CS 395T: Sublinear Algorithms	Fall 2014
Lecture 15 — Oct 16, 2014	
Prof. Eric Price	*Scribe: Chi-Kit Lam*

1 Overview

In this lecture we will talk about adaptive sparse recovery.

2 Adaptivity in group testing

In sparse recovery, we have

$$y = Ax = \begin{pmatrix} \langle v_1, x \rangle \\ \langle v_2, x \rangle \\ \vdots \\ \langle v_m, x \rangle \end{pmatrix}.$$

In our previous non-adaptive setting, v_i's were chosen independently. Intuitively, it seems we may be able to do better if they are not independent. So, in the adaptive setting we talk about today, v_i is chosen dependent on $\langle v_1, x \rangle, \langle v_2, x \rangle, \ldots, \langle v_{i-1}, x \rangle$. The idea is, for some architecture like the single-pixel camera, linear measurements are being taken sequentially. So, you can choose what measurement to take based on what you have found previously.

2.1 Non-adaptive group testing

Consider the example of blood testing, where you have n people and some small fraction of them are carrying disease. You want to know which of them has the disease. This problem was first studied in World War II to test syphilis in the army. Group testing is also used by the Orthodox Jewish community for pre-dating genetic testing.

The trivial solution is to test every single person. Alternatively, to make it faster, we can mix together blood samples. For example, you can mix together the blood samples of 10 people, and test whether any of these 10 people got the disease. If the chance of the disease is $1/1000$, you can make your poll size ≈ 1000. Then the chance of this pool has a positive or negative is $\approx 1/2$. So, you can get ≈ 1 bit of information per test (either yes or no). And if you do $\approx k \log n$ such trial, using a union bound, with high probability, you can find out which person has the disease.

Let $x \in \{0,1\}^n$ be the indicator vector of the set with disease. Let $v \in \{0,1\}^n$ be the indicator vector of the set that you mix together. Then, each test observes whether $\langle v, x \rangle \neq 0$. If x is k-sparse, we can choose $m = O(k \log n)$ random v's, where each v has n/k ones and each person is in $d = O(\log n)$ tests. We can thinking of the design of the group testing as a bipartite graph similiar to Figure 1.

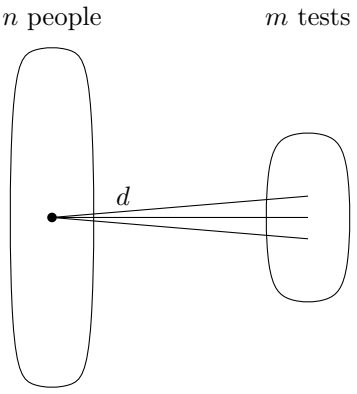

Figure 1: Bipartite graph corresponding to the design of group testing. The n left vertices correspond to the n people and the m right vertices correspond to the m tests.

For each person, suppose their tests are $t_1, \ldots, t_d$. If that person has disease, then all tests t_i are positive. Otherwise,

$$\begin{aligned}\mathbb{P}[\text{test } t_i \text{ is positive}] =& \mathbb{P}[\text{some sample } x_j \text{ in test } t_i \text{ is non-zero}] \\ =& 1 - \mathbb{P}[\text{all samples } x_j \text{ in test } t_i \text{ is zero}] \\ \approx& 1 - (1 - k/n)^{n/k} \approx 1 - 1/e.\end{aligned}$$

So,

$$\mathbb{P}[\text{all tests } t_i \approx (1 - 1/e)^d \leq 1/n^{10}$$

for some $d = O(\log n)$.

Therefore, non-adaptive random test uses $O(k \log n)$ measurements, and for all such x, with high probability it returns the correct answer.

2.2 Adaptive group testing

In most cases, you can actually tell whether you are certain about the result. Consider the example in Figure 2. We know for certain that people 2, 4 and 6 do not carry disease, but people 1 and 6 do. The result for person 3 is ambiguous. In general, we can first cross out all people who are in at least one negative test. For the remainng people, if they are the only remaining participant of some positive test, then we are certain that they carry disease. Otherwise, they are ambiguous.

