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 $P_1(x)$ και $P_2(x)$ βαθμού d, και πολυώνυμο $P_3(x)$ βαθμού $2d$, όλα 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 $2d$ και όχι (ταυτοτικά) 0.
    Τότε, $Pr_{r \in F}[Q(r) = 0] \leq 2d/|F|$.
  - Για $|F| = 200d$ και 3 ανεξ. δείγματα, πιθαν. λάθους $\leq 10^{-6}$.
  - Χρόνος πολ/μού: $\Theta(d^2)$ (ή $\Theta(d \log d)$). 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. NAI ανν $A(Br) = Cr$.
  - Ισοδύναμα ανν $Dr = 0$, όπου $D = (AB - C)$.
  - Αν $D \neq 0$, D έχει μη μηδενικά στοιχεία.
    Χβτγ., κάποια στην 1η γραμμή του D, ένα στην 1η στήλη.
  - Για κάθε επιλογή των $r_2, \ldots, r_n$,
    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, πιθ. λάθους $< 10^{-6}$.

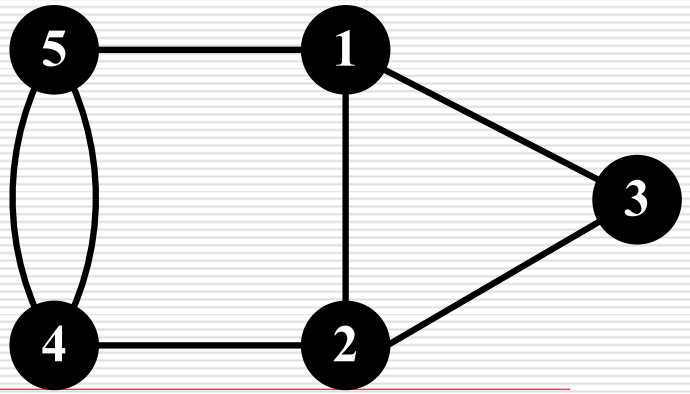
Γινόμενο Πινάκων: Εργαλείο

- 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, .., E_n μια διαμέριση του δειγματοχώρου σε γεγονότα. Τότε:

$$\mathbb{P}\mathrm{r}[B] = \sum_{i=1}^{n} \mathbb{P}\mathrm{r}[B \cap E_i] = \sum_{i=1}^{n} \mathbb{P}\mathrm{r}[B|E_i]\,\mathbb{P}\mathrm{r}[E_i]$$

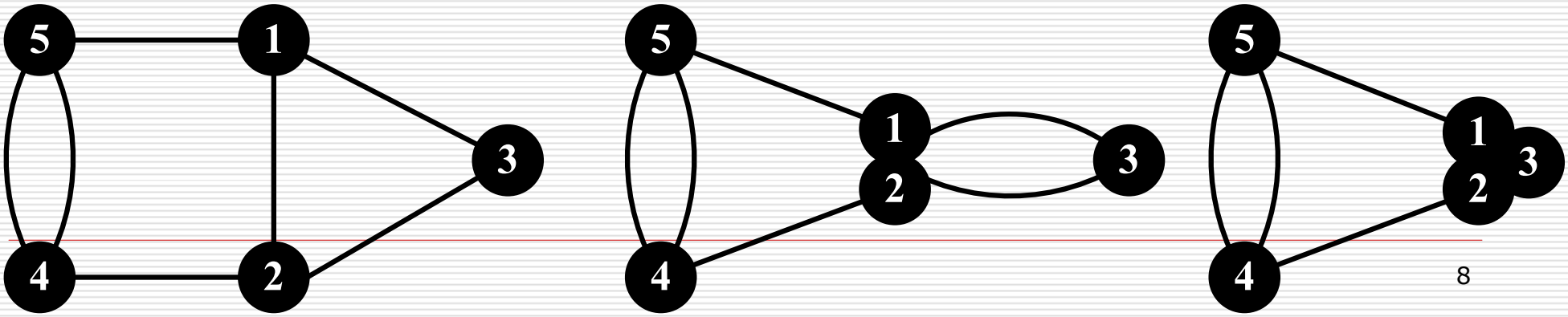
Ελάχιστη Τομή

- Μη κατευθυνόμενο συνεκτικό πολυγράφημα G(V, E).
  - Πολλαπλές ακμές, όχι χωρητικότητες / βάρη.
- Τομή: διαμέριση κορυφών (S, V \ 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 $O(n^3)$.



