A Proofs from Section 2

Algorithm 4: Output $\hat{\alpha} \in \left[G^{-1}(1-\eta_1) - \frac{\varepsilon}{3}, G^{-1}\left(1-\eta_1+\eta_2\right) + \frac{\varepsilon}{3}\right]$ with probability $1-\frac{\delta}{2}$

input: arm set $\mathcal{S} = (a_1, a_2, \dots)$ and parameters $(\eta_1, \eta_2, \varepsilon, \delta) \in (0,1)$ with $\eta_2 < \eta_1$.
initialize: $K = \frac{C\eta_1 \log(1/\delta)}{\eta_2^2}$.
for $i = 1, 2, \dots, K$ **do**
  Collect $n = \frac{C\log(1/\eta_2)}{\varepsilon^2}$ samples of arm i. Set $\hat{p}_i = \hat{p}_i(n)$ to be the average observed reward.
end
Let $\hat{\alpha}$ be the k-th largest value in $\{\hat{p}_1, \dots, \hat{p}_K\}$ for $k = \lceil K\left(\eta_1 - \frac{\eta_2}{2}\right)\rceil$.
Return $\hat{\alpha}$

We show the following generalization of Proposition 2.1.

Proposition A.1. *Fix $0 \le \eta_1, \eta_2, \varepsilon, \delta \le 1$ with $\eta_2 \le \eta_1$. With probability at least $1-\frac{\delta}{2}$, the output $\hat{\alpha}$ of Alg. 4 satisfies*

$$\hat{\alpha} \in \left[G^{-1}(1-\eta_1) - \frac{\varepsilon}{3}, G^{-1}\left(1-\eta_1+\eta_2\right) + \frac{\varepsilon}{3}\right].$$

Moreover, Alg. 4 has sample complexity

$$O\left(\frac{\eta_1 \log(1/\eta_2)\log(1/\delta)}{\eta_2^2\varepsilon^2}\right).$$

Proof. The sample complexity is clear so we focus on the first statement. First observe that by a Chernoff estimate, for each $i \in [K]$,

$$\mathbb{P}\left[|p_i - \hat{p}_i| \ge \frac{\varepsilon}{3}\right] \le \frac{\eta_2}{8}. \tag{A.1}$$

Let $N(\varepsilon)$ be the number of $i \in [K]$ such that $|p_i - \hat{p}_i| \ge \frac{\varepsilon}{3}$. Applying a second Chernoff estimate (of multiplicative form, see e.g. Theorem 4.5 in [MU17]) on these events as i varies and noting that $K\eta_2 \ge C\log(1/\delta)$, (A.1) implies

$$\mathbb{P}\left[N(\varepsilon) \le \frac{K\eta_2}{6}\right] \ge 1 - \frac{\delta}{8}. \tag{A.2}$$

We next show that with probability at least $1-\frac{\delta}{4}$,

$$\hat{\alpha} \le \overline{\alpha} + \frac{\varepsilon}{3} \equiv G^{-1}\left(1-\eta_1+\eta_2\right) + \frac{\varepsilon}{3}. \tag{A.3}$$

With p_i the (true) mean reward from arm a_i, let

$$N_{\overline{\alpha}} \equiv |\{i \in [K] \;:\; p_i > \overline{\alpha}\}|$$

denote the number of the K tested arms which satisfy $p_i > \overline{\alpha}$. By definition, $N_{\overline{\alpha}}$ is stochastically dominated by a $\mathrm{Bin}\left(K, \eta_1 - \frac{9\eta_2}{10}\right)$ random variable, and $\eta_1 - \frac{3\eta_2}{4} = \Theta(\eta_1)$ since $\eta_2 \le \eta_1$. Note that

$$\eta_1 - \frac{9\eta_2}{10} \asymp \eta_1 - \frac{3\eta_2}{4} \asymp \eta_1,$$
$$\frac{\eta_1 - \frac{9\eta_2}{10}}{\eta_1 - \frac{3\eta_2}{4}} \ge 1 + \frac{\eta_2}{20\eta_1}.$$

Therefore another multiplicative Chernoff estimate implies

$$\mathbb{P}\left[N_{\overline{\alpha}} \le K\left(\eta_1 - \frac{3\eta_2}{4}\right)\right] \ge e^{-\Omega(K\eta_2^2/\eta_1)} \ge 1 - \frac{\delta}{8}.$$

When both $N(\varepsilon) \le \frac{K\eta_2}{6}$ and $N_{\overline{\alpha}} \le K\left(\eta_1 - \frac{3\eta_2}{4}\right)$ hold, it follows by definition that $\hat{\alpha} \le \overline{\alpha} + \frac{\varepsilon}{3}$. Hence recalling (A.2) above, we conclude that

$$\mathbb{P}\left[\hat{\alpha} \le \overline{\alpha} + \frac{\varepsilon}{3}\right] \ge 1 - \frac{\delta}{4},$$

establishing (A.3). The other direction is similar. With $\alpha = G^{-1}(1-\eta_1)$ as usual, we set

$$N_\alpha \equiv |\{i \in [K] : p_i \geq \alpha\}|\,. \tag{A.4}$$

This time, N_α stochastically dominates a $\mathrm{Bin}(K,\eta_1)$ random variable. Yet another Chernoff estimate yields

$$\mathbb{P}\left[N_\alpha \geq K\left(\eta_1 - \frac{\eta_2}{4}\right)\right] \geq 1 - \frac{\delta}{8}.$$

Using (A.2) in the same way as above, we find

$$\mathbb{P}\left[\hat{\alpha} \geq \alpha - \frac{\varepsilon}{3}\right] \geq 1 - \frac{\delta}{4}.$$

This concludes the proof. □

Proof of Theorem 2.1. First we analyze the expected sample complexity. On the event that

$$\hat{\alpha} \in \left[G^{-1}(1-\eta) - \frac{\varepsilon}{3}, G^{-1}\left(1 - \frac{\eta}{2}\right) + \frac{\varepsilon}{3}\right] \tag{A.5}$$

we claim that Alg. 2 terminates with probability $\eta/4$ for each a_i. Indeed, if

$$\hat{p}_i \geq G^{-1}\left(1 - \frac{\eta}{2}\right)$$

then termination always happens by definition. This has probability at least $1/4$ if $p_i \geq G^{-1}\left(1-\frac{\eta}{2}\right)$ by Theorem 1 in [GM14], and the latter condition has probability at least $\eta/2$ by definition. It follows that when (A.5) holds, the expected sample complexity of Alg. 2 is $O\left(\frac{\log(1/\eta\delta)}{\eta\varepsilon^2}\right)$. On the other hand, (A.5) fails to hold with probability less than δ. Because of the explicit termination condition in Alg. 2, this yields a additional sample complexity contribution of smaller order $O\left(\delta\log(1/\delta)\frac{\log(1/\eta\delta)}{\eta\varepsilon^2}\right)$. Finally Alg. 4 has sample complexity

$$O\left(\frac{\log(1/\eta)\log(1/\delta)}{\eta\varepsilon^2}\right)$$

which clearly forms the dominant contribution. This completes the proof of the sample complexity bound and we now turn to proving correctness with probability $1-\delta$. First, it is easy to see that Alg. 4 outputs some arm a_i with probability at least $1-\frac{\delta}{2}$. It therefore suffices to show that for any fixed $\hat{\alpha}$ satisfying (A.5), conditioned on the event $\hat{p}_i \geq \hat{\alpha} - \frac{\varepsilon}{3}$, the conditional probability that $p_i \geq \alpha - \varepsilon$ is at least $1-\frac{\delta}{2}$.

We do this using Bayes' rule. If $p_i \geq G^{-1}(1-\frac{\eta}{2})$, then as above Theorem 1 in [GM14] implies

$$\mathbb{P}\left[\hat{p}_i \geq \hat{\alpha} - \frac{\varepsilon}{3}\right] \geq \mathbb{P}[\hat{p}_i \geq p_i] \geq 1/4.$$

This event hence contributes probability at least $\eta/4$ to the event $p_i \geq G^{-1}(1-\eta)$. On the other hand, if $p_i \leq G^{-1}(1-\eta) - \varepsilon \leq \hat{\alpha} - \frac{2\varepsilon}{3}$, then

$$\mathbb{P}\left[\hat{p}_i \geq \hat{\alpha} - \frac{\varepsilon}{3}\right] \leq \mathbb{P}\left[\hat{p}_i \geq p_i + \frac{\varepsilon}{3}\right] \leq \eta\delta/8$$

for an absolute constant C. Combining these via Bayes' rule implies the desired result. □

B Lower Bound for Fixed Budget

Fixed Budget with Unknown α

Before giving the proof, we give some qualtiative discussion of the role of unknown α. We consider Theorem 3.2 to be a definitive lower bound, since e.g. being given the value of α only makes the result stronger. When α is unknown, it is possible to give an essentially matching algorithm, but more care is required when stating the result. This is inherent and stems from the fact that the value $\alpha = G_\mu^{-1}(1-\eta)$ can be difficult or even impossible to estimate, yet determines the constant $c_{\alpha,\beta}$ in the desired rate.

Let us illustrate the issue by a counterexample. Consider μ_N defined by:

$$\begin{aligned}\mathbb{P}^{p\sim\mu_N}[p=0.4] &= \frac{1}{2}+e^{-10N},\\ \mathbb{P}^{p\sim\mu_N}[p=0.2] &= \frac{1}{2}-e^{-10N}.\end{aligned}\tag{B.1}$$

Similarly define $\tilde{\mu}_N$ by:

$$\begin{aligned}\mathbb{P}^{p\sim\tilde{\mu}_N}[p=0.4] &= \frac{1}{2}-e^{-10N},\\ \mathbb{P}^{p\sim\tilde{\mu}_N}[p=0.3] &= 2e^{-10N},\\ \mathbb{P}^{p\sim\tilde{\mu}_N}[p=0.2] &= \frac{1}{2}-e^{-10N}.\end{aligned}\tag{B.2}$$

Then μ_N and $\tilde{\mu}_N$ are not distinguishable using N samples, yet $G_{\mu}^{-1}(1/2)=0.4$ while $G_{\tilde{\mu}}^{-1}(1/2)=0.3$. Using non-distinguishability it follows that the lower bound of Theorem 3.2 applies to $\tilde{\mu}_N$ with threshold $\alpha=G_{\mu_N}^{-1}(1/2)=0.4$, as opposed to the direct application using $G_{\tilde{\mu}_N}^{-1}(1/2)=0.3$. It is not hard to show using monotonicity of $\frac{1}{\sqrt{x(1-x)}}$ that

$$c_{0.4,0.4-\varepsilon}<c_{0.3,0.3-\varepsilon}$$

for all $\varepsilon\le 0.3$. As a result, it is information-theoretically **impossible** to achieve the rate (3.1) for $\tilde{\mu}_N$ if the target quantile value α is not given. The core reason is that the value $G_{\tilde{\mu}}^{-1}(1/2)=0.3$ is too sensitive to the choice $\eta=1/2$ of quantile.

