The Algorithmic Foundations of Adaptive Data Analysis May 9, 2018

Lecture 20: Lower Bounds via Tracing

Lecturer: Adam Smith *Scribe: Adam Smith*

In this lecture, we complete the proof of the lower bound showing that no adaptive mechanism can answer $\omega(n^2 \log n)$ queries with nontrivial accuracy in the worst case over distributions.

1 Restricted Mechanisms for Adaptive Data Analysis

We will follow the broad proof outline from last lecture but, in order to allow for a simpler proof, we will use an object that is quite different from a fingerprinting code. (See below for an explanation of the relationship between the two objects.)

We will consider a universe $\{1, ..., N\}$ from which the mechanism receives a multiset $S^1 \subset [N]$ of size n, sampled with replacement from the uniform distribution on $[N]$. The mechanism answers statistical queries $q^1, ..., q^k$ posed adaptively by an analyst, and aims to return answers a^j close to the population mean $q^j(U_N) = \frac{1}{N}\sum_{i=1}^{N} q^j(i)$.

The crucial difference with the model considered previously is that the mechanism does not see a complete description of the queries. Instead, as with fingerprinting codes, we assume that the analyst can "name" possible users of being in the mechanism's sample. In each round j, the mechanism sees only the values $\{q^j(i)\}_{i \in S_j}$ corresponding to the set $S^j \subseteq S^1$ of currently unnamed sample points.

We summarize the game concisely here:

Algorithm 1: Adaptive Statistical Queries with a Restricted Mechanism

for *each round* $j = 1, ...k$ **do**
- A selects a *bias* p_j uniformly at random in $[0, 1]$;
- A selects a query $q^j : [N] \to \{0, 1\}$ by choosing each entry to be 1 independently with probability p_j;
- Let $q^j\big|_{S_j}$ be the restriction of q^j to the elements in S^j. This is given to M ;
  `/* The mechanism does not see the values of` q^j `outside of` S^j. `*/`
- M outputs $a^j \in [0, 1]$. The output is given to A ;
- A names a set of individuals $I^j \subseteq [N]$. Set $S^{j+1} := S^j \setminus I^j$.

As with fingerprinting codes, we will eventually remove the restriction on the mechanism by encrypting query values (and augmenting data points with their corresponding secret keys).

Measuring accuracy The main accuracy requirement on M is that it returns a value a^j close to p_j for most rounds j. Namely, we consider the mechanism's *mean (absolute) population error*. For a given round j, define

$$\mathrm{err}_p^j \stackrel{\text{def}}{=} \mathbb{E}\left|a^j - p_j\right| \tag{1}$$

We require that this be bounded at all rounds, namely that for all analyst strategies and for all i.i.d distributions over data points,

$$\max_{j=1,...k} \mathrm{err}_p^j \leq \alpha\,. \tag{2}$$

Our goal will be to find an analyst strategy that prevents the mechanism from answering accurately by accusing many users in the sample, and none outside of it.

We will also consider the *mean (absolute) empirical error*. In a given round j, this is defined the error in estimating the *observed* proportion of 1's in the set S_j (that is, the average of the query values it sees):

$$\mathrm{err}^j_{emp} \stackrel{\text{def}}{=} \mathbb{E}\left| a^j - \underbrace{\frac{1}{|S_j|}\sum_{i\in S^j} q^j(i)}_{\text{empirical mean over } S^j} \right| \tag{3}$$

Relation to fingerprinting codes The object we construct in this lecture is not directly comparable to an interactive fingerprinting code. Recall that, in an interactive fingerprinting code, in each round there is a codeword c^j chosen. The pirate receives the restriction c^j of the codeword to S^j (the set of currently unnamed users in the pirate's coalition), and must answer with a single bit that is *consistent* with $c^j_{S_j}$: if all the entries of $c^j_{S_j}$ are the same, the pirate must answer with that value.

