Algorithmic Foundations of Learning **Lecture 16**

Minimax Lower Bounds and Hypothesis Testing

Lecturer: Patrick Rebeschini *Version: December 1st 2019*

16.1 Introduction

At the introductory level, statistics is traditionally taught using *asymptotic* tools, such as the Law of Large Numbers and the Central Limit Theorem, and the latter is used to build confidence intervals and perform hypothesis testing. In this setting one is interested in understanding what happens when the amount of data points n at our disposal goes to infinity.

In this course we have developed corresponding *non-asymptotic* tools for the case when n is finite. In fact, much of this course was about showing how non-asymptotic tools are essential both for establishing upper bounds on the rate of convergence of algorithms and for designing algorithms (e.g., UCB). In particular, we developed non-asymptotic uniform Laws of Large Numbers and concentration inequalities for functions of many random variables which lead, respectively, to notions of complexity to bound the expected (uniform) generalization error of the empirical risk minimization rule and to establish confidence intervals that captures the behavior of algorithms with high probability.

Today we complete the picture and look at hypothesis testing from a non-asymptotic point of view. We show that hypothesis testing is a key tool to prove *lower bounds* (in the minimax sense) for estimators. As an example, we will derive lower bounds for the Multi-Armed Bandit problem introduced last time, and confirm that UCB yields quasi-optimal error rates (optimal modulo a log term) in the worst case scenario of distribution-independent bounds.

In a way, today's discussion on hypothesis testing completes the picture and shows that the non-asymptotic version of classical statistical tools are at the foundation of the design and analysis of the modern algorithmic paradigms used in machine learning. Indeed, statistical thinking permeates machine learning is many other ways. The very use of randomness in the design of algorithms can be traced back to the field of statistics (although these days the field of "randomized algorithm" is more and more embraced by computer scientists). For instance, the Robbins and Monro's 1951 seminar paper [1] on stochastic gradient descent (which is one of the most widely used randomized algorithms in machine learning) appeared in The Annals of Mathematical Statistics, a statistics journal.

16.2 Binary Hypothesis Testing and Neyman Pearson Lemma

Minimax lower bounds are based on information theory and on the basic fact that a certain amount of information is *needed* to distinguish probability distributions from samples. Consider the following classical setup of binary hypothesis testing in statistics. Assume that you are given a random variable $X \in \mathcal{X}$ that either comes from a distribution $\mathbf{P}$ (this forms the so-called null hypothesis H_0) or from a distribution $\mathbf{Q}$ (the alternative hypothesis H_1). Let us define a *test* as any function $f : \mathcal{X} \to \{0, 1\}$ that, given the sample X, indicate which hypothesis should be true $f(X) \in \{0, 1\}$. Any such test can commit two types of error. A type I error if $f(X) = 1$ when $X \sim \mathbf{P}$, and a type II error if $f(X) = 0$ when $X \sim \mathbf{Q}$. Henceforth, we say that $\mathbf{P}$ and $\mathbf{Q}$ have densities p and q with respect to a measure ρ if the following holds for any measurable

event E:

$$\mathbf{P}(E) = \int \rho(\mathrm{d}x) p(x) \mathbf{1}_E(x),$$
$$\mathbf{Q}(E) = \int \rho(\mathrm{d}x) q(x) \mathbf{1}_E(x).$$

This general way of writing integrals (with respect to the measure ρ) is useful as it can be used to describe the case of standard integrals, or sums, or both! If ρ is the Lebesgue measure we write $\rho(\mathrm{d}x) = \mathrm{d}x$ and we recover the classical integrals we are used to. On the other hand, if we have $\rho(\mathrm{d}x) = \sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)$ for a collection of points $x_1, \dots, x_n \in \mathcal{X}$, where δ_{x_i} is the Dirac measure sitting at x_i, we recover sums:

$$\mathbf{P}(E) = \int \rho(\mathrm{d}x) p(x) \mathbf{1}_E(x) = \int \sum_{i=1}^n \delta_{x_i}(\mathrm{d}x) p(x) \mathbf{1}_E(x) = \sum_{i=1}^n \int \delta_{x_i}(\mathrm{d}x) p(x) \mathbf{1}_E(x)$$
$$= \sum_{i=1}^n p(x_i) \mathbf{1}_E(x_i) = \sum_{i : x_i \in E} p(x_i).$$

The Neyman Pearson Lemma shows that *any* test f commits one of the two types of error (type I and type II) with strictly positive probability unless $\mathbf{P}$ and $\mathbf{Q}$ have disjoin support under the reference measure ρ.