Fortunately, this issue is more of an annoyance than a real difficulty. It can be fixed in several ways. In Theorems B.1, B.2, and B.3 below we give three concrete formulations under which the guarantee (3.1) can be achieved, as mentioned in the main body.

Theorem B.1. *For fixed $\eta_1,\eta_2,\varepsilon$, there is a sequence $(\mathcal{A}_N)_{N\ge1}$ of N-sample algorithms outputting a_{i^*} such that the following holds for any sequence $(\mu_N)_{N\ge1}$ of reservoir distributions. Letting*

$$\alpha_N=\frac{1}{\eta_1-\eta_2}\cdot\int_{1-\eta_1}^{1-\eta_2}G_{\mu_N}^{-1}(x)dx$$

be a quantile average of μ_N, we have

$$\limsup_{N\to\infty}\frac{(-\log\mathbb{P}[p_{i^*}<\alpha_N-\varepsilon])\cdot\log^2 N}{c_{\alpha_N,\alpha_N-\varepsilon}\,N}\ge 1.\tag{B.3}$$

Theorem B.2. *For fixed η,ε, there is a sequence $(\mathcal{A}_N)_{N\ge1}$ of N-sample algorithms outputting a_{i^*} such that for any sequence of reservoir distributions μ_N satisfying*

$$\alpha_N\equiv G_{\mu_N}^{-1}(1-\eta)\ge\frac{1+\varepsilon}{2},$$

we have

$$\limsup_{N\to\infty}\frac{(-\log\mathbb{P}[p_{i^*}<G_{\mu_N}^{-1}(1-\eta)-\varepsilon])\cdot\log^2 N}{c_{\alpha_N,\alpha_N-\varepsilon}\,N}\ge 1.\tag{B.4}$$

Theorem B.3. *For any fixed $\varepsilon_1>\varepsilon$, there is a sequence $(\mathcal{A}_N)_{N\ge1}$ of N-sample algorithms outputting a_{i^*} such that for any fixed reservoir distribution μ with $\mu^*>\varepsilon$,*

$$\limsup_{N\to\infty}\frac{(-\log\mathbb{P}[p_{i^*}<\mu^*-\varepsilon_1)\cdot\log^2 N}{N}\ge c_{\mu^*,\mu^*-\varepsilon}.\tag{B.5}$$

We emphasize that the rate (3.1) is optimal in all cases since the lower bound of Theorem 3.2 is for an easier problem. The first formulation above may be the most principled choice. The idea is that an averaged quantile depends continuously on μ, and can in fact be estimated by applying Proposition A.1 for several pairs (η_1,η_2) and computing a Riemann sum. The second formulation requires only the mild condition that $\alpha\ge\frac{1+\varepsilon}{2}$ and uses monotonicity of $c_{\alpha,\alpha-\varepsilon}$ on this set. (In other words, if the average reward values p appearing in (B.1), (B.2) were larger than 0.5, there would be no counterexample.) The third formulation allows us to almost send η all the way down to 0. It uses the fact that

$$\mu^*-(\varepsilon_1-\varepsilon)\le G_{\mu}^{-1}(1-\eta')$$

for some $\eta'=\eta'(\mu,\varepsilon_1,\varepsilon)>0$. These results show that (3.1) is achievable even without knowledge of α, up to a choice of technical modification to sidestep the counterexample discussed above.

Remark B.1. *In fact uniformity in (α, β) holds in the following sense. For any sequence $(\alpha_N, \beta_N)_{N\geq 1}$ of pairs with $\min\left(\beta_N, \alpha_N - \beta_N, 1-\alpha_N\right)$ uniformly bounded below, there is a sequence $(\mathcal{A}_N)_{N\geq 1}$ of N-sample algorithms such that for any $\eta \in (0,1)$ and any sequence of reservoir distributions μ_N with $G^{-1}_{\mu_N}(1-\eta) \geq \alpha_N$,*

$$\limsup_{N\to\infty} \frac{(-\log \mathbb{P}[p_{i^*} < \beta_N]) \cdot \log^2 N}{c_{\alpha_N,\beta_N} N} \leq 1. \tag{B.6}$$

This can be shown identically to Theorem 3.1, though we don't give the proof in this generality. It is useful for the reduction arguments in Theorems B.1, B.2, and B.3.

B.1 Preparation for the Proof

Here we prove Theorem 3.2. For any $\alpha, \beta, \eta, \varrho > 0$ we construct a reservoir $\mu = \mu_{\alpha,\beta,\eta,\varrho}$ such that

$$\liminf_{N\to\infty} \frac{(-\log \mathbb{P}^{\mu}[p_{i^*} < \beta]) \cdot \log^2 N}{N} \leq c_{\alpha,\beta} + \lambda(\varrho) \tag{B.7}$$

holds for any sequence of N-sample algorithms $\mathcal{A}_N$, and where $\lim_{\varrho\to 0}\lambda(\varrho) = 0$ for fixed α, β, η.

B.2 Admissible Reservoirs and Bayesian Perspective

In proving Theorem 3.2, we will use reservoir distributions μ of a specific form. Namely, we require each μ to be supported on an interval $[\underline{\gamma}, \overline{\gamma}]$, where

$$0 < \beta - \varrho < \underline{\gamma} < \beta < \alpha < \overline{\gamma} < \alpha + \varrho < 1.$$

In fact we define $\underline{\gamma}, \overline{\gamma}$ explicitly (recall that $\varrho > 0$ is a small constant which we eventually send to 0) by

$$\begin{aligned} \theta(\underline{\gamma}) &= \theta(\beta) - \varrho^2; \\ \theta(\overline{\gamma}) &= \theta(\alpha) + \varrho^2. \end{aligned} \tag{B.8}$$

We say μ is $(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f})$ *admissible* if μ has density $\mu(dx) = f(x)dx$ for a Borel measurable function f and satisfies for constants $0 < \underline{f} < \overline{f} < \infty$,

$$f(x) \in [\underline{f}, \overline{f}], \quad \forall x \in [\underline{\gamma}, \overline{\gamma}].$$

Towards proving Theorem 3.2, we fix throughout this section some $(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f})$ admissible μ such that $G^{-1}_{\mu}(\alpha) = \eta$ holds, for appropriate constants $(\underline{f}, \overline{f})$ depending only on $(\eta, \varepsilon, \alpha, \beta, \underline{\gamma}, \overline{\gamma})$. It is easy to see that this is always possible.

An admissible μ is roughly comparable to the uniform distribution on an interval. Using admissible reservoirs gives each a_i the potential to slowly degrade in observed quality over time. We remark that while it is more convenient to work with reservoirs supported away from the boundaries, i.e. in $[\underline{\gamma}, \overline{\gamma}] \subseteq (0,1)$, we do not expect this to be essential.

It will be helpful throughout this section to take a Bayesian point of view. We treat μ_N as known to $\mathcal{A}_N$, since $\mathcal{A}_N$ is in fact allowed to depend on μ_N. Thus at each time t, each p_i has a posterior probability distribution which we denote by $\mu_{i,t}$. Note that each $\mu_{i,t}$ depends only on $(n_{i,t}, \hat{p}_{i,t})$ and is initialized at $\mu_{i,0} = \mu$. We denote by

$$\boldsymbol{\mu}^t = (\mu_{1,t}, \mu_{2,t}, \dots) \tag{B.9}$$

the sequence of posterior distributions $\mu_{i,t}$. Since arms are independent, $\boldsymbol{\mu}^t$ is the full time-t posterior of the algorithm.

B.3 Batched Algorithms and Adversaries

In pure exploration problems, it is possible to significantly simplify the structure of any algorithm at the cost of a small multiplicative increase in the sample complexity. We carry this out using the notion of a batch-compressed algorithm.

Definition B.1. *Given an increasing sequence $B = (b_1, b_2, \dots)$ of positive integers, an algorithm $\mathcal{A}$ is B-**batch-compressed** if $\mathcal{A}$ can only act by increasing the number of times n_i that a_i has been sampled from b_k to b_{k+1}, so that $n_i \in B$ holds at all times. B is ϱ-**slowly increasing** if*

$$\frac{b_{k+1}}{b_k + 1} \leq 1 + \varrho, \quad \forall k \geq 1.$$

Finally if $\mathcal{A}$ is B-batch-compressed and B is ϱ-slowly increasing, we say that $\mathcal{A}$ is ϱ-batch-compressed.

Unlike the batched algorithms studied in [PRCS16, GHRZ19], batch-compression is only important for us as an analysis technique. Indeed the following proposition shows that it does not fundamentally affect pure exploration algorithms.

Proposition B.2. *If B is ϱ-slowly increasing, then for any N-sample algorithm $\mathcal{A}$, there exists an B-batch-compressed $\lfloor N(1+\varrho) \rfloor$ algorithm $\mathcal{A}'$ with the same output.*

Proof. We show how to simulate $\mathcal{A}$ using the B-batch-compressed $\mathcal{A}'$, assuming that the sequence of rewards for each a_i is fixed. Each time $\mathcal{A}$ samples arm i for the $n_i = (a_k + 1)$-st time for $a_k \in A$, $\mathcal{A}'$ samples arm i until $n_i = a_{k+1}$. Then $\mathcal{A}'$ has all the information of $\mathcal{A}$ at all times, hence can simulate the behavior and output of $\mathcal{A}$. Moreover by the definition of ϱ-slowly increasing, the sample complexity of $\mathcal{A}'$ is larger than that of $\mathcal{A}$ by at most a factor $(1+\varrho)$. ☐

We will use the above with $\varrho \to 0$ slowly as $N \to \infty$. Then the sample complexity increase $1 + \varrho$ is absorbed into the $1 + o(1)$ factor in Theorem 3.2. As a result it suffices to establish (B.7) under the additional assumption that $\mathcal{A}_N$ is ϱ-batch-compressed.

