

Πιθανότητες και Αλγόριθμοι

Δημήτρης Φωτάκης

Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών

Εθνικό Μετσόβιο Πολυτεχνείο



Πιθανοτικοί Αλγόριθμοι

- **Πιθανοτικός αλγόριθμος** κάνει **τυχαίες επιλογές** και εξαρτά **εξέλιξή του** από αυτές.
 - **Κατανομή πιθανότητας** πάνω σε ντετερμινιστικούς αλγόριθμους.
- Πλεονεκτήματα πιθανοτικών αλγόριθμων:
 - **Απλότητα** και κομψότητα (π.χ. quickselect, primality).
 - Συνήθως **ταχύτεροι** από ντετερμινιστικούς.
 - Όταν έχουμε μερική γνώση, περιορισμένη μνήμη, κλπ., πρακτικά αποτελούν **μόνη αποδοτική λύση**.
- Μειονεκτήματα:
 - **Λάθος** απάντηση (με μικρή πιθανότητα).
 - Κυμαινόμενος **χρόνος** εκτέλεσης.
 - Δύσκολο **debugging**.

Πώς τα Καταφέρνουν;

- ❑ Εκμεταλλεύονται «εργαλεία» της πιθανότητας.
- ❑ «Αδυνατίζει» (και γίνεται πιο ρεαλιστική) η χειρότερη περίπτωση (π.χ. quicksort).
- ❑ Τυχαία δειγματοληψία: αντιπροσωπευτικό δείγμα και λύση (π.χ. clustering, sublinear algs).
- ❑ Ικανό πλήθος πιστοποιητικών (βλ. property testing).
- ❑ Τυχαία μοιρασιά εργασιών: ισορροπημένη και με ελάχιστο κόστος (υπολογιστικό, επικοινωνιακό).
- ❑ Fingerprinting και hashing.
- ❑ «Σπάσιμο» συμμετρίας (π.χ. Ethernet, leader election).
- ❑ Προσομοίωση διαδικασιών και rapid mixing.

Γινόμενο Πολυωνύμων

- Πολυώνυμα $P_1(x), P_2(x), P_3(x)$ ορισμένα σε field F .
- Έλεγχος αν $P_1(x) \times P_2(x) = P_3(x)$
 - ... σε χρόνο (σημαντικά) μικρότερο του πολλαπλασιασμού;
- Ελέγχουμε αν $Q(x) = P_1(x) \times P_2(x) - P_3(x)$ είναι (ταυτ.) 0.
 - Έστω $Q(x)$ βαθμού d και όχι (ταυτοτικά) 0.
Για κάθε $S \subseteq F$, $\Pr_{r \in S}[Q(r) = 0] \leq d/|S|$.
 - Για $|S| = 100d$ και 3 ανεξ. δείγματα, πιθαν. λάθους $\leq 10^{-6}$.
 - Χρόνος πολ/μού: $\Theta(d^2)$. Χρόνος ελέγχου: $\Theta(d)$.
- Επεκτείνεται σε πολυώνυμα **πολλών μεταβλητών**, όπου αντίστοιχη πιθανότητα ορίζεται με **συνολικό βαθμό**.
 - Θεώρημα **Schwartz-Zippel**.

Γινόμενο Πινάκων

- Δίνονται A, B, C πίνακες $n \times n$.
 - Έλεγχος αν $AB = C$ σε χρόνο $O(n^2)$.
- Τυχαίο διάνυσμα $r \in \{0, 1\}^n$. Απαντ. ΝΑΙ αν $A(Br) = Cr$.
 - Ισοδύναμα αν $Dr = 0$, όπου $D = (AB - C)$.
 - Αν $D \neq 0$, D έχει μη μηδενικά στοιχεία.
Χβτγ., κάποια στην 1^η γραμμή του D , ένα στην 1^η στήλη.
 - Για κάθε επιλογή των $r_2, \dots, r_n$,
υπάρχει μια (το πολύ) επιλογή για το r_1 τ.ω. $\sum_{j=1}^n D_{1j}r_j = 0$
 - Άρα πιθανότητα λάθους $\leq 1/2$.
 - Με π.χ. 30 ανεξάρτητες επαναλήψεις, πιθαν. λάθους $< 10^{-6}$.

