33 Pure Exploration

All the policies proposed in this book so far were designed to maximize the cumulative reward. As a consequence, the policies must carefully balance exploration against exploitation. But what happens if there is no price to be paid for exploring? Imagine, for example, that a researcher has K configurations of a new drug and the budget to test the drugs on n mice. The researcher wants to find the most promising drug configuration for subsequent human trials, but is not concerned with the outcomes for the mice. Problems of this nature are called **pure exploration** problems. Although there are similarities to the cumulative regret setting, there are also differences. This chapter outlines a variety of pure exploration problems and describes the basic algorithmic ideas.

Notation
To keep things simple we restrict our attention to K-armed Gaussian bandits with unit variance, but all upper bounds generalize easily to the subgaussian case or with natural modifications to exponential families and other well-behaved distributions. Unless otherwise specified, $\mathcal{E} = \mathcal{E}_{\mathcal{N}}^K(1)$ is the class of Gaussian bandits with unit variance. Let $\nu \in \mathcal{E}$ be a Gaussian bandit. Recall that ν_i is the reward distribution of the ith arm, which has mean $\mu_i(\nu)$. The suboptimality gap of the ith arm is $\Delta_i(\nu) = \max_j \mu_j(\nu) - \mu_i(\nu)$. When convenient we treat $\mu(\nu)$ and $\Delta(\nu)$ as K-dimensional vectors. The policy and bandit interact sequentially to produce a sequence of outcomes $A_1, X_1, \dots, A_n, X_n$. For policy π and Gaussian bandit ν remember that $\mathbb{P}_{\nu\pi}$ is the measure on outcomes induced by the interaction of π and ν and $\mathbb{E}_{\nu\pi}$ the expectation with respect to this measure. We also let $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$. Also recall that $T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_s = i\}$ and $\hat{\mu}_i(t) = \sum_{s=1}^t \mathbb{I}\{A_s = i\} X_s / T_i(t)$. When the context is obvious we write Δ_i.

33.1 Simple regret

One way to model the pure exploration problem is to assume a horizon of n rounds. The policy π is expected to output an action A_{n+1} and the loss on bandit $\nu \in \mathcal{E}$ is the **simple regret**, which is the expected suboptimality gap of the last action:

$$R_n^{\text{SIMPLE}}(\pi, \nu) = \mathbb{E}_{\nu\pi}\left[\Delta_{A_{n+1}}(\nu)\right] .$$

In order to get a handle on this new objective we investigate the explore-then-commit algorithm introduced in Chapter 6. Because we only care about choosing a good arm in the final round it makes sense to explore for the first n rounds and then choose the empirically best arm in the last round. Since the commitment is only for the last round, this algorithm is often referred to as the **uniform exploration** (UE) policy.

1: **for** $t = 1, 2, \ldots, n$ **do**
2: Choose $A_t = 1 + (t \bmod K)$
3: **end for**
4: Choose $A_{n+1} = \operatorname{argmax}_{i \in [K]} \hat{\mu}_i(n)$

Algorithm 18: Uniform exploration

THEOREM 33.1 *Let π be the policy of Algorithm 18 and $\nu \in \mathcal{E}$ be a Gaussian bandit. Then*

$$R_n^{\text{SIMPLE}}(\pi, \nu) \leq \min_{\Delta \geq 0} \left(\Delta + \sum_{i : \Delta_i(\nu) > \Delta} \Delta_i(\nu) \exp\left(-\frac{\lfloor n/K \rfloor \, \Delta_i(\nu)^2}{4} \right) \right) .$$

Proof Let $\Delta_i = \Delta_i(\nu)$ and $\mathbb{P} = \mathbb{P}_{\nu\pi}$. Assume without loss of generality that $\Delta_1 = 0$ so the first arm is optimal. Let i be a suboptimal arm with $\Delta_i > \Delta$ and observe that $A_{n+1} = i$ implies that $\hat{\mu}_i(n) \geq \hat{\mu}_1(n)$. Now $T_i(n) \geq \lfloor n/K \rfloor$ is not random, so by Theorem 5.1 and Lemma 5.2,

$$\mathbb{P}\left(\hat{\mu}_i(n) \geq \hat{\mu}_1(n)\right) = \mathbb{P}\left(\hat{\mu}_i(n) - \hat{\mu}_1(n) \geq 0\right) \leq \exp\left(-\frac{\lfloor n/K \rfloor \, \Delta_i^2}{4} \right) .$$

The definition of the simple regret yields

$$R_n^{\text{SIMPLE}}(\pi, \nu) = \sum_{i=1}^{K} \Delta_i \mathbb{P}\left(A_{n+1} = i\right) \leq \Delta + \sum_{i : \Delta_i > \Delta} \Delta_i \mathbb{P}\left(A_{n+1} = i\right) .$$

The proof is completed by taking the minimum over all $\Delta \geq 0$. □

The theorem highlights some important differences between the simple regret and the cumulative regret. If ν is fixed and n tends to infinity, then the simple regret converges to zero exponentially fast. On the other hand, if n is fixed and ν is allowed to vary, then we are in a worst-case regime. Theorem 33.1 can be used to derive a bound in this case by choosing $\Delta = 2\sqrt{\log(K)/\lfloor n/K \rfloor}$, which after a short algebraic calculation shows that for $n \geq K$ there exists a universal constant $C > 0$ such that

$$R_n^{\text{SIMPLE}}(\text{UE}, \nu) \leq C \sqrt{\frac{K \log(K)}{n}} \qquad \text{for all } \nu \in \mathcal{E} . \tag{33.1}$$

In Exercise 33.1 we ask you to use the techniques of Chapter 15 to prove that for all policies there exists a bandit $\nu \in \mathcal{E}$ such that $R_n^{\text{SIMPLE}}(\pi, \nu) \geq C\sqrt{K/n}$ for

some universal constant $C > 0$. It turns out the logarithmic dependence on K in Eq. (33.1) is tight for uniform exploration (Exercise 33.2), but there exists another policy for which the simple regret matches the aforementioned lower bound up to constant factors. There are several ways to do this, but the most straightforward is via a reduction from algorithms designed for minimizing cumulative regret.

