15 Minimax Lower Bounds

After the short excursion into information theory, let us return to the world of K-armed stochastic bandits. In what follows we fix the horizon $n > 0$ and the number of actions $K > 1$. This chapter has two components. The first is an exact calculation of the relative entropy between measures in the canonical bandit model for a fixed policy and different bandits. In the second component we prove a minimax lower bound that formalizes the intuitive arguments given in Chapter 13.

15.1 Relative entropy between bandits

The following result will be used repeatedly. Some generalizations are provided in the exercises.

LEMMA 15.1 (Divergence decomposition) *Let $\nu = (P_1, \dots, P_K)$ be the reward distributions associated with one K-armed bandit, and let $\nu' = (P'_1, \dots, P'_K)$ be the reward distributions associated with another K-armed bandit. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the measures on the canonical bandit model (Section 4.4) induced by the interconnection of π and ν (respectively, π and ν'). Then*

$$\mathrm{D}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{K} \mathbb{E}_\nu[T_i(n)]\ \mathrm{D}(P_i, P'_i)\,. \tag{15.1}$$

Proof Assume that $\mathrm{D}(P_i, P'_i) < \infty$ for all $i \in [K]$. From this it follows that $P_i \ll P'_i$. Define $\lambda = \sum_{i=1}^K P_i + P'_i$, which is the measure defined by $\lambda(A) = \sum_{i=1}^K (P_i(A) + P'_i(A))$ for any measurable set A. Recalling that ρ is the counting measure on $[K]$, Theorem 14.1 shows that

$$\mathrm{D}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \mathbb{E}_\nu\left[\log\left(\frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}\right)\right].$$

The Radon-Nikodym derivative of $\mathbb{P}_\nu$ with respect to the product measure $(\rho \times \lambda)^n$ is given in Eq. (4.7) as

$$p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^{n} \pi_t(a_t \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{a_t}(x_t)\,.$$

The density of $\mathbb{P}_{\nu'}$ is identical except that p_{a_t} is replaced by p'_{a_t}. Then

$$\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(a_1, x_1, \ldots, a_n, x_n) = \sum_{t=1}^{n} \log \frac{p_{a_t}(x_t)}{p'_{a_t}(x_t)} ,$$

where we used the chain rule for Radon-Nikodym derivatives and the fact that the terms involving the policy cancel. Taking expectations of both sides:

$$\mathbb{E}_\nu \left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(A_t, X_t) \right] = \sum_{t=1}^{n} \mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right] ,$$

and

$$\mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right] = \mathbb{E}_\nu \left[\mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \,\middle|\, A_t \right] \right] = \mathbb{E}_\nu \left[\mathrm{D}(P_{A_t}, P'_{A_t}) \right] ,$$

where in the second equality we used that under $\mathbb{P}_\nu(\cdot|A_t)$ the distribution of X_t is $dP_{A_t} = p_{A_t} d\lambda$. Plugging back into the previous display,

$$\begin{aligned}
\mathbb{E}_\nu \left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(A_1, X_1, \ldots, A_n, X_n) \right] &= \sum_{t=1}^{n} \mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right] \\
&= \sum_{t=1}^{n} \mathbb{E}_\nu \left[\mathrm{D}(P_{A_t}, P'_{A_t}) \right] = \sum_{i=1}^{K} \mathbb{E}_\nu \left[\sum_{t=1}^{n} \mathbb{I}\left\{A_t = i\right\} \mathrm{D}(P_{A_t}, P'_{A_t}) \right] \\
&= \sum_{i=1}^{K} \mathbb{E}_\nu \left[T_i(n) \right] \mathrm{D}(P_i, P'_i) .
\end{aligned}$$

When the right-hand side of (15.1) is infinite, by our previous calculation, it is not hard to see that the left-hand side will also be infinite. □

15.2 Minimax lower bounds

Recall that $\mathcal{E}_{\mathcal{N}}^K(1)$ is the class of Gaussian bandits with unit variance, which can be parameterized by their mean vector $\mu \in \mathbb{R}^K$. Given $\mu \in \mathbb{R}^K$ let ν_μ be the Gaussian bandit for which the ith arm has reward distribution $\mathcal{N}(\mu_i, 1)$.

