Concentration of measure (tail bounds)

Today
- quick recap of tail bounds
- application to balls in bins
- Poisson approx for analyzing balls in bins
- estimating reachability

The more you know about r.v. X, the better the bounds

① If only thing you know is mean & $X \geq 0$ (but not $=0$)

Markov's Inequality

$$\Pr(X \geq t) \leq \frac{E(X)}{t} \quad t>0 \qquad \left[\Pr(X \geq c\,E(X)) \leq \frac{1}{c}\right] \; c>0$$

Corollary

If $0 \leq X \leq B \Rightarrow \Pr(X \leq t) \leq \frac{E(B-X)}{B-t}$

② If you know mean μ and variance σ^2

Chebychev's Inequality

$$\forall t>0 \quad \Pr(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2}$$

std deviation

③ Sums of indep r.v.'s

Theorem $X = X_1 + X_2 + \dots + X_n$ where X_i's indep & $X_i \in [0,1]$ with $E(X_i) = p_i$

$$\mu = E(X) = \sum_{i=1}^{n} p_i \qquad p = \frac{\mu}{n}$$

$$\boxed{\Pr(X \geq \mu+\lambda) \leq e^{-n H(p+\frac{\lambda}{n} \| p)}} \quad (1)$$

$$\boxed{\Pr(X \leq \mu-\lambda) \leq e^{-n H(1-p+\frac{\lambda}{n} \| 1-p)}} \quad (2)$$

$$\boxed{H(x \| p) = x \ln\left(\frac{x}{p}\right) + (1-x)\ln\left(\frac{1-x}{1-p}\right)} \quad \text{relative entropy}$$

More useful forms (corollaries)

$$\boxed{\begin{matrix} \Pr(X \geq \mu + \lambda) \\ \Pr(X \leq \mu - \lambda) \end{matrix} \leq e^{-\frac{2\lambda^2}{n}}} \quad (3)$$

given a CLT

$X \sim \mathrm{Bin}(n,p)$

$\sigma^2 = np(1-p)$

$$\Pr(X \geq \mu + k\sigma) \leq e^{-\frac{2(k\sigma)^2}{n}}$$

$$= e^{-k^2 \text{const.}}$$

$$\Pr\left(X \leq (1-\delta)\mu\right) \leq e^{-\frac{\delta^2 \mu}{2}} \qquad 0 \leq \delta \leq 1 \qquad \Leftarrow \mu_L \leq \mu$$

$$\Pr\left(X \geq (1+\delta)\mu\right) \leq e^{-\mu\left[(1+\delta)\ln(1+\delta) - \delta\right]} \leq \begin{cases} e^{-\frac{\delta^2 \mu}{2+\delta}} & \delta > 0 \\ e^{-\frac{\delta^2 \mu}{3}} & 0 < \delta \leq 1 \end{cases} \quad (4)$$

$\mu \leq \mu_H$

probably use most frequently.

$\mu_L \leq \mu$

$\mu \leq \mu_H$

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

X_i's indep. X_i takes values in $[a_i, b_i]$ $\quad E(X) = \mu$

$$\boxed{\begin{matrix} \Pr(X \leq \mu - \lambda) \\ \Pr(X \geq \mu + \lambda) \end{matrix} \leq e^{-\frac{2\lambda^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}} \quad (5)$$

Comments:

- many different versions (hold for Poisson r.v.s, Gaussian etc); we'll see more later.
- also hold for "negatively correlated r.v.s" [negatively associated]

$$E\left[e^{t(X+Y)}\right] \leq E\left[e^{tX}\right] E\left[e^{tY}\right]$$

Example

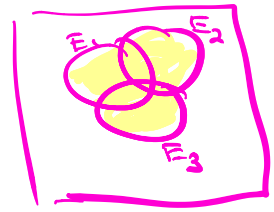
- edges in random spanning trees

sampling random spanning trees.

$$X_e = \begin{cases} 1 & e \in T \\ 0 & \text{o.w.} \end{cases}$$

One more simple probabilistic tool

Union Bound



Let $E_1, E_2, \ldots, E_k$ be any collection of events in prob space

Then $\Pr(E_1 \vee E_2 \vee \ldots \vee E_k) \le \sum_{i=1}^{k} \Pr(E_i)$

- Chernoff bounds
- union bound
- cleverness

Application: Balls in Bins

jobs → machines

items → hash table

requests → servers

n bins

Throw n balls, one at a time, independently at random into a bin. Prob → i^{th} bin $= \frac{1}{n}$

$\mathrm{Bin}(n,p)$

$X_i \sim \mathrm{Bin}(n, \frac{1}{n})$

$X_i := \#$ balls in i^{th} bin at end (after n balls) $E(X_i) = 1$

X_i's not indep.

