Προηγούμενη φορά :

Markov's Inequality

Αν X μη-αρνητική τ.μ. $\forall t \quad P[X \geq t] \leq \frac{E[X]}{t}$

Chebyshev's Inequality

Αν X έχει $Var[X] = \sigma^2$

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

Chernoff Bound

Αν $X = \sum X_i$ με X_i ανεξάρτητες τ.μ.

στο $\{0, 1\}$ τότε

$$Pr[X \geq (1+\varepsilon)\mu] \leq e^{-\frac{\varepsilon^2 \mu}{2+\varepsilon}}$$

όπου $\mu = E[X]$

Hoeffding Inequality

Αν $X = \sum X_i$ με X_i ανεξάρτητες τμ.

στο $[a_i, b_i]$ τότε

$$\Pr[|X-\mu| \geq t] \leq 2e^{-\frac{2t}{\sum (b_i - a_i)^2}}$$

όπου $\mu = E[X]$

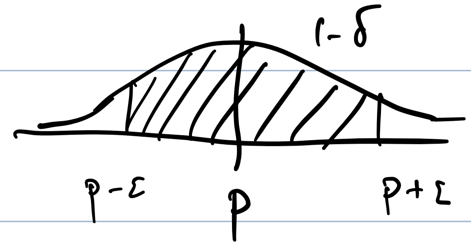
Πιθανότητα ενός κέρματος p

Αν $\hat{p} = \dfrac{\sum_{i=1}^{m} X_i}{m}$

$$\Pr[|\hat{p} - p| > \varepsilon] \leq e^{-\Omega(\varepsilon^2 m)}$$

με
$m = \Theta\left(\frac{1}{\varepsilon^2}\log(1/\delta)\right)$ δείγματα

$|\hat{p} - p| < \varepsilon$ με πιθανότητα $1-\delta$.



Εκτίμηση Πλήθους Διαφορετικών Στοιχείων

Πρόβλημα: Μεγάλο stream από δεδομένα

$x_1, \dots, x_N$

Πόσα είναι διαφορετικά;

Έστω n το πλήθος διαφορετικών

με $O(n)$ μνήμη

μπορώ να τα υπολογίσω

Τεχνική : Hash κάθε στοιχείο

$x_i \to h(x_i) \sim U[0,1]$

$h(x_1) = h(x_8)$ αν

το $x_1 = x_8$

$$Y = \min_{i=1}^{N} h(x_i) \Leftrightarrow$$

$$Y = \min_{i=1}^{n} s_i$$ όπου $s_i \sim U[0,1]$

Claim: $E[Y] = \frac{1}{n+1}$

$$n = \frac{1}{E[Y]} - 1$$

Proof. $\Pr[Y \geq t] = (1-t)^n$

$$E[Y] = \int_0^1 \Pr[Y=t]\, t \, dt$$

$$= \int_0^1 (-\Pr[Y \geq t])' \, t \, dt$$

$$= \int_0^1 \Pr[Y \geq t] \, dt = \int_0^1 (1-t)^n dt$$

$$= \int_0^1 t^n dt = \frac{1}{n+1}$$

Claim: $\mathrm{Var}[Y] \leq \frac{1}{(n+1)^2}$

$$\mathrm{Var}[Y] = E[Y^2] - E[Y]^2$$

Proof.

$$E[Y^2] = \int_0^1 \Pr[Y=t]\, t^2\, dt$$

$$= \int_0^1 n(1-t)^{n-1} t^2\, dt$$

$$= \frac{2}{(n+1)(n+2)} \le \frac{2}{(n+1)^2}$$

$$\mathrm{Var}[Y] \le \frac{2}{(n+1)^2} - \frac{1}{(n+1)^2} = \frac{1}{(n+1)^2} \quad \square$$

Από Chebyshev

$$\Pr\left[\,\left|Y - \frac{1}{n+1}\right| \geq \frac{k}{n+1}\right] \leq \frac{1}{k^2}$$