THEOREM 15.1 *Let $K > 1$ and $n \geq K - 1$. Then for any policy π there exists a mean vector $\mu \in [0,1]^K$ such that*

$$R_n(\pi, \nu_\mu) \geq \frac{1}{27} \sqrt{(K-1)n} .$$

Since $\nu_\mu \in \mathcal{E}_K$, it follows that the minimax regret for $\mathcal{E}_K$ is lower bounded by the right-hand side of the above display as soon as $n \geq K - 1$:

$$R_n^*(\mathcal{E}_K) \geq \frac{1}{27} \sqrt{(K-1)n} .$$

The idea of the proof is illustrated in Fig. 15.1.

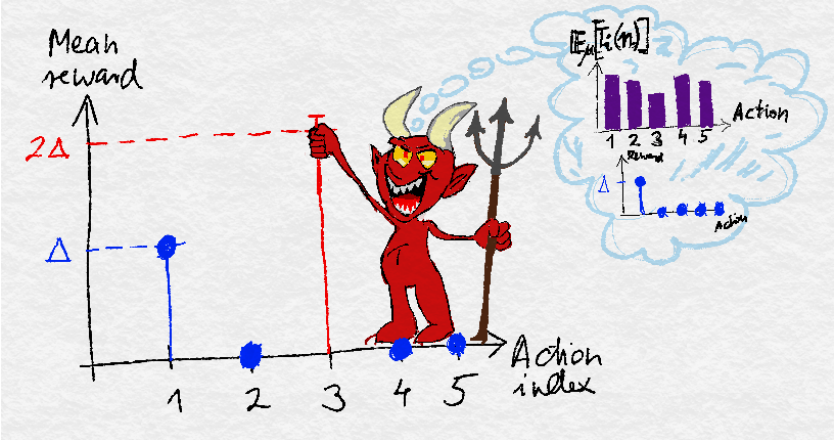

Figure 15.1 The idea of the minimax lower bound. Given a policy and one environment, the evil antagonist picks another environment so that the policy will suffer a large regret in at least one environment

Proof Fix a policy π. Let $\Delta \in [0, 1/2]$ be some constant to be chosen later. As suggested in Chapter 13 we start with a Gaussian bandit with unit variance and mean vector $\mu = (\Delta, 0, 0, \dots, 0)$. This environment and π gives rise to the distribution $\mathbb{P}_{\nu_\mu, \pi}$ on the canonical bandit model $(\mathcal{H}_n, \mathcal{F}_n)$. For brevity we will use $\mathbb{P}_\mu$ in place of $\mathbb{P}_{\nu_\mu, \pi}$ and expectations under $\mathbb{P}_\mu$ will be denoted by $\mathbb{E}_\mu$. To choose the second environment, let

$$i = \operatorname{argmin}_{j>1} \mathbb{E}_\mu[T_j(n)] \,.$$

Since $\sum_{j=1}^K \mathbb{E}_\mu[T_j(n)] = n$, it holds that $\mathbb{E}_\mu[T_i(n)] \leq n/(K-1)$. The second bandit is also Gaussian with unit variance and means

$$\mu' = (\Delta, 0, 0, \dots, 0, 2\Delta, 0, \dots, 0) \,,$$

where specifically $\mu'_i = 2\Delta$. Therefore $\mu_j = \mu'_j$ except at index i and the optimal optimal arm in ν_μ is the first arm and in i in $\nu_{\mu'}$. We abbreviate $\mathbb{P}_{\mu'} = \mathbb{P}_{\nu_{\mu'}, \pi}$. Lemma 4.2 and a simple calculation leads to

$$R_n(\pi, \nu_\mu) \geq \mathbb{P}_\mu(T_1(n) \leq n/2) \frac{n\Delta}{2} \quad \text{and} \quad R_n(\pi, \nu_{\mu'}) > \mathbb{P}_{\mu'}(T_1(n) > n/2) \, \frac{n\Delta}{2} \,.$$

