Description Lenght Bounds III

Philipp Goebels

May 28, 2019

Goals

Construct transcript-compressible statistical estimators, that:

- answer arbitrary statistical queries accurately
- yield only polylogarithmic dependence on k in error bounds
- prevent analysts from overfitting

Remark: Above Threshold

AboveThreshold$(T, q_1, q_2, \ldots)$:

```
AllDone ← FALSE
while not AllDone do
    Accept the next query q_i
    Compute a_i ← q_i(S)
    if a_i < T then
        Return ⊥
    else
        Return ⊤
        AllDone ← TRUE.
    end if
end while
```

Let g_i be a guess to query q_i.
Given a fixed cutoff η and a sequence of tuples $(q_1, g_1), ..., (q_k, g_k)$ initialize an instance of AboveThreshold$(\eta, \hat{q}_1, ..., \hat{q}_k)$, with

$$\hat{q}_i = |q_i(\mathcal{D}) - g_i|.$$

If we get the answer $\perp$ we know g_i is sample accurate with accuracy η.

OneWrongGuess

OneWrongGuess$(\eta, (q_1, g_1), (q_2, g_2), \ldots)$

Start an instance of **AboveThreshold** with threshold η.
while AboveThreshold has not halted **do**
  Accept the next query (q_i, g_i).
  Feed **AboveThreshold** the query $\hat{q}_i(S) = |q_i(S) - g_i|$.
  if AboveThreshold returns $\perp$ **then**
    Return the answer $a_i = g_i$
  end if
end while
Return the answer $a_i = \mathcal{O}_b^T(q_i)$ for $b = \log(1/\eta)$.

Theorem 1

*For any threshold $0 < \eta \leq 1$, **OneWrongGuess** is $(\eta, 0)$-sample accurate and transcript compressible to $b(n, k)$ bits where $b(n, k) = \log(k + 1) + \log(1/\eta)$.*

Proof of transcript compressibility

Proof

Let f be a post processing function which replaces (q_i, g_i) with $\hat{q}_i(S) = |q_i(S) - g_i|$ and answers $a_i = \perp$ with $a_i = g_i$. Then **OneWrongGuess** is a composition of f(**AboveThreshold**) and $\mathcal{O}_b^T(q)$. We know that **AboveThreshold** is $\log(k+1)$-transcript compressible, by the postprocessing Theorem, so is f(**AboveThreshold**).

$\mathcal{O}_b^T(q)$ is transcript compressible to $\log(1/\eta)$ for $b = \log(1/\eta)$. By the composition theorem **OneWrongGuess** is transcript compressible to $b(n,k) = \log(k+1) + \log(1/\eta)$ bits.

Proof of accuracy

Every guess g_i which does not exceed the threshold η is by definition of **AboveThreshold** $(\eta, 0)$-accurate.
For the one query we cannot guess, we use the truncated estimator. We already know that $\mathcal{O}_b^T(q)$ is $(1/2^b, 0)$-accurate, which is $(\eta, 0)$-accurate for our choice of b. $\square$

GuessAndCheck

GuessAndCheck$(\eta, m, (q_1, g_1), (q_2, g_2), \ldots)$

TimesWrong $\leftarrow 0$
while TimesWrong $< m$ **do**
  Start an instance of **AboveThreshold** with threshold η.
  while AboveThreshold has not halted **do**
    Accept the next query (q_i, g_i).
    Feed **AboveThreshold** the query $\hat{q}_i(S) = |q_i(S) - g_i|$.
    if AboveThreshold returns $\perp$ **then**
      Return the answer $a_i = g_i$
    end if
  end while
  Return the answer $a_i = \mathcal{O}_b^T(q_i)$ for $b = \log(1/\eta)$.
  TimesWrong $\leftarrow$ TimesWrong $+ 1$
end while

Theorem 2

*For any η, m, **GuessAndCheck** is $(\eta, 0)$-sample accurate and transcript compressible to $b(n, k)$ bits where $b(n, k) = m(\log(k + 1) + \log(1/\eta))$.*

Proof

GuessAndCheck is just a composition of **OneWrongGuess** with itself, m times. The result follows from the composition theorem. $\square$

Theorem 3

*Fix a value of m and a value of δ. Setting $\eta = \sqrt{\frac{m}{n}}$, **GuessAndCheck**$(\eta, m)$ is (ϵ, δ)-accurate for any sequence of compound queries (q_i, g_i) until it halts, where q_i can be any $1/n$-sensitive query, for:*

$$\epsilon = O\left(\sqrt{\frac{m(\log(k) + \log(n/m)) + \log(k/\delta)}{n}}\right).$$

Proof

We have shown compressibility to $b(n,k) = m(\log(k+1) + \log(1/\eta))$ bits, and $(\eta, 0)$-sample accuracy.

