The Algorithmic Foundations of Adaptive Data Analysis September, 2017

Lecture 4

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Description Length Bounds

In this class, we try for the first time to provide *upper bounds* on the widths of confidence intervals we can endow the answers to statistical queries with, when the queries are selected adaptively. Rather than assuming something about the class of analyses to be performed (as we would if we wanted to take the "uniform convergence" approach), our goal will be to provide confidence intervals around statistical queries that may have been chosen by an *arbitrary* adaptive adversary. Because we know that if such an adversary has access to the *exact* empirical answers to statistical queries $\mathbb{E}_S[q]$ he can arbitrarily overfit after only 2 queries, in order to give non-trivial guarantees in this setting, we will need to further limit the kind of access the analyst has to the data. The goal will to do so in a way that both allows us to control how much the analyst overfits the data, while allowing him to still perform useful analyses.

We formalize the setting by modeling a *game* between an (arbitrary) data analyst $\mathcal{A}$ and a statistical estimator $\mathcal{O}$. We think of $\mathcal{A}$ and $\mathcal{O}$ as both being *stateful* algorithms — i.e. their behavior at round t can depend on the transcript of their interaction up through round $t-1$. We state the game in a more general setting, allowing the analyst to choose queries from an abstract space $\mathcal{Q}$ (not necessarily statistical queries) so that we will later be able to generalize the statements we prove beyond statistical queries.

GenerateTranscript$_{n,k}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})$

S is given to $\mathcal{O}$.
for $i = 1$ to k **do**
  $\mathcal{A}$ chooses a query $q_i \in Q$. q_i is given to $\mathcal{O}$.
  $\mathcal{O}$ generates an answer $a_i \in [0,1]$. a_i is given to $\mathcal{A}$.
end for
The *transcript* $T = (q_1, a_1, \ldots, q_k, a_k)$ is output

---

We will elide n, k, and $\mathcal{Q}$ when they are clear from context, and shorten "GT" to GT, simply writing $\mathbf{GT}(\mathcal{A}, S, \mathcal{O})$

With this interaction formally defined, we can now define what we mean by endowing a statistical estimator $\mathcal{O}$ with confidence intervals in the adaptive setting.

Definition 1 *A statistical estimator $\mathcal{O}$ (a stateful algorithm, parameterized by a dataset S, mapping queries to answers) is (ϵ, δ)-accurate for a class of queries $\mathcal{Q}$ if for* every *algorithm $\mathcal{A}$, and for* every *distribution $\mathcal{D}$, with probability $1-\delta$, when $S \sim \mathcal{D}^n$ consists of n samples drawn i.i.d. from $\mathcal{D}$, the transcript T generated by $\boldsymbol{GT}(\mathcal{A}, \mathcal{D}, \mathcal{O}, \mathcal{Q})$ has the property that:*

$$\max_{i=1}^{k} |\text{val}(q_i, \mathcal{D}) - a_i| \leq \epsilon$$

where $\text{val}(q_i, \mathcal{D})$ *denotes the "true value" of the query q_i on the underlying distribution $\mathcal{D}$. When $\mathcal{Q}$ is the set of statistical queries,* $\text{val}(q_i, \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[q_i(x)]$.

We start by making a basic observation about randomness. In this class, we will sometimes want to design statistical estimators $\mathcal{O}$ that employ randomness. We also want our theorems to hold for data analysts $\mathcal{A}$ which are randomized. But when we are proving that a statistical estimator $\mathcal{O}$ is (ϵ, δ) accurate, then without loss of generality it suffices to consider *deterministic* analysts $\mathcal{A}$. Suppose we have a randomized algorithm $\mathcal{A}$ that makes use of b bits of randomness. We can write such an algorithm as a deterministic algorithm that takes as input a uniformly random b bit string $r \in \{0,1\}^b$: $\mathcal{A}(r)$. Now

suppose that a statistical estimator $\mathcal{O}$ is not (ϵ, δ)-accurate. That means that there is some analyst $\mathcal{A}(r)$ such that

$$\Pr_{S\sim\mathcal{D}^n, r\sim\{0,1\}^b \; T\sim\mathbf{GT}(\mathcal{A}(r),S,\mathcal{O})}[\max_{i=1}^{k} |q_i(\mathcal{D}) - a_i| > \epsilon] \geq \delta.$$

