Lecture 2 Jan 11, 2024

- Sortedness Testing (cont.)
- Concentration of random variables

  (Markov, chebyshev, chernoff, Hoeffding)
- Running example: estimating coin bias.

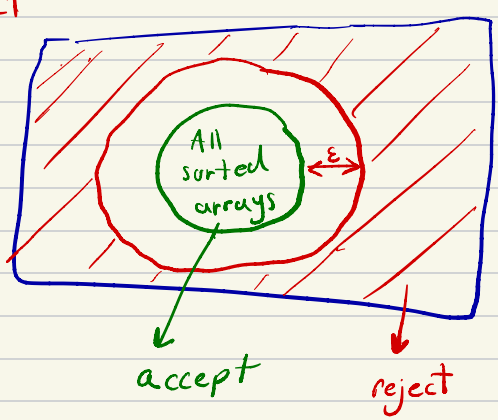
Testing sortedness (general case)

Recall def

Give an array A,

Design an algorithm A s.t. with prob. $1-\delta$

- if A is sorted, A outputs accept
- if A is ε-far from being sorted, A outputs reject



what does ε-far mean here?

distance between two array of size n:

$$dist(A, A') = \frac{\text{entries we need to change to change } A \text{ to } A'}{n}$$

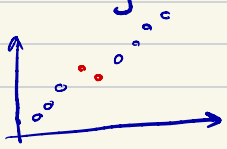
$$P = \{ \text{all sorted arrays} \}$$

$$dist(A, P) = \min_{A' \in P} dist(A, A')$$

ε-far from sortedness = We need to change $\geq \varepsilon \cdot n$ entries in A to get a sorted array.

Why randomly throwing darts in the dark won't work in general case?

unable to detect local changes



New algorithm

Binary search base algorithm.

Sorted $\Rightarrow$ binary search works.

" $\overset{?}{\Leftarrow}$ "

Assumption: WLOG, entries of A are distinct.

Try s times

pick a random $i \in [n]$

$\ell \leftarrow$ Binary search$(A, A[i])$

if $(\ell \neq i)$

return reject

return accept.

* If A is sorted $\Rightarrow$

all calls to binary search work correctly

$\Rightarrow$ the algorithm returns accept w. prob 1.

* If A is ε-far from sorted $\overset{?}{\Rightarrow}$

$p :=$ the probability of binary search fails when we are ε-far

what we need:

$$\Pr[\text{outputting accept} \mid \varepsilon\text{-far}] = (1-p)^s \leq \delta$$

by setting $s = \frac{\log(1/\delta)}{p}$

if $p < \varepsilon$ $\Rightarrow$ $(1-\varepsilon) \cdot n$ many entries are nice

$\Downarrow$ Lemma 1

$(1-\varepsilon) \cdot n$ many entries are sorted

$\Downarrow$

A is not ε-far from being sorted

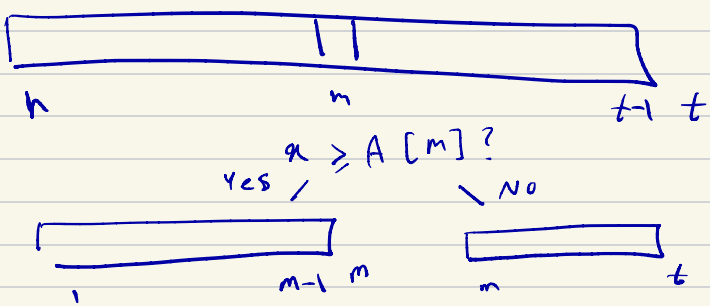
$\Rightarrow$ $p \geq \varepsilon$ $\Rightarrow$ $S = \frac{\log 1/\delta}{\varepsilon}$ would be enough.

```
Binary_search (array A, value x, indices h, t)

    if (t < h)
        return h

    m ← ⌊(h+t)/2⌋

    if (A[m] = x)
        return m

    if (A[m] > x)
        return binary_search (A, x, h, m-1)

    if (A[m] < x)
        return binary_search (A, x, m+1, t)
```

Binary search on a sorted A returns the smallest i such that

$$A[i] \geq x$$

Binary search $(A, x, 1, n+1)$ could return $n+1$.