PROPOSITION 33.1 *Let π be a policy for which the $(n+1)$th action, A_{n+1} is chosen randomly with $\mathbb{P}(A_{n+1} = i) = T_i(n)/n$, then its simple regret satisfies*

$$R_n^{\text{SIMPLE}}(\pi, \nu) = \frac{R_n(\pi, \nu)}{n},$$

where $R_n(\pi, \nu)$ is the cumulative regret of policy π when executed on bandit ν.

Proof By the regret decomposition identity (4.5),

$$R_n(\pi, \nu) = n\mathbb{E}\left[\sum_{i=1}^{K} \Delta_i \frac{T_i(n)}{n}\right] = n\mathbb{E}\left[\Delta_{A_{n+1}}\right] = nR_n^{\text{SIMPLE}}(\pi, \nu),$$

where the first equality follows from the definition of the cumulative regret, the third from the definition of the policy in the $(n+1)$th round and the last the definition of the simple regret. □

COROLLARY 33.1 *Fix $K \in \mathbb{N}$, $n \in \mathbb{N}$. Then, there exist a policy π such that for any $\nu \in \mathcal{E}$ with $\Delta(\nu) \in [0,1]^K$ the simple regret is bounded by $R_n^{\text{SIMPLE}}(\pi, \nu) \le C\sqrt{K/n}$, where $C > 0$ is a universal constant.*

Proof Combine the previous result with Theorem 9.1. □

Proposition 33.1 raises our hopes that policies designed for minimizing the cumulative regret might also have well-behaved simple regret. Unfortunately this is only true in the intermediate regimes where the best arm is hard to identify. Policies with small cumulative regret spend most of their time playing the optimal arm and play suboptimal arms just barely enough to ensure they are not optimal. In pure exploration this leads to a highly suboptimal policy for which the simple regret is asymptotically polynomial while we know from Theorem 33.1 that even for the 'clueless' uniform exploration algorithm, the simple regret decreases exponentially fast.

33.2 Best arm identification

Let $\delta \in (0,1)$ be a known confidence level. The objective in **fixed confidence best arm identification** is to design a policy π and $\mathcal{F}_t$-stopping time τ such that $\mathbb{E}_{\nu\pi}[\tau]$ is as small as possible while ensuring that

$$\mathbb{P}_{\nu\pi}\left(\Delta_{A_\tau+1}(\nu) > 0\right) \le \delta \qquad \text{for all } \nu \in \mathcal{E}. \tag{33.2}$$

Like the cumulative regret, minimizing $\mathbb{E}_{\nu\pi}[\tau]$ is a multi-objective criteria and it is not immediately clear that the same policy and stopping rule should minimize

$\mathbb{E}_{\nu\pi}[\tau]$ for all $\nu \in \mathcal{E}$ simultaneously. Conveniently, however, the condition that the policy and stopping rule must satisfy Eq. (33.2) plays the role of the consistency assumption in the asymptotic lower bounds in Chapter 16 and for small δ there is a single policy and stopping rule that essentially minimizes $\mathbb{E}_{\nu\pi}[\tau]$ for all ν simultaneously.

Lower bound

We start with the lower bound, which serves as a target for the upper bound to follow. For $\nu \in \mathcal{E}$ define $i^*(\nu) = \operatorname{argmin}_{i\in[K]} \Delta_i(\nu)$ to be the set of optimal arms and

$$\mathcal{E}_{\text{alt}}(\nu) = \{\nu' \in \mathcal{E} : i^*(\nu') \cap i^*(\nu) = \emptyset\} \ ,$$

which is the set of Gaussian bandits with different optimal arms than ν.

THEOREM 33.2 *Let $\delta \in (0,1)$ and suppose that π is a policy and τ a stopping time such that for all $\nu \in \mathcal{E}$ with a unique optimal arm, $\mathbb{P}_{\nu\pi}\left(A_{\tau+1} \notin i^*(\nu)\right) \leq \delta$. Then $\mathbb{E}_{\nu\pi}[\tau] \geq c^*(\nu) \log\left(\frac{4}{\delta}\right)$, where*

$$c^*(\nu)^{-1} = \sup_{\alpha\in\Delta^{K-1}} \left(\inf_{\nu'\in\mathcal{E}_{alt}(\nu)} \left(\sum_{i=1}^{K} \alpha_i \, \mathrm{D}(\nu_i, \nu_i') \right) \right) . \tag{33.3}$$