Then applying the high probability Pinsker inequality from the previous chapter (Theorem 14.2),

$$\begin{aligned} R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) &> \frac{n\Delta}{2} \left(\mathbb{P}_\mu(T_1(n) \leq n/2) + \mathbb{P}_{\mu'}(T_1(n) > n/2)\right) \\ &\geq \frac{n\Delta}{4} \exp(-\operatorname{D}(\mathbb{P}_\mu, \mathbb{P}_{\mu'})) \,. \end{aligned} \tag{15.2}$$

It remains to upper bound $\operatorname{D}(\mathbb{P}_\mu, \mathbb{P}_{\mu'})$. For this, we use Lemma 15.1 and the

definitions of μ and μ' to get

$$\mathrm{D}(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) = \mathbb{E}_\mu[T_i(n)]\,\mathrm{D}(\mathcal{N}(0,1), \mathcal{N}(2\Delta, 1)) = \mathbb{E}_\mu[T_i(n)]\,\frac{(2\Delta)^2}{2} \leq \frac{2n\Delta^2}{K-1}\,.$$

Plugging this into the previous display, we find that

$$R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) \geq \frac{n\Delta}{4}\,\exp\left(-\frac{2n\Delta^2}{K-1}\right).$$

The result is completed by choosing $\Delta = \sqrt{(K-1)/4n} \leq 1/2$, where the inequality follows from the assumptions in the theorem statement. The final steps are lower bounding $\exp(-1/2)$ and using $2\max(a,b) \geq a+b$. □

We encourage readers to go through the alternative proof outlined in Exercise 15.1, which takes a slightly different path.

15.3 Notes

1 We used the Gaussian noise model because the KL divergences are so easily calculated in this case, but all that we actually used was that $\mathrm{D}(P_i, P_i') = O((\mu_i - \mu_i')^2)$ when the gap between the means $\Delta = \mu_i - \mu_i'$ is small. While this is certainly not true for *all* distributions, it very often is. Why is that? Let $\{P_\mu : \mu \in \mathbb{R}\}$ be some parametric family of distributions on Ω and assume that distribution P_μ has mean μ. Assuming the densities are twice differentiable and that everything is sufficiently nice that integrals and derivatives can be exchanged (as is almost always the case), we can use a Taylor expansion about μ to show that

$$\begin{aligned}
\mathrm{D}(P_\mu, P_{\mu+\Delta}) &\approx \frac{\partial}{\partial\Delta}\,\mathrm{D}(P_\mu, P_{\mu+\Delta})\Big|_{\Delta=0}\Delta + \frac{1}{2}\,\frac{\partial^2}{\partial\Delta^2}\,\mathrm{D}(P_\mu, P_{\mu+\Delta})\Big|_{\Delta=0}\Delta^2 \\
&= \frac{\partial}{\partial\Delta}\int_\Omega \log\left(\frac{dP_\mu}{dP_{\mu+\Delta}}\right) dP_\mu\Big|_{\Delta=0}\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\int_\Omega \frac{\partial}{\partial\Delta}\log\left(\frac{dP_{\mu+\Delta}}{dP_\mu}\right)\Big|_{\Delta=0} dP_\mu\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\int_\Omega \frac{\partial}{\partial\Delta}\frac{dP_{\mu+\Delta}}{dP_\mu}\Big|_{\Delta=0} dP_\mu\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\frac{\partial}{\partial\Delta}\int_\Omega \frac{dP_{\mu+\Delta}}{dP_\mu} dP_\mu\Big|_{\Delta=0}\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\frac{\partial}{\partial\Delta}\int_\Omega dP_{\mu+\Delta}\Big|_{\Delta=0}\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= \frac{1}{2}I(\mu)\Delta^2\,,
\end{aligned}$$

where $I(\mu)$, introduced in the second line, is called the **Fisher information** of the family $(P_\mu)_\mu$ at μ. Note that if λ is a common dominating measure for

$(P_{\mu+\Delta})$ for Δ small, $dP_{\mu+\Delta} = p_{\mu+\Delta} d\lambda$ and we can write

$$I(\mu) = -\int \frac{\partial^2}{\partial \Delta^2} \log p_{\mu+\Delta}\Big|_{\Delta=0} p_\mu d\lambda \,,$$

which is the form that is usually given in elementary texts. The upshot of all this is that $\mathrm{D}(P_\mu, P_{\mu+\Delta})$ for Δ small is indeed quadratic in Δ, with the scaling provided by $I(\mu)$, and as a result the worst-case regret is always $O(\sqrt{nK})$, provided the class of distributions considered is sufficiently rich and not too bizarre.