We have shown that for all such x, the number of ambiguity is 0 with probability $1 - n^{-\Omega(1)}$. In fact, we can show that for all such x, the number of ambiguity is less than t with probability $1 - n^{-\Omega(t)}$. Take $t = O(k)$ and apply union bound over $\binom{n}{k}$ possible values of x. We can show that with high probability, for all such x, the number of ambiguity is less than $O(k)$. This uniform guarantee is analogous to the result that a Gaussian matrix with high probability satisfies RIP, but it does not give us an explicit construction.

So, after one round, with high probability, for all such x, there are only $O(k)$ ambiguity. This means that in two rounds, it can eliminate all ambiguity uniformly with $O(k \log n)$ tests. However, with one round, uniformity requires $\Omega(k^2 \log_k n)$ tests, and $O(k^2 \log n)$ tests suffice. Therefore, adaptivity helps group testing.

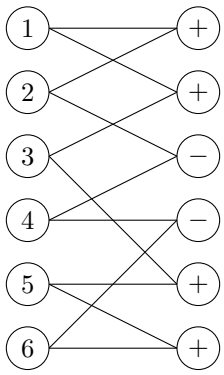

Figure 2: Example of ambiguous group testing result. The left vertices correspond to 6 people and the right vertices correspond to 6 tests labeled by the test results.

3 Adaptivity in sparse recovery

In group testing, we get at most 1 bit of information per measurement, so we need $O(k \log \frac{n}{k})$ measurements to distinguish all $\binom{n}{k}$ possible inputs. The difference of sparse recovery and group testing is that we do not just learn whether $\langle v, x \rangle = 0$ or not, but also the value of $\langle v, x \rangle$. In the case of real measurement, we can in principle get a lot of information per measurement. However, if we need to tolerate noise, we can show that we can only get $O(1)$ effective bit of information per measurement. So, the question is whether adaptivity can help in sparse recovery.

The result is that adaptivity provides no improvement in the uniform setting. In particular, for ℓ_2/ℓ_2 sparse recovery, $\Omega(\frac{k}{\epsilon} \log \frac{n}{k})$ measurements are necessary [FR13].

In contrast, adaptivity can help in the non-uniform setting. It has been shown that adaptive ℓ_2/ℓ_2 sparse recovery in the non-uniform setting is possible with $O(\frac{k}{\epsilon} + k \log \log \frac{n}{k})$ measurements [IPW11] and $\Omega(\frac{k}{\epsilon} + \log \log n)$ measurements are required [ACD13, PW13].

3.1 Lower bound of non-adaptive sparse recovery

Consider $k = 1$ and $\epsilon = 1$. We will prove a lower bound of $m = \Omega(\log n)$ measurements for non-adaptive sparse recovery in the non-uniform setting.

Suppose $A \sim \mathcal{P}_A$ such that for all x, it works with high probability. Let $\mathcal{P}_x$ be a distribution of x. Then,

$$\begin{aligned} & \forall x \mathop{\mathbb{P}}_{A \sim \mathcal{P}_A}[\text{correct}] > 1 - n^{-10} \\ \implies & \mathop{\mathbb{P}}_{x \sim \mathcal{P}_x, A \sim \mathcal{P}_A}[\text{correct}] > 1 - n^{-10} \\ \implies & \exists A \mathop{\mathbb{P}}_{x \sim \mathcal{P}_x}[\text{correct}] > 1 - n^{-10}. \end{aligned}$$

This is known as Yao's minimax principle. Using this, it suffices to exhibit some distribution $\mathcal{P}_x$ such that for all fixed deterministic A, if the $\mathbb{P}_{x \sim \mathcal{P}_x}[\text{correct}] > 1 - n^{-10}$, then $m \geq \Omega(\log n)$.

Consider $x = e_j + w$, where e_j has a one in a random position j and w is drawn according to $\mathcal{N}(0, I/100n)$. Since $\|w\|_2 \approx 0.1$, ℓ_2/ℓ_2 recovery gets x to ± 0.2. So, rounding x recovers the value

of j, which contains $\log n$ bits of information. Therefore, the communication

$$j \to x \to Ax \to \hat{x} \to \hat{j} = j$$

must cost $\Omega(\log n)$ bits. If we can show that each measurement $\langle v, x\rangle$ contains information $I(\langle v, e_j\rangle ; \langle v, x\rangle) \lesssim 1$, then by Lemma 1 which we will argue, we can show that $\log n = I(j; Ax) \leq I(Ae_j; Ax) \lesssim m$.