Lemma 16.1 (Neyman Pearson) *For any function $f : \mathcal{X} \to \{0, 1\}$ we have*

$$\boxed{\mathbf{P}(f(X) = 1) + \mathbf{Q}(f(X) = 0) \geq \int \rho(\mathrm{d}x) \min\{p(x), q(x)\}}$$

and the equality is achieved by the Likelihood Ratio Test $f^\star := \mathbf{1}_{q \geq p}$.

Proof: First of all, we prove the equality for the Likelihood Ratio Test:

$$\begin{aligned}
\mathbf{P}(f^\star(X) = 1) + \mathbf{Q}(f^\star(X) = 0) &= \int_{q \geq p} \rho(\mathrm{d}x) p(x) + \int_{q < p} \rho(\mathrm{d}x) q(x) \\
&= \int_{q \geq p} \rho(\mathrm{d}x) \min\{p(x), q(x)\} + \int_{q < p} \rho(\mathrm{d}x) \min\{p(x), q(x)\} \\
&= \int \rho(\mathrm{d}x) \min\{p(x), q(x)\}.
\end{aligned}$$

For any given test f, let $R = \{f = 1\} \equiv \{z \in \mathcal{X} : f(X) = 1\}$ be its so-called rejection region. Let $R^\star = \{f^\star = 1\} = \{q \geq p\}$. We have

$$\begin{aligned}
\mathbf{P}(f(X) = 1) + \mathbf{Q}(f(X) = 0) &= 1 + \mathbf{P}(R) - \mathbf{Q}(R) \\
&= 1 + \int_R \rho(\mathrm{d}x)(p(x) - q(x)) \\
&= 1 + \int_{R \cap R^\star} \rho(\mathrm{d}x)(p(x) - q(x)) + \int_{R \cap (R^\star)^\mathsf{c}} \rho(\mathrm{d}x)(p(x) - q(x)) \\
&= 1 - \int_{R \cap R^\star} \rho(\mathrm{d}x)|p(x) - q(x)| + \int_{R \cap (R^\star)^\mathsf{c}} \rho(\mathrm{d}x)|p(x) - q(x)| \\
&= 1 + \int \rho(\mathrm{d}x)|p(x) - q(x)|(\mathbf{1}_{R \cap (R^\star)^\mathsf{c}}(x) - \mathbf{1}_{R \cap R^\star}(x)).
\end{aligned}$$

The inequality in the statement of the lemma follows as the right-hand side of the previous identity is minimized by the choice $R = R^\star$ (so that the function $\mathbf{1}_{R\cap(R^\star)^\mathsf{c}} - \mathbf{1}_{R\cap R^\star}$ is negative $-\mathbf{1}_{R^\star}$), which corresponds to the choice $f = f^\star$. ■

The Neyman Pearson Lemma is remarkable as it gives a lower bound for the sum of the two types of error that holds for *any* test function f: no matter how we choose the decision rule f, we can not make a decision for which the probability of error on either $\mathbf{P}$ or $\mathbf{Q}$ is smaller than $\int \rho(\mathrm{d}x) \min\{p(x), q(x)\}$. That is, this result gives a structural *limitation* on what one can hope to achieve statistically based on the "amount of information" in the problem, as captured by properties of the probability model at hand (namely, the overlap of the densities p and q with respect to the measure ρ). This result forms the basis of the lower bounds that we are going to develop, which will hold for *any* choice of estimators/algorithms.

The lower bound in the Neyman Pearson Lemma holds uniformly over the choice of tests, and it is expressed by a quantity that is inversely proportional to a notion of distance between the distributions $\mathbf{P}$ and $\mathbf{Q}$. The greater the overlap between the two distributions is, the bigger the quantity $\min\{p(x), q(x)\}$ is; this, in turns, yields a bigger lower-bound reflecting that the hypothesis testing problem is more difficult. The precise notion of distance between probability distributions that is behind the scenes is the *total variation* distance, of which we now give different characterizations.