In contrast, the query here plays the role of the codeword. The pirate/mechanism answers a real number that is close on average to the average of the codeword's entries. It is easier to construct the tracer/analyst for this model, and such a model suffices for our needs. See Steinke and Ullman [5] for how similar ideas can be used to construct full interactive fingerprinting codes.

1.1 Analyst's Goals

Our requirements on the analyst's accusation strategy will be:

1. **Soundness:** No user is ever falsely named (i.e. named when it is not in the sample). For all mechanisms M:
$$\Pr\left(\exists j \in [k] \text{ s.t. } I^j \not\subseteq S^j\right) \leq 0.01$$
2. **Completeness for mechanisms with bounded empirical error:** Suppose M satisfies the property that for every round j, and conditioned on every partial transcript $q^1, a^1, I^1, ..., q^{j-1}, a^{j-1}, I^{j-1}$ for which $S^j \neq 0$, we have $\mathrm{err}^j_{emp} \leq \delta$ where δ is a constant to be set later. Then, after k rounds, all users in the sample will have been named:
$$\Pr(S^k \neq \emptyset) \leq 0.01$$

.

Theorem 1 *For any $\delta < \frac{1}{6}$, there exists an analyst strategy that uses $k = O(n^2 \log n)$ rounds and satsifes both soudness and completeness for mechanisms with empirical error δ.*

This will allow us to prove the main result:

Theorem 2 *For every $\alpha < \frac{1}{6}$, there is an analyst strategy A asking $k = O(n^2 \log n)$ queries for which there is no mechanism that achieves population error less than α.*

2 Designing the Analyst Strategy

The main idea for the analyst will be to maintain a score, for each possible data point $i \in [N]$ that measures how much evidence for i's presence has been revealed so far by the mechanism. Specifically, at round ℓ, let

$$\mathsf{score}_\ell(i) \stackrel{\text{def}}{=} \sum_{j=1}^{\ell} (a^j - p_j)(q^j(i) - p_j) \tag{4}$$

The analyst will name those points whose score rises above a threshold

$$\tau \stackrel{\text{def}}{=} \sqrt{k \ln(kN/0.01)}. \tag{5}$$

Algorithm 2: The analyst's accusation strategy

Set $\tau = \sqrt{k \ln(kN/0.01)}$. **for** *each round* j **do**
  Select p_j, q^j as in Game 1 ;
  Receive a^j from M;
  for $i \in [N]$ **do**
    $$\mathsf{score}_j(i) = \begin{cases} \mathsf{score}_{j-1}(i) + (a^j - p_j)(q^j(i) - p_j) & \text{if } i \notin I^j \\ \mathsf{score}_{j-1}(i) & \text{if } i \in I^j \end{cases};$$
  Output $I^j = \{i \ : \ \mathsf{score}_j(i) \geq \tau\}$;

The scores of named users are frozen at the round they are named. The idea will be to show that the scores of users in the sample eventually grow so large that they must be named (or the mechanism must give a highly inaccurate answer).

Suppose we have a mechanism with expected absolute empirical error at most δ.

Recall that a real-valued random variable X is σ^2-subgaussian if its moment-generating function $\mathbb{E}(e^{tX})$ is bounded above by the MGF of a normally-distributed random variable with mean 0 and variance σ^2, that is, $E(e^{tX}) \leq \exp(\sigma^2 t^2/2)$.Such random variables are concentrated about 0: for all $x \geq 0$, we have $\Pr(|X| > x) \leq \exp(-(x/\sigma)^2/2)$. Furthermore, if X and Y are independent random variables that are (respectively) ρ^2- and σ^2-subgaussian, then their sum $X + Y$ is $(\sigma^2 + \rho^2)$-sugaussian.

We show two main lemmas:

Lemma 3 *For every user i not in the sample ($i \notin S^1$), $\mathsf{score}_j(i)$ is σ^2-subgaussian with $\sigma \leq \sqrt{j}$.*

In particular, the probability that there exists a round j and user $i \notin S^1$, for which $\mathsf{score}_j(i) \geq \tau$ is at most 0.01.

