17 High Probability Lower Bounds

The lower bounds proven in the last two chapters were for stochastic bandits. In this chapter we prove high probability lower bounds for both stochastic and adversarial bandits. Recall that for adversarial bandit $x \in [0,1]^{nK}$ and policy π the random regret is

$$\hat{R}_n(\pi, x) = \max_{i \in [K]} \sum_{t=1}^{n} x_{ti} - x_{tA_t}$$

and the (expected) regret is $R_n(\pi, x) = \mathbb{E}[\hat{R}_n(\pi, x)]$. To set expectations, remember that in Chapter 12 we proved two high probability upper bounds on the regret of Exp3-IX. In the first we showed there exists a policy π such that for all adversarial bandits $x \in [0,1]^{nK}$ and $\delta \in (0,1)$ it holds with probability at least $1 - \delta$ that

$$\hat{R}_n(\pi, x) = O\left(\sqrt{Kn \log(K)} + \sqrt{\frac{Kn}{\log(K)}} \log\left(\frac{1}{\delta}\right)\right). \tag{17.1}$$

We also gave a version of the algorithm that depended on $\delta \in (0,1)$ for which with probability at least $1 - \delta$,

$$\hat{R}_n(\pi, x) = O\left(\sqrt{Kn \log\left(\frac{K}{\delta}\right)}\right). \tag{17.2}$$

The important difference is the order of quantifiers. In the first we have a single algorithm and a high-probability guarantee that holds simultaneously for any confidence level. The second algorithm needs the confidence level to be specified in advance. The price for using the generic algorithm appears to be $\sqrt{\log(1/\delta)/\log(K)}$, which is usually quite small but not totally insignificant. We will see that both bounds are tight up to constant factors, which implies that knowing the desired confidence level in advance really does help. One reason why choosing the confidence level in advance is not ideal is that the resulting high-probability bound cannot be integrated to prove a bound in expectation. For algorithms satisfying (17.1) the expected regret can be bounded by

$$R_n(\pi, x) \leq \int_0^\infty \mathbb{P}\left(\hat{R}_n \geq x\right) dx = O(\sqrt{Kn \log(K)}). \tag{17.3}$$

On the other hand, if the high-probability bound only holds for a single δ as in (17.2), then it seems hard to do much better than

$$R_n \le n\delta + O\left(\sqrt{Kn \log\left(\frac{K}{\delta}\right)}\right),$$

which with the best choice of δ leads to a bound of $O(\sqrt{Kn\log(n)})$. It turns out that this argument cannot be strengthened and algorithms with the strong high-probability regret cannot be near-optimal in expectation. For simplicity we start with the stochastic setting before explaining how to convert the arguments to the adversarial model.

17.1 Stochastic bandits

There is no randomness in the expected regret, so in order to derive a high probability bound we define the **random pseudo regret** by

$$\tilde{R}_n = \sum_{i=1}^{K} T_i(n)\Delta_i\,,$$

which is a random variable through the pull counts $T_i(n)$.

For all results in this section we let $\mathcal{E}^K \subset \mathcal{E}^K_{\mathcal{N}}$ denote the set of K-armed Gaussian bandits with suboptimality gaps bounded by one. For $\mu \in [0,1]^d$ we let $\nu_\mu \in \mathcal{E}^K$ be the Gaussian bandit with means μ.

THEOREM 17.1 *Let $n \ge 1$ and $K \ge 2$ and $B > 0$ and π be a policy such that for any $\nu \in \mathcal{E}^K$,*

$$R_n(\pi,\nu) \le B\sqrt{(K-1)n}\,. \tag{17.4}$$

Let $\delta \in (0,1)$. Then there exists a bandit ν in $\mathcal{E}^K$ such that

$$\mathbb{P}\left(\tilde{R}_n(\pi,\nu) \ge \frac{1}{4}\min\left\{n, \frac{1}{B}\sqrt{(K-1)n}\log\left(\frac{1}{4\delta}\right)\right\}\right) \ge \delta\,.$$