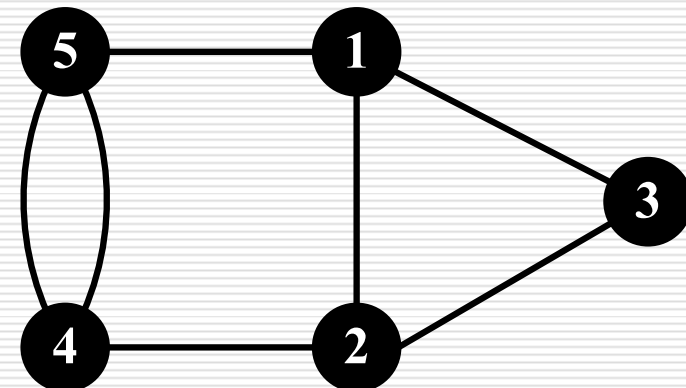
Γινόμενο Πινάκων: Εργαλείο

- Ανάλυση βασίζεται σε **αρχή αναβολής τυχαίων αποφάσεων** (principle of deferred decisions):
 - «Φιξάρουμε» μέρος των **τυχαίων** επιλογών (συνήθως σε **αυθαίρετες** τιμές).
 - Υπολογίζουμε **πιθανότητα**, δεδομένων αυτών των τιμών.
 - Τεχνικά, υπολογίζουμε την **πιθανότητα υπό συνθήκη**.
Επειδή ισχύει για αυθαίρετη συνθήκη, **ισχύει χωρίς συνθήκη**.
- Γενικότερα, έστω $E_1, \dots, E_n$ μια **διαμέριση** του δειγματοχώρου σε γεγονότα. Τότε:

$$\mathbb{Pr}[B] = \sum_{i=1}^n \mathbb{Pr}[B \cap E_i] = \sum_{i=1}^n \mathbb{Pr}[B|E_i] \mathbb{Pr}[E_i]$$

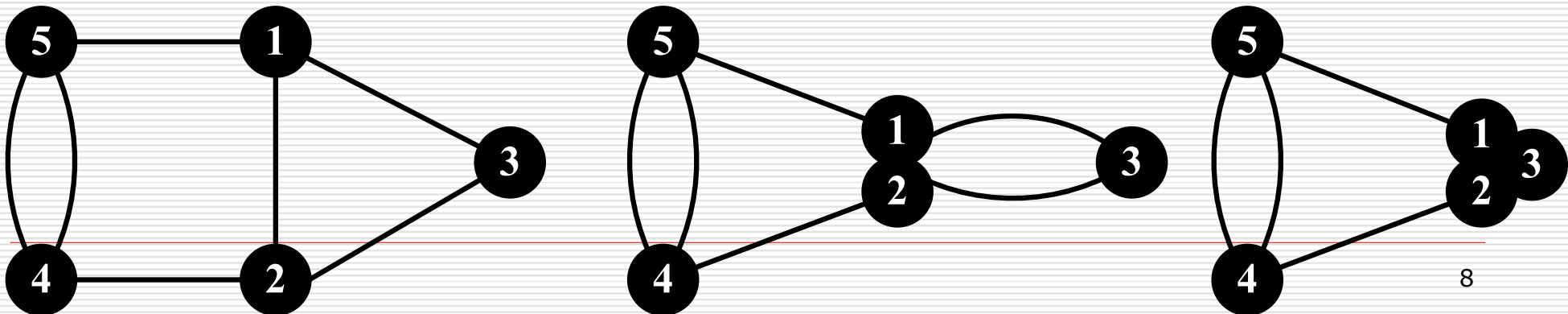
Ελάχιστη Τομή

- Μη κατευθυνόμενο συνεκτικό **πολυγράφημα** $G(V, E)$.
 - Πολλαπλές ακμές, όχι χωρητικότητες / βάρη.
- **Τομή**: διαμέριση κορυφών $(S, V \setminus S)$ με $\emptyset \neq S \subset V$.
 - Σύνολο ακμών που **αφαίρεσή** τους δημιουργεί τουλ. 2 συνεκτικές **συνιστώσες**.
 - Μέγεθος τομής $b(S, V \setminus S) = |\{\{u, v\} \in E : u \in S, v \notin S\}|$
- Πρόβλημα: υπολογισμός μιας **ελάχιστης τομής**.
 - Λύνεται σε χρόνο $O(n^4)$ με διαδοχικές εφαρμογές αλγόριθμου μέγιστης ροής.
 - Υπάρχουν εξειδικευμένοι αλγόριθμοι με χρόνο $O(n^3)$.