Definition 16.2 (Total variation distance) *Let $\mathbf{P}$ and $\mathbf{Q}$ be two probability distributions defined on the same measurable space, with respective densities p and q relative to the same measure ρ. The total variation distance between them is defined as*

$$\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}} = \sup_E |\mathbf{P}(E) - \mathbf{Q}(E)| = \frac{1}{2}\int \rho(\mathrm{d}x)|p(x) - q(x)| = 1 - \int \rho(\mathrm{d}x)\min\{p(x), q(x)\},$$

where the supremum is over all measurable sets.

Proof: Omitted. ■

The total variation distance $\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}}$ is a quantity bounded in the interval $[0, 1]$. It is equal to zero if and only if $\mathbf{P} = \mathbf{Q}$, and it is equal to 1 if and only if $\mathbf{P}$ and $\mathbf{Q}$ have disjoint support under ρ.

Proving lower bounds using the Neyman Pearson Lemma reduces to computing upper bounds on the total variation distance $\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}}$. In statistics, a useful way to upper bound the total variation distance is to upper bound the Kullback-Leibler divergence and use Pinsker's inequality to relate the two quantities.

Definition 16.3 (Kullback-Leibler divergence) *Let $\mathbf{P}$ and $\mathbf{Q}$ be two probability distributions defined on the same measurable space, with respective densities p and q relative to the same measure ρ. The Kullback-Leibler divergence between them is given by*

$$\mathtt{KL}(\mathbf{P}, \mathbf{Q}) = \begin{cases} \int \rho(\mathrm{d}x) p(x) \log \frac{p(x)}{q(x)} & \text{if } \mathbf{P} \ll \mathbf{Q} \\ +\infty & \text{otherwise} \end{cases}$$

where the notation $\mathbf{P} \ll \mathbf{Q}$ (to be read as "$\mathbf{P}$ is absolutely continuous with respect to $\mathbf{Q}$") means that whenever $\mathbf{Q}(E) = 0$ for a measurable event E, then also $\mathbf{P}(E) = 0$.

Note that if ρ is the sum of Dirac measures sitting at $\{x_1, \ldots, x_n\}$, this definition reduces to the one given in Lecture 10 (see Section 10.4.2) for discrete random variables taking values in $\{x_1, \ldots, x_n\}$, namely,

$$\mathtt{KL}(\mathbf{P}, \mathbf{Q}) = \begin{cases} \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} & \text{if } \mathbf{P} \ll \mathbf{Q} \\ +\infty & \text{otherwise} \end{cases}$$

The usefulness of the Kullback-Leibler divergence in statistics stems from the fact that this quantity factorizes with respect to product measures representing independent random variables. In statistics, in fact, we are often dealing with data in the form of i.i.d. samples from the same distribution. That is, we care about the case when $X = \{X_1, \dots, X_n\}$ for a given collection of n i.i.d. variables, so that the measures $\mathbf{P}$ and $\mathbf{Q}$ in the hypothesis test considered above are product measures. The Kullback-Leibler divergence is always non-negative, but, contrarily to the total variation distance, it is not a distance (for instance, it is not symmetric).

Proposition 16.4 (Properties of Kullback-Leibler divergence) *Let $\mathbf{P}$ and $\mathbf{Q}$ be two probability distributions defined on the same measurable space. Then,*

1. ***Gibbs' inequality.*** $\boxed{\mathtt{KL}(\mathbf{P}, \mathbf{Q}) \geq 0}$ *with equality if and only if $\mathbf{P} = \mathbf{Q}$.*
2. ***Chain rule for product distributions.*** *If $\mathbf{P}$ and $\mathbf{Q}$ are product measures, i.e., $\mathbf{P} = \bigotimes_{i=1}^n \mathbf{P}_i$ and $\mathbf{Q} = \bigotimes_{i=1}^n \mathbf{Q}_i$, then*
$$\boxed{\mathtt{KL}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^n \mathtt{KL}(\mathbf{P}_i, \mathbf{Q}_i)}$$
3. ***Pinsker's inequality.*** *For any measurable event E, we have*
$$\boxed{\mathbf{P}(E) - \mathbf{Q}(E) \leq \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{P}, \mathbf{Q})}} \tag{16.1}$$
which yields, in particular,
$$\boxed{\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}} \leq \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{P}, \mathbf{Q})}}$$

Proof: See **Problem 4.3** in the Problem Sheets. ■

Remark 16.5 *Note that Pinker's inequality also holds if we swap $\mathbf{P}$ and $\mathbf{Q}$ in the KL divergence (just consider* (16.1) *with E^{C}) so that*
$$\mathbf{P}(E) - \mathbf{Q}(E) \leq \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{Q}, \mathbf{P})}$$
and
$$\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}} \leq \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{Q}, \mathbf{P})}.$$
As the Kullback-Leibler divergence is not symmetric, these inequalities are truly different from the ones listed in Proposition 16.4. In what follows we only state inequalities with respect to $\mathtt{KL}(\mathbf{P}, \mathbf{Q})$, but the symmetry of the problem makes it clear that these inequalities also hold if we swap $\mathbf{P}$ and $\mathbf{Q}$.

