6.5240 Sublinear Time Algorithms	November 7, 2022

Lecture 16

Lecturer: Ronitt Rubinfeld *Scribe: Anson Hu*

Today: hypothesis testing, the cover method.
Previously covered: given samples of a distribution p of domain size n, it is possible to check if

- $p = q$ for known q or ϵ-far in $O(\sqrt{n})$ samples
- p is ϵ-close to known q in L_1 distance or ϵ-far in L_2 distance in $O(n/\log n)$ samples
- $p = q$ for q given via samples or ϵ-far in L_2 in $O(n^{2/3})$ samples
- p is ϵ-close to q given via samples in L_1 distance or ϵ-far in L_2 distance in $O(n/\log n)$ samples

1 Hypothesis testing

Tool: Given a collection of distributions H, of which you have complete knowledge, and samples of a distribution p such that there exists q in H for which $\text{dist}(p, q)$ is small, the goal is to output h in H such that $\text{dist}(p, h)$ is small. Our metric is the number of samples in terms of H and the domain size.

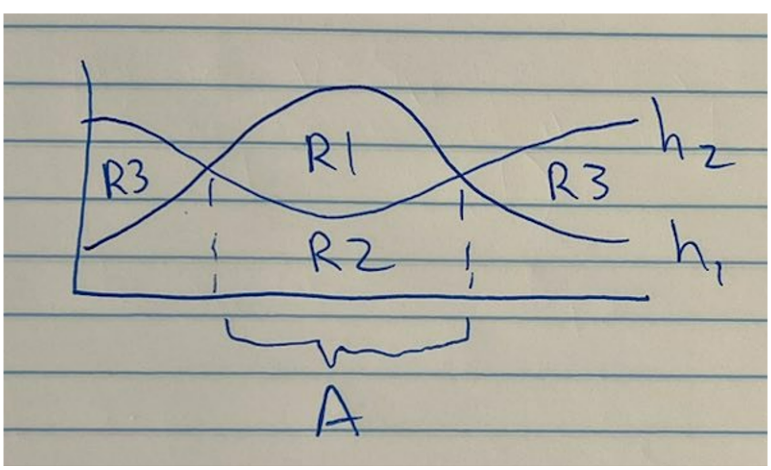
Start with the simple case with $|H| = 2$. h_1, h_2 are given explicitly and p is taken via samples. The goal is to output whichever h_i is closer to p. If $||h_1 - h_2||_1 \leq \epsilon$, either one can be output.

Theorem 1 *: Given p via samples, h_1, h_2 explicitly, an ϵ' parameter for accuracy, and a δ' confidence parameter, there is an algorithm "Choose" which takes $O(\log(\frac{1}{\delta'})/\epsilon'^2)$ samples and outputs one of $\{h_1, h_2\}$ which satisfies that if one of $\{h_1, h_2\}$ has $||h_i - p||_1 \leq \epsilon'$, then with probability $\geq 1 - \epsilon'$ the output is h_j such that $||h_j - p||_1 \leq 12\epsilon'$.*

We will use $\epsilon' \approx \epsilon/12$. ($\delta'$ is used because down the line it will be needed to pass all tests in a union bound.)

1.1 Algorithm "Choose"

First, define $A = \{x | h_1(x) > h_2(x)\}$. Think about a simplified example where h_1 and h_2 only cross twice:



Call these regions R_1, R_2, R_3. Let $a_1 = h_1(A)$ and $a_2 = h_2(A)$. We can see that $a_1 = R_1 + R_2$, $a_2 = R_2$, and $R_1 = R_3 = a_1 - a_2$. Notice that $R_1 = R_3$ because the sum of probabilities is equal to 1 for h_1 and h_2, and therefore the "additional" probability R_1 gained by h_1 over A must be gained by h_2 over the remainder of the domain.

The L_1 distance between h_1 and h_2 is $R_1 + R_3 = 2R_1 = 2(a_1 - a_2)$.

The algorithm "Choose" does the following:

1. if $a_1 - a_2 \leq 5\epsilon'$, declare a tie and return h_1. (No samples are taken.)
2. draw $m = \frac{2\log(1/\delta')}{\epsilon'^2}$ samples $S_1 \dots S_m$ from p.
3. let $\alpha = \frac{1}{m}|i|S_i \in A|$. (In other words α is the fraction of samples in A.)
4. if $\alpha > a_1 - (3/2)\epsilon'$, return h_1, else if $\alpha < a_2 + (3/2)\epsilon'$ return h_2, else there is a tie and return h_1.

We need that $a_1 - (3/2)\epsilon' > a_2 + (3/2)\epsilon'$ to make these regions exclusive, which means that $a_1 > a_2 + 3\epsilon'$. This is enforced by step 1.

Behavior

If h_1 or h_2 is ϵ' close to p, then if there is a tie in step 1, the L_1 distance between the two is at most $10\epsilon'$ and then $||p - H_i||_1 \leq 11\epsilon'$, so we are good.

