1 Probability and random variables

https://stella.uniba.sk/texty/FMFI_HF_zaklady_pravdepodobnosti.pdf

- Kap. 1 a 2: pravdepodobnostný model, udalosti, podmienená pravdepodobnosť, nezávislé udalosti.
- Kap. 3: náhodné premenné.
- Kap. 4: diskrétna náhodná premenná (DRV), probability mass function (PMF).
- **U:** Nájdite všetky nezáporné DRV X také, že $E(X) = 0$.
- Môže nezáporná DRV mať strednú hodnotu ľubovoľne blízku nule? (Riešiť osobitne pre konečný a nekonečný support.)
- **U:** Nájdite DRV, ktorá nemá konečnú strednú hodnotu.
- Nájdite DRV X, ktorá má konečnú strednú hodnotu, a funkciu g takú, že $g(X)$ nemá konečnú strednú hodnotu. Existuje funkcia g ku každej X?
- Veta 4.12: linearita strednej hodnoty.
- Def. 4.15: rozptyl, disperzia, variancia.
- Veta 4.20: lineárna transformácia rozptylu.
- **U:** Vypočítajte E pre DRV s geometrickým rozdelením. **DU:** aj D.

$$E(X) = \sum_{n=1}^{\infty} nP[X = n] = \sum_{n=1}^{\infty} np(1-p)^{n-1} = p/p^2 = 1/p, \qquad \text{using } \sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$$

$$E\left(X^2\right) = \sum_{n=1}^{\infty} n^2 P[X = n] = \frac{2-p}{p^2}, \qquad \text{using } \sum_{n=1}^{\infty} n^2 x^{n-1} = \frac{1+x}{(1-x)^3}$$

$$D(X) = E\left(X^2\right) - (E(X))^2 = (1-p)/p^2$$

- **U:** Nájdite DRV, ktorá má E, ale nemá D.
- Kap. 4.2: pozrite si bežné diskrétne rozdelenia (DÚ).
- Kap. 6: nezávislosť náhodných premenných, iid, vety 6.10 a 6.14 o E a D nezávislých DRV.
- **DU:** For two independent DRVs X and Y, prove that $D(X - Y) = D(X) + D(Y)$.

2 Markov and Chebyshev inequalities

Markov inequality. $P[X \geq a] \leq \dfrac{E(X)}{a}$ (*non-negative* DRV X)

- Proof: if large values have high probability, they sum to too much in $E(X)$ calculation.
- Carefully examine the proof to determine all cases where equality holds.

- If $E(X)$ is all we know about X, Markov bound is best possible.
- In general, the bound is very loose except in cases near equality. Especially for light-tailed distributions: the Markov bound falls only linearly, while probability can fall exponentially or even faster.
- **DU:** Compare Markov's bound with true probability for geometric distribution.
- **U:** Suppose the expected runtime of QuickSort is $2n\log(n)$ operations. Use Markov's inequality to bound the probability that QuickSort runs for longer than $20n\log(n)$.

  Solution: $P \leq 2n\log n/20n\log n = 0.10$. Nice: no assumption about distribution of running times.

Chebyshev inequality. $P[|X - E(X)| \geq a] \leq \dfrac{D(X)}{a^2}$ (any DRV X that has finite $D(X)$)

- Proof: apply Markov inequality to DRV $Z = (X - E(X))^2$.
- **DU:** Carefully examine the proof to determine all cases where equality holds (a distribution with only 3 values with non-zero probability).
- **U:** Prove the classical form of Chebyshev inequality (using the variance form given above): $P[|X - E(X)| \geq k\sigma] \leq \frac{1}{k^2}$ where σ is standard deviation, i.e. $D(X) = \sigma^2$.
- Assume X is a DRV with $E(X) = 10$, $D(X) = 25$.
  U: What does the Chebyshev inequality tell us about $P[X \geq c]$ for $c \in \{12, 16, 30\}$?

  Solution: Since $[X \geq c] \subseteq [|X - 10| \geq c - 10]$, we have $P[X \geq c] \leq 25/(c-10)^2$. Only useful for $c = 30$. **DU:** Construct a DRV X such that $E(X) = 10$, $D(X) = 25$, and $P[X \geq 12] = 0$.
- Let's prove Cantelli's inequality:

  $$P[X - E(X) \geq a] \leq \frac{D(X)}{a^2 + D(X)}.$$

  Solution: Let $Y = X - E(X)$; then $E(Y) = 0$, $D(Y) = D(X)$, $E(Y^2) = D(Y) + E(Y)^2 = D(X)$. For any $t > 0$, consider the non-negative DRV $(Y+t)^2$ and apply Markov:

  $$P[Y \geq a] \leq P\left[(Y+t)^2 \geq (a+t)^2\right] \leq \frac{E((Y+t)^2)}{(a+t)^2} = \frac{D(X) + t^2}{(a+t)^2}.$$

  Minimizing this over t (e.g. by taking derivative equal to zero) we get minimum for $t = D(X)/a$.
- **U:** Check equality for Bernoulli distribution (binomial with $n = 1$) and $a = 1 - p$.
- **DU:** determine all cases of equality in Cantelli's inequality.
- Is Cantelli better or worse than Chebyshev for bounding just one tail, i.e. $P[X \geq a]$?
- There are other variations of Chebyshev inequality that involve more information about X, e.g. supremum of probability density.

- Observation: while Markov bound is loose, we can get better bounds by applying it to some modified DRV in case we know enough about $D(X)$.
- Markov vs. Chebyshev and Chebyshev vs. Cantelli advantage: simplicity of expressions. If we need complicated subsequent work with the bound, like minimizing it, it might be preferable to use Markov despite being much weaker. Example: existential proofs where we only care about some $P > 0$.
- **DU:** Using Chebyshev's inequality, prove the weak law of large numbers (Veta 6.22 in Harman).

Solve the following using both Markov and Chebyshev inequality and compare the strength of the obtained bounds.

- **U:** A coin is weighted so that its probability of landing on heads is 20%, independently of other flips. Suppose the coin is flipped 20 times. Bound the probability it lands on heads at least 15 times.

  Solution: $X \sim Binomial(n = 20, p = 0.2)$, thus $E(X) = np = 4$ and $D(X) = np(1-p) = 3.2$.
  Markov: $P[X \geq 15] \leq 4/15 = 0.27$.
  Chebyshev: $P[|X - 4| \geq 11] \leq 3.2/11^2 = 0.026$ and $[X \geq 15] \subset [|X - 4| \geq 11]$.
  Cantelli: $P[X - 4 \geq 11] \leq 3.2/(11^2 + 3.2) = 0.026$.
  The actual probability is 10^{-7}.

- **U:** A system is processing a random stream of requests; let X be the number of requests per hour. Assume $E(X) = 100$ and $D(X) = 100$. What is the probability that the number of requests over the next hour is at least 250?

  Solution: Markov: $P[X \geq 250] \leq 100/250 = 0.4$; Chebyshev $P[|X - 100| \geq 150] \leq 100/150^2 = 0.004$. If $X \sim Poisson(100)$, actual probability would be almost zero.

- **U:** What is the probability that a random permutation has at least 2 fixed points?