But we have:

$$\Pr_{S\sim\mathcal{D}^n, r\sim\{0,1\}^b T\sim\mathbf{GT}(\mathcal{A}(r),S,\mathcal{O})}[\max_{i=1}^{k} |q_i(\mathcal{D}) - a_i| > \epsilon] \leq \max_r \Pr_{S\sim\mathcal{D}^n, T\sim\mathbf{GT}(\mathcal{A}(r),\mathcal{O})}[\max_{i=1}^{k} |q_i(\mathcal{D}) - a_i| > \epsilon]$$

So, we must also have that there exists some r such that $\Pr_{S\sim\mathcal{D}^n, T\sim\mathbf{GT}(\mathcal{A}(r),\mathcal{O})}[\max_{i=1}^{k} |q_i(\mathcal{D}) - a_i| > \epsilon] \geq \delta$. But by definition, $\mathcal{A}(r)$ is a deterministic analyst. The contrapositive is that if we can prove (ϵ, δ) accuracy for all deterministic analysts, we must have proven it also for randomized analysts.

Thus, for the remainder of the course, we will without loss of generality constrain our attention to deterministic analysts $\mathcal{A}$.

We will show that if the *transcript* generated by the interaction between the statistical estimator $\mathcal{S}$ and the analyst $\mathcal{A}$ is *compressible* for any $\mathcal{A}$, then the estimator $\mathcal{S}$ must be *accurate*, with parameters relating to how compressible the transcript is.

Definition 2 *We say that a statistical estimator $\mathcal{O}$ for a class of queries $\mathcal{Q}$ is* transcript compressible *to $b(n,k)$ bits if for every analyst $\mathcal{A}$ there is a set of transcripts $H_\mathcal{A}$ of size $|H_\mathcal{A}| \leq 2^{b(n,k)}$ such that for every dataset $S \in \mathcal{X}^n$*

$$\Pr[\mathbf{GT}_{n,k}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q}) \in H_\mathcal{A}] = 1$$

Note that the set of transcripts $H_\mathcal{A}$ which we must bound to show that an estimator $\mathcal{O}$ is compressible is allowed to depend on the analyst $\mathcal{A}$, and does not have to be uniformly bounded over all $\mathcal{A}$. This will be important. It will mean, in particular, that a given transcript $h = (q_1, a_1, \ldots, q_k, a_k)$ can be described more compactly as $(a_1, \ldots, a_k)$, because (together with our ability to assume WLOG that $\mathcal{A}$ is deterministic), each query q_i is a deterministic function of $(a_1, \ldots, a_{i-1})$ and $\mathcal{A}$.

Theorem 3 *Any $b(n,k)$-compressible estimator $\mathcal{O}$ for statistical queries will have the property that for every data analyst $\mathcal{A}$ and every distribution $\mathcal{D}$, with probability $1-\delta$ over the sampled dataset $S \sim \mathcal{D}^n$ and generated transcript $h = \mathbf{GT}_{n,k}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})$:*

$$\max_{i=1}^{k} |\mathbb{E}_S[q_i] - \mathbb{E}_\mathcal{D}[q_i]| \leq \sqrt{\frac{(b(n,k)+1)\ln(2) + \ln(k/\delta)}{2n}}$$

In fact, Theorem 3 will follow as a corollary from a more general theorem:

Theorem 4 *Fix a distribution $\mathcal{D}$. Fix a history h, and let $R(h) \subseteq \mathcal{X}^n$ be an arbitrary set of datasets of size n. Suppose for each h, $\Pr_{S'\sim\mathcal{D}^n}[S' \in R(h)] \leq \delta$. Then, for any $\mathcal{A}$:*

$$\Pr_{S\sim\mathcal{D}^n, h\sim\mathbf{GT}_{n,k}(\mathcal{A},S,\mathcal{O},\mathcal{Q})}[S \in R(h)] \leq 2^{b(n,k)}\delta$$

Proof Once we fix a data analyst $\mathcal{A}$ we know that there is a set $H_\mathcal{A}$ of size $|H_\mathcal{A}| \leq 2^{b(n,k)}$ such that for all S, $\Pr[\mathbf{GT}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q}) \in H_\mathcal{A}] = 1$. So, we know that:

$$\begin{aligned}
\Pr_{S\sim\mathcal{D}^n, h\sim\mathbf{GT}(\mathcal{A},S,\mathcal{O},\mathcal{Q})}[S \in R(h)] &\leq \Pr_{S\sim\mathcal{D}^n}[S \in \cup_{h'\in H_\mathcal{A}} R(h')] \\
&\leq \sum_{h'\in H_\mathcal{A}} \Pr_{S\sim\mathcal{D}^n}[S \in R(h')] \\
&\leq |H_\mathcal{A}|\delta \\
&\leq 2^{b(n,k)}\delta
\end{aligned}$$

■

We can now prove Theorem 3 as a corollary of Theorem 4.