$A[m]$ is called pivot

We say i is nice if the binary search on $x = A[i]$ returns i

Lemma 1 Suppose we have two nice indices i and $j \in [n]$. If $i < j$ then $A[i] < A[j]$

Proof

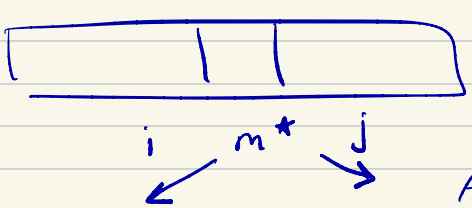
pivots of i

$$m_1^{(i)}, m_2^{(i)}, \ldots, m_{k_i}^{(i)} = i$$

pivots of j

$$m_1^{(j)}, m_2^{(j)}, \ldots, m_{k_j}^{(j)} = j$$

m^* = last mutual pivot



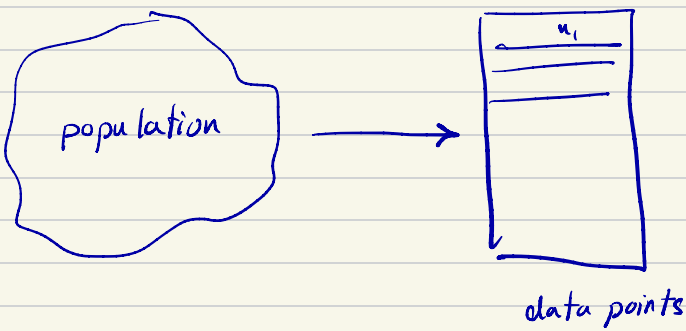
$A[i] < A[m^*] < A[j]$

Hypothesis testing (property testing of distions)

We used randomness to model the world.

data points are random samples from

an unknown data distribution



distribution p $\quad x_1, \ldots, x_m$

$x_i \sim p$

Estimating coin bias

$p := \Pr[\text{head}]$

Testing a coin is fair:

- if $p = \frac{1}{2}$, output accept w. prob. $1-\delta$.
- if $|p - \frac{1}{2}| > \varepsilon$, output reject w. prob $1-\delta$.

Algorithm

Flip a coin m times

$X \leftarrow$ # heads

if $\left|\frac{X}{m} - \frac{1}{2}\right| \leq \varepsilon$

return accept

else

return reject

Question: How well $\frac{X}{m}$ approximate p?

what should be m?

boils down → How well $\frac{X}{m}$ concentrate around p?

Concentration of random variables.

Questions: { Estimating average height of students
exit polls

n samples:

$X_1, X_2, \ldots, X_n \sim P$

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu := \mathop{\mathbb{E}}_{X \sim P}[X]$$

Goal: measure how much $\bar{X}_n$ deviates from μ

Law of Large numbers

(weak) $\forall \varepsilon \quad \lim_{n \to \infty} \Pr[|\bar{X}_n - \mu| < \varepsilon] = 1$

(strong) $\Pr\left[\lim_{n \to \infty} \bar{X}_n = \mu\right] = 1$

Central Limit Theorem:

$\mathrm{Var}_{X\sim P}[X]$

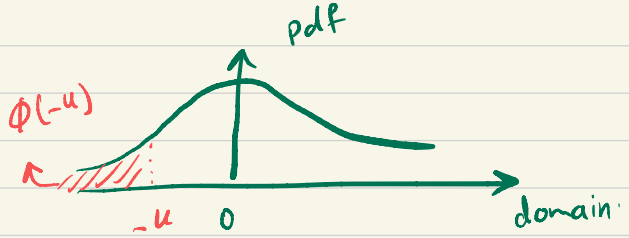
$$\sqrt{n}\,(\bar{X}_n - \mu) \to \mathcal{N}(0, \sigma^2)$$

$Z \sim \mathcal{N}(0,1)$

$$\Pr\left[\frac{\sqrt{n}\,|\bar{X}_n - \mu|}{\sigma} > u\right] \approx \Pr[|Z| > u]$$

$$= 2\Phi(-u)$$

where Φ is the cdf of the standard normal dist.



Look up table

$u = 1.96 \quad \to \quad 2\Phi(-u) \simeq 95\%$

Hence: with prob. 0.95

$$\mu \in \left[\bar{X}_n - 1.96\,\sigma/\sqrt{n},\ \bar{X}_n + 1.96\,\sigma/\sqrt{n}\right]$$

- Quality of Approximation varies depending on P.