  Solution: Let $X = \sum X_i$, where X_i is an indicator DRV for i being a fixed point. We want $P[X \geq 2]$, while $E(X) = \sum_{i=1}^{n} 1/n = 1$. Markov inequality: $P[X \geq 2] \leq 1/2 = 0.50$.
  Variance:

  $$E(X^2) = \sum E(X_i^2) + \sum_{i \neq j} E(X_i X_j) = 1 + n(n-1) \cdot \frac{(n-2)!}{n!} = 2,$$

  so $D(X) = E(X^2) - E(X)^2 = 2 - 1 = 1$. Chebyshev inequality:

  $$P[|X - 1| \geq 1] \leq 1/1^2 = 1, \quad \text{so} \quad P[X = 0] + P[X \geq 2] \leq 1.$$

  $P[X = 0]$ is known to be $1/e$, so $P[X \geq 2] \leq 0.63$.
  Cantelli inequality: $P[X - 1 \geq 1] \leq 1/(1+1) = 0.50$.

- **U:** The score distribution of an exam is modelled by a random variable X with range $[0, 110]$ and average 50. Give an upper bound on the proportion of students who score at least 100. What are the possible values of variance? For each of the values, determine if Cantelli bound is better than Markov bound.

  Solution: Markov: $P[X \geq 100] \leq 50/100 = 0.50$.
  Minimum variance is 0, for one-point distribution. Maximum variance: we maximize

$E\left((X-50)^2\right)$. The function $(x-50)^2$ on $[0,110]$ is convex, thus we want to push probability mass to the endpoints of the range. Thus $P[X=0]=p$ and $P[X=110]=1-p$, in which case $E(X)=50$ implies $p=6/11$ and $d=D(X)=3000$. Cantelli: $P[X-50 \geq a] \leq \frac{d}{d+a^2}$, so $P[X \geq 100] \leq \frac{d}{d+2500}$. Thus Cantelli would be better if $d<2500$. Note how Markov becomes better than Chebyshev/Cantelli when we get close to a two-outcome distribution (the one maximizing variance).

- **DU:** Suppose the length of time to complete a process T is the sum of three known independent distributions:
  - $X \sim Uniform(0,12)$, with $E(X)=6$ and $D(X)=12$.
  - $Y \sim Exponential(rate=0.5)$, with $E(Y)=5$ and $D(Y)=25$.
  - $Z \sim Normal(10,4)$, with $E(Z)=10$ and $D(Z)=4$.

  Verify the calculations of E and D and prove that $E(T)=21$ and $D(T)=41$. It is hard to find density or distribution function of T, but we can easily bound tail probabilities. Find $P(T>35)$. (The actual answer is ~ 0.03. Note that T could be negative, so Markov is not allowed, but you can still apply it to get a bound better than a wild guess because it is negative just a bit.)

- **DU:** Chebyshev's inequality uses variance to bound deviation from expectation. Variance is essentially about the second moment of a distribution. We can also use higher moments. Prove that if X is a DRV such that for some even k, $E\left((X-E(X))^k\right)$ is finite, then
$$P\left[|X-E(X)| \geq t\sqrt[k]{E\left((X-E(X))^k\right)}\right] \leq \frac{1}{t^k}.$$
  Why would it be difficult to derive a similar inequality for odd k?

- **DU:** Prove that if a k-th moment of a DRV X, i.e. $E\left(X^k\right)$, is finite, then all its smaller moments are finite.

3 Cauchy-Schwarz and Jensen inequalities

Let's discuss $E(XY)$ assuming that $E(X)$, $E(X^2)$, $E(Y)$, $E(Y^2)$ are finite. We can derive an upper bound on $E(XY)$:

$$E(XY)^2 \leq E\left(X^2\right)E\left(Y^2\right).$$

Proof: For any real t, $E\left((X-tY)^2\right) \geq 0$, hence

$$E\left(X^2\right) - 2tE(XY) + t^2E\left(Y^2\right) \geq 0.$$

Since the left-hand side is non-negative, the discriminant of this quadratic equation in t must not be positive, i.e.

$$4E(XY)^2 - 4E\left(X^2\right)E\left(Y^2\right) \leq 0.$$

Equality is achieved when discriminant is zero, i.e. $Y=cX$ for some constant c *almost surely* (i.e. everywhere except on some set that has probability zero).

For two DRVs X and Y, their covariance is defined as

$$\mathrm{Cov}(X,Y) = E\big((X-E(X))(Y-E(Y))\big) = E(XY) - E(X)E(Y).$$

(the last equality holds because after expanding, $E(XE(Y)) = E(Y)E(X)$ since $E(Y)$ is a constant and expectation is linear).

Applying the above upper bound to DRVs $X - E(X)$ and $Y - E(Y)$, we obtain an upper bound on covariance:

$$\text{Cov}(X,Y)^2 \leq D(X)\,D(Y)\,.$$

Consequently,

$$\left|E(XY) - E(X)\,E(Y)\right| \leq \sqrt{D(X)\,D(Y)}.$$

Equality holds when $X - E(X)$ and $Y - E(Y)$ are almost surely linearly dependent, i.e. when $Y = aX + b$ for some real a, b.

U: Let X, Y be DRVs such that $E(X) = 1, D(X) = 4, E(Y) = 2, D(Y) = 1$. Find maximum possible value of $E(XY)$.

Solution: Using the first bound, $E(XY) \leq 5$ because $E(X^2) = E(Y^2) = 5$. But equality cannot be achieved: if $Y = cX$, then $c = E(Y)/E(X) = 2 \neq \sqrt{D(Y)/D(X)} = \sqrt{1/4} = 1/2$.
Using the second bound,

$$E(XY) \leq E(X)\,E(Y) + \sqrt{D(X)\,D(Y)} = 1 \cdot 2 + \sqrt{4 \cdot 1} = 4.$$

Equality for $Y = aX + b$ such that $2 = E(Y) = a + b$ and $D(Y) = a^2 \cdot 4$, i.e. $a = 1/2$, $b = 3/2$.

Note how the simple shift from X to $X - E(X)$ before applying an inequality (Cauchy-Schwarz this time) helped to improve the bound.

Since $D(X) \geq 0$, we know that $E(X^2) \geq E(X)^2$. In general, according to Jensen's inequality for a convex function g,

$$E(g(X)) \geq g(E(X))$$

(assuming that both expectations are finite).

DU: Let X be a positive DRV with $E(X) = 10$. What can we say about

$$(a)\ E\left(\frac{1}{1+X}\right), \qquad (b)\ E\left(\ln\sqrt{X}\right)?$$

Solution: (a) Since $\frac{1}{1+x}$ is convex, $E\left(\frac{1}{1+X}\right) \geq \frac{1}{1+E(X)} = \frac{1}{11}$.

(b) Since ln is concave, $E\left(\ln\sqrt{X}\right) \leq \frac{1}{2}\ln E(X) \approx 1.15$.

4 Birthday paradox

Let $X_1, \ldots X_n$ be $n \geq 2$ iid DRVs with range $1, 2, \ldots, d$, where X_i denotes the birthday of the person i. A collision occurs if $X_i = X_j$ for some $i \neq j$.

(i) If $n \leq \sqrt{d}$, then $P[\text{collision}] < 1/2$.

(ii) If $n \geq c\sqrt{d}$, then $P[\text{collision}] \geq 1 - \frac{4}{c^2}$.

This is our first example where a *threshold phenomenon* occurs: below some threshold, the probability is small, but above the threshold, it is asymptotically close to 1. In our case, P is not close to zero, but the typical thresholds are sharper (e.g. in statistical physics, water suddenly changes from liquid to gas at a certain temperature; SAT instances with the number of clauses below the threshold are almost surely satisfiable, while above the threshold unsatisfiable etc.).