Theorem $\Pr\left(\max_i X_i \ge \boxed{e \frac{\ln n}{\ln\ln n}}\right) \le \frac{1}{n^c}$ for some $c > 0$

$o(1)$

proof idea: prove that prob that a particular bin so unlikely to exceed this bound that a union bound gives result

$E(\max_i X_i) = \Theta\left(\frac{\ln n}{\ln\ln n}\right)$

Pf Fix bin X_1 $X_i \sim \mathrm{Bin}(n, \frac{1}{n})$. $E(X_1) = 1$

X_1 & X_2 dep. $X_1 = B_1 + B_2 + \ldots + B_n$ $B_i \sim \mathrm{Ber}(\frac{1}{n})$

$\Pr(X \ge (1+\delta)\mu) \le e^{-\mu[(1+\delta)\ln(1+\delta) - \delta]}$
$\qquad c \ln c \quad c-1$

$\Pr(X_1 \ge \boxed{e\frac{\ln n}{\ln\ln n}} \underbrace{E(X_1)}_{1})$

$c = (1+\delta)$

$\boxed{\propto \frac{\ln n}{\ln\ln n}}$

$\le e^{-[c \ln \ln c - (c-1)]} \le e^{-c(\ln c - 1)}$

$\ln c = 1 + \ln\ln n - \underbrace{\ln\ln\ln n} \ge 1 + \frac{\ln\ln n}{2}$

$e^v = \underbrace{1 + v} + \frac{v^2}{2} + \ldots$

$\ln\ln n > 2\ln\ln\ln n \xLeftarrow{v = \ln\ln\ln n} e^v > 2v$ $v > 1$

$$\Pr(X_1 \geq c) \leq e^{-c\left[1+\frac{\ln\ln n}{2}-1\right]} = e^{-\frac{e}{2}\ln\ln n}$$

$$= e^{-\frac{e}{2}\ln n} = n^{-\frac{e}{2}} \leq \underline{n^{-1.35}}$$

plug c back in

needs $o\left(\frac{1}{n}\right)$

$\forall$ bins i: $\quad \Pr\left(X_i \geq \frac{e \ln n}{\ln\ln n}\right) \leq n^{-1.35} \qquad n^{-(1+c)}$

$$\Pr\left(\exists \text{ bin } i \text{ with } X_i \geq \frac{e\ln n}{\ln\ln n}\right) = \Pr(E_1 \vee E_2 \vee \dots \vee E_n)$$

$E_i = 1 \quad X_i \geq e\frac{\ln n}{\ln\ln n}$

$$\leq \sum_{i=1}^{n} \Pr(E_i) \leq n \cdot n^{-1.35}$$

$$= n^{-0.35} \longrightarrow 0 \quad \text{as } n \to \infty$$

$$\frac{1}{n^{0.35}}$$

Poisson approx $\qquad$ m balls $\to$ n bins

$X_1, X_2, \dots, X_n$

final loads in bins

X_i: # balls in bin i at end.

$$X_i \sim \text{Bin}\left(m, \frac{1}{n}\right) \xrightarrow[m\to\infty]{\text{in limit}} \text{Poi}\left(m \cdot \frac{1}{n}\right)$$

$\frac{m}{n}$ is const. $\qquad \frac{m}{n}$.

$\text{Bin}(n,p) \longrightarrow \text{Poi}(np)$

$\text{Poi}(\lambda_1)\ \text{Poi}(\lambda_2) \dots \text{Poi}(\lambda_n)$

$Z_1 \quad Z_2 \quad \dots \quad Z_n$

$\sum Z_i \sim \text{Poi}\left(\sum_{i=1}^{n} \lambda_i\right)$

indep $\text{Poi}\left(\frac{m}{n}\right)$ $\quad \boxed{Y_1 \quad Y_2 \quad \dots \quad Y_n}$

Observation.

$$\Pr(X_1 = k_1, X_2 = k_2, \dots, X_n = k_n) = \frac{\Pr(Y_1 = k_1)\Pr(Y_2 = k_2)\cdots P(Y_n = k_n)}{\Pr\left(\sum_{i=1}^{n} Y_i = m\right)}$$

