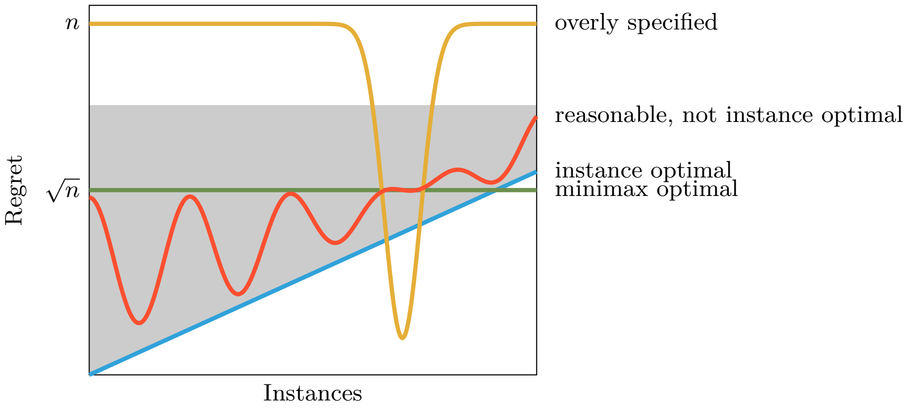
16 Instance Dependent Lower Bounds

In the last chapter we proved a lower bound on the minimax regret for subgaussian bandits with suboptimality gaps in $[0, 1]$. Such bounds serve as a useful measure of the robustness of a policy, but are often excessively conservative. This chapter is devoted to understanding **instance-dependent** lower bounds, which try to capture the optimal performance of a policy on a specific bandit instance.

Because the regret is a multi-objective criteria, an algorithm designer might try and design algorithms that perform well on one kind of instance or another. An extreme example is the policy that chooses $A_t = 1$ for all t, which suffers zero regret when the first arm is optimal and linear regret otherwise. This is a harsh tradeoff with the price for reducing the regret from logarithmic to zero on just a few instances being linear regret on the remainder. Surprisingly, this is the nature of the game in bandits. One can assign a measure of difficulty to each instance such that policies performing overly well relative to this measure on some instances pay a steep price on others. The situation is illustrated in the figure below.



On the x-axis the instances are ordered according to the measure of difficulty and the y-axis shows the regret (on some scale). In the previous chapter we proved that no policy can be entirely below the horizontal 'minimax optimal' line. The results in this chapter show that if the regret of a policy is below the 'instance optimal' line at any point, then it must have regret above the shaded region for other instances. For example, the 'overly specified' policy. In finite-time the situation is a little messy, but if one pushes these ideas to the limit, then for many classes of bandits one can define a precise notion of instance-dependent optimality.

16.1 Asymptotic bounds

We need to define exactly what is meant by a reasonable policy. If one is only concerned with asymptotics, then a rather conservative definition suffices.

Definition 16.1 A policy π is called **consistent** over a class of bandits $\mathcal{E}$ if for all $\nu \in \mathcal{E}$ and $p > 0$ it holds that

$$\lim_{n\to\infty} \frac{R_n(\pi, \nu)}{n^p} = 0\,. \tag{16.1}$$

The class of consistent policies over $\mathcal{E}$ is denoted by $\Pi_{\text{cons}}(\mathcal{E})$.

Theorem 7.1 shows that UCB is consistent over $\mathcal{E}^K_{\text{SG}}(1)$. The strategy that always chooses the first action is not consistent on any class $\mathcal{E}$ unless $K = 1$ or $\mathcal{E}$ is so restrictive that the first arm is optimal action for every $\nu \in \mathcal{E}$.

Consistency is an asymptotic notion. A policy could be consistent and yet play $A_t = 1$ for all $t \leq 10^{100}$. For this reason an assumption on consistency is insufficient to derive nonasymptotic lower bounds. Later we introduce a finite-time version of consistency that allows us to prove finite-time instance-dependent lower bounds.

