CS6781 - Theoretical Foundations of Machine Learning

Lecture 15: Sequential Experimentation: Theory and Principles

March 12, 2020

Lecturer: Robert Kleinberg *Readings: n/a*
Scribe: Sloan Nietert

This lecture will examine several categories of problems related to sequential experimentation.

- **Sequential Experimental Design:** Given two or more hypotheses, and one or more experiments whose outcome distributions differ under the hypotheses, design a procedure to test which hypothesis is true. For example, Hodgkin or non-Hodgkin lymphoma?
- **Multi-Armed Bandit:** Given two or more actions, each of which produces stochastic payoffs sampled from an unknown stationary distribution, design a procedure to maximize average payoff over time. For example, choose a color for the "donate" button on our site.
- **Best Arm Identification:** Given two or more actions as in the multi-armed bandit problem, design a procedure to find the one with the highest average payoff. For example, which of these drugs is most effective at treating high blood pressure?

There are several standard modes of analysis for these problems.

- **Bayesian:** Optimize average-case performance under some prior distribution on the true state of the world.
- **Stochastic:** Optimize worst-case performance under the assumption that payoffs/experiment results are sampled independently from a fixed distribution.
- **Adversarial:** Optimize worst-case performance under the assumption that payoffs/experiment results are chosen by an oblivious or adaptive adversary.

The following techniques will be key to our analysis.

- The familar Chernoff bound will be used to justify that certain procedures have high probability of success.
- Kullback-Leibler (KL) divergence is an information theoretic measure of distance between probability distributions which will be used to prove that certain procedures are (nearly) optimal.

These scribe notes are based on slides for a guest lecture given to Yale's Econ 421 class in October 2017.

Example #1: Biased Coin Testing

As a first example, we consider a scenario in which we toss a single coin and wish to determine whether:

(A) the coin is fair, or

(B) $\Pr(\text{heads}) = \frac{1+\varepsilon}{2}$.

We seek to design a procedure to toss the coin until eventually stopping and guessing A or B, such that $\Pr(\text{error}) < \delta$ in both cases. A natural goal is to minimize the expected number of coin tosses. In particular, we will consider the *fixed design procedure* which tosses the coin s times and guesses A unless the empirical frequency of heads exceeds $1/2 + \varepsilon/4$ (i.e. maximum likelihood estimation). An error (under either hypothesis) requires empirical frequency to differ from its expected value by more than $\varepsilon/4$, so the Chernoff-Hoeffding bound gives that

$$\Pr(\text{error}) < \exp\left(-\frac{\varepsilon^2 s}{8}\right).$$

To make this less than δ, set $s > 8\log(1/\delta)/\varepsilon^2$. This example exhibits two quantitative hallmarks of optimal experimentation.

- $1/\varepsilon^2$ independent samples suffice to distinguish distributions that differ by ε,
- Inflating sample complexity by $\log(1/\delta)$ boosts confidence to $1 - \delta$.

Next, we'll develop some machinery to prove that no procedure can admit the same error rate with $o\left(\log(1/\delta)/\varepsilon^2\right)$ expected samples.

1.1 Kullback-Leibler Divergence

Definition 1.1 (KL Divergence)**.** If p, q are two distributions on a finite set Ω, we define the *Kullback-Leibler divergence* between p and q as

$$\mathrm{D}(p \,\|\, q) := \sum_{x \in \Omega} p(x) \log\left(\frac{p(x)}{q(x)}\right).$$

More generally, if p and q are probability measures on a measure space Ω,

$$\mathrm{D}(p \,\|\, q) = \int_{\Omega} \log\left(\frac{\mathrm{d}p}{\mathrm{d}q}\right) \mathrm{d}p,$$

where $\frac{\mathrm{d}p}{\mathrm{d}q}$ denotes the Radon-Nikodym derivative.

KL divergence exhibits several important properties.