Let Y_{ij} be the indicator variable for $X_i = X_j$ and $Y = \sum_{i<j} Y_{ij}$.
For (i), we want $P[Y \geq 1] < 1/2$. Let's try Markov.

$$E(Y) = \sum E(Y_{ij}) = \sum P[Y_{ij} = 1] = \frac{1}{d}\binom{n}{2},$$

so

$$P[Y \geq 1] \leq \frac{E(Y)}{1} = \frac{1}{d}\binom{n}{2} = \frac{n(n-1)}{2d} < \frac{n^2}{2d} \leq \frac{1}{2}.$$

For (ii), we would need a lower bound on $P[Y \geq 1]$, or equivalently, an upper bound on $P[Y = 0]$. Markov bound cannot give us a useful estimate: Y could be 0 with probability $1-\varepsilon$ and $E(Y)/\varepsilon$ with probability ε. Since $P[Y=0] \leq P[|Y - E(Y)| \geq E(Y)]$, we hope that Chebyshev inequality plus an upper bound on $D(Y)$ will help us.

The variables Y_{ij} are pairwise independent, but triples of them are not independent (because $Y_{ij} = Y_{jk} = 1$ implies $Y_{ik} = 1$).

$$D(Y) = \sum D(Y_{ij}) = \sum E\left(Y_{ij}^2\right) - E(Y_{ij})^2 = \sum \frac{1}{d} - \frac{1}{d^2} < \sum \frac{1}{d} = \frac{1}{d}\binom{n}{2}.$$

Note that $D(Y) \leq E(Y)$; if Y was a sum of iid variables (why not?), then the distribution of deviations from $E(Y)$ would be approaching normal for large n (by central limit theorem). Hence, most probability mass would be concentrated around $E(Y) \pm 3\sqrt{D(Y)}$, and thus $Y = 0$ would be very unlikely. Now more precisely with Chebyshev (for $a = E(Y)$) that can be applied even if Y_{ij}'s are not iid:

$$P[Y=0] \leq P[|Y - E(Y)| \geq E(Y)] \leq \frac{D(Y)}{E(Y)^2} \leq \frac{\frac{1}{d}\binom{n}{2}}{\left(\frac{1}{d}\binom{n}{2}\right)^2} = \frac{d}{\binom{n}{2}} = \frac{2d}{n^2 - n} \leq \frac{2d}{\frac{1}{2}n^2} \leq \frac{4}{c^2}.$$

From this, (ii) follows immediately.

In (ii), we could employ the Paley-Zygmund inequality which states that for a non-negative DRV X with finite D and any $\theta \in [0,1]$,

$$P[X > \theta E(X)] \geq (1-\theta)^2 \frac{E(X)^2}{E(X^2)}.$$

In our case, for $\theta = 0$,

$$P[Y > 0] \geq \frac{E(Y)^2}{E(Y^2)} = \frac{E(Y)^2}{D(Y) + E(Y)^2} \geq \frac{E(Y)^2}{E(Y) + E(Y)^2} = \frac{E(Y)}{1 + E(Y)}.$$

For large $n \geq c\sqrt{d}$, we have $E(Y) = n(n-1)/2d \geq c^2/4$. Thus

$$P[Y > 0] \geq \frac{c^2/4}{1 + c^2/4} = \frac{1}{1 + 4/c^2} \geq 1 - \frac{4}{c^2}.$$

U: Let X be a non-negative integer-valued DRV with positive expectation. Prove that

$$\frac{E(X)^2}{E(X^2)} \leq P[X \neq 0] \leq E(X).$$

(The lower bound is a variant of Paley-Zygmund and its use is sometimes called *second moment method*. We want a full proof, not just an application of a more generalized form of the same theorem.)

Solution: Consider an indicator DRV $Y = I[X > 0]$. Then $XY = X$, $Y^2 = Y$, and $E(Y^2) = E(Y) = P[X \neq 0]$. Hence

$$E(X)^2 = E(XY)^2 \leq E\left(X^2\right) E\left(Y^2\right) = E(X)^2 P[X \neq 0].$$

For the upper bound, we just apply Markov's inequality: $P[X \geq 1] \leq E(X)/1$ (such usage is known as *first moment method*).

5 More concentration bounds

Markov and Chebyshev inequalities give us some bounds on how likely it is that a random variable deviates from its expectation. There are many more inequalities of this kind, known as concentration bounds; we discuss a few examples below.

In the following, let us assume that $X_1, X_2, \ldots, X_n$ are iid DRVs with finite E and $D = \sigma^2$ and $X = \frac{1}{n}\sum X_i$.

Weak law of large numbers (WLLN).

$$P[|X - E(X)| \geq \varepsilon] \leq \frac{D(X)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Proof: the formula above already applies Chebyshev, we only need
$D(X) = \frac{1}{n^2}\sum D(X_i) = \frac{1}{n}D(X_i)$ (we used independence of X_i).

Hoeffding's inequality. If each $X_i \in [a, b]$,

$$P[|X - E(X)| \geq \varepsilon] \leq 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

Note how this bound is exponential in terms of n vs. just linear for the bound derived via Chebyshev. It is commonly used in machine learning:
`https://people.cs.umass.edu/~domke/courses/sml2010/10theory.pdf`

Notice the difference between the two bounds: while WLLN gives a bound in terms of variance, Hoeffding gives a bound that only requires knowing $E(X)$, the variance is implicitly covered by the assumption $X_i \in [a, b]$. We can venture a guess that the Hoeffding's bound corresponds to the worst case (if we knew more about $D(X_i)$, we could perhaps improve the bound). If we wanted a bound independent of $D(X_i)$ from WLLN, we would need some upper bound on $D(X)$. Let's try it in the following exercise.

U: Assume that the X_i above are Bernoulli, i.e. $P[X_i = 1] = p$ and $P[X_i = 0] = 1 - p$. Use WLLN and Hoeffding's inequality to determine how large should n be so that $P[|X - E(X)| \geq \varepsilon] \leq \delta$ for a given δ.

Solution: First, we derive a bound on $D(X_i)$ independent of p: $D(X_i) = p(1-p) \leq 1/4$. For WLLN, we get

$$P[|X - E(X)| \geq \varepsilon] \leq \frac{D(X)}{\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

If this is to be at most δ, we need $n = \Omega\left(\frac{1}{\varepsilon^2\delta}\right)$.

For Hoeffding's inequality, we have $a = 0$, $b = 1$, and thus

$$P[|X - E(X)| \geq \varepsilon] \leq 2e^{-2n\varepsilon^2}.$$

If this is to be at most δ, we need $n = \Omega\left(\frac{\ln(2/\delta)}{\varepsilon^2}\right)$. This is obviously better than the bound derived from WLLN; the probability of failure decreases exponentially with the number of samples n. For asymptotic purposes (proving that $P \to 0$ as $n \to \infty$, without regard for rate of convergence), both bounds could be used with the same effect.

Let's take a look at how ε depends on n (assuming fixed δ): the error band size only decreases as the square root of the number of samples, both for Chebyshev and Hoeffding's bound since they both contain the term $n\varepsilon^2$.

