Algorithmic Foundations of Learning

Lecture 16
Minimax Lower Bounds and Hypothesis Testing

Patrick Rebeschini

Department of Statistics
University of Oxford

Introduction

Traditionally, **STATISTICS** is taught via **asymptotic** results, for $n \to \infty$:

- **Law of Large Numbers**
- Central Limit Theorem, yielding
  - **Confidence bounds**
  - **Hypothesis testing**

In this course we have developed **non-asymptotic** results, for $n < \infty$:

- **Uniform Law of Large Numbers**
  $\Rightarrow$ **Notions of complexity to bound generaliz. error of ERM algorithm**
- **Confidence bounds**
  $\Rightarrow$ **Analysis of algorithms (upper bounds with high-probability)**
  $\Rightarrow$ **Design of algorithms (UCB)**
- **Hypothesis testing** (Today's lecture)
  $\Rightarrow$ **Lower bounds holding for any algorithm**

STATISTICS lays the foundation of **ALGORITHMS** for machine learning

Hypothesis Testing and Lower Bounds

- **Data:** random variable $X \in \mathcal{X}$
- **Hypotheses:**
  - $X \sim \mathbf{P}$ (**null hypothesis** H_0)
  - $X \sim \mathbf{Q}$ (**alternative hypothesis** H_1)
- **Test:** any function $f : \mathcal{X} \to \{0, 1\}$
- **Errors:**
  - Type I: if $f(X) = 1$ when $X \sim \mathbf{P}$
  - Type II: if $f(X) = 0$ when $X \sim \mathbf{Q}$

Any test commits one type of error with strictly positive probability unless $\mathbf{P}$ and $\mathbf{Q}$ have disjoin support under the reference measure ρ

Neyman Pearson (Lemma 16.1)

For any function $f : \mathcal{X} \to \{0, 1\}$ we have

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

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

Proof of Lemma 16.1

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

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

- For a test f, let $R=\{f=1\}\equiv\{x\in\mathcal{X}:f(x)=1\}$, $R^\star=\{f^\star=1\}=\{q\geq p\}$

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

- 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 $1_{R\cap(R^\star)^\mathsf{C}}-1_{R\cap R^\star}$ is negative $-1_{R^\star}$), which corresponds to the choice $f=f^\star$

Total Variation Distance

Neyman Pearson Lemma:

- No matter how we choose the decision rule f, we can not make a decision with probability of error on either $\mathbf{P}$ or $\mathbf{Q}$ smaller than $\int \rho(dx) \min\{p(x), q(x)\}$
- **Structural limitation** of what we can hope to achieve statistically based on the **"amount of information"** in the problem
- The greater the overlap between $\mathbf{P}$ and $\mathbf{Q}$, the more difficult the problem is
- There is a notion of distance behind the scenes...

Total variation distance (Definition 16.2)

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

To prove lower bounds on sum of errors, enough to upper bound $\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}}$

Kullback-Leibler Divergence

- In statistics, often data is $X_1, \ldots, X_n$ i.i.d. ($\mathbf{P} = \bigotimes_{i=1}^n \mathbf{P}_i$ and $\mathbf{Q} = \bigotimes_{i=1}^n \mathbf{Q}_i$)
- The total variation distance does not factorize under product measures
- The Kullback-Leibler divergence (**not a distance!**) does factorize instead

Kullback-Leibler divergence (Definition 16.3)

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

Properties of Kullback-Leibler divergence (Proposition 16.4)

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 measures:** $\boxed{\mathtt{KL}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^n \mathtt{KL}(\mathbf{P}_i, \mathbf{Q}_i)}$
3. **Pinsker's inequality:** $\boxed{\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{tv}} \leq \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{P}, \mathbf{Q})}}$

Lower Bound with Independent Samples

Corollary 16.6

- **Data:** Let $X_1, \ldots, X_n \in \mathcal{X}$
- **Hypotheses:** $\mathbf{P}$ (null H_0) or $\mathbf{Q}$ (alternative H_1)
- **Test:** $f : \mathcal{X}^n \to \{0, 1\}$

$$\mathbf{P}(f(X_1, \ldots, X_n) = 1) + \mathbf{Q}(f(X_1, \ldots, X_n) = 0) \geq 1 - \sqrt{\frac{1}{2}\mathtt{KL}(\mathbf{P}, \mathbf{Q})}$$