A class $\mathcal{E}$ of stochastic bandits is **unstructured** if $\mathcal{E} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_K$ with $\mathcal{C}_1, \ldots, \mathcal{C}_K$ sets of distributions. The main theorem of this chapter is a generic lower bound that applies to any unstructured class of stochastic bandits. After the proof we will see some applications to specific classes. Let $\mathcal{C}$ be a set of distributions with finite means and let $\mu : \mathcal{C} \to \mathbb{R}$ be the function that maps $P \in \mathcal{C}$ to its mean. Let $\alpha \in \mathbb{R}$ and $P \in \mathcal{C}$ have $P(\mu) < \alpha$ and define

$$d_{\mathcal{C}}(P, \alpha) = \inf_{P' \in \mathcal{C}} \left\{ \mathrm{D}(P, P') : \mu(P') > \alpha \right\}\,.$$

Recall that $P_i(\nu)$ is the distribution of rewards for the ith arm of bandit ν and $\mu_i(\nu)$ is its mean and $\mu^*(\nu) = \max_i \mu_i(\nu)$ and $\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)$.

Theorem 16.1 *Let $\mathcal{E} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_K$ and $\pi \in \Pi_{cons}(\mathcal{E})$ be a consistent policy over $\mathcal{E}$. Then for all $\nu \in \mathcal{E}$ it holds that*

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \geq c^*(\nu, \mathcal{E}) = \sum_{i:\Delta_i(\nu)>0} \frac{\Delta_i(\nu)}{d_{\mathcal{C}_i}(P_i(\nu), \mu^*(\nu))}\,. \tag{16.2}$$

Proof Abbreviate $P_i = P_i(\nu)$ and $\mu_i = \mu_i(\nu)$ and $\Delta_i = \Delta_i(\nu)$ and $\mu^* = \mu^*(\nu)$ and $d_i = d_{\mathcal{C}_i}(P_i, \mu^*)$. The result will follow from Lemma 4.2 and by showing that for any suboptimal arm i it holds that

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu\pi}[T_i(n)]}{\log(n)} \geq \frac{1}{d_i}\,.$$

Fix a suboptimal arm i and let $\varepsilon > 0$ be arbitrary and $\nu' \in \mathcal{E}$ be a bandit with $P_j(\nu') = P_j$ for $j \neq i$ and $P_i(\nu')$ be such that $\mathrm{D}(P_i, P_i') \leq d_i + \varepsilon$ and $\mu(P_i') > \mu^*$, which exists by the definition of d_i. Then by Lemma 15.1 we have $\mathrm{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi}) \leq \mathbb{E}_{\nu\pi}[T_i(n)](d_i + \varepsilon)$ and by Theorem 14.2 for any event A

$$\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu'\pi}(A) \geq \frac{1}{2}\exp\left(-\mathrm{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi})\right) \geq \frac{1}{2}\exp\left(-\mathbb{E}_{\nu\pi}[T_i(n)](d_i+\varepsilon)\right)\,.$$

Now choose $A = \{T_i(n) > n/2\}$ and let $R_n = R_n(\pi, \nu)$ and $R_n' = R_n(\pi, \nu')$. Then

$$\begin{aligned} R_n + R_n' &\geq \frac{n}{2}\left(\mathbb{P}_{\nu\pi}(A)\Delta_i + \mathbb{P}_{\nu'\pi}(A^c)(\mu_i' - \mu^*)\right) \\ &\geq \frac{n}{2}\min\left\{\Delta_i, \mu_i' - \mu^*\right\}\left(\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu'\pi}(A^c)\right) \\ &\geq \frac{n}{2}\min\left\{\Delta_i, \mu_i' - \mu^*\right\}\exp\left(-\mathbb{E}_{\nu\pi}[T_i(n)](d_i+\varepsilon)\right)\,. \end{aligned}$$