Lemma 4 *For every round j, for every partial transcript $T_{j-1} = (q^1, a^1, I^1, ..., q^{j-1}, a^{j-1}, I^{j-1})$ for which $S^j \neq \emptyset$, the sum of scores of unnamed users in the sample increases by a constant in expectation for sufficiently accurate mechanisms. Specifically,*

$$\mathbb{E}\left(\sum_{i \in S^j} (\mathsf{score}_j(i) - \mathsf{score}_{j-1}(i)) \,\middle|\, T_{j-1} \right) \geq \frac{1}{6} - \delta,$$

where $\delta \geq \mathrm{err}^j_{emp}$ is the mechanism's empirical mean absolute error guarantee.

In particular, for $\delta < 1/6$, the sum of scores increases by a constant in expectation at each round (no matter what happened at previous rounds).

Combining this lemma with Azuma's inequality, we get the following corollary:

Corollary 5 *For every $\beta > 0$ and $j \in \{1, ..., k\}$: at round j, with probability at least $1 - \beta$, the sum of scores of users in the sample is at least $k(\frac{1}{6} - \delta) - \sqrt{j \ln(1/\beta)}$.*

Why are these lemmas enough? We can prove the main result (Theorems 1 and 2) by combining the previous lemmas:

1. Lemma 3 shows that the scores of users outside of the sample will stay below $\tau \approx \sqrt{k}$, so they will never be named.
2. Corollary 5 says that the sum of scores of unnamed members of the sample will increase over time at a rate of $\Omega(1)$ per round as long as some unnamed user remains.

Once the score rises above $n(\tau + 1) \approx n\sqrt{k}$, then all users will have been named (since the scores of named users get frozen at τ). Selecting $k \gg n^2$ is sufficient to guarantee that all users will be named.

3 Analyzing the Scores

The first lemma follows from the fact that a restricted mechanism sees only the values of the query at the points in the sample.

3.1 Out-of-sample users (Lemma 3)

Proof of Lemma 3: The key observation is that, for users i outside of the sample, *the restricted mechanism's view is independent of the query values at i once we condition on the population mean p_j.* Thus, we can imagine selecting p_j and a^j before user i's value $q^j(i)$. Conditioned on p_j and a^j, the increment in user i's score $(a^j - p_j)(q^j(i) - p_j)$ has expectation 0 and absolute value at most 1, so it is 1-subgaussian.

If we condition on the entire sequence of p_j's and a^j's, the score at round j is a sum of independent 1-subgaussian terms, and so is j-subgaussian. A convex combination of subgaussian random variables is also subgaussian, and so we get the lemma statement. ■

3.2 In-sample users (Lemma 4)

Proof of Lemma 4: The proof of the second lemma is more challenging. To develop some intuition, let's fix a partial transcript T_{j-1} for the first $j-1$ rounds, and let n_j denote $|S_j|$, the number of unaccused users (which we assume to be at least 1).

The information seen by the attacker at round j is $q^j\big|_{S_j}$. By symmetry, the information conveyed about p_j can be summarized by the number number of ones in $q^j\big|_{S_j}$, which we denote by t. Let e_t be the expected empirical error of the mechanism on a uniformly chosen input with t ones, that is,

$$e_t \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\substack{x\in\{0,1\}^{n_j}\\ weight(x)=t}} \left(a^j(x) - \tfrac{t}{n}\right).$$

To make the notation more compact, we write $x \sim t$ to denote selecting a uniform string in $\{0,1\}^{n_j}$ of weight t, so that $e_t \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{x\sim t} \left(a^j(x) - \frac{t}{n}\right)$.