If $X_1, \ldots, X_n$ are independent, then

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

- **"Amount of information":** Function of n and $\mathtt{KL}(\mathbf{P}_i, \mathbf{Q}_i)$, $i \in [n]$

Back to the Multi-Armed Bandit Problem

At every time step $t = 1, 2, \dots, n$:

1. Choose an action $A_t \in \mathcal{A}$
2. A data point Z_t is sampled independently from an unknown distribution
   - **Bandit:** Z_t is not revealed
3. Suffer a loss $\ell(A_t, Z_t) = -Z_{t,A_t}$

Vectors Z_t's are indep., but observed data $(A_1, Z_{1,A_1}), \dots, (A_n, Z_{n,A_n})$ are not!

Proposition 16.8

- Two bandit models (μ and ν): rewards for arm a either $\mathbf{P}_{\mu,a}$ or $\mathbf{P}_{\nu,a}$
- Fix an algorithm $A_1, \dots, A_n$
- $\mathbf{P}_\mu$ and $\mathbf{P}_\nu$ probab. each model assigns to $(A_1, Z_{1,A_1}), \dots, (A_n, Z_{n,A_n})$

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

Distribution-Independent Lower Bound

Theorem 16.7

Let $n \geq k-1$. For any algorithm there exists a k-armed bandit problem with

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

where c is a universal constant

- UCB achieves **quasi-optimal** **distribution-independent** pseudo-regret.
- Using similar ideas (but more involved), one can prove that UCB achieves **optimal** **distribution-dependent** pseudo-regret.
- Ideas can be generalized to **multiple hypothesis testing**...

Fano's Inequality (Theorem 16.10)

Let $\mathbf{P}_1, \ldots, \mathbf{P}_m$ be probability measures such that $\mathbf{P}_\mu \ll \mathbf{P}_\nu$ for any $\mu, \nu \in [m]$

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

Proof of Theorem 16.7 (Part I)

- Fix any algorithm/policy $A_1, \dots, A_n$.
- We will construct two bandit problems with Bernoulli mean reward vectors given by μ and ν, respectively, and corresponding pseudo-regrets defined as

$$(R_\mu)_n = n\mu^\star - \sum_{t=1}^{n} \mu_{A_t} \qquad\qquad (R_\nu)_n = n\nu^\star - \sum_{t=1}^{n} \nu_{A_t}$$

  where $\mu^\star := \operatorname{argmax}_{i \in [k]} \mu_i$ and $\nu^\star := \operatorname{argmax}_{i \in [k]} \nu_i$.
- 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 theorem:

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

  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.

Proof of Theorem 16.7 (Part II)

- **First bandit problem** (for a fix $\Delta \in (0, 1/4)$):

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

- To define the second bandit problem, find the sub-optimal arm that is played the least (in expectation) by our algorithm in the first problem:

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

- **Second bandit problem**:

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

In this model, arm b is optimal with mean reward $\frac{1}{2} + 2\Delta$

Proof of Theorem 16.7 (Part III)

- By the law of total expectations we have

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

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

- Analogously, we find

$$
\mathbf{E}_\nu(R_\nu)_n > \frac{\Delta n}{2}\mathbf{P}_\nu\left(N_{n,1} > \frac{n}{2}\right)
$$

Proof of Theorem 16.7 (Part IV)

- By the Neyman Pearson Lemma and Pinsker's inequality, we find

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

- Proposition 16.8 yields

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

- As $\sum_{a\in[k]} \mathbf{E}_\mu N_{n,a} = n$ and by definition of b we have $\mathbf{E}_\mu N_{n,b} \le \frac{n}{k-1}$
- Using that $-\log(1-x) \le 2x$ for any $0 \le x \le 1/2$, we have

$$\mathtt{KL}(\mathsf{Bern}(1/2), \mathsf{Bern}(1/2 + 2\Delta)) = \frac{1}{2}\log\frac{1/2}{1/2 - 2\Delta} + \frac{1}{2}\log\frac{1/2}{1/2 + 2\Delta}$$

$$= \frac{1}{2}\log\frac{1/4}{1/4 - 4\Delta^2} = -\frac{1}{2}\log(1 - 16\Delta^2) \le 16\Delta^2$$

Proof of Theorem 16.7 (Part V)

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

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

- 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 $c = \frac{1}{16\sqrt{2}}$

"New science is based on maximum likelihood rather than certainty"

Arthur C. Clarke and Gentry Lee, Rama Series Book 2, 1989