Rearranging and taking the limit inferior leads to

$$\begin{aligned} \liminf_{n\to\infty} \frac{\mathbb{E}_{\nu\pi}[T_i(n)]}{\log(n)} &\geq \frac{1}{d_i+\varepsilon}\liminf_{n\to\infty}\frac{\log\left(\frac{n\min\{\Delta_i, \mu_i'-\mu^*\}}{2(R_n+R_n')}\right)}{\log(n)} \\ &= \frac{1}{d_i+\varepsilon}\left(1 - \limsup_{n\to\infty}\frac{\log\left(R_n + R_n'\right)}{\log(n)}\right) = \frac{1}{d_i+\varepsilon}\,, \end{aligned}$$

where the last equality follows from the definition of consistency, which says that for any $p > 0$ there exists a constant C_p such that for sufficiently large n, $R_n + R_n' \leq C_p n^p$, which implies that

$$\limsup_{n\to\infty}\frac{\log\left(R_n+R_n'\right)}{\log(n)} \leq \limsup_{n\to\infty}\frac{p\log(n) + \log(C_p)}{\log(n)} = p\,,$$

which gives the result since $p > 0$ was arbitrary and by taking the limit as ε tends to zero. □

The next theorem gives $d_{\mathcal{C}}(P, \alpha)$ for common choices of $\mathcal{C}$.

THEOREM 16.2 *The following hold:*

(a) Let $\sigma^2 > 0$ *and* $\mathcal{C} = \left\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\right\}$*, then*

$$d_{\mathcal{C}}(\mathcal{N}(\mu, \sigma^2), \alpha) = \frac{(\mu-\alpha)^2}{2\sigma^2}\,.$$

(b) Let $\mathcal{C} = \left\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\right\}$*, then*

$$d_{\mathcal{C}}(\mathcal{N}(\mu, \sigma^2), \alpha) = \frac{1}{2}\log\left(1 + \frac{(\mu-\alpha)^2}{\sigma^2}\right)\,.$$

(c) Let $\mathcal{C} = \{\mathcal{B}(\mu) : \mu \in [0,1]\}$, then

$$d_{\mathcal{C}}(\mathcal{B}(\mu), \alpha) = \mu \log\left(\frac{\mu}{\alpha}\right) + (1-\mu)\log\left(\frac{1-\mu}{1-\alpha}\right) .$$

(d) Let $\mathcal{C} = \{\mathcal{U}(a,b) : a, b \in \mathbb{R}\}$, then

$$d_{\mathcal{C}}(\mathcal{U}(a,b), \alpha) = \log\left(1 + \frac{2((a+b)/2 - \alpha)^2}{b-a}\right) .$$

Proof of part (a) Fix $\sigma^2 > 0$ and note that the class $\mathcal{C} = \left\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\right\}$ is parameterised by the mean. Therefore for any $\mu \in \mathbb{R}$ and $\alpha > \mu$ we have

$$d_{\mathcal{C}}(\mathcal{N}(\mu, \sigma^2), \alpha) = \inf_{\theta > \alpha} \mathrm{D}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\theta, \sigma^2)) = \inf_{\theta > \alpha} \frac{(\mu-\theta)^2}{2\sigma^2} = \frac{(\mu-\alpha)^2}{2\sigma^2} . \quad \square$$

The reader is asked to complete the remaining parts in Exercise 16.1. It appears that the lower bound and definition of $c^*(\nu, \mathcal{E})$ are quite fundamental quantities in the sense that for most classes $\mathcal{E}$ it appears there exists a policy π for which

$$\lim_{n\to\infty} \frac{R_n(\pi, \nu)}{\log(n)} = c^*(\nu, \mathcal{E}) \qquad \text{for all } \nu \in \mathcal{E} . \tag{16.3}$$

This justifies calling a policy **asymptotically optimal** on class $\mathcal{E}$ if Eq. (16.3) holds. For example, UCB from Chapter 8 and KL-UCB from Chapter 10 are asymptotically optimal for $\mathcal{E}_{\mathcal{N}}^K(1)$ and $\mathcal{E}_{\mathcal{B}}^K$ respectively.