Proof Let $\nu' \in \mathcal{E}_{\text{alt}}(\nu)$. By assumption we have $\mathbb{P}_{\nu\pi}(A_{\tau+1} \notin i^*(\nu)) \leq \delta$ and $\mathbb{P}_{\nu'\pi}(A_{\tau+1} \notin i^*(\nu')) \leq \delta$. The high probability Pinsker's inequality (Theorem 14.2) and the stopping time version of Lemma 15.1 (see Exercise 15.6) show that for any $\mathcal{F}_\tau$-measurable event E,

$$\mathbb{P}_{\nu\pi}(E) + \mathbb{P}_{\nu'\pi}(E^c) \geq \frac{1}{2} \exp\left(-\sum_{i=1}^{K} \mathbb{E}_{\nu\pi}[T_i(\tau)] \, \mathrm{D}(\nu_i, \nu_i') \right) .$$

Choosing $E = \mathbb{I}\{A_{\tau+1} \notin i^*(\nu)\}$ leads to

$$\begin{aligned} 2\delta &\geq \mathbb{P}_{\nu\pi}(A_{\tau+1} \notin i^*(\nu)) + \mathbb{P}_{\nu'\pi}(A_{\tau+1} \notin i^*(\nu')) \\ &\geq \frac{1}{2} \exp\left(-\sum_{i=1}^{K} \mathbb{E}_{\nu\pi}\left[T_i(\tau)\right] \mathrm{D}(\nu_i, \nu_i') \right) . \end{aligned} \tag{33.4}$$

Using the definition of $c^*(\nu)$ and the above display we have

$$\begin{aligned} \frac{\mathbb{E}_{\nu\pi}[\tau]}{c^*(\nu)} &= \mathbb{E}_{\nu\pi}[\tau] \sup_{\alpha\in\Delta^{K-1}} \inf_{\nu'\in\mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{K} \alpha_i \, \mathrm{D}(\nu_i, \nu_i') \\ &\geq \mathbb{E}_{\nu\pi}[\tau] \inf_{\nu'\in\mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{K} \frac{\mathbb{E}_{\nu\pi}[T_i(\tau)]}{\mathbb{E}_{\nu\pi}[\tau]} \mathrm{D}(\nu_i, \nu_i') \\ &= \inf_{\nu'\in\mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{K} \mathbb{E}_{\nu\pi}[T_i(\tau)] \, \mathrm{D}(\nu_i, \nu_i') \geq \log\left(\frac{4}{\delta}\right) , \end{aligned} \tag{33.5}$$

where the last inequality follows from Eq. (33.4). Rearranging completes the proof. □

We will shortly show that the lower bound is tight asymptotically as δ tends to zero, but first it is worth examining the value of $c^*(\nu)$. Suppose that $\alpha^*(\nu) \in \Delta^{K-1}$ satisfies

$$c^*(\nu)^{-1} = \inf_{\nu' \in \mathcal{E}_{\rm alt}(\nu)} \sum_{i=1}^{K} \alpha_i^*(\nu) \, \mathrm{D}(\nu_i, \nu_i') \,.$$

A few observations about this optimization problem:

(a) Provided that ν has a unique optimal arm, then the value of $\alpha^*(\nu)$ is unique. Uniqueness continues to hold when $\mathcal{E}$ is an unstructured bandit with distributions from an exponential family.

(b) The inequality in Eq. (33.5) is tightest when $\mathbb{E}_{\nu\pi}[T_i(\tau)]/\mathbb{E}_{\nu\pi}[\tau] = \alpha_i^*(\nu)$, which shows a policy can only match the lower bound by playing arm i exactly in proportion to $\alpha_i^*(\nu)$ in the limit as δ tends to zero.

(c) When $\mathcal{E} = \mathcal{E}_{\mathcal{N}}^2(1)$ and $\nu \in \mathcal{E}$ has a unique optimal arm, then

$$\begin{aligned} c^*(\nu)^{-1} &= \frac{1}{2} \sup_{\alpha \in [0,1]} \inf_{\nu' \in \mathcal{E}_{\rm alt}(\nu)} \left(\alpha(\mu_1(\nu) - \mu_1(\nu'))^2 + (1-\alpha)(\mu_2(\nu) - \mu_2(\nu'))^2 \right) \\ &= \frac{1}{2} \sup_{\alpha \in [0,1]} \left((1-\alpha)^2 + \alpha^2 \right) (\mu_1(\nu) - \mu_2(\nu))^2 = \frac{1}{4} (\mu_1(\nu) - \mu_2(\nu))^2 \,. \end{aligned}$$

In this case we observe that $\alpha_1^*(\nu) = \alpha_2^*(\nu) = 1/2$.

Policy, stopping rule and upper bounds

Both the stopping rule and policy are derived almost directly from insights provided by the lower bound. For the policy we would like it to choose action i in proportion to $\alpha_i^*(\nu)$, which must be estimated from data. The stopping rule is motivated by recalling from the proof of Theorem 33.2 that for all $\nu' \in \mathcal{E}_{\rm alt}(\nu)$,

$$\begin{aligned} \mathbb{P}_{\nu\delta}(A_{\tau_\delta+1} \notin i^*(\nu)) + \mathbb{P}_{\nu'\delta}(A_{\tau_\delta+1} \notin i^*(\nu')) &\geq \frac{1}{2} \exp\left(-\mathrm{D}(\mathbb{P}_{\nu\delta}, \mathbb{P}_{\nu'\delta})\right) \qquad (33.6) \\ &= \frac{1}{2} \exp\left(- \sum_{i=1}^{K} \mathbb{E}[T_i(\tau_\delta)] \, \mathrm{D}(\nu_i, \nu_i') \right) . \end{aligned}$$