B.4 Fisher Information Distance

Determining the tight constant $c_{\alpha,\beta}$ requires significant care. In particular the adversary must decrease the empirical average rewards $\hat{p}_{i,t}$ at a precise rate depending on $n_{i,t}$. This rate turns out to involve the *Fisher information distance*. For $a, b \in [0,1]$ we define the Fisher information distance $d_F(a,b)$ between a and b to be

$$d_F(a,b) = \left| \int_a^b \frac{dx}{\sqrt{x(1-x)}} \right|.$$

This agrees with the more general Fisher information metric when each $a \in [0,1]$ is identified with the corresponding Bernoulli distribution. We refer the reader to [Nie20] for a survey on information geometry. In short, the Fisher information yields a natural Riemannian metric on families of probability distributions which are parametrized by smooth manifolds. However we will use only elementary properties of d_F.

We parametrize $[0,1]$ using the function $\theta : [0,1] \to [0,\pi]$ defined by

$$\theta(a) = d_F(0,a) = \int_0^a \frac{dx}{\sqrt{x(1-x)}} = \arccos(1-2a). \tag{B.10}$$

In particular,

$$d_F(a,b) = |\arccos(1-2a) - \arccos(1-2b)| \geq 2|a-b|$$

and so $d_F(0,1) = \pi$. The main property of θ that we will use is the resulting differential equation

$$\theta'(a) = \frac{1}{\sqrt{\theta(a)(1-\theta(a))}}. \tag{B.11}$$

In our case, θ^{-1} parametrizes a "constant speed" path through the space of Bernoulli variables, viewing the Fisher information. Correspondingly, our adversary will ensure that $\theta(\hat{p}_i(n_{i,t}))$ decreases linearly in $\log(n_{i,t})$.

B.5 Preliminary Lemmas from Moderate Deviations

Recall that for positive integers a and b, the $\mathrm{Beta}(a,b)$ distribution has probability density function

$$\frac{(a+b-1)!}{(a-1)!(b-1)!}x^{a-1}(1-x)^{b-1}$$

for $x \in [0,1]$. We now recall a moderate deviations principle for the binomial distribution and a central limit theorem for the beta distribution.

Lemma 4 (Theorem 2.2 in [DA92]). *For any $0 < \underline{q} < \overline{q} < 1$ and constant $\varrho > 0$ there exists $\Delta_0(\underline{q}, \overline{q}, \varrho)$ and $M_0(\underline{q}, \overline{q}, \varrho)$ such that the following holds for all $p \in [\underline{q}, \overline{q}]$. For $n \geq n_0()$ sufficiently large and any $\frac{1}{\Delta_0\sqrt{n}} \leq \Delta \leq \Delta_0$ we have*

$$e^{\left(-\frac{\Delta^2}{2p(1-p)}-\varrho\right)n} \leq \mathbb{P}\left[\frac{Bin(n,p)}{n} \leq p-\delta\right] \leq e^{\left(-\frac{\Delta^2}{2p(1-p)}+\varrho\right)n}.$$

Lemma 5 (Lemma A.1 in [MNS16]). *Let $\{a_n\}_{n\geq n_0}$ be a sequence satisfying*

$$\underline{\gamma} \leq \frac{a_n}{n} \leq \overline{\gamma}.$$

Then the $\mathrm{Beta}(n-a_n+1, a_n+1)$ distribution on $[0,1]$ obeys a central limit theorem with mean $\frac{a_n}{n}$ and standard deviation $\sqrt{\frac{(a_n/n)(1-(a_n/n))}{n}}$ in the sense that for any bounded sequence $(w_n)_{n\geq n_0}$ of real numbers and with Φ the normal CDF,

$$\lim_{n\to\infty}\left|\Phi(w_n) - \mathbb{P}^{x\sim \mathrm{Beta}(n-a_n+1,a_n+1)}\left[\left(x-(a_n/n)\right)\cdot\sqrt{\frac{n}{(a_n/n)(1-(a_n/n))}} \leq w_n\right]\right| = 0.$$

In the next two lemmas, we lower bound the probability that $\hat{p}_{i,t}$ changes significantly when the number $n_{i,t}$ of samples for a_i increases by a factor $(1+\varrho)$.

Lemma 6. *Assume μ is $(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f})$-admissible. Suppose that arm i's average reward $\hat{p}_{i,t}$ after $n = n_{i,t}$ samples satisfies*

$$\hat{p}_{i,t} \in [\beta, \overline{\gamma}]. \tag{B.12}$$

Then for $n \geq C(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f}, \beta)$ sufficiently large,

$$\mathbb{P}^{x\sim\mu_{i,n}}\left[x \leq \hat{p}_{i,t}\right] \geq \frac{\underline{f}}{3\overline{f}}. \tag{B.13}$$

Proof. Let $R_{i,t} = n\hat{p}_{i,t}$ be the total reward from arm i so far. The posterior distribution $\mu_{i,t}$ for p_i takes the form

$$\mu_{i,t}(dx) = \frac{x^{R_{i,t}}(1-x)^{n-R_{i,t}}f(x)dx}{\int_{\underline{\gamma}}^{\overline{\gamma}} x^{R_{i,t}}(1-x)^{n-R_{i,t}}f(x)dx}.$$

For $x \in [\underline{\gamma}, \overline{\gamma}]$ we estimate

$$\frac{x^{R_{i,t}}(1-x)^{n-R_{i,t}}f(x)}{\int_{\underline{\gamma}}^{\overline{\gamma}} x^{R_{i,t}}(1-x)^{n-R_{i,t}}f(x)dx} \geq (\underline{f}/\overline{f})\cdot\frac{x^{R_{i,t}}(1-x)^{n-R_{i,t}}}{\int_0^1 x^{R_{i,t}}(1-x)^{n-R_{i,t}}dx}.$$

The right-hand side is the density of a beta variable with parameters $(R_{i,t}+1, n-R_{i,t}+1)$. We conclude that

$$\mathbb{P}^{x\sim\mu_{i,t}}\left[x \in [\underline{\gamma}, \hat{p}_{i,t}]\right] \geq (\underline{f}/\overline{f})\cdot\mathbb{P}^{z\sim \mathrm{Beta}(n-R_{i,t}+1,R_{i,t}+1)}\left[z \in [\underline{\gamma}, \hat{p}_{i,t}]\right]$$

For n sufficiently large, it follows from Lemma 5 and (B.12) that

$$\mathbb{P}^{z\sim \mathrm{Beta}(n-R_{i,t}+1,R_{i,t}+1)}\left[z \in [\underline{\gamma}, \hat{p}_{i,t}]\right] \geq \frac{1}{3}.$$

Therefore $\mathbb{P}^{\mu_{i,t}}[p_i \leq \hat{p}_{i,t}] \geq \frac{1}{3}$, proving (B.13). $\square$

Lemma 7. *Assume μ is $(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f})$-admissible and that (B.12) holds. For $n = n_{i,t}$, let $\tilde{n} \geq 1$ satisfy $|\tilde{n} - \varrho n| \leq 2$. Let*

$$\tilde{p}_i = \frac{R_{i,n+\tilde{n}} - R_{i,n}}{\tilde{n}}$$

be the average reward from the $(n+1)$-th through $(n+\tilde{n})$-th samples of arm i. Then as $n \to \infty$, for any sequence $\Delta_n = \Theta(1/\log n)$,

$$\mathbb{P}^t[\tilde{p}_i \leq \theta^{-1}(\theta(\hat{p}_{i,t}) - \delta)] \geq \exp\left(-\frac{n\varrho\Delta_n^2(1+o_n(1))}{2}\right). \tag{B.14}$$

Proof. Stochastic monotonicity implies that

$$\mathbb{P}\left[\frac{\mathrm{Bin}(\tilde{n}, p)}{\tilde{n}} \leq \theta^{-1}\big(\theta(\hat{p}_{i,t}) - \Delta_n\big)\right]$$

is a decreasing function of $p \in [0,1]$. Combining with Lemma 6, it follows that

$$\begin{aligned}
\mathbb{P}^t[E] &= \int \mathbb{P}\left[\frac{\mathrm{Bin}(\tilde{n}, x)}{\tilde{n}} \leq \theta^{-1}\big(\theta(\hat{p}_{i,t}) - \Delta_n\big)\right] d\mu_{i,t}(x) \\
&\geq \mathbb{P}^{\mu_{i,t}}[p_i \leq \hat{p}_{i,t}] \cdot \mathbb{P}\left[\frac{\mathrm{Bin}(\tilde{n}, \hat{p}_{i,t})}{\tilde{n}} \leq \theta^{-1}\big(\theta(\hat{p}_{i,t}) - \Delta_n\big)\right] \\
&\geq \frac{\underline{f}}{3\overline{f}} \cdot \mathbb{P}\left[\frac{\mathrm{Bin}(\tilde{n}, \hat{p}_{i,t})}{\tilde{n}} \leq \theta^{-1}\big(\theta(\hat{p}_{i,t}) - \Delta_n\big)\right].
\end{aligned}$$

Since θ is smooth with smooth inverse on $[\underline{\gamma}, \overline{\gamma}]$ and $\Delta_n \leq o_n(1)$, we have

$$\begin{aligned}
\hat{p}_{i,t} - \theta^{-1}\big(\theta(\hat{p}_{i,t}) - \Delta_n\big) &= (1 \pm o_n(1))\Delta_n \cdot (\theta^{-1})'\big(\theta(\hat{p}_{i,t})\big) \\
&= \frac{(1 \pm o_n(1)) \cdot \Delta_n}{\theta'(\theta^{-1}(\hat{p}_{i,t}))} \\
&= (1 \pm o_n(1)) \cdot \Delta_n \sqrt{\hat{p}_{i,t}(1 - \hat{p}_{i,t})}.
\end{aligned}$$

The result now follows from Lemma 4, where we absorb the factor $\underline{f}/(3\overline{f})$ into the $o_n(1)$. $\square$

B.6 Proof of Theorem 3.2

Recall the definition (B.8) of $\underline{\gamma}$ and $\overline{\gamma}$. We require $\mathcal{A}$ to be B-batch-compressed for $B = B(N, \varrho)$ containing:

1. All positive integers at most $N^{2\varrho}$.
2. All positive multiples of $\lfloor N^{\varrho} \rfloor$ at most $N^{6\varrho}$.
3. Integers of the form $\lfloor N^{6\varrho}(1+\varrho)^j \rfloor$ for $j \geq 0$.

It is easy to see that B thus defined is ϱ-slowly increasing for any $\varrho > 0$ and N sufficiently large. We denote $b_k = \lfloor N^{6\varrho}(1+\varrho)^k \rfloor$ so that $|b_{k+1} - (1+\varrho)b_k| \leq 2$. (This choice of indexing differs from that of Definition B.1, which will not be used in the sequel.)

We next construct our randomness distorting adversary $\mathbb{A} = \mathbb{A}(N, \varrho)$. For each arm i, the adversary $\mathbb{A}$ acts as follows depending on the current number of samples $n_{i,t}$.

