ANNOUNCEMENTS

➤ HW 3 is due tomorrow!

➤ Send project topics

➤ Send email to utah-algo-ta@googlegroups.com, with subject “Project topic”; one email per group; names and UIDs

\[
(1 - x) \leq e^{-x} \quad \text{(easy calculus)}
\]

\[
\left(1 - \frac{1}{10}\right)^2 \leq \left(e^{-\frac{1}{10}}\right)^2 = e^{-2}
\]
LAST CLASS

➤ Hashing

➤ place $n$ balls into $n$ bins, independently and uniformly at random

➤ expected size of a bin $= 1$

➤ number of bins with $k$ balls $\sim = \frac{n}{k!}$

➤ max size of bin $= \mathcal{O}(\log n/\log \log n)$
MAIN IDEAS

➤ Random variables as sums of “simple” random variables

➤ Linearity of expectation

Theorem. Let $X$ be a non-negative random variable. For any $t > 0$,

$$\Pr[X > t \cdot \mathbb{E}[X]] \leq \frac{1}{t}$$

➤ Markov’s inequality is usually not tight

➤ Union bound
MAIN IDEAS

➤ Random variables as sums of “simple” random variables
➤ Linearity of expectation

**Theorem.** Let $X$ be a non-negative random variable. For any $t > 0$,

$$\Pr[X > t \cdot \mathbb{E}[X]] \leq \frac{1}{t}$$

➤ Markov’s inequality is usually not tight
➤ Union bound

**Theorem.** Suppose $E_1, E_2, \ldots, E_n$ are $n$ events in a probability space. Then

$$\Pr[E_1 \cup E_2 \cup \ldots \cup E_n] \leq \Pr[E_1] + \Pr[E_2] + \ldots + \Pr[E_n]$$
When hashing $n$ balls to $n$ bins, outcomes not “as uniform” as one likes

Max “load” of a bin $\approx \frac{\log n}{\log \log n}$.

Many empty bins (HW)

What happens if there are more balls? hash $m$ balls, where $m \gg n/\log n$

“Power of two choices” (Broder et al. 91)
Question: suppose each person votes R or B. Can we predict the winner without counting all votes?

Answer: sample \( n \) of the people; ask who they will vote for, and output the winner in the sample
Things that “ought to” matter:
- sampling to be truly uniform.
- everyone answering truthfully
- the number "n" of samples.
- the "margin" matters — how close the true # votes are . . .
- "confidence" in our prediction.
ANALYZING SAMPLING

Natural formalism:

➤ Choose \( n \) people uniformly at random.
➤ Let \( X_i \) (0/1) be outcome of \( i \)'th person

\( n_0 \) : # in sample voting 0
\( n_1 \) : # in sample voting 1

Predicted winner: \( \{ 0 \} \) if \( n_0 > n_1 \), \( \{ 1 \} \) o/wise.

\[
\Pr[X_i = 1] = \frac{N_1}{N} ; \quad \Pr[X_i = 0] = \frac{N_0}{N}
\]

Each person \( \in \) has a choice of 0 or 1.

\( N \) : entire population
\( N_0 \) : # that vote 0
\( N_1 \) : # that vote 1.
\[ n_1 = X_1 + X_2 + \ldots + X_n \]
\[ n_0 = n - (X_1 + X_2 + \ldots + X_n) = \sum_i (1 - X_i) \]

**What is \( \mathbb{E}[X_i] \)?**

\[ \mathbb{E}[X_i] = n \cdot \frac{N_i}{N} \]

\[ \mathbb{E}[n_1] = n \cdot \frac{N_1}{N} \]

\[ \mathbb{E}[n_0] = n \left( \frac{N_0}{N} \right) \]

What we need: a way to argue that

\[ n_1 \text{ is actually close to } \mathbb{E}[n_1] \]

**Estimation error:** fraction of votes \( n_1 \) received in the sample — fraction of votes \( 1 \) in the population —

\[ \frac{n_1}{n} - \left( \frac{N_1}{N} \right) = \frac{n_1}{n} - \mathbb{E} \left[ \frac{n_1}{n} \right] \]

(1)
We just argue. (from the corr. exprs...

If \( n_0 = E[n_0] \) and \( n_1 = E[n_1] \),
then sample winner \( \equiv \) true winner.

1) What if \( n_0 = 0.4N; \ n_1 = 0.6N \)?

- Our prediction is right iff \( n_0 < n_1 \).
- \( E[n_0] = 0.4 \cdot n \).