If the inequality is tight, then we might guess that a reasonable stopping rule might be the first round t when

$$\sum_{i=1}^{K} T_i(t) \, \mathrm{D}(\nu_i, \nu_i') \gtrsim \log\left(\frac{1}{\delta}\right) .$$

There are two problems: *(a)* ν is unknown, so the expression cannot be evaluated and *(b)* we have replaced the expected pull counts with their realizations, which may invalidate the expression. Still, let us persevere. To deal with the first problem we can try replacing ν by the Gaussian bandit with mean vector $\hat{\mu}(t)$, which we

denote by $\hat{\nu}(t)$. Then let

$$Z_t = \inf_{\nu' \in \mathcal{E}_{\text{alt}}(\hat{\nu}(t))} \sum_{i=1}^{K} T_i(t)\, \mathrm{D}(\hat{\nu}_i(t), \nu'_i) = \frac{1}{2} \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\hat{\mu}(t))} \sum_{i=1}^{K} T_i(t)(\hat{\mu}_i(t) - \mu_i(\nu'))^2 \,.$$

We will show there exists a choice of $\beta_t(\delta)$ such that if $\tau_\delta = \min\{t : Z_t \geq \beta_t(\delta)\}$, then the empirically optimal arm at τ_δ is the best arm with probability at least $1 - \delta$. As we remarked earlier, if the policy is to match the lower bound it should play arm i approximately in proportion to $\alpha_i^*(\nu)$. This suggests estimating $\alpha^*(\nu)$ by $\hat{\alpha}(t) = \alpha^*(\hat{\nu}(t))$ and then playing the arm for which $t\hat{\alpha}_i(t) - T_i(t)$ is maximized. If $\hat{\alpha}(t)$ is inaccurate, then perhaps the samples collected will not allow the algorithm to improve its estimates. To overcome this last challenge the policy includes enough forced exploration to ensure that eventually $\hat{\alpha}(t)$ converges to $\alpha^*(\nu)$ with high probability. Combining all these ideas leads to the Track-and-Stop policy (Algorithm 19).

Input δ and $\beta_t(\delta)$
Choose each arm once
while $Z_t \leq \beta_t(\delta)$ **do**
    if $\operatorname{argmin}_{i \in [K]} T_i(t-1) \leq \sqrt{t}$ **then**
        Choose $A_t = \operatorname{argmin}_{i \in [K]} T_i(t-1)$
    else
        Choose $A_t = \operatorname{argmax}_{i \in [K]} (t\hat{\alpha}_i^*(t-1) - T_i(t-1))$
    end if
end while

Algorithm 19: Track-and-Stop

THEOREM 33.3 *Let π_δ and τ_δ be the policy/stopping rule of Algorithm 19. There exists a choice of $\beta_t(\delta)$ such that for all $\nu \in \mathcal{E}$ with $|i^*(\nu)| = 1$ it holds that*

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\nu \pi_\delta}[\tau_\delta]}{\log(1/\delta)} = c^*(\nu) \,.$$

Furthermore, $\mathbb{P}_{\nu \pi_\delta}(i^*(\hat{\nu}(\tau_\delta)) \neq i^*(\nu)) \leq \delta$.

The proof takes a little work. First we show the stopping rule is sound in the sense that indeed the algorithm outputs the optimal arm with probability at least $1 - \delta$.

LEMMA 33.1 *Let $f : [K, \infty) \to \mathbb{R}$ be given by $f(x) = \exp(K - x)(x/K)^K$ and $\beta_t(\delta) = K \log(t^2 + t) + f^{-1}(\delta)$. Then for $\tau = \min\{t : Z_t \geq \beta_t(\delta)\}$ it holds that $\mathbb{P}\left(i^*(\hat{\nu}(\tau)) \neq i^*(\nu)\right) \leq \delta$.*

Basic calculus shows that f is monotone decreasing on $[K, \infty)$ so the inverse is well defined. In fact the inverse has a closed form solution in terms of

the Lambert W function. By staring at the form of f one can check that $\lim_{\delta\to 0} f^{-1}(\delta)/\log(1/\delta) = 1$ or equivalently that $f^{-1}(\delta) = (1+o(1))\log(1/\delta)$.

Proof of Lemma 33.1 Abbreviate $\mu = \mu(\nu)$ and $\Delta = \Delta(\nu)$ and assume without loss of generality that $\Delta_1 = 0$. By the definition of τ, if $\nu \in \mathcal{E}_{\text{alt}}(\hat{\nu}(\tau))$, then

$$\frac{1}{2}\sum_{i=1}^{K} T_i(\tau)(\hat{\mu}_i(\tau) - \mu_i)^2 \geq \beta_\tau(\delta)\,.$$

Using the definition of $\mathcal{E}_{\text{alt}}(\hat{\nu}(\tau))$ yields

$$\mathbb{P}\left(1 \neq i^*(\hat{\nu}(\tau))\right) = \mathbb{P}\left(\nu \in \mathcal{E}_{\text{alt}}(\hat{\nu}(\tau))\right) \leq \mathbb{P}\left(\frac{1}{2}\sum_{i=1}^{K} T_i(\tau)(\hat{\mu}_i(\tau) - \mu_i)^2 \geq \beta_\tau(\delta)\right).$$

Then apply Lemma 33.2 and Proposition 33.2 from Section 33.2.1. □

Below we sketch the proof of Theorem 33.3. A more complete outline is given in Exercise 33.6.