16.2 Finite-time bounds

The proofs that follow use the same technique as what we already saw. For future reference we extract the common part, which summarizes what can be obtained by chaining the high-probability Pinsker inequality with the divergence decomposition lemma.

LEMMA 16.1 *Let $\nu = (P_i)$ and $\nu' = (P_i')$ be K-action stochastic bandits that differ only in the distribution of the reward for action $i \in [K]$. Assume that i is suboptimal in ν and uniquely optimal in ν'. Let $\lambda = \mu_i(\nu') - \mu_i(\nu)$. Then for any policy π,*

$$\mathbb{E}_{\nu\pi}[T_i(n)] \geq \frac{\log\left(\frac{\min\{\lambda - \Delta_i(\nu), \Delta_i(\nu)\}}{4}\right) + \log(n) - \log(R_n(\nu) + R_n(\nu'))}{\mathrm{D}(P_i, P_i')} . \tag{16.4}$$

The lemma holds for finite n and any ν and can be used to derive finite-time instance-dependent lower bounds for any environment class $\mathcal{E}$ that is rich enough. The following result provides a finite-time instance-dependence bound for Gaussian bandits where the asymptotic notion of consistency is replaced by an assumption that the minimax regret is not too large. This assumption alone is enough to show that no policy that is remotely close to minimax optimal can be much better than UCB on any instance.

THEOREM 16.3 *Let $C, p > 0$ and π be a policy such that $R_n(\pi, \nu) \leq Cn^p$ for all $\nu \in \mathcal{E}_{\mathcal{N}}^K$. Then for any $\nu \in \mathcal{E}_{\mathcal{N}}^K$ and $\varepsilon \in (1, 2)$ it holds that*

$$R_n(\pi, \nu) \geq 2 \sum_{i:\Delta_i > 0} \left(\frac{(1-p)\log(n) + \log(\frac{\varepsilon \Delta_i}{8C})}{\Delta_i} \right)^+ , \tag{16.5}$$

where $(x)^+ = \max(x, 0)$ is the positive part of $x \in \mathbb{R}$.

Proof Let i be suboptimal in ν and choose $\nu' \in \mathcal{E}_{\mathcal{N}}^K$ such that $\mu_j(\nu') = \mu_j(\nu)$ for $j \neq i$ and $\mu_j(\nu') = \mu_i + \Delta_i(1+\varepsilon)$. Then by Lemma 16.1 with $\lambda = \Delta_i(1+\varepsilon)$,

$$\begin{aligned}\mathbb{E}_{\nu\pi}[T_i(n)] &\geq \frac{2}{\Delta_i^2(1+\varepsilon)^2} \left(\log\left(\frac{n}{2Cn^p}\right) + \log\left(\frac{\min\{\lambda - \Delta_i, \Delta_i\}}{4}\right)\right) \\ &= \frac{2}{\Delta_i^2(1+\varepsilon)^2} \left((1-p)\log(n) + \log\left(\frac{\varepsilon\Delta_i}{8C}\right)\right) .\end{aligned}$$

Plugging this into the basic regret decomposition identity (Lemma 4.2) gives the result. □

When $p = 1/2$ the leading term in this lower bound is approximately half that of the asymptotic bound. This effect may be real: the class of policies considered is larger than in the asymptotic lower bound and so there is the possibility that the policy that is best tuned for a given environment achieves a smaller regret.

16.3 Notes

1 We mentioned that for most classes $\mathcal{E}$ there is a policy satisfying Eq. (16.3). Its form is derived from the lower bound, and by making some additional assumptions on the underlying distributions. For details, see the article by Burnetas and Katehakis [1996], which is also the original source of Theorem 16.1.

