ECS 189A Sublinear Algorithms for Big Data **Fall 2024**

Lecture 13: Instance Optimal Testing, Tolerant Testing, and Relations with Estimating Parameters

Lecturer: Jasper Lee *Scribe: Yuhao Wang*

Instance Optimality

In the context of property testing and theoretical computer science, we often analyze algorithms under the assumption of worst-case optimality. For instance, consider the problem of identity testing against a distribution q over $[n]$ compared to being ϵ-far from q. Typically, we derive upper and lower bounds for the number of samples needed by an algorithm to succeed in these scenarios.

Upper Bound: For any distribution q over $[n]$, identity testing requires only $c_1 \cdot f(n, \epsilon)$ samples, where $f(n, \epsilon) = \frac{\sqrt{n}}{\epsilon^2}$.

Lower Bound: There exists a distribution q over $[n]$ such that identity testing requires at least $c_2 \cdot f(n, \epsilon)$ samples, where $f(n, \epsilon) = \frac{\sqrt{n}}{\epsilon^2}$.

The above analysis focuses on the "worst-case" scenario, where we consider the performance of the tester under the most challenging conditions. However, it is natural to ask if we can develop an algorithm that performs better for distributions that are "easier" to test than the worst-case scenario. This leads to the concept of instance optimality, where we replace the complexity measure $f(n, \epsilon)$ with a more refined function $f'(q, \epsilon)$ that is tailored to the specific distribution being tested. This type of optimality provides a more nuanced understanding of algorithm performance, as it considers the particular characteristics of the distribution instead of a blanket worst-case assumption.

Definition 13.1

Given a distribution q, define q_ϵ as the vector remaining after removing:

- The element with the highest probability.
- The elements with the smallest mass, until ϵ mass is removed.

Theorem 13.2 (Valiant & Valiant 2017)

There exists a tester $\mathcal{A}$, universal constants c_1, c_2 such that for any $\epsilon > 0$, known distribution q over $[n]$, $\mathcal{A}$ tests identity to q (versus ϵ-far from q) with probability $\geq \frac{2}{3}$ when run on

$$c_1 \cdot \max\left(\frac{1}{\epsilon}, \frac{\left\|q^{\text{-max}}_{-\epsilon/16}\right\|^{2/3}}{\epsilon^2}\right)$$

No tester can test identity to q (versus ϵ-far from q) with probability $\geq \frac{2}{3}$ when run on fewer than

$$c_2 \cdot \max\left(\frac{1}{\epsilon}, \frac{\|q^{\text{-max}}_{-\epsilon}\|_{2/3}}{\epsilon^2}\right) \text{ samples.}$$

- Standard statistics: iid samples from distribution D.
- Typically: makes assumptions on D.

What if:

1. Assumptions hold only approximately?
   E.g., assume D is Gaussian, but get samples from D' only close to Gaussian.
2. Data corrupted adversarially?
   The possibility of data poisoning attack

Q: How to robustly perform statistical tasks despite (small) deviation from iid or distributional assumptions?

Consider a set $\mathcal{D}$ of distributions, which encodes certain assumptions on $\mathcal{D}$ (e.g., $\mathcal{D}$ could be the set of all Gaussian distributions).

1. Generate m samples from an unknown distribution $D \in \mathcal{D}$.
2. An adversary then modifies an ϵ-fraction of these samples arbitrarily.

The second step is somewhat ambiguous, and there are numerous ways to formalize this model. For example:

- Does the adversary have the ability to observe the samples before making edits? (i.e., Adaptive versus Data-oblivious)
- Regarding the changes made by the adversary, are they restricted to adding new data? Removing data? Or possibly both?

To build some intuition, let us consider a simple case. Assume $\mathcal{D}$ is the set of one-dimensional Gaussian distributions. How can we estimate its mean in a robust way?

Answer: Use the median.

Fact 13.3 (Coupling/Simulation View of Total Variation (TV) Distance)

The TV distance is defined as:

where X has marginal distribution p and Y has marginal distribution q.

In simple terms, TV distance represents the minimum probability that the random variables X and Y differ. This perspective helps to understand how to simulate and match different distributions.

TV Distance Corruption Model (Model 13.4)

In this model, we consider a scenario where part of the data may be corrupted:

1. **Ground Truth Distribution** D:
   - Statistician S specifies the number of samples m and a corruption parameter η, where $\eta \leq \frac{1}{2}$.
2. **Adversary Chooses Distribution**:
   - Adversary (Adv) chooses a distribution D' such that $d_{\text{TV}}(D', D) \leq \eta$.
3. **Data Generation**:
   - Nature draws m i.i.d. samples from D'.
4. **Data Reception**:
   - S gets the m samples

Observation 13.5

Even as the sample size $m \to \infty$, and even if the distribution is known to be $W(\mu, 1)$, no estimator can get arbitrarily close to zero error.

This observation indicates that in the presence of adversarial corruption, the error lower bound does not vanish, even with an infinite number of samples.

Strong Contamination Model (Model 13.6)

In the strong contamination model, we consider direct replacement of data samples by an adversary:

1. **Statistician** S specifies the number of samples m and a corruption budget η.
2. **Data Generation**:
   Nature generates m i.i.d. samples from distribution D.
3. **Adversary Intervention**:
   The adversary inspects D and arbitrarily replaces up to ηm samples.
4. **Data Reception**:
   S receives the η-corrupted, unlabeled samples.

This model is used to analyze how adversarial attacks affect the overall properties of the data.

Fact 13.7

For m samples from a normal distribution $N(\mu, \sigma^2)$, with probability at least $1 - \delta$, the sample median $\hat{\mu}$ achieves:

$$|\hat{\mu} - \mu| \leq O\left(\sigma\sqrt{\frac{\log f}{m}}\right) + o(\sigma\eta)$$

This fact indicates that even in the presence of adversarial corruption, as long as there are enough samples, the sample median can still provide a certain level of accuracy.