1. If $n_{i,t} \leq N^{2\varrho}$, then $\mathbb{A}$ does nothing.
2. When $N^{2\varrho} \leq n_{i,t} < N^{6\varrho}$ increases by N^{ϱ}, $\mathbb{A}$ declares that the average reward of this batch of N^{ϱ} samples is at most $\overline{\gamma} - N^{-\varrho}$.
3. When $n_{i,t}$ increases from $b_k \geq N^{6\varrho}$ to b_{k+1}:
   (a) If $\hat{p}_i(b_k) > \beta$ holds, then $\mathbb{A}$ declares that

$$\theta(\hat{p}_i(b_{k+1})) \leq \theta(\hat{p}_i(b_k)) - \frac{\varrho(1+10\varrho)d_F(\alpha,\beta)}{\log N}. \tag{B.15}$$

(b) If $\hat{p}_i(b_k) \leq \beta$ holds, then $\mathbb{A}$ declares that

$$\hat{p}_i(b_{k+1}) \leq \beta.$$

4. When the $\mathcal{A}$ chooses the arm a_{i^*} to output, $\mathbb{A}$ declares that $p_{i^*} < \beta$.

Due to step 4, the declarations made by $\mathbb{A}$ ensure that $p_{i^*} < \beta$. Recalling Lemma 4 and Proposition B.2, it remains to show the upper bound

$$\text{strength}(\mathbb{A}) \leq \frac{(c_{\alpha,\beta} + C_* \varrho)N}{\log^2(N)}$$

for a constant $C_* = C_*(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f}, \beta, \alpha)$ independent of ϱ (and N). We show this bound in several parts. Recalling (3.3), we refer to the *cost* of a step above as the contribution to Cost from the corresponding declarations by $\mathbb{A}$. The most important parts are Lemmas 10 and 11, which bound the cost of the main step 3a and form the dominant contribution to Cost. Note that throughout the analysis below, all cost upper bounds hold almost surely and we **assume that all of $\mathbb{A}$'s declarations hold true**.

Lemma 8. *The total cost from step 2 is at most $C_* N^{1-\varrho}$, for $N \geq C(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f}, \beta, \alpha, \varrho)$ sufficiently large.*

Proof. The probability for each such declaration by $\mathbb{A}$ is at least

$$\mathbb{P}[\text{Bin}(N^{2\varrho}, \overline{\gamma}) \leq \overline{\gamma} N^{2\varrho} - N^{\varrho}] \tag{B.16}$$

since $p_i \leq \overline{\gamma}$ almost surely. Recall that a $\text{Bin}(N^{2\varrho}, \overline{\gamma})$ random variable obeys a central limit theorem centered at $\overline{\gamma} N^{2\varrho}$ with standard deviation at least $C(\overline{\gamma}) N^{\varrho}$. Therefore the probability in (B.16) is at least $\frac{1}{3}$ for N is sufficiently large depending on ϱ. Hence each such declaration costs at most C_* for N sufficiently large. Moreover such declarations can occur only $N^{1-\varrho}$ times because each one involves N^{ϱ} samples, and the base algorithm $\mathcal{A}$ is an N-sample algorithm. This completes the proof. $\square$

Lemma 9. *The total cost from step 3b is at most $C_* N^{1-6\varrho}$ as long as $N \geq C(\underline{\gamma}, \overline{\gamma}, \underline{f}, \overline{f}, \varrho)$.*

Proof. It suffices to show that the cost per step 3b declaration is at most C_*. This follows from (B.13) and stochastic monotonicity. $\square$

Lemma 10. *The total cost from step 3a is at most*

$$\frac{N}{\log^2(N)} \cdot (c_{\alpha,\beta} + C_* \varrho + o_N(1)).$$

Proof. We claim that the cost from a single instance of step 3a when increasing from b_k to b_{k+1} samples is at most

$$\left(\frac{(b_{k+1} - b_k)}{\log^2(N)}\right)(c_{\alpha,\beta} + C_* \varrho + o_N(1)).$$

This implies the desired result since $\mathcal{A}_N$ is an N-sample algorithm. Taking $\Delta = (1 + 10\varrho) d_F(\alpha, \beta)/\log(N)$ in Lemma 7, we find that the declared event has probability at least

$$\exp\left(-\frac{(b_{k+1} - b_k)(1 + 10\varrho)^2 d_F(\alpha, \beta)^2 (1 + o_N(1))}{2\log^2(N)}\right) \geq \exp\left(-\frac{(b_{k+1} - b_k)}{\log^2(N)}\big(c_{\alpha,\beta} + C_* \varrho + o_N(1)\big)\right).$$

This implies the desired claim and completes the proof. $\square$

Lemma 11. *For any a_i sampled $b_0 = \lfloor N^{6\varrho} \rfloor$ times, $\hat{p}_i(b_0) \leq \overline{\gamma}$.*

Proof. By definition of $\mathbb{A}$,

$$\begin{aligned}
\hat{p}_i(b_0) &\leq \frac{N^{2\varrho} + (N^{6\varrho} - N^{2\varrho})(\overline{\gamma} - N^{-\varrho})}{N^{6\varrho}} \\
&= \overline{\gamma} - \frac{1}{N^{\varrho}} + \frac{(1 - \overline{\gamma})}{N^{4\varrho}} + \frac{1}{N^{5\varrho}} \\
&\leq \overline{\gamma}.
\end{aligned}$$

In the last step we used the fact that

$$\frac{1}{N^{\varrho}} \geq \frac{(1-\overline{\gamma})}{N^{4\varrho}} + \frac{1}{N^{5\varrho}}$$

for any $\varrho > 0$ if N is sufficiently large. $\square$

Lemma 12. *For $\varrho \in (0, 1/100)$, if $n_{i,t} \geq N^{1-\varrho}$ and the declarations of $\mathbb{A}$ hold, then $\hat{p}_{i,t} \leq \beta$.*

Proof. We analyze the rate at which the adversary forces $\theta(\hat{p}_i(b_k))$ to decrease. From (B.15) and (11) it follows that for k with $b_k \geq N^{1-\varrho}$, we have

$$\begin{aligned}
\theta(\hat{p}_i(b_k)) &\leq \theta(\overline{\gamma}) - \frac{\varrho(1+10\varrho)d_F(\alpha,\beta)\log_{1+\varrho}(N^{1-8\varrho})}{\log N} \\
&= \theta(\overline{\gamma}) - \frac{\varrho(1+10\varrho)(1-8\varrho)d_F(\alpha,\beta)}{\log(1+\varrho)} \\
&\leq \theta(\overline{\gamma}) - (1+\varrho)d_F(\alpha,\beta) \\
&\overset{(B.8)}{<} \theta(\beta).
\end{aligned}$$

Here we used the fact that $\log(1+\varrho) \leq \varrho$ and $(1+10\varrho)(1-8\varrho) \geq 1$ for $\varrho \in (0, 1/100)$. Since θ is increasing, this shows that $\hat{p}_{i,t} = \hat{p}_i(b_k) < \beta$ for $b_k \geq N^{1-\varrho}$, completing the proof. $\square$

Lemma 13. *The cost from step 4 is at most $C_*\big(N^{1-\varrho}+1\big)$.*

Proof. First, if $\hat{p}_{i^*,N} \leq \beta$ then the cost from step 4 is at most C_*. On the other hand if $\hat{p}_{i^*,N} > \beta$, then Lemma 11 implies $n_{i^*,N} \leq N^{1-\varrho}$. Since the prior μ is supported in $[\underline{\gamma}, \overline{\gamma}]$, the likelihood ratio of updates from $N^{1-\varrho}$ samples is almost surely bounded by $e^{C_* N^{1-\varrho}}$. Therefore

$$\begin{aligned}
\mathbb{P}^{x\sim\mu_{i,N}}[x<\beta] &\geq e^{-C_* N^{1-\varrho}}\mathbb{P}^{x\sim\mu}[x<\beta] \\
&\geq e^{-C_* N^{1-\varrho}}\frac{(\beta-\underline{\gamma})\underline{f}}{\overline{f}}.
\end{aligned}$$

This completes the proof. $\square$

We now combine the lemmas above to conclude Theorem 3.1 via (B.7).

Proof of Theorem 3.1. Let C_*' be a larger constant depending on the same parameters. Then by Lemmas 8, 9, and 13, the total cost from Steps 2, 3b, 4 combines to $C_*' N^{1-\varrho}) \leq o_N(N/\log^2 N)$. The main cost contribution of

$$\frac{N}{\log^2 N}(c_{\alpha,\beta} + C_*\varrho + o_N(1)).$$

comes from Lemma 10, and all other terms are of strictly smaller order. We have thus constructed a reservoir sequence $(\mu_N(\varrho))_{N\geq 1}$ satisfying (B.7) for arbitrary $\varrho > 0$, completing the proof. $\square$

C An Optimal Algorithm with Fixed Budget

Here we provide an asymptotically optimal algorithm which establishes Theorems B.1, B.2, and B.3. In the next subsection in which we show how to reduce the other results mentioned to Theorem 3.1 (in which α is given) using Proposition A.1. Our main focus will then be to prove Theorem 3.1.

We will fix $\varrho > 0$ small and construct a sequence of N-sample algorithms $(\mathcal{A}(N,\varrho))$ satisfying the slightly relaxed guarantee

$$\liminf_{N\to\infty} \frac{(-\log(\mathbb{P}^{\mu_N(\varrho)}[p_{i^*} < \beta]))\cdot \log^2 N}{N} \geq c_{\alpha,\beta} - \lambda(\varrho) \tag{C.1}$$

for a (possibly different) function λ satisfying $\lim_{\varrho\to 0}\lambda(\varrho) = 0$ (for fixed α,β,η). Here $(\mu_N)_{N\geq 1}$ is any sequence of reservoir distributions satisfying $G_{\mu_N}^{-1}(1-\eta) = \alpha$. An elementary diagonalization argument then implies Theorem 3.1. Thus it suffices to construct algorithms satisfying (C.1) for any desired $\varrho > 0$.

