I. Proof of Theorem 1

For a sequence of numbers $\{r\}_{r\in\mathcal{R}}$, let $\min2(\{r\}_{r\in\mathcal{R}})$ be the magnitude of the smallest difference between any two numbers in $\mathcal{R}$. Let

$$\Delta_{\min,1} := \min_{\boldsymbol{f}_0 \in \bar{\mathcal{F}}_0} \min2(\{y_{\boldsymbol{a},\boldsymbol{f}_0}\}_{\boldsymbol{a}\in\bar{\mathcal{A}}_1})$$

and

$$\Delta_{\min,t} := \\ \min_{x\in\mathcal{X}} \min_{(\boldsymbol{a}[t-1],\boldsymbol{f}[t-1])\in S_{t-1}(x)} \left(\min2(\{y_{(\boldsymbol{a}[t-1],\boldsymbol{a}),\boldsymbol{f}[t-1]}\}_{\boldsymbol{a}\in\bar{\mathcal{A}}_t})\right)$$

for $1 < t \leq l_{\max}$, where $S_t(x) = \{\boldsymbol{a}[t] \in \boldsymbol{\mathcal{A}}[t], \boldsymbol{f}[t] \in \boldsymbol{\mathcal{F}}(\boldsymbol{a}[t]) : \phi(\boldsymbol{a}[t], \boldsymbol{f}[t]) = x\}$. By Assumptions 1 and 2 it is guaranteed that $\Delta_{\min,t} > 0$ for $1 \leq t \leq l_{\max}$. Let

$$\Delta_{\min} := \min_{t=1,\dots,l_{\max}} \Delta_{\min,t}$$

denote the *minimum gap*.

The proof involves showing that when FAL estimates the expected rewards for the action sequences it selects such that they are within $\Delta_{\min}/2$ of the true expected rewards, then it will always select the same sequence of actions as the BF benchmark does. The regret of FAL can be written as

$$\mathrm{E}[R(n)] = \mathrm{E}[R_e(n)] + \mathrm{E}[R_s(n)]$$

where $R_e(n)$ is the total (random) regret in rounds when FAL explores, and $R_s(n)$ is the total (random) regret in rounds when FAL exploits, where the expectation is taken with respect to the distribution of feedbacks given the past action and feedback sequences.

Let $x_t^\rho = \phi(\boldsymbol{a}^\rho[t], \boldsymbol{f}^\rho[t]) = \boldsymbol{f}_t^\rho$ be the state at the end of stage t of round ρ. To proceed, we define the following sets of rounds. Let $E_1(n)$ be the set of rounds in $\{1, \dots, n\}$ for which FAL explores the action selected in the first stage of that round, i.e., $\rho \in \{1, \dots, n\}$ for which

$$T_{x_0^\rho,1,\boldsymbol{a}}(\rho) < D\log(\rho/\delta)$$

for some $\boldsymbol{a} \in \mathcal{A}_1^\rho(B_1^\rho)$ such that after randomly selecting an under-explored action $\boldsymbol{a}$, remaining actions are taken according to a predefined rule. Let $E_t(n)$, $1 < t \leq l_{\max}$ be the set of rounds $\rho \in \{1, \dots, n\}$ for which FAL explores in the tth stage of that round, i.e., the set of rounds for which FAL exploited up to stage $t-1$ and

$$T_{x_{t-1}^\rho,t,\boldsymbol{a}} < D\log(\rho/\delta)$$

for some $\boldsymbol{a} \in \mathcal{A}_t^\rho(B_t^\rho)$ such that after randomly selecting an under-explored action $\boldsymbol{a}$, the remaining actions are taken according to the predefined rule. Let $\tau_1(n)$ be the set of rounds $\rho \in \{1, \dots, n\}$ for which FAL exploits for the first stage of the round, i.e.,

$$T_{x_0^\rho,1,\boldsymbol{a}} \geq D\log(\rho/\delta)$$

for all $\boldsymbol{a} \in \mathcal{A}_1^\rho(B_1^\rho)$. Let $\tau_t(n)$ be the set of rounds $\rho \in \{1, \dots, n\}$ for which FAL exploits for the tth stage in that round, i.e.,