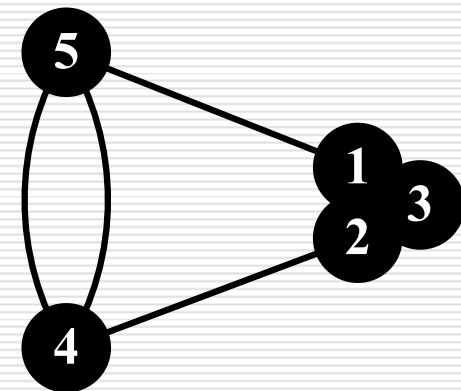
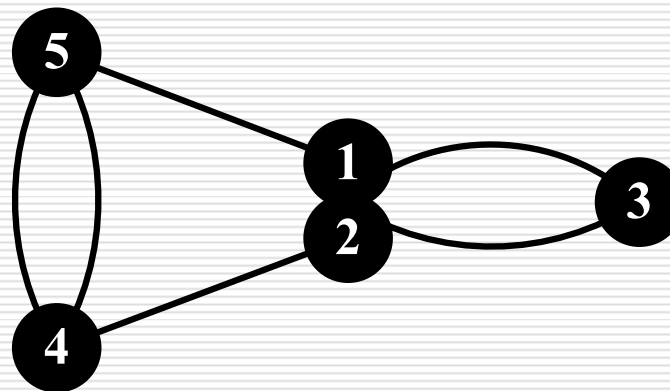
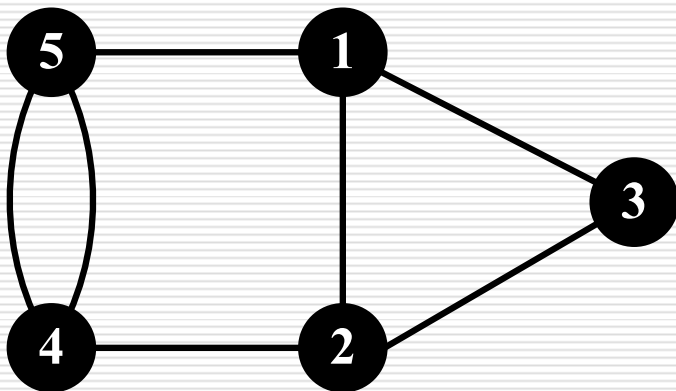
Σύμπτυξη Κορυφών

- **Σύμπτυξη** κορυφών **u και v**:
 - Αντικατάσταση u, v από μία **νέα κορυφή uv** .
 - Κάθε ακμή $\{x, u\} / \{x, v\}$ αντικαθίσταται από ακμή $\{x, uv\}$.
 - Ακμές $\{u, v\}$ παραλείπονται.
 - Διαδοχικές συμπτύξεις κορυφών 1, 2 και 12, 3.
- **Τομή** σε γράφημα **μετά από διαδοχικές συμπτύξεις** αντιστοιχεί σε **τομή σε αρχικό** γράφημα.
 - Λειτουργία σύμπτυξης **δεν** μειώνει ελάχιστη τομή.



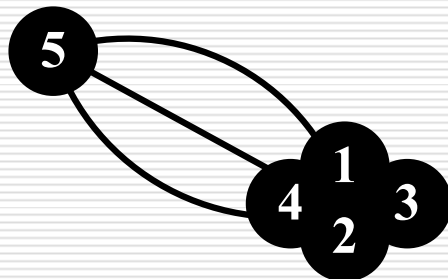
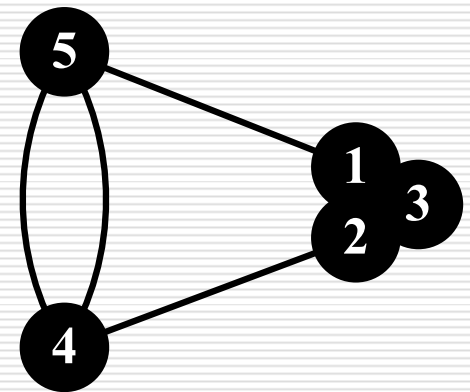
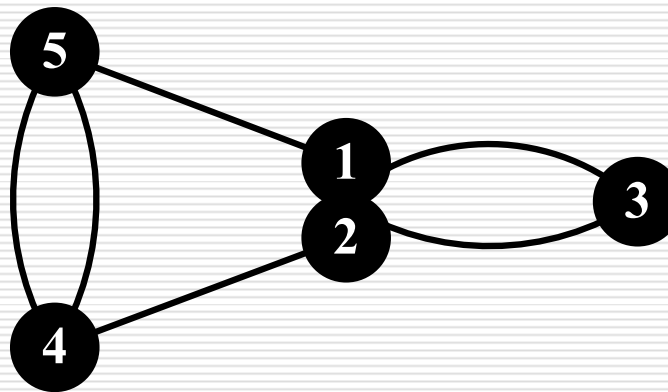
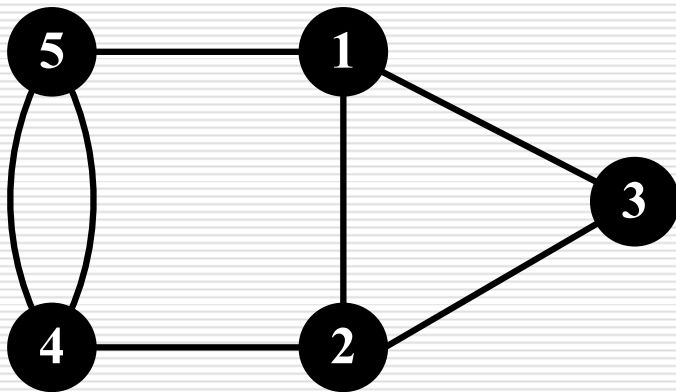
Πιθανοτικός Αλγόριθμος [Karger, 93]