C.1 Reduction to Known α

We explain why Theorems B.1, B.2, and B.3 all follow from Theorem 3.1 (more precisely, the uniform statement given in Remark B.1). We begin with Theorem B.1, where

$$\alpha_N = \frac{1}{\eta_1 - \eta_2} \cdot \int_{1-\eta_1}^{1-\eta_2} G_{\mu_N}^{-1}(x)dx.$$

Let $J = \lceil \frac{6}{\varepsilon(\eta_1-\eta_2)} \rceil$ and define

$$\eta^{(j)} = \frac{(J-j)\eta_1 + j\eta_2}{J}, \quad j \in [J].$$

It is easy to see that $\eta^{(j+1)} - \eta^{(j)} \leq \eta^{(j)}$ for all j. We next apply Alg. 4 on $(\eta^{(j)}, \eta^{(j+1)} - \eta^{(j)}, \varepsilon', \delta')$ for $0 \leq j \leq J-1$, with:

$$\begin{aligned}
\varepsilon' &= \log^{-1/3}(N), \\
\delta' &= e^{-\frac{10N}{\log^2(N)}}/J.
\end{aligned}$$

This requires sample complexity

$$N_A \leq \frac{C(\eta_1,\eta_2) N \log\log(N)}{\log(N)} \leq o_N(N). \tag{C.2}$$

Let $\hat{\alpha}_j$ be the resulting output. With probability $1 - J\delta$, we have for each $0 \leq j \leq J-1$,

$$\hat{\alpha}_j \in \left[G^{-1}(1-\eta^{(j)}) - \frac{\varepsilon}{3}, G^{-1}\left(1-\eta^{(j+1)}\right) + \frac{\varepsilon}{3}\right]. \tag{C.3}$$

Note that the function G_{μ}^{-1} is increasing and $[0,1]$-valued. Therefore if (C.3) holds for each j, then

$$\left|\frac{1}{J} \cdot \sum_{j=0}^{J-1} \hat{\alpha}_j - \frac{1}{\eta_1-\eta_2} \cdot \int_{1-\eta_1}^{1-\eta_2} G_{\mu_N}^{-1}(x)dx\right| \leq \frac{\varepsilon}{3} + \frac{1}{J} \leq \frac{\varepsilon}{2}.$$

Therefore the estimator

$$\hat{\alpha}_A = \frac{1}{J} \cdot \sum_{j=0}^{J-1} \hat{\alpha}_j$$

satisfies

$$\mathbb{P}\left[\left|\hat{\alpha}_A - \frac{1}{\eta_1-\eta_2} \cdot \int_{1-\eta_1}^{1-\eta_2} G_{\mu_N}^{-1}(x)dx\right| \leq \varepsilon/2\right] \geq 1 - J\delta' = 1 - e^{-\frac{10N}{\log^2(N)}}.$$

Finally, $c_{\alpha,\alpha-\varepsilon} \leq \pi < 10$ for any $\alpha, \varepsilon \in [0,1]$ (see (B.10)). Therefore the $\delta' = e^{-\frac{10N}{\log^2(N)}}$ failure probability above has a negligible contribution in Theorem B.1. It follows that applying Theorem 3.1 with $\alpha = \hat{\alpha}_A$ as above and $N' = N - N_A$ implies Theorem B.1.

We now turn to Theorem B.2, where μ_N is required to satisfy $G_{\mu_N}^{-1}(1-\eta) \geq \frac{1+\varepsilon}{2}$. We run Alg. 4 with parameters

$$\begin{aligned}
\eta_1 &= \eta, \\
\eta_2 &= \log^{-1/3}(N), \\
\varepsilon' &= \log^{-1/3}(N), \\
\delta' &= e^{-\frac{10N}{\log^2(N)}}.
\end{aligned}$$

The sample complexity N_B again satisfies $N_B \leq o(N)$ exactly as in (C.2). Let $\hat{\alpha}_B + \varepsilon'$ be the resulting output. Then with probability at least $1 - e^{-\frac{10N}{\log^2(N)}}$,

$$\hat{\alpha}_B \geq G_{\mu_N}^{-1}(1-\eta) - 2\varepsilon'$$

and so with $\varepsilon'' = \varepsilon - 2\varepsilon'$, we have

$$\hat{\alpha}_B - \varepsilon'' \geq G_{\mu_N}^{-1}(1-\eta) - \varepsilon.$$

Moreover, also with probability at least $1 - e^{-\frac{10N}{\log^2(N)}}$,

$$\hat{\alpha}_B \le G_{\mu_N}^{-1}(1 - \eta + \eta_2).$$

It follows that applying the algorithm of Theorem 3.1 with

$$(N, \alpha, \eta, \varepsilon) = (N - N_B, \hat{\alpha}_B, \eta - \eta_2, \varepsilon - 2\varepsilon')$$

suffices to recover Theorem B.2, since η_2 and ε' tend to 0 as $N \to \infty$. As in our discussion of Theorem B.1 above, the failure probability $e^{-\frac{10N}{\log^2(N)}}$ is negligible compared to the relevant rate in Theorem B.2.

Finally, Theorem B.3 relies on the simple fact

$$\lim_{\eta \to 0} G_\mu^{-1}(1 - \eta) = \mu^{\cdot} \tag{C.4}$$

Recall that $\mu^* \in [0, 1]$ denotes the maximum value in the support of μ. We run Alg. 4 on $(\eta_1, \eta_2, \varepsilon', \delta')$ where:

$$\begin{aligned}
\eta_1 &= \log^{-1/3}(N), \\
\eta_2 &= \eta_1/2, \\
\varepsilon' &= \varepsilon_1 - \varepsilon, \\
\delta' &= e^{-\frac{10N}{\log^2(N)}}.
\end{aligned}$$

It follows from Proposition A.1 that the resulting output $\hat{\alpha}_C + \frac{\varepsilon_1 - \varepsilon}{2}$ is computed using $O\left(\frac{N \log\log(N)}{\log(N)}\right) \le o(N)$ samples as in the previous cases. Moreover for N sufficiently large:

$$\begin{aligned}
\mathbb{P}\left[\hat{\alpha}_C + \frac{\varepsilon_1 - \varepsilon}{2} \ge \mu^* - \frac{\varepsilon'}{3} - o_N(1)\right] &\overset{(C.4)}{\ge} \mathbb{P}\left[\hat{\alpha}_C + \frac{\varepsilon_1 - \varepsilon}{2} \ge G_\mu^{-1}(1 - \eta_1) - \frac{\varepsilon'}{3}\right] \\
&\ge 1 - \delta' \\
&= 1 - e^{-\frac{10N}{\log^2(N)}}.
\end{aligned}$$

Since $\varepsilon_1 > \varepsilon$, this means for $N \ge N_0(\mu, c', \dots)$ large enough,

$$\mathbb{P}\left[\hat{\alpha}_C \ge \mu^* - (\varepsilon_1 - \varepsilon)\right] \ge 1 - e^{-\frac{10N}{\log^2(N)}}.$$

Note that Alg. 4 also ensures that with probability $1 - e^{-\frac{10N}{\log^2(N)}}$,

$$\begin{aligned}
\hat{\alpha}_C \le \mu^* + \frac{\varepsilon'}{3} - \frac{\varepsilon_1 - \varepsilon}{2} &= \mu^* - \frac{\varepsilon_1 - \varepsilon}{6} \\
&\le G_\mu^{-1}(1 - \eta')
\end{aligned}$$

for some $\eta'(\mu, \varepsilon_1, \varepsilon) > 0$. It follows that applying Theorem 3.1 with

$$(N, \alpha, \eta, \varepsilon) = (N - N', \hat{\alpha}_C, \eta', \varepsilon)$$

implies Theorem B.3.

C.2 The Fixed Budget Algorithm

We now present Algorithm 3 for the fixed budget problem (recall the informal discussion in Section 3). Algorithm 3 studies one arm a_i at a time, moving to a_{i+1} if a_i is rejected. Similarly to the previous section, some details are needed while $n_{t,i}$ is small, since large deviation asymptotics may not have kicked in yet. As explained at the start of the section, we choose a small constant $\varrho > 0$. In fact, we will eventually choose small constants

$$0 < \varrho \ll \varrho_1 \ll \varrho_2 \ll \varrho_3 \ll \varrho_4 \ll \varrho_5 \ll 1$$

which all tend to 0 as $\varrho \to 0$. These constants will be defined throughout the proof. More formally, these values can be obtained by choosing $\varrho_5 > 0$ arbitrarily small, then $\varrho_4 > 0$ sufficiently small depending on ϱ_5, and so on.

Algorithm 3 operates in a batch-compressed way, for a sequence $(b_1, b_2, \dots)$ defined as follows:

$$
\begin{aligned}
b_0 &= \lceil \varrho_1 \log^2(N) \rceil, \\
k_0 &= \lceil \log_{1+\varrho} \left(\log^4(N)/b_0 \right) \rceil \\
b_k &= b_0 (1+\varrho)^k, \quad k \le k_0 \\
b_{k_0+j} &= \lceil (1+\varrho)^j b_{k_0} \rceil, \quad j \ge 1 \\
\tau_k &= \alpha - \varrho - \frac{k}{\sqrt{\log N}}, \quad k \le k_0 \\
\tau_{k_0+j} &= \theta(\alpha - 2\varrho) - j \cdot \frac{d_F(\alpha,\beta)\varrho(1-\varrho_2)}{\log N}, \quad j \ge 1.
\end{aligned}
$$

Note in particular that $b_{k_0} \ge \log^4(N)$. We denote by $\hat{p}_{i,t}$ the empirical average reward collected by a_i from its first t samples.

Algorithm 5: Output arm with $p_i \ge \beta$ using N samples with high probability

---

input: an infinite sequence of arms $i = 1, 2, \dots$
initialize: $i = 0$
while *fewer than N samples have been collected* **do**
  $i \leftarrow i + 1$
  Collect b_0 samples of arm i.
  if $\hat{p}_{i,b_0} \le \alpha - \varrho$ **then**
    Reject arm i
  end
  for $k = 1, 2, \dots, k_0$ **do**
    Collect $b_k - b_{k-1}$ samples of arm i for a total of b_k samples.
    if $\hat{p}_{i,b_k} \le \alpha - \varrho - \frac{k}{\sqrt{\log N}}$ **then**
      Reject arm i;
    end
  end
  for $j = 1, 2, \dots$ **do**
    Collect $b_{k_0+j} - b_{k_0+j-1}$ samples of arm i for a total of b_{k_0+j}.
    if $\theta(\hat{p}_{i,b_{k_0+j}}) \le \theta(\alpha - 2\varrho) - j \cdot \frac{d_F(\alpha,\beta)\varrho(1-\varrho_2)}{\log N}$ **then**
      Reject arm i
    end
  end
end
Return arm i.

The role of the values b_j is as follows. When an arm a_i reaches b_k samples for some $k \ge 0$, it is checked for possible rejection by comparing its empirical average reward to the threshold τ_k. Algorithm 3 rejects arm i and moves to arm a_{i+1} if the empirical average $\hat{p}_{i,b_k}$ of arm a_i drops below a moving threshold τ_k. The threshold τ_k begins close to α and gradually decreases until reaching $\beta + \varrho$ by the time $\tau_k \ge \Omega(N)$.