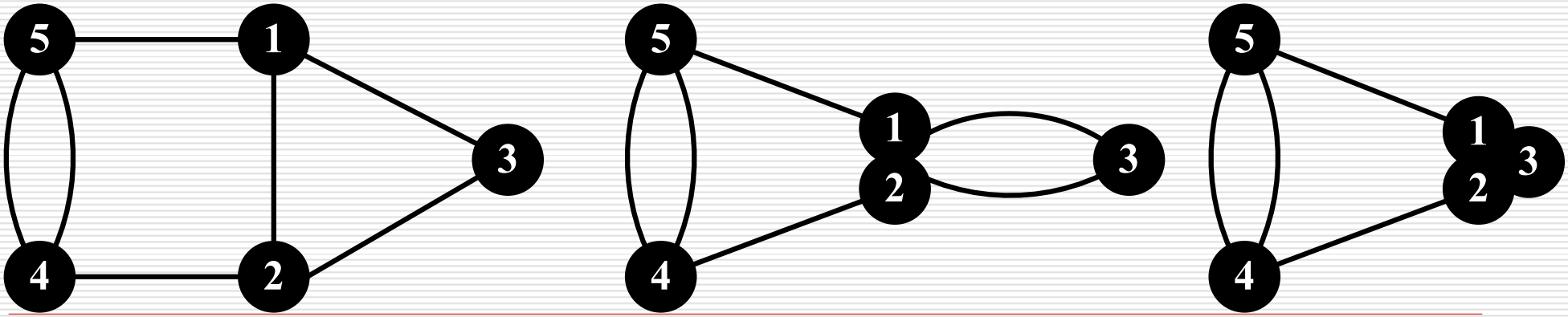
Σύμπτυξη Κορυφών

- **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.



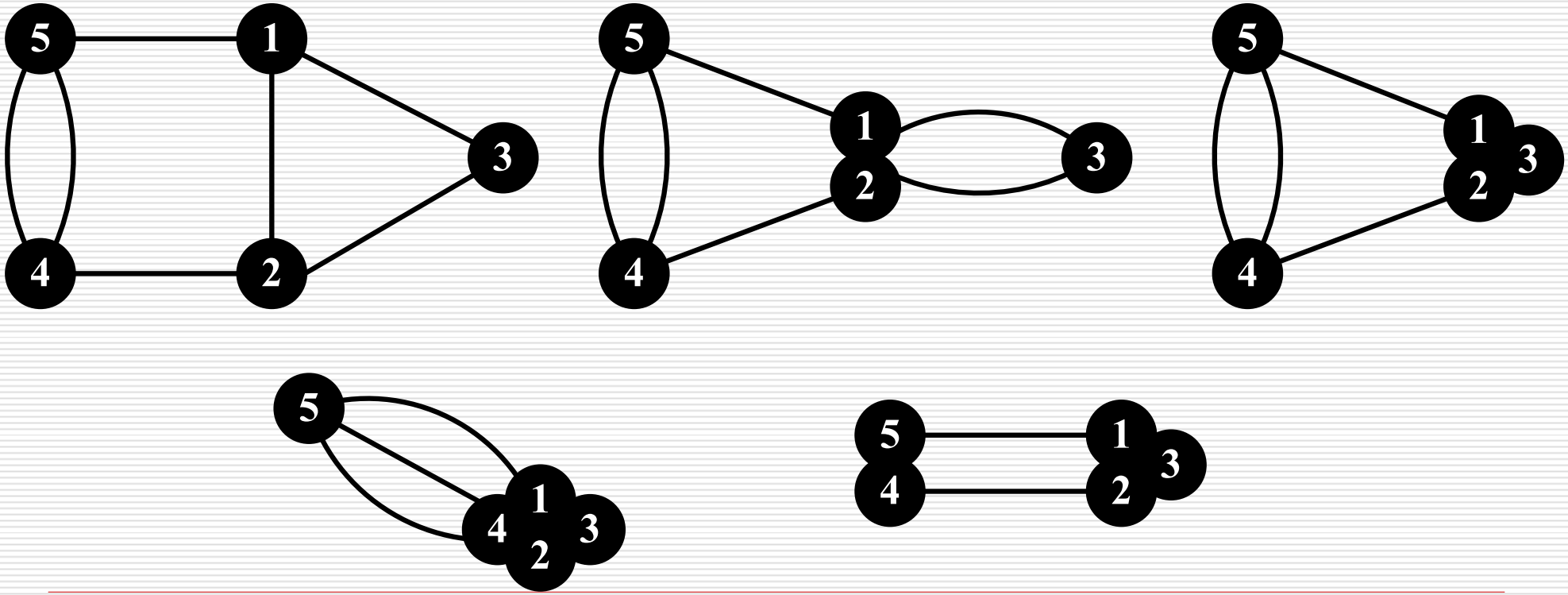
Πιθανοτικός Αλγόριθμος [Karger, 93]

- 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.



