15-853:Algorithms in the Real World

Announcements:

HW2 is released. Due next Friday noon.

HW1 grades will be released today.

Hashing:

- Concentration bounds
- Load balancing: balls and bins
- Hash functions

Markov's Inequality

The most basic concentration bound.

Let X be a **non-negative** R.V. with mean μ then

$$P(X \geq \alpha) \leq \frac{\mu}{\alpha}$$

Proof: (Did last class)

In other terms,

$$P(X \geq k\mu) \leq \frac{1}{k}$$

Uses expectation only

Chebyshev's Inequality

More powerful than Markov's

Let X be a R.V. with mean μ and variance σ^2

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof: Ideas ?

In other terms,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Stronger since it uses variance information

Smaller the variance more concentrated the R.V. around mean

Central limit theorem

Sum of R.V.'s arises often in analysis of algorithms

Central limit theorem (CLT)

For n **iid** R.V.s with mean μ and variance σ^2

$$\lim_{n \to \infty} \left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \right) \sim \mathcal{N}(0,1)$$

Standard Normal distribution is highly concentrated

- Probability of being a few standard deviations away from mean drops **exponentially**

Very strong result — note there is not much restriction on the form of distribution of Xi's

Chernoff Bound

For any R.V. X, for any t>0

$$P(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}}$$

$$\Rightarrow P(X \geq a) \leq \min_{t>0} \frac{E[e^{tX}]}{e^{ta}}$$

There are many different variants of Chernoff bounds applied to various different distributions

Chernoff Bounds for Binomial

Binomial = sum of Bernoulli (i.e. Binary valued) R.V.s

Let X = $\sum_{i=1}^{n} x_i$

Where X_i's = Bernoulli (p) and independent.

$\mu = E[X] = np$

Then for all $\delta > 0$

$$P(X - np \geq \delta) \leq e^{-\frac{2\delta^2}{n}}$$

$$P(X - np \leq -\delta) \leq e^{-\frac{2\delta^2}{n}}$$

Chernoff Bounds for Binomial

Binomial = sum of Bernoulli (I.e. Binary valued) R.V.s

Let X = $\sum_{i=1}^{n} x_i$

Where X_i's = Bernoulli (p_i) and independent (more general)

μ = E[X] = $\sum_{i=1}^{n} p_i$

Then for all $\delta > 0$

$$P(X \geq (1+\delta)\mu) \leq c_\delta^{\mu}$$

where $$c_\delta = \frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}$$

Chernoff/Hoeffding Bounds

Hoeffding bound is a generalization of Chernoff bound

Hoeffding bound:

Let Xi's be independent R.V.s taking values in [0,1].

Let X = X1 + X2 + … + Xn

Let μ = E[X]

$$P(X > \mu + \lambda) \leq e^{-\frac{\lambda^2}{2\mu+\lambda}}$$

$$P(X < \mu - \lambda) \leq e^{-\frac{\lambda^2}{3\mu}}$$

Chernoff/Hoeffding Bounds

Hoeffding bound:

Let Xi's be independent R.V.s taking values in [0,1].

Let X = X1 + X2 + … + Xn

Let μ = E[X]

$$P(X > \mu + \lambda) \leq e^{\frac{-\lambda^2}{2\mu + \lambda}}$$

Q: Put $\lambda = c\mu$. How does it compare with Markov and Chebysev?

Exponential decay! Much much stronger.

Applications: Sampling and Opinion Polls

Let there be n arbitrary binary numbers in {0,1}

Pick s of them randomly with replacement

Show that the sample mean is within $(1 \pm \varepsilon)$ of the true mean with probability at least $1 - \delta$ if

$$s \geq \Omega\left(\frac{1}{\epsilon^2} \log(1/\delta)\right)$$

<board>

Chernoff bounds imply that large deviations from the mean happen with exponentially small probability.

Applications: Load balancing

The famous balls and bins problem:

N balls and N bins

Randomly put balls into bins

Q: Expected number of balls in each bin?

1

For studying load imbalance we need to understand the number of balls in the **maximally loaded** bin.

Load balancing

Theorem: The max-loaded bin has O($\frac{\log N}{\log \log N}$) balls with probability at least 1 – 1/N.

Proof. High level steps:

1. We will first look at probability of any particular bin receiving more than O($\frac{\log N}{\log \log N}$) balls.
2. Then we will look at the probability of there being a (i.e., at least one) bin with more than these many balls.

Q: What should the probability for Step 2 be?

At most 1/N

Load balancing

Theorem: The max-loaded bin has O($\frac{\log N}{\log \log N}$) balls with probability at least 1 – 1/N.

Proof. High level steps:

1. First look at probability of any particular bin receiving more than O($\frac{\log N}{\log \log N}$) balls.
2. Then look at the probability of there being at least one bin with these many balls.

   (Want this to be at most 1/N)

Q: What should the answer to Step 1 be?

Hint: Union bound

At least 1/N^2 (Can use union bound over all bins)

Load balancing

Theorem: The max-loaded bin has O($\frac{\log N}{\log\log N}$) balls with probability at least 1 – 1/N.

Proof. High level steps:

1. First look at probability of any particular bin receiving more than O($\frac{\log N}{\log\log N}$) balls.

(Want this to be at most 1/N^2)

P (bin i has at least k balls) is $\leq \binom{N}{k}\left(\frac{1}{N}\right)^k$

$$= \frac{N!}{(N-k)!\,k!} \quad \frac{1}{N^k}$$

$$\leq \frac{N^k}{k!} \cdot \frac{1}{N^k} \leq \frac{1}{k!}$$

Load balancing

Theorem: The max-loaded bin has $O(\frac{\log N}{\log \log N})$ balls with probability at least $1 - 1/N$.

Proof.

P (bin i has at least k balls) is $\leq \frac{1}{k!}$

(Want this to be at most 1/N^2)

Using Sterling's approximation:

$$k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

and choosing $k = O(\frac{\log N}{\log \log N})$ gives the desired result