Advanced Algorithms

Lecture 18: Sampling

Announcements

- HW 3 graded
- **HW 4 due on Friday**

Randomized alg recap

- Expected running time (boost via ~~"amplification~~ repetition)
- Hashing — analyzed via "balls and bins"
- Key ideas:
  - define appropriate random variables
  - linearity of expectation, Markov's inequality, Union bound, ...

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Recap

Markov's inequality: let X be a <u>non-negative</u> random variable with expectation C. Then prob[X > tC] <= 1/t.

Union bound: let $E_1, E_2, \ldots E_n$ be any collection of events in a probability space. Then we have

$$\Pr(E_1 \text{ -or- } E_2 \text{ -or- } \ldots \text{ -or- } E_n) \leq \Pr(E_1) + \Pr(E_2) + \ldots + \Pr(E_n)$$

becomes equality if

E_i are all disjoint.

Sampling / estimation

Average of elements in array

[-1, 1]

Problem: let *A* be an array with *n* elements, each in interval [~~0,1~~]. Find the average of all elements.

- Twist: suppose we are OK with a little bit of error (+/- 0.01)

- Basic sampling algorithm:
  - sample *k* of the elements (with replacement), find their "empirical average" (sum/*k*)

Trade-offs

<u>Key quantities:</u>

- Number of samples (k)
- Error in result (+/- "true average")
- Confidence in result (error bound holds w.p. ...) $\geq 90\%$.

Formalizing sampling

X_i's are all independent r.v.s.

Right random variable?

$A[1], \ldots, A[n]$

X_i : value of the i'th sample [could be -ve].

$\hat{X}$: empirical average of samples.

$$\hat{X} := \frac{X_1 + X_2 + \cdots + X_k}{k}$$

What is $\mathbb{E}[X_i]$? $= \sum_{j=1}^{n} \Pr[\text{i'th sample was } A[j]] \cdot A[j] = \mu$

$\Pr[\text{i'th sample was } A[j]] = 1/n$

(average of the full sample)

Vanilla Markov

What is $\mathbb{E}[\hat{X}]$? $= \sum_{i=1}^{k} \frac{\mathbb{E}[X_i]}{k} = \mu$.

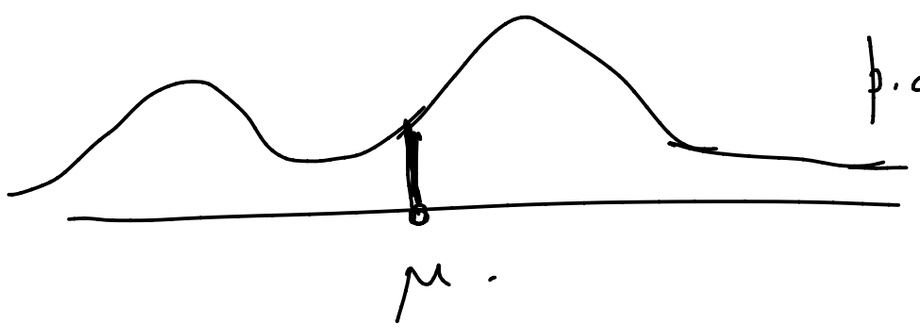
But if $A[i]$ were all in $\underline{[0,1]}$, you can use Markov,

and get

$$\Pr[\hat{X} \geq t\underbrace{\mu}_{0.5}] \leq \frac{1}{t} \qquad [\text{'say } t = 4].$$

Variance

$$\mathrm{Var}(\hat{X}) := \mathbb{E}\left[\left(\hat{X} - \underbrace{\mathbb{E}(\hat{X})}_{\mu}\right)^2\right] = \mathbb{E}\left[(\hat{X} - \mu)^2\right].$$



$$\hat{X} = \frac{X_1 + X_2 + \cdots + X_k}{k}.$$

intuitively, the "average squared distance" to the mean.

$$\mathbb{E}\left[(\hat{X} - \mu)^2\right] = \mathbb{E}\left[\left(\frac{X_1 + X_2 + \cdots + X_k}{k} - \mu\right)^2\right] = \mathbb{E}\left[\left(\frac{(X_1 - \mu) + \cdots + (X_k - \mu)}{k}\right)^2\right]$$

Variance of a sum

$$V = \mathbb{E}\left[\left(\frac{(X_1-\mu)+(X_2-\mu)+\cdots+(X_k-\mu)}{k}\right)^2\right]$$

$$(a_1+a_2+\cdots+a_k)^2 = \sum_i a_i^2 + \sum_{i\neq j} a_i a_j$$

$$= \frac{1}{k^2}\left\{\sum_i \mathbb{E}\left[(X_i-\mu)^2\right] + \sum_{i\neq j} \mathbb{E}\left[(X_i-\mu)(X_j-\mu)\right]\right\}$$

