A. Proof of Theorem 1

Sampling from the posterior as proportional to exponential weights: Let $N_t(a)$ be the number of times action $a$ has been played up to (and including) time $t$. At any time $t$, the posterior distribution $\pi_t$ over $\Theta$ is given by Bayes’ rule:

$$\forall S \subseteq \Theta: \pi_t(S) = \frac{W_t(S)}{W_t(\Theta)} = \int_S W_t(\theta) \pi(d\theta),$$

(4)

with the weight $W_t(\theta)$ of each $\theta$ being the likelihood of observing the history under $\theta$:

$$W_t(\theta) := \prod_{i=1}^t \frac{l(Y_i; A_i, \theta)}{l(Y_i; A_i, \theta^*)} = \prod_{a \in A} \prod_{y \in Y} \prod_{i=1}^t \left[ \frac{l(y; a, \theta)}{l(y; a, \theta^*)} \right]^{1 \{A_i = a, Y_i = y\}}$$

$$= \exp \left( - \sum_{a \in A} \sum_{y \in Y} \sum_{i=1}^t 1 \{A_i = a, Y_i = y\} \log \frac{l(y; a, \theta^*)}{l(y; a, \theta)} \right)$$

$$= \exp \left( - \sum_{a \in A} N_t(a) \sum_{y \in Y} \frac{\sum_{i=1}^t 1 \{A_i = a, Y_i = y\}}{N_t(a)} \log \frac{l(y; a, \theta^*)}{l(y; a, \theta)} \right),$$

where we set $N_t(a) := \sum_{i=1}^t 1 \{A_i = a\}$. Let $Z_t(a, y) := \sum_{i=1}^t 1 \{A_i = a, Y_i = y\}$, and $Z_t(a) := (Z_t(a, y))_{y \in Y} \in \mathbb{R}^{|Y|}$. Thus $Z_t(a)$ is the empirical distribution of the observations from playing action $a$ up to time $t$. The expression for $W_t(\theta)$ above becomes

$$W_t(\theta) = \exp \left( - \sum_{a \in A} N_t(a) D(\theta^* || \theta_a) - \sum_{a \in A} N_t(a) \sum_{y \in Y} (Z_t(a, y) - l(y; a, \theta^*)) \log \frac{l(y; a, \theta^*)}{l(y; a, \theta)} \right).$$

(5)

Note that by definition, $W_t(\theta^*) = 1$ at all times $t$ – a fact that we use often in the analysis.

Instead of observing $Y_t = f(X_t, A_t)$ at each round $t$, consider the following alternative probability space for the stochastic bandit in a time horizon $1, 2, \ldots$ with probability measure $\tilde{\mathbb{P}}$. First, for each action $a \in A$ and each time $k = 1, 2, \ldots$, an independent random variable $Q_a(k) \in Y$, is drawn with $\mathbb{P}(Q_a(k) = y) = l(y; a, \theta^*)$. Denote by $Q := \{Q_a(k)\}_{a \in A, k \geq 1}$ the $|A| \times \infty$ matrix of these independent random variables. Next, at each round $t = 1, 2, \ldots$, playing action $A_t = a$ yields the observation $Y_t = Q(a, N_a(t) + 1)$. Thus, in this space,

$$Z_t(a, y) = U_{N_t(a)}(a, y), \text{ where } U_{j}(a, y) := \frac{1}{j} \sum_{k=1}^j 1 \{Q_a(k) = y\}.$$

The following lemma shows that the distribution of sample paths seen by a bandit algorithm in both probability spaces (i.e., associated with the measures $\mathbb{P}$ and $\tilde{\mathbb{P}}$) is identical. This allows us to equivalently work in the latter space to make statements about the regret of an algorithm.

**Lemma 1.** For any action-observation sequence $(a_t, y_t), t = 1, \ldots, T$ of a bandit algorithm,

$$\tilde{\mathbb{P}} \{1 \leq t \leq T, (A_t, Y_t) = (a_t, y_t)\} = \mathbb{P} \{1 \leq t \leq T, (A_t, Y_t) = (a_t, y_t)\}.$$

Henceforth, we will drop the tilde on $\tilde{\mathbb{P}}$ and always work in the latter probability space, involving the matrix $Q$.

**Lemma 2.** For any suboptimal action $a \neq a^*$,

$$\delta_a = \min_{\theta \in S_a^*} D(\theta^*_a || \theta_a) > 0.$$
Let \( N'_i(a) \) (resp. \( N''_i(a) \)) be the number of times that a parameter has been drawn from \( S'_a \) (resp. \( S''_a \)), so that \( N_i(a) = N'_i(a) + N''_i(a) \).

The following self-normalized, uniform deviation bound controls the empirical distribution of each row \( Q_a(\cdot) \) of the random reward matrix \( Q \). It is a version of a bound proved in (Abbasi-Yadkori et al., 2011).