- **Ενόσω** το γράφημα που απομένει έχει > 2 κορυφές:
 - Διάλεξε μια **τυχαία ακμή** $\{u, v\}$.
 - Αντικατέστησε γράφημα με αυτό που προκύπτει από **σύμπτυξη** κορυφών u και v .
- **Ακμές τομής** αυτές **μεταξύ 2 κορυφών** που απομένουν.
- **Τομή** ορίζεται από **κορυφές που συμπτύχθηκαν στις 2 κορυφές** που απομένουν.



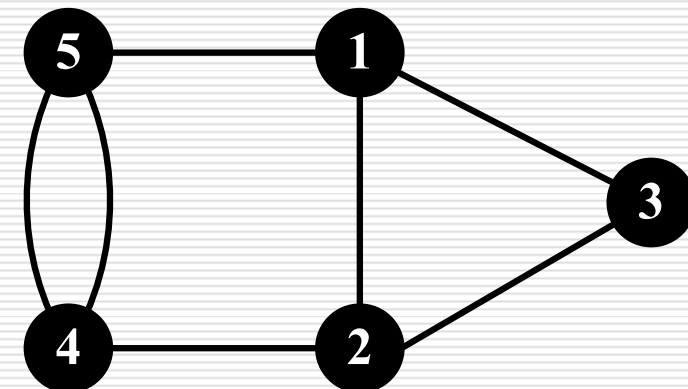
Παράδειγμα

- Αρχικές συμπτώξεις 1, 2, και 12, 3.
 - Σύμπτυξη 123, 4.
 - Σύμπτυξη 5, 4.



Πιθανοτικός Αλγόριθμος [Karger, 93]

- Βασικές ιδιότητες:
 - Πάντα **τερματίζει** έπειτα από $n - 2$ συμπτώξεις.
 - Υπολογίζει μία τομή, μπορεί **όχι** ελάχιστη.
 - Ποια πιθανότητα p να καταλήξει σε ελάχιστη τομή;
 - Αν p όχι αμελητέα, **μεγαλώνει γρήγορα με επαναλήψεις**.
 - Αν $p \geq 2/n^2$, πιθανότητα τουλ. μία από $n^2 \ln n$ επαναλήψεις να καταλήξει σε ελάχιστη τομή $\geq 1 - 1/n^2$.
- Έστω ελάχιστη τομή $C = \{e_1, \dots, e_k\}$ μεγέθους k .
 - Αλγ. **επιστρέφει C** ανν **καμία από ακμές C δεν επιλεγεί για σύμπτυξη**.