Proof [Theorem 3] Fix any distribution $\mathcal{D}$ and transcript $h = (q_1, a_1, \ldots, q_k, a_k)$. Define the event:

$$R(h) = \{S' : \max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| \geq \sqrt{\frac{\ln(2k/\delta')}{2n}}\}$$

for a value of δ' to be determined. Recall from several lectures ago: a Chernoff bound tells us that for any *fixed* set of k queries:

$$\Pr_{S \sim \mathcal{D}^n}[\max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| \geq \sqrt{\frac{\ln(2k/\delta')}{2n}}] \leq \delta'$$

. Thus, we have that for every fixed history h, $\Pr_{S' \sim \mathcal{D}^n}[S' \in R(h)] \leq \delta'$. Applying Theorem 4, we therefore know that:

$$\begin{aligned} \Pr_{S \sim \mathcal{D}^n, h \sim \mathbf{GT}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})}[\max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| \geq \sqrt{\frac{\ln(2k/\delta')}{2n}}] &= \Pr_{S \sim \mathcal{D}^n, h \sim \mathbf{GT}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})}[S \in R(h)] \\ &\leq 2^{b(n,k)}\delta' \end{aligned}$$

Setting $\delta' = \delta/2^{b(n,k)}$ and plugging this value in above, we have:

$$\Pr_{S \sim \mathcal{D}^n, h \sim \mathbf{GT}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})}\left[\max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| \geq \sqrt{\frac{\ln(2^{b(n,k)+1}k/\delta)}{2n}}\right] \leq \delta$$

Nothing that $\ln(2^{b(n,k)+1}k/\delta) = (b(n,k)+1)\ln(2) + \ln(k/\delta)$ yields the theorem. ■

Great! But we are not done yet. There are a couple of things to complain about the theorem we just proved:

1. All we have argued is that $|\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]|$ is small: not that $|a_i - \mathbb{E}_{\mathcal{D}}[q_i]|$ is small. In particular, we know that our queries haven't substantially overfit the dataset S, but we don't know if our answers a_i are close to accurate. In particular, it is easy to give a statistical estimator $\mathcal{O}$ that is transcript compressible to 0 bits: $\mathcal{O}$ just ignores queries q_i, and always answers with $a_i = 0$. This is not useful. In this case, the bound we proved in 3 exactly recovers the bound on $|\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]|$ we proved for non-adaptively chosen queries using the Chernoff bound. This is good: when the estimator $\mathcal{O}$ is providing no information about the dataset, the transcript is non-adaptively chosen.

2. Are there any non-trivial statistical estimators that are non-trivially transcript compressible? Or is this kind of theorem vacuous?

We'll address the first question now, and the second one next lecture.

Lets define a notion of accuracy with respect to the sample S (rather than with respect to the distribution $\mathcal{D}$):

Definition 5 *A statistical estimator $\mathcal{O}$ is (ϵ, δ)-sample accurate for a class of queries $\mathcal{Q}$ if for* every *algorithm $\mathcal{A}$, and for* every *dataset of size n $S \in \mathcal{X}^n$, with probability $1 - \delta$, the transcript T generated by $\mathbf{GT}_{n,k}(\mathcal{A}, S, \mathcal{O}, \mathcal{Q})$ has the property that:*

$$\max_{i=1}^{k} |\text{val}(q_i, S) - a_i| \leq \epsilon$$

where $\text{val}(q_i, S)$ *denotes the "true value" of the query q_i on the dataset S. When $\mathcal{Q}$ is the set of statistical queries,* $\text{val}(q_i, S) = \mathbb{E}_S[q_i]$.