(ϵ, δ)-accuracy for

$$\epsilon = \eta + \sqrt{\frac{(m(\log(k+1) + \log(1/\eta) + 1)\log(2) + \log(k/\delta)}{2n}}$$

follows from the transfer theorem for transcript compressibility. $\square$

Lemma 4

For any $\epsilon > 0$*, any* k *statistical queries* $\Phi_1, ..., \Phi_k$ *and for any dataset* $S \in \mathcal{X}^n$*, there is an* $S' \in \mathcal{X}^{n'}$ *with* $n' = \frac{\log(4k)}{2\epsilon^2}$ *such that:*

$$\max_i |\mathbb{E}_S[\Phi_i] - \mathbb{E}_{S'}[\Phi_i]| \leq \epsilon$$

Remark: Chernoff Bound

Theorem

Fix any distribution $\mathcal{D}$, and any statistical query Φ. Let $S \sim \mathcal{D}^n$ consist of a set of n points sampled i.i.d from $\mathcal{D}$, with probability $1-\delta$ over the sample:

$$|\mathbb{E}_S[\Phi] - \mathbb{E}_{\mathcal{D}}[\Phi]| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

Proof

Generate S' by subsampling m points from S with replacement.
Under this sampling distirbution, $\mathbb{E}[\Phi_i] = \mathbb{E}_S[\Phi_i]$ for each i.
Apply a Chernoff bound with $\delta = 1/2$ to follow:

$$\max_i |\mathbb{E}_S[\Phi_i] - \mathbb{E}'_S[\Phi_i]| \leq \sqrt{\frac{log(4k)}{2m}} \leq \epsilon.$$

$\square$

MedianOracle

MedianOracle$(q_1, \ldots, q_k)$

Initialize an instance of **GuessAndCheck**(η, m) with $m = \sqrt{\frac{n \log |\mathcal{X}| \ln(4k)}{2}}$ and $\eta = \sqrt{\frac{m}{n}}$.

Initialize a version space $\mathcal{S}_0 = \mathcal{X}^{n'}$ where $n' = \frac{\ln(4k)}{2\eta^2}$

for $i = 1$ to k **do**

Given query q_i, construct a guess $g_i = \text{median}\left(\{q_i(S') : S' \in \mathcal{S}_{i-1}\}\right)$

Feed the query (q_i, g_i) to **GuessAndCheck** and receive answer a_i.

if $\hat{a}_i = g_i$ **then**

$\mathcal{S}_i \leftarrow \mathcal{S}_{i-1}$

else

$\mathcal{S}_i \leftarrow \mathcal{S}_{i-1} \setminus \{S' \in \mathcal{S}_{i-1} : |q_i(S') - a_i| > \eta\}$

end if

Return answer a_i.

end for

Theorem 5
For any $\delta > 0$, ***MedianOracle*** *is (ϵ, δ)-accurate for any sequence of k statistical queries where:*

$$\epsilon = O\left(\frac{\log(|\mathcal{X}| \log(k))^{1/4}\sqrt{\log(k) + \log(n)}}{n^{1/4}}\right)$$

Proof

MedianOracle is a postprocessing of **GuessAndCheck**. So (ϵ, δ)-accuracy *for the queries asked before the algorithm halts* follows from the accuracy of **GuessAndCheck**.

We need to show that **MedianOracle** will answer all k queries and never halt, this is equivalent to showing that $|q_i(S) - g_i| \leq \eta$ for all but m rounds.

tracking $|S_i|$

- by construction $|S_0| = |\mathcal{X}|^{n'}$
- in every round i we make a mistake, $|S_i| \leq |S_{i-1}|/2$, because on these round $|g_i - q_i(S)| > \eta$ and all sets S' such that $|q_i(S') - a_i| > \eta$ are removed from S_i.
- by definition $g_i = \text{median}(\{q_i(S') : S' \in S_{i-1}\})$, so at least half of the S' in S_i are removed.
- by *Lemma* 4 we know that there is at least one S' such that $|q_i(S') - q_i(S)| \leq \eta$. Hence $|S_i| \geq 1$ for every i.
- with that the number of mistaken guesses can be at most $\log(|S_0|) = n' \log(|\mathcal{X}|) = m$

$\square$

ReusableHoldout

ReusableHoldout$(m, q_1, \ldots, q_k)$

Randomly split the dataset S into two equal parts: a training set S_T and a holdout set S_H, each of size $n/2$.

Initialize an instance of **GuessAndCheck**(η, m) on S_H with $\eta = \sqrt{\frac{2m}{n}}$.

for $i = 1$ to k **do**

Given query q_i, construct a guess $g_i = q_i(S_T)$

Feed the query (q_i, g_i) to **GuessAndCheck** and receive answer a_i.

Return answer a_i.

end for

Theorem 6

Fix a value of m and a value of $\delta > 0$. ***ReusableHoldout*** *is (ϵ, δ)-accurate for any sequence of $1/n$ sensitive queries q_i until it halts, for:*

$$\epsilon = O\left(\sqrt{\frac{m(\log(k) + \log(n/m)) + \log(k/\delta)}{n}}\right).$$

Previous definitions and theorems

truncated estimator

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

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. ■

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:

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

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

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