**Theorem 3.** Let \( a \in \mathcal{A}, y \in \mathcal{Y} \) and \( \delta \in (0,1) \). Then, with probability at least \( 1 - \delta \sqrt{2} \),

\[
\forall k \geq 1 \quad |U_k(a,y) - l(y;a,\theta^*)| \leq 4 \sqrt{\frac{1}{k} \log \left( \frac{\sqrt{k}}{\delta} \right)}.
\]

Put \( c := \log \frac{|\mathcal{Y}||\mathcal{A}|}{\delta} \), and \( \rho(x) \equiv \rho_c(x) := 4 \sqrt{c + \log x} \) for \( x > 0 \). It follows that the following “good data” event occurs with probability at least \( (1 - \delta \sqrt{2}) \):

\[
G \equiv G(c) := \left\{ \forall a \in \mathcal{A} \quad \forall y \in \mathcal{Y} \forall k \geq 1 \quad |U_k(a,y) - l(y;a,\theta^*)| \leq \rho(k) \sqrt{k} \right\}.
\]

**Lemma 3.** Fix \( \epsilon \in (0,1) \). There exist \( \lambda, n^* \geq 0 \), not depending on \( T \), so that the following is true. For any \( \theta \in \Theta, a \in \mathcal{A} \) and \( y \in \mathcal{Y} \), under the event \( G \),

1. At all times \( t \geq 1 \),

\[
N_i(a)D(\theta^*_a||\theta_a) + N_i(a) \sum_{y \in \mathcal{Y}} (Z_i(a,y) - l(y;a,\theta^*)) \log \frac{l(y;a,\theta^*)}{l(y;a,\theta)} \geq -\lambda.
\]

2. If \( N_i(a) \geq n^* \), then

\[
N_i(a)D(\theta^*_a||\theta_a) + N_i(a) \sum_{y \in \mathcal{Y}} (Z_i(a,y) - l(y;a,\theta^*)) \log \frac{l(y;a,\theta^*)}{l(y;a,\theta)} \geq (1 - \epsilon)N_i(a)D(\theta^*_a||\theta_a).
\]

**Proof.** Under \( G \), we have

\[
N_i(a)D(\theta^*_a||\theta_a) + N_i(a) \sum_{y \in \mathcal{Y}} (Z_i(a,y) - l(y;a,\theta^*)) \log \frac{l(y;a,\theta^*)}{l(y;a,\theta)} \\
\geq N_i(a)D(\theta^*_a||\theta_a) - N_i(a) \sum_{y \in \mathcal{Y}} |Z_i(a,y) - l(y;a,\theta^*)| \left| \log \frac{l(y;a,\theta^*)}{l(y;a,\theta)} \right| \\
\geq N_i(a)D(\theta^*_a||\theta_a) - \rho(N_i(a)) \sqrt{N_i(a)} \sum_{y \in \mathcal{Y}} \left| \log \frac{l(y;a,\theta^*)}{l(y;a,\theta)} \right|.
\]

(6)

For a fixed \( \theta \in \Theta, a \in \mathcal{A} \), the expression above diverges to \( +\infty \), viewed as a function of \( N_i(a) \), as \( N_i(a) \to \infty \) (except when \( \theta_a = \theta^*_a \), in which case the expression is identically 0.) Hence, the expression achieves a finite minimum \( -\lambda_{\theta,a} \) (not depending on \( T \)) over non-negative integers \( N_i(a) \in \mathbb{Z}^+ \). Since there are only finitely many parameters \( \theta \in \Theta \), it follows that if we set \( \lambda := \max_{\theta \in \Theta, a \in \mathcal{A}} \lambda_{\theta,a} \), then the above expression is bounded below by \( -\lambda \), uniformly across \( \Theta \). This proves the first part of the lemma.

To show the second part, notice again that for fixed \( \theta \in \Theta \) and \( a \in \mathcal{A} \), there exists \( n^*_{\theta,a} \geq 0 \) such that

\[
\rho(x) \sqrt{x} \sum_{y \in \mathcal{Y}} \left| \log \frac{l(y;a,\theta^*)}{l(y;a,\theta)} \right| \leq cxD(\theta^*_a||\theta_a), \quad x \geq n^*_{\theta,a}
\]

since \( \rho(x) = o(x) \). Setting \( n^* := \max_{\theta \in \Theta, a \in \mathcal{A}} n^*_{\theta,a} \) then completes the proof of the second part. \( \square \)
A.1. Regret due to sampling from \( S_a'' \)

The result of Lemma 3 implies that under the event \( G \), and at all times \( t \geq 1 \):

\[
\pi_t(\theta) = \frac{W_t(\theta^*) \pi(\theta^*)}{\int_{[0,1]} W_t(\theta) \pi(d\theta)} = \frac{\pi(\theta^*)}{\int_{[0,1]} W_t(\theta) \pi(d\theta)} \geq \frac{1}{\int_{[0,1]} \exp(\lambda |A|) \pi(d\theta)} = \pi(\theta^*) e^{-\lambda |A|} \equiv p^*, \text{ say.} \tag{7}
\]