$$T_{x_{t-1}^\rho,t,\boldsymbol{a}} \geq D\log(\rho/\delta)$$

for all $\boldsymbol{a} \in \mathcal{A}_t^\rho(B_t^\rho)$ and FAL also exploited in all stages prior to stage t (FAL has not followed the predefined rule before stage t). The set of all rounds for which FAL explores until the nth round is equal to

$$E(n) := \bigcup_{t=1}^{l_{\max}} E_1(n)$$

where $E_t(n) \cap E_{t'}(n) = \emptyset$ for $t \neq t'$.

In the following we will bound $R_s(n)$. We define the events which correspond to the case that the estimated reward of the action that will be selected by the BF benchmark is always greater than the estimated rewards of the other actions. Hence, when these events happen, FAL operates exactly the same way as the BF benchmark. Let $a^*(\mathcal{A}_t^\rho(B_t^\rho), \boldsymbol{a}^\rho[t-1], \boldsymbol{f}^\rho[t-1])$ be the action chosen by BF benchmark at stage t of round ρ when previous sequence of feedbacks and actions are $\boldsymbol{a}^\rho[t-1], \boldsymbol{f}^\rho[t-1]$. Let

$$\mathrm{Perf}_1(n) := \{a_1^\rho = a^*(\mathcal{A}_1^\rho(B_1^\rho), \emptyset, \boldsymbol{f}_0^\rho), \forall \rho \in \tau_1(n)\}$$

and

$$\begin{aligned}\mathrm{Perf}_t(n) := \{&a_t^\rho = a^*(\mathcal{A}_t^\rho(B_t^\rho), \boldsymbol{a}^\rho[t-1], \boldsymbol{f}^\rho[t-1]), \\ &\forall \rho \in \tau_t(n)\}\,.\end{aligned}$$

We have

$$\begin{aligned}&\mathrm{Perf}_1(n) \\ &\supset \{|\hat{y}_{\boldsymbol{f}_0^\rho,1,\boldsymbol{a}}(\rho) - y_{\boldsymbol{a},\boldsymbol{f}_0^\rho}| < \Delta_{\min}/2, \forall \boldsymbol{a} \in \bar{\mathcal{A}}_1, \forall \rho \in \tau_1(n)\}.\end{aligned} \tag{1}$$

Let $\mathrm{Perf}(n) = \bigcap_{t=1}^{l_{\max}} \mathrm{Perf}_t(n)$. On event $\mathrm{Perf}(n)$, FAL selects sequence of actions in the same way as the BF benchmark does. Hence, the contribution to the regret given in Equation 1 of the manuscript on event $\mathrm{Perf}(n)$ is zero.

Next, we lower bound the probability of event $\mathrm{Perf}(n)$. Using the chain rule we can write

$$\begin{aligned}\mathrm{P}(\mathrm{Perf}(n)) &= \mathrm{P}(\mathrm{Perf}_{l_{\max}}(n), \mathrm{Perf}_{l_{\max}-1}(n), \dots, \mathrm{Perf}_1(n)) \\ &= \mathrm{P}(\mathrm{Perf}_{l_{\max}}(n) | \mathrm{Perf}_{l_{\max}-1}(n), \dots, \mathrm{Perf}_1(n)) \\ &\times \mathrm{P}(\mathrm{Perf}_{l_{\max}-1}(n) | \mathrm{Perf}_{l_{\max}-2}(n), \dots, \mathrm{Perf}_1(n)) \\ &\times \dots \times \mathrm{P}(\mathrm{Perf}_2(n) | \mathrm{Perf}_1(n)) \times \mathrm{P}(\mathrm{Perf}_1(n)).\end{aligned} \tag{2}$$