- For all p, q, $\mathrm{D}(p \,\|\, q) \geq 0$ with equality if and only if $p = q$.
- If $\mathrm{Ber}(r)$ denotes the Bernoulli distribution with parameter r, then

$$\mathrm{D}\left(\mathrm{Ber}\left(\frac{1}{2}\right) \,\middle\|\, \mathrm{Ber}\left(\frac{1+\varepsilon}{2}\right)\right) < \varepsilon^2$$

  for $\varepsilon < 8/9$.

Less trivially, we also have the following.

Proposition 1.2 (High-Probability Pinsker Inequality)**.** *If $\Omega = A \cup B$, then*

$$p(B) + q(A) \geq \frac{1}{2}\exp(-\mathrm{D}(p \,\|\, q)).$$

Proposition 1.3 (Chain Rule for KL Divergence)**.** *If p, q are probability distributions over sequences $\boldsymbol{x} = (x_1, \ldots, x_n) \in \Omega_1 \times \cdots \times \Omega_n$, then*

$$\mathrm{D}(p \,\|\, q) = \sum_{k=1}^{n} \mathbb{E}_{\boldsymbol{x}\sim p}[\mathrm{D}(p(x_k|x_1, \ldots, x_{k-1}) \,\|\, q(x_k|x_1, \ldots, x_{k-1}))],$$

where $p(x_k|x_1, \ldots, x_{k-1})$ is shorthand for the distribution r over Ω_k with

$$r(x) = \Pr_{\boldsymbol{y}\sim p}[y_k = x \mid y_1 = x_1, \ldots, y_{k-1} = x_{k-1}].$$

Now, let $\mathfrak{I}$ be a set of experiments, and let p_i, q_i denote the distribution of outcomes for experiment $i \in \mathfrak{I}$ under hypotheses p, q respectively. Let π be a sequential experimentation protocol, and let p^π, q^π denote the distributions of outcome sequences produced under p, q, respectively. Then, we have the following.

Lemma 1.4 (Divergence Decomposition Lemma)**.** *If $S_i(\pi)$ is the random variable denoting the number of times experiment i is performed under protocol π, then*

$$\mathrm{D}(p^\pi \,\|\, q^\pi) = \sum_{i\in\mathfrak{I}} \mathbb{E}_p[S_i(\pi)] \cdot \mathrm{D}(p_i \,\|\, q_i).$$

Proof. Let $i_k = i_k(\pi)$ be the random variable denoting the experiment selected by π at step k. Then, by the chain rule, we have

$$\begin{aligned}
\mathrm{D}(p^\pi \,\|\, q^\pi) &= \sum_k \mathbb{E}_{\boldsymbol{x}\sim p^\pi}[\mathrm{D}(p^\pi(x_k|x_1, \ldots, x_{k-1}) \,\|\, q^\pi(x_k|x_1, \ldots, x_{k-1})] \\
&= \sum_k \mathbb{E}_{\boldsymbol{x}\sim p^\pi}\left[\sum_{i\in\mathfrak{I}} 1(i_k = i)\,\mathrm{D}(p_i \,\|\, q_i)\right] \\
&= \sum_{i\in\mathfrak{I}} \mathbb{E}_p[S_i(\pi)] \cdot \mathrm{D}(p_i \,\|\, q_i).
\end{aligned}$$

□

Now, if we know that an experimentation protocol has a low error probability, then we can use Pinkser's inequality to bound $\mathrm{D}(p^\pi \,\|\, q^\pi)$ from below. From there, we can use divergence decomposition to lower bound $\mathbb{E}_p[S_i(\pi)]$. To make this concrete, we return to the coin tossing experiment.