Take one term of the second kind:

$$\mathbb{E}\left[(X_i-\mu)(X_j-\mu)\right] \to \mathbb{E}\left[(X_i-\mu)\right]\cdot\mathbb{E}\left[X_j-\mu\right].$$

<u>Fact</u>: if X, Y are independent r.v.s,

then $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

$$\mathbb{E}[X_i - \mu] = \underbrace{\mathbb{E}[X_i]}_{\overset{\shortparallel}{\mu}} - \mu = 0.$$

$$V = \frac{1}{k^2} \cdot \sum_{i=1}^{k} \underbrace{\mathbb{E}\left[(X_i - \mu)^2\right]}_{\leq \underline{\underline{4}}} \leq \frac{4k}{k^2} = \frac{4}{k}.$$

$\mathbb{E}[|\hat{x} - \mu|]$

Sample size and variance

If we take k samples, define $\hat{X} = \dfrac{X_1 + \cdots + X_k}{k}$,

then $\mathrm{var}(\hat{X}) \leq \dfrac{4}{k}$.

$$\mathrm{var}(\hat{X}) = \underbrace{\mathbb{E}\left[(\hat{X} - \mu)^2\right]}_{\Pr\left[|\hat{X} - \mu| > \frac{4}{\sqrt{k}}\right] \leq \frac{1}{4}}$$

$$\Pr\left[(\hat{X} - \mu)^2 > 4V\right] \leq \frac{1}{4}$$

$$\Downarrow$$

$$\Pr\left[|\hat{X} - \mu| > \frac{4}{\sqrt{k}}\right] \leq \frac{1}{4}. \iff \Pr\left[(\hat{X} - \mu)^2 > \frac{16}{k}\right] \leq \frac{1}{4}.$$

Chebychev's inequality

<u>Theorem</u>: Let X be any random variable with and mean $= \mu$ variance $= \sigma^2$. Then for all $t \geq 1$,

$$\Pr\left[|X - \mu| \geq t\sigma \right] \leq \frac{1}{t^2}.$$

$\sigma^2 = \frac{4}{k}$

or $\sigma = \frac{2}{\sqrt{k}}$.

σ: std-deviation

<u>Proof</u>: $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$ [defn of variance being σ^2]

By Markov: $\Pr\left[(X-\mu)^2 > t^2 \cdot \sigma^2\right] \leq \frac{1}{t^2}$

$\Updownarrow$

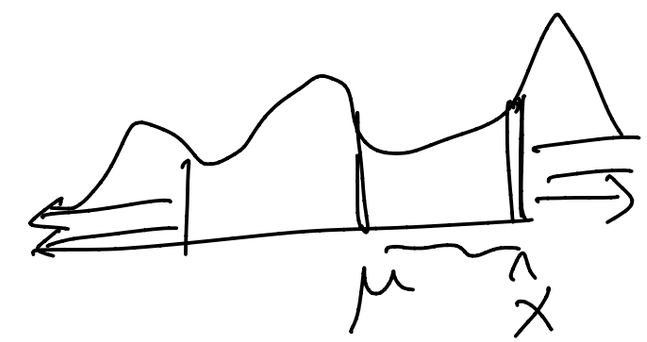
$|X - \mu| > t\sigma$

Confidence bounds

If we an algorithm that outputs $\hat{X}$, then
(k samples)

$t=2$

$$\rightarrow \boxed{\Pr\left[\,|\hat{X}-\mu| > \frac{4}{\sqrt{k}}\right] < \frac{1}{4}}$$



Suppose we take $k = \underline{16 \cdot 10^4}$ samples

$$\Pr\left[\,|\hat{X}-\mu| > 0.01\right] < \frac{1}{4} \implies \text{w.p.} \geq \frac{3}{4},$$

$$|\hat{X}-\mu| \leq 0.01$$

$$\Pr\left[|\hat{x} - \mu| > \frac{2t}{\sqrt{k}} \right] \leq \frac{1}{t^2}$$ [Came from Chebychev's ineq.].

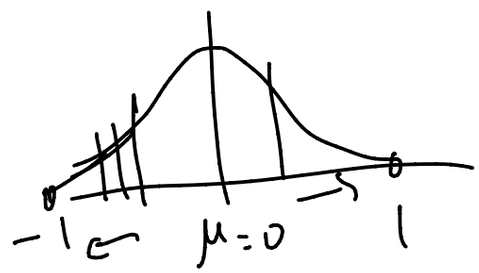
Suppose we want confidence $= \frac{99}{100}$

we will plug in $t = 10$

and $k = 4 \cdot \underline{100 \cdot 10^4}$; $\frac{2t}{\sqrt{k}} = 0.01$

$$\Pr\left[|\hat{x} - \mu| > 0.01 \right] < \frac{1}{100}$$

How "tight" is variance bound?