Note that we have encountered the Pinsker's inequality before, in Lecture 10. Together, the above results yields the following corollary of the Neyman Pearson Lemma.

Corollary 16.6 *Let $X = \{X_1, \dots, X_n\} \in \mathcal{X}^n$ be distributed according to either $\mathbf{P}$ or $\mathbf{Q}$ on $\mathcal{X}^n$. For any test function $f : \mathcal{X}^n \to \{0, 1\}$ we have*
$$\boxed{\mathbf{P}(f(X_1, \dots, X_n) = 1) + \mathbf{Q}(f(X_1, \dots, X_n) = 0) \geq 1 - \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{P}, \mathbf{Q})}}$$

If, furthermore, each X_i is independently distributed either according to $\mathbf{P}_i$ or according to $\mathbf{Q}_i$, then $\mathbf{P} = \bigotimes_{i=1}^n \mathbf{P}_i$ and $\mathbf{Q} = \bigotimes_{i=1}^n \mathbf{Q}_i$ and

$$\boxed{\mathbf{P}(f(X_1, \ldots, X_n) = 1) + \mathbf{Q}(f(X_1, \ldots, X_n) = 0) \geq 1 - \sqrt{\frac{1}{2} \sum_{i=1}^n \mathtt{KL}(\mathbf{P}_i, \mathbf{Q}_i)}} \tag{16.2}$$

Hence, in the setting of i.i.d. data ($\mathbf{P}_i = \widetilde{\mathbf{P}}$ and $\mathbf{Q}_i = \widetilde{\mathbf{Q}}$ for all $i \in [n]$), the Neyman Pearson Lemma along with the Kullback-Leibler divergence gives a convenient lower bound for the statistical limitation of *any* decision rule. This lower bound explicitly express the "amount of information" in the probabilistic model as a function of the number of data points n and the Kullback-Leibler divergence of the individual distributions: $\mathtt{KL}(\widetilde{\mathbf{P}}, \widetilde{\mathbf{Q}})$.

16.3 Back to Bandits: Distribution-Independent Lower Bounds

We now apply the results so far developed to establish a lower bound for the multi-armed bandit problem introduced in the last lecture. In this setting, the data is given by a collection of n i.i.d. random vectors $Z_1, \ldots, Z_n$ on $[0,1]^k$, where each component $a \in [k] \equiv \mathcal{A}$ (corresponding to an arm to be played) is sampled i.i.d. from a probability distribution on $[0,1]$ with mean μ_a (unknown to the player). In the bandit setting, at each time step $t \in [n]$, only the component associated to A_t (the action played at time t) is revealed to the player. Namely, only Z_{t,A_t} is revealed. An algorithm (also known as a policy) is a set of actions $A_1, \ldots, A_n \in \mathcal{A}$ such that each action A_t can depend only on the information available prior to time t, namely, on $(A_s, Z_{s,A_s})_{s \in [t-1]}$. We now show that for any algorithm, there exists a bandit problem (i.e., a choice of arm distributions) such that the expected pseudo-regret incurred by the algorithm grows at least as $\sqrt{kn}$.

Theorem 16.7 (Distribution-independent lower bound for multi-armed bandit) *Let $n \geq k - 1$. For any algorithm $A_1, \ldots, A_n$, there exists a k-armed bandit problem such that*

$$\boxed{\mathbf{E}R_n \geq c\sqrt{(k-1)n}}$$

where c is a universal constant.