Μείωση του Variance

Επαναλαμβάνω το πείραμα t φορές

$Y_1, Y_2, \ldots, Y_t$

$$\bar{Y} = \frac{\sum_{i=1}^{t} Y_i}{t}$$

$$E[\bar{Y}] = \frac{1}{t}\sum_{i=1}^{t} E[Y_i] = \frac{1}{n+1}$$

$$Var[\bar{Y}] = \frac{1}{t^2} Var[\Sigma Y_i] = \frac{\sum_{i=1}^{t} Var[Y_i]}{t^2} = \frac{1}{t}\frac{1}{(n+1)^2}$$

$$Pr\left[\, \left|\bar{Y} - \frac{1}{n+1}\right| \geq k \frac{1}{\sqrt{t}}\frac{1}{n+1} \right] \leq \frac{1}{k^2}$$

Av $k=2$ $\frac{1}{\sqrt{\delta}}$ kal $\boxed{t = \frac{4}{\varepsilon^2}}$ $\frac{1}{\delta \varepsilon^2}$

$$Pr\left[\, \left|\bar{Y} - \frac{1}{n+1}\right| \geq \varepsilon \frac{1}{n+1} \right] \leq \frac{1}{4}$$

Άρα με πιθανότητα $\geq \frac{3}{4}$

$$\frac{1}{n+1}(1-\varepsilon) \leq \overline{Y} \leq (1+\varepsilon)\frac{1}{n+1}$$

$$\hat{n} = \frac{1}{\overline{Y}} - 1$$

$$\frac{n+1}{1+\varepsilon} \leq \underbrace{\frac{1}{\overline{Y}}}_{\hat{n}+1} \leq \frac{n+1}{1-\varepsilon}$$

[Flajolet- Martin algorithm]

Πιθανότητα $1-\delta$

Γενική τεχνική για boosting της πιθανότητας

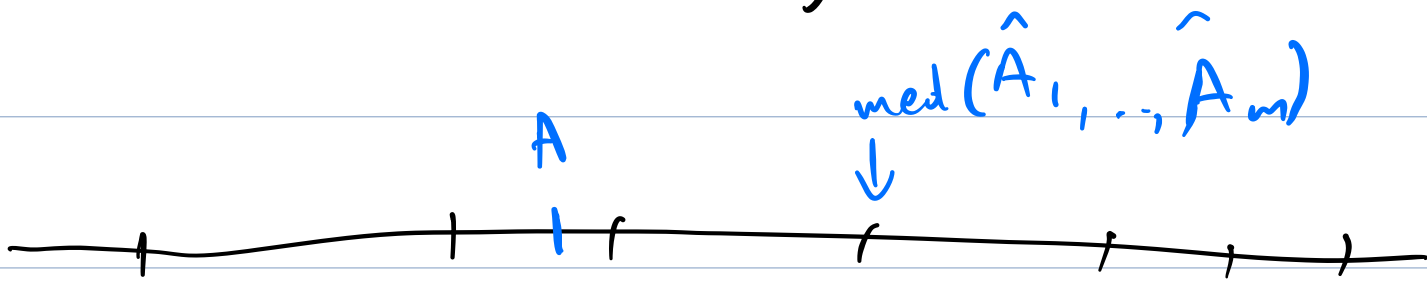
Έστω ότι έχω έναν εκτιμητή $\hat{A}$ ενός μεγέθους A ώστε

$|\hat{A} - A| \leq \varepsilon$ με πιθανότητα $\frac{3}{4}$

Μπορώ να αυξήσω την πιθανότητα επιτυχίας με ανεξάρτητα πειράματα του εκτιμητή

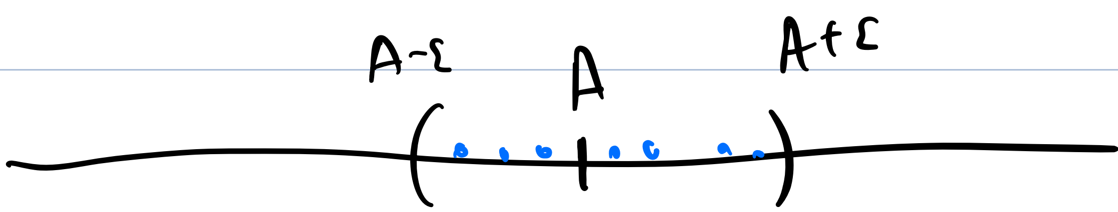
$\hat{A}_1, \hat{A}_2, \ldots, A_m$

Έστω ότι υπολογίζω το median



Claim: Αν πάνω από τα $\frac{m}{2}$ $\hat{A}_i$ ικανοποιούν $|\hat{A}_i - A| \leq \varepsilon$