Πιθανότητα Επιτυχίας

- Συγκεκριμένη ελάχιστη τομή $C = \{e_1, \dots, e_k\}$ μεγέθους k .
 - Πιθανότητα **καμία** από ακμές C **δεν** επιλέγεται για σύμπτυξη.
 - Ελάχιστος βαθμός κορυφής $\geq$ ελάχιστη τομή.
 - $G(V, E)$ έχει **ελάχιστο βαθμό** κορυφής $\geq k$.
 - G έχει **#ακμών** $\geq nk/2$.
 - Πιθανότητα **δεν** επιλέγεται ακμή του C στην **1^η** σύμπτυξη:
$$p_1 \geq \frac{\frac{nk}{2} - k}{\frac{nk}{2}} = \frac{n-2}{n}$$
 - Μετά από t συμπτώξεις, γράφημα έχει **ελάχιστο βαθμό** $\geq k$.
 - **#ακμών** $\geq (n-t)k/2$.
 - Πιθανότητα **δεν** επιλέγεται ακμή C του **ούτε** στην **$(t+1)^{\text{η}}$** σύμπτυξη:
$$p_{t+1} \geq \frac{\frac{(n-t)k}{2} - k}{\frac{(n-t)k}{2}} = \frac{n-t-2}{n-t}$$

Πιθανότητα Επιτυχίας

- Συγκεκριμένη ελάχιστη τομή $C = \{e_1, \dots, e_k\}$ μεγέθους k .
 - Πιθανότητα **καμία** από ακμές C **δεν επιλέγεται** για σύμπτυξη:

$$p = p_1 \cdot p_2 \cdots p_{n-2} \geq \frac{n-2}{n} \cdot \frac{n-3}{n-1} \cdot \frac{n-4}{n-2} \cdots \frac{2}{4} \cdot \frac{1}{3} = \frac{2}{n(n-1)}$$

- Άρα $p \geq 2/n^2$, και πιθανότητα τουλ. **μία** από $n^2 \log n$ επαναλήψεις να καταλήξει σε **ελάχιστη τομή** $\geq 1 - 1/n^2$.
 - Χρόνος εκτέλεσης $O(n^2)$ / επανάληψη.
 - Συνολικός χρόνος $O(n^4 \log n)$.

Χρόνος Εκτέλεσης

- Όμως (σχετικά) μικρή πιθανότητα αποτυχίας στις πρώτες μισές συμπτώξεις!
 - Π.χ. πιθανότητα να μην συμπτυχθεί καμία ακμή C στις πρώτες $(n-3)/2$ συμπτώξεις $\geq 1/4$.
 - «Ακριβές» συμπτώξεις είναι «επιτυχημένες».
- Αναδρομική υλοποίηση σε φάσεις:
 - Εκτέλεση βασικού αλγόριθμου για $n/2$ συμπτώξεις 4 φορές.
 - Συνεχίσουμε αναδρομικά για καθένα από τα αποτελέσματα.
- Χρόνος εκτέλεσης $O(n^2 \log^3 n)$ για πιθανότητα επιτυχίας $= 1 - O(1/n)$.

Monte Carlo vs Las Vegas

- Monte Carlo αλγόριθμοι (π.χ. min-cut, max-cut):
 - Μπορεί να δώσουν **λάθος απάντηση** (με μικρή πιθανότητα), χρόνος εκτέλεσης **ντετερμινιστικός** (συνήθως!).
 - Πιθανότητα λάθους μπορεί να γίνει **πολύ-πολύ μικρή** με ανεξάρτητες επαναλήψεις.
 - Προβλήματα απόφασης: **one-sided error** και **two-sided error**.
 - Πολυωνυμικοί one-sided error αλγόριθμοι: **RP** και **coRP**.
 - Πολυωνυμικοί two-sided error αλγόριθμοι: **BPP**.
- Las Vegas αλγόριθμοι (π.χ. quicksort, quickselect):
 - **Πάντα σωστή απάντηση**, χρόνος εκτέλεσης **τυχαία μεταβλητή**.
 - Πολυωνυμικοί αλγόριθμοι: **ZPP**.

Βασικές Έννοιες Πιθανότητας

- Δειγματοχώρος, γεγονός και πιθανότητα, τυχαία μεταβλητή.
 - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$
(γενικεύεται με μέθοδο εγκλεισμού – αποκλεισμού).
 - Union bound: $\Pr[\cup_{i=1}^n A_i] \leq \sum_{i=1}^n \Pr[A_i]$
 - Πιθανότητα A υπό συνθήκη B: $\Pr[A|B] = \Pr[A \cap B] / \Pr[B]$
Γενίκευση: $\Pr[\cap_{i=1}^n A_i] = \Pr[A_1] \Pr[A_2|A_1] \cdots \Pr[A_n | \cap_{i=1}^{n-1} A_i]$
 - Ανεξάρτητα γεγονότα: $\Pr[A \cap B] = \Pr[A] \Pr[B]$.
 - Αρνητικά σχετιζόμενα γεγονότα.