Proof sketch of Theorem 33.3 Lemma 33.1 shows that the stopping procedure and selection rule of Track-and-Stop are valid in the sense that the probability of the arm selected being suboptimal is at most δ. It remains to control the expectation of the stopping time. The intuition is straightforward. As more samples are collected we expect that $\hat{\alpha}(t) \approx \alpha^*(\nu)$ and $\hat{\mu} \approx \mu$ and

$$\begin{aligned}
Z_t &= \inf_{\tilde{\nu}\in\mathcal{E}_{\text{alt}}(\hat{\nu}(t))} \sum_{i=1}^{K} \frac{T_i(t)(\hat{\mu}_i(t) - \mu_i(\tilde{\nu}))^2}{2} \\
&\approx \inf_{\tilde{\nu}\in\mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{K} \frac{\alpha_i^*(\nu)(\mu_i(\nu) - \mu_i(\tilde{\nu}))^2}{2} \\
&= \frac{t}{c^*(\nu)}\,.
\end{aligned}$$

Provided the approximation is reasonably accurate, the algorithm should halt once

$$\frac{t}{c^*(\nu)} \geq \beta_t(\delta) = (1+o(1))\log(1/\delta)\,,$$

which occurs once $t \geq (1+o(1))c^*(\nu)\log(1/\delta)$. □

33.2.1 Concentration

The first concentration theorem follows from Corollary 5.1 and a union bound.

LEMMA 33.2 *Let $X_1, X_2, \ldots$ be a sequence of independent Gaussian random variables with mean μ and unit variance. Let $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$. Then*

$$\mathbb{P}\left(\textit{exists } n \in \mathbb{N}^+ : \frac{n}{2}(\hat{\mu}_n - \mu)^2 \geq \log(1/\delta) + \log(n(n+1))\right) \leq \delta\,.$$

PROPOSITION 33.2 *Let $g : \mathbb{N} \to \mathbb{R}$ be monotone nondecreasing and for each $i \in [K]$ let $S_{i1}, S_{i2}, \dots$ be an infinite sequence of random variables such that for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\textit{exists } s \in \mathbb{N} : S_{is} \geq g(s) + \log(1/\delta)\right) \leq \delta\,.$$

Then provided that $(S_i)_{i=1}^K$ are independent and $x \geq 0$,

$$\mathbb{P}\left(\textit{exists } s \in \mathbb{N}^K : \sum_{i=1}^K S_{is_i} \geq Kg\left(\sum_{i=1}^K s_i\right) + x\right) \leq \left(\frac{x}{K}\right)^K \exp(K - x)\,.$$

Proof For $i \in [d]$ let $W_i = \max\{w \in [0,1] : S_{is} < g(s) + \log(1/w) \text{ for all } s \in \mathbb{N}\}$. Then for any $s \in \mathbb{N}^d$,

$$\sum_{i=1}^d S_{is_i} \leq \sum_{i=1}^d g(s_i) + \sum_{i=1}^d \log(1/W_i) \leq dg\left(\sum_{i=1}^d s_i\right) + \sum_{i=1}^d \log(1/W_i)\,.$$

By assumption $(W_i)_{i=1}^d$ are independent and satisfy $\mathbb{P}(W_i \leq x) \leq x$ for all $x \in [0,1]$. The proof is completed by using Exercise 5.18. □

33.3 Best arm identification with a budget

The setting in the previous section is called the fixed confidence version of best arm identification because the learner should minimize the exploration time in order to satisfy a constraint on the confidence. In the fixed budget variant the learner is given a constraint on the horizon and should minimize the probability of choosing a suboptimal arm.

This reframing of the problem makes algorithm design and analysis a little more nuanced and the results are not as clean. A naive option would be to use the explore-then-commit policy, but as discussed in Section 33.1 this approach leads to poor results when the suboptimality gaps are not close. To overcome this problem the Sequential Halving algorithm divides the budget into $L = \lceil \log_2(K) \rceil$ phases. In the first phase the algorithm chooses each arm $\lfloor n/(KL) \rfloor$ times. The bottom half of the arms are eliminated and the process is repeated.

Let $\nu \in \mathcal{E}$ have mean vector μ and assume that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Define $H_1(\mu)$ and $H_2(\mu)$ by

$$H_1(\mu) = \sum_{i=2}^K \frac{1}{\Delta_i^2} \qquad\qquad H_2(\mu) = \max_{i>1} \frac{i}{\Delta_i^2}\,.$$

For bandits where the arms are not in order the value of $H_i(\mu)$ is defined as above after permuting the arms. The quantity $H_2(\mu)$ looks a bit unusual, but we will see it arises quite naturally in the analysis. The following also holds:

$$H_2(\mu) \leq H_1(\mu) \leq \frac{H_2(\mu)}{1 + \log(K)}\,. \tag{33.7}$$

Input n and K
Set $L = \lceil \log_2(K) \rceil$ and $\mathcal{A}_1 = [K]$.
for $\ell = 1, \dots, L$ **do**
    Let $T_\ell = \left\lfloor \frac{n}{L|\mathcal{A}_\ell|} \right\rfloor$.
    Choose each arm in $\mathcal{A}_\ell$ exactly T_ℓ times
    For each $i \in \mathcal{A}_\ell$ compute $\hat{\mu}_i^\ell$ as the empirical mean of arm i based on the last T_ℓ samples
    Let $\mathcal{A}_{\ell+1}$ contain the top $\lceil |\mathcal{A}_\ell|/2 \rceil$ arms in $\mathcal{A}_\ell$
end for
Output the arm in $\mathcal{A}_{L+1}$