So for, our informal description of Alg. 3 also applies to the algorithm proposed in [GM20]. We now highlight two important differences. The first is that our algorithm is defined more carefully during the "early" phases when an arm has been sampled at most $N^{O(\varrho)}$ times. This is crucial for carrying out a rigorous analysis. The second difference is that in the main phase, we increase the sample size for a given arm in powers of $1+\varrho$ rather than powers of 2, and also move the rejection thresholds τ_k based on the Fisher information distance via the function θ. The latter ingredients allow us to obtain the optimal constant factor.

We begin the analysis of Alg 3 by proving Lemma 3.

Proof of Lemma 3. Let $M_j = \prod_{1 \le i \le j} Y_i$ and observe that M_j^c is a positive supermartingale with $M_0 = 0$. The result follows by Doob's maximal inequality. $\square$

We will apply Lemma 3 in the following way. Let X_i be the number of samples used by arm a_i before rejection, and $I_i \in \{0,1\}$ be the indicator of the event that a_i is ever rejected, even if Algorithm 3 were to continue past time N and sample arm i an infinite number of times. We set

$$Y_i = e^{X_i} \cdot I_i,$$

With M defined from $(Y_i)_{i \geq 1}$ as in Lemma 3, it follows that $\log(M)$ is at most the amount of time spent on eventual rejections before the first eventually accepted arm. Therefore if $\log(M) \leq N(1-\varrho)$, we conclude that the last arm to be studied was sampled at least $N\varrho$ times. Since it was not rejected during that time, we can conclude this arm has $p_i \geq \beta$ with probability $1 - e^{-\Omega_\varrho(N)}$. The main contribution to the failure probability of Algorithm 3 comes from the event $\{M \geq A\}$ above, for suitable A. Correspondingly, the main work will be to verify $\mathbb{E}[Y_i^c] \leq 1$ for suitable c.

Note that $Y_i \in \{0\} \cup [1, \infty)$ almost surely for each i. Therefore a necessary first step in showing $\mathbb{E}[Y_i^c] \leq 1$ is to lower bound $\mathbb{P}[Y_i = 0]$, the probability that Algorithm 3 never rejects a_i. We now give a sufficient lower bound from the event $p_i \geq \alpha$.

Proposition C.1. *Let $x_1, x_2, \ldots$ be an i.i.d. Bernoulli(p) sequence for $p \geq \alpha$, and let $S_k = \sum_{i=1}^k x_i$ and set*

$$\underline{S} = \inf_{k \geq 1} S_k/k.$$

Then $\underline{S} \geq \alpha - \varrho$ holds with probability at least $c(\alpha, \varrho) > 0$. Thus $\mathbb{E}[I_i] \leq 1 - c(\alpha, \varrho)$.

Proof. Since the probability that $\underline{S} \geq \alpha - \varrho$ is increasing in p it suffices to take $p = \alpha$ and show the probability is positive for any $\varrho > 0$. Assume not. Then by restarting the indexing every time $S_k \leq k(\alpha - \varrho)$ holds, we find that

$$\lim \inf_{n \to \infty} S_n/n \leq \alpha - \varrho.$$

This contradicts the strong law of large numbers, thus completing the proof of the first assertion. The second assertion follows since if $S_k/k \geq \alpha - \varrho$ for all k where $x_1, \ldots$ are the rewards of arm i, then arm i will never be rejected by Algorithm 3. $\square$

Based on Proposition C.1 above, to show

$$\mathbb{E}\left[e^{X_i \cdot \frac{c_{\alpha,\beta} - \varrho_3}{\log^2 N}} \cdot I_i\right] \leq 1$$

(which is essentially what we want in light of Lemma 3), it suffices to show that

$$\mathbb{E}\left[\left(e^{X_i \cdot \frac{c_{\alpha,\beta} - \varrho_3}{\log^2 N}} - 1\right) \cdot I_i\right] \leq c(\alpha, \varrho). \tag{C.5}$$

We let $I_i^t = I_i \cdot 1_{X_i = t}$ be the event that arm i was rejected after exactly t steps. Since Alg 3 can only reject after b_j samples, we have

$$I_i = \sum_{j=0}^{\infty} I_i^{b_j}$$

We use this to break the left-hand side of (C.5) into three separate parts and estimate the parts separately. The parts correspond to b_0, b_1 through b_{k_0}, and b_{k_0+1} onward. The first two parts are easier and handled in Subsection C.3 below. The final term is the main contribution and is handled in Subsection C.4.

C.3 Analysis of Algorithm 3 in the Small and Medium Sample Phases

Proposition C.2 bounds the contribution to (C.5) from the *small sample phase*, i.e. the first rejection condition in line 7 of Alg 3.

Proposition C.2. *For any α, ϱ there is $\varrho_1 > 0$ sufficiently small that with b_0 as defined above, and with N sufficiently large,*

$$\mathbb{E}\left[\left(e^{X_i \cdot \frac{c_{\alpha,\beta} - \varrho_3}{\log^2 N}} - 1\right) \cdot I_i^{b_0}\right] \leq c(\alpha, \varrho)/4$$

Proof. It suffices to observe that for fixed α, ϱ and ϱ_1 small and N sufficiently large, we have

$$e^{b_0 \cdot \frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N}} - 1 \leq e^{\varrho_1} - 1 \leq 2\varrho_1.$$

$\square$

Proposition C.3 bounds the contribution to (C.5) from the *medium sample phase*, i.e. the second rejection condition in line 12 of Alg 3.

Proposition C.3. *For any $\alpha, \varrho, \varrho_1$ and for N sufficiently large,*

$$\sum_{k=1}^{k_0} \mathbb{E}\left[\left(e^{X_i \cdot \frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N}} - 1\right) \cdot I_i^{b_k}\right] \leq c(\alpha, \varrho)/4$$

Proof. The event $I_i^{b_k}$ requires $|\hat{p}_{i,b_k} - \hat{p}_{i,b_{k-1}}| \geq \frac{1}{\sqrt{\log N}}$. Hence by a standard Chernoff estimate, regardless of the true reward probability p_i,

$$\mathbb{E}[I_i^{b_k}] \leq e^{-\Omega_{\alpha,\varrho,\varrho_1}(b_k/\log N)}.$$

Since by construction $b_0 \geq \varrho_1 \log^2 N$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(e^{X_i \cdot \frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N}} - 1\right) \cdot I_i^{b_k}\right] &\leq e^{b_k \frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N} - \Omega_{\alpha,\varrho,\varrho_1}(b_k/\log N)} \\
&\leq e^{-\Omega_{\alpha,\varrho,\varrho_1}(\log N)} \\
&= N^{-\Omega_{\alpha,\varrho,\varrho_1}(1)}.
\end{aligned}$$

Since $k_0 \leq O(\log N)$, summing gives the desired conclusion. $\square$

Propositions C.2 and C.3 imply that the total contribution from rejections in the small and medium sample phases is at most $c(\alpha, \varrho)/2$. It remains to analyze the large sample phase in the following subsection.

C.4 Analysis of Algorithm 3 in the Large Sample Phase

Similarly to the previous section, the main part of the analysis concerns the large sample phases b_{k_0+j} for $j \geq 1$. Our goal is to precisely estimate the rejection probability at each time b_{k_0+j}. Note that these estimates should not depend on the true average rewards p_i.

Our approach is based on exchangeability and avoids any consideration of p_i. For a given value j and a large constant $L = L(\varrho)$, consider the sequence of times

$$b_{k_0+j-L},\; b_{k_0+j-L+1},\; \ldots,\; b_{k_0+j}$$

and the associated sequence of empirical average rewards

$$\hat{p}_{i,b_{k_0+j-L}},\; \hat{p}_{i,b_{k_0+j-L+1}},\; \ldots,\; \hat{p}_{i,b_{k_0+j}}. \tag{C.6}$$

It follows from the algorithm description that for $I_i^{b_{k_0+j}}$ to occur, we must have

$$\hat{p}_{i,b_{k_0+j}} - \hat{p}_{i,b_{k_0+j-\ell}} \geq \ell \cdot \frac{d_F(\alpha,\beta)\varrho(1-\varrho_2)}{\log N}, \quad \forall\, 1 \leq \ell \leq L. \tag{C.7}$$

This is clear for $j > L$, but it holds also for $0 \leq j \leq L$ as for N sufficiently large,

$$\alpha - \varrho - \frac{k_0}{\sqrt{\log N}} - L \cdot \frac{d_F(\alpha,\beta)\varrho(1-\varrho_2)}{\log N} \geq \alpha - 2\varrho.$$

By exchangeability, conditioned on the future values $\hat{p}_{i,b_{k_0+j}}, \ldots, \hat{p}_{i,b_{k_0+j-\ell}}$ the law of $\hat{p}_{i,b_{k_0+j-\ell-1}}$ depends only on $\hat{p}_{i,b_{k_0+j-\ell}}$ and is given explicitly by a hypergeometric variable. Recalling that

$R_{i,t} = n_{i,t}\hat{p}_{i,t}$ is the total reward from the first $n_{i,t}$ samples of arm i, $R_{i,b_{k_0+j-\ell-1}}$ has hypergeometric conditional law given by:

$$\mathbb{P}\Big[R_{i,b_{k_0+j-\ell-1}} = k \mid \big(\hat{p}_{i,b_{k_0+j}}, \dots, \hat{p}_{i,b_{k_0+j-\ell}}\big)\Big] = \mathbb{P}\big[R_{i,b_{k_0+j-\ell-1}} = k \mid \hat{p}_{i,b_{k_0+j-\ell}}\big]$$
$$= \frac{\binom{b_{k_0+j-\ell-1}}{k}\binom{b_{k_0+j-\ell}-b_{k_0+j-\ell-1}}{R_{k_0+j-\ell}-k}}{\binom{b_{k_0+j-\ell}}{R_{k_0+j-\ell}}}. \tag{C.8}$$

We will refer to this as the $\mathrm{HyperGeom}\big(b_{k_0+j-\ell}, b_{k_0+j-\ell-1}, R_{k_0+j-\ell}\big)$ distribution. Importantly, this distribution is independent of μ. We exploit this below to control the probability of a given sequence $\big(\hat{p}_{i,b_{k_0+j-L}},\ \hat{p}_{i,b_{k_0+j-L+1}},\ \dots,\ \hat{p}_{i,b_{k_0+j}}\big)$ of empirical average rewards. The following useful result states that hypergeometric variables automatically inherit tail bounds from the corresponding binomial random variables.

