Table 1: Major notation

<table>
<thead>
<tr>
<th>symbol</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>number of the arms</td>
</tr>
<tr>
<td>$T$</td>
<td>number of the rounds</td>
</tr>
<tr>
<td>$B$</td>
<td>number of the batches</td>
</tr>
<tr>
<td>$T_B$</td>
<td>$= T/(B + K - 1)$</td>
</tr>
<tr>
<td>$T'$</td>
<td>$= T - (B + K - 1)K$</td>
</tr>
<tr>
<td>$I(t)$</td>
<td>arm selected at round $t$</td>
</tr>
<tr>
<td>$X(t)$</td>
<td>reward at round $t$</td>
</tr>
<tr>
<td>$J(T)$</td>
<td>recommendation arm at the end of round $T$</td>
</tr>
<tr>
<td>$P$</td>
<td>hypothesis class of $P$</td>
</tr>
<tr>
<td>$Q$</td>
<td>distribution of estimated parameter of $Q$</td>
</tr>
<tr>
<td>$P \in \mathcal{P}^K$</td>
<td>true parameters</td>
</tr>
<tr>
<td>$P_i \in \mathcal{P}$</td>
<td>$i$-th component of $P$</td>
</tr>
<tr>
<td>$\mathcal{I}^*(P)$</td>
<td>set of best arms under parameter $P$</td>
</tr>
<tr>
<td>$i^*(P)$</td>
<td>one arm in $\mathcal{I}^*(P)$ (taken arbitrary in a deterministic way)</td>
</tr>
<tr>
<td>$Q \in \mathcal{Q}^K$</td>
<td>estimated parameters of $P$</td>
</tr>
<tr>
<td>$Q_i \in \mathcal{Q}$</td>
<td>$i$-th component of $Q$</td>
</tr>
<tr>
<td>$Q_b \in \mathcal{Q}^K$</td>
<td>estimated parameters of $b$-th batch</td>
</tr>
<tr>
<td>$Q_{b,i} \in \mathcal{Q}$</td>
<td>$i$-th component of $Q_b$</td>
</tr>
<tr>
<td>$Q_B \in \mathcal{Q}^K$</td>
<td>stored parameters (in Algorithm 2)</td>
</tr>
<tr>
<td>$Q'_b \in \mathcal{Q}$</td>
<td>$i$-th component of $Q'_b$</td>
</tr>
<tr>
<td>$D(Q|P)$</td>
<td>KL divergence between $Q$ and $P$</td>
</tr>
<tr>
<td>$\Delta_K$</td>
<td>probability simplex in $K$ dimensions</td>
</tr>
<tr>
<td>$r \in \Delta_K$</td>
<td>allocation (proportion of arm draws)</td>
</tr>
<tr>
<td>$r_i \in \Delta_K$</td>
<td>$i$-th component of $r$</td>
</tr>
<tr>
<td>$r_b \in \Delta_K$</td>
<td>allocation at $b$-th batch</td>
</tr>
<tr>
<td>$r_{b,i} \in \Delta_K$</td>
<td>$i$-th component of $r_b$</td>
</tr>
<tr>
<td>$r^b = (r_1, r_2, \ldots, r_b)$</td>
<td></td>
</tr>
<tr>
<td>$n_b$</td>
<td>numbers of draws of Algorithm 2 at $b$-th batch</td>
</tr>
<tr>
<td>$n_{b,i}$</td>
<td>$i$-th component of $n_b$. Note that $n_{b,i} \geq r_{b,i}(T_B - K)$ holds.</td>
</tr>
<tr>
<td>$J(Q_B)$</td>
<td>recommendation arm given $Q_B$</td>
</tr>
<tr>
<td>$(r^B,<em>,J^</em>)$</td>
<td>$\epsilon$-optimal allocation</td>
</tr>
<tr>
<td>$H(\cdot)$</td>
<td>complexity measure of instances</td>
</tr>
<tr>
<td>$R({\pi_T})$</td>
<td>worst-case rate of PoE of sequence of algorithms ${\pi_T}$ in (1)</td>
</tr>
<tr>
<td>$R_{go}$</td>
<td>best possible $R({\pi_T})$ for oracle algorithms in (2)</td>
</tr>
<tr>
<td>$R_{go}^B$</td>
<td>best possible $R({\pi_T})$ for $B$-batch oracle algorithms in (3)</td>
</tr>
<tr>
<td>$R_{g}^\infty$</td>
<td>$\lim_{B \to \infty} R_{B}^g$. Limit exists (Theorem 7)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>model parameter of the neural network</td>
</tr>
<tr>
<td>$r_\theta$</td>
<td>allocation by a neural network with model parameters $\theta$</td>
</tr>
<tr>
<td>$r_{\theta,i}$</td>
<td>$i$-th component of $r_\theta$</td>
</tr>
</tbody>
</table>

A Notation table

Table 1 summarizes our notation.

B Uniform optimality in the fixed-confidence setting

For sufficiently small $\delta > 0$, the asymptotic sample complexity for the FC setting is known.
Namely, any fixed-confidence $\delta$-PAC algorithm require at least $C_{\text{conf}}(P) \log \delta^{-1} + o(\log \delta^{-1})$ samples, where
\[
C_{\text{conf}}(P) = \left( \sup_{r(P) \in \Delta_2} \inf_{P' := P \setminus \{i\}} \sum_{i=1}^{K} r_i D(P_i \| P'_i) \right)^{-1}. \tag{8}
\]

Garivier and Kaufmann (2016) proposed $C$-Tracking and $D$-Tracking algorithms that have a sample complexity bound that matches Eq. (8). This achieveability bound implies that there is no tradeoff between the performances for different instances $P$, and sacrificing the performance for some $P$ never improves the performance for another $P'$. To be more specific, for example, even if we consider a $(\delta$-correct) algorithm that has a suboptimal sample complexity of $2C_{\text{conf}}(P) \log \delta^{-1} + o(\log \delta^{-1})$ for some instance $P$, it is still impossible to achieve sample complexity better than $C_{\text{conf}}(Q) \log \delta^{-1} + o(\log \delta^{-1})$ for another instance $P'$ as far as the algorithm is $\delta$-PAC.

C Suboptimal performance of fixed-confidence algorithms in view of fixed-budget setting

This section shows that an optimal algorithm for the FC-BAI can be arbitrarily bad for the FB-BAI.

For a small $\epsilon \in (0, 0.1)$, consider a three-armed Bernoulli bandit instance with $P^{(1)} = (0.6, 0.5, 0.5 - \epsilon)$ and $P^{(2)} = (0.4, 0.5, 0.5 - \epsilon)$. Here, the best arm is arm 1 (resp. arm 2) in the instance $P^{(1)}$ (resp. $P^{(2)}$).