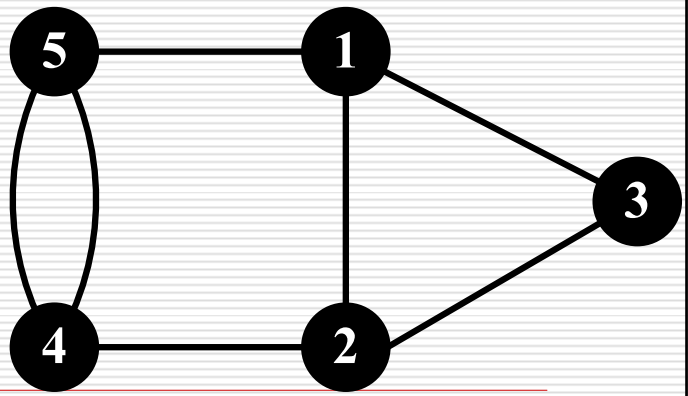
Παράδειγμα

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



Πιθανοτικός Αλγόριθμος [Karger, 93]

- Βασικές ιδιότητες:
  - Πάντα τερματίζει έπειτα από $n - 2$ συμπτύξεις.
  - 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 $C = \{e_1, \ldots, e_k\}$ μεγέθους k.
  - Αλγ. επιστρέφει C ανν καμία από ακμές C δεν επιλεγεί για σύμπτυξη.



Πιθανότητα Επιτυχίας

- Συγκεκριμένη ελάχιστη τομή $C = \{e_1, \ldots, e_k\}$ μεγέθους k.
  - Πιθανότητα καμία από ακμές C **δεν** επιλέγεται για σύμπτυξη.
  - Ελάχιστος βαθμός κορυφής ≥ ελάχιστη τομή.
  - G(V, E) έχει ελάχιστο βαθμό κορυφής ≥ k.
    - G έχει #ακμών ≥ $nk/2$.
    - Πιθανότητα δεν επιλέγεται ακμή του C στην 1η σύμπτυξη:

$$p_1 \geq \frac{\frac{nk}{2} - k}{\frac{nk}{2}} = \frac{n-2}{n}$$

  - Μετά από t συμπτύξεις, γράφημα έχει ελάχιστο βαθμό ≥ k.
    - #ακμών ≥ $(n-t)k/2$.
    - Πιθανότητα δεν επιλέγεται ακμή C του ούτε στην $(t+1)$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 $C = \{e_1, \ldots, e_k\}$ 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 συμπτύξεις!
  - Π.χ. πιθανότητα να μην συμπτυχθεί καμία ακμή C στις πρώτες (n–3)/2 συμπτύξεις $\geq 1/4$.
  - «Ακριβές» συμπτύξεις είναι «επιτυχημένες».
- Αναδρομική υλοποίηση σε φάσεις:
  - Εκτέλεση βασικού αλγόριθμου για n/2 συμπτύξεις 4 φορές.
  - Συνεχίσουμε αναδρομικά για καθένα από τα αποτελέσματα.
  - Χρόνος εκτέλεσης: $O(n^2 \log n)$ ($O(\log n)$ επίπεδα, $O(n^2)$ / επίπεδο)
  - Έστω $P(n)$ πιθανότητα επιτυχίας για γράφημα n κορυφών:
    - $P(n) = 1 - (1 - P(n)/4)^4$, με λύση $P(n) = \Omega(1/\log n)$
- Χρόνος εκτέλεσης συνολικά $O(n^2 \log^3 n)$ για πιθανότητα επιτυχίας $= 1 - O(1/n)$.

Monte Carlo vs Las Vegas

- Monte Carlo αλγόριθμοι (π.χ. min-cut):
  - 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 Α υπό συνθήκη Β: $\Pr[A|B] = \Pr[A \cap B]/\Pr[B]$
    Γενίκευση: $\Pr[\cap_{i=1}^{n} A_i] = \Pr[A_1] \Pr[A_2|A_1] \cdots \Pr[A_n| \cap_{i=1}^{n} A_i]$
  - Ανεξάρτητα γεγονότα: $\Pr[A \cap B] = \Pr[A] \Pr[B]$.
  - Αρνητικά σχετιζόμενα γεγονότα.

Βασικές Έννοιες Πιθανότητας

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

$$G_X(z) = \sum_{k=0}^{\infty} \mathbb{P}\mathrm{r}[X = k]\, z^k$$

$$\mathbb{E}[X] = G'(1)$$
$$\mathrm{Var}[X] = G''(1) + G'(1) - G'(1)^2$$

Παραδείγματα Κατανομών

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

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

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

  - Συνεπώς $\mathbb{Pr}\left[L_i \geq \frac{3\ln n}{\ln\ln n}\right] \leq n^{-2}$
  - ... και (από union bound) $\mathbb{Pr}\left[\exists i : L_i \geq \frac{3\ln n}{\ln\ln n}\right] \leq \frac{n}{n^2} = \frac{1}{n}$
  - Πιο ακριβής ανάλυση είναι εφικτή [Gonnet].
- Ndo με πιθανότητα $\geq 1 - 1/n$, υπάρχει κουτί με $\Omega(\ln n / \ln\ln n)$ μπάλες.

Δυο Μπάλες στο Ίδιο Κουτί

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

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

- 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 = #μπαλών όταν για πρώτη φορά #γεμάτων κουτιών = k.
  - $X_k = Z_{k+1} - Z_k$: #μπαλών για να γεμίσει το k+1 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. $\mathbb{Pr}[Y_{j,k}] = \left(1 - \frac{1}{n}\right)^k \leq e^{-k/n}$
  - Για κάθε β > 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: $\leq e^{-e^{-c}}$