Lemma 1 ([LP14, Hoe94])**.** *Fix non-negative integers $A \geq B, C$ and let $X \sim \mathrm{HyperGeom}(A,B,C)$ and $Y \sim \mathrm{Bin}(B, C/A)$. Then for any convex function $f : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)].$$

Lemma 2. *For any $0 < \underline{q} < \overline{q} < 1$ and constants $\varrho > 0$ there exists $\Delta_0(\underline{q}, \overline{q}, \varrho)$ and $N_0(\underline{q}, \overline{q}, \varrho)$ such that the following holds for all $p \in [\underline{q}, \overline{q}]$. For $n \geq n_0$ sufficiently large and $\frac{1}{\Delta_0\sqrt{n}} \leq \Delta \leq \Delta_0$,*

$$\mathbb{P}\left[\frac{\mathrm{HyperGeom}(n(1+\varrho), n, np(1+\varrho))}{n} \leq p - \Delta\right] \leq e^{\left(-\frac{\Delta^2}{2p(1-p)}+\varrho\right)n}.$$

Proof. The corresponding binomial result Lemma 4 is proved in Theorem 2.2 in [DA92] by upper bounding an exponential moment. The same proof applies here by Lemma 1. □

It will be convenient to define a restricted set of *good* sequences $(q_L, q_{L-1},\ \dots, q_0)$. These satisfy the key properties of empirical average reward sequences (C.6) for which $I_i^{b_{k_0+j}}$ holds. We say such a length $L+1$ sequence is good if the following conditions are satisfied:

1. $q_0 \in [\underline{q}, \overline{q}] \subseteq (0,1)$ for constants $0 < \underline{q} < \overline{q} < 1$ depending only on ϱ, L.
2. $$\max_{\ell_1,\ell_2} |q_{\ell_1} - q_{\ell_2}| \leq O\big(1/\sqrt{\log N}\big). \tag{C.9}$$
3. For each $1 \leq \ell \leq L$:
$$\begin{aligned}\theta(q_0) &\leq \theta(\alpha - 2\varrho) - j \cdot \frac{d_F(\alpha,\beta)\varrho(1-\varrho_2)}{\log N}\\ &\leq \theta(\alpha - 2\varrho) - (j-\ell) \cdot \frac{d_F(\alpha,\beta)\varrho(1-\varrho_2)}{\log N}\\ &\leq \theta(q_\ell).\end{aligned}$$

The third condition above is necessary for $I_i^{b_{k_0+j},i} = 1$, and these together imply the first condition. Indeed for fixed $\underline{q}, \overline{q}$ and small $\varrho \in (0, 1/10)$ one always has

$$\frac{\hat{p}_{i,b_{k_0+j-1}}}{\hat{p}_{i,b_{k_0+j}}}, \frac{1-\hat{p}_{i,b_{k_0+j-1}}}{1-\hat{p}_{i,b_{k_0+j}}} \in \big[1-2\varrho, (1-2\varrho)^{-1}\big]$$

for large enough N and any j. Hence it suffices to take $\underline{q} = \beta(1-2\varrho)^L$ and $\overline{q} = 1-(1-\alpha)(1-2\varrho)^L$. With this choice, if

$$\hat{p}_{i,b_{k_0+j-L}},\ \hat{p}_{i,b_{k_0+j-L+1}},\ \dots,\ \hat{p}_{i,b_{k_0+j}}.$$

is **not** good and $I_i^{b_{k_0+j}} = 1$, then the second condition must be the only violated one. The following easy lemma controls the failure probability of the second condition. Recall from (C.8) that conditioning on $\hat{p}_{i,b_{k_0+j}}$ determines the joint conditional law of the previous conditional rewards, regardless of μ.

Lemma 3. *All sequences violating only the second condition* (C.9) *above have probability at most*

$$e^{-\Omega_{L,\varrho}(b_{k_0+j}/\log N)},$$

even after conditioning on an arbitrary value for $\hat{p}_{i,b_{k_0+j}}$.

Proof. The claim follows by an elementary Chernoff estimate for hypergeometric variables, which hold just as for binomial variables by Lemma 1. Indeed the assumption implies that some adjacent difference $|\hat{p}_{i,b_{k_0+j-\ell}} - \hat{p}_{i,b_{k_0+j-\ell+1}}|$ has size $\Omega(1/\sqrt{\log N})$. (Note for applying the Chernoff bound that L is a constant independent of N, and so $b_{k_0+j-L} \geq \Omega_{L,\varrho}(b_{k_0+j})$.) □

We now focus on upper-bounding the probability of any good sequence $(q_L, \ldots, q_0)$ appearing, conditionally on q_0.

Lemma 4. *For any good sequence* $(q_L, q_{L-1}, \ldots, q_0)$ *and* $j \geq 0$,

$$\mathbb{P}\Big[\big(\hat{p}_{i,b_{k_0+j-L}},\ \hat{p}_{i,b_{k_0+j-L+1}},\ \ldots,\ \hat{p}_{i,b_{k_0+j}}\big) = (q_L, q_{L-1}, \ldots, q_0)\ \big|\ p_{i,b_{k_0+j}} = q_0\Big]$$
$$\leq \exp\left(-\frac{(1-O(\varrho))}{2q_0(1-q_0)\varrho}\sum_{\ell=0}^{L-1} b_{k_0+j-\ell}(q_\ell - q_{\ell+1})^2\right).$$

Proof. It suffices to show that

$$\mathbb{P}[\hat{p}_{i,b_{k_0+j-\ell-1}} = q_{\ell+1} \mid q_\ell] \leq \exp\left(-\frac{(1-O(\varrho))}{2q_0(1-q_0)\varrho} b_{k_0+j-\ell}(q_\ell - q_{\ell+1})^2\right)$$

This follows by applying Lemma 2 to the hypergeometric random variable

$$\hat{p}_{i,b_{k_0+j-\ell}} \cdot b_{k_0+j-\ell} - \hat{p}_{i,b_{k_0+j-\ell-1}} \cdot b_{k_0+j-\ell-1} = R_{b_{k_0+j-\ell}} - R_{b_{k_0+j-\ell-1}}.$$

The fact that

$$b_{k_0+j-\ell+1} - b_{k_0+j-\ell} = \varrho \cdot b_{k_0+j-\ell} \pm O(1)$$

leads to the factor of ϱ in the denominator of the desired result. □

Lemma 5. *For fixed problem parameters and* N *large, any good sequence* $(q_L, \ldots, q_0)$ *satisfies*

$$q_\ell \geq q_0 + \frac{\ell \cdot d_F(\alpha,\beta)\varrho(1-2\varrho_2) \cdot \sqrt{q_0(1-q_0)}}{(\log N)}$$

Proof. Recall that $\theta'(q) = \frac{1}{\sqrt{q(1-q)}}$ and that θ is smooth on $[\underline{q}, \overline{q}] \subseteq (0,1)$. By Item 2 above, all q_ℓ are within $o_N(1)$ of each other, so the result follows from the inverse function theorem. (Notice that the factor $(1-\varrho_2)$ changed to $(1-2\varrho_2)$ above.) □

Lemma 6. *For* $1 \leq m \leq L$ *and any good sequence* $(q_L, \ldots, q_0)$, *we have*

$$\sum_{\ell=0}^{m-1}(q_\ell - q_{\ell+1})^2 \geq \frac{m \cdot d_F(\alpha,\beta)^2\varrho^2(1-4\varrho_2)\cdot q_0(1-q_0)}{\log^2 N}.$$

Proof. The result follows from Lemma 5 and Cauchy-Schwarz in the form

$$\sum_{\ell=0}^{m-1}(q_\ell - q_{\ell+1})^2 \geq m^{-1}\left(\sum_{\ell=0}^{m-1}|q_\ell - q_{\ell+1}|\right)^2.$$

□

Lemma 7. *For any good sequence* $(q_L, \ldots, q_0)$ *and* $j \geq 0$, *we have*

$$\sum_{\ell=0}^{L-1} b_{k_0+j-\ell}(q_\ell - q_{\ell+1})^2 \geq (1-O(\varrho_2)) \cdot \frac{b_{k_0+j}\,\varrho\, d_F(\alpha,\beta)^2 \cdot q_0(1-q_0)}{\log^2 N}.$$

Proof. We break the sum into parts and apply Lemma 6 to each one. We have:

$$\sum_{\ell=0}^{L-1} b_{k_0+j-\ell}(q_\ell - q_{\ell+1})^2 = b_{k_0+j-L+1}\sum_{\ell=0}^{L-1}(q_\ell - q_{\ell+1})^2 + \sum_{m=1}^{L-1}(b_{k_0+j-m+1} - b_{k_0+j-m})\sum_{\ell=0}^{m-1}(q_\ell - q_{\ell+1})^2$$
$$\geq \sum_{m=1}^{L-1} b_{k_0+j}\cdot\frac{\varrho}{(1+\varrho)^{m+10}}\cdot(1-4\varrho_2)\frac{m\varrho^2 d_F(\alpha,\beta)^2\cdot q_0(1-q_0)}{\log^2 N}$$
$$\geq (1-O(\varrho+\varrho_2))\cdot b_{k_0+j}\cdot\frac{\varrho^3 d_F(\alpha,\beta)^2\cdot q_0(1-q_0)}{\log^2 N}\cdot\sum_{m=1}^{L-1}\frac{m}{(1+\varrho)^m}.$$

For $L = L(\varrho) = O\big(\varrho^{-1}\log(\varrho^{-1})\big)$ sufficiently large,

$$\sum_{m=1}^{L-1}\frac{m\varrho}{(1+\varrho)^m} \geq (1-\varrho)\sum_{m=1}^{\infty}\frac{m}{(1+\varrho)^m}.$$
$$= (1-\varrho)\left(\sum_{m=1}^{\infty}\frac{1}{(1+\varrho)^m}\right)^2$$
$$= \frac{1-\varrho}{\varrho^2}.$$

Substituting and recalling that $\varrho \ll \varrho_2$ completes the proof. $\square$

Combining with Lemma 4 yields the second inequality below (the first is trivial).