Let $r_{\text{conf}}(P) = (r_1^{\text{conf}}(P), r_2^{\text{conf}}(P), r_3^{\text{conf}}(P))$ be the optimal FC allocation of Eq. (8). The following characterizes the optimal allocation for $P^{(1)}, P^{(2)}$.

Lemma 8. The optimal solution of Eq. (8) for instance $P^{(1)}$ satisfies the following:
\[
r_1^{\text{conf}}(P^{(1)}), r_2^{\text{conf}}(P^{(1)}), r_3^{\text{conf}}(P^{(1)}) \geq 0.07 = \Theta(1).
\]

Lemma 9. The optimal solution of Eq. (8) for instance $P^{(2)}$ satisfies the following:
\[
r_1^{\text{conf}}(P^{(2)}), r_2^{\text{conf}}(P^{(2)}), r_3^{\text{conf}}(P^{(2)}) = \Theta(\epsilon^2), \Theta(1), \Theta(1).
\]

These two lemmas are derived in Section C.1.

Assume that we run an FC algorithm that draws arms according to allocation $r_{\text{conf}}$ in an FB problem with $T$ rounds. Under the parameters $P^{(2)}$, it draws arm 1 for $O(\epsilon^2) + o(T)$ times. Letting $\delta = P^{(1)}[J(T) = 2], \text{Lemma 1 in Kaufmann et al. (2016)}$ implies that
\[
(TO(\epsilon^2) + o(T)) D(0.4\|0.6) \geq d(P^{(2)}[J(T) = 2], P^{(1)}[J(T) = 2]) \geq d(1/2, P^{(1)}[J(T) = 2]) \quad (\text{assuming the consistency of algorithm})
\]
\[
= \frac{1}{2} \left( \log \left( \frac{1}{23} \right) + \log \left( \frac{1}{2(1-\delta)} \right) \right) \\
\geq \frac{1}{2} \log \left( \frac{1}{23} \right),
\]

which implies
\[
P^{(1)}[J(T) = 2] = \delta \geq \frac{1}{2} \exp \left( -2 \left( \frac{T O(\epsilon^2) + o(T)}{D(0.4\|0.6)} \right) \right). \tag{9}
\]

The exponent of Eq. (9) can be arbitrarily small as $\epsilon \to +0$. In other words, the rate of this algorithm can be arbitrarily close to 0, while the complexity is $H_1(P^{(1)}) = \Theta(1)$. This fact implies that the optimal algorithm for the FC-BAI has an arbitrarily bad performance in terms of the minimax rate of the FB-BAI.
C.1 Proofs of Lemmas 8 and 9

Proof of Lemma 8. For \( r = (1/3, 1/3, 1/3) \), we have

\[
P_i: i^\ast(P) \notin \mathcal{I}^\ast(P_i^{(1)}) \sum_{i=1}^{K} r_i D(P_i^{(1)} || P_i') > \frac{1}{3} \min(D(0.6||0.55), D(0.5||0.55))
\]

(by \( i^\ast(P') \notin \mathcal{I}^\ast(P_i^{(1)}) \) implies \( P_1' < 0.55 \) or \( P_2' > 0.55 \) or \( P_3' > 0.55 \))

\[\geq 1/600.\]

We have

\[
P_i: i^\ast(P') \notin \mathcal{I}^\ast(P_i^{(1)}) \sum_{i=1}^{K} r_i^{\text{conf}}(P_i^{(1)}) D(P_i^{(1)} || P_i') \leq r_1^{\text{conf}}(P_i^{(1)}) D(0.6||0.5)
\]

(on instance \( P' = (0.5, 0.5, 0.5 - \epsilon) \))

\[\leq 0.021 r_1,
\]

which implies \( r_1^{\text{conf}}(P^{(1)}) \geq (1/600) \times (1/0.021) \geq 0.07 \) for the optimal allocation \( r_1^{\text{conf}}(P^{(1)}) \).

Similar discussion yields \( r_2, r_3 \geq 0.07. \)

Proof of Lemma 9. For \( r = (1/3, 1/3, 1/3) \), we have

\[
P_i: i^\ast(P) \notin \mathcal{I}^\ast(P_i^{(2)}) \sum_{i=1}^{K} r_i D(P_i^{(2)} || P_i')
\]

\[
> \frac{1}{3} \min(D(0.5||0.5 - \epsilon/2), D(0.5 - \epsilon||0.5 - \epsilon/2)),
\]

(by \( P' \notin \mathcal{I}^\ast(P_i^{(2)}) \) implies \( P_2' < 0.5 - \epsilon/2 \) or \( P_1' > 0.5 - \epsilon/2 \) or \( P_3' > 0.5 - \epsilon/2 \))

\[\geq \frac{\epsilon^2}{6}.
\]

(by Pinsker’s inequality)

We have

\[
P_i: i^\ast(P') \notin \mathcal{I}^\ast(P_i^{(2)}) \sum_{i=1}^{K} r_i^{\text{conf}}(P_i^{(2)}) D(P_i^{(2)} || P_i') \leq r_2^{\text{conf}}(P_i^{(2)}) D(0.5||0.5 - \epsilon/2),
\]

(on instance \( P' = (0.4, 0.5 - \epsilon/2, 0.5 - \epsilon/2) \))

which implies \( r_2^{\text{conf}}(P_i^{(2)}) = \Omega(1) \) for the optimal allocation. Similar discussion yields \( r_3^{\text{conf}}(P_i^{(2)}) = \Omega(1) \).

In the rest of this proof, we show \( r_1^{\text{conf}}(P_i^{(2)}) = O(\epsilon^2) \). For the ease of exposition, we drop \( P_i \) to denote \( P : (r_1^{\text{conf}}, r_2^{\text{conf}}, r_3^{\text{conf}}) \). Lemma 4 in Garivier and Kaufmann (2016) states that the optimal solution satisfies:

\[
(r_2^{\text{conf}} + r_1^{\text{conf}}) I_{r_2^{\text{conf}} + r_1^{\text{conf}}} (P_2^{(2)}, P_1^{(2)}) = (r_2^{\text{conf}} + r_3^{\text{conf}}) I_{r_2^{\text{conf}} + r_3^{\text{conf}}} (P_2^{(2)}, P_3^{(2)}),
\]

where

\[
I_\alpha(P_2^{(2)}, P_1^{(2)}) = \alpha D\left(P_2^{(2)}, \alpha P_1^{(2)} + (1 - \alpha)P_1^{(2)}\right) + (1 - \alpha) D\left(P_1^{(2)}, \alpha P_1^{(2)} + (1 - \alpha)P_1^{(2)}\right).
\]