Claim:
- As long as \( n_0 < \frac{n}{2} \Leftrightarrow \) our prediction is right.

\[ n_0 - E[n_0] < 0.1 \cdot n \]

2) What if \( n_0 = 0.49N \) and \( n_1 = 0.51N \)?

- Our prediction is right iff

\[ n_0 - E[n_0] < 0.01 \cdot n \]

Goal: ask: if we take 'n' samples, what is the prob. that \( |n_0 - E[n_0]| < 0.01n \)?
ANALYZING SAMPLING

Natural formalism:

➤ Choose $n$ people uniformly at random.
➤ Let $X_i (0/1)$ be outcome of $i$’th person

➤ Error in estimation: $||\text{empirical mean} - \text{true expectation}||$?
➤ “Confidence”

Ideal guarantee: $||\text{empirical mean} - \text{true expectation}|| < 0.001 \text{ w.p. } 0.999$
MARKOV?

\[
\begin{cases}
N_0 = 0.4N \\
N_1 = 0.6N
\end{cases}
\]

Want: \( \eta_0 - \mathbb{E}[\eta_0] < 0.1 \eta_0 \)

\[
P_r \left[ \eta_0 > t \mathbb{E}[\eta_0] \right] \leq \frac{1}{t}.
\]

\[
\frac{\mathbb{E}[\eta_0]}{0.4} < \mathbb{E}[\eta_0] \left( 1 + \frac{0.1}{0.4} \right)
\]

\[
t = \frac{5}{4} \quad \text{failure prob.} \leq \frac{4}{5},
\]

\[
\eta = 10^6; \quad n = 1
\]
CAN WE USE THE "NUMBER OF SAMPLES"?

Variance of random variable $X$

Definition:

$$\text{Variance (}X\text{)} = \text{var}(X)$$

Basic property:

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

$$\text{var}(X) = \text{var}(X_1) + \text{var}(X_2) + \ldots + \text{var}(X_n) \text{ if } X_i \text{ are independent}.$$
Message: If a random variable has "low variance", then Markov can be improved.

Theorem: Let \( X \) be a random variable whose variance \( \sigma^2 \). Then

\[
\Pr \left[ |X - \mathbb{E}X| > t \cdot \sigma \right] \leq \frac{1}{t^2}.
\]
Back to sampling: we wanted

\[ n_0 - \mathbb{E}[n_0] \leq 0.1 \pi \]

\[ \Pr \left[ |n_0 - \mathbb{E}[n_0]| > \frac{\alpha \cdot \sigma^2}{\pi} \right] \leq \frac{1}{\alpha^2} . \]

\[ 0.1 \pi \equiv \alpha \cdot \sigma^2 \]

Idea: use compute this \( \alpha \).
VARIANCE OF AVERAGE
BOUND VIA CHEBYCHEV
WHAT IF WE TAKE HIGHER POWERS?

\[ \mathbb{E}[(X - \mathbb{E}X)^4] \leq \ldots \]

“Moment methods”

- Usually get improved bounds
**Theorem 3.2.** [Chernoff 81] Let $X_1, \ldots, X_n$ be independent random variables with

$$\Pr(X_i = 1) = p_i, \quad \Pr(X_i = 0) = 1 - p_i.$$ 

We consider the sum $X = \sum_{i=1}^{n} X_i$, with expectation $E(X) = \sum_{i=1}^{n} p_i$. Then, we have

- **(Lower tail)** \( \Pr(X \leq E(X) - \lambda) \leq e^{-\lambda^2 / 2E(X)} \),

- **(Upper tail)** \( \Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(E(X) + \lambda / 3)}} \).
INTERPRETING THE CHERNOFF BOUND
**INTERPRETING THE CHERNOFF BOUND**

*Useful heuristic:*

- Sums of independent random variables don’t deviate much more than the variance
Theorem 3.4. [McDiarmid 98] Let $X_i$ (1 ≤ $i$ ≤ $n$) be independent random variables satisfying $X_i \leq E(X_i) + M$, for 1 ≤ $i$ ≤ $n$. We consider the sum $X = \sum_{i=1}^{n} X_i$ with expectation $E(X) = \sum_{i=1}^{n} E(X_i)$ and variance $\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i)$. Then, we have

$$\Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\text{Var}(X) + M \lambda/3)}}.$$
ESTIMATING THE SUM OF NUMBERS
ESTIMATING THE SUM OF NUMBERS