# Lecture 23: Closeness Testing for Product Distributions, Introduction to Robust Statistics

Daniel Kane
Scribe: Ken Hoover

May 24, 2017

## Closeness Testing for Product Distributions

Last lecture, we proved that for two binary product distributions $\mathcal{P}$ and $\mathcal{Q}$ over $\{0,1\}^n$, testing $\mathcal{P} = \mathcal{Q}$ versus $d_{TV}(\mathcal{P}, \mathcal{Q}) > \epsilon$ requires $\Omega(\max\{n^{3/4}/\epsilon, \sqrt{n}/\epsilon^2\})$ samples. Today, we give an algorithm with sample complexity matching those bounds.

Suppose we are given binary product distributions $\mathcal{P}$ and $\mathcal{Q}$. First, we can assume that $p_i, q_i \leq 2/3$ for each coordinate $i$. This is because we can take $\log(n)$ samples from both, and flip any coordinates where 1s occur much more frequently than 0s. Now consider the case of $p_i$ and $q_i$ being very small ($\approx 0$). We can use the fact that if $d_{TV}(\mathcal{P}, \mathcal{Q}) > \epsilon$, then since

$$||\mathbf{p} - \mathbf{q}||_1 \geq d_{TV}(\mathcal{P}, \mathcal{Q}) \ ,$$

we have that

$$||\mathbf{p} - \mathbf{q}||_2^2 > \epsilon^2/n \ .$$

This lends itself to the $L^2$ closeness testing approach introduced at the beginning of the course.

Let $X_i \sim \text{Poi}(p_i m)$ and $Y_i \sim \text{Poi}(q_i m)$ be samples from $2n$ independent Poisson distributions. Define

$$Z = \sum_i [(X_i - Y_i)^2 - X_i - Y_i] \ .$$

Then

$$\mathbb{E}[Z] = m^2 ||\mathbf{p} - \mathbf{q}||_2^2 \ ,$$

and

$$\begin{aligned} \text{Var}(Z) &= O\left( m^3 \sum_i [(p_i - q_i)^2 (p_i + q_i)] + m^2 \sum_i (p_i + q_i)^2 \right) \\ &= O(m^2 \sqrt{\mathbb{E}[Z]} \cdot ||\mathbf{p} + \mathbf{q}||_2^2 + m^2 ||\mathbf{p} + \mathbf{q}||_2^2) \ . \end{aligned}$$

If we have $m >> n||\mathbf{p} + \mathbf{q}||_2/\epsilon^2$, then this algorithm would work as desired. However, $||\mathbf{p} + \mathbf{q}||_2$ could be as large as $\sqrt{n}$, making the algorithm non-optimal. What we then want to do is as follows; for coordinates where $|p_i + q_i|$ is small, we use the above algorithm. For coordinates where $|p_i + q_i|$

is large, we will take a different approach, shown later. We will call coordinates of the former type light coordinates, and coordinates of the latter type heavy.

Since we don't know which coordinates are light and which are heavy, we take $\text{Poi}(k)$ samples from $\mathcal{P}$ and $\mathcal{Q}$, for some $k$, and define light coordinates as those where the value of each sample in that coordinate is 0. Note that

$$d_{TV}(\mathcal{P},\mathcal{Q}) << \sum_{\text{light coords } i} |p_i - q_i| + \sqrt{\sum_{\text{heavy coords } i} \frac{(p_i-q_i)^2}{p_i+q_i}} \ .$$

If $||\mathbf{p}-\mathbf{q}||_{1,\text{light}} >> \epsilon$, then only $n\sqrt{n/k^2}/\epsilon^2$ samples are required. If $m = k$, then $m > n^{3/2}/m\epsilon^2$, or $m > n^{3/4}/\epsilon$ samples are needed.

Now suppose instead that

$$\sum_{\text{heavy coords } i} \frac{(p_i-q_i)^2}{p_i+q_i} >> \epsilon^2 \ .$$

Let $a_i$ the total number of samples (from the $2\cdot\text{Poi}(k)$ taken from $\mathcal{P}$ and $\mathcal{Q}$) where the $i$th coordinate is 1. We then have that $a_i \approx k(p_i+q_i)$; stated differently, we have that $a_i \sim \text{Poi}(k(p_i+q_i))$. Let $H$ be the set of heavy coordinates. Define

$$Z' = \sum_{i\in H} \frac{(X_i-Y_i)^2 - X_i - Y_i}{a_i/k} \ .$$

Then

$$\begin{aligned}\mathbb{E}[Z'] &= \sum_{i\in H} \frac{(p_i-q_i)^2}{a_i/k} \\ &\geq \sum_{i|p_i+q_i\geq 1/k} \frac{(p_i-q_i)^2}{p_i+q_i} \ .\end{aligned}$$

Since the heavy coordinates are probably approximately those coordinates for which $p_i+q_i \geq 1/k$, this inequality holds. We also have that

$$\text{Var}(Z') = O\left(m^3 \sum_{i\in H} \frac{(p_i-q_i)^2(p_i+q_i)}{(a_i/k^2)} + m^2 \sum_{i\in[n]} \frac{(p_i+q_i)^2}{(a_i/k)^2}\right) \ .$$

The first term is exactly $m\mathbb{E}_{X_i,Y_i}[Z']$, while the second is bounded above by $m^2 n$. Also note that the latter term dominates, thus, to get $\mathbb{E}[Z'] >> \sqrt{\text{Var}(Z')}$, it is roughly sufficient to have

$$\begin{aligned}\mathbb{E}[Z'] &>> \sqrt{m^2 n} \\ m^2\epsilon^2 &>> \sqrt{m^2 n} \\ m &>> \sqrt{n}/\epsilon^2 \ .\end{aligned}$$

So, by using this tester on the heavy coordinates, and using the previous tester $Z$ on the light coordinates, we have a closeness testing algorithm using $\max\{\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon\}$ samples, which is optimal.