DU: Study the idea behind **Chernoff bounds**, another strengthening of Markov's inequality. `https://web.stanford.edu/class/archive/cs/cs109/cs109.1218/files/student_drive/6.2.pdf` `https://www.probabilitycourse.com/chapter6/6_2_3_chernoff_bounds.php`

6 Monte Carlo method and the median trick

We say that a RV X is an unbiased estimator of μ if $E(X) = \mu$. The relative variance of an unbiased estimator X is $t = \frac{D(X)}{\mu^2}$.

Let $X_1, X_2, \ldots, X_k$ be independently chosen samples of X and let $Y = \frac{1}{k}\sum X_i$. Since X is an unbiased estimator of μ, $E(X_i) = \mu$, and thus $E(Y) = \mu$. In addition,

$$D(Y) = \frac{1}{k^2} \cdot D\left(\sum X_i\right) = \frac{1}{k^2} \cdot k\, D(X) = \frac{D(X)}{k} = \frac{t\mu^2}{k}.$$

By Chebyshev's inequality,

$$P[(1-\varepsilon)\mu \leq Y \leq (1+\varepsilon)\mu] = P[|Y - \mu| \leq \varepsilon\mu] \geq 1 - \frac{D(Y)}{\varepsilon^2\mu^2} = 1 - \frac{t}{k\varepsilon^2}. \tag{1}$$

If we are aiming for $P = 1 - \delta$, we take $k = \frac{t}{\varepsilon^2} \cdot \frac{1}{\delta}$ and get a $(1 \pm \varepsilon)$-approximation of μ.

We want to investigate how many dart throws a Monte Carlo method needs to determine the area of a complicated shape. Let A be the shape, B a rectangle in which A lies, and s denote the surface area function. Let W be a Bernoulli RV indicating whether a uniformly randomly chosen point from B belongs to A or not; then $E(W) = P[W = 1] = s(A)/s(B)$ and

$$D(W) = P[W = 1]\big(1 - P[W = 1]\big) = \frac{s(A)}{s(B)}\left(1 - \frac{s(A)}{s(B)}\right).$$

Since we precisely know $s(B)$, we can use $X = s(B)W$ as an unbiased estimator of $s(A)$. The relative variance is

$$t = \frac{D(X)}{\mu^2} = \frac{s(B)^2 D(W)}{s(A)^2} = \frac{s(B)}{s(A)} - 1.$$

This relative variance t is fixed (the shapes A and B are given, they are not part of the random process). Thus in the inequality (1) it is sufficient to take $k = \Theta(\frac{1}{\varepsilon^2} \cdot \frac{1}{\delta})$ to get a $(1 \pm \varepsilon)$-approximation of $s(A)$ with probability $1 - \delta$ (the asymptotics in Θ with respect to $\delta \to 0$). This holds *independently of the distribution of X (or W).*

It would be possible to improve the bound on k by using Hoeffding's inequality instead of Chebyshev inequality directly. Instead, we show a general "*median trick*" used to increase the probability of having a good unbiased estimator.

Let us fix k such that the probability in inequality (1) is 3/4 (or more). We repeat the estimation ℓ times, obtaining $Y_1, Y_2, \dots, Y_\ell$ as samples of Y, and output the median of the ℓ obtained values. The median is in the interval $I_\varepsilon = [(1-\varepsilon)\mu, (1+\varepsilon)\mu]$ if at least half of the Y_i's are in I_ε.

We show that the probability of the opposite event (most Y_i's outside of the interval) is very small. Let Z_i be the indicator DRV for $Y_i \in I_\varepsilon$ and $Z = \sum Z_i$. Clearly $P[Z_i = 1] \geq 3/4$, thus $E(Z) = \sum E(Z_i) \geq 3\ell/4$. By Hoeffding's inequality, using $Z \in [0, \ell]$,

$$P\left[Z \leq \frac{\ell}{2}\right] \leq P\left[Z \leq E(Z) - \frac{\ell}{4}\right] \leq P\left[\left|Z - E(Z)\right| \geq \frac{\ell}{4}\right] \leq 2\exp\left(\frac{-\ell}{8}\right).$$

We can pick $\ell = \Theta(\ln \frac{1}{\delta})$. That way, we only need $O(\frac{t}{\varepsilon^2} \cdot \ln \frac{1}{\delta})$ samples (the O notation hides k which is treated as a constant since it is independent of δ). So instead of a bound linear in $1/\delta$ we have one logarithmic in $1/\delta$.

This median trick is especially useful for randomized algorithms — starting from a mediocre algorithm giving the right answer barely above half of the time, we can boost the probability arbitrarily close to 1 without blowing up the time complexity.

7 Random walk

Consider a *random walk*: a person walks along integers, starting at 0. In each step, he moves right or left, each with probability 1/2, independently of what he did before. Let X_i be ± 1; the position after n steps is $S_n = \sum_{i=1}^{n} X_i$.

U: Prove all of the following:

- $E(S_n^2) = n$;
- $E(|S_n|) \leq \sqrt{n}$;
- $\lim_{c\to\infty} P\left[|S_n| \geq c\sqrt{n}\right] = 0$.

Solution: Obviously $E(X_i) = 0$, $D(X_i) = 1$ and $E(S_n) = 0$. For the investigation of the distance from origin, which is $|S_n|$, it seems easier to look at S_n^2 instead. We have

$$E\left(S_n^2\right) = n + \sum_{i \neq j} E(X_i X_j) = n.$$

The sum is zero because $E(X_i X_j) = E(X_i)\, E(X_i) = 0$ (X_i's are independent). Alternatively,

$$E\left(S_n^2\right) = E\left(S_n^2\right) - E(S_n)^2 = D(S_n) = \sum_i D(X_i) = n.$$

If we are interested in the distance $|S_n|$ and not its square S_n^2, we cannot just naively take the square root: in general, $E(X)^2 \neq E(X^2)$. But we can use the second moment to at least get an upper bound via Cauchy-Schwarz inequality:

$$E(|S_n|) = E(|S_n| \cdot 1) \leq \sqrt{E(S_n^2)\,E(1^2)} = \sqrt{n}.$$

Then Markov's inequality yields

$$P\left[|S_n| \geq c\sqrt{n}\right] \leq \frac{E(|S_n|)}{c\sqrt{n}} \leq \frac{1}{c}.$$

For $c \to \infty$, $P \to 0$.

Alternative solution: For $c > 0$, Hoeffding's inequality gives us

$$P\left[|S_n| \geq c\sqrt{n}\right] \leq 2\exp\left(-\frac{c^2}{2}\right).$$

For $c \to \infty$, $P \to 0$. If we choose c to be any unbounded function of n (albeit growing very slowly), then $P \to 0$ as $n \to \infty$.

A careful analysis shows that $E(|S_n|) \sim \sqrt{2n/\pi} \approx 0.8\sqrt{n}$ (check WolframAlpha). This value could also be obtained from normal approximation, i.e. by applying the Central Limit Theorem to the binomial distribution arising in our random walk. We don't do it here.

U: Use the Paley-Zygmund inequality to derive a lower bound on $P[S_n^2 \geq cn]$ for some constant c.

Solution: In order to use Paley-Zygmund, we need to compute variance of S_n^2, or its second moment, i.e. $E(S_n^4)$. In the expansion of $(X_1 + \cdots + X_n)^4$, we can apply E to each term individually because E is linear. But all terms that have at least one odd exponent are zero since $E(X_i) = 0$ (our X_i's are independent, so $E(X_iX_j) = E(X_i)E(X_j)$). Thus we are only concerned with terms where all exponents are even. They are of two kinds:

- $E(X_i^4) = E(1) = 1$; there are n such terms.
- $E\left(X_i^2X_j^2\right) = E(X_i^2)\,E\left(X_j^2\right) = 1$; there are $3n^2 - 3n$ such terms because there are $\binom{n}{2}$ ways to choose the pair of indices and $\binom{4}{2}$ terms for a fixed pair of indices (we are choosing 2 positions for index i out of 4 possible ones).