Corollary C.4. *For any μ and q_0, we have*

$$\mathbb{P}^{p_i\sim\mu}\Big[\big(\hat{p}_{i,b_{k_0+j-L}},\ \hat{p}_{i,b_{k_0+j-L+1}},\ \ldots,\ \hat{p}_{i,b_{k_0+j}}\big) = \big(q_L, q_{L-1},\ldots,q_0\big)\Big]$$
$$\leq \mathbb{P}\Big[\big(\hat{p}_{i,b_{k_0+j-L}},\ \hat{p}_{i,b_{k_0+j-L+1}},\ \ldots,\ \hat{p}_{i,b_{k_0+j}}\big) = \big(q_L, q_{L-1},\ldots,q_0\big)\ \big|\ p_{i,b_{k_0+j}} = q_0\Big]$$
$$\leq \exp\Big(-\big(1-O(\varrho_2)\big)\frac{b_{k_0+j}d_F(\alpha,\beta)^2}{2\log^2 N}\Big).$$

Lemma 8. *Let j_0 be the largest j such that $b_{k_0+j} \leq N$. Then for N sufficiently large,*

$$\sum_{j=1}^{j_0}\mathbb{E}\left[e^{X_i\cdot\frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N}}\cdot I_i^{b_{k_0+j}}\right] \leq c(\alpha,\varrho)/4.$$

Proof. Recall that $c_{\alpha,\beta} = \frac{d_F(\alpha,\beta)^2}{2}$, and observe that the number of total sequences $(q_L,\ldots,q_0) \in [0,1]^{L+1}$ with $b_{k_0+j+\ell}q_\ell \in \mathbb{Z}$ is at most N^{L+1} for each $j \leq j_0$. Combining Lemma 3 and Corollary C.4 and noting that the latter always gives the main contribution, we find for each $j \leq j_0$,

$$\mathbb{E}\left[e^{X_i\cdot\frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N}}\cdot I_i^{b_{k_0+j}}\right] \leq N^{L+1}\exp\left(\frac{b_{k_0+j}}{\log^2 N}\cdot\big((c_{\alpha,\beta}-\varrho_3) - (1-O(\varrho_2))c_{\alpha,\beta}\big)\right)$$
$$\leq \exp\left(-\Omega\left(\frac{\varrho_3 b_{k_0+j}}{\log^2 N}\right)\right)$$

so long as ϱ_3 is chosen so that $\varrho_3 \gg \max(\varrho,\varrho_2)$. In the last line we used the fact that $b_{k_0+j} \geq b_{k_0} \geq \log^4 N$ to absorb the factor $N^{L+1} \leq e^{\varrho\log^{3/2} N}$ for large N. Summing over j gives the desired result, since for $\varrho_4 = \Omega(\varrho_3)$ and N sufficiently large,

$$\begin{aligned}\sum_{j=1}^{\infty} e^{-\Omega\left(\frac{\varrho_3 b_{k_0+j}}{\log^2 N}\right)} &\leq \sum_{m=1}^{\infty} e^{-\frac{\varrho_4(m+b_{k_0})}{\log^2 N}} \\ &= e^{-\varrho_4 \log^2 N} \sum_{m=1}^{\infty} e^{-\frac{\varrho_4 m}{\log^2 N}} \\ &\leq e^{-\varrho_4 \log^2 N} \cdot O\left(\frac{\log^2 N}{\varrho_4}\right) \\ &\leq e^{-\frac{\varrho_4 \log^2 N}{2}} \\ &\leq c(\alpha, \varrho)/4.\end{aligned}$$

□

We now use Lemma 3 to conclude.

Proof that Algorithm 3 achieves the guarantee of Theorem 3.1. By combining Lemma 8 with the previous Propositions C.2 and C.3, it follows that

$$\mathbb{E}\left[e^{X_i \cdot \frac{c_{\alpha,\beta}-\varrho_3}{\log^2 N}} \cdot I_i\right] \leq 1.$$

Lemma 3 now implies that the total amount of time spent on eventually rejected arms is at most $N(1-\varrho)$ with probability

$$e^{-\frac{(c_{\alpha,\beta}-\varrho_3)(1-\varrho)N}{\log^2 N}}.$$

On this event, the output arm i^* satisfies $n_{i^*,N} \geq \varrho N$ by definition. Since i^* was not rejected, for j_1 be the largest value such $b_{k_0+j_1} \leq \varrho N$ we have

$$\hat{p}_{i^*, b_{k_0+j_1}} \geq \beta + \varrho.$$

The probability for this to hold if $p_i \leq \beta$ is at most $e^{-\Omega_\varrho(N)}$. Altogether we find that

$$\mathbb{P}[p_{i^*} \geq \beta] \geq 1 - \exp\left(-\frac{(c_{\alpha,\beta}-\varrho_5)N}{\log^2 N}\right) - e^{-\Omega_\varrho(N)} \tag{C.10}$$

for ϱ_5 arbitrarily small. This concludes the analysis of Algorithm 3 (since the last error term is negligible). □

C.5 Finding Many Good Arms with a Fixed Budget

In this final subsection we observe that Algorithm 3 can be modified to output as many as $\log N$ distinct arms each of which satisfies the same $(\eta, \varepsilon, \delta)$-PAC guarantee[2], with no degradation in the asymptotic failure probability. With other parameters fixed, we denote the N-sample version of Algorithm 3 by $\mathcal{A}_N$ to emphasize the dependence on N. In particular, N both equals the number of steps in $\mathcal{A}_N$ and appears (via its logarithm) in the description of $\mathcal{A}_N$'s individual steps.

Let $\tilde{N} = N + \lceil \frac{2N}{\log^{1/2}(N)} \rceil$. We consider a modified algorithm $\tilde{\mathcal{A}}_{\tilde{N}}$ which mimicks the behavior of $\mathcal{A}_N$ with two changes:

1. $\tilde{\mathcal{A}}_{\tilde{N}}$ is a $\tilde{N}$-sample algorithm.
2. If an arm a_i has not yet been rejected after $M = \lceil N/\log^{3/2}(N) \rceil$ samples, then $\tilde{\mathcal{A}}_{\tilde{N}}$ accepts a_i and continues to a_{i+1}. In particular, $\tilde{\mathcal{A}}_{\tilde{N}}$ may accept several arms instead of just one.

Theorem C.9. *With probability* $1-\exp\left(-\frac{(c_{\alpha,\beta}-\varrho_5-o_N(1))N}{\log^2 N}\right)$, $\tilde{\mathcal{A}}_{\tilde{N}}$ *accepts at least* $\log(N)$ *distinct arms* a_i, *all of which satisfy* $p_i \geq \beta$.

[2] In fact $\log N$ can be replaced by anything $o_N(\log^2 N)$ by more precisely defining M and $\tilde{N}$.

The change from N to $\tilde{N}$ is almost irrelevant in the actual statement of Theorem C.9 since $\log(N) \geq \log(\tilde{N}) - o_N(1)$. In particular, $\tilde{\mathcal{A}}_{\tilde{N}}$ is a $\tilde{N}$-sample algorithm which outputs at least $\log(\tilde{N}) - 1$ arms with probability $1 - \exp\left(-\frac{(c_{\alpha,\beta} - \varrho_5 - o_{\tilde{N}}(1))\tilde{N}}{\log^2 \tilde{N}}\right)$. It is certainly not really necessary to use the value $\log(N)$ rather than $\log(\tilde{N})$ to describe the individual steps taken by $\tilde{A}_{\tilde{N}}$. However introducing $\tilde{N}$ streamlines the proof below by letting us treat $\mathcal{A}_N$ as a blackbox.

Proof. To show that all accepted arms a_i satisfy $p_i \geq \beta$ with sufficiently high probability, it suffices to consider (C.10) with the final term replaced by $e^{-\Omega_{\varrho}(N/\log^{3/2}(N))}$. In particular, observe that the main term does not change, even after multiplying the failure probability by $O\big(\log^{3/2}(N)\big)$ (the maximum possible number of arms accepted by $\tilde{\mathcal{A}}_{\tilde{N}}$. Thus we focus on showing that $\tilde{\mathcal{A}}_{\tilde{N}}$ outputs at least $\log(N)$ arms with high probability.

Consider yet another N-sample algorithm $\widehat{\mathcal{A}}_N$ which deletes each arm independently with probability $1/N$ and follows $\mathcal{A}_N$ on the set of non-deleted arms in order of increasing index. (Like $\mathcal{A}_N$, $\widehat{\mathcal{A}}_N$ never accepts arms before time N.) We simulate $\tilde{\mathcal{A}}_{\tilde{N}}$ and $\widehat{\mathcal{A}}_N$ on the same reward sequences, i.e. we couple them so that the t-th sample of arm a_i always gives the same result for each (t, i). We **claim** that in this coupling, conditioned on $\tilde{\mathcal{A}}_{\tilde{N}}$ failing to accept $\log(N)$ arms within the first $\tilde{N}$ samples, $\widehat{\mathcal{A}}_N$ has probability $\Omega(N^{-\log(N)})$ to fail (i.e. output a_i with $p_i < \beta$) when run for N samples.

First let us assume the claim and deduce Theorem C.9. Denote by $p(N)$ the probability for $\mathcal{A}_N$ to fail. Note that $\widehat{\mathcal{A}}_N$ has the same failure probability $p(N)$, having in fact the same behavior as $\mathcal{A}_N$ in distribution (as the set of deleted arms is independent of everything else). Moreover let $\tilde{p}(\tilde{N}, k)$ denote the probability that $\tilde{\mathcal{A}}_{\tilde{N}}$ fails to accept at least k arms. The claim above implies that

$$\begin{aligned}
\tilde{p}(\tilde{N}, \log N) &\leq O\big(N^{\log N}\big) \cdot p(N, 1) \\
&\leq e^{o_N(N/\log^2 N)} \cdot p(N, 1) \\
&\leq \exp\left(-\frac{(c_{\alpha,\beta} - \varrho_5 - o_N(1))N}{\log^2 N}\right).
\end{aligned}$$

It remains to prove the above claim. Let us say the infinite i.i.d. reward sequence $(r_{i,n})_{n\geq 1}$ of arm a_i is **acceptable** if $\mathcal{A}_N$ would not reject a_i within M samples, i.e. $\tilde{\mathcal{A}}_{\tilde{N}}$ will either accept a_i or run out of samples before doing so. We take the point of view that each a_i is either acceptable or not (by randomly fixing the reward sequences at the start). Then with probability $\Omega(N^{-\log(N)})$, the first $\log(N)$ acceptable arms are skipped by $\widehat{\mathcal{A}}$, and the first $\hat{N}$ unacceptable arms are not skipped. On this event, the first $\hat{N} - M \geq N$ samples obtained by $\widehat{\mathcal{A}}_N$, i.e. all N of its samples, are drawn from unacceptable arms. On this event, $\widehat{\mathcal{A}}_N$ fails with constant probability, which establishes the claim and completes the proof. $\square$