Observe that each measurement is $\langle v, x\rangle = \langle v, e_j\rangle + \frac{\|v\|_2}{10\sqrt{n}}\mathcal{N}(0, I)$. By Shannon-Hartley Theorem, $I(\langle v, e_j\rangle, \langle v, x\rangle) \leq 0.5 \log(1 + \text{SNR})$, where the signal-to-noise ratio in this case equals

$$\text{SNR} = \frac{\mathbb{E}[\text{signal}^2]}{\mathbb{E}[\text{noise}^2]} = \frac{\sum_j \langle v, e_j\rangle^2 / n}{\|v\|_2^2/100n} = \frac{\|v\|_2^2/n}{\|v\|_2^2/100n} = 100.$$

So, $I(\langle v, e_j\rangle, \langle v, x\rangle) \lesssim 1$. Notice that non-adaptivity is required for $\mathbb{E}[\text{signal}^2]$ above.

3.2 Entropy and mutual information

For a discrete random variable $x \sim p$, the entropy of x defined as

$$H(x) = \sum_i p_i \log \frac{1}{p_i}.$$

is the expected/asymptotic description length. If x is continuous, we write $h(p)$ and the summation becomes integration.

For example, if

$$x = \begin{cases} 0 & \text{with probability } 999/1000 \\ 1 & \text{with probability } 1/1000 \end{cases},$$

then

$$H(x) = \frac{999}{1000} \log \frac{1000}{999} + \frac{1}{1000} \log 1000 \approx \frac{999}{1000} 0.001 + \frac{1}{1000} \log 1000 \approx \frac{\log 1000}{1000}.$$

The mutual information between a and b is defined as

$$I(a; b) = H(a) - H(a, b).$$

3.3 Subadditivity of mutual information

In general, mutual information is not subadditive. For example, consider x and b drawn independently and uniformly from $\{0, 1\}$. Let $y_1 = x \oplus b$ and $y_2 = b$. Then, $I(x; y_1) = I(x; y_2) = 0$ but $I(x; (y_1, y_2)) = 1$.

We can assume matrix A has orthonormal rows. (If $A = U\Sigma V^{\mathrm{T}}$ is the SVD factorization of A, then $\Sigma^{-1}U^{\mathrm{T}}A$ has orthonormal rows and this is an invertible transformation.)

Lemma 1. *$I(Ae_j; Ax) \leq \sum_i I(\langle v_i, e_j\rangle ; \langle v_i, x\rangle)$.*

Proof. Consider $Ax = Ae_j + Aw$. Since $w \sim \mathcal{N}(0, I/100n)$ and A has orthonormal rows, entries in Aw are independent Ae_i and each other. So, it suffices to prove the following lemma. □

Lemma 2. *If $y = \overline{y} + w \in \mathbb{R}^n$ such that w_i's are independent of $\overline{y}$ and each other, then $I(\overline{y}; y) \leq \sum_i I(\overline{y}_i; y_i)$.*

Proof.

$$\begin{aligned}
I(\overline{y}; y) =& h(y) - h(y|\overline{y}) \\
=& h(y) - h(w) \\
=& \sum_i h(y_i|y_1, \ldots, y_{i-1}) - \sum_i h(w_i|w_1, \ldots, w_{i-1}) \\
=& \sum_i h(y_i|y_1, \ldots, y_{i-1}) - \sum_i h(w_i) \\
\leq& \sum_i h(y_i) - \sum_i h(w_i) \\
=& \sum_i h(y_i) - \sum_i h(y_i|\overline{y}_i) \\
=& \sum_i I(\overline{y}_i; y_i)
\end{aligned}$$

□

References

[ACD13] Ery Arias-Castro, Emmanuel J. Candès, Mark A. Davenport. On the Fundamental Limits of Adaptive Sensing. *IEEE Transaction on Information Theory* 59(1): 472-481 (2013)

[FR13] Simon Foucart, Holger Rauhut. A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis, Birkhaser.

[PW13] Eric Price, David P. Woodruff. Lower Bounds for Adaptive Sparse Recovery. *SODA* 2013: 652-663.

[IPW11] Piotr Indyk, Eric Price, David P. Woodruff. On the Power of Adaptivity in Sparse Recovery. *FOCS* 2011: 285-294.