With this definition in hand, we obtain the following simple theorem, instantiated here for statistical queries:

Theorem 6 (Transcript Compressibility Transfer Theorem) *For any $\delta'' > 0$, a statistical estimator $\mathcal{O}$ for statistical queries that is:*

1. *$b(n,k)$-compressible and*

2. *(ϵ', δ')-sample accurate*

is (ϵ, δ) accurate, where $\delta = \delta' + \delta''$ and

$$\epsilon = \epsilon' + \sqrt{\frac{(b(n,k)+1)\ln(2) + \ln(k/\delta'')}{2n}}$$

Proof Because $\mathcal{O}$ is (ϵ', δ')-sample accurate, we know that:

$$\Pr_{T \sim \mathbf{GT}(\mathcal{A},S,\mathcal{O})}\left[\max_i |\mathbb{E}_S[q_i] - a_i| > \epsilon'\right] \leq \delta'$$

We also know from Theorem 3 and the fact that $\mathcal{O}$ is $b(n,k)$-compressible, that for any distribution $\mathcal{D}$:

$$\Pr_{S \sim \mathcal{D}^n, T \sim \mathbf{GT}(\mathcal{A},S,\mathcal{O})}\left[\max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| > \sqrt{\frac{(b(n,k)+1)\ln(2) + \ln(k/\delta'')}{2n}}\right] \leq \delta''$$

A union bound therefore gives that:

$$\Pr_{S \sim \mathcal{D}^n, T \sim \mathbf{GT}(\mathcal{A},S,\mathcal{O})}\left[\max_i |\mathbb{E}_S[q_i] - a_i| + \max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| > \epsilon' + \sqrt{\frac{(b(n,k)+1)\ln(2) + \ln(k/\delta'')}{2n}}\right] \leq \delta + \delta''$$

But by the triangle inequality, we know that:

$$\begin{aligned}\max_i |a_i - \mathbb{E}_{\mathcal{D}}[q_i]| &\leq \max_i |a_i - \mathbb{E}_S[q_i]| + |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]| \\ &\leq \max_i |a_i - \mathbb{E}_S[q_i]| + \max_i |\mathbb{E}_S[q_i] - \mathbb{E}_{\mathcal{D}}[q_i]|\end{aligned}$$

which completes the proof. ■

Actually, this theorem extends beyond statistical queries to any *low sensitivity* query.

Definition 7 *A query $q : \mathcal{X}^n \to \mathbb{R}$ has sensitivity c if for all $x_1, \ldots, x_n \in \mathcal{X}$, all indices i, and all $x_i' \in \mathcal{X}$:*

$$|q(x_1, \ldots, x_n) - q(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c$$

Note that statistical queries are c-sensitive for $c = 1/n$.

For a low sensitivity query, we will write $q(S)$ to denote its empirical value on S, and $\mathbb{E}_{\mathcal{D}}[q] = \mathbb{E}_{S \sim \mathcal{D}^n}[q(S)]$ to denote its expected value on a data set S with entries drawn independently from $\mathcal{D}$.

It turns out that the same bound we derived to prove confidence intervals around statistical queries (the Chernoff bound) still holds for arbitrary sensitivity $1/n$ queries. The generalization is called McDiarmid's inequality:

Theorem 8 *Fix any distribution $\mathcal{D}$ over $\mathcal{X}$. Let $S \sim \mathcal{D}^n$. Let q be a sensitivity c query. Then for any $t > 0$:*

$$\Pr\left[|q(S) - \mathbb{E}_{\mathcal{D}}[q]| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{n \cdot c^2}\right)$$

Plugging in $c = 1$ we see that this is exactly the bound we proved for sums of independent random variables taking values in $[0,1]$. Since this is the only fact we used about statistical queries, our transfer theorem also extends to arbitrary $1/n$-sensitive queries.

Theorem 9 (Generalized Transcript Compressibility Transfer Theorem) *For any $\delta'' > 0$, a statistical estimator $\mathcal{O}$ for $1/n$ sensitive queries that is:*

1. *$b(n,k)$-compressible and*

2. (ϵ', δ')-*sample accurate*

is (ϵ, δ) *accurate, where* $\delta = \delta' + \delta''$ *and*

$$\epsilon = \epsilon' + \sqrt{\frac{(b(n,k)+1)\ln(2) + \ln(k/\delta'')}{2n}}$$

Next lecture, we will start thinking about how to design useful transcript-compressible estimators.

Bibliographic Information Description length bounds for adaptively answering queries first appeared in [DFH+15]. The game-like formalization **GT** of the interaction between an analyst and a statistical estimator is adapted from [BNS+16].

References

[BNS+16] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1046–1059. ACM, 2016.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2350–2358, 2015.