We can confirm that

\[
(r_2^{\text{conf}} + r_3^{\text{conf}}) I_{r_2^{\text{conf}} + r_3^{\text{conf}}} (P_2^{(2)}, P_3^{(2)}) = \Theta(1) \times \Theta(\epsilon^2),
\]

15
and

\[(r_2^{\text{conf}} + r_1^{\text{conf}}) \geq r_2^{\text{conf}} = \Theta(1),\]

which, combined with Eq.(10), implies that

\[I_{\frac{r_2^{\text{conf}}}{r_2^{\text{conf}} + r_1^{\text{conf}}}}(P_2^{(2)}, P_1^{(2)}) = \Theta(\epsilon^2),\]

which implies \(r_1^{\text{conf}} = \Theta(\epsilon^2).\)

\[\square\]

**D Extension to wider models**

In the main body of the paper, we assumed that \(P \in \mathcal{P}\) and \(Q \in \mathcal{Q}\) are Bernoulli or Gaussian distributions. Many parts of the results of the paper can be extended to exponential families or distributions over a support set \(S \subset \mathbb{R}\).

Let us consider an exponential family of form

\[dP(x|\theta) = \exp(\theta^T T(x) - A(\theta)) dF(x),\]

where \(F\) is a base measure and \(\theta \in \Theta \subset \mathbb{R}^d\) is a natural parameter. We assume that \(A'(\theta) = \mathbb{E}_{X \sim F(\theta)}[T(X)]\) has the inverse \((A')^{-1} : \text{im}(T) \rightarrow \Theta\), where \(\text{im}(T)\) is the image of \(T\).

Let \(\mathcal{P}\) be a class of reward distributions. \(\mathcal{P}\) can be the family of distributions over a known support \(S \subset \mathbb{R}\). We can also consider the case where \(\mathcal{P}\) is the above exponential family with a possibly restricted parameter set \(\Theta' \subset \Theta\). For example, \(\mathcal{P}\) can be the set of Gaussian distributions with mean parameters in \([0, 1]\) and variances in \((0, \infty)\).

When we derive the lower bounds and construct algorithms, we introduce \(\mathcal{Q}\) as a class of distributions corresponding to the estimated reward distributions of the arms. We set \(\mathcal{Q} = \mathcal{P}\) when \(\mathcal{P}\) is a family of distributions over a known support \(S \subset \mathbb{R}\). When we consider a natural exponential family with parameter set \(\Theta' \subset \Theta\), we set \(\mathcal{Q}\) as this exponential family with parameter set \(\Theta',\) so that the estimator of \(P_1\) is always within \(\mathcal{Q}\). For example, if we consider \(\mathcal{P}\) as a class of Gaussians with means in \([0, 1]\) and variances in \((0, \infty)\), \(\mathcal{Q}\) is the class of all Gaussians with means in \((-\infty, \infty)\) and variances in \((0, \infty)\).

In Algorithm 2, we use a convex combination of distributions \(Q\) and \(Q'\). The key property used in the analysis is the convexity of KL divergence between distributions. When we consider the family \(\mathcal{P}\) of distributions over support set \(S\), the convexity

\[D(\alpha Q + (1 - \alpha)Q' || P) \leq \alpha D(Q || P) + (1 - \alpha) D(Q' || P)\]

holds for any \(P, Q, Q' \in \mathcal{Q}\) when we define \(\alpha Q + (1 - \alpha)Q'\) as the mixture of \(Q\) and \(Q'\) with weight \(\alpha, 1 - \alpha\). When \(\mathcal{P}\) is the exponential family, the convexity of the KL divergence holds when \(\alpha Q + (1 - \alpha)Q'\) is defined as the distribution in this family such that the expectation of the sufficient statistics \(T(X)\) is equal to \(\alpha \mathbb{E}_{X \sim Q}[T(X)] + (1 - \alpha) \mathbb{E}_{X \sim Q'}[T(X)]\). Note that this corresponds to taking the convex combination of the empirical means when we consider Bernoulli distributions or Gaussian distributions with a known variance.

By the convexity of the KL divergence, most parts of the analysis apply to \(\mathcal{P}\) in this section and we straightforwardly obtain the following result.

**Proposition 10.** Theorems 1 and 2, Corollary 3, and Lemma 4 hold under the models \(\mathcal{P}\) with the definition of the convex combination in this section.

The only part where the analysis is limited to Bernoulli or Gaussian is Theorem 5 on the PoE upper bound of the DOT algorithm. The subsequent results immediately follow if Theorem 5 is extended to the models in this section. Since the key property of the DOT algorithm in Lemma 4 on the trackability of the empirical divergence is still valid for these models, we expect that Theorem 5 can also be extended though it remains as an open question.

**E Computational resources**

We used a modern laptop (Macbook Pro) for learning \(\theta\). It took less than one hour to learn \(\theta\). For conducting a large number of simulations (i.e., Run TNN and existing algorithms for
We adopt the formulation of random rewards such that every
10^5 times), we used a 2-CPU Xeon server of sixteen cores. It took less than twelve hours to complete simulations. We did not use a GPU for computation.

F Implementation details

To speed up computation, the same Q was used for each P with the same optimal arm i^*(P)
in the mini-batches.

The final model θ of the neural network is chosen as follows. We stored sequence of models
\theta^{(1)}, \theta^{(2)}, \ldots during training (Algorithm 3). Among these models, we chose the one with
the maximum objective function arg\ max_{P,Q \in (P_{emp}, Q_{emp})} \text{E}(P,Q,\theta^{(l)})$. Here, the
minimum is taken over a finite dataset of size \mid P_{emp} \mid = 32 and \mid Q_{emp} \mid = 10^5.

The black lines in Figure 1 (a)–(c) representing
which is polynomial in

\text{G Proofs}

G.1 Proofs of Theorems 1

In this section, we prove Theorem 1. This theorem as well as its proof is a special case of
Theorem 2, but we solely prove Theorem 1 here since it is easier to follow.

In this proof, we write candidates of the true distributions and empirical distributions by
P = (P_1, P_2, \ldots, P_K) and Q = (Q_1, Q_2, \ldots, Q_K), respectively. In this Sections G.1 and
G.2, we write P[A] and Q[A] to denote the probability of the event A when the reward of
each arm i follows P_i and Q_i, respectively. The entire history of the drawn arms and
observed rewards is denoted by H = ((I(1), X(1)), (I(2), X(2)), \ldots, (I(T), X(T))). We write
X_{i,m} to denote the reward of the \(n\)-th draw of arm \(i\). We define \(n = (n_1, n_2, \ldots, n_K)\)
and \(r = (r_1, r_2, \ldots, r_K) = n/T\) as the numbers of draws of \(K\) arms and their fractions,
respectively, for which we write \(n(H)\) and \(r(H)\) when we emphasize the dependence on the
history \(H\).