Put together, $E(S_n^4) = 3n^2 - 2n$. Now we use Paley-Zygmund with $t \in [0, 1]$:

$$P\left[S_n^2 \geq tE\left(S_n^2\right)\right] \geq (1-t)^2\frac{E(S_n^2)^2}{E(S_n^4)} = (1-t)^2\frac{n^2}{3n^2 - 2n}$$

Thus

$$P\left[S_n^2 \geq tn\right] \geq (1-t)^2\frac{n^2}{3n^2 - 2n} \geq (1-t)^2 \cdot \frac{1}{3}.$$

and

$$P\left[|S_n| \geq c\sqrt{n}\right] \geq \tfrac{1}{3}(1 - c^2)^2.$$

Consider now $M_n = \max_{0 \leq k \leq n} |S_k|$, i.e. the maximum distance achieved during n steps of our random walk. *Kolmogorov's maximal inequality* yields

$$P[M_n \geq m] \leq \frac{D(S_n)}{m^2} = \frac{n}{m^2}.$$

In order to calculate $E(M_n)$, we use a trick applicable to any non-negative integer-valued DRV X:

$$E(X) = \sum_{a=1}^{\infty} aP[X = a] = \sum_{a=1}^{\infty}\sum_{b=1}^{a} P[X = a] = \sum_{b=1}^{\infty}\sum_{a=b}^{\infty} P[X = a] = \sum_{b=1}^{\infty} P[X \geq b].$$

In our case,

$$E(M_n) = \sum_{m=1}^{\infty} P[M_n \geq m] \leq \sum_{m=1}^{\infty} \frac{n}{m^2} = n\frac{\pi^2}{6} = O(n).$$

This is disappointing — trivially, $M_n \leq n$. But notice that the first terms of the sum, i.e. $n/1^2, n/2^2, \ldots$, are much larger than 1, so they are very poor bounds for probabilities. We can do better if we split our sum at the point where the Kolmogorov's bound gets below 1:

$$E(M_n) \leq \sum_{m=1}^{\sqrt{n}} 1 + \sum_{m>\sqrt{n}} \frac{n}{m^2} = \sqrt{n} + n \sum_{m\geq\sqrt{n}} \frac{1}{m^2} = \sqrt{n} + n \cdot O\left(\frac{1}{\sqrt{n}}\right) = O(\sqrt{n})$$

because $\sum \frac{1}{k^2} \leq \sum \frac{1}{k(k-1)} = \sum \frac{1}{k-1} - \frac{1}{k}$ and this telescopic series can be bounded by its first term $O(1/\sqrt{n})$. Hence even the maximum distance achieved during the random walk does not deviate much from the average distance of $\approx \sqrt{n}$.

Random walks (and Brownian motion) have been researched in depth, whole books exist on them.

8 Triangles in random graphs

The basic Erdős-Rényi model $G(n, p)$ generates a graph on n vertices where each edge is present with probability p independently of other edges.

Let's investigate how the value of p affects certain graph properties. We are interested in what happens for $n \to \infty$ if p is a given function of n.

Having a triangle. The probability that a specific triangle is present in $G(n, p)$ is p^3. However, the triangles overlap, so it is not easy to compute the probability that there is at least one triangle in a graph. Instead, we will explore moments such as E and D to derive bounds on the probability that is difficult to compute precisely.

Let $X = \sum_T I_T$, where the sum is over all triangles T in $G(n, p)$ and I_T is the indicator DRV for the presence of T. Then,

$$E(X) = \binom{n}{3} p^3 \sim \frac{n^3 p^3}{6}.$$

If $p \prec 1/n$, then $E(X) \to 0$, thus by Markov's inequality (first moment method),

$$P[X \geq 1] \leq \frac{E(X)}{1} \to 0.$$

Next, we focus on the case $p \succ 1/n$. We use the second moment method to derive a lower bound on $P[X \geq 1]$.

$$E\left(X^2\right) = \sum_{T,T'} E(I_T I_{T'}).$$

The pairs (T, T') fall into four categories:

- *Triangles sharing no vertices.* For disjoint triangles, the events are independent, so $E(I_T I_{T'}) = p^6$. There are

$$\binom{n}{3}\binom{n-3}{3} = \frac{n!}{(n-3)!\,3!} \cdot \frac{(n-3)!}{(n-6)!\,3!} = \frac{1}{6^2} n(n-1)\cdots(n-5) = \frac{1}{36}\left(n^6 + o(n^6)\right)$$

  such pairs (we are not dividing by 2 because in the sum the pairs T, T' and T', T are both present). Notice that while we could estimate the contribution of these pairs to $E(X^2)$ as $\Theta(n^6p^6)$, we do not want to do it because we want to show that $P[X \geq 1] \to 1$, not just some constant — hence we cannot hide the constant in Θ.

- *Triangles sharing exactly one vertex.* These pairs also have no common edge, so the six edges (three for each triangle) are distinct, and again $E(I_T I_{T'}) = p^6$. The number of such pairs is $\Theta(n^5)$ (we pick 5 vertices out of n and there is a constant number of ways how to arrange them into a pair of triangles). Since this is negligible compared to $\Theta(n^6p^6)$ from the previous case, we don't need to determine the constant.

- *Triangles sharing exactly two vertices (i.e. an edge).* In this case, the two triangles together have five distinct edges; hence $E(I_T I_{T'}) = p^5$. There are $\Theta(n^4)$ such pairs (thanks to $p^5 = O(p^6 n)$, we don't need the constant for the same reason as above).

- *Triangles sharing three vertices.* This covers the case $T = T'$, so we have $\binom{n}{3}$ such pairs and each of them has probability p^3. Their contribution is thus $\frac{1}{6}(n^3 + o(n^3))p^3 = \Theta(n^3p^3) = o(n^6p^6)$ because $n^3p^3/n^6p^6 = (np)^{-3} = \omega(1)^{-3} \to 0$.

Put together,

$$E\left(X^2\right) \sim \frac{1}{36}\left(n^6 + o(n^6)\right)p^6 + \Theta(n^5p^6) + \Theta(n^4p^5) + \Theta(n^3p^3) = \frac{1}{36}\left(n^6 + o(n^6)\right)p^6.$$

Thus

$$P[X \geq 1] \geq \frac{(E(X))^2}{E(X^2)} \sim \frac{\frac{1}{36}n^6p^6}{\frac{1}{36}\left(n^6 + o(n^6)\right)p^6} \to 1.$$

If we choose $p \prec 1/n$, then the probability of having a triangle in G goes to 0.
If we choose $p \succ 1/n$, then the probability of having a triangle in G goes to 1.
This transition from "almost surely no triangles" to "almost surely some triangles" occurring at $p = 1/n$ is called a *threshold phenomenon.* In this case, however, the threshold is *coarse* because the transition does not occur entirely suddenly. Let's investigate it.

U: What is the probability that a random graph $G(n, p)$ contains a triangle if $p = c/n$ for a positive constant c?