Βασικές Έννοιες Πιθανότητας

- Μέση τιμή: $\mathbb{E}[X] = \sum_{k=0}^{\infty} \mathbb{Pr}[X = k] k$
 - Ισοδύναμα (ακέραιες τυχαίες μεταβλ.): $\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{Pr}[X \geq k]$
 - Γραμμικότητα: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
 - Ανισότητα Jensen: Αν f κυρτή συνάρτηση, $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.
 - Αν X και Y ανεξάρτητες: $\mathbb{E}[X Y] = \mathbb{E}[X] \mathbb{E}[Y]$.
- Διακύμανση (variance):
 - $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
 - Τυπική απόκλιση (std deviation): $\sigma_X = \text{Var}(X)^{1/2}$
 - Αν X και Y ανεξάρτητες: $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$.
- Probability generating function:

$$G_X(z) = \sum_{k=0}^{\infty} \mathbb{Pr}[X = k] z^k$$

$$\begin{aligned}\mathbb{E}[X] &= G'(1) \\ \text{Var}[X] &= G''(1) + G'(1) - G'(1)^2\end{aligned}$$

Παραδείγματα Κατανομών

- Bernoulli μεταβλητή X : 1 με **πιθ. p** , και 0 διαφ.
 - $E[X] = p$, $\text{Var}[X] = p(1 - p)$, $G_X(z) = 1 - p + pz$.
- Δυωνυμική κατανομή **$\text{Pr}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$**
 $\text{Bin}(n, p)$:
 - **Αριθμός επιτυχιών** σε n «ρίψεις» με πιθανότητα επιτυχίας p .
 - Άθροισμα n Bernoulli μεταβλητών με παράμετρο p .
 - $E[X] = np$, $\text{Var}[X] = np(1 - p)$, $G_X(z) = (1 - p + pz)^n$
- Γεωμετρική κατανομή **$\text{Geo}(p)$: $\text{Pr}[X = k] = (1 - p)^{k-1} p$**
 - **Αριθμός «ρίψεων» μέχρι την πρώτη επιτυχία** (waiting time).
 - $E[X] = 1/p$, $\text{Var}[X] = (1 - p)/p^2$, $G_X(z) = pz / (1 - z + pz)$
 - Αμνησία: **$\text{Pr}[X = n+k \mid X > k] = \text{Pr}[X = n]$**

Μπάλες και Κουτιά

- Έχουμε m μπάλες και n κουτιά. Κάθε μπάλα επιλέγει το κουτί της ισοπίθανα και ανεξάρτητα.
 - Απλό μοντέλο, πλήθος εφαρμογών(!).
 - Μέγιστος #μπαλών σε κάποιο κουτί;
 - Load balancing. Hashing with chains.
 - Ελάχιστο m ώστε να εμφανιστεί κουτί με ≥ 2 μπάλες;
 - Birthday paradox.
 - Ελάχιστο m ώστε κανένα κουτί άδειο;
 - Coupon collecting.

Μέγιστος #Μπαλών

□ Πιθανότητα να βρεθεί κουτί με $\geq 3 \ln n / \ln \ln n$ μπάλες είναι $\leq 1/n$.

■ $L_i = \# \text{μπαλών σε κουτί } i: \mathbb{P}[L_i \geq k] \leq \binom{n}{k} \left(\frac{1}{n}\right)^k \quad k! \geq (k/e)^k$

$$\leq \frac{n^k e^k}{k^k n^k} = \left(\frac{e}{k}\right)^k$$

■ Συνεπώς $\mathbb{P}\left[L_i \geq \frac{3 \ln n}{\ln \ln n}\right] \leq n^{-2}$

■ ... και (από union bound) $\mathbb{P}\left[\exists i : L_i \geq \frac{3 \ln n}{\ln \ln n}\right] \leq \frac{n}{n^2} = \frac{1}{n}$

■ Πιο ακριβής ανάλυση είναι εφικτή [Gonnet].

□ Νδο με πιθανότητα $\geq 1/n$, υπάρχει κουτί με $\Omega(\ln n / \ln \ln n)$.

