The Algorithmic Foundations of Adaptive Data Analysis September, 2017

Lecture 5

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Description Length Bounds II

In this lecture, we will start to develop useful estimators for statistical queries that are transcript compressible. We will start with simple estimators, and then build more sophisticated ones. First, we will prove a basic "postprocessing" and "composition theorem" that will allow us to analyze algorithms which are built from compressible sub-routines. For a stateful algorithm $f : \mathcal{Q} \cup \mathcal{R} \to \mathcal{Q} \cup \mathcal{R}$, we write the "postprocessing" of an estimator $\mathcal{O}$ as $f \circ \mathcal{O}$ to denote the following interaction:

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

S is given to $\mathcal{O}$.
for $i = 1$ to k **do**
  $\mathcal{A}$ chooses a query $q_i \in Q$. $\hat{q}_i = f(q_i)$ is given to $\mathcal{O}$.
  $\mathcal{O}$ generates an answer $a_i \in [0, 1]$. $\hat{a}_i = f(a_i)$ is given to $\mathcal{A}$.
end for
The *transcript* $T = (\hat{q}_1, \hat{a}_1, \ldots, \hat{q}_k, \hat{a}_k)$ is output

---

Theorem 1 (Postprocessing for Transcript Compressibility) *Suppose $\mathcal{O} : \mathcal{Q} \to \mathcal{R}$ is b-transcript compressible. Let $f : \mathcal{Q} \cup \mathcal{R} \to \mathcal{Q} \cup \mathcal{R}$ be an arbitrary stateful algorithm. Then, $f \circ \mathcal{O}$ is also b-transcript compressible.*

Proof First, observe that the transcript $T' = (\hat{q}_1, a_1, \ldots, \hat{q}_k, a_k)$ is compressible to b bits, because we may view this as the outcome of an interaction between $\mathcal{O}$ and an analyst $\mathcal{A}'$ that responds to query q_i as $\mathcal{A}$ responds to query $\hat{q}_i$. Since compressibility is quantified over all data analysts $\mathcal{A}'$, we know in particular that for every S, there exists a set $H_{\mathcal{A}'}$ of size $|H_{\mathcal{A}'}| \leq 2^b$ such that:

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

Now define a set $H_{f,\mathcal{A}} = \{h' = (\hat{q}_1, f(a_1), \ldots, \hat{q}_k, f(a_k)) : h \in H_{\mathcal{A}'}\}$. Note that $|H_{f,\mathcal{A}}| \leq |H_{\mathcal{A}'}| \leq 2^b$, and $\mathbf{GT}_{n,k}(\mathcal{A}, S, f \circ \mathcal{O}, \mathcal{Q}) \in H_{f,\mathcal{A}}$ if $\Pr[\mathbf{GT}_{n,k}(\mathcal{A}', S, \mathcal{O}, \mathcal{Q}) \in H_{\mathcal{A}'}]$. So,

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

as desired. ■

We now define what it means to compose two estimators. Together with an analyst $\mathcal{A}$, the composition of two estimators $\mathcal{O}_1, \mathcal{O}_2$ designed to answer k_1 and k_2 queries respectively, generates a transcript according to the following interaction: We write the composition of two estimators as $(\mathcal{O}_1, \mathcal{O}_2)$.

Theorem 2 (Composition for Transcript Compressibility) *Suppose $\mathcal{O}_1 : \mathcal{Q} \to \mathcal{R}$ is transcript compressible to $b_1(n, k_1)$ bits, and $\mathcal{O}_2 : \mathcal{Q} \to \mathcal{R}$ is transcript compressible to $b_2(n, k_2)$ bits. Then the composition $(\mathcal{O}_1, \mathcal{O}_2)$ is transcript compressible to $b(n, k_1 + k_2) = b_1(n, k_1) + b_2(n, k_2)$ bits.*

Proof Since $\mathcal{O}_1$ is $b_1(n, k_1)$-transcript compressible, for any analyst $\mathcal{A}$, we know there is a set $H_{\mathcal{A}}$ of size $|H_{\mathcal{A}}| \leq 2^{b_1(n,k_1)}$ such that for every S, $\Pr[\mathbf{GT}_{n,k_1}(\mathcal{A}, S, \mathcal{O}_1, \mathcal{Q}) \in H_{\mathcal{A}}] = 1$. Write $T_1 = (q_1, a_1, \ldots, q_{k_1}, a_{k_1})$ to denote the fraction of the transcript that has been generated after $\mathcal{A}$ interacts with $\mathcal{O}_1$, and write $\mathcal{A}_{T_1}$ to denote analyst $\mathcal{A}$ at its internal state after it has finished interacting with $\mathcal{O}_1$. Since $\mathcal{O}_2$ is $b_2(n, k_2)$-transcript compressible, for any analyst $\mathcal{A}_{T_1}$, there is a set $H_{\mathcal{A}_{T_1}}$ of size