Also, under the event \( G \), the posterior probability of \( \theta \in S_a'' \) at all times \( t \) can be bounded above using Lemma 3 and the basic bound in (6):

\[
\pi_t(\theta) = \frac{W_t(\theta) \pi(\theta)}{\int_{[0,1]} W_t(\psi) \pi(d\psi)} \leq \frac{W_t(\theta^*) \pi(\theta^*)}{\int_{[0,1]} W_t(\theta) \pi(d\theta)} = \frac{\pi(\theta^*)}{\int_{[0,1]} \exp(\lambda |A|) \pi(d\theta)} = \frac{\pi(\theta^*)}{\int_{[0,1]} \exp(\lambda |A|) \pi(d\theta)} \cdot \exp \left( -\sum_{a \in A} N_t(a) D(\theta^*_a || \theta_a) - \sum_{a \in A} N_t(a) \sum_{y \in Y} (Z_t(a, y) - l(y; a, \theta^*)) \log \frac{l(y; a, \theta^*)}{l(y; a, \theta)} \right) \leq \frac{\pi(\theta^*)}{\pi(\theta^*)} \cdot \exp \left( -N_t(a^*) D(\theta^* || \theta_{a^*}) + \rho(N_t(a)) \sqrt{N_t(a^*)} \sum_{y \in Y} \left| \log \frac{l(y; a^*, \theta^*)}{l(y; a^*, \theta)} \right| \right).
\]

In the above, the penultimate inequality is by Lemma 3 applied to all actions \( a \neq a^* \), and the final inequality follows in a manner similar to (6), for action \( a^* \). Letting \( d := \frac{\lambda |A|}{\pi(\theta^*)} \), we have that under the event \( G \), for \( a \neq a^* \) and \( \theta \in S_a'' \),

\[
\pi_t(\theta) \leq d \pi(\theta) \exp \left( -N_t(a^*) D(\theta^* || \theta_{a^*}) + \rho(N_t(a)) \sqrt{N_t(a^*)} \sum_{y \in Y} \left| \log \frac{l(y; a^*, \theta^*)}{l(y; a^*, \theta)} \right| \right). \tag{8}
\]

Recall that by definition, any \( \theta \in S_a'' \) with \( a \neq a^* \) can be resolved apart from \( \theta^* \) in the action \( a^* \), i.e., \( D(\theta^* || \theta_{a^*}) \geq \xi \). Moreover, the discrete prior assumption (Assumption 2) implies that \( \xi > 0 \). Using this, we can bound the right-hand side of (8) further under the event \( G \):

\[
\pi_t(\theta) \leq d \pi(\theta) \exp \left( -\xi N_t(a^*) + 2 \rho(N_t(a)) \sqrt{N_t(a^*)} \log \frac{1 - \Gamma}{\Gamma} \right). \tag{9}
\]

Integrating (9) over \( \theta \in S_a'' \) and noticing that \( \pi(S_a'') \leq 1 \) gives, under \( G \),

\[
\pi_t(S_a'') \leq d \exp \left( -\xi N_t(a^*) + 2 \rho(N_t(a)) \sqrt{N_t(a^*)} \log \frac{1 - \Gamma}{\Gamma} \right). \tag{10}
\]

We can now estimate, using the conditional version of Markov’s inequality, the number of times that parameters from \( S_a'' \) are sampled under “good data” \( G \):

\[
P \left[ \sum_{t=1}^{\infty} 1 \{ \theta_t \in S_a'' \} > \eta \mid G \right] \leq \eta^{-1} \sum_{t=1}^{\infty} \mathbb{E} \left[ 1 \{ \theta_t \in S_a'' \} \mid G \right] = \eta^{-1} \sum_{t=1}^{\infty} \mathbb{E} \left[ \pi_t(S_a'') \mid G \right] \leq \eta^{-1} \sum_{t=1}^{\infty} \left( 1 \wedge \mathbb{E} \left[ d \exp \left( -\xi N_t(a^*) + 2 \rho(N_t(a)) \sqrt{N_t(a^*)} \log \frac{1 - \Gamma}{\Gamma} \right) \mid G \right] \right), \tag{11}
\]

where the final inequality is by (10) and the fact that \( \pi_t(S_a'') \leq 1 \).\(^{13}\)

\(^{13}\) \( \wedge \) denotes the minimum of \( a \) and \( b \).
At each time $t$, if we let $\mathcal{F}_t$ denote the $\sigma$-algebra generated by the random variables $\{(\theta_i, A_i, Y_i) : i \leq t\}$, then

$$
\mathbb{E}
\left[
\xi_{N_t(a^*)} \mid \mathcal{G}
\right] = \mathbb{E}
\left[
\mathbb{E}
\left[
\xi_{N_t(a^*)} \mid \mathcal{F}_{t-1}, \mathcal{G}
\right] \mid \mathcal{G}
\right]
= \mathbb{E}
\left[
\mathbb{E}
\left[
\xi_{N_{t-1}(a^*)} \mid \mathcal{F}_{t-1}, \mathcal{G}
\right] \mid \mathcal{G}
\right]
\leq \mathbb{E}
\left[
\mathbb{E}
\left[
\xi_{N_{t-1}(a^*)} \mid \mathcal{F}_{t-1}, \mathcal{G}
\right] \mid \mathcal{G}
\right]
(\theta_t = \theta \Rightarrow A_t = a^*)
= \mathbb{E}
\left[
\xi_{N_{t-1}(a^*)} (\pi_t(\theta^*)e^{-\xi} - \mathbb{E}[\xi_{N_{t-1}(a^*)}]) \mid \mathcal{G}
\right]
\leq \mathbb{E}
\left[
\xi_{N_{t-1}(a^*)} (p^*e^{-\xi} + 1 - p^*) \mid \mathcal{G}
\right]
= (p^*e^{-\xi} + 1 - p^*) \mathbb{E}
\left[
\xi_{N_{t-1}(a^*)} \mid \mathcal{G}
\right],
$$

