13 Lower Bounds: Basic Ideas

We start the block on lower bounds by considering stochastic bandits. Let $\mathcal{E}$ be a set of stochastic bandits and π be a policy. The **worst case regret** of policy π on environment class $\mathcal{E}$ is

$$R_n(\pi, \mathcal{E}) = \sup_{\nu \in \mathcal{E}^K} R_n(\pi, \nu) \,.$$

Let Π be the set of all policies. The **minimax regret** is

$$R_n^*(\mathcal{E}) = \inf_{\pi \in \Pi} R_n(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu) \,.$$

A policy is called **minimax optimal** for $\mathcal{E}$ if $R_n(\pi, \mathcal{E}^K) = R_n^*(\mathcal{E})$. The value $R_n^*(\mathcal{E})$ is of interest by itself. A small value of $R_n^*(\mathcal{E})$ indicates that the underlying bandit problem is less challenging in the worst-case sense. A core activity in bandit theory is to understand what makes $R_n^*(\mathcal{E})$ large or small, often focusing on its behavior as a function of the number of rounds n.

Minimax optimality is not a property of a policy alone. It is a property of a policy together with a set of environments and a horizon.

Finding a minimax policy is generally too computationally expensive to be practical. For this reason we almost always settle for a policy that is nearly minimax optimal. One of the main results of this part is a proof of the following theorem, which together with Theorem 9.1 shows that Algorithm 6 from Chapter 9 is minimax optimal up to constant factors for 1-subgaussian bandits with suboptimality gaps in $[0, 1]$.

THEOREM 13.1 *Let $\mathcal{E}^K$ be the set of K-armed Gaussian bandits with unit variance and means $\mu \in [0, 1]^K$. Then there exists a constant $c > 0$ such that for all $K > 1$ and $n \geq K$ it holds that*

$$R_n^*(\mathcal{E}^K) \geq c\sqrt{(K-1)n} \,.$$

We will prove this theorem in Chapter 15, but first we given an informal justification. Let $X_1, X_2, \ldots, X_n$ be an observed sequence of independent Gaussian random variables with unknown mean μ and known variance 1. Assume you are told that μ takes on one of two values: $\mu = 0$ or $\mu = \Delta$ for some known $\Delta > 0$.

Your task is to guess the value of μ. Let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean, which is Gaussian with mean μ and variance $1/n$. While it is not immediately obvious how easy this task is, intuitively we expect the optimal decision is to predict that $\mu = 0$ if $\hat{\mu}$ is closer to 0 than to Δ, and otherwise to predict $\mu = \Delta$. For large n we expect our prediction will probably be correct. Supposing that $\mu = 0$ (the other case is symmetric), then the prediction will be wrong only if $\hat{\mu} \geq \Delta/2$. Using the fact that $\hat{\mu}$ is Gaussian with mean $\mu = 0$ and variance $1/n$, combined with known bounds on the Gaussian tail probabilities (see Abramowitz and Stegun, 1964) leads to

$$\frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 4}} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{n\Delta^2}{8}\right) \leq \mathbb{P}\left(\hat{\mu} \geq \frac{\Delta}{2}\right) \leq \frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 8/\pi}} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{n\Delta^2}{8}\right).$$

The upper and lower bounds are incredibly close, differing only in the constant in the square root of the denominator. One might believe that the decision procedure could be improved, but the symmetry of the problem makes this seem improbable. The formula exhibits the expected behaviour, which is that once n is large relative to $8/\Delta^2$, then the probability that this procedure fails drops exponentially with further increases in n. But the lower bound also shows that if n is small relative to $8/\Delta^2$, then the procedure fails with constant probability.

The problem described is called hypothesis testing and the ideas underlying the argument above are core to many impossibility result in statistics. The next task is to reduce our bandit problem to hypothesis testing. The high level idea is to select two bandit problem instances in such a way that the following two conditions hold simultaenously:

1 *Competition*: A sequence of actions that is good for one bandit is not good for the other.
2 *Similarity*: The instances are 'close' enough that the policy interacting with either of the two instances cannot statistically identify the true bandit with reasonable statistical accuracy.