GenerateTranscript$_{n,k_1+k_2}(\mathcal{A}, S, (\mathcal{O}_1, \mathcal{O}_2), \mathcal{Q})$

S is given to $\mathcal{O}$.
for $i = 1$ to k_1 **do**
  $\mathcal{A}$ chooses a query $q_i \in Q$. q_i is given to $\mathcal{O}_1$.
  $\mathcal{O}_1$ generates an answer $a_i \in [0,1]$. a_i is given to $\mathcal{A}$.
end for
for $i = k_1 + 1$ to $k_1 + k_2$ **do**
  $\mathcal{A}$ chooses a query $q_i \in Q$. q_i is given to $\mathcal{O}_2$.
  $\mathcal{O}_2$ generates an answer $a_i \in [0,1]$. a_i is given to $\mathcal{A}$.
end for
The *transcript* $T = (q_1, a_1, \ldots, q_{k_1+k_2}, a_{k_1+k_2})$ is output

$|H_{\mathcal{A}_{T_1}}| \leq 2^{b_2(n,k_2)}$ such that for every S, $\Pr[\mathbf{GT}_{n,k_2}(\mathcal{A}_{T_1}, S, \mathcal{O}_2, \mathcal{Q}) \in H_{\mathcal{A}_{k_1}}] = 1$. Thus, we have that $T = (T_1, T_2)$ where $T_1 \in H_{\mathcal{A}}$, and $T_2 \in H_{\mathcal{A}_{T_1}}$. The number of such transcripts is at most:

$$\sum_{T_1 \in H_{\mathcal{A}}} |H_{\mathcal{A}_{T_1}}| \leq 2^{b_1(n,k_1)} \cdot 2^{b_2(n,k_2)} = 2^{b_1(n,k_1)+b_2(n,k_2)}$$

■

Ok. So lets build some compressible estimators. We will start with a trivial (and not very impressive) estimator that simply reports empirical averages of statistical queries, to a truncated number of digits of precision.

Definition 3 *Given a dataset S, the b-bit truncated estimator $\mathcal{O}_b^T(q)$ returns $q(S)$ truncated to b bits of binary precision.*

Observation 4 *Trivially, on a single query, $\mathcal{O}_b^T$ is transcript-compressible to b bits. By the composition theorem, it can be composed to answer k queries, and is $b(n,k)$-transcript compressible for $b(n,k) = b \cdot k$. Similarly, $\mathcal{O}_b^T$ is $(1/2^b, 0)$-sample accurate.*

Now that we know the sample accuracy and compressibility of a simple estimator, by the Transcript Compressibility Transfer Theorem from last lecture, we can derive its accuracy:

Theorem 5 *Fix any $k < n$ and $\delta > 0$. When $b = \log\left(\sqrt{\frac{n}{k}}\right)$, the b-bit truncated estimator $\mathcal{O}_b^T$ is (ϵ, δ)-accurate for k $1/n$-sensitive queries, where:*

$$\epsilon = \sqrt{\frac{k}{n}} + \sqrt{\frac{k \cdot \log\left(\sqrt{\frac{n}{k}}\right) + 1)\ln(2) + \ln(k/\delta)}{2n}} = \tilde{O}\left(\sqrt{\frac{k + \ln(1/\delta)}{n}}\right)$$

Proof We simply invoke the Transcript Compressibility transfer theorem, that says that 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

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

For us, $\delta' = 0, \epsilon' = 1/2^b$, and $b(n,k) = k \cdot b$. Note that we could get a slightly better bound by optimizing this expression over b, but the optimal solution does not have a closed form. ■

Although not terribly impressive, note that the truncated estimator does have accuracy that tightly matches the feature-selection and linear classification "attack" we saw several lectures ago, and defeats the 2-query attack (which relies on encoding the dataset in the low order bits of a query answer). So perhaps we have gained something just by hiding the low-order bits of our empirical estimates. Now lets see if we can go beyond this.

We'll start with a basic building block, that does not directly provide numeric answers to queries. Instead, it takes as input a numeric valued query q and a *threshold* T, and reports a single bit: whether or not the value of the query $q(S)$ is *above* or *below* the threshold on the sample. (We write $q(S)$ here since we are not restricting attention to statistical queries: for statistical queries, $q(S) = \mathbb{E}_S[q]$). Moreover, it will keep accepting queries while the answers returned so far have all been "below threshold". After the first "above threshold" query, it stops.