where, in the penultimate step, we use $\pi_t(\theta^*) \geq p^* \cdot 1_G$ from (7). Iterating this estimate and using it in (11) together with the trivial bound $\sqrt{N_t(a^*)} \leq \sqrt{t}$ gives

$$
\mathbb{P}
\left[
\sum_{t=1}^{\infty} \mathbb{1}\{\theta_t \in S_a^{\prime \prime} \} > \eta \mid \mathcal{G}
\right] \leq \mathbb{P}
\left[
\sum_{t=1}^{\infty} (1 \land d(p^*e^{-\xi} + 1 - p^*)) \mathbb{1}\{\theta_t \in S_a^{\prime \prime} \} > \eta \mid \mathcal{G}
\right] \exp \left(2\rho(t)\sqrt{t} \log \frac{1 - \Gamma}{\delta}\right).
$$

Since $p^*e^{-\xi} + 1 - p^* < 1$ and $\rho(t)\sqrt{t} = o(t)$, the sum above is dominated by a geometric series after finitely many $t$, and is thus a finite quantity $\alpha < \infty$, say. (Note that $\alpha$ does not depend on $T$.) Replacing $\delta$ by $\frac{\delta}{\alpha}$ and taking a union bound over all $a \neq a^*$, this proves

**Lemma 4.** There exists $\alpha < \infty$ such that

$$
\mathbb{P}
\left[
G, \exists a \neq a^* \sum_{t=1}^{\infty} \mathbb{1}\{\theta_t \in S_a^{\prime \prime} \} > \frac{\alpha|A|}{\delta}
\right] \leq \delta.
$$

### A.2. Regret due to sampling from $S_a^{\prime}$

For $\theta \in \Theta$, $a \in A$, define $b_{\theta, a} : \mathbb{R}^+ \rightarrow \mathbb{R}$ by

$$
b_{\theta, a}(x) := \begin{cases} 
-\lambda, & x < n^*, \\
(1 - \epsilon) x D(\theta_n^a || \theta_a), & x \geq n^*, 
\end{cases}
$$

where $\lambda$ and $n^*$ satisfy the assertion of Lemma 3. Thus, by Lemma 3, under $G$, and for all $\theta \in \Theta$,

$$
W_t(\theta) \leq e^{-\sum_{a \in A} b_{\theta, a}(N_t(a))} \leq e^{-\sum_{a \in A} b_{\theta, a}(N'_t(a))},
$$

where the last inequality is because $N_t(a) = N'_t(a) + N''_t(a)$, and because $b_{\theta, a}(x)$ is monotone non-decreasing in $x$.

*Note:* In what follows, we assume that $T > 0$ is large enough such that $\log T \geq \frac{\lambda|A|}{\epsilon}$ holds.

We proceed to define the following sequence of non-decreasing stopping times, and associated sets of actions, for the time horizon $1, 2, \ldots, T$.

Let $\tau_0 := 1$ and $A_0 := \emptyset$. For each $k = 1, \ldots, |A| - 1$, let

$$
\tau_k := \min_{\tau_{k-1} \leq t \leq T} \left\{ a \in A_{k-1} \cup \{ a^* \}, \right. \left. \min_{\theta \in S_{a^*}} \sum_{m=1}^{k-1} \frac{N'_t(a_m) D(\theta_{a_m} || \theta_a) + \sum_{a \in A_{k-1}} N'_t(a) D(\theta_{a} || \theta_a)}{1 - \epsilon} \log T. \right\}
$$

(12)

In other words, for each $k$, $A_k$ represents a set of “eliminated” suboptimal actions. $\tau_k$ is the first time after $\tau_{k-1}$, when some suboptimal action (which is not already eliminated) gets eliminated in the sense of satisfying the inequality in (12).
Essentially, the inequality checks whether the condition

\[ \sum_{a \neq a^*} N'_t(a)D(\theta^*_a||\theta_a) \approx \log T \]

is met for all particles \( \theta \in S'_t \) at time \( t \), with a slight modification in that the play count \( N'_t(a) \) is “frozen” to \( N_{\tau_m}(a_m) \) if action \( a \) has been eliminated at an earlier time \( \tau_m \leq t \), and the introduction of the factor \( \frac{1 + \epsilon}{1 - \epsilon} \) to the right hand side.

In case more than one suboptimal action is eliminated for the first time at \( \tau_k \), we use a fixed tie-breaking rule in \( \mathcal{A} \) to resolve the tie. We then put

\[ \mathcal{A}_k := \mathcal{A}_{k-1} \cup \{a_k\} \]

Thus, \( \tau_0 \leq \tau_1 \leq \ldots \leq \tau_{|\mathcal{A}|-1} \leq T \), and \( \mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \ldots \subseteq \mathcal{A}_{|\mathcal{A}|-1} = \mathcal{A} \).