As the arms' rewards are i.i.d. random vectors, it would seem natural that the proof of Theorem 16.7 relies on the bound (16.2) in Corollary 16.6. However, recall that we are in the bandit setting (i.e., limited information setting), and at each time step we only see the component of the reward vector associated with the arm we choose to pull. That is, the data we see is of the form $(A_1, Z_{1,A_1}), \ldots, (A_n, Z_{n,A_n})$, which is not i.i.d.. In fact, the algorithm $A_1, \ldots, A_n$ can depend (in a possibly very complicated way) on all the information available in the past. However, things are not as bad as they might seem, as the dependency introduced by the algorithm is model-independent, in the sense that the arm-choosing policy can only depend on the *observed* data and not on the bandit model itself (as the bandit model is defined by the unknown reward distributions, which is not observable). For this reason, the Kullback-Leibler divergence between the probability distributions that encode two different bandit models has a simple expression: it is a weighted sum of the Kullback-Leibler divergence of the arms' rewards in the two models, weighted by the mean expected number of times each arm is played under one model (recall that the Kullback-Leibler divergence is not symmetric).

Proposition 16.8 *Consider two k-armed bandit models. The first one (labeled by μ) is defined by a reward probability $\mathbf{P}_{\mu,a}$ for each arm $a \in [k]$. The second one (labeled by ν) is defined by a reward probability $\mathbf{P}_{\nu,a}$*

for each arm $a \in [k]$. Assume that for any $a \in [k]$, $\mathbf{P}_{\mu,a}$ and $\mathbf{P}_{\nu,a}$ have densities $p_{\mu,a}$ and $p_{\nu,a}$ relative to the same measure ρ. Fix an algorithm $A_1, \dots, A_n$, and let $\mathbf{P}_\mu$ and $\mathbf{P}_\nu$ be, respectively, the probability that each bandit model assigns to the random variables $(A_1, Z_{1,A_1}), \dots, (A_n, Z_{n,A_n})$. We have

$$\boxed{\mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu) = \sum_{a=1}^{k} \mathtt{KL}(\mathbf{P}_{\mu,a}, \mathbf{P}_{\nu,a}) \mathbf{E}_\mu N_{n,a}}$$

Remark 16.9 *Related to the discussion above, note that $\mathbf{P}_\mu \neq \bigotimes_{a=1}^k \mathbf{P}_{\mu,a}$ (analogously, $\mathbf{P}_\nu \neq \bigotimes_{a=1}^k \mathbf{P}_{\nu,a}$) as each $\mathbf{P}_{\mu,a}$ refers only to the corresponding arm reward, whereas $\mathbf{P}_\mu$ is the probability of the* entire *model, which also includes the randomness of the algorithm/policy.*

Proof: Let $\lambda(\mathrm{d}x) = \sum_{a=1}^k \delta_a(\mathrm{d}x)$, where δ_a is the Dirac measure sitting at a. The probability density function associated to the bandit problem with mean reward vector μ, with respect to the reference measure $(\lambda \times \rho)^n$, can be decomposed as follows

$$\begin{aligned}
&p_\mu(a_1, x_1, \dots, a_n, x_n) \\
&\quad := p_\mu(A_1 = a_1, Z_{1,A_1} = x_1, \dots, A_n = a_n, Z_{n,A_n} = x_n) \\
&\quad = p(A_1 = a_1) p_\mu(Z_{1,A_1} = x_1 | A_1 = a_1) \times \\
&\qquad \times \prod_{s=2}^{n} p(A_s = a_s | A_1 = a_1, Z_{1,A_1} = x_1, \dots, A_{s-1} = a_{s-1}, Z_{s-1,A_{s-1}} = x_{s-1}) p_\mu(Z_{s,A_s} = x_s | A_s = a_s) \\
&\quad = p(A_1 = a_1) p_{\mu,a_1}(x_1) \prod_{s=2}^{n} p(A_s = a_s | A_1 = a_1, Z_{1,A_1} = x_1, \dots, A_{s-1} = a_{s-1}, Z_{s-1,A_{s-1}} = x_{s-1}) p_{\mu,a_s}(x_s),
\end{aligned}$$

and an analogous expression holds for p_ν. Note that the transition probabilities of the arm's choice at time s given all the information available up to time $s-1$ do *not* depend on the bandit model; these probabilities are only a function of the chosen algorithm/policy and this is the reason why we use the general notation p (instead of p_μ or p_ν). As a consequence, we have

$$\frac{p_\mu(a_1, x_1, \dots, a_n, x_n)}{p_\nu(a_1, x_1, \dots, a_n, x_n)} = \prod_{s=1}^{n} \frac{p_{\mu,a_s}(x_s)}{p_{\nu,a_s}(x_s)}.$$