The two requirements are clearly conflicting. The first makes us want to choose instances with means $\mu, \mu' \in [0,1]^K$ that are far from each other, while the second requirement makes us want to choose them to be close to each other. The lower bound will follow by optimizing this tradeoff.

Let us start to make things concrete by choosing bandits $\nu = (P_i)_{i=1}^K$ and $\nu' = (P'_i)_{i=1}^K$ where $P_i = \mathcal{N}(\mu_i, 1)$ and $P'_i = \mathcal{N}(\mu'_i, 1)$ are Gaussian and $\mu, \mu' \in [0,1]^K$. In order to prove a lower bound it suffices to show that for every strategy π there exists a choice of μ and μ' such that

$$\max\{R_n(\pi, \nu), R_n(\pi, \nu')\} \geq c\sqrt{Kn}\,,$$

where $c > 0$ is a universal constant. Let $\Delta > 0$ be a constant to be tuned

subsequently and choose $\mu = (\Delta, 0, 0, \dots, 0)$, which means that the first arm is optimal in instance ν and

$$R_n(\pi, \nu) = (n - \mathbb{E}[T_1(n)])\Delta\,, \tag{13.1}$$

where the expectation is taken with respect to the induced measure on the sequence of outcomes when π interacts with ν. Now we need to choose μ' to satisfy the two requirements above. Since we want ν and ν' to be hard to distinguish and yet have different optimal actions, we should make μ' as close to μ except in a coordinate where π expects to explore the least. To this end, let

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

be the suboptimal arm in ν that π expects to play least often. By the pigeonhole principle and the fact that $\sum_i \mathbb{E}[T_i(n)] = n$, it must hold that

$$\mathbb{E}[T_i(n)] \leq \frac{n}{K-1}\,.$$

Then define $\mu' \in \mathbb{R}^K$ by

$$\mu'_j = \begin{cases} \mu_j & \text{if } j \neq i \\ 2\Delta & \text{otherwise}\,. \end{cases}$$

The regret in this bandit is

$$R_n(\pi, \nu') = \Delta\mathbb{E}'[T_1(n)] + \sum_{j \notin 1,i} 2\Delta\mathbb{E}'[T_j(n)] \geq \Delta\mathbb{E}'[T_1(n)]\,, \tag{13.2}$$

where $\mathbb{E}'[\cdot]$ is the expectation operator on the sequence of outcomes when π interacts with ν'. So now we have the following situation: The strategy π interacts with either ν or ν' and when interacting with ν it expects to play arm i at most $n/(K-1)$ times. But the two instances only differ when playing arm i. The time has come to tune Δ. Because the strategy expects to play arm i only about $n/(K-1)$ times, taking inspiration from the previous discussion on distinguishing samples from Gaussian distributions with different means, we will choose

$$\Delta = \sqrt{\frac{1}{\mathbb{E}[T_i(n)]}} \geq \sqrt{\frac{K-1}{n}}\,.$$

If we are prepared to ignore the fact that $T_i(n)$ is a random variable and take for granted the claims in the first part of the chapter, then with this choice of Δ the strategy cannot distinguish between instances ν and ν' and in particular we expect that $\mathbb{E}[T_1(n)] \approx \mathbb{E}'[T_1(n)]$. If $\mathbb{E}[T_1(n)] < n/2$, then by Eq. (13.1) we have

$$R_n(\pi, \nu) \geq \frac{n}{2}\sqrt{\frac{K-1}{n}} = \frac{1}{2}\sqrt{n(K-1)}\,.$$

On the other hand, if $\mathbb{E}[T_1(n)] \geq n/2$, then

$$R_n(\pi, \nu') \geq \Delta\mathbb{E}'[T_1(n)] \approx \Delta\mathbb{E}[T_i(n)] \geq \frac{1}{2}\sqrt{n(K-1)}\,,$$

which completes our heuristic argument that there exists a universal constant $c > 0$ such that

$$R_n^*(\mathcal{E}^K) \geq c\sqrt{nK}\,.$$

We have been sloppy in many places: The claims in the first part of the chapter have not been proven yet and $T_i(n)$ is a random variable. Before we can present the rigorous argument we need a chapter to introduce some ideas from information theory. Readers already familiar with these concepts can skip to Chapter 15 for the proof of Theorem 13.1.