For each action \( a \neq a^* \), by definition, there exists a unique \( \tau_k \) for which \( a \) is first eliminated at \( \tau_k \), i.e., \( \mathcal{A}_k \setminus \mathcal{A}_{k-1} = a \).

Let \( \tau(a) := \tau_k \).

The following lemma states that after an action \( a \) is eliminated, the algorithm does not sample from \( S'_a \) more than a constant number of times.

**Lemma 5.** If \( \log T \geq \lambda|\mathcal{A}| \), then

\[ \Pr \left[ G, \forall k, \sum_{t=\tau_k+1}^{T} 1 \{ \theta_t \in S'_{a_k} \} > \frac{|\mathcal{A}|}{\delta \pi(\theta^*)} \right] \leq \delta. \]

**Proof.** Observe that under \( G \), whenever \( T \geq t > \tau_k \), every \( \theta \in S'_{a_k} \) satisfies

\[
W_t(\theta) \leq \exp \left( -\sum_{a \in \mathcal{A}} b_{\theta,a}(N'_t(a)) \right) \\
\leq \exp \left( -\sum_{a \in \mathcal{A}} \left( (1 - \epsilon)N'_t(a)D(\theta^*_a||\theta_a) - \lambda \right) \right) = \exp \left( -(1 - \epsilon) \sum_{a \in \mathcal{A}} N'_t(a)D(\theta^*_a||\theta_a) + \lambda|\mathcal{A}| \right) \\
\leq \exp \left( -(1 - \epsilon) \sum_{m=1}^{k-1} N'_{\tau_m}(a_m)D(\theta^*_{a_m}||\theta_{a_m}) - (1 - \epsilon) \sum_{a \notin \mathcal{A}_{k-1}} N'_t(a)D(\theta^*_a||\theta_a) + \lambda|\mathcal{A}| \right) \\
\leq \exp \left( -(1 - \epsilon) \frac{1 + \epsilon}{1 - \epsilon} \log T + \epsilon \log T \right) = \frac{1}{T}.
\]

The second inequality above is because the definition of \( b_{\theta,a}(x) \) implies that \( \forall x \geq 0 \ (1 - \epsilon) x D(\theta^*_a||\theta_a) - b_{\theta,a}(x) \leq \lambda \). The penultimate inequality above is due to the fact that for any \( m \leq k \), we have \( \tau_m \leq \tau_k \leq t \), implying that \( N'_{\tau_m}(a_m) \geq N'_{\tau_m}(a_m) \).

We now estimate

\[
\mathbb{E} \left[ 1 \{ t > \tau_k \} 1 \{ \theta_t \in S'_{a_k} \} \mid G \right] = \mathbb{E} \left[ \mathbb{E} \left[ 1 \{ t > \tau_k \} 1 \{ \theta_t \in S'_{a_k} \} \mid G, \mathcal{F}_t \right] \mid G \right] \\
= \mathbb{E} \left[ 1 \{ t > \tau_k \} \pi_t(S'_{a_k}) \mid G \right] = \mathbb{E} \left[ 1 \{ t > \tau_k \} \frac{\int_{S'_{a_k}} W_t(\theta)\pi(\theta)}{\int_{\theta} W_t(\theta)\pi(\theta)} \mid G \right] \\
\leq \mathbb{E} \left[ 1 \{ t > \tau_k \} \frac{T-1}{\pi(\theta^*)} \mid G \right] \leq \frac{T-1}{\pi(\theta^*)},
\]

which implies that

\[
\mathbb{E} \left[ \sum_{t=\tau_k+1}^{T} 1 \{ \theta_t \in S'_{a_k} \} \mid G \right] = \sum_{t=\tau_k+1}^{T} \mathbb{E} \left[ 1 \{ t > \tau_k \} 1 \{ \theta_t \in S'_{a_k} \} \mid G \right] \leq \frac{1}{\pi(\theta^*)}.
\]

Thus,

\[ \Pr \left[ \sum_{t=\tau_k+1}^{T} 1 \{ \theta_t \in S'_{a_k} \} > \frac{1}{\delta \pi(\theta^*)} \mid G \right] \leq \delta. \]
Replacing \( \delta \) by \( \frac{\delta}{|A|} \) and taking a union bound over \( k = 1, 2, \ldots, |A| - 1 \) proves the lemma.