Algorithm 20: Sequential Halving

THEOREM 33.4 *If $\nu \in \mathcal{E}$ with mean vector $\mu \in \mathbb{R}^K$ and π is sequential halving, then*

$$\mathbb{P}_{\nu\pi}(\Delta_{A_{n+1}} > 0) \leq 3 \log_2(K) \exp\left(-\frac{n}{16 H_2(\mu) \log_2(K)}\right) .$$

In Exercise 33.7 the reader is guided through the proof of this theorem. Let's see how the bound compares to explore-then-commit, which is the same as Algorithm 18. Like in the proof of Theorem 33.1, the probability that uniform exploration selects a suboptimal arm is easily controlled using Theorem 5.1 and Lemma 5.2:

$$\mathbb{P}_{\nu,UE}(\Delta_{A_{n+1}} > 0) \leq \sum_{i=2}^{K} \mathbb{P}\left(\hat{\mu}_i(n) \geq \hat{\mu}_1(n)\right) \leq \sum_{i=2}^{K} \exp\left(-\frac{\lfloor n/K \rfloor \Delta_i^2}{4}\right) .$$

Suppose that $\Delta = \Delta_2 = \Delta_K$ so that all suboptimal arms have the same suboptimality gap. Then $H_2 = K/\Delta^2$ and terms in the exponent for sequential halving and uniform exploration are $O(n\Delta^2/(K \log K))$ and $O(n\Delta^2/K)$ respectively, which means that uniform exploration is actually moderately better than sequential halving, at least if n is sufficiently large. On the other hand, if Δ_2 is small, but $\Delta_i = 1$ for all $i > 2$, then $H_2 = O(1/\Delta_2^2)$ and the exponents are $O(n\Delta^2)$ and $O(n\Delta^2/K)$ respectively and sequential halving is significantly better. The reason for the disparity is the non-adaptivity of uniform exploration, which wastes many samples on arms $i > 2$. On the other hand, with high probability the sequential halving algorithm spends one quarter of its budget sampling from arm two.

33.4 Notes

1 The problems studied in this chapter belong to the literature on **stochastic optimization**, where the simple regret is called the **expected suboptimality** (gap). Many other variants of pure exploration problems exist. Staying with

the example from the preamble of the chapter, a medical researcher may be interested in getting the most reliable information about differences between treatments. This falls into the class of pure information seeking problems, the subject of optimal experimental design from statistics, which we have met earlier.

2 We mentioned briefly that algorithms with logarithmic cumulative regret are not well suited for pure exploration. Suppose that π is a policy such that for each $i \in [K]$ it holds that

$$\mathbb{E}_{\nu\pi}[T_i(n)] = \frac{2}{\Delta_i^2}\log(n) + o(\log(n)) \qquad \text{for all } \nu \in \mathcal{E}\,.$$

We showed that such policies exist in Chapter 8 and that one cannot do better in Chapter 16. Let $\nu \in \mathcal{E}$ be a bandit with mean vector μ and for which there is a unique optimal arm. Then let $\nu' \in \mathcal{E}_{\text{alt}}(\nu)$ be the alternative bandit that has the same mean rewards as μ for all arms except $\mu'_i = \mu_i + (1+\varepsilon)\Delta_i$. Then by Theorem 14.2 and Lemma 15.1,

$$\begin{aligned}\mathbb{P}_{\nu\pi}(A_{n+1} \neq 1) + \mathbb{P}_{\nu'\pi}(A_{n+1} \neq i) &\geq \frac{1}{2}\exp\left(-\operatorname{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi})\right) \\ &\geq \frac{1}{2}\exp\left(-(\log(n) + o(\log(n)))(1+\varepsilon)^2\right) = \frac{1}{2}\left(\frac{1}{n}\right)^{(1+o(1))(1+\varepsilon)^2}\,.\end{aligned}$$

This shows that using an asymptotically optimal policy for cumulative regret minimization leads to a best arm identification policy for which the probability of selecting a suboptimal arm decays only polynomially with n. Note that here we did not make any restrictions on the selection rule that determines A_{n+1}, only that the first n samples were collected by an algorithm that is asymptotically optimal for the cumulative regret.

3 Although there is no exploration/exploitation dilemma in the pure exploration setting, there is still an 'exploration dilemma' in the sense that the optimal exploration policy depends on an unknown quantity. This means the policy must balance (to some extent) the number of samples dedicated to learning the how to explore relative to those actually exploring.

4 The forced exploration in the Track-and-Stop algorithm is good enough for asymptotic optimality, but the fact that the proof would go through with almost any sublinear amount of exploration should cause a little unease. We do not currently know of a principled way to tune the amount of forced exploration, or indeed if there is better algorithm design for best arm identification.

5 The choice of $\beta_t(\delta)$ significantly influences the practical performance of Track-and-Stop. We believe the analysis given here is mostly tight except that the naive concentration bound given in Lemma 33.2 can be improved significantly.

6 Perhaps the most practical setup in pure exploration has not yet received any attention, which is upper and lower instance-dependent bounds on the simple regret. Even better, an analysis of the distribution of $\Delta_{A_{n+1}}$.