2 The analysis in this chapter only works for unstructured classes. Without this assumption a policy can potentially learn about the reward from one arm by playing other arms and this greatly reduces the regret. Lower bounds for structured bandits are more delicate and will be covered on a case-by-case basis in subsequent chapters.

3 The classes analyzed in Theorem 16.2 are all parametric, which makes the calculation possible analytically. There has been relatively little analysis in the non-parametric case, but we know of three exceptions for which we simply refer the reader to the appropriate source. The first is the class of distributions with bounded support: $\mathcal{C} = \{P : \text{Supp}(P) \subseteq [0,1]\}$, which has been analyzed exactly [Honda and Takemura, 2010]. The second is the class of distributions with semi-bounded support, $\mathcal{C} = \{P : \text{Supp}(P) \subseteq (-\infty, 1]\}$ [Honda and Takemura, 2015]. The third is the class of distributions with bounded kurtosis, $\mathcal{C} = \{P : \text{Kurt}_{X\sim P}[X] \leq \kappa\}$. For details see Lattimore [2017].

16.4 Bibliographic remarks

Asymptotic optimality via a consistency assumption first appeared in the seminal paper by Lai and Robbins [1985], which was later generalized by Burnetas and Katehakis [1996]. In terms of upper bounds, there now exist policies that are asymptotic optimal for single-parameter exponential families [Cappé et al., 2013]. Until recently, there were no results on asymptotic optimality for multi-parameter classes of reward distributions. There has been some progress on this issue recently for the Gaussian distribution with unknown mean and variance [Cowan et al., 2015] and for the uniform distribution [Cowan and Katehakis, 2015]. There are plenty of open questions related to asymptotically optimal strategies for nonparametric classes of reward distributions. When the reward distributions are discrete and finitely supported an asymptotically optimal policy is given by Burnetas and Katehakis [1996], though the precise constant is hard to interpret. A relatively complete solution is available for classes with bounded support [Honda and Takemura, 2010]. Already for the semi-bounded case things are getting murky [Honda and Takemura, 2015]. One of the authors thinks that classes with bounded kurtosis are quite interesting, but here things are only understood up to constant factors [Lattimore, 2017]. An asymptotic variant of Theorem 16.3 is by Salomon et al. [2013]. Finite-time instance-dependent lower bounds have been proposed by several authors including Kulkarni and Lugosi [2000] for two arms and Garivier et al. [2016c], Lattimore [2018] for the general case.

16.5 Exercises

16.1 Prove parts (b), (c) and (d) of Theorem 16.2.

16.2 Let $\mathcal{R}(\mu)$ be the shifted Rademacher distribution, which for $\mu \in \mathbb{R}$ and $X \sim \mathcal{R}(\mu)$ is characterized by $\mathbb{P}(X = \mu + 1) = \mathbb{P}(X = \mu - 1) = 1/2$.

(a) Show that $d_{\mathcal{C}}(\mathcal{R}(\mu), \alpha) = \infty$ for any $\mu < \alpha$.

(b) Design a policy π for bandits with shifted Rademacher rewards such that the regret is bounded by

$$R_n(\pi, \nu) \leq CK \qquad \text{for all } n \text{ and } \nu \in \times\mathcal{C}\,,$$

where $C > 0$ is a universal constant.

(c) The results from parts (a) and (b) seem to contradict the heuristic analysis in Note 1 at the end of Chapter 15. Explain.

16.3 Let π be a consistent policy for a single parameter exponential family as explained in Exercise 10.4 in Chapter 10. Prove the upper bound given in part (h) is tight.

16.4 Let $\mathcal{C} = \left\{P : \text{there exists a } \sigma^2 \geq 0 \text{ such that } P \text{ is } \sigma^2\text{-subgaussian}\right\}$.

(a) Find a distribution P such that $P \notin \mathcal{C}$.

(b) Suppose that $P \in \mathcal{C}$ has mean $\mu \in \mathbb{R}$. Prove that $d_{\mathcal{C}}(P, \alpha) = 0$ for all $\alpha > \mu$.