Δυο Μπάλες στο Ίδιο Κουτί

- Πιθανότητα όλες οι m ($< n$) μπάλες σε διαφορετικό κουτί:

$$P_m = \frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \dots \frac{n-m+1}{n} = \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right)$$
$$\leq \prod_{k=1}^{m-1} e^{-k/n} = e^{-m(m-1)/(2n)}$$

- Πιθανότητα τουλ. 2 μπάλες στο ίδιο κουτί $\geq 1 - P_m$
 - Για $n = 365$ και $m = 28$: πιθανότητα σε 28 ανθρώπους, κάποιοι να έχουν γενέθλια την ίδια μέρα $> 1 - e^{-1}$

Συλλογή Κουπονιών

- Ελάχιστο m ώστε κανένα κουτί άδειο.
 - $Z_k = \# \text{μπαλών όταν για πρώτη φορά } \# \text{γεμάτων κουτιών} = k.$
 - $X_k = Z_{k+1} - Z_k$: $\# \text{μπαλών για να γεμίσει το } k+1 \text{ κουτί.}$
 - X_k ακολουθεί **γεωμετρική κατανομή** με παράμετρο $1 - k/n$, και έχει $E[X_k] = n/(n - k).$
 - Γραμμικότητα μέσης τιμής:
$$\mathbb{E}[Z_n] = \sum_{k=0}^{n-1} \mathbb{E}[X_k] = \sum_{k=0}^{n-1} \frac{n}{n - k} = nH_n$$
- Εμφανίζει **ισχυρή συγκέντρωση** γύρω από την μέση τιμή:
 - $Y_{j,k}$: κουτί j είναι άδειο μετά τις πρώτες k μπάλες.
$$\Pr[Y_{j,k}] = \left(1 - \frac{1}{n}\right)^k \leq e^{-k/n}$$
 - Για κάθε $\beta > 1$, πιθανότητα κάποιο κουτί άδειο μετά από $\beta n \ln n$ μπάλες:
$$\leq n e^{-\beta \ln n / n} = n^{1-\beta}$$
 - Μπορεί ν.δ.ο. για κάθε c , πιθανότητα κάποιο κουτί άδειο μετά από $n(\ln n + c)$ μπάλες:
$$\leq e^{-e^{-c}}$$

Συγκέντρωση στη Μέση Τιμή

- (Πραγματική τιμή) «ομαλών» συναρτήσεων μεγάλου αριθμού ανεξάρτητων τυχαίων μεταβλητών «κινείται» σε ένα μικρό διάστημα γύρω από την μέση τιμή.
 - Βλ. [Dubhashi and Panconessi, Concentration of Measure for the Analysis of Randomized Algorithms, 2007].
- **Ανισότητα Markov** (γενική, αλλά όχι ιδιαίτερα ισχυρή):
 - X μη-αρνητική τυχαία μεταβλητή. $\Pr[X \geq t \mathbb{E}[X]] \leq 1/t$
Για κάθε $t > 0$, $\Pr[X \geq t] \leq \mathbb{E}[X]/t$
- **Ανισότητα Chebysev** (γενική, ισχυρότερη):
 - Για κάθε $t > 0$, $\Pr[|X - \mathbb{E}[X]| \geq t \sigma_X] \leq 1/t^2$
 - Απόδειξη εύκολα από ορισμό $\text{Var}[X]$ και ανισότητα Markov.

Chernoff Bounds

$$\forall \varepsilon \in (0, 0.7), \quad \frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}} \leq 1 - \frac{\varepsilon^2}{e}$$

- Έστω $X_1, \dots, X_n$ **ανεξάρτητες** Bernoulli τ.μ. με $E[X_k] = p_k$, $X = X_1 + \dots + X_n$, και $E[X] = \mu$. Για κάθε $\varepsilon > 0$,

$$\Pr[X > (1 + \varepsilon)\mu] \leq \left[\frac{e^\varepsilon}{(1 + \varepsilon)^{1+\varepsilon}} \right]^\mu$$

- Για κάθε $t > 0$, και χρησιμοποιώντας **ανισότητα Markov**:

$$\Pr[X > (1 + \varepsilon)\mu] = \Pr[e^{tX} > e^{t(1+\varepsilon)\mu}] \leq \mathbb{E}[e^{tX}] / e^{t(1+\varepsilon)\mu}$$

$$\mathbb{E}[e^{tX}] = \prod_{k=1}^n \mathbb{E}[e^{tX_k}] \leq \prod_{k=1}^n e^{p_k(e^t - 1)} = e^{(e^t - 1)\mu}$$

$$\Pr[X > (1 + \varepsilon)\mu] \leq \left[\frac{e^{e^t - 1}}{e^{t(1+\varepsilon)}} \right]^\mu$$

$$\stackrel{t=\ln(1+\varepsilon)}{\Rightarrow} \Pr[X > (1 + \varepsilon)\mu] \leq \left[\frac{e^\varepsilon}{(1 + \varepsilon)^{(1+\varepsilon)}} \right]^\mu$$