33.5 Bibliographical remarks

Pure exploration for bandits seems to have been first studied by Even-Dar et al. [2002], Mannor and Tsitsiklis [2004], Even-Dar et al. [2006] in the 'Probability Approximately Correct' setting where the objective is to find an ε-optimal with as few samples as possible. After a dry spell the field was restarted by Bubeck et al. [2009], Audibert and Bubeck [2010b]. The asymptotically optimal algorithm for the fixed confidence setting of Section 33.2 was introduced by Garivier and Kaufmann [2016], who also provide results for exponential families as well as in-depth intuition and historical background. The same problem is studied in a Bayesian setting by Russo [2016], where the focus is on understanding the posterior probability that a suboptimal arm is optimal and designing policies to minimize this quantity. Even more recently Qin et al. [2017] designed a policy that is optimal in both the frequentist and Bayesian settings. The stopping rule used by Garivier and Kaufmann [2016] is inspired by similar rules by Chernoff [1959]. The sequential halving algorithm is by Karnin et al. [2013]. Besides this there have been many other approaches: Jamieson and Nowak [2014]. The negative result showing that policies for minimizing the cumulative regret do not explore enough in the pure exploration setting is due to Bubeck et al. [2009]. For lower bounds in the fixed budget problem we refer the reader to the recent paper by Carpentier and Locatelli [2016]. Pure exploration has recently become a hot topic and is expanding beyond the finite-armed case. For example, to linear bandits [Soare et al., 2014] and continuous-armed bandits and tree search [Garivier et al., 2016a, Huang et al., 2017a]. Best arm identification has also been considered in the adversarial setting [Jamieson and Talwalkar, 2016, Li et al., 2018, Abbasi-Yadkori et al., 2018].

33.6 Exercises

33.1 Show there exists a universal constant $C > 0$ such that for all $n \geq K > 1$ and all policies π there exists a $\nu \in \mathcal{E}$ such that

$$R_n^{\text{SIMPLE}}(\pi, \nu) \geq C\sqrt{\frac{K}{n}}\,.$$

33.2 Show there exists a universal constant $C > 0$ such that for all $n \geq K > 1$ there exists a $\nu \in \mathcal{E}$ such that

$$R_n^{\text{SIMPLE}}(\text{UE}, \nu) \geq C\sqrt{\frac{K \log(K)}{n}}\,.$$

33.3 Prove both inequalities in Eq. (33.7).

33.4 This exercise is about designing (ε, δ)-PAC algorithms.

(a) For each $\varepsilon > 0$ and $\delta \in (0,1)$ and number of arms $K > 1$ design a policy π and stopping time τ such that for all $\nu \in \mathcal{E}$,

$$\mathbb{P}_{\nu\pi}(\Delta_{A_\tau} \geq \varepsilon) \leq \delta \qquad \text{and} \qquad \mathbb{E}_{\nu\pi}[\tau] \leq \frac{CK}{\varepsilon^2} \log\left(\frac{K}{\delta}\right),$$

for universal constant $C > 0$.

(b) It turns out the logarithmic dependence on K can be eliminated. Design a policy π and stopping time τ such that for all $\nu \in \mathcal{E}$,

$$\mathbb{P}_{\nu\pi}(\Delta_{A_\tau} \geq \varepsilon) \leq \delta \qquad \text{and} \qquad \mathbb{E}_{\nu\pi}[\tau] \leq \frac{CK}{\varepsilon^2} \log\left(\frac{1}{\delta}\right).$$

(c) Prove a lower bound showing that the bound in part (b) is tight up to constant factors in the worst case.

Part (b) of the above exercise is a challenging problem. The simplest approach is to use an elimination algorithm that operates in phases where at the end of each phase the bottom half of the arms (in terms of their empirical estimates) are eliminated. For details see the paper by Even-Dar et al. [2002].

33.5 Let $K = 2$ and suppose a bandit policy π has a cumulative regret of $R_{n-1}(\pi, \nu) \leq C_n(\nu) \log(n)$ where $C_n : \mathcal{E} \to [0, \infty)$ is an instance-dependent constant. Suppose this policy is run for $n-1$ steps and subsequently the empirically best arm is played.

(a) Show there exists a $\nu' \in \mathcal{E}$ such that

$$\mathbb{P}_{\nu\pi}(\Delta_{A_n} > 0) + \mathbb{P}_{\nu'\pi}(\Delta_{A_n} > 0) \geq \frac{1}{2} \exp\left(-\frac{1}{2} C_n(\nu) \Delta(\nu) \log(n)\right),$$

where $\Delta(\nu) = |\mu_1(\nu) - \mu_2(\nu)|$.

(b) Suppose that π is asymptotically optimal in the sense that $\lim_{n\to\infty} C_n(\nu) = 2/\Delta(\nu)$. Show that

$$\lim_{n\to\infty} \sup_{\nu\in\mathcal{E}} n\mathbb{P}_{\nu\pi}(\Delta_{A_n} > 0) \geq 1.$$

33.6 In this exercise you will complete the proof of Theorem 33.3. Define

$$\Phi(\nu, \alpha) = \inf_{\tilde{\nu}\in\mathcal{E}_{\text{alt}}(\nu)} \sum_{i=1}^{K} \alpha_i (\mu_i(\nu) - \mu_i(\tilde{\nu}))^2 .$$

$$M(\varepsilon) = \min\{t : \sup_{s\geq t} |\Phi(\hat{\mu}_t, T(t)/t) - \Phi(\nu, \alpha^*)| \leq \varepsilon\} .$$

$$t^*(\varepsilon, \delta) = \min\{t : \sup_{s\geq t} s(\Phi(\nu, \alpha^*) - \varepsilon) - \beta_s(\delta) \geq 0\} .$$

Let $F_{t,\varepsilon}$ be the event that $\|\hat{\nu}_t - \mu(\nu)\|_\infty \leq \varepsilon$.