Solution: Let X be the number of triangles in G. For an upper bound on P, we revisit our first moment argument.

$$P[X > 0] \leq E(X) \sim \frac{c^3}{6}$$

For a lower bound on P, we revisit the second moment argument. It will be handy that $np = c$.

$$E\left(X^2\right) \sim \frac{1}{36}\left(n^6 + o(n^6)\right)p^6 + \Theta(n^5p^6) + \Theta(n^4p^5) + \frac{1}{6}(n^3 + o(n^3))p^3 = \frac{c^6}{36} + \frac{c^3}{6}.$$

Hence

$$P[X > 0] \geq \frac{(E(X))^2}{E(X^2)} \sim \frac{\frac{c^6}{36}}{\frac{c^6}{36} + \frac{c^3}{6}} = \frac{c^3}{c^3 + 6}.$$

We can conclude that P belongs to an interval dependent on c and is neither approaching 0 nor 1 (the precise value is $P \approx 1-\exp(-c^3/6)$, but we don't have tools to prove it). That is why our threshold is not sharp.

Let's try a variant of the second moment method that uses $D(X)$. If we can prove that $\frac{D(X)}{E(X)^2} \to 0$, then

$$\frac{E(X^2)}{E(X)^2} = \frac{D(X)+E(X)^2}{E(X)^2} = \frac{D(X)}{E(X)^2} + 1 \to 1,$$

and thus $P[X>0] \geq \frac{E(X)^2}{E(X^2)} \to 1$.

The variance can be calculated by considering individual pairs of triangles:

$$D(X) = E\left(X^2\right) - E(X)^2 = \sum_{T,T'} E\left(I_T I_{T'}\right) - \sum_{T,T'} E\left(I_T\right) E\left(I_{T'}\right) = \sum_{T,T'} E\left(I_T I_{T'}\right) - E\left(I_T\right) E\left(I_{T'}\right).$$

This proves that the variance of a sum of DRVs can also be expressed as a sum over their covariances:

$$D(X) = D\left(\sum_T I_T\right) = \sum_{T,T'} \operatorname{Cov}(I_T, I_{T'}).$$

We split the calculation according to the number of edges shared by T and T'.

- 0 edges — then I_T and $I_{T'}$ are independent and the contribution of this pair is zero.
- 1 edge — then $E\left(I_T I_{T'}\right) = p^5$ while $E\left(I_T\right)E\left(I_{T'}\right) = p^6$. There are $\binom{n}{4}$ such pairs and each contributes $p^5 - p^6 \leq p^5$, so in total we have $O(n^4p^5)$.
- ≥ 2 edges — then $T = T'$ and the contribution of such pairs is $\binom{n}{3}(p^3 - p^6) = O(n^3p^3)$.

If we assume $p \succ 1/n$, then

$$\frac{D(X)}{E(X)^2} = O\left(\frac{n^4p^5 + n^3p^3}{n^6p^6}\right) \to 0.$$

It seems easier to calculate expectations (which are linear and easily bounded) than probabilities (which would require some kind of convoluted inclusion-exclusion). Note how in the second moment calculation, it turned out that the overlapping triangles are negligible. This is not always the case, as we will see below.

Notation: Consider an event A_n that depends on n somehow (e.g. A_n could be "$G(n,p(n))$ contains a triangle"). If $P(A_n) \to 1$ for $n \to \infty$, we say that A_n happens *with high probability* (w.h.p.) or *asymptotically almost surely (a.a.s.)*.

Our method could be used for other subgraphs of G, not just the triangle. But if overlapping is common, the second moment method will fail; this tends to happen if the graph we're looking at has a subgraph which is more dense than the original graph. Let's consider a $(4,1)$-lollipop graph H, i.e. a K_4 with one pendant vertex attached. The K_4 is the dense part here.

U: Use the first moment method to determine a bound on p such that $G(n,p)$ does not contain the lollipop H w.h.p.

Solution: There are 5 vertices and 7 edges in H. The number of ways to choose H is $\Theta(n^5)$ and for a fixed copy of H, it is present with probability p^7, so $E(X) = \Theta(n^5p^7)$. This goes to 0 for $p \prec n^{-5/7}$.

U: For $p \sim n^{-5/7}$, compare the contribution to $E(X^2)$ from a pair of disjoint copies of H and from a pair of copies that share the K_4 contained in H.

Solution: For independent copies, there are $\Theta(n^{10})$ of them, each with probability p^{14}, thus the contribution is $\Theta(n^{10}p^{14}) = \Theta(1)$. For copies overlapping in K_4, the probability is p^8 and there are $\Theta(n^6)$ such pairs, so the contribution is $\Theta(n^6p^8) = \Theta(n^{2/7})$, significantly larger than that of the independent copies.

U: Suppose that $p = d/n$ where d is a constant. Prove that w.h.p., no vertex belongs to more than one triangle in $G(n,p)$.

Solution: Consider the probability P_v that two triangles share a vertex v. Triangles that only share v and no other vertices contribute at most $p^6\binom{n-1}{4} = O(n^{-2})$ to P_v (upper bound because the probability of a union of events is at most the sum of their probabilities). Triangles that share another vertex contribute at most $p^5\binom{n-1}{3} = O(n^{-2})$. So for a fixed v, $P_v = O(n^{-2})$. Summing over all v, we get an upper bound on the desired probability equal to $n \cdot O(n^{-2}) = O(n^{-1})$, which goes to zero.

DU: Exercises from BOOK: 1.4.1 to 1.4.9.

9 Isolated vertices in random graphs

Let's investigate the property $S = \{G \in \mathcal{G}(n,p) \mid G \text{ does not contain isolated vertices}\}$.

Let $X = \sum_v I_v$, where the sum is over all vertices v in $G(n,p)$ and I_v is the indicator DRV for v being isolated. A vertex is isolated with probability $(1-p)^{n-1}$, so

$$E(X) = \sum_v E(I_v) = n(1-p)^{n-1}.$$

U: Prove that if p is a constant from $(0,1)$, $E(X) \to 0$. On the other hand, if $p = 1/n$, $E(X) \to \infty$.

Solution: In the first case, $E(X) = nc^{n-1} \to 0$. In the second case,

$$E(X) = n\left(1 - \frac{1}{n}\right)^n \left(1 - \frac{1}{n}\right)^{-1} \sim ne^{-1} \to \infty.$$

In the following, let $p = \dfrac{c\ln n}{n}$ for a positive constant c.

U: Prove that if $c > 1$, then $E(X) \to 0$ and $G \in S$ w.h.p.

Thanks to $p \to 0$, we have $(1-p) \sim e^{-p}$ and thus

$$E(X) \sim n \cdot e^p \cdot e^{-np} = n\, e^p\, e^{-c\ln n} = n^{1-c} \cdot e^p \to 0.$$

From the first moment method we get $P[X > 0] \to 0$, so $P[X = 0] \to 1$, i.e. $G \in S$ w.h.p.

U: Prove that if $c < 1$, then $E(X) \to \infty$. Can we conclude that $G \notin S$ w.h.p.?

Solution: Again $p \to 0$, so $E(X) = n^{1-c} \cdot e^p \to \infty$. Infinite expectation does not guarantee that graphs with $X > 0$ occur often because the isolated vertices contributing to $E(X)$ could be concentrated in a small number of graphs.

U: Use the second moment method to prove that for $c < 1$, $G \notin S$ w.h.p.