Therefore, if $A_1, Z_{1,A_1}, \dots, A_n, Z_{n,A_n}$ are random variables distributed according to $\mathbf{P}_\mu$, we have

$$\mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu) = \mathbf{E}_\mu \log \frac{p_\mu(A_1, Z_{1,A_1}, \dots, A_n, Z_{n,A_n})}{p_\nu(A_1, Z_{1,A_1}, \dots, A_n, Z_{n,A_n})} = \sum_{s=1}^{n} \mathbf{E}_\mu \log \frac{p_{\mu,A_s}(Z_{s,A_s})}{p_{\nu,A_s}(Z_{s,A_s})}.$$

Note that, by the tower property of conditional expectations,

$$\mathbf{E}_\mu \log \frac{p_{\mu,A_s}(Z_{s,A_s})}{p_{\nu,A_s}(Z_{s,A_s})} = \mathbf{E}_\mu \mathbf{E}_\mu \left[\log \frac{p_{\mu,A_s}(Z_{s,A_s})}{p_{\nu,A_s}(Z_{s,A_s})} \middle| A_s \right] = \mathbf{E}_\mu \mathtt{KL}(\mathbf{P}_{\mu,A_s}, \mathbf{P}_{\nu,A_s})$$

so we find, using that $N_{n,a} = \sum_{s=1}^{n} \mathbf{1}_{A_s=a}$,

$$\begin{aligned}
\mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu) &= \sum_{s=1}^{n} \mathbf{E}_\mu \mathtt{KL}(\mathbf{P}_{\mu,A_s}, \mathbf{P}_{\nu,A_s}) \\
&= \sum_{s=1}^{n} \mathbf{E}_\mu \left[\mathtt{KL}(\mathbf{P}_{\mu,A_s}, \mathbf{P}_{\nu,A_s}) \sum_{a=1}^{k} \mathbf{1}_{A_s=a} \right] \\
&= \sum_{a=1}^{k} \mathbf{E}_\mu \left[\mathtt{KL}(\mathbf{P}_{\mu,a}, \mathbf{P}_{\nu,a}) \sum_{s=1}^{n} \mathbf{1}_{A_s=a} \right] \\
&= \sum_{a=1}^{k} \mathtt{KL}(\mathbf{P}_{\mu,a}, \mathbf{P}_{\nu,a}) \, \mathbf{E}_\mu N_{n,a}.
\end{aligned}$$

■

Proof:[Proof of Theorem 16.7] As Theorem 16.7 refers to a *negative* result, it is enough to consider a specific class of multi-armed bandit problems. We choose the class of Bernoulli arms, in which the rewards of each arm are modelled as i.i.d. from the Bernoulli distribution with a certain mean. In this setting, the vector of mean rewards in $\{0,1\}^k$ defines the probability distribution of the rewards, hence it defines the bandit problem. Fix any algorithm/policy $A_1, \ldots, A_n$ (recall that apart from the randomness in the data, this algorithm can depend on external sources of randomness such as coin flips to favour exploration as in the ε-Greedy algorithm). Given the chosen policy, we will construct two specific bandit problems with mean reward vectors given by μ and ν, respectively, and corresponding pseudo-regrets defined as

$$\begin{aligned}
(R_\mu)_n &= n\mu^\star - \sum_{t=1}^{n} \mu_{A_t}, \\
(R_\nu)_n &= n\nu^\star - \sum_{t=1}^{n} \nu_{A_t},
\end{aligned}$$

where $\mu^\star := \operatorname{argmax}_{i\in[k]} \mu_i$ and $\nu^\star := \operatorname{argmax}_{i\in[k]} \nu_i$. Let us denote by $\mathbf{P}_\mu$ (respectively, $\mathbf{E}_\mu$) and $\mathbf{P}_\nu$ (respectively, $\mathbf{E}_\nu$) the probabilities (respectively, expectations) with respect to all the randomness in the model (i.e., the random variables $A_1, Z_{1,A_1}, \ldots, A_n, Z_{n,A_n}$) when the rewards follow a Bernoulli vector with mean μ and ν, respectively. We will prove that in at least one of these two problems the policy attains an expected pseudo-regret that is lower-bounded as in the statement of the theorem. Specifically, we will show that

$$\max\{\mathbf{E}_\mu(R_\mu)_n, \mathbf{E}_\nu(R_\nu)_n\} \geq \frac{1}{2}(\mathbf{E}_\mu(R_\mu)_n + \mathbf{E}_\nu(R_\nu)_n) \geq c\sqrt{(k-1)n},$$