Proof Let $\Delta \in (0, 1/2]$ be a constant to be tuned subsequently and $\nu = \nu_\mu$ where the mean vector $\mu \in \mathbb{R}^d$ is defined by $\mu_1 = \Delta$ and $\mu_i = 0$ for $i > 1$. Abbreviate $R_n = R_n(\pi,\nu)$ and $\mathbb{P} = \mathbb{P}_{\nu\pi}$ and $\mathbb{E} = \mathbb{E}_{\nu\pi}$. Let $i = \operatorname{argmin}_{i>1} \mathbb{E}[T_i(n)]$. Then by Lemma 4.2 and the assumption in Eq. (17.4),

$$\mathbb{E}[T_i(n)] \le \frac{R_n}{\Delta} \le \frac{B}{\Delta}\sqrt{\frac{n}{K-1}}\,. \tag{17.5}$$

Define alternative bandit $\nu' = \nu_{\mu'}$ where $\mu' \in \mathbb{R}^d$ is equal to μ except $\mu'_i = \mu_i + 2\Delta$. Abbreviate $\mathbb{P}' = \mathbb{P}_{\nu'\pi}$ and $\tilde{R}_n = \tilde{R}_n(\pi,\nu)$ and $\tilde{R}'_n = \tilde{R}_n(\pi,\nu')$. By Lemma 4.2

and Pinsker's inequality (Theorem 14.2) and the divergence decomposition (Lemma 15.1) we have

$$\begin{aligned}\mathbb{P}\left(\tilde{R}_n \geq \frac{\Delta n}{2}\right) + \mathbb{P}\left(\tilde{R}'_n \geq \frac{\Delta n}{2}\right) &\geq \mathbb{P}\left(T_i(n) \geq \frac{n}{2}\right) + \mathbb{P}\left(T_i(n) < \frac{n}{2}\right) \\ &\geq \frac{1}{2}\exp\left(-\operatorname{D}(\mathbb{P}, \mathbb{P}')\right) \geq \frac{1}{2}\exp\left(-2B\Delta\sqrt{\frac{n}{K-1}}\right) \geq 2\delta\,,\end{aligned}$$

where the last line follows by choosing

$$\Delta = \min\left\{\frac{1}{2}, \frac{1}{2B}\sqrt{\frac{K-1}{n}}\log\left(\frac{1}{4\delta}\right)\right\}.$$

The result follows since $\max\{a, b\} \geq (a+b)/2$. □

COROLLARY 17.1 *Let $n \geq 1$ and $K \geq 2$. Then for any policy π and $\delta \in (0,1)$ such that*

$$n\delta \leq \sqrt{n(K-1)\log\left(\frac{1}{4\delta}\right)} \tag{17.6}$$

there exists a bandit problem $\nu \in \mathcal{E}^K$ such that

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq \frac{1}{4}\min\left\{n, \sqrt{\frac{n(K-1)}{2}\log\left(\frac{1}{4\delta}\right)}\right\}\right) \geq \delta\,. \tag{17.7}$$

Proof We prove the result by contradiction. Assume that the conclusion does not holds for π and let $\delta \in (0,1)$ satisfy (17.6). Then for any bandit problem $\nu \in \mathcal{E}^K$ the expected regret of π is bounded by

$$R_n(\pi, \nu) \leq n\delta + \sqrt{\frac{n(K-1)}{2}\log\left(\frac{1}{4\delta}\right)} \leq \sqrt{2n(K-1)\log\left(\frac{1}{4\delta}\right)}\,.$$

Therefore π satisfies the conditions of Theorem 17.1 with $B = \sqrt{2\log(1/(4\delta))}$, which implies that there exists some bandit problem $\nu \in \mathcal{E}^K$ such that (17.7) holds, contradicting our assumption. □

COROLLARY 17.2 *Let $K \geq 2$ and $p \in (0,1)$ and $B > 0$. Then there does not exist a policy π such that for all $n \geq 1$, $\delta \in (0,1)$ and $\nu \in \mathcal{E}^K$,*