We adopt the formulation of random rewards such that every \(X_{i,m}\), the \(m\)-th reward of arm \(i\)
is randomly generated before the game begins, and if an arm is drawn, then this reward is
revealed to the player. Then \(X_{i,m}\) is well defined even if the arm \(i\) is not drawn \(m\) times.

Fix an arbitrary \(\epsilon > 0\). We define sets of “typical” rewards under \(Q\): we write \(T_{r}(Q)\) to
denote the event such that the rewards (some of which might not be revealed as noted above)
satisfy

\[
\sum_{i=1}^{K} \left| n_i D(Q_i \| P_i) - \frac{1}{n} \log \frac{d Q_i}{d P_i}(X_{i,m}) \right| \leq \epsilon T. \tag{11}
\]

By the strong law of large numbers, \(\lim_{T \to \infty} Q(T_{r}(Q)) = 1\).

Let \(\mathcal{R}_T \subset \Delta_K\) be the set of all possible \(r = n/T\). Since \(n_i \in \{0,1,\ldots,T\}\) we have

\[
|\mathcal{R}_T| \leq (T+1)^K,
\]

which is polynomial in \(T\).

Consider an arbitrary algorithm \(\pi\) and define the “typical” allocation \(r(Q,\pi,\epsilon)\) and decision
\(J(Q,\pi,\epsilon)\) of the algorithm for distributions \(Q\) as

\[
r(Q,\pi,\epsilon) = \arg\ max_{r \in \mathcal{R}_T} Q \left[ r(H) = r|T_{r}(Q) \right],
\]

\[
J(Q,\pi,\epsilon) = \arg\ max_{i \in [K]} Q \left[ J(T) = i \big| r(H) = r(Q,\pi,\epsilon), T_{r}(Q) \right].
\]

Then we have

\[
Q \left[ r(H) = r(Q,\pi,\epsilon) \big| T_{r}(Q) \right] \geq \frac{1}{|\mathcal{R}_T|}, \tag{12}
\]
\[ Q \left[ J(T) = J(Q; \pi, \epsilon) \right| r(H) = r(Q; \pi, \epsilon), T_i(Q) \right] \geq \frac{1}{K}. \] (13)

**Lemma 11.** Let \( \epsilon > 0 \) and algorithm \( \pi \) be arbitrary. Then, for any \( P, Q \) such that \( J(Q; \pi, \epsilon) \neq T^*(P) \) it holds that

\[
\frac{1}{T} \log P[J(T) \notin T^*(P)] \geq - \sum_{i=1}^{K} r_i(Q; \pi, \epsilon) D(Q_i || P_i) - \epsilon - \delta_{P,Q,\epsilon}(T)
\]

for a function \( \delta_{P,Q,\epsilon}(T) \) satisfying \( \lim_{T \to \infty} \delta_{P,Q,\epsilon}(T) = 0. \)

**Proof.** For arbitrary \( Q \) we obtain by a standard argument of a change of measures that

\[
P[J(T) \notin T^*(P)] \geq P[T_i(Q), r(H) = r(Q; \pi, \epsilon), J(T) = J(Q; \pi, \epsilon)]
\]

\[
= P[T_i(Q), r(H) = r(Q; \pi, \epsilon)] | P[J(T) = J(Q; \pi, \epsilon) \ | T_i(Q), r(H) = r(Q; \pi, \epsilon)]
\]

\[
= P[T_i(Q), r(H) = r(Q; \pi, \epsilon)] | Q[J(T) = J(Q; \pi, \epsilon) \ | T_i(Q), r(H) = r(Q; \pi, \epsilon)]
\]

\[
\geq \frac{1}{K} P[T_i(Q), r(H) = r(Q; \pi, \epsilon)] \quad \text{by (13))}
\]

\[
= \frac{1}{K} \mathbb{E}_P \left[ 1[H \in T_i(Q), r(H) = r(Q; \pi, \epsilon)] \right]
\]

\[
= \frac{1}{K} \mathbb{E}_P \left[ 1[H \in T_i(Q), r(H) = r(Q; \pi, \epsilon)] \prod_{t=1}^{T} \frac{dP(t)}{dQ(t)}(X(t)) \right]
\]

\[
\geq \frac{1}{K} \mathbb{E}_P \left[ 1[H \in T_i(Q), r(H) = r(Q; \pi, \epsilon)] \exp \left( -T \sum_{i=1}^{K} r_i(Q; \pi, \epsilon) D(Q_i || P_i) - \epsilon T \right) \right]
\]

\[
= \frac{1}{K} Q \left[ T_i(Q), r(H) = r(Q; \pi, \epsilon) \right] \exp \left( -T \sum_{i=1}^{K} r_i(Q; \pi, \epsilon) D(Q_i || P_i) - \epsilon T \right)
\]

\[
\geq \frac{Q \left[ T_i(Q) \right]}{K |R_T|} \exp \left( -T \sum_{i=1}^{K} r_i(Q; \pi, \epsilon) D(Q_i || P_i) - \epsilon T \right), \quad \text{by (12)}
\]

where (14) holds since \( J(T) \) does not depend on the true distribution \( P \) given the history \( H \).

The proof is completed by letting \( \delta_{P,Q,\epsilon} = \log \frac{Q[H \in T_i(Q)]}{K |R_T|}. \) \( \square \)

**Proof of Theorem 1.** For each \( Q \), let \( r(Q; \{ \pi_T \}, \epsilon), J(Q; \{ \pi_T \}, \epsilon) \) be such that there exists a subsequence \( \{ T_n \} \subset \mathbb{N} \) satisfying

\[
\lim_{n \to \infty} r(Q; \pi_{T_n}, \epsilon) = r(Q; \{ \pi_T \}, \epsilon),
\]

\[
J(Q; \pi_{T_n}, \epsilon) = J(Q; \{ \pi_T \}, \epsilon), \quad \forall n.
\]

Such \( r(Q; \{ \pi_T \}, \epsilon) \in \Delta_K \) and \( J(Q; \{ \pi_T \}, \epsilon) \in [K] \) exist since \( \Delta_K \) and \( [K] \) are compact. By Lemma 11, for any \( J(Q; \{ \pi_T \}, \epsilon) \notin T^*(P) \) we have

\[
\lim_{T \to \infty} \frac{1}{T} \log 1/P[J(T) \notin T^*(P)] \leq \lim_{n \to \infty} \frac{1}{T_n} \log 1/P[J(T_n) \notin T^*(P)]
\]

\[
\leq \sum_{i=1}^{K} r_i(Q; \{ \pi_T \}, \epsilon) D(Q_i || P_i) + \epsilon. \quad (15)
\]