Now consider how the sum of the scores of players in S_j increases at round j. To simplify notation, we will drop the subscript and superscript j. Given a particular query q, we have

$$\begin{aligned}
\sum_{i\in S^j} (\mathsf{score}_j(i) - \mathsf{score}_{j-1}(i)) &= \sum_{i\in S} (a(q|_S) - p)(q(i) - p) && (6)\\
&= n(a(q|_S) - p)(\tfrac{1}{n}\sum_{i\in S_j} q(i) - p) && (7)\\
&= n(a(q|_S) - p)\left(\tfrac{t}{n} - p\right) && (8)\\
&= n\underbrace{\left(a(q|_S) - \tfrac{t}{n}\right)}_{\substack{\text{expectation}\\ e_t \text{ for}\\ \text{fixed } t}}\left(\tfrac{t}{n} - p\right) + n\underbrace{\left(\tfrac{t}{n} - p\right)^2}_{\substack{\text{expectation}\\ p(1-p) \text{ for}\\ \text{fixed } p}} && (9)
\end{aligned}$$

We can now take the expectation over the choice of $p \sim uniform([0,1])$ (the bias of the coin) and $t \sim B(n,p)$ (the number of ones seen by the mechanism) and the mechanism's random coins. We will take these expectations in different orders to bound the two terms in (9).

To bound the first term, we first sample t and then sample p from its conditional distribution given t. For p drawn uniformly in $[0,1]$, the conditional distribution on p given t is known exactly.

For any positive real numbers α, β, let $Beta(\alpha,\beta)$ be the continuous distribution on $[0,1]$ with density $f(x)$ proportional to $x^{\alpha-1}(1-x)^{\beta-1}$.

Lemma 6 *If $p \sim uniform([0,1])$ and $t \sim B(n,p)$, then for each possible value $t \in \{0,...,n\}$, the conditional distribution of $p|t$ is $Beta(t+1, n-t+1)$.*

Proof Consider an observed value t of ones, and a candidate value p for the coin's bias. By Bayes' rule, the density of p given t scales as

$$\frac{d}{dx}\Pr(p \le x|t) = \underbrace{\binom{n}{t}x^t(1-x)^{n-t}}_{\Pr(t|p=x)} \times \frac{d}{dx}\Pr(p \le x) \times \frac{1}{\Pr(t)}$$

Since the prior on p is uniform on $[0,1]$, the conditional probability of p is $p^t(1-p)^{n-t}$ times a normalizing constant. This is the definition of $Beta(t+1, n-t+1)$. ■

If a random variable follows the $Beta(\alpha,\beta)$ distribution, then its mean is $\frac{\alpha}{\alpha+\beta}$. We have the following immediate corollary:

Corollary 7 *If $p \sim uniform([0,1])$ and $t \sim B(n,p)$, then for each possible value $t \in \{0,...,n\}$, the conditional mean $\mathbb{E}(p|t)$ is $\frac{t+1}{n+2}$.*

We can now bound the expectation of the first term in (9), by first selecting t and then selecting p from the conditional distribution. Once we condition on t, the mechanism's error $\left(a(q|_S) - \frac{t}{n}\right)$ becomes independent of p and of the sampling error $\left(\frac{t}{n} - p\right)$.

$$\begin{aligned} \mathop{\mathbb{E}}_{p,t,coins}\left(n\left(a(q|_S) - \tfrac{t}{n}\right)\left(\tfrac{t}{n} - p\right)\right) &= \mathop{\mathbb{E}}_{t}\left(n \cdot \underbrace{\mathop{\mathbb{E}}_{coins}\left(a(q|_S) - \tfrac{t}{n}\right)}_{e_t} \cdot \underbrace{\mathop{\mathbb{E}}_{p|t}\left(\tfrac{t}{n} - p\right)}_{\frac{t}{n} - \frac{t+1}{n+2}}\right) \quad (10) \\ &= \mathop{\mathbb{E}}_{t}\left(n \cdot e_t \cdot \left(\frac{t}{n} - \frac{t+1}{n+2}\right)\right) \quad (11) \end{aligned}$$

where the second line follows from the definition of e_t, and Corollary 7. Now the term $\left(\frac{t}{n} - \frac{t+1}{n+2}\right)$ equals $\frac{2t-1}{n(n+2)}$, which is at most $1/n$ in absolute value. So we have

$$\mathop{\mathbb{E}}_{p,t,coins}\left(n\left(a(q|_S) - \tfrac{t}{n}\right)\left(\tfrac{t}{n} - p\right)\right) \geq -\mathbb{E}_t\left|e_t\right| . \quad (12)$$