Chernoff Bounds

- Έστω $X_1, \dots, X_n$ **ανεξάρτητες** Bernoulli τ.μ.,
 $X = X_1 + \dots + X_n$, και $E[X] = \mu$.
 - Για κάθε $1 \geq \varepsilon \geq 0$, $\Pr[X > (1 + \varepsilon)\mu] \leq e^{-\varepsilon^2 \mu / 3}$
 $\Pr[X < (1 - \varepsilon)\mu] \leq e^{-\varepsilon^2 \mu / 2}$
 - Εξαιρετικά ισχυρή συγκέντρωση γύρω από την μέση τιμή!
 - Απαιτούν σύγκριση X με **λογαριθμική ποσότητα** για να «δουλέψουν καλά».
 - Αντίστοιχα φράγματα για τ.μ. X_k με πεδίο τιμών το $[0, w_k]$.
 - Απαιτούν **ανεξαρτησία** (ή αρνητική εξάρτηση).
 - Αντίστοιχα bounds για τ.μ. με **περιορισμένη εξάρτηση**.
 - Πολύ σημαντικά για την ανάλυση πιθανοτικών αλγόριθμων.

Παραδείγματα

□ Αν μοιράζουμε $m = n \ln n$ μπάλες σε n κουτιά, πιθανότητα προκύψει κουτί με $> 3 \ln n$ μπάλες είναι $\leq 1/n$.

□ Set Balancing:

■ $A_1, \dots, A_n$ υποσύνολα U , $|U| = n$ και για κάθε j , $|A_j| = n/2$.

■ Ζητείται διαμέριση U σε B και W που ελαχιστοποιεί:

$$\max_j ||A_j \cap B| - |A_j \cap W||$$

■ Τυχαία διαμέριση B, W : $\max_j ||A_j \cap B| - |A_j \cap W|| \leq 3\sqrt{n \ln n}$ με πιθανότητα $\geq 1 - 2/n^2$.

■ Για κάθε j , $X_j = |A_j \cap W|$ με $E[X_j] = n/4$. Έχουμε:

$$\Pr \left[||A_j \cap B| - |A_j \cap W|| > 3\sqrt{n \ln n} \right] = \Pr \left[|n/2 - 2|A_j \cap W|| > 3\sqrt{n \ln n} \right]$$

$$= \Pr \left[|\mathbb{E}[X_j] - X_j| > \frac{3}{2}\sqrt{n \ln n} \right]$$

$$\underbrace{\frac{\mathbb{E}[X_j]}{n}}_{\frac{1}{4}} \cdot \underbrace{\frac{6\sqrt{\ln n}}{\sqrt{n}}}_{\varepsilon} = \frac{3}{2}\sqrt{n \ln n}$$

$$\leq 2e^{-\frac{n}{12} \left(\frac{6\sqrt{\ln n}}{\sqrt{n}} \right)^2} = 2/n^3$$

Τυχαία Δειγματοληψία

- Σύνολο A , $|A| = n$, (άγνωστου μεγέθους) σύνολο $X \subseteq A$ με στοιχεία A που έχουν κάποια ιδιότητα.
 - Έστω $|X| = p n$. Θα υπολογίσουμε εκτίμηση p' για p .
 - Επιλέγουμε «δείγμα» A' , $|A'| \geq 3 \ln(2/\delta)/\varepsilon^2$, και υπολογίζουμε $p' = |A' \cap X|$.
 - Με πιθανότητα $\geq 1 - \delta$, εκτίμηση $p' \in [p - \varepsilon, p + \varepsilon]$.
- Σύνολο A , $|A| = n$, με διαμέριση $A_1, \dots, A_k$, $|A_j| = a_j n \geq \gamma n$, που ορίζεται από κάποιες ιδιότητες (π.χ. τι ψηφίζουν).
 - Θα εκτιμήσουμε όλα τα a_j γνωρίζοντας μόνο ότι είναι $\geq \gamma$.
 - Επιλέγουμε «δείγμα» B , $|B| \geq 3 \ln(2/(\delta \gamma))/(\gamma \varepsilon^2)$.
 - Έστω $B_j = A_j \cap B$ και $\beta_j = |B_j|/|B|$.
 - Με πιθανότητα $\geq 1 - \delta$, για όλα τα A_j , $(1 - \varepsilon) a_j \leq \beta_j \leq (1 + \varepsilon) a_j$