AboveThreshold$(T, q_1, q_2, \ldots)$:
  AllDone $\leftarrow$ **FALSE**
  while not **AllDone do**
    Accept the next query q_i
    Compute $a_i \leftarrow q_i(S)$
    if $a_i < T$ **then**
      Return $\perp$
    else
      Return $\top$
      AllDone $\leftarrow$ **TRUE**.
    end if
  end while

Theorem 6 *For any threshold T,* **AboveThreshold**(T) *is transcript compressible to $b(n,k)$ bits where $b(n,k) = \log(k+1)$.*

Proof The sequence of answers generated by **AboveThreshold** takes the form $\perp^i \top$ for $0 \leq i \leq k-1$ or $\perp^k$. There are $k+1$ such strings. ■

Lets start with a simple application of the AboveThreshold technique: the ability to repeatedly re-use a testbed dataset to validate machine learning models, and to keep a running score of the most accurate model tried so far. This functionality is called maintaining a "leaderboard" in machine learning competitions, e.g. like those hosted at Kaggle. Recall the basic setting: the distribution $\mathcal{D}$ is over *labeled examples* $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and a machine learning model is some mapping $f : \mathcal{X} \to \mathcal{Y}$. There is some *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, and we say that the loss of a classifier f on an example (x, y) is $\ell(f(x), y)$. A common loss function is "0/1" loss $\ell(y', y) = \mathbb{1}(y' \neq y)$. The loss of a classifier f_i on a distribution is just the statistical query $q_i = \ell(f_i(x), y)$.

We can define the "Ladder Mechanism" with step size $0 < \eta \leq 1$ as follows.

The semantics of the ladder algorithm are that the value output at round i is intended to represent the accuracy of the most accurate classifier submitted up to round i. Call such a query a "leader query" $\text{Best}_i(S) = \max_{j \leq i} q_j(S)$. Note that although $\text{Best}_i(S)$ is not a statistical query, it is a $1/n$ sensitive query, and so our transfer theorems apply. In this setting, (ϵ, δ) sample accuracy will mean that with probability $1 - \delta$, for all i, $|\text{Best}_i(S) - \text{BestAccuracy}_i| \leq \epsilon$.

Theorem 7 *Setting $\eta = \left(\frac{\log(k/\delta)}{n}\right)^{1/3}$, for any $\delta > 0$* **Ladder** *is (ϵ, δ)-accurate for any set of k leader queries, where:*

$$\epsilon = O\left(\left(\frac{\log(k/\delta)}{n}\right)^{1/3}\right).$$

Ladder$(\eta, f_1, f_2, \ldots)$:
  Output BestAccuracy$_0 \leftarrow 0$
  for $m = 1$ to $1/\eta$ **do**
    Start an instance of **AboveThreshold** with threshold $T_m = \text{BestAccuracy} + \eta$.
    while AboveThreshold has not halted **do**
      Accept the next classifier f_i.
      Feed **AboveThreshold** the query $q_i(S) = 1 - \ell(f_i(x), y)$.
      if AboveThreshold returns $\perp$ **then**
        Output BestAccuracy$_i \leftarrow$ BestAccuracy$_{i-1}$
      end if
    end while
    Output BestAccuracy$_i = \mathcal{O}_b^T(q_i)$ for $b = \log(1/\eta)$.
  end for

Proof We make two observations, and then apply our transcript-compressibility transfer theorem for $1/n$ sensitive queries. First, the Ladder mechanism is a composition of at most $1/eta$ copies of a post-processing of **AboveThreshold**. Thus, by our composition and post-processing theorems for transcript-compressibility, together with our bound for the transcript compressibility of **AboveThreshold**, we find that **Ladder** is transcript compressible to $b(n,k) = \frac{\log(k+1)}{\eta}$ bits. Second, the ladder mechanism is $(\max(\eta, 1/2^b), 0) = (\eta, 0)$-sample accurate. (There are two potential sources of sample error — the discretization of the threshold, which introduces at most η error, and the truncation of the b-truncated sample estimator, which introduces error at most $1/2^b$. Since $b = \log(1/\eta)$, this is also η.) Our transfer theorem for $1/n$ sensitive queries tells us that the Ladder mechanism is therefore (ϵ, δ)-accurate for:

$$\epsilon = \eta + \sqrt{\frac{(\frac{\log(k+1)}{\eta} + 1)\ln(2) + \ln(k/\delta)}{2n}} = O\left(\eta + \sqrt{\frac{\log(k/\delta)}{n\eta}}\right)$$

for any δ. We chose η to equalize the two summands in the last term — plugging in the value of η yields the theorem. Note that a more careful optimization of η can yield somewhat better bounds. ■

So... Not bad! At least in this one specialized setting (maintaining a leader board) we can obtain accuracy bounds that scale with only poly-logarithmically with the number of queries k asked, just as was possible in the non-adaptive setting. (Still, the bounds are not as good as we could obtain in the non-adaptive setting, where we can obtain accuracy $O\left(\sqrt{\frac{\log(k/\delta)}{n}}\right)$.) But what about if we want the ability to answer arbitrary statistical queries? We will think about this in the next lecture.

Bibliographic Information Composition theorems for description length bounds were given (following a different presentation) in [DFH+15]. The AboveThreshold technique is adapted from the differential privacy literature [DR+14] and was also applied in [DFH+15]. The Leaderboard application of these techniques is from [BH15]

References

[BH15] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pages 1006–1014, 2015.

[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.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.