1.2 Approximate Optimality of Fixed Design

Theorem 1.5. *To distinguish a fair coin from an ε-biased coin with error probability at most δ, $O(\log(1/\delta)/\varepsilon^2)$ samples are necessary and sufficient.*

Proof. We have already proven sufficiency. For the lower bound, let p, q denote the outcome distributions under Hypothesis A (fair coin) and Hypothesis B (bias $1/2 + \varepsilon/2$) respectively, and let A, B denote the events "guess A", "guess B". Then, if π is a procedure satisfying $p^\pi(B) < \delta$ and $q^\pi(A) < \delta$, we have

$$2\delta > p^\pi(B) + q^\pi(A) \geq \frac{1}{2}\exp(-\mathrm{D}(p^\pi \,\|\, q^\pi)),$$

so $\mathrm{D}(p^\pi \,\|\, q^\pi) \geq \log(1/4\delta)$. If $S(\pi)$ denotes the number of coin tosses, then

$$\begin{aligned}\log\left(\frac{1}{4\delta}\right) \leq \mathrm{D}(p^\pi \,\|\, q^\pi) &= \mathbb{E}_p[S(\pi)] \cdot \mathrm{D}\left(\mathrm{Ber}\left(\frac{1}{2}\right) \,\middle\|\, \mathrm{Ber}\left(\frac{1+\varepsilon}{2}\right)\right) \\ &< \mathbb{E}_p[S(\pi)] \cdot \varepsilon^2.\end{aligned}$$

Hence, $\log(1/4\delta)/\varepsilon^2$ samples are required in expectation under Hypothesis A. Because the Pinsker inequality is symmetric, this lower bound actually holds under both hypotheses. □

Example #2: Best Arm Identification

Next, we will design a procedure to select one out of n coins with maximum bias. Each step consists of picking a coin and tossing it. We require that the bias of the selected coin is ε-close to maximum bias, with probability at least $1-\delta$, i.e. the procedure is (ε, δ)-PAC, and aim to minimize the expected number of steps under such requirements.

Algorithm 1 Best Arm Identification - Fixed Design Procedure

```
procedure BestArmFixed(s)
    Toss each coin s times
    return coin with highest empirical frequency
```

Theorem 2.6. *If $s = 2\log(n/\delta)/\varepsilon^2$, then* BestArmFixed$(s)$ *is (ε, δ)-PAC and has total sample complexity $O_{\varepsilon,\delta}(n \log n)$.*

Proof. To make an incorrect selection, either the best coin or the selected coin must deviate from its expected frequency by at least $\varepsilon/2$. Hoeffding's inequality tells us that the probability of any single coin deviating by $\varepsilon/2$ is less than $\exp(-\varepsilon^2 s/2)$, so the union bound gives that $s = 2\log(n/\delta)/\varepsilon^2$ suffices (n bad events, each occuring with probability less than δ/n). Thus, the total sample complexity is $2n\log(n/\delta)/\varepsilon^2 = O_{\varepsilon,\delta}(n \log n)$. □

Now, we move to a lower bound.

Theorem 2.7. *Any (ε, δ)-PAC algorithm for best arm identification requires $\Omega\left(n \log(1/\delta)/\varepsilon^2\right)$ samples.*

Proof. Define the null model p such that coin 1 has bias $1/2 + \varepsilon$ and all others have bias $1/2$. For $i = 2, \ldots, n$, define the alternative model q_i to be the same as p except with coin i having bias $1/2 + 2\varepsilon$. For $i = 2, \ldots, n$, let B_i denote the event "i selected", and let $A_i =$ "i not selected". Pinsker's inequality requires that

$$2\delta > p^\pi(B_i) + q_i^\pi(A_i) \geq \frac{1}{2}\exp(-\mathrm{D}(p^\pi \,\|\, q_i^\pi)),$$

so divergence decomposition gives

$$\begin{aligned}
\log\left(\frac{1}{4\delta}\right) &\leq \mathrm{D}(p^\pi \,\|\, q_i^\pi) \\
&= \frac{1}{2}\mathbb{E}_p[S_i(\pi)] \cdot \mathrm{D}\left(\mathrm{Ber}\left(1/2\right) \,\|\, \mathrm{Ber}\left(1/2 + 2\varepsilon\right)\right) \\
&< 8\varepsilon^2 \mathbb{E}_p[S_i(\pi)].
\end{aligned}$$