(a) Assume that ν has a unique optimal arm. Show that Φ is continuous at ν where the topology is induced by the bijection $\mu : \mathcal{E} \leftrightarrow \mathbb{R}^K$.

(b) Show that if $\cup_{s\geq t} F_{s,\varepsilon}$, then

$$\|\alpha_s - \alpha_s^*\|_\infty \leq 3\varepsilon\,.$$

(c) Show that

$$\|T(t)/t - \alpha_t^*\|_\infty$$

(d) Let $\varepsilon > 0$ and $t_\delta^* = \min\{t : tc^*(\nu) \geq \beta_t(\delta)\}$. Show that

$$\mathbb{E}[\tau_\delta] \leq (1+\varepsilon)t_\delta^* + \sum_{t=\lceil (1+\varepsilon)t_\delta^* \rceil}^{\infty} \mathbb{P}\left(|\Phi(\hat{\nu}_t, \hat{\alpha}_t) - \Phi(\nu, \alpha^*)| \geq \varepsilon c^*(\nu)\right)\,.$$

(e) Use the continuity of Φ to show there exists a function $\zeta : [0,\infty) \to \mathbb{R}$ with $\lim_{\varepsilon\to 0}\zeta(\varepsilon) = 0$ such that

$$\mathbb{P}\left(|\Phi(\hat{\nu}_t, \hat{\alpha}_t) - \Phi(\nu, \alpha^*)| \geq c^*(\nu)\varepsilon\right) \leq \mathbb{P}\left(\|\hat{\mu}_t - \mu(\nu)\|_\infty \geq \zeta(\varepsilon)\right)\,.$$

(f) Prove that $\mathbb{P}\left(\|\hat{\mu}_t - \mu(\nu)\|_\infty \geq \zeta(\varepsilon)\right) \leq K \exp\left(-\frac{\left\lfloor \sqrt{t/K} \right\rfloor \zeta(\varepsilon)^2}{2}\right)$.

(g) Show that $\lim_{\delta\to 0} \frac{t_\delta^*}{\log(1/\delta)} = c^*(\nu)$.

(h) Combine the previous parts to complete the proof of Theorem 33.3 by showing that

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq c^*(\nu)\,.$$

Part (b) is by far the hardest step. Use the forced exploration to prove reasonably fast convergence of $\hat{\mu}_t$ to μ and then continuity arguments. For more details see the article by Garivier and Kaufmann [2016].

33.7 The purpose of this exercise is to prove Theorem 33.4. Assume without loss of generality that $\mu = \mu(\nu)$ satisfies $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$. Given a set $A \subset [K]$ let

$$\textsc{TopK}(A,k) = \left\{ i \in [K] : \sum_{j\leq i} \mathbb{I}\{j \in A\} \leq k \right\}$$

be the top k arms in A. To make life easier you may also assume that K is a power of two so that $|\mathcal{A}_\ell| = K2^{1-\ell}$ and $T_\ell = n2^{\ell-1}/\log_2(K)$.

(a) Prove that $|\mathcal{A}_{L+1}| = 1$.

(b) Let i be a suboptimal arm in $\mathcal{A}_\ell$ and suppose that $1 \in \mathcal{A}_\ell$. Show that

$$\mathbb{P}\left(\hat{\mu}_1^\ell \leq \hat{\mu}_i^\ell \;\middle|\; i \in \mathcal{A}_\ell, 1 \in \mathcal{A}_\ell\right) \leq \exp\left(-\frac{T_\ell \Delta_i^2}{4}\right)\,.$$

(c) Let $\mathcal{A}'_\ell = \mathcal{A}_\ell \setminus \textsc{TopK}(\mathcal{A}_\ell, \lceil |\mathcal{A}_\ell|/4 \rceil)$ be the bottom three quarters of the arms in round ℓ. Show that if the optimal arm is eliminated after the ℓth phase, then

$$N_\ell = \sum_{i \in \mathcal{A}'_\ell} \mathbb{I}\left\{\hat{\mu}_i^\ell \geq \hat{\mu}_1^\ell\right\} \geq \frac{1}{3}|\mathcal{A}'_\ell| \,.$$

(d) Let $i_\ell = \min \mathcal{A}'_\ell$ and show that

$$\mathbb{E}[N_\ell \mid \mathcal{A}_\ell] \leq |\mathcal{A}'_\ell| \max_{i \in \mathcal{A}'_\ell} \exp\left(-\frac{\Delta_i^2 n 2^{\ell-1}}{4\log_2(K)}\right) \leq |\mathcal{A}'_\ell| \exp\left(-\frac{n\Delta_{i_\ell}^2}{16 i_\ell \log_2(K)}\right) .$$

(e) Combine the previous two parts with Markov's inequality to show that

$$\mathbb{P}\left(1 \notin \mathcal{A}_{\ell+1} \mid 1 \in \mathcal{A}_\ell\right) \leq 3\exp\left(-\frac{T\Delta_{i_\ell}^2}{16\log_2(K) i_\ell}\right) .$$

(f) Join the dots to prove Theorem 33.4.