Solution:

$$E\left(X^2\right) = E\left(\sum_v I_v^2 + \sum_{u\neq v} I_u I_v\right) = E\left(X\right) + n(n-1)E\left(I_u I_v\right),$$

where

$$E\left(I_u I_v\right) = P\left[u \text{ and } v \text{ are isolated}\right] = (1-p)^{2n-3}.$$

We would like to prove that $\lim_{n\to\infty} \frac{E(X)^2}{E(X^2)} = 1$. The fraction is easier to deal with when inverted:

$$\frac{E\left(X^2\right)}{E\left(X\right)^2} = \frac{n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}}{n^2(1-p)^{2(n-1)}} = \frac{1+(n-1)(1-p)^{n-2}}{n(1-p)^{n-1}}.$$

Thus

$$\lim_{n\to\infty} \frac{E\left(X^2\right)}{E\left(X\right)^2} = \lim_{n\to\infty} \frac{1}{n(1-p)^{n-1}} + \lim_{n\to\infty} \frac{n-1}{n} \cdot \frac{1}{1-p} = 0 + 1 = 1.$$

(the first limit uses that $E\left(X\right) \to \infty$ in the denominator). The limit of the inverted fraction is thus also 1.

The behavior of $G \in S$ with respect to changes in c is corresponds to a *sharp threshold* at $p = \frac{\ln n}{n}$. For p below the threshold, there are isolated vertices in G almost surely, while after crossing the threshold by just a bit, isolated vertices completely disappear. We also say that a *phase transition* occurs at the threshold probability (this term comes from statistical physics and describes e.g. how ice suddenly melts into water at a specific temperature).

U: Prove that the threshold for isolated vertices in a random bipartite graph $G(n, n, p)$ is $p = \frac{\ln n}{n}$. Specifically, let $p = \frac{\ln n + f(n)}{n}$;

$$P = P\left[G(n,n,p) \text{ does not have an isolated vertex}\right] = \begin{cases} 0 & \text{if } f(n) \to -\infty, \\ \exp(-2e^{-c}) & \text{if } f(n) \to c, \\ 1 & \text{if } f(n) \to \infty. \end{cases}$$

Solution: Let X be the number of isolated vertices in $G(n, n, p)$;

$$E(X) = 2n(1-p)^n,$$

$$E\left(X^2\right) = E\left(\sum_v I_v^2 \quad + \quad 2\sum_{\substack{u\neq v \\ u,v \text{ in the same part}}} I_u I_v \quad + \quad 2\sum_{\substack{u\neq v \\ u,v \text{ in different parts}}} I_u I_v\right)$$

$$= E\left(X\right) + 2n(n-1)(1-p)^{2n} + 2n^2(1-p)^{2n-1}.$$

Then

$$E(X) \sim 2ne^{-np} = 2ne^{-\ln n - f(n)} = 2e^{-f(n)}.$$

If $f(n) \to \infty$, $E(X) \to 0$, thus first moment method yields $P\left[X > 0\right] \to 0$ and thus $P \to 1$. If $f(n) \to -\infty$, $E(X) \to \infty$, and the second moment method yields $P\left[X > 0\right] \to 1$ and thus $P \to 0$:

$$\lim_{n\to\infty} \frac{E\left(X^2\right)}{E\left(X\right)^2} = \lim_{n\to\infty} \frac{2n(1-p)^n + 2n(n-1)(1-p)^{2n} + 2n^2(1-p)^{2n-1}}{4n^2(1-p)^{2n}}$$

$$= \lim_{n\to\infty} \frac{1}{E(X)} + \frac{2n(n-1)}{4n^2} + \frac{1}{2(1-p)} = 0 + \frac{1}{2} + \frac{1}{2} = 1.$$

The remaining case, $f(n) \to c$, is the hardest; we deal with it in the next section.

10 Isolated vertices: Poisson approximation

The Chen-Stein method is mentioned in several books and articles. It can be used in situations where a sum of indicator variables approaches the Poisson distribution, as is the case for our isolated vertices. Similar methods can be used for normal distribution (which was the original result of Stein).

- Frieze, Karonski: Introduction to Random Graphs, Chapter 26.3, Theorem 26.12
- Janson, Luczak, Rucinski: Random Graphs, Chapter 6.2, Example 6.28
- Notes on the Chen-Stein method for Poisson convergence, Theorem 3.2

Let $(X_n)_{n=0}^{\infty}$ be a sequence of non-negative integer-valued random variables and $(\lambda_n)_{n=0}^{\infty}$ the sequence of their expectations, i.e. $\lambda_n = E(X_n)$. We say that (X_n) is Poisson-convergent if the *total variation distance* d_{TV} between the distribution of X_n and Poisson(λ_n) tends to 0 as $n \to \infty$. (Note that we did not define total variation distance because we do not need to actually calculate it. We are only interested in applying theorems providing upper bounds on d_{TV}. See the references for the definition and the proofs of the theorems given below.) Assume that $\lambda_n \to \lambda$.

- If $\lambda = 0$, (X_n) converges to a distribution degenerate at 0, i.e. $P[X = 0] = 1$ and $P[X = k] = 0$ for any $k > 0$.
- If $\lambda > 0$, (X_n) converges to Poisson(λ).
- If $\lambda = \infty$, then CLT implies that $(X_n - \lambda_n)/\sqrt{\lambda_n} \to \text{Normal}(0, 1)$.

In order to prove this convergence, we employ the following theorem.

Theorem Chen-Stein 1. Let $\{I_i\}_{i\in\Gamma}$ be a set of indicator variables. For each i, dependency neighborhood Γ_i is chosen so that the collection $\{I_j : j \notin \Gamma_i \cup \{i\}\}$ is independent of I_i. Let $X = \sum_{i\in\Gamma} I_i$ and $\lambda = E(X)$. Then the total variation distance between X and Poisson(λ) is

$$d_{TV}(X, \text{Poisson}(\lambda)) \leq \min(1, \lambda^{-1}) \left(\sum_{i\in\Gamma} E(I_i)^2 + \sum_{i\in\Gamma} \sum_{j\in\Gamma_i} \Big(E(I_i) E(I_j) + E(I_i I_j) \Big) \right).$$

Let $G(n, n, p)$ be a random bipartite graph with parts A and B, each of size n, and edge probability $p = \frac{\ln n + f(n)}{n}$. Let I_v be the indicator variable for $v \in A \cup B$ being isolated;

$$P[I_v = 1] = (1-p)^n \qquad \text{and} \qquad E(I_v) = (1-p)^n \sim e^{-pn} = \frac{e^{-f(n)}}{n}.$$

Let $X = \sum_{v\in A\cup B} I_v$; we computed previously that

$$\lambda = E(X) = 2n(1-p)^n \sim 2e^{-f(n)}.$$

The dependency neighbourhoods are non-empty because I_u and I_v are independent only if u and v belong to the same part. For I_v, we can pick Γ_v as the part not containing v.