# Robust Statistics

In all of our algorithms so far, we have assumed that we are perfectly sampling from the distributions in question. In practical applications, however, there's often some degree of error in our sample generator. Robust statistics deals with handling these errors.

The first point we make is that we assume we have some class $\mathcal{C}$ of hypothesis distributions, which will be the ones our algorithms attempt to, e.g., learn or test closeness between. Our model will be as follows; an $\epsilon$ fraction of samples are "bad," where bad is defined in one of several ways:

- Additive error (the Huber model); a $(1-\epsilon)$ fraction of our samples are taken from the intended distribution in $\mathcal{C}$, and an $\epsilon$ fraction are taken from some other arbitrary distribution.
- General $L^1$ error; for an intended distribution $\mathcal{D} \in \mathcal{C}$, our samples are taken from some $\mathcal{D}' \in \mathcal{C}$ such that $d_{TV}(\mathcal{D}, \mathcal{D}') < \epsilon$.
- Strong adversarial error; given $N$ independent samples from $\mathcal{D}$, an adversary is allowed to change an arbitrary $\epsilon N$ selection of them arbitrarily.

These models are arranged in terms of increasing difficulty, i.e. an algorithm which works in the presence of strong adversarial error will still work in the presence of general $L^1$ or additive error, but the converse is not necessarily true.

For our first example of a problem in robust statistics, we shall look at the problem of learning the mean of a Gaussian with the identity covariance matrix over $\mathbb{R}^n$.

## Learning a Gaussian Mean

Our class $\mathcal{C}$ here will be the set $\{\mathcal{N}(\mu, \mathsf{I}) | \mu \in \mathbb{R}^n\}$. The goal is as follows: given some $\mathcal{D} \in \mathcal{D}$, learn $\mathcal{D}$ to a total variation distance of at most $\epsilon$. Note that it is impossible, in general, to learn $\mathcal{D}$ to distance better than $\epsilon$ under general $L^1$ errors. An information theoretic argument for that is as follows. Let $\mathcal{D}'$ be another Gaussian in $\mathcal{C}$ such that $d_{TV}(\mathcal{D}, \mathcal{D}') = 2\epsilon$. Consider the distribution $\frac{1}{2}\mathcal{D} + \frac{1}{2}\mathcal{D}'$. The best thing any algorithm can do is return $\mathcal{N}(\frac{\mu_{\mathcal{D}} + \mu_{\mathcal{D}'}}{2}, \mathsf{I})$, as otherwise we could arbitrarily relabel $\mathcal{D}$ and $\mathcal{D}'$ such that the algorithm returned a distribution with total variation distance larger than $\epsilon$.

With this bound on the performance, we give a first attempt at an algorithm. Note that this algorithm only works if we have restricted the set of possible mean vectors to some bounded region in $\mathbb{R}^n$; this is usually a reasonable assumption. In this case, we find an $\epsilon$-cover of that bounded region. Note that the mean vector of the true distribution has distance at most $2\epsilon$ between it and the nearest point in the cover. Also note that the size of this cover is $\text{poly}(n/\epsilon)^n$. If we then take a random sample from the distribution, we note that it probably lies within a $\sqrt{n}$ radius ball around the true mean. Using this, we can conclude that the distribution is learnable to $O(\epsilon)$ error with $n \log(n/\epsilon)/\epsilon^2$ samples. However, the algorithm achieving this takes exponential time.

A second approach would be to take the sample mean, as in the non-noisy case. Here, however, a single corrupted sample can move the sample mean an arbitrary distance away from the true mean. A different approach would be to take the coordinate-wise median. In the one-dimensional case, this works; the median of our samples with errors, with high probability, is within an $O(\epsilon)$ distance from the true mean. In the $n$-dimensional case, however, this error grows to $O(\epsilon\sqrt{n})$.

We can fix this by looking at projections of our sample points, minus the candidate mean. If our mean is correct, then projecting in any direction should result in a roughly equal number of points

on either side of the plane normal to the projection vector. This gives rise to the Tukey median of a set of points $X$,

$$\widehat{\mu}_T = \arg\min_{\widehat{\mu}} \sup_{\mathbf{v}:||\mathbf{v}||_2=1} (|\{\mathbf{x} \in X : \mathbf{v} \cdot (\mathbf{x} - \widehat{\mu}) > 0\}| - |\{\mathbf{x} \in X : \mathbf{v} \cdot (\mathbf{x} - \widehat{\mu}) < 0\}|) \quad .$$

We can prove that, with high probability, $||\mu - \widehat{\mu}_T||_2 = O(\epsilon)$ as follows; suppose that $||\mu - \widehat{\mu}_T||_2 >> \epsilon$. Then there must be some $\mathbf{v} \in \mathbb{R}^n$ such that $||\mathbf{v}||_2 = 1$ and $|\mathbf{v} \cdot (\widehat{\mu}_T - \mu)| >> \epsilon$. This then means that

$$|\{\mathbf{x} \in X : \mathbf{v} \cdot (\mathbf{x} - \mu) > 0\}| - |\{\mathbf{x} \in X : \mathbf{v} \cdot (\mathbf{x} - \mu) < 0\}| >> \epsilon \ ,$$

which occurs with low probability.

There is a problem with this approach; for arbitrary point-sets $X$, computing $\widehat{\mu}_T$ is NP-hard. Thus, even this simple problem of learning a Gaussian mean was considered difficult. Some recent developments have occurred, however, which give algorithms for solving the problem with improved performance.