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq B\sqrt{(K-1)n}\log^p\left(\frac{1}{\delta}\right)\right) < \delta$$

Proof We proceed by contradiction. Suppose that such a policy exists. Choosing δ sufficiently small and n sufficiently large ensures that

$$\frac{1}{B}\log\left(\frac{1}{4\delta}\right) \geq B\log^p\left(\frac{1}{\delta}\right) \quad \text{and} \quad \frac{1}{B}\sqrt{n(K-1)}\log\left(\frac{1}{4\delta}\right) \leq n\,.$$

Now by assumption for any $\nu \in \mathcal{E}^K$ we have

$$R_n(\pi, \nu) \leq \int_0^\infty \mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq x\right) dx$$
$$\leq B\sqrt{n(K-1)} \int_0^\infty \exp\left(-x^{1/p}\right) dx \leq B\sqrt{n(K-1)}\,.$$

Therefore by the Theorem 17.1 there exists a bandit $\nu \in \mathcal{E}^K$ such that

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq B\sqrt{n(K-1)} \log\left(\frac{1}{\delta}\right)\right)$$
$$\geq \mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq \frac{1}{4} \min\left\{n, \frac{1}{B}\sqrt{n(K-1)} \log\left(\frac{1}{4\delta}\right)\right\}\right) \geq \delta\,,$$

which contradicts our assumption and completes the proof. □

We suspect there exists a policy π and universal constant $B > 0$ such that for all $\nu \in \mathcal{E}^K$,

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq B\sqrt{Kn} \log\left(\frac{1}{\delta}\right)\right) \leq \delta\,.$$

Except for the issue of unbounded rewards we would have this for Exp3-IX and suspect the analysis of that algorithm could more-or-less be adapted to this setting. Care would be required to deal with the unbounded rewards, but we expect the math to go through with minor adaptations to the algorithm.

17.2 Adversarial bandits

We now explain how to translate the ideas in the previous section to the adversarial model. Throughout we assume a fixed policy π. Let $\Omega = [0,1]^{nK}$ and let $x \in \Omega$ be an adversarial bandit environment. Recall the random regret is

$$\hat{R}_n(x) = \max_{i \in [K]} \sum_{t=1}^n \left(x_{ti} - x_{tA_t}\right)\,,$$

where the randomness in $\hat{R}_n$ is due to the policy only.

THEOREM 17.2 *Let $c, C > 0$ be sufficiently small/large universal constants and $K \geq 2$, $n \geq 1$ and $\delta \in (0,1)$ be such that $n \geq CK\log(1/(2\delta))$. Then there exists a reward sequence $x \in [0,1]^{nK}$ such that*

$$\mathbb{P}\left(\hat{R}_n(x) \geq c\sqrt{nK \log\left(\frac{1}{2\delta}\right)}\right) \geq \delta\,.$$

The proof is technical and messy, but also contains some nuggets of interest. For the sake of brevity we explain only the high level ideas and refer the

reader elsewhere for the gory details. There are two difficulties in translating the arguments in the previous section to the adversarial model. First, in the adversarial model we need the rewards to be bounded in $[0, 1]$. The second difficulty is we now analyse the adversarial regret rather than the random pseudo-regret.

Suppose we sample $X \in \Omega$ from distribution Q on $(\Omega, \mathfrak{B}(\Omega))$ and let $\mathbb{P}_Q$ be the distribution of $\hat{R}_n(X)$.