$\sum_{i=1}^{n} k_i = m$

real loads behave like indep Poissons conditioned on $\sum_{i=1}^{n} Y_i = m$

$$\Pr\left(Y_1 = k_1 \dots Y_n = k_n \,\middle|\, \sum_{i=1}^{n} Y_i = m\right)$$

Theorem m balls $\to$ n bins

Let $\mathcal{E}$ be any event that depends only on $(X_1, X_2, \ldots, X_n)$

s.t. $\Pr(\mathcal{E})$ is monotone $\uparrow$ in # balls (m)

or monotone $\searrow$ (m)

$\Pr(\mathcal{E})$ monotone $\uparrow$ max $X_i \geq 1000$

$\Pr(\mathcal{E}')$ monotone $\searrow$ $\forall i \; X_i \leq 100$

read Alistair Sinclair's notes

$$\Pr(\mathcal{E}(X_1, \ldots, X_n)) \leq 4 \Pr(\mathcal{E}(Y_1, \ldots, Y_n))$$

unlikely $\Leftarrow$ unlikely.

---

we showed w.h.p. n balls $\to$ n bins max load $O\left(\frac{\ln n}{\ln \ln n}\right)$

now w.h.p. max load $\Omega\left(\frac{\ln n}{\ln \ln n}\right)$

$1 - o(1)$

$o(1)$ goes to 0 as $n \to \infty$

show very unlikely that every bin has load $\leq (1-\beta)\frac{\ln n}{\ln \ln n}$

pos const.

indep Poi(1)

$$\Pr\left(\max(Y_1, \ldots, Y_n) < \underbrace{\boxed{(1-\beta)\frac{\ln n}{\ln \ln n}}}_{k}\right)$$

$$= \prod_{i=1}^{n} \Pr(Y_i < k) = \prod_{i=1}^{n} \left(1 - \underbrace{\Pr(Y_i \geq k)}_{\geq \frac{e^{-1}}{k!}}\right)$$

Poi(1)

$$\Pr(Y_i \geq k) = \sum_{j=k}^{\infty} \frac{e^{-1}}{j!} \geq \frac{e^{-1}}{k!}$$

$$\leq \left(1 - \frac{e^{-1}}{k!}\right)^n$$

$1 - x \leq e^{-x}$

$0 < x < 1$

$$\leq e^{-\frac{ne^{-1}}{k!}} \stackrel{?}{=} o(1)$$

$\Pr(Z = k) = \frac{e^{-\lambda}\lambda^k}{k!}$

Proof? to show this $\frac{ne^{-1}}{k!} \to \infty$

$$\ln n - 1 - \ln k! \to \infty$$

$$\underbrace{\boxed{(1-\beta)\frac{\ln n}{\ln\ln n}}}_{k}$$

$$\sim k \ln k$$

$$k\ln k \sim (1-\beta)\frac{\ln n}{\cancel{\ln\ln n}}\left[\cancel{\ln\ln n}\,(1+o(1))\right]$$

$$(1-\beta)\ln n\,(1+o(1))$$

$$\ln n - (1-\beta)\ln n \sim \underline{\beta}\ln n$$

$$k! = \sqrt{2\pi k}\;\frac{k^k}{e^k}$$

Stirling's formula

$$\ln k! = \boxed{k\ln k} - k - \ln(\sqrt{2\pi k})$$

Power of 2 choices

standard balls in balls

$n \to n$

$$E[\text{max load}] = \frac{\ln n}{\ln\ln n}$$

n bins

throw balls one at time.

each ball picks 2 random bins

. puts itself into the less loaded of 2 bins

$$E[\text{max load after } n \text{ balls}] = \Theta(\ln\ln n)$$

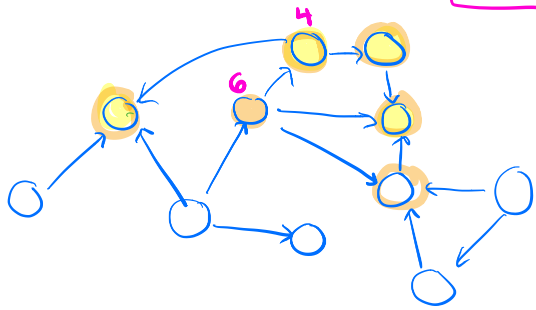
A cool application: reachability

n vertices m edges $m \geq n$.

Input: directed graph $G=(V,E)$

Output: $\forall v \in V$, $|S(v)|$ where

$S(v)$: set of nodes reachable from v



do BFS from every single node.

$O(nm)$

$\underline{O(n^{2.37...})}$ best matrix multiplic

$\underline{O(m \log^2 n)}$