CS5314
Randomized Algorithms

Lecture 12: Chernoff Bounds
(More Application)

Objectives

- Parameter Estimation
- Chernoff bounds of some special RVs
- Set Balancing

Parameter Estimation

- Suppose we want to know the probability that a person in Taiwan has a particular gene mutated
- Given a DNA sample, a lab test can check if there is a mutation
- However, the test is expensive

Can we obtain a "relatively" reliable estimate based on small # of samples ?

Parameter Estimation

Let p = the unknown probability that we want to estimate (p == parameter)

Assume we have n samples.

- After the lab test, let X be the number of samples that contain mutations
- By setting $q = X/n$, we can treat q as an estimate of p

Idea: when n is large, q "should be" very close to p

Parameter Estimation

Some possible questions we may ask:

1. How many samples should we use so that the unknown p is 99.9% likely to be within $q \pm 0.001$?
2. If we just have 1000 samples. What is the probability that p is within $q \pm 0.001$?
3. If we now have 5000 samples. What should be the value of δ so that we can say at least 85% of time p is within $q \pm \delta$?

Parameter Estimation

We define the concept of confidence as follows:

> If Pr(p is not within $q \pm \delta$) $\leq \gamma$,
>
> we say p is within the interval $[q-\delta, q+\delta]$
>
> with confidence $1 - \gamma$

Common Question:

Can we derive a relationship for n, δ, and γ?

Parameter Estimation

Firstly, X is actually a binomial random variable $\mathrm{Bin}(n,p)$ ➔ $E[X] = np$

Now, suppose that p is not within $q \pm \delta$

This implies that either:

(case 1) $p < q - \delta$

So, $nq > n(p + \delta)$ ➔ $X > E[X](1 + (\delta/p))$

(case 2) $p > q + \delta$

So, $nq < n(p - \delta)$ ➔ $X < E[X](1 - (\delta/p))$

Parameter Estimation

So,

$$\Pr(\ p \text{ not within } q \pm \delta\ )$$

$$= \Pr(X > E[X](1 + (\delta/p))\ )$$

$$+ \Pr(X < E[X](1 - (\delta/p))\ )$$

$$< e^{-np(\delta/p)^2/3} + e^{-np(\delta/p)^2/2}$$

$$= e^{-n\delta^2/(3p)} + e^{-n\delta^2/(2p)}$$

$$< e^{-n\delta^2/3} + e^{-n\delta^2/2}$$

Parameter Estimation

By setting $\gamma = e^{-n\delta^2/3} + e^{-n\delta^2/2}$, we thus have

$$\Pr(\ p \text{ is not within } q \pm \delta\ ) < \gamma$$

➔ we have a relationship for n, δ, and γ !!!

Chernoff Bounds for Some Other RVs

We shall look at two simple examples:

1. Sum of RVs, each RV has value +1 or −1 with equal probability 0.5

2. Bin(n,0.5) :

   This is a special case of Sum of Poisson, and we will give tighter bounds than the ones in Lecture 11 (Page 13 and Page 19)

Sum of +1/-1 Random Variables

Theorem: Let $X_1, X_2, \ldots, X_n$ be independent random variables such that

$$\Pr(X_i = +1) = \Pr(X_i = -1) = 0.5$$

Let $X = X_1 + X_2 + \ldots + X_n$.

Then, for all $R > 0$,

$$\Pr(X \geq R) \leq e^{-R^2/(2n)}$$

How to prove?

Sum of +1/-1 Random Variables

To apply Chernoff bound, let us obtain the Moment Generating Function of X

Let M_X be the MGF of X, and
M_{X_i} be the MGF of X_i

Since X_i's are independent, we have

$$M_X(t) = \prod_i M_{X_i}(t) \qquad \text{... (why?)}$$

Sum of +1/-1 Random Variables

Question: What is $M_{X_i}(t)$?

By definition,

$$M_{X_i}(t) = E[e^{tX_i}]$$

$$= 0.5\, e^{t(1)} + 0.5\, e^{t(-1)}$$

$$= \sum_{k \geq 0} t^{2k}/(2k)! \qquad \text{(Taylor series)}$$

$$\leq \sum_{k \geq 0} (t^2/2)^k/k! \qquad \text{(why?)}$$

$$= e^{t^2/2} \qquad \text{(Taylor series)}$$

Sum of +1/-1 Random Variables

So, $M_X(t) = \prod_i M_{X_i}(t) \leq e^{t^2 n/2}$

Thus, for any $t > 0$,

$$\Pr(X \geq R) = \Pr(e^{tX} \geq e^{tR})$$

$$\leq E[e^{tX}] / e^{tR}$$

$$= M_X(t) / e^{tR}$$

$$\leq e^{t^2 n/2 - tR}$$

Sum of +1/-1 Random Variables

By calculus,

$e^{t^2 n/2 - tR}$ is minimized when $t = R/n$

Substituting $t = R/n$ to previous inequality:

$$\Pr(X \geq R) \leq e^{-R^2/(2n)}$$

Remark: By symmetry, we can show that

$$\Pr(X \leq -R) \leq e^{-R^2/(2n)}$$

So, we have:

Sum of +1/-1 Random Variables

Corollary: Let $X_1, X_2, \ldots, X_n$ be independent random variables such that

$$\Pr(X_i = +1) = \Pr(X_i = -1) = 0.5$$

Let $X = X_1 + X_2 + \ldots + X_n$.