2 We have now shown a lower bound that is $\Omega(\sqrt{nK})$, while many of the upper bounds were $O(\log(n))$. There is no contradiction because the logarithmic bounds depended on the inverse suboptimality gaps, which may be very large.

3 Our lower bound was only proven for $n \geq K - 1$. In Exercise 15.2 we ask you to show that when $n < K - 1$ there exists a bandit such that

$$R_n \geq \frac{n(2K - n - 1)}{2K} > \frac{n}{2} \,.$$

4 The method that we arrive at the lower bound can be seen as a generalization (and strengthening) of what is known as **Le Cam's method** in statistics. To see this recall that Eq. (15.2) establishes that for any μ and μ',

$$\inf_\pi \sup_\nu R_n(\pi, \nu) \geq \frac{n\Delta}{8} \exp(-\mathrm{D}(\mathbb{P}_\mu, \mathbb{P}_{\mu'})) \,.$$

Now, Le Cam's method is concerned with a minimax lower bound on the expected error of an estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$ of the a function $\theta(P)$ of a probability distribution P over $\mathcal{X}$, when the estimator is fed with an independent sample $(X_1, \ldots, X_n) \sim P^n$ and the error is measured by a metric $d : \Theta^2 \to [0, \infty)$. In particular, Le Cam's method is to choose $P_0, P_1 \in \mathcal{P}$ to maximize $d(\theta(P_0), \theta(P_1)) \exp(-n\,\mathrm{D}(P_0, P_1))$, on the basis that for *any* $P_0, P_1 \in \mathcal{P}$,

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}[d(\hat{\theta}, \theta(P))] \geq \frac{\Delta}{8} \exp(-n\,\mathrm{D}(P_0, P_1)) \,,$$

where $\Delta = d(\theta(P_0), \theta(P_1))$ and the expectation is with respect to the product measure P^n. Now, there are two differences to this: *(i)* we deal with the sequential setting and *(ii)* having chosen P_0 we choose P_1 based on what a given algorithm (here, estimator) does. This gives the method a much needed extra boost: Without this, the method would in general be unable to capture how the characteristics of $\mathcal{P}$ are reflected in the minimax risk (or regret, in our case).

15.4 Bibliographic remarks

The first work on lower bounds that we know of was the remarkably precise minimax analysis of two-armed Gaussian bandits by Vogel [1960]. The high

probability Pinsker inequality (Theorem 14.2) was first used for bandits by Bubeck et al. [2013b]. As mentioned in the notes, the use of this inequality for proving lower bounds is known as Le Cam's method in statistics Le Cam [1973]. As far as we can tell, the earliest proof of the high probability Pinsker inequality is due to Bretagnolle and Huber [1979], but we also recommend the book by Tsybakov [2008]. The proof of Theorem 15.1 uses the same ideas as Gerchinovitz and Lattimore [2016], while the alternative proof in Exercise 15.1 is essentially due to Auer et al. [1995], who analyzed the more difficult case where the rewards are Bernoulli (see Exercise 15.3). Yu [1997] describes some alternatives to Le Cam's method for the passive, statistical setting. These alternatives can be (and often are) adapted to the sequential setting.

15.5 Exercises

15.1 Here you will prove Theorem 15.1 with a different method. Let $c > 0$ and $\Delta = 2c\sqrt{K/n}$ and for each $i \in \{0, 1, \dots, K\}$ let $\mu^{(i)} \in \mathbb{R}^K$ satisfy $\mu_k^{(i)} = \mathbb{I}\{i = k\}\Delta$. Further abbreviate the notation in the proof of Theorem 15.1 by letting $\mathbb{E}_i[\cdot] = \mathbb{E}_{\mu^{(i)}}[\cdot]$.