According to the theorem above,

$$d_{TV}(X, \text{Poisson}(\lambda)) \leq \min(1, \lambda^{-1}) \left(\sum_{v\in A\cup B} E(I_v)^2 + \sum_{v\in A\cup B} \sum_{u\in\Gamma_v} \Big(E(I_v) E(I_u) + E(I_v I_u) \Big) \right).$$

We need upper bounds for the two sums. First,

$$\sum_{v\in A\cup B} E\left(I_v\right)^2 \sim 2n\left(\frac{e^{-f(n)}}{n}\right)^2 = \frac{2e^{-2f(n)}}{n}.$$

For the second sum, we need

$$E\left(I_u I_v\right) = P\left[\text{both } I_u,\, I_v \text{ are isolated}\right] = (1-p)^{2n-1} \sim (1-p)^{2n} \sim e^{-2pn} = \frac{e^{-2f(n)}}{n^2}.$$

Then

$$\sum_{v\in A\cup B}\sum_{u\in\Gamma_v}\Big(E\left(I_v\right)E\left(I_u\right) + E\left(I_v I_u\right)\Big) = 2n\cdot n\cdot\left(\left(\frac{e^{-f(n)}}{n}\right)^2 + \frac{e^{-2f(n)}}{n^2}\right) = 4e^{-2f(n)}$$

Put together,

$$d_{TV}(X, P_\lambda) \le \min\left(1, \lambda^{-1}\right)\left(\frac{2e^{-2f(n)}}{n} + 4e^{-2f(n)}\right).$$

If $f(n) \to \infty$, then $d_{TV} \to 0$ and $\lambda \to 0$, and thus $P\left[X = 0\right] \to 1$. In other words, there are no isolated vertices in $G(n,n,p)$ w.h.p. Note that our usage of Theorem Chen-Stein 1 replaced the first moment method calculations in this case.

However, if $f(n) \to c$, then d_{TV} does not approach 0, and our bound is too weak too prove Poison convergence. We cannot derive any useful conclusion. Let's try another version of Chen-Stein.

For certain sets of indicator random variables we say that they are *positively related*. The notion is somewhat technical and we don't want to explore that too deeply (see Section 6.2 in the references above for more details). Our indicator variables for isolated vertices do have this property, as explained in Example 6.28, so we can use the following theorem.

Theorem Chen-Stein 2. Let $\{I_i\}_{i\in\Gamma}$ be a set of *positively related* indicator variables. Then

$$d_{TV}(X, \text{Poisson}(\lambda)) \le \min(1, \lambda^{-1})\left(D(X) - E(X) + 2\sum_{i\in\Gamma} E\left(I_i\right)^2\right).$$

We already did the calculations required for this bound:

$$\begin{aligned} D(X) - E(X) &= E(X^2) - E(X)^2 - E(X) \\ &= 2n(n-1)(1-p)^{2n} + 2n^2(1-p)^{2n-1} - \left(2n(1-p)^n\right)^2 \\ &= 2n(1-p)^{2n}\left(n - 1 + \frac{n}{1-p} - 2n\right) = O(n\, e^{-2pn})o(n) = o(e^{-2f(n)}), \end{aligned}$$

$$\sum_{v\in A\cup B} E\left(I_v\right)^2 \sim \frac{2e^{-2f(n)}}{n} = o\left(e^{-2f(n)}\right).$$

Put together,

$$d_{TV}(X, \text{Poisson}(\lambda)) \le \min(1, \lambda^{-1})\; o\left(e^{-2f(n)}\right).$$

If $f(n) \to c$, then $d_{TV} \to 0$ (notice that while the small o guarantees the convergence, it says nothing about how fast it is). Consequently, the number X of isolated vertices converges to Poisson(λ), and we know $\lambda = E(X) = 2e^{-c}$, and thus $P[X = 0] \to \exp(-2e^{-c})$.

11 Random bipartite graphs and perfect matchings

Consider a random bipartite graph $G(n,n,p)$ with parts of size n and edge probability p. Existence of perfect matchings (PMs) is covered by the following theorem (Erdös, Rényi 1964). Let $p = \frac{\ln n + f(n)}{n}$.

$$P\left[G(n,n,p) \text{ has a perfect matching}\right] = \begin{cases} 0 & \text{if } f(n) \to -\infty, \\ \exp(-2e^{-c}) & \text{if } f(n) \to c, \\ 1 & \text{if } f(n) \to \infty. \end{cases}$$

Notice that this threshold coincides with the one for isolated vertices. That's great: a graph with an isolated vertex does not have a PM, so for $f(n) \to -\infty$, we immediately get G has no PM w.h.p., i.e. $P \to 0$ (and we don't need to employ the second moment method). In the rest of the chapter, we will prove that if $f(n) \to \infty$, G has a PM w.h.p.

A bipartite graph has a PM if and only if it does not contain an obstacle according to Hall's marriage theorem, i.e. a set S of vertices in one part coupled with its neighbourhood $N(S)$ such that $|N(S)| < |S|$. Among all the obstacles, we can look at a minimal one that satisfies

(i) $|S| = |N(S)| + 1$,

(ii) $|S| \leq n/2$,

(iii) every vertex in $N(S)$ is adjacent to at least two vertices in S.

U: Prove that such S exists iff G does not have a PM.

Solution: If $u \in N(S)$ only had one neighbour v in S, then $S - v$ is also an obstacle because its neighbourhood is $N(S) - u$. If $|S| > |N(S)| + 1$, we just remove some vertices from S. If $|S| > n/2$, then the complement of $N(S)$ in the part not containing S is also an obstacle, and a smaller one.

Next, we show that the expected number $E(X)$ of obstacles is $o(1)$, which by first moment method would imply that $P[X > 0] \to 0$, i.e. a random bipartite graph $G(n,n,p)$ has a PM w.h.p.

Assume S is an obstacle satisfying the three condition given above. If $|S| = 1$, S is an isolated vertex; we already know that those disappear above the threshold, so they contribution to $E(X)$ is $o(1)$.

If $|S| = 2$, then S consists of two vertices joined to a single vertex in $N(S)$; we call this structure a cherry. The number of cherries is $O(n^3)$ and a fixed cherry exists with the probability $p^2(1-p)^{2n-2}$. The contribution of cherries to $E(X)$ is thus at most

$$T(n) = O(n^3 p^2 (1-p)^{2n-2}) = O(n \cdot (np)^2 e^{-2(np)}) = O\left(nx^2 e^{-2x}\right)$$

where $x = np = \ln n + f(n) \to \infty$. Since the exponential function dominates any polynomial, we have $x^2 e^{-2x} = o(e^{-1.5x})$, and thus $T(n) = O(n \cdot n^{-1.5} \cdot e^{-1.5f(n)}) = o(1)$.

Remark: Without the condition (iii), we can only assume one edge between S and $N(S)$, and thus $T(n) \sim n^2 x e^{-2x}$. Since $e^{-2x} = n^{-2} e^{-2f(n)}$, we have $T(n) = (\ln n + f(n)) e^{-2f(n)}$ and a slowly growing $f(n)$ is insufficient to overcome $\ln n$ and make it $o(1)$. It actually helps for $|S|$ somewhere up to $n/\ln n$: if we assumed just one edge from every vertex of $N(S)$, which we can thanks to each vertex being neighbour of a vertex in S, we would get

$s^{s-1}p^{s-1}$; if we assume two edges, we get $\binom{s}{2}^{s-1}p^{2(s-1)}$ (see the relevant bullet point below). So the additional terms are roughly $(\frac{s-1}{2})^{s-1}p^{s-1} \approx (sp/2)^s$, and for $|S| = 2n/\ln n$, we have $sp/2 \approx 1$ (note that $\approx$ is only to give a rough idea, we cannot use $\sim$ since the exponent depends on n, too, and it is not correct to just estimate the basis of the exponentiation if the exponent is also a function of n and not fixed). For large $|S|$, applying (iii) increases the upper bound on the contribution of such S, but not too much, and it simplifies the calculations somewhat, as we will see, so we still apply (iii).