By taking the worst case we have

\[
R(\{ \pi_T \}) = \inf_P H(P) \lim_{T \to \infty} \frac{1}{T} \log 1/P[J(T) \notin T^*(P)]
\]

\[
\leq \inf_{P \in P_K, Q \in Q_K: J(Q; \{ \pi_T \}, \epsilon) \notin T^*(P)} H(P) \sum_{i=1}^{K} r_i(Q; \{ \pi_T \}, \epsilon) D(Q_i || P_i) + \epsilon.
\]
By optimizing \( \{\pi_T\} \) we have
\[
R(\{\pi_T\}) \leq \sup_{\{\pi_T\}} \inf_{P \in \mathcal{P}^K} H(P) \lim \inf_{T \to \infty} \frac{1}{T} \log \frac{1}{P[J(T) \neq I^*(P)]} = \sup_{r(\cdot), J(\cdot)} \sup_{\{\pi_T\} : r(\cdot ; \{\pi_T\}) = r(\cdot)} \inf_{P \in \mathcal{P}^K} H(P) \lim \inf_{T \to \infty} \frac{1}{T} \log \frac{1}{P[J(T) \neq I^*(P)]} \leq \sup_{r(\cdot), J(\cdot)} \sup_{P \in \mathcal{P}^K, Q \in \mathcal{Q}^K : J(Q) \not\in I^*(P)} \inf_{P \in \mathcal{P}^K} H(P) \sup_{K} \sum_{i=1}^{K} r_i(Q) D(Q_i || P_i) + \epsilon \quad \text{(by (15))}
\]
We obtain the desired result since \( \epsilon > 0 \) is arbitrary. \( \square \)

G.2 Proof of Theorem 2

Theorem 2 is a generalization of Theorem 1, and we consider different candidates of empirical distributions depending on the batch.

As in the case of the proof of Theorem 1, we write \( P = (P_1, P_2, \ldots, P_t) \) and \( P[A] \) to denote a candidate of the true distributions and the probability of the event under \( P \). We divide \( T \) rounds into \( B \) batches, and the \( b \)-th batch corresponds to \((t_b, t_b + 1, \ldots, t_{b+1} - 1)\)-th rounds for \( b \in [B] \) and \( t_b = [(b - 1)T/B] + 1 \). The entire history of the drawn arms and observed rewards is denoted by \( \mathcal{H} = ((I(1), X(1)), (I(2), X(2)), \ldots, (I(T), X(T))) \). We write \( X_{b,i} \) to denote the reward of the \( n \)-th draw of arm \( i \) in the \( b \)-th batch. We define \( n_b = (n_{b,1}, n_{b,2}, \ldots, n_{b,K}) \) and \( r = (r_{b,1}, r_{b,2}, \ldots, r_{b,K}) = n_b / T \) as the numbers of draws of \( K \) arms and their fractions in the \( b \)-th batch, respectively, for which we write \( n_b(\mathcal{H}) \) and \( r_b(\mathcal{H}) \) when we emphasize the dependence on the history \( \mathcal{H} \).

We adopt the formulation of the random rewards such that every \( X_{b,i,m} \), the \( m \)-th reward of arm \( i \) in the \( b \)-th batch, is randomly generated before the game begins, and if an arm is drawn then this reward is revealed to the player. Then \( X_{b,i,m} \) is well-defined even if arm \( i \) is not drawn \( m \) times in the \( b \)-th batch.

Fix an arbitrary \( \epsilon > 0 \). We define sets of “typical” rewards under \( Q^B \): we write \( \mathcal{R}_T(Q^B) \) to denote the event such that the rewards (a part of which might be unrevealed as noted above) satisfy
\[
\sum_{i=1}^{K} \left| \left( \frac{n_{b,i} D(Q_{b,i} || P_i)}{P_i} - \sum_{m=1}^{n_{b,i}} \log \frac{dQ_{b,i}}{dP_i}(X_{b,i,m}) \right) \right| \leq cT/B \quad (16)
\]
for any \( b \in [B] \). By the strong law of large numbers, \( \lim_{T \to \infty} Q^B[\mathcal{T}_B(Q^B)] = 1 \), where \( Q^B[\cdot] \) denotes the probability under which \( X_b(t) \) follows distribution \( Q_{b,i} \) for \( t \in \{t_b, t_b + 1, \ldots, t_{b+1} - 1\} \).

Let \( \mathcal{R}_{T,B} \subset (\Delta_K)^B \) be the set of all possible \( r^B(\mathcal{H}) \). Since \( n_{b,i} \in \{0, 1, \ldots, t_{b+1} - t_b\} \) and \( t_{b+1} - t_b \leq T/B + 1 \), we see that
\[
|\mathcal{R}_{T,B}| \leq (T/B + 2)^K,
\]
which is polynomial in \( T \).

Consider an arbitrary algorithm \( \pi \) and define the “typical” allocation \( r^b(Q^b; \pi, \epsilon) \) and decision \( J(Q^b; \pi, \epsilon) \) of the algorithm for distributions \( Q^b = (Q_1, Q_2, \ldots, Q_b) \) as
\[
r_1(Q^1; \pi, \epsilon) = \arg \max_{r \in \mathcal{R}_{T,1}} Q^1[r_1(\mathcal{H}) = r | J_1(Q^B)],
\]
\[
r_b(Q^b; \pi, \epsilon) = \arg \max_{r \in \mathcal{R}_{T,b}} Q^b[r_b(\mathcal{H}) = r | r^{b-1}(\mathcal{H}^{b-1})] = r^{b-1}(Q^{b-1}; \pi, \epsilon), \quad J_b(Q^B), \quad b = 2, 3, \ldots, B,
\]
Then we have
\[
Q^B \left[ r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon) \middle| T_c(Q^B) \right] \geq \frac{1}{|R_{T,B}|},
\]
and
\[
Q^B \left[ J(T) = J(Q^B; \pi, \epsilon) \middle| r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon), T_c(Q^B) \right] \geq \frac{1}{K}.
\]

Lemma 12. Let \( \epsilon > 0 \) and algorithm \( \pi \) be arbitrary. Then, for any \( P, Q^B \) such that \( J(Q^B; \pi, \epsilon) \neq I^*(P) \) it holds that
\[
\frac{1}{T} \log P[J(T) \neq I^*(P)] \geq \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \pi, \epsilon) D(Q_{b,i}||P_i) - \epsilon - \delta_{P, Q^B, \epsilon}(T)
\]
for a function \( \delta_{P, Q^B, \epsilon}(T) \) satisfying \( \lim_{T \to \infty} \delta_{P, Q^B, \epsilon}(T) = 0. \)

