpha}_t(a^*)d(a,a^*)\right)$$

$$\overset{(g)}{\ge}\sum_{a\in\mathcal{K}}\left(\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')\right)\boldsymbol{\alpha}_t(a)d(a,a) \tag{34}$$

$$=\sum_{a\in\mathcal{K}}\left(\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')\right)\boldsymbol{\alpha}_t(a)\left(x(a)\right)^2, \tag{35}$$

where (e) follows from Proposition 1 of Liu *et al.* [2018], (f) follows from the KL divergence form of mutual information and (g) follows by dropping some nonnegative terms.

As shorthand, we let $\zeta(a)=\frac{\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')}{\boldsymbol{\alpha}_t(a)}$. Now we are ready to bound the first term in equation (32).

$$\sum_{a\in\mathcal{K}}\boldsymbol{\alpha}_t(a)x(a)=\sum_{a\in\mathcal{K}}\frac{1}{\sqrt{\zeta(a)}}\sqrt{\zeta(a)}\boldsymbol{\alpha}_t(a)x(a) \tag{36}$$

$$\overset{(h)}{\le}\sqrt{\sum_{a\in\mathcal{K}}\frac{1}{\zeta(a)}}\sqrt{\sum_{a\in\mathcal{K}}\zeta(a)(\boldsymbol{\alpha}_t(a)x(a))^2} \tag{37}$$

$$\overset{(i)}{\le}\sqrt{\sum_{a\in\mathcal{K}}\frac{\boldsymbol{\alpha}_t(a)}{\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')}}\sqrt{\boldsymbol{g}_t^T\boldsymbol{\pi}_t} \tag{38}$$

$$\overset{(j)}{\le}\sqrt{\frac{1}{1-\epsilon}\sum_{a\in\mathcal{K}}\frac{\boldsymbol{\pi}_t(a)}{\sum_{a':a'\xrightarrow{t}a}\boldsymbol{\pi}_t(a')}}\sqrt{\boldsymbol{g}_t^T\boldsymbol{\pi}_t} \tag{39}$$

$$=\sqrt{\frac{1}{1-\epsilon}Q_t(\boldsymbol{\pi}_t)}\sqrt{\boldsymbol{g}_t^T\boldsymbol{\pi}_t}, \tag{40}$$

where (h) follows from Cauchy-Schwartz inequality, (i) follows from equation (35) and (j) follows from the fact that $\boldsymbol{\pi}_t\ge(1-\epsilon)\boldsymbol{\alpha}_t$.

Now, we are ready to bound the regret.

$$\mathbb{E}[R(T,\boldsymbol{\pi})]=\mathbb{E}\sum_{t=1}^T\boldsymbol{\Delta}_t^T\boldsymbol{\pi}_t \tag{41}$$

$$\overset{(k)}{\le}\mathbb{E}\sum_{t=1}^T\left((1-\epsilon)\sqrt{\frac{1}{2}}\sum_{a\in\mathcal{K}}\left(\boldsymbol{\alpha}_t(a)x(a)\right)+\epsilon\right)$$

$$\overset{(l)}{\le}\epsilon T+\sqrt{\frac{1}{2}}\mathbb{E}\sum_{t=1}^T\sqrt{(1-\epsilon)Q_t(\boldsymbol{\pi}_t)}\sqrt{\boldsymbol{g}_t^T\boldsymbol{\pi}_t}$$

$$\overset{(m)}{\le}\epsilon T+\sqrt{\frac{1}{2}\mathbb{E}\sum_{t=1}^TQ_t(\boldsymbol{\pi}_t)}\sqrt{\mathbb{E}\sum_{t=1}^T\boldsymbol{g}_t^T\boldsymbol{\pi}_t}$$

$$\overset{(n)}{\le}\epsilon T+\sqrt{\frac{1}{2}\sum_{t=1}^T\mathbb{E}[Q_t(\boldsymbol{\pi}_t)]H(\boldsymbol{\alpha}_1)}, \tag{42}$$

where (k) follows from equation (32), (l) follows from equation (40), (m) follows from Holder's inequality and (n) follows from equation (26).