(a) Use Pinsker's inequality (Eq. 14.9) and Lemma 15.1 and the result of Exercise 14.1 to show

$$\mathbb{E}_i[T_i(n)] \le \mathbb{E}_0[T_i(n)] + n\sqrt{\frac{1}{4}\Delta^2 \mathbb{E}_0[T_i(n)]} = \mathbb{E}_0[T_i(n)] + c\sqrt{nK\mathbb{E}_0[T_i(n)]}\,.$$

(b) Using the previous part, Jensen's inequality and the identity $\sum_{i=1}^K \mathbb{E}_0[T_i(n)] = n$, show that

$$\sum_{i=1}^K \mathbb{E}_i[T_i(n)] \le n + c\sum_{i=1}^K \sqrt{nK\mathbb{E}_0[T_i(n)]} \le n + cKn\,.$$

(c) Let $R_i = R_n(\pi, G_{\mu^{(i)}})$. Find a choice of $c > 0$ for which

$$\sum_{i=1}^K R_i = \Delta\sum_{i=1}^K (n - \mathbb{E}_i[T_i(n)]) \ge \Delta_i\,(nK - n - cKn)$$
$$= 2c\sqrt{\frac{K}{n}}\,(nK - n - cKn) \ge \frac{nK}{8}\sqrt{\frac{K}{n}}$$

(d) Conclude there exists an $i \in [K]$ such that

$$R_i \ge \frac{1}{8}\sqrt{Kn}\,.$$

The method used in this exercise is borrowed from Auer et al. [2002b] and is closely related to the lower bound technique known as Assouad's method in statistics [Yu, 1997].

15.2 Let $K > 1$ and $n < K$. Prove that for any policy π there exists a Gaussian bandit with unit variance and means $\mu \in [0,1]^K$ such that $R_n(\pi, \nu_\mu) \geq n(2K - n - 1)/(2K) > n/2$.

15.3 Recall from Table 4.1 that $\mathcal{E}_{\mathcal{B}}^K$ is the set of K-armed Bernoulli bandits. Show that there exists a universal constant $c > 0$ such that for any $2 \leq K \leq n$ it holds that:

$$R_n^*(\mathcal{E}_{\mathcal{B}^K}) = \inf_\pi \sup_{\nu \in \mathcal{E}_{\mathcal{B}}^K} R_n(\pi, \nu) \geq c\sqrt{nK}\,.$$

Use the fact that KL divergence is upper bounded by the χ^2-distance (14.12).

15.4 In Chapter 9 we proved that if π is the MOSS policy and $\nu \in \mathcal{E}_{\text{SG}}^K(1)$, then

$$R_n(\pi, \nu) \leq C\left(\sqrt{Kn} + \sum_{i:\Delta_i > 0} \Delta_i\right)\,,$$

where $C > 0$ is a universal constant. Prove that the dependence on the sum cannot be eliminated.

You will have to use that $T_i(t)$ is an integer for all t.

15.5 Let ETC_{nm} be the Explore-Then-Commit policy with inputs n and m respectively (Algorithm 1). Prove that for all m there exists a $\mu \in [0,1]^K$ such that

$$R_n(\text{ETC}_{nm}, \nu_\mu) \geq c \min\left\{n,\, n^{2/3}K^{1/3}\right\}\,,$$

where $c > 0$ is a universal constant.

15.6 Consider the setting of Lemma 15.1 and let X be a random variable and $\mathbb{P}_{\nu X}$ and $\mathbb{P}_{\nu' X}$ be the distributions of X induced by $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$ respectively. Let $\mathcal{F}_t = \sigma(A_1, X_1, \ldots, A_t, X_t)$ and τ be an $(\mathcal{F}_t)$-measurable stopping time. Show that if X is $\mathcal{F}_\tau$-measurable, then

$$\text{D}(\mathbb{P}_{\nu X}, \mathbb{P}_{\nu' X}) \leq \sum_{i=1}^K \mathbb{E}_\nu[T_i(\tau)]\,\text{D}(P_i, P_i')\,.$$