Proof. For arbitrary \( Q^B \) we obtain by a standard argument of a change of measures that
\[
P[J(T) \neq I^*(P)]
\]
\[
\geq P[T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon), J(T) = J(Q^B; \pi, \epsilon)]
\]
\[
= P[T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)]
\]
\[
\times P[J(T) = J(Q^B; \pi, \epsilon) \mid T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)]
\]
\[
= P[T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)]
\]
\[
\times Q^B[J(T) = J(Q^B; \pi, \epsilon) \mid T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)]
\]
\[
\geq \frac{1}{K} P[T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)]
\]
\[
= \frac{1}{K} \mathbb{E}_P \left[ 1[\mathcal{H} \in T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)] \right]
\]
\[
= \frac{1}{K} \mathbb{E}_Q^B \left[ 1[T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)] \prod_{b=1}^{B} \prod_{t=t_b}^{t_{b+1}-1} \frac{dP_i(t)}{dQ_{b,i}(t)}(X(t)) \right]
\]
\[
\geq \frac{1}{K} \mathbb{E}_Q^B \left[ 1[\mathcal{H} \in T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon)] \right]
\]
\[
\times \exp \left( -\frac{T}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \pi, \epsilon) D(Q_{b,i}||P_i) - \epsilon T \right)
\]
\[
= \frac{1}{K} Q^B \left[ T_c(Q^B), r^B(\mathcal{H}) = r^B(Q^B; \pi, \epsilon) \right]
\]
\[
\times \exp \left( -\frac{T}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \pi, \epsilon) D(Q_{b,i}||P_i) - \epsilon T \right)
\]
\[
\geq \frac{Q^B[T_c(Q^B)]}{K|R_{T,B}|} \exp \left( -\frac{T}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \pi, \epsilon) D(Q_{b,i}||P_i) - \epsilon T \right),
\]
where (19) holds since \( J(T) \) does not depend on the true distribution \( P \) given the history \( \mathcal{H} \). The proof is completed by letting \( \delta_{P, Q^B, \epsilon} = \log \frac{Q^B[T_c(Q^B)]}{K|R_{T,B}|}. \)

Proof of Theorem 2. For each \( Q^B \), let \( r^B(Q^B; \{\pi_T\}, \epsilon), J(Q^B; \{\pi_T\}, \epsilon) \) be such that there exists a subsequence \( \{T_n\}_n \subset \mathbb{N} \) satisfying
\[
\lim_{n \to \infty} r^B(Q^B; \pi_{T_n}, \epsilon) = r^B(Q^B; \{\pi_T\}, \epsilon),
\]
\[
J(Q^B; \pi_{T_n}, \epsilon) = J(Q^B; \{\pi_T\}, \epsilon), \quad \forall n.
\]
Such \( r^B(Q^B; \{\pi_T\}, \epsilon) \in (\Delta_B)^B \) and \( J(Q^B; \{\pi_T\}, \epsilon) \in [K] \) exist since \((\Delta_B)^B \) and \([K] \) are compact. By Lemma 12, for any \( J(Q^B; \{\pi_T\}, \epsilon) \notin I^*(P) \) we have

\[
\liminf_{T \to \infty} \frac{1}{T} \log 1/P[J(T) \notin I^*(P)] \leq \liminf_{n \to \infty} \frac{1}{T_n} \log 1/P[J(T_n) \notin I^*(P)] 
\leq \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \{\pi_T\}, \epsilon) D(Q_{b,i}||P_i) + \epsilon.
\]

By taking the worst case we have

\[
R(\{\pi_T\}) = \inf_{P} \frac{1}{T} H(P) \liminf_{T \to \infty} \frac{1}{T} \log 1/P[J(T) \notin I^*(P)] 
\leq \inf_{P \in P^K, Q^B \in Q^K_B, J(Q^B; \{\pi_T\} \notin I^*(P)} \frac{H(P)}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \{\pi_T\}, \epsilon) D(Q_{b,i}||P_i) + \epsilon.
\]

By optimizing \( \{\pi_T\} \) we have

\[
R(\{\pi_T\}) \leq \sup_{\{\pi_T\} \in P^K, Q^B \in Q^K_B, J(Q^B; \{\pi_T\} \notin I^*(P)} \frac{1}{T} \log 1/P[J(T) \notin I^*(P)] 
\leq \sup_{P \in P^K, Q^B \in Q^K_B, J(Q^B; \{\pi_T\} \notin I^*(P)} \frac{H(P)}{B} \sum_{b=1}^{B} \sum_{i=1}^{K} r_{b,i}(Q^b; \{\pi_T\}, \epsilon) D(Q_{b,i}||P_i) + \epsilon
\]

(by 20)

We obtain the desired result since \( \epsilon > 0 \) is arbitrary.

\[ \Box \]

G.3 Proof of Corollary 3

Proof of Corollary 3. We have

\[
R^{\text{oo}}_B := \sup_{r^B(Q^B), J(Q^B)} \inf_{Q^B, Q_{1}, Q_{2}, \ldots, Q_{B} \in [K], b \in [B]} \frac{H(P)}{B} \sum_{i \in [K], b \in [B]} r_{b,i} D(Q_{b,i}||P_i)
\]

\[
\leq \sup_{r^B(Q^B), J(Q^B)} \inf_{Q^B} \frac{H(P)}{B} \sum_{i \in [K], b \in [B]} r_{b,i} D(Q_{b,i}||P_i)
\]

(by inf over a subset).

\[
= \sup_{r^B(Q), J(Q)} \inf_{Q, \pi, J(Q) \notin I^*(P)} \frac{H(P)}{B} \sum_{i \in [K]} \left( \frac{1}{B} \sum_{b \in [B]} r_{b,i} \right) D(Q_i||P_i)
\]

(by denoting \( Q = Q_1 = Q_2 = \ldots = Q_B \))

\[
= R^{\text{oo}} \quad \text{(by definition)}.
\]
G.4 Additional lemmas

The following lemma is used to derive the regret bound.