(c) Let $\mathcal{E} = \{(P_i) : P_i \in \mathcal{C} \text{ for all } 1 \leq i \leq K\}$. Prove that if $K > 1$, then for all consistent policies π,

$$\liminf_{n \to \infty} \frac{R_n(\pi, \nu)}{\log(n)} = \infty \qquad \text{for all } \nu \in \mathcal{E}\,.$$

(d) Let $f : \mathbb{N} \to [0, \infty)$ be any monotone increasing function with $\lim_{n\to\infty} f(n)/\log(n) = \infty$. Prove there exists a policy π such that

$$\limsup_{n \to \infty} \frac{R_n(\pi, \nu)}{f(n)} = 0 \qquad \text{for all } \nu \in \mathcal{E}\,,$$

where $\mathcal{E}$ is as in the previous part.

(e) Conclude there exists a consistent policy for $\mathcal{E}$.

16.5 Use Lemma 16.1 to prove Theorem 15.1, possibly with different constants.

16.6 Let $K = 2$ and for $\nu \in \mathcal{E}_{\mathcal{N}}^2$ let $\Delta(\nu) = \max\{\Delta_1(\nu), \Delta_2(\nu)\}$. Suppose that π is a policy such that for all $\nu \in \mathcal{E}_{\mathcal{N}}^2$ with $\Delta(\nu) \leq 1$ it holds that

$$R_n(\pi, \nu) \leq \frac{C \log(n)}{\Delta(\nu)}\,. \tag{16.6}$$

(a) Give an example of a policy satisfying Eq. (16.6).

(b) Assume that $i = 2$ is suboptimal for ν and $\alpha \in (0, 1)$ be such that $\mathbb{E}_{\nu\pi}[T_2(n)] = \frac{1}{2\Delta(\nu)^2} \log(\alpha)$. Let ν' be the alternative environment where $\mu_1(\nu') = \mu_1(\nu)$ and $\mu_2(\nu') = \mu_1(\nu) + 2\Delta(\nu)$. Show that

$$\exp(-\operatorname{D}(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi})) = \frac{1}{\alpha}\,.$$

(c) Let A be the event that $T_2(n) \geq n/2$. Show that

$$\mathbb{P}_{\nu\pi}(A) \leq \frac{2C\log(n)}{n\Delta^2} \quad \text{and} \quad \mathbb{P}_{\nu'\pi}(A) \geq \frac{1}{2\alpha} - \frac{2C\log(n)}{n\Delta^2}\,.$$

(d) Show that

$$R_n(\pi, \nu') \geq \frac{n\Delta}{2}\left(\frac{1}{2\alpha} - \frac{2C\log(n)}{n\Delta^2}\right).$$

(e) Show that $\alpha \geq \frac{n\Delta^2}{8C\log(n)}$ and conclude that

$$R_n(\pi, \nu) \geq \frac{1}{2\Delta(\nu)} \log\left(\frac{n\Delta^2}{8C\log(n)}\right).$$

(f) Generalize the argument to an arbitrary number of arms.

16.7 Let $K > 1$ and $p \in [0, 1)$ and π be a policy such that for all $\mathcal{E}_{\mathcal{N}}^K$ so that for all $\nu \in \mathcal{E}_{\mathcal{N}}^K$ it holds that

$$\limsup_{n \to \infty} \frac{R_n(\pi, \nu)}{\log(n)} = \sum_{i:\Delta_i > 0} \frac{2(1+p)}{\Delta_i}\,.$$

Let $\hat{R}_n(\pi, \nu) = n\mu^*(\nu) - \sum_{t=1}^n \mu_{A_t}(\nu)$ be the random regret and prove that

$$\limsup_{n \to \infty} \sup_{\nu \in \mathcal{E}_n} \frac{\log(\mathbb{V}[\hat{R}_n(\pi, \nu)])}{(1-p)\log(n)} \geq 1\,.$$