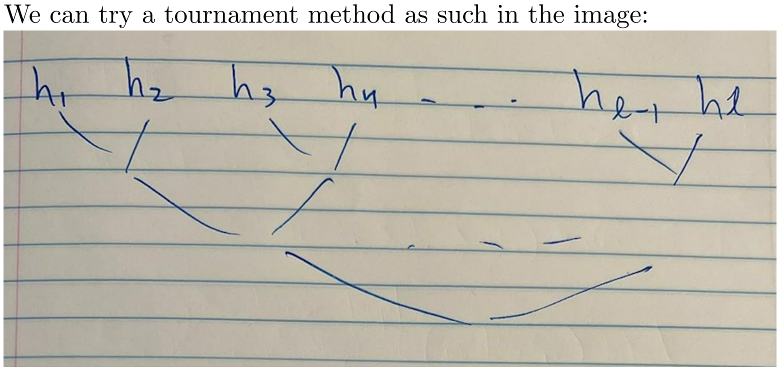
(Side note: total variation distance is used in some papers; it just means half of L_1 distance.)

Otherwise, we reach step 2, and L_1 distance between the two is $> 10\epsilon'$. $E[\alpha] = Pr_{x \in p}[x \in A] = p(A)$. By Chernoff bound on the number of samples, with high probability $|\alpha - E[\alpha]| < \epsilon'/2$. h_1 assigns a_1 weight to A, and h_2 assigns a_2 weight to A. If p is ϵ'-close to h_1, it assigns $\geq a_1 - \epsilon'$ weight to A, which implies $\alpha > a_1 - \epsilon' - \epsilon'/2 = a_1 - (3/2)\epsilon'$. Therefore h_1 is output with high probability. The same argument holds for h_2 in the other direction. We have demonstrated that the algorithm has correct behavior.

1.2 A first attempt at arbitrary-size $|H|$

We will try to run this as a subroutine where we reuse samples when plugging into "Choose". The plan is to use union bound since the runs are dependent. The probability of a run being bad is at most δ', therefore we need $k\delta'$ to be small, where k is the number of times we run it. Therefore we need $\delta' \approx 1/k$.

We can try a tournament method as such in the image:



However, this is not good, since at each level we gain a factor of 11 of error (for example, if $p = h_1$ but h_2 passes, the distance of h_2 could be up to about $11\epsilon'$. A similar argument holds as we advance down the tournament tree, so the final winner of the tournament could have as far as $||p - h_{winner}||_1 \leq 11^{\log l}\epsilon'$.

Now, we instead try to test all pairs. Then we can see that the distribution closest to p never loses, and we want to show that things 11 apart will lose to the winner. We will modify the choose spec: if $h_i > 12\epsilon'$-far from p, then it will likely lose, and if $h_i > 10\epsilon'$-far, then it is likely to tie or lose.

2 The cover method

Definition 2 *C is an ϵ-cover of D, where both C and D are collections of distributions and C is smaller, if $\forall p \in D$, $\exists q \in C$ such that $||p-q||_1 \leq \epsilon$.*

Theorem 3 *Given a cover C of D, there exists an algorithm, given $p \in D$, which takes $O(\frac{1}{\epsilon^2}\log|C|)$ samples of p and outputs $h \in C$ such that $||h-p_1||_1 \leq 12\epsilon$ with probability $\geq 9/10$.*

Proof Run "Choose" on p with every pair $(q_1, q_2) \in C$, the best q_{opt} ties or wins all matches. If $q' \geq 12\epsilon$-far from p, then it is at least 11ϵ-far from q_{opt}. ■

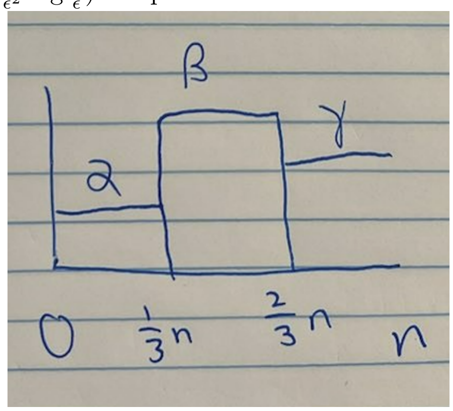
2.1 Examples

Finding the bias of a coin. The coin has domain $\{0,1\}$ and $D = [0,1]$. Use $C = \{0, 1/k, 2/k \ldots k/k\}$.

$\forall p$, use $\tilde{p} \leftarrow$ closest i/k, so $||p-\tilde{p}||_1 < 1/k$.

$k = \Theta(1/\epsilon) \rightarrow ||p-\tilde{p}||_1 \leq \epsilon, |C| = k+1 = \Theta(1/\epsilon)$, and therefore the number of samples taken by the cover method is $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$.

3-bucket distributions. $|C| = \Theta(1/\epsilon^2)$, since we have to pick pairs of (α, β), and the algorithm takes $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$ samples.



Monotone distributions. By Birge's theorem, $C = \{i_1/k \ldots i_{\log n/\epsilon}/k\}$, where the is are in $\{0 \ldots k\}$. $|C| = \Theta(\frac{1}{e^{\log n/\epsilon}})$, so the number of samples is $O(\frac{1}{\epsilon^3} \cdot \log n \cdot \log\frac{1}{\epsilon})$.

Poisson binomial distribution. $X = \sum x_i$, where x_i is an indicator variable for a coin with bias p_i. The p_i are independent but not identically distributed. For example, where $p_1 = 1/2, p_2 = 1, p \ldots = 0$, $\Pr[x=0] = 0, \Pr[x=1] = 1/2$.