$\mathbb{E}[(X_i - \mu)^2] \to \leq 4 \to \frac{1}{2}$

- I.e., is the error in estimation *truly* around 1/\sqrt k ?
- For "bounded" random variables, variance calculation is fairly tight
- Central limit theorem

if variance of each $X_i \approx 1$ then $\pm \frac{1}{\sqrt{k}}$

if we want error $\approx \varepsilon$, then # samples $\approx \frac{1}{\varepsilon^2}$.

Higher “moments”?

- What if we compute fourth powers?

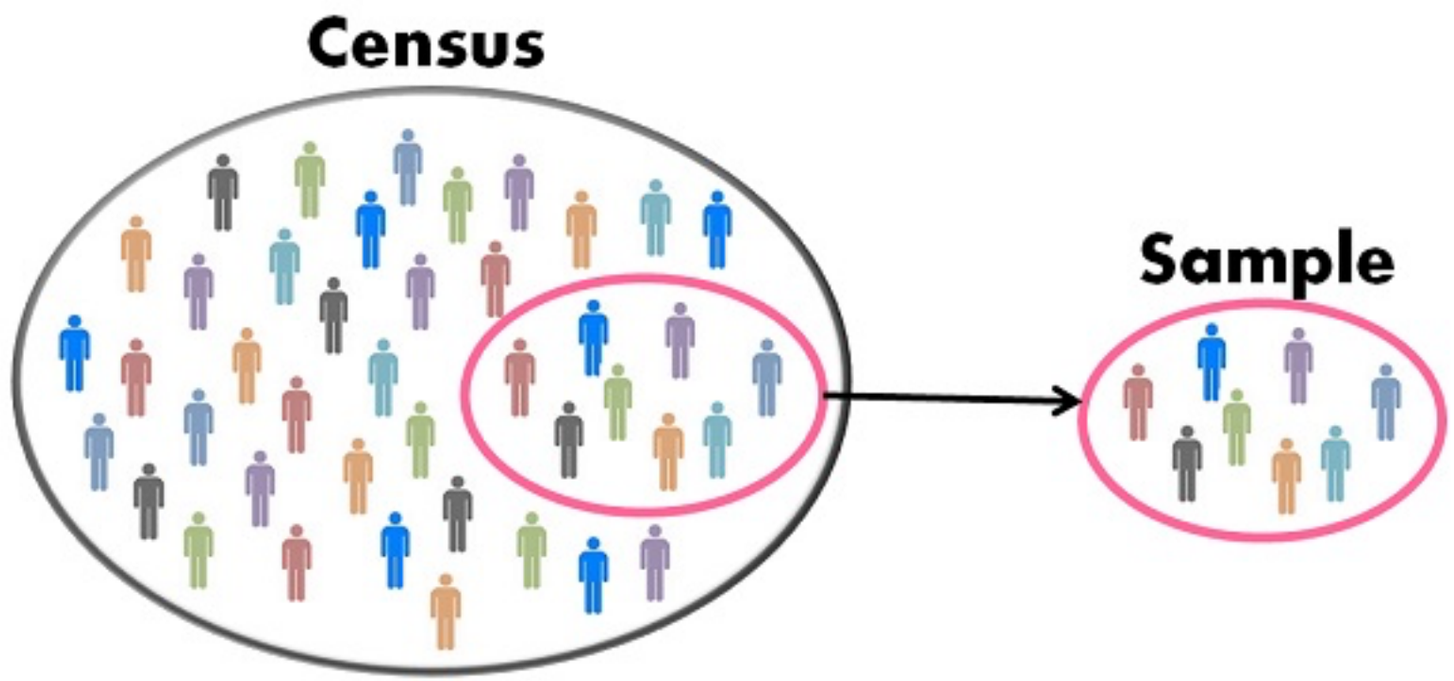
- Chernoff bound

Average of elements in array

Problem: let A be an array with n elements, <u>each in interval [0,1]</u>. Find the average of all elements.

- Bound on entries is important …

Applications of sampling



Problem: predicting an election; say everyone votes *R* or *B*

Procedure, trade-offs

Trade-offs

- Number of samples (k)
- Error in result (+/- “true average”)
- Confidence in result (error bound holds w.p. ...)
- How close is the margin in the **true population**?