CLAIM 17.1 *Let $\mathbb{P}_x$ be the distribution of $\hat{R}_n(x)$ and $u > 0$. If $\mathbb{P}_Q(\hat{R}_n(X) \geq u) \geq \delta$, then there exists an $x \in \Omega$ such that $\mathbb{P}_x(\hat{R}_n(x) \geq u)$.*

The next step is to choose Q and argue that $\mathbb{P}(\hat{R}_n(X) \geq u) \geq \delta$ for sufficiently large u. To do this we need a truncated normal distribution. Defining clipping function

$$\operatorname{clip}_{[0,1]}(x) = \begin{cases} 1 & \text{if } x > 1 \\ 0 & \text{if } x < 0 \\ x & \text{otherwise}\,. \end{cases}$$

Let $\sigma, \Delta > 0$ be constants that we'll tune later and $\eta_1, \ldots, \eta_n$ a sequence of independent random variables with $\eta_t \sim \mathcal{N}(1/2, \sigma^2)$. For each $i \in [K]$ let Q_i be the distribution of $X \in \Omega$ where

$$X_{tj} = \begin{cases} \operatorname{clip}_{[0,1]}(\eta_t + \Delta) & \text{if } j = 1 \\ \operatorname{clip}_{[0,1]}(\eta_t + \Delta) & \text{if } j = i \text{ and } i \neq 1 \\ \operatorname{clip}_{[0,1]}(\eta_t) & \text{otherwise}\,, \end{cases}$$

Notice that under any Q_i for fixed t the random variables $X_{t1}, \ldots, X_{tK}$ are not independent, but for fixed j the random variables $X_{1j}, \ldots, X_{tj}$ are independent and identically distributed. We will let the reader justify for themselves that this is equivalent to a stochastic bandit model.

CLAIM 17.2 *If $\sigma > 0$ and $\Delta = \sigma\sqrt{\frac{K-1}{2n} \log\left(\frac{1}{6\delta}\right)}$, then there exists an arm i such that*

$$\mathbb{P}_{Q_1}(T_i(n) < n/2) \geq 2\delta\,.$$

The proof of this claim follows along the same lines as the theorems in the previous section. All that changes is the calculation of the relative entropy. The last step is to relate $T_i(n)$ to the random regret. In the stochastic model this was straightforward, but for adversarial bandits there is an additional step. Notice that under Q_i it holds that $X_{ti} - X_{tA_t} \geq 0$ and that if $X_{ti}, X_{tA_t} \in (0, 1)$, then $X_{ti} - X_{tA_t} = \Delta$. In other words, if no clipping occurs, then $X_{ti} - X_{tA_t} = \Delta$. The following claim upper bounds the number of rounds in which clipping occurs with high probability.

CLAIM 17.3 *If* $\sigma = 1/10$ *and* $\Delta < 1/8$ *and* $n \geq 32 \log(2/\delta)$*, then*

$$\mathbb{P}_{Q_i}\left(\sum_{t=1}^{n} \mathbb{I}\left\{X_{ti}, X_{tA_t} \in (0,1)\right\} \geq \frac{3n}{4}\right) \geq 1 - \delta\,.$$

By combining the first two claims with a union bound we know there exists an arm i such that

$$\mathbb{P}_{Q_i}\left(\hat{R}_n \geq \frac{n\Delta}{4}\right) \geq \delta\,,$$

which by the definition of Δ and Claim 17.1 implies the first part of the theorem.

17.3 Notes

1 The adversarial bandits used in Section 17.2 had the interesting property that the same arm has the best reward in every round (not just the best mean). It is perhaps a little surprising that algorithms cannot exploit this fact.

2 In Theorem 17.2 we did not make any assumptions on the algorithm. If we had assumed the algorithm enjoyed an expected regret bound of $R_n \leq B\sqrt{Kn}$, then we could conclude that for each sufficiently small $\delta \in (0,1)$ there exists an adversarial bandit such that

$$\mathbb{P}\left(\hat{R}_n \geq \frac{c}{B}\sqrt{Kn}\log\left(\frac{1}{2\delta}\right)\right) \geq \delta\,,$$

which shows that our high probability upper bounds for Exp3-IX are nearly tight.

17.4 Bibliographic remarks

Though none of the results are terribly surprising, we do not know of any references except the recent paper by Gerchinovitz and Lattimore [2016].

17.5 Exercises

17.1 Prove each of the claims in Section 17.2.