**Lemma 13.** Assume that we run Algorithm 2. Then, for any $B_C \in K, K + 1, \ldots, B$, it follows that

$$\sum_{i, b \in [B_C]} r_{b,i} D(Q_b || P_i) \geq \sum_{i, a \in [B_C - K]} r_{a,i}^* D(Q_a' || P_i) + \sum_{i \in [K]} D(Q_{B_C - K + 1,i} || P_i). \tag{21}$$

\[Proof\ of\ Lemma 13.\] We use induction over $B_C \geq K$. (i) It is trivial to derive Eq. (21) for $B_C = K$. (ii) Assume that Eq. (21) holds for $B_C$. In batch $B_C + 1$, the algorithm draws arms in accordance with allocation $r_{B_C+1} = r_{B_C - K + 1}$. We have,

$$\sum_{i \in [K], b \in [B_C+1]} r_{b,i} D(Q_{b,i} || P_i)$$

$$\geq \sum_{i \in [K], a \in [B_C - K]} r_{a,i}^* D(Q_a' || P_i) + \sum_{i \in [K]} D(Q_{B_C - K + 1,i} || P_i) + \sum_{i} r_{B_C+1,i} D(Q_{B_C+1,i} || P_i)$$

(by the assumption of the induction)

$$= \sum_i \left( \sum_{a \in [B_C - K]} r_{a,i}^* D(Q_a' || P_i) + r_{B_C - K + 1,i}^* D(Q_{B_C - K + 1,i} || P_i) \right) + \sum_i \left( 1 - r_{B_C - K + 1,i}^* \right) D(Q_{B_C - K + 1,i} || P_i)$$

$$+ \sum_i r_{B_C+1,i} D(Q_{B_C+1,i} || P_i)$$

(by definition)

$$= \sum_i \left( \sum_{a \in [B_C - K]} r_{a,i}^* D(Q_a' || P_i) + r_{B_C - K + 1,i}^* D(Q_{B_C - K + 1,i} || P_i) \right) + \sum_i D(Q_{B_C - K + 2,i} || P_i)$$

(by Jensen’s inequality and $Q_{B_C - K + 2,i} = r_{B_C+1,i} Q_{B_C+1,i} + (1 - r_{B_C+1,i}) Q_{B_C - K + 1,i}$)

$$= \sum_i \sum_{a \in [B_C - K + 1]} r_{a,i}^* D(Q_a' || P_i) + \sum_i D(Q_{B_C - K + 2,i} || P_i).$$

\[\square\]

G.5 Proof of Lemma 4

**Proof of Lemma 4.**

$$\sum_{i, b \in [B + K - 1]} r_{b,i} D(Q_b || P_i) \geq \sum_{i, b \in [B - 1]} r_{b,i}^* D(Q_b' || P_i) + \sum_i D(Q_{B,i} || P_i). \tag{21}$$

$$\geq \sum_{i, b \in [B]} r_{b,i}^* D(Q_b' || P_i)$$

$$\geq \frac{B(R_{\beta}^{\text{opt}} - \epsilon)}{H(P)} \quad \text{(by definition of $\epsilon$-optimal solution).}$$
G.6 Proof of Theorem 5

Proof of Theorem 5, Bernoulli rewards. Since the reward is binary, the possible values that $Q_{b,i}$ lie in a finite set

$$V = \left\{ \frac{l}{m} : l \in \mathbb{N}, m \in \mathbb{N}^+ \right\},$$

where it is easy to prove $|V| \leq \frac{T}{(B + K - 1) + 2} \leq \frac{T}{B + 2}$.

We have

$$\mathbb{P}[J(T) = J^*(P)] = \sum_{V_1, \ldots, V_B \in \mathcal{V}} \mathbb{P} \left[ J(T) = J^*(P), \bigcap_{b} \{Q_b = V_b\} \right]$$

$$= \sum_{V_1, \ldots, V_B \in \mathcal{V} : J^*(V_1, \ldots, V_B) = J^*(P)} \mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \right].$$

By using the Chernoff bound, we have

$$\mathbb{P} \left[ Q_{b,i} = V_{b,i} \left| \bigcap_{b' \in \{b-1\}} \{Q_{b'} = V_{b'}\} \right. \right] \leq e^{-\frac{2T}{B+K}D(V_{b,i}||P_i)}, \quad (22)$$

and thus

$$\mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \right]$$

$$= \prod_{b} \mathbb{P} \left[ Q_b = V_b \left| \bigcap_{b' \in \{b-1\}} \{Q_{b'} = V_{b'}\} \right. \right]$$

$$\leq \prod_{b} e^{-\frac{2T}{B+K} \sum_{i,b \in [B+K-1]} r_{b,i} D(V_{b,i}||P_i)} \quad \text{(by Eq. (22))}$$

$$= e^{-\frac{2T}{B+K} \sum_{b,i} r_{b,i} D(V_{b,i}||P_i)}. \quad (23)$$

Furthermore,

$$\mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \right]$$

$$= \mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\}, \sum_{i,b \in [B+K-1]} r_{b,i} D(Q_{b,i}||P_i) \geq \frac{B(R^B_{\infty} - \epsilon)}{H(P)} \right]$$

(by Lemma 4).

$$= \mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \bigg| \sum_{i,b \in [B+K-1]} r_{b,i} D(Q_{b,i}||P_i) \geq \frac{B(R^B_{\infty} - \epsilon)}{H(P)} \right] \bigcap_{b} \{Q_b = V_b\}$$

$$= \mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \bigg| \sum_{i,b \in [B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R^B_{\infty} - \epsilon)}{H(P)} \right] \bigcap_{b} \{Q_b = V_b\}$$

$$= \mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \bigg| \sum_{i,b \in [B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R^B_{\infty} - \epsilon)}{H(P)} \right] \bigcap_{b} \{Q_b = V_b\}$$

$$= \mathbb{P} \left[ \bigcap_{b} \{Q_b = V_b\} \right] \mathbb{E} \left[ 1 \left| \sum_{i,b \in [B+K-1]} r_{b,i} D(V_{b,i}||P_i) \geq \frac{B(R^B_{\infty} - \epsilon)}{H(P)} \right] \bigcap_{b} \{Q_b = V_b\} \right].$$
\[
\leq e^{-\frac{2T'}{\sigma^2 + K}} \sum_{b_i} r_{b_i} D(V_{b_i} || P_i) \mathbb{E} \left[ 1 \left( \sum_{i,b \in [B+K-1]} r_{b_i} D(V_{b_i} || P_i) \geq \frac{B(R_B^{R_B} - \epsilon)}{H(P)} \right) \right]
\]
(by Eq. (23))
\[
= E \left[ e^{-\frac{2T'}{\sigma^2 + K}} \sum_{b_i} r_{b_i} D(V_{b_i} || P_i) \mathbb{1} \left( \sum_{i,b \in [B+K-1]} r_{b_i} D(V_{b_i} || P_i) \geq \frac{B(R_B^{R_B} - \epsilon)}{H(P)} \right) \right]
\leq E \left[ e^{-\frac{2T'}{\sigma^2 + K}} \frac{B(R_B^{R_B} - \epsilon)}{H(P)} \right]
= e^{-\frac{2T'}{\sigma^2 + K}} \frac{B(R_B^{R_B} - \epsilon)}{H(P)}.
\]
(24)