Now we bound the number of plays of suboptimal actions under the event

\[
H := G \cap \left\{ \exists a \neq a^* \sum_{t=1}^{\infty} 1\{\theta_t \in S''_a\} \leq \frac{\alpha|A|}{\delta} \right\} \cap \left\{ \forall k \sum_{t=\tau_k+1}^{T} 1\{\theta_t \in S'_{a_k}\} \leq \frac{|A|}{\delta \pi(\theta^*)} \right\},
\]

which, according to the results of Theorem 3, Lemma 4 and Lemma 5, occurs with probability at least \( 1 - (\delta \sqrt{2} + 2\delta) \). Under the event \( H \), we have

\[
\sum_{a \neq a^*} N'_T(a) = \sum_{k=1}^{|A|-1} N'_T(a_k)
\]

\[
= \sum_{k=1}^{|A|-1} N'_T(a_k) + \sum_{k=1}^{|A|-1} (N'_T(a_k) - N'_T(a_k))
\]

\[
= \sum_{k=1}^{|A|-1} N'_T(a_k) + \sum_{k=1}^{|A|-1} \sum_{t=\tau_k+1}^{T} 1\{\theta_t \in S'_{a_k}\}
\]

\[
\leq \sum_{k=1}^{|A|-1} N'_T(a_k) + \frac{|A|^2}{\delta \pi(\theta^*)}.
\]

**Lemma 6.** Under \( H \), \( \sum_{k=1}^{|A|-1} N'_T(a_k) \leq C_T \), where \( C_T \) solves

\[
C(\log T) := \max \sum_{k=1}^{|A|-1} z_k(a_k)
\]

s.t. \( z_k \in \mathbb{Z}_{+}^{|A|-1} \times \{0\}, a_k \in A \setminus \{a^*\}, 1 \leq k \leq |A| - 1, \)

\( z_i \geq z_k, z_i(a_k) = z_k(a_k), i \geq k, \)

\( \forall 1 \leq j, k \leq |A| - 1 : \)

\( \min_{\theta \in S'_{a_k}} \langle z_k, D_h \rangle \geq \frac{1+\epsilon}{1-\epsilon} \log T, \)

\( \min_{\theta \in S'_{a_k}} \langle z_k - c^{(j)}, D_h \rangle < \frac{1+\epsilon}{1-\epsilon} \log T. \)

**Proof.** With regard to the definition of the \( \tau_k \) and \( a_k \) in (12), if we take

\( a_k = a_k, \ 1 \leq k \leq |A| - 1, \)

and

\( z_k(a) = \begin{cases} N'_T(a), & \tau(a) \leq \tau_k, \\ N''_T(a), & \tau(a) > \tau_k, \end{cases} \)

then it follows, from (12), that the \( z_k \) and \( a_k \) satisfy all the constraints of the optimization problem (13). We also have

\( \sum_{k=1}^{|A|-1} z_k(k) = \sum_{k=1}^{|A|-1} N'_T(a_k). \) This proves the lemma.

**B. Proof of Corollary 1**

The optimal action (in this case a subset) is \( a^* = \{N - M + 1, \ldots, N\} \). It can be checked that the assumptions 1-3 are verified, thus the bound (3) applies and we will be done if we estimate \( C(\log T) \).

The essence of the proof is to first partition the space of suboptimal actions (subsets) according to the least-index basic arm that they contain, i.e., for \( i = 1, 2, \ldots, N - M, \) let

\( A_i := \{ a \subset [N] : a \neq a^*, \min \{ j \in a \} = i \} \)
be all the actions whose least-index (or “weakest”) arm is \( i \).

Take any sequence \( \{ z_k \}_{k=1}^{\lfloor |A| - 1 \rfloor} \), \( \{ a_k \}_{k=1}^{\lfloor |A| - 1 \rfloor} \) feasible for (3). Fix \( 1 \leq i \leq N - M \) and consider the sum \( \sum_{k:a_k \in A_i} z_k(a_k) \).

We claim that this does not exceed \( 1 + \left( \frac{1}{1 - \epsilon} \right) \frac{1}{D(\mu_i || \mu_{N-M+1})} \log T \). If, on the contrary, it does, then put \( \hat{k} := \max \{ k : a_k \in A_i \} \). Take any model \( \theta \in S'_{a_i} \). We must have \( D(\mu_\hat{k} || \theta) = 0 \). Since the KL divergence due to observing a tuple of \( M \) independent rewards is simply the sum of the \( M \) individual (binary) KL divergences, we get that \( \theta_j = \mu_j \) for all \( j \geq N - M + 1 \). However, the optimal action for \( \theta \) is \( a_i^* \) containing the basic arm \( i \). Hence, we get that \( \theta_i \geq \mu_{N-M+1} \geq \mu_i \), which implies that \( D(\mu_i || \theta_i) \geq D(\mu_i || \mu_{N-M+1}) \).

It now remains to estimate
\[
(z_k - e^{(k)} \nu(D), \nu(D)) = \sum_{j=1}^{N} \left( \sum_{a_j \in a} z_k(a) - \delta_{j \in a_k} \right) D(\mu_j || \theta_j)
\]
\[
\geq \left( \sum_{a_j \in a} z_k(a) - 1 \right) D(\mu_j || \theta_j)
\]
\[
\geq \left( \sum_{a_j \in a} z_k(a) - 1 \right) D(\mu_j || \mu_{N-M+1})
\]
\[
= \left( \sum_{k:a_k \in A_i} z_k(a_k) - 1 \right) D(\mu_i || \mu_{N-M+1})
\]
\[
> \log T,
\]
by hypothesis. This violates the final inequality of (3) and yields the desired contradiction. Since the above argument is valid for any \( 1 \leq i \leq N - M \), summing over all such \( i \) completes the proof.