where the first inequality follows from $x + y \leq 2\max\{x, y\}$ and the second inequality follows from Corollary 16.6, as we will see. Let us define the first bandit problem by the following vector of mean rewards:

$$\mu = \left(\frac{1}{2} + \Delta, \frac{1}{2}, \ldots, \frac{1}{2}\right),$$

for a fix $\Delta \in (0, 1/4)$. That is, arm 1 is optimal and follows a Bernoulli distribution with mean $1/2 + \Delta$, and all the other sub-optimal arms follow a Bernoulli distribution with mean $1/2$. By definition of this model, Δ is the sub-optimality gap of all the sub-optimal arms. To define the second bandit problem, we first find the sub-optimal arm that is played the least (in expectation) by our algorithm in the first bandit problem, namely,

$$b = \operatorname*{argmin}_{a\in\{2,\ldots,k\}} \mathbf{E}_\mu N_{n,a}.$$

We define the second bandit problem by the following vector of mean rewards ν, which shares the same components as μ apart from the component associated to arm b:

$$\nu = \left(\frac{1}{2} + \Delta, \frac{1}{2}, \dots, \frac{1}{2}, \frac{1}{2} + 2\Delta, \frac{1}{2}, \dots, \frac{1}{2}\right).$$

In this model, arm b is optimal with mean reward $\frac{1}{2} + 2\Delta$. By the law of total expectations we have

$$\begin{aligned}
\mathbf{E}_\mu(R_\mu)_n &= \mathbf{E}_\mu\left[(R_\mu)_n \middle| N_{n,1} \le \frac{n}{2}\right]\mathbf{P}_\mu\left(N_{n,1} \le \frac{n}{2}\right) + \mathbf{E}_\mu\left[(R_\mu)_n \middle| N_{n,1} > \frac{n}{2}\right]\mathbf{P}_\mu\left(N_{n,1} > \frac{n}{2}\right) \\
&\ge \mathbf{E}_\mu\left[(R_\mu)_n \middle| N_{n,1} \le \frac{n}{2}\right]\mathbf{P}_\mu\left(N_{n,1} \le \frac{n}{2}\right) \\
&\ge \frac{\Delta n}{2}\mathbf{P}_\mu\left(N_{n,1} \le \frac{n}{2}\right),
\end{aligned}$$

where the last inequality follows by the fact that the event $N_{n,1} \le n/2$ is equivalent to the event that an arm different than 1 (sub-optimal for the bandit model μ) is played more than $n/2$ times, and each times this happens we are adding a Δ term to the pseudo-regret for model μ. Analogously, we find

$$\begin{aligned}
\mathbf{E}_\nu(R_\nu)_n &= \mathbf{E}_\nu\left[(R_\nu)_n \middle| N_{n,1} \le \frac{n}{2}\right]\mathbf{P}_\nu\left(N_{n,1} \le \frac{n}{2}\right) + \mathbf{E}_\nu\left[(R_\nu)_n \middle| N_{n,1} > \frac{n}{2}\right]\mathbf{P}_\nu\left(N_{n,1} > \frac{n}{2}\right) \\
&\ge \mathbf{E}_\nu\left[(R_\nu)_n \middle| N_{n,1} > \frac{n}{2}\right]\mathbf{P}_\nu\left(N_{n,1} > \frac{n}{2}\right) \\
&> \frac{\Delta n}{2}\mathbf{P}_\nu\left(N_{n,1} > \frac{n}{2}\right),
\end{aligned}$$

as the event $N_{n,1} > n/2$ tells that arm 1 (sub-optimal for the bandit model ν) is played more than $n/2$ times, and each times this happens we are adding a Δ term to the pseudo-regret for model ν. Hence, by the Neyman Pearson Lemma and Pinsker's inequality (i.e., Corollary 16.6) we find

$$\mathbf{E}_\mu(R_\mu)_n + \mathbf{E}_\nu(R_\nu)_n > \frac{\Delta n}{2}\left(\mathbf{P}_\mu\left(N_{n,1} \le \frac{n}{2}\right) + \mathbf{P}_\nu\left(N_{n,1} > \frac{n}{2}\right)\right) \ge \frac{\Delta n}{2}\left(1 - \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu)}\right).$$