Therefore, we have
\[
P[J(T) \notin \mathcal{I}^*(P)] \\
\leq \sum_{V_1, \ldots, V_B \in V^K} \sum_{b_i} e^{-\frac{2T'}{\sigma^2 + K}} \frac{B(R_B^{R_B} - \epsilon)}{H(P)}
\]
(by Eq. (24))
\[
\leq (T/B + 2)^{2KB} e^{-\frac{2T'}{\sigma^2 + K}} \frac{B(R_B^{R_B} - \epsilon)}{H(P)}.
\]

Here, \(\log((T/B + 2)^{2KB}) = o(T)\) to \(T\) when we consider \(K, B\) as constants.

\[\square\]

**Proof of Theorem 5, Normal rewards.** For the ease of discussion, we assume unit variance \(\sigma = 1\). Extending it to the case of common known variance \(\sigma\) is straightforward. Let
\[
\mathcal{B} = \bigcup_{i,b} \{Q_{b,i} \geq T\}.
\]

Then, it is easy to see
\[
P[\mathcal{B}] = T^{2KB}O(e^{-T^2/2}),
\]
which is negligible because \(\log(1/P[\mathcal{B}])/T\) diverges.

The PoE is bounded as
\[
P[J(T) \notin \mathcal{I}^*(P)] = P[J(T) \notin \mathcal{I}^*(P), \mathcal{B}^c] + P[\mathcal{B}]
\]

We have,
\[
P[J(T) \notin \mathcal{I}^*(P), \mathcal{B}^c]
= \int_{-T}^{T} \cdots \int_{-T}^{T} \mathbb{1}[J(T) \notin \mathcal{I}^*(P)] p(Q_B | Q_{B-1} \ldots Q_1) dQ_B \ldots p(Q_B | Q_{B-1} \ldots Q_1) dQ_b \ldots p(Q_1) dQ_1.
\]
(25)

Here,
\[
p(Q_b | Q_{b-1} \ldots Q_1) = \prod_{i \in [K]} \frac{n_{b,i}}{\sqrt{2\pi}} \exp \left( -\frac{n_{b,i}(Q_{b,i} - P_i)^2}{2} \right)
= \prod_{i \in [K]} \frac{n_{b,i}}{\sqrt{2\pi}} \exp \left( -n_{b,i} D(Q_{b,i} || P_i) \right)
\leq \prod_{i \in [K]} T \exp \left( -n_{b,i} D(Q_{b,i} || P_i) \right).
\]

24
Finally, we have
\[(25) \leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} 1[J(T) \notin \mathcal{I}^*] \prod_{i \in \{K\}} \prod_{b \in [B+K-1]} \exp \left( -n_{b,i} D(Q_{b,i}||P_i) \right) dQ_B \cdots dQ_1 \]
\[ \leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} 1[J(T) \notin \mathcal{I}^*] \prod_{i \in \{K\}} \prod_{b \in [B+K-1]} \exp \left( -\frac{B}{B+K-1} \frac{(R^0_B - \epsilon)T'}{H(P)} \right) dQ_B \cdots dQ_1 \]
\[ \leq T^{BK} \int_{-T}^{T} \cdots \int_{-T}^{T} \exp \left( -\frac{B}{B+K-1} \frac{(R^0_B - \epsilon)T'}{H(P)} \right) dQ_B \cdots dQ_1 \]
\[ \leq T^{BK} (2T)^{BK} \exp \left( -\frac{B}{B+K-1} \frac{(R^0_B - \epsilon)T'}{H(P)} \right). \]

\[ \square \]

G.7 Proof of Theorem 7

Proof of Theorem 7. We first show that the limit
\[ R^0_{\infty} = \lim_{B \to \infty} R^0_B \]
exists. Namely, for any \( \eta > 0 \) there exists \( B_0 \in \mathbb{N} \) such that for any \( B_1 > B_0 \) we have
\[ |R^0_{B_1} - R^0_{B_0}| \leq \eta. \]

Theorem 5 implies that Algorithm 2 with \( B = B_0 \) and \( \epsilon = \eta/2 \) satisfies\(^{15}\)
\[ \liminf_{T \to \infty} \frac{\log(1/P[J(T) \notin \mathcal{I}^*(P)])}{T} \geq \frac{B_0}{B_0 + K - 1} \frac{R^0_{B_0} - \eta/2}{H(P)}, \]
and thus
\[ \inf H(P) \liminf_{T \to \infty} \frac{\log(1/P[J(T) \notin \mathcal{I}^*(P)])}{T} \geq \frac{B_0}{B_0 + K - 1} \left( R^0_{B_0} - \frac{\eta}{2} \right). \]

Moreover, Theorem 2 implies that any algorithm satisfies
\[ \inf H(P) \limsup_{T \to \infty} \frac{\log(1/P[J(T) \notin \mathcal{I}^*(P)])}{T} \leq R^0_{B_1}. \]

Combining Eq. (26) and Eq. (27), we have
\[ \frac{B_0}{B_0 + K - 1} \left( R^0_{B_0} - \frac{\eta}{2} \right) \leq R^0_{B_1} \]
and thus
\[ R^0_{B_0} \leq R^0_{B_1} + \frac{\eta}{2} + \frac{K - 1}{B_0 + K - 1} R^0_{B_0} \]
\[ \leq R^0_{B_1} + \frac{\eta}{2} + \frac{K - 1}{B_0 + K - 1} R^0_{B_0} \quad \text{(by Corollary 3)} \]
\[ \leq R^0_{B_1} + \frac{\eta}{2} + \frac{\eta}{2} \quad \text{(by } K \geq 2, \text{ by taking } B_0 \geq 2KR^0_{B_1}/\eta) \]

\(^{15}\)Strictly speaking, Algorithm 2 depends on \( T \), and we take sequence of the algorithm \( (\pi_{dot,T})_{T=1,2,...} \).
\[ \leq R_{B_1}^{\infty} + \eta. \]

By swapping \( B_0, B_1 \), it is easy to show that
\[ R_{B_1}^{\infty} \leq R_{B_0}^{\infty} + \eta, \]
and thus
\[ |R_{B_0}^{\infty} - R_{B_1}^{\infty}| \leq \eta, \]
which implies that the limit exists. It is easy to confirm that the performance of Algorithm 2 with any \( B \geq 2KR^{\infty}/\eta \) and \( \epsilon = \eta/2 \) satisfies Eq. (6). \( \Box \)