Summing over i, we find that

$$\sum_{i=2}^{n} \mathbb{E}_p[S_i(\pi)] > \frac{\log\left(\frac{1}{4\delta}\right)}{8\varepsilon^2}(n-1).$$

Thus, for any sequential protocol, at least $\Omega\left(n \log(1/\delta)/\varepsilon^2\right)$ samples are necessary. □

We just saw that $O\left(n \log(n/\delta)/\varepsilon^2\right)$ samples sufficed via the fixed design procedure – what about the $\log(n)$ factor? It turns out that there is an (ε, δ)-PAC sequential procedure which avoids the $\log(n)$: *median elimination* (Even-Dar et al. [2006]).

2.3 Median Elimination

The main idea of this algorithm is to run in phases, with each phase eliminating half of the remaining arms (based on empirical frequency), until only 1 remains. The goal of phase j is to decrease the bias of the best remaining arm by no more than ε_j with probability at least $1 - \delta_j$. Setting $\delta_j = \delta/2^j$ and $\varepsilon_j = 1/3\,(3/4)^j$ will ensure the (ε, δ)-PAC property.

Algorithm 2 Best Arm Identification - Median Elimination

```
procedure MEDIANELIMINATION
    while multiple coins remain do
        Toss each coin s_j = ⌈2 log(3/δ_j)/ε_j^2⌉ times
        Eliminate half of the coins according to empirical frequency
    return remaining coin
```

Theorem 2.8. MEDIANELIMINATION *is* (ε, δ)*-PAC and requires* $O\left(n \log(1/\delta)/\varepsilon^2\right)$ *samples.*

Proof. By our choice of s_j, we can bound

$$\Pr\left(\text{arm } i \text{ error} > \frac{\varepsilon_j}{2}\right) < \exp\left(-\frac{\varepsilon_j^2 s_j}{2}\right) \leq \frac{\delta_j}{3}.$$

Eliminating *every* ε_j-good arm requires one of two bad events:

1. Error on best arm: probability $\delta_j/3$.
2. Fraction of errors on other arms exceeds $1/2$: since the expected fraction of arms with errors is at most $\delta_j/3$, Markov's inequality gives that the probability of this many errors is no greater than $(\delta_j/3)/(1/2) \leq 2\delta_j/3$.

Hence, a ε_j-good arm remains with probability δ_j, as desired. Next, we observe that

$$s_j = \left\lceil \frac{2}{\varepsilon^2} \log\left(\frac{3}{\delta_j}\right) \right\rceil = O\left(\frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \cdot j \left(\frac{4}{3}\right)^{2j}\right),$$

so the total sample complexity is at most

$$\sum_{j=1}^{\log_2 n} s_j \cdot \frac{n}{2^j} = O\left(\frac{n \log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \cdot \sum_{j=1}^{\infty} j \left(\frac{4}{3}\right)^{2j} 2^{-j}\right) = O\left(\frac{n \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}\right),$$

matching the information-theoretic lower bound up to constant factors. □

In summary, to select an arm that is ε-close to optimal with probability $1 - \delta$, $O\left(n \log(1/\delta)/\varepsilon^2\right)$ samples are necessary and sufficient. Fixed design procedures, which sample each arm a pre-specified number of times, must inflate the number of samples by a factor of $\log(n)$ in order to mitigate selection bias. Median elimination culls unpromising arms, drawing more samples from the surviving ones to improve estimation accuracy. This mitigates selection bias in a more sample-efficient manner.