13.1 Notes

1 The worst-case regret has a game-theoretic interpretation. Imagine a game between a protagonist and an antagonist that works as follows: For $K > 1$ and $n \geq K$ the protagonist proposes a bandit policy π. The antagonist looks at the policy and chooses a bandit ν from the class of environments considered. The utility for the antagonist is the expected regret and for the protagonist it is the negation of the expected regret, which makes this a zero-sum game. Both players aim to maximizing their payoffs. The game is completely described by n and $\mathcal{E}$. One characteristic value in a game is its minimax value. As described above, this is a sequential game (the protagonist moves first, then the antagonist). The minimax value of this game from the perspective of the antagonist is exactly $R_n^*(\mathcal{E})$, while for the protagonist is $\sup_\pi \inf_\nu(-R_n(\pi, \nu)) = -R_n^*(\mathcal{E})$.

2 We mentioned that finding the minimax optimal policy is usually computationally infeasible. In fact it is not clear we should even try. In classical statistics it often turns out that minimizing the worst case leads to a flat risk profile. In the language of bandits this would mean that the regret is the same for every bandit (where possible). What we usually want in practice is to have low regret against 'easy' bandits and larger regret against 'hard' bandits. The analysis in Part II suggests that easy bandits are those where the suboptimality gaps are large or very small. There is evidence to suggest that the exact minimax optimal strategy may not exploit these easy instances, so in practice one might prefer to find a policy that is nearly minimax optimal and has much smaller regret on easy bandits. We will tackle questions of this nature in Chapter 16.

3 The regret on a class of bandits $\mathcal{E}$ is a multi-objective criteria. Some policies will be good for some instances and bad on others, and there are clear trade-offs. One way to analyze the performance in a multi-objective setting is called **Pareto optimality**. A policy is Pareto optimal if there does not exist another policy that is a strict improvement. More precisely, if there does not exist a π' such that $R_n(\pi', \nu) \leq R_n(\pi, \nu)$ for all $\nu \in \mathcal{E}$ and $R_n(\pi', \nu) < R_n(\pi, \nu)$ for at least one instance $\nu \in \mathcal{E}$.

4 When we say a policy is minimax optimal up to constant factors for finite-armed 1-subgaussian bandits with suboptimality gaps in $[0, 1]$ we mean there exists a $C > 0$ such that

$$\frac{R_n(\pi, \mathcal{E}^K)}{R_n^*(\mathcal{E}^K)} \leq C \text{ for all } K \text{ and } n\,,$$

where $\mathcal{E}^K$ is the set of K-armed 1-subgaussian bandits with suboptiality gaps in $[0, 1]$. We often say a policy is minimax optimal up to logarithmic factors, by which we mean that

$$\frac{R_n(\pi, \mathcal{E}^K)}{R_n^*(\mathcal{E}^K)} \leq C(n, K) \text{ for all } K \text{ and } n\,,$$

where $C(n, K)$ is logarithmic in n and K. We hope the reader will forgive us for not always specifying in the text exactly what is meant and promise that statements of theorems will always be precise.

13.2 Exercises

13.1 Let $\mathbb{P}_\mu = \mathcal{N}(\mu, 1)$ be the Gaussian measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ with mean $\mu \in \{0, \Delta\}$ and unit variance. Let $X : \mathbb{R} \to \mathbb{R}$ be the identity random variable $(X(\omega) = \omega)$. For decision rule $d : \mathbb{R} \to \{0, \Delta\}$ define risk

$$R(d) = \max_{\mu \in \{0,\Delta\}} \mathbb{P}_\mu(d(X) \neq \mu)\,,$$

Prove that $R(d)$ is minimized by $d(x) = \operatorname{argmin}_{\tilde{\mu} \in \{0,\mu\}} |X - \tilde{\mu}|$.

13.2 Let $K > 1$ and $\mathcal{E} = \mathcal{E}_{\mathcal{N}}^K(1)$ be the set of Gaussian bandits with unit variance. Find a Pareto optimal policy for this class.

Think about simple policies (not necessarily good ones) and use the definition.