For an event E, let E^c denote its complement. Note that we have

$$\begin{aligned}&\mathrm{P}(\mathrm{Perf}_1(n)^c) \\ &\leq \sum_{\rho\in\tau_1(n)} \sum_{\boldsymbol{f}\in\bar{\mathcal{F}}_0} \sum_{\boldsymbol{a}\in\bar{\mathcal{A}}_1} \mathrm{P}(|\hat{y}_{\boldsymbol{f},1,\boldsymbol{a}}(\rho) - y_{\boldsymbol{a},\boldsymbol{f}}| \geq \Delta_{\min}/2) \\ &\leq \sum_{\rho\in\tau_1(n)} 2F_{\max}A_{\max}\exp(-2D\log(\rho/\delta)\Delta_{\min}^2/4) \\ &\leq \sum_{\rho\in\tau_1(n)} 2F_{\max}A_{\max}\delta^2/\rho^2 \leq 2F_{\max}A_{\max}\beta\delta^2\end{aligned}$$

where the first inequality follows from (1) and union bound, the second inequality follows from the Chernoff-Hoeffding bound and the third inequality follows from the fact that $D \geq 4/\Delta_{\min}^2$ and $\beta = \sum_{\rho=1}^\infty 1/\rho^2$. Hence, we have

$$\mathrm{P}(\mathrm{Perf}_1(n)) \geq 1 - 2F_{\max}A_{\max}\beta\delta^2.$$

On event $\mathrm{Perf}_1(n)$, it is always the case that the first selected action by FAL is chosen according to the BF benchmark, independent of whether the FAL explores or exploits in the second stage of those rounds. Hence given $\mathrm{Perf}_1(n)$, in all rounds ρ', which contribute to the estimation of the sample mean ex-ante reward $\hat{y}_{x_1^\rho,2,a^*(\mathcal{A}_2^\rho(B_2^\rho),\boldsymbol{a}^\rho[1],\boldsymbol{f}^\rho[1])}(\rho)$ (exploration in stage 2), the reward $Y(\rho')$ comes from a distribution whose expectation is at least $\Delta_{\min}$ greater than the expected reward of any other action $\boldsymbol{a} \in \mathcal{A}_2^\rho(B_2^\rho)$ (due to Assumptions 1 and 2). Due to this fact, we can use the Chernoff-Hoeffding inequality to bound the probability of $\mathrm{Perf}_2(n)$ given $\mathrm{Perf}_1(n)$ by $\mathrm{P}(\mathrm{Perf}_2(n)|\mathrm{Perf}_1(n)) \geq 1 - 2F_{\max}A_{\max}\beta\delta^2$. Similarly, it can be shown that $\mathrm{P}(\mathrm{Perf}_t(n)|\mathrm{Perf}_{t-1}(n), \ldots, \mathrm{Perf}_1(n)) \geq 1 - 2F_{\max}A_{\max}\beta\delta^2$ for other stages t. Combining all of this and using (2) we get

$$\begin{aligned}\mathrm{P}(\mathrm{Perf}(n)) \geq & (1 - 2F_{\max}A_{\max}\beta\delta^2)^{l_{\max}} \\ & \geq 1 - 2F_{\max}A_{\max}l_{\max}\beta\delta^2 = 1 - \epsilon\end{aligned}$$

for $\delta = \sqrt{\epsilon/(2\beta F_{\max}A_{\max}l_{\max})}$.

Next we bound $R_e(n)$. From the definition of $E_t(n)$, $t = 1, \ldots, l_{\max}$, we know that $|E_t(n)| \leq 2F_{\max}A_{\max}DX\log(n/\delta)$ for $t = 1, \ldots, l_{\max}$. Hence, we have $|E(n)| \leq 2l_{\max}F_{\max}A_{\max}DX\log(n/\delta)$. Since the worst-case reward loss due to a suboptimal sequence of actions is at most 1, we have $R_e(n) \leq 2l_{\max}F_{\max}A_{\max}DX\log(n/\delta)$. Finally, the regret bound on $\mathrm{E}[R(n)]$ holds by setting $\epsilon = 1/n$ and taking the expectation.