The second term in (9), $n\,\mathbb{E}_{p,t}(\frac{t}{n} - p)^2$, can be computed explicitly, since $\mathbb{E}_t(\frac{t}{n} - p) = Var(\frac{t}{n}|p)$ for each particular p. Since t is binomial, it's variance is $np(1-p)$. We therefore obtain:

$$n\,\mathbb{E}_{p,t}\left(\tfrac{t}{n} - p\right)^2 = n\,\mathbb{E}_p\left(\tfrac{1}{n^2} \cdot np(1-p)\right) = \mathbb{E}_p\, p(1-p) = 1/6\,. \quad (13)$$

Plugging our bounds on the two terms ((12) and (13)) into the formula for the score increase (9), we get that

$$\mathbb{E}\left(\sum_{i \in S^j}\left(\mathsf{score}_j(i) - \mathsf{score}_{j-1}(i)\right)\Big| T_{j-1}\right) \geq \frac{1}{6} - \mathbb{E}_t\left|e_t\right| \geq \frac{1}{6} - \mathrm{err}^j_{emp}$$

as desired. ■

3.3 Completing the Main Results

We leave the complete proofs of Theorems 1 and 2) as an exercise.

Exercise 1 *Use Lemmas 3 and 4 to prove Theorems 1 and 2, using the outline at the end of Section 2.*

Remark Lemma 4 shows that analyst's tracing strategy works whenever the mean absolute error is less than $1/6$ and the mechanism answers $\tilde{\Omega}(n^2)$ queries. This may be a little loose: a mechanism which ignores the data and always returns $\frac{1}{2}$ as its answer has mean absolute error $1/4$ against this particular analyst (since the poplulation mean is chosen uniformly at random).

One can in fact get a tight result by considering root mean squared error instead. The trivial mechanism which always answers $1/2$ has root mean squared error $1/\sqrt{12}$. One can modify Lemma 4 to work for any mechanism with root mean squared error strictly less than $1/\sqrt{12}$. We leave this as an exercise for the reader.

Exercise 2 *Show that if the mechanism has root mean squared empirical error $\delta_{RMSE} = \sqrt{\mathbb{E}_t(e_t^2)}$ at each round, then the expected increase in the sum of in-sample points' scores (as in Lemma 4) is a positive constant whenever $\delta^2_{RMSE} < 1/\sqrt{12}$.*

Bibliographic Notes

The theorems we sketch in this lecture are due to Steinke and Ullman [5], who also analyze the interactive fingerprinting code used as the core of the reduction. The simplified proof we give comes from work on tracing attacks in data privacy (Bun et al. [1], Dwork et al. [3], Steinke and Ullman [6]) and is based on conversations with Jon Ullman. Lemma 4 is a simplification of Lemma A.5 in Bun et al. [2]. Earlier work by Hardt and Ullman [4] proved analogous impossibility results for answering $O(n^3)$ adaptively chosen queries, basing their reductions on n iterations of a similar attack based on non-interactive fingerprinting codes, with code-length $O(n^2)$. Their proof in particular implies that only $O(n)$ rounds of adaptivity are necessary for computational hardness.

References

[1] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, pages 1–10. ACM, May 31 – June 3 2014.

[2] Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1306–1325, 2017. doi: 10.1137/1.9781611974782.85. URL `https://doi.org/10.1137/1.9781611974782.85`.

[3] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 650–669, Oct 2015. doi: 10.1109/FOCS.2015.46.

[4] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 454–463. IEEE, 2014.

[5] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on Learning Theory*, pages 1588–1628, 2015.

[6] Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 552–563, 2017. doi: 10.1109/FOCS.2017.57. URL `https://doi.org/10.1109/FOCS.2017.57`.