τότε $|\mathrm{med}(\hat{A}_1, \ldots, \hat{A}_m) - A| \leq \varepsilon$



Ποια είναι η πιθανότητα σφάλματος;

Έστω Z_i είναι η δείκτρια τ.μ.

του γεγονότος ότι το $\hat{A}_i$

είναι κακό

$$Z_i = \begin{cases} 1 & \text{αν } |\hat{A}_i - A| > \varepsilon \\ 0 & \text{διαφορετικά} \end{cases}$$

$$E[Z_i] \leq \frac{1}{4}$$

$$\Pr\left[\sum_{i=1}^{m} Z_i \geq \frac{m}{2}\right] \leq$$

$$\Pr\left[\sum_{i=1}^{m} Z_i \geq 2\, E\left[\sum_{i=1}^{m} Z_i\right]\right]$$

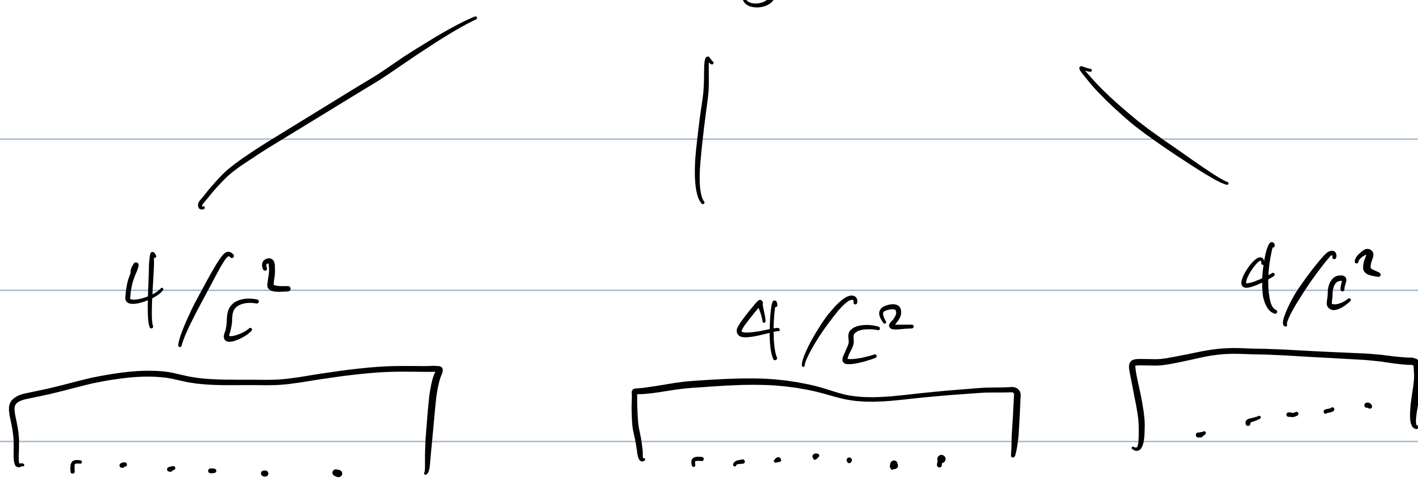
(Chernoff bound)

$$\leq e^{-\frac{\mu}{3}} = e^{-\frac{m}{12}} \leq \delta$$

$$m = O\left(\log \frac{1}{\delta}\right)$$

Συνολικά : Μνήμη $O\left(\frac{1}{\varepsilon^2} \log(1/\delta)\right)$

$\log(1/\delta)$



Median - of - Means