Proposition 16.8 yields

$$\mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu) = \sum_{a=1}^{k} \mathtt{KL}(\mathrm{Bern}(\mu_a), \mathrm{Bern}(\nu_a))\, \mathbf{E}_\mu N_{n,a} = \mathtt{KL}(\mathrm{Bern}(1/2), \mathrm{Bern}(1/2 + 2\Delta))\, \mathbf{E}_\mu N_{n,b}.$$

Note that as $\sum_{a \in [k]} \mathbf{E}_\mu N_{n,a} = n$ and by definition of b we have

$$\mathbf{E}_\mu N_{n,b} \le \frac{n}{k-1}.$$

At the same time, using that $-\log(1-x) \le 2x$ for any $0 \le x \le 1/2$, we have

$$\begin{aligned}
\mathtt{KL}(\mathrm{Bern}(1/2), \mathrm{Bern}(1/2 + 2\Delta)) &= \frac{1}{2}\log\frac{1/2}{1/2 - 2\Delta} + \frac{1}{2}\log\frac{1/2}{1/2 + 2\Delta} = \frac{1}{2}\log\frac{1/4}{1/4 - 4\Delta^2} \\
&= -\frac{1}{2}\log(1 - 16\Delta^2) \le 16\Delta^2.
\end{aligned}$$

Hence, $\mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu) \le \frac{16\Delta^2 n}{k-1}$ and

$$\mathbf{E}_\mu(R_\mu)_n + \mathbf{E}_\nu(R_\nu)_n \ge \frac{\Delta n}{2}\left(1 - \sqrt{\frac{8\Delta^2 n}{k-1}}\right).$$

The proof follows by taking the maximum of the right-hand side of this inequality with respect to Δ, which yields $\Delta^\star = \frac{1}{4}\sqrt{\frac{k-1}{2n}}$ and

$$\frac{\Delta^\star n}{2}\left(1 - \sqrt{\frac{8(\Delta^\star)^2 n}{k-1}}\right) = c\sqrt{(k-1)n},$$

with $c = \frac{1}{16\sqrt{2}}$. ■

16.4 Minimax Lower Bounds. Multiple Hypothesis Testing

The lower bound in Theorem 16.7 is of the type

$$\min_{A_1,\dots,A_n} \sup_{w\in\mathcal{W}} \mathbf{E}_w(R_w)_n \geq c\sqrt{(k-1)n},$$

where the minimum is over all possible algorithms $A_1, \dots, A_n$ and the supremum is over a class of bandit models indexed by a parameter $w \in \mathcal{W}$. Here, $\mathbf{E}_w$ refers to the expectation with respect to the bandit model indexed by w and $(R_w)_n$ refers to the corresponding pseudo-regret (recall that the pseudo-regret is defined as a function of $\mathbf{E}_w$). This type of lower bounds is called "minimax". Theorem 16.7 was proven by setting up a binary hypothesis testing problem and using the Neyman Pearson Lemma. This idea can be generalized to multiple hypothesis testing using a tool from information theory: Fano's inequality.

Theorem 16.10 (Fano's inequality) *Let $\mathbf{P}_1, \dots, \mathbf{P}_m$ be probability measures such that $\mathbf{P}_\mu \ll \mathbf{P}_\nu$ for any $\mu, \nu \in [m]$. Then,*

$$\boxed{\inf_f \max_{\mu\in[m]} \mathbf{P}_\mu(f(X) \neq \mu) \geq 1 - \frac{\frac{1}{m^2}\sum_{\mu,\nu=1}^m \mathtt{KL}(\mathbf{P}_\mu, \mathbf{P}_\nu) + \log 2}{\log(m-1)}}$$

A direct comparison with Corollary 16.6, using that

$$\max_{\mu\in[2]} \mathbf{P}_\mu(f(X) \neq \mu) \geq \frac{1}{2}\mathbf{P}_1(f(X) \neq 1) + \frac{1}{2}\mathbf{P}_2(f(X) \neq 2)$$

shows that for $m = 2$ Fano's inequality is not as good as the Neyman Pearson Lemma (indeed, Fano's inequality yields a trivial bound for $m = 2$). However, Fano's inequality is more general and holds for any $m > 2$. The following is a general strategy to prove minimax lower bounds:

1. Reduce the supremum over $\mathcal{W}$ to a sum over finitely-many terms.
2. Use Fano's Lemma.

For the first step, the notion of covering numbers is typically used.

References

[1] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.