C. Proof of Proposition 2 & Corollary 2

**Lemma 7.** Let \( T \) be large enough such that \( \max_{\theta \in \Theta, a \in A} D(\theta^* || \theta_a) \leq \frac{1 + \epsilon}{1 - \epsilon} \log T \). Then, the optimization problem (3) admits the following upper bound:

\[
(C(\log T) \leq \max_{\text{s.t. } z \in \mathbb{R}^{\lfloor |A| - 1 \rfloor} \times \{0\}, a \in A, a \neq a^*} ||z||_1
\]

\[
\min_{\theta \in S'_{a}} \langle z, D_\theta \rangle \leq \frac{2(1 + \epsilon)}{1 - \epsilon} \log T,
\]

\[
0 \leq z(\hat{a}) \leq \frac{2}{\delta_{\hat{a}}} \left( \frac{1 + \epsilon}{1 - \epsilon} \right) \log T, \quad \forall \hat{a} \in A, \hat{a} \neq a^*.
\]

**Proof.** Take a feasible solution \( \{ z_k, a_k \}_{k=1}^{\lfloor |A| - 1 \rfloor} \) for the optimization problem (3). We will show that \( z = z_{\lfloor |A| - 1 \rfloor} \) and \( a = a_{\lfloor |A| - 1 \rfloor} \) satisfy the constraints (14) above and yield the same objective function value in both optimization problems.

First,
\[
||z||_1 = \sum_{\hat{a} \in A, \hat{a} \neq a^*} z(\hat{a}) = \sum_{k=1}^{\lfloor |A| - 1 \rfloor} z_{\lfloor |A| - 1 \rfloor}(a_k) = \sum_{k=1}^{\lfloor |A| - 1 \rfloor} z_k(a_k),
\]
as \( z_{\lfloor |A| - 1 \rfloor}(a_k) \geq z_k(a_k) \), for all \( k \leq |A| - 1 \), by (3). This shows that the objective functions of both (3) and (14) are equal at \( \{ z_k, a_k \}_{k=1}^{\lfloor |A| - 1 \rfloor} \) and \( (z, a) \) respectively.

\(^{14}\text{This covers all of } A \setminus \{ a^* \} \text{ since every suboptimal set must contain a basic arm of index } N - M \text{ or lesser.}\)
Next, for any \(1 \leq j \leq |A| - 1\) and the unit vector \(e^{(j)}\), we have

\[
\min_{\theta \in S'_a} \langle z, D_\theta \rangle = \min_{\theta \in S'_a} \langle z_k, D_\theta \rangle 
\leq \min_{\theta \in S'_a} \langle z_k - e^{(j)}, D_\theta \rangle + \max_{\theta \in \Theta, a \in A} D(\theta_a^* || \theta_a) 
\leq \frac{1 + \epsilon}{1 - \epsilon} \log T + \frac{1 + \epsilon}{1 - \epsilon} \log T = \frac{2(1 + \epsilon)}{1 - \epsilon} \log T.
\]

This shows that the penultimate constraint in (14) is satisfied. For the final constraint in (14), fix \(1 \leq j \leq |A| - 1\), so that we have

\[
\delta_a, \cdot z(a_j) = \delta_a, \cdot z_j(a_j) \leq \min_{\theta \in S'_a} \langle z_j, D_\theta \rangle \leq \frac{2(1 + \epsilon)}{1 - \epsilon} \log T,
\]

exactly as in the preceding derivation. This implies that \(z(\hat{a}) \leq \frac{2}{\delta_a} \left( \frac{1 + \epsilon}{1 - \epsilon} \right) \log T\) for all \(\hat{a} \neq a^*\).

**Proposition 2.** Let \(T\) be large enough such that \(\max_{\theta \in \Theta, a \in A} D(\theta_a^* || \theta_a) \leq \frac{1 + \epsilon}{1 - \epsilon} \log T\). Suppose

\[
\Delta \leq \min_{a \neq a^*, \theta \in S'_a} D(\theta_a^* || \theta_a).
\]

Suppose also that \(L \in \mathbb{Z}^+\) is such that for every \(a \neq a^*\) and \(\theta \in S'_a\),

\[
|\{\hat{a} \in A : \hat{a} \neq a^*, D(\theta_a^* || \theta_a) \geq \Delta\}| \geq L,
\]

i.e., at least \(L\) coordinates of \(D_\theta\) (excluding the \(|A|\)-th coordinate \(a^*\)) are at least \(\Delta\). Then,

\[
C(\log T) \leq \left( \frac{|A| - L}{\Delta} \right) \frac{2(1 + \epsilon)}{1 - \epsilon} \log T.
\]

**Proof of Proposition 2.** Consider a solution \((z, a)\) to a relaxation of the optimization problem (14) obtained by replacing \(\delta_a\) with \(\Delta\) and \(D_\theta\) with \(D'_\theta := \min(D_\theta, \Delta \cdot 1) \leq D_\theta\)\(^{15}\). We claim that \(|z|_1 \equiv \langle 1, z \rangle \leq \left( \frac{|A| - L}{\Delta} \right) \chi\) where \(\chi := \frac{2(1 + \epsilon)}{1 - \epsilon} \log T\). If not, let \(y = \chi \left( \frac{1}{\Delta}, \ldots, \frac{1}{\Delta}, 0 \right)\), and observe that

\[
\langle D'_\theta, y - z \rangle = \langle D'_\theta, y \rangle - \langle D'_\theta, z \rangle 
\geq \chi \cdot L \cdot \frac{1}{\Delta} - \chi = \chi(L - 1).
\]