Example # 3: Multi-Armed Bandits

In this scenario, we are given n arms, where arm i produces random payoff $R_{i,t} \in [0, 1]$ at step t by sampling from an unknown stationary distribution F_i. At each step, we pull an arm and observe its reward. Let μ_i denote the expected payoff of arm i, and let $\mu_* = \max\{u_i\}$. Then, the (pseudo) regret of policy π at time T is

$$R(\pi, T) = \mu_* T - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{\pi(t)}\right].$$

Standard regret compares the expected cumulative payoff to the best possible payoff, but this formulation is a more realistic baseline for comparison. Our goal, as usual, is to minimize regret. This setup distills the omnipresent "exploration vs exploitation" trade-off to its core: pulling suboptimal arms has an opportunity cost but is necessary in order to confidently identify the best arm.

3.4 The UCB1 Algorithm

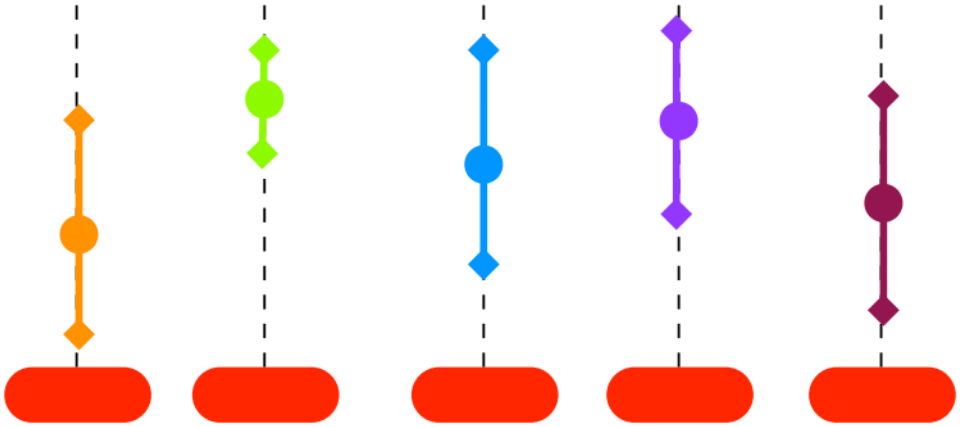
We will analyze the following algorithm due to Auer [2002].

Algorithm 3 Multi-Armed Bandits - Upper Confidence Bound

1: **procedure** UCB1
2: Play each arm once
3: Maintain a confidence interval for each arm, centered at its empirical average (with radius specified below)
4: Always pull arm with highest upper confidence bound

Specifically, an arm sampled s times in the first t steps is given confidence radius $\sqrt{\log(t)/s}$.

Figure 1: State of UCB1 midway through algorithm



Theorem 3.9. *The regret of* UCB1 *is bounded by*

$$R(\text{UCB1}, T) < \frac{\pi^2}{6} n + 4 \log(T) \sum_{i:\Delta_i > 0} \frac{1}{\Delta_i},$$

where $\Delta_i = \mu_* - \mu_i$.

Proof. By our selection of confidence radius, Hoeffding's inequality implies that

$$\Pr(\text{true mean outside confidence interval}) = O(t^{-2}).$$

We call a time step "weird" if at least one arm violates its confidence interval, so that

$$\mathbb{E}[\#\text{ weird steps}] < n \sum_{t=1}^{\infty} t^{-2} = \frac{\pi^2}{6} n.$$

Excluding weird time steps, arm i with $\mu_* - \mu_i = \Delta_i > 0$ is only pulled if its confidence interval is wide enough to bridge the gap to the best arm, i.e. $2\sqrt{\log(t)/s_i} \geq \Delta_i$. Among the first T time steps, there are at most $4\log(T)/\Delta_i^2$ non-weird pulls of arm i. Thus,

$$R(\text{UCB1}, T) < \frac{\pi^2}{6} n + 4\log(T) \sum_{i:\Delta_i > 0} \frac{1}{\Delta_i}. \qquad \square$$

As before, a KL divergence argument (omitted) establishes that this is optimal, up to constant factors.

References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.