Then, for all $R > 0$,

$$\Pr(|X| \geq R) \leq 2e^{-R^2/(2n)}$$

Example: Set Balancing

- Suppose we have a group of m students
- We try to classify them by checking whether they have a particular feature or not
- E.g., Feature 1: Good at Baseball ?

  Feature 2: Good at Maths ?

  Feature 3: Well-behaved ?

  …

Example: Set Balancing

Let n be the number of features

One day, your boss (the headmaster) gives you a difficult task: Can you try to divide the m students into two groups G_1 and G_2, such that

for each k,

# of students with Feature k in G_1

$\approx$ # of students with Feature k in G_2

Example: Set Balancing

Most likely, we cannot find a partition such that for each k,

# of students with Feature k in G_1

is exactly equal to

# of students with Feature k in G_2

However, we can target to find a partition so as to minimize

$\max_k$ { difference in # for Feature k }

Example: Set Balancing

- Formally, we want to find a way of "partition", described by B, as follows:

Given an n x m matrix A, all entries are either 0 or 1,

find an m x 1 vector B, all entries are either +1 or –1, such that

$$\|AB\|_\infty = \max_k |(AB)_k|$$

is minimized

Example: Set Balancing

Yet, we are very lazy… So, we don't want to try all possible partitions …

(^o^ We can try to get a random partition, and fool our boss that this is the best …)

(@@" However, we don't want the result to look very bad… Will it be very bad?

Note: In the worst case,

the difference will be $\Theta(m)$)

Set Balancing

Theorem: For a random m x 1 vector B such that each entry is chosen with equal probability from +1 and -1,

$$\Pr\left(\|AB\|_\infty \geq \sqrt{4m \log_e n}\right) \leq 2/n$$

How to prove?

Proof

Let us examine a particular row, say k, of A

Suppose the there are j ones in row k

Case 1: $j \leq (4m \log_e n)^{0.5}$

Then, $|(AB)_k| \leq (4m \log_e n)^{0.5}$ ... (why?)

Case 2: $j > (4m \log_e n)^{0.5}$

Then, these j ones each has equal chance of contributing +1 or –1 to the sum $(AB)_k$

Proof [Case 2 (cont.)]

By setting $R = (4m \log_e n)^{0.5}$,

$$\Pr(|(AB)_k| \geq R) \leq 2e^{-R^2/(2j)}$$

$$= 2e^{(-4m \log_e n)/2j}$$

$$\leq 2e^{(-4m \log_e n)/2m}$$

$$= 2/n^2$$

Then, by union bound,

$$\Pr(\|AB\|_\infty \geq R) \leq \Sigma_k \Pr(|(AB)_k| \geq R) \leq 2/n$$

Tail of Bin(n,0.5)

Theorem: Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables such that

$$\Pr(Y_i = 1) = \Pr(Y_i = 0) = 0.5$$

Let $Y = Y_1 + Y_2 + \ldots + Y_n$ and $\mu = E[Y] = n/2$. Then,

(1) for all $a > 0$,

$$\Pr(Y \geq \mu + a) \leq e^{-2a^2/n}$$

(2) for all $\delta > 0$,

$$\Pr(Y \geq (1+\delta)\mu) \leq e^{-\mu\delta^2}$$

Tail of Bin(n,0.5)

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that

$$\Pr(X_i = +1) = \Pr(X_i = -1) = 0.5$$

➔ $Y_i = 0.5\, X_i + 0.5$

Let $X = X_1 + X_2 + \ldots + X_n$.

Then,

$$Y = 0.5X + n/2 = 0.5X + \mu$$

Tail of Bin(n,0.5)

In other words,

$$\Pr(Y \geq \mu + a) = \Pr(X \geq 2a) \quad \text{... (why?)}$$

By the previous theorem on Sum of +1/-1 random variables, we have

$$\Pr(Y \geq \mu + a) = \Pr(X \geq 2a)$$

$$\leq e^{-(2a)^2/(2n)}$$

$$= e^{-2a^2/n} \quad \text{... (proof of (1) done)}$$

Tail of Bin(n,0.5)

Next, we set $a = \delta\mu$

Then,

$$\Pr(Y \geq (1+\delta)\mu) = \Pr(Y \geq \mu + a)$$
$$\leq e^{-2a^2/n}$$
$$= e^{-2(\delta\mu)^2/n}$$
$$= e^{-2\delta^2\mu^2/(2\mu)} \quad \text{... since } \mu = n/2$$
$$= e^{-\delta^2\mu}$$

... which completes the proof of (2).

Head of Bin(n,0.5)

Theorem: Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables such that

$$\Pr(Y_i = 1) = \Pr(Y_i = 0) = 0.5$$

Let $Y = Y_1 + Y_2 + \ldots + Y_n$ and $\mu = E[Y] = n/2$. Then,

(1) for all $0 < a < \mu$,

$$\Pr(Y \le \mu - a) \le e^{-2a^2/n}$$

(2) for all $0 < \delta < 1$,

$$\Pr(Y \le (1-\delta)\mu) \le e^{-\mu\delta^2}$$