But then,

\[
\langle 1, y - z \rangle = \langle 1, y \rangle - \langle 1, z \rangle 
\leq \frac{\chi(|A| - 1)}{\Delta} - \frac{\chi(|A| - L)}{\Delta} = \frac{\chi(L - 1)}{\Delta}

\leq \frac{\Delta \cdot \langle 1, y - z \rangle}{\Delta} = \langle 1, y - z \rangle,
\]

since \(D'_\theta \leq \Delta \cdot 1\) by definition and \(z \leq y\) by hypothesis. This is a contradiction. \(\square\)

**Playing Subsets with Max reward:** Let \(\beta \in (0, 1)\), and suppose that \(\Theta = \{1 - \beta R, 1 - \beta R^{-1}, \ldots, 1 - \beta^2, 1 - \beta\}^N\), for positive integers \(R\) and \(N\). Consider an \(N\) armed Bernoulli bandit with arm parameters \(\mu \in \Theta\). The complex actions are all size \(M\) subsets of the \(N\) basic arms, \(M \leq \frac{N^{\beta - 1}}{2}\). Let \(\mu_{\text{min}} := \min_{a \in A} \prod_{i \in a} (1 - \mu_i)\).

\(^{15}\)Here \(\chi\) represents an all-ones vector of dimension \(\mathcal{A}\), and the minimum is taken coordinatewise. Also, a solution exists since the objective is continuous and the feasible region is compact.
**Proof of Corollary 2.** Since the reward from playing a subset \( a \) is the maximum (equivalently, the Boolean OR) value, the marginal KL divergence along action \( a \) is simply the Bernoulli KL divergence for the product of the parameters:

\[
D(\theta^*_a || \theta_a) = D(\mu_a || \theta_a) = D \left( \prod_{i \in a} (1 - \mu_i) \right| \prod_{i \in a} (1 - \theta_i))
\]

Let us estimate

\[
\Delta := \min \{ D(\mu_a || \theta_a) : \theta \in \Theta, a \in A, D(\mu_a || \theta_a) > 0 \}
\]

If \( \mu_i = 1 - \beta r_i \) and \( \theta_i = 1 - \theta^* s_i \) for integers \( r_i, s_i, i = 1, 2, \ldots, N \), then Pinsker’s inequality yields

\[
D(\mu_a || \theta_a) \geq \frac{2}{\log 2} \left( \prod_{i \in a} (1 - \mu_i) - \prod_{i \in a} (1 - \theta_i) \right)^2
\]

\[
= \frac{2}{\log 2} \left( \beta \sum_{i \in a} r_i - \beta \sum_{i \in a} s_i \right)^2
\]

\[
= \frac{2}{\log 2} \beta^2 \sum_{i \in a} r_i \left( 1 - \beta \sum_{i \in a} s_i - \sum_{i \in a} r_i \right)^2.
\]

\( D(\mu_a || \theta_a) > 0 \) if and only if \( | \sum_{i \in a} s_i - \sum_{i \in a} r_i | \geq 1 \). This implies, together with the above, that

\[
\Delta \geq \frac{2 \mu_{\min}^2 (1 - \beta)}{\log 2}.
\]

Next, we claim that for any \( \mu \neq \theta \in \Theta, D(\mu_a || \theta_a) > 0 \) for at least \( L = \binom{N-1}{M-1} \) size \( M \) subsets/actions \( a \). This is because otherwise, then \( \sum_{i \in a} r_i = \sum_{i \in a} s_i \) for at least \( \binom{N}{M} - L = \binom{N}{M} - \binom{N-1}{M-1} + 1 = \binom{N-1}{M-1} + 1 \) subsets \( a \). However, a combinatorial result (Ahlswede et al., 2003) states that the maximum number of weight \( M \) vertices of the \( N \) dimensional hypercube (in our case, a size \( M \) subset corresponds to a weight \( M \) vertex) that do not span \( N \) dimensions is \( \binom{N-1}{M-1} \). This forces \( r_i = r_i \) for all \( i \in [N] \) and hence \( \mu = \theta \), a contradiction.

Now, we can apply Proposition 2 with \( \Delta \) and \( L \) as above. This gives us that for \( T \) large enough, the total number of arm plays is bounded above, with probability at least \( 1 - \delta \), by

\[
B_3 + (\log 2) \left( \frac{1 + \epsilon}{1 - \epsilon} \right) \left[ \binom{N}{M} - \binom{N - 1}{M - 1} + 1 \right] \frac{\log T}{\mu_{\min}^2 (1 - \beta)}
\]

\[
= B_3 + (\log 2) \left( \frac{1 + \epsilon}{1 - \epsilon} \right) \left[ \binom{N - 1}{M} + 1 \right] \frac{\log T}{\mu_{\min}^2 (1 - \beta)}.
\]

\( \square \)