These are asymptotic results. Very general, but

- work in the limit,
- Do not indicate the relationship among the parameters.

n, d, ε, δ ?

d → dimension

ε → error

δ → confidence (in our example δ was $1-0.95 = 0.05$)

what about finite sample setting ?

Usefull tools to show concentration (tail bounds)

Markov's inequality:

For non-negative random variable X, and $a > 0$:

$$\Pr[X \geq a] \leq \frac{E[X]}{a}$$

proof.

$$E[X] = \int_0^\infty x \Pr[X = x] \, dx \quad \text{(pdf)}$$

$$= \int_0^a x \Pr[X = x] \, dx + \int_a^\infty x \Pr[X = x] \, dx$$

$$\geq 0 + \int_a^\infty a \Pr[X = x] \, dx$$

$$\geq a \Pr[X \geq a]$$

$$\Rightarrow \Pr[X \geq a] \leq \frac{E[X]}{a} \qquad \square$$

→ back to coin example

works well for small p

if $p \leq 0.01$

$$\Pr\left[\frac{X}{m} > 0.1\right] \leq \frac{E[X]}{0.1} \leq 0.1$$

not very meaningful when $p = \frac{1}{2}$

Chebyshev's inequality

For a random variable with finite mean and variance, and $k > 0$:

$$\Pr[\,|X - \mathbb{E}[X]| \geq k\sigma\,] \leq \frac{1}{k^2}$$

(σ: standard deviation of X)

proof:

$$\Pr[\,|X - \mathbb{E}[X]| \geq k\sigma\,]$$

$$= \Pr[\,(X - \mathbb{E}[X])^2 \geq k^2\sigma^2\,]$$

$$\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

(≤: Markov) ∎

→ back to coin example

$$E\left[\frac{X}{m}\right] = p$$

$$\mathrm{Var}\left[\frac{X}{m}\right] = \frac{p(1-p)}{m}$$

$$\Pr\left[\left|\frac{X}{m} - p\right| > \varepsilon\right] \leq \frac{\mathrm{Var}[X/m]}{\varepsilon^2} \leq \frac{1}{m\varepsilon^2} \leq \delta$$

$$m = \frac{1}{\delta \cdot \varepsilon^2}$$

right dependencies to ε

but not δ

Chernoff bound:

m Bernoulli random variable: $X_1, X_2, \dots X_m$

$$X_i \sim \text{Ber}(p_i) \qquad X_i = \begin{cases} 1 & \text{with prob } p_i \\ 0 & \text{" } \sim 1-p_i \end{cases}$$

empirical mean $\quad X := \frac{1}{m} \sum_{i=1}^{m} X_i$

and true mean $\quad \mu := \frac{1}{m} \sum_{i=1}^{m} p_i$

$$\Pr[X - p > \varepsilon p] \leq e^{-mp\varepsilon^2/3}$$

$$\Pr[p - X > \varepsilon p] \leq e^{-mp\varepsilon^2/2}$$

general structure of the proof:

(can be applied to any random variable)

For all $\varepsilon > 0$, $t > 0$:

$$\Pr[X \geq \varepsilon] = \Pr[e^{tX} \geq e^{t\varepsilon}]$$

$$\leq \frac{\mathbb{E}[e^{tX}]}{e^{t\varepsilon}} = e^{-t\varepsilon} M_X(t)$$

↙ Markov

↓ moment generating func

since the bound holds for any t, we can conclude:

$$\Pr[X \geq \varepsilon] \leq \inf_{t>0} e^{-t\varepsilon} M_X(t) \quad \square$$

→ back to coin example

$$\Pr[\,|X - p| \geq \varepsilon\,] \leq 2\exp(-mp\varepsilon^2)$$

$$m = \frac{1}{p\varepsilon^2}\log\frac{2}{\delta} \qquad \leq \delta$$

works when $p = 1/2$. ✓

Hoeffding bound:

$$\Pr\left[\, X - \mu > \varepsilon \,\right] < e^{-2m\varepsilon^2}$$

$$\Pr\left[\, \mu - X < \varepsilon \,\right] < e^{-2m\varepsilon^2}$$

→ back to coin example

$$m = \frac{\log(2/\delta)}{2\varepsilon^2} \Rightarrow$$

$$\Pr\left[\, |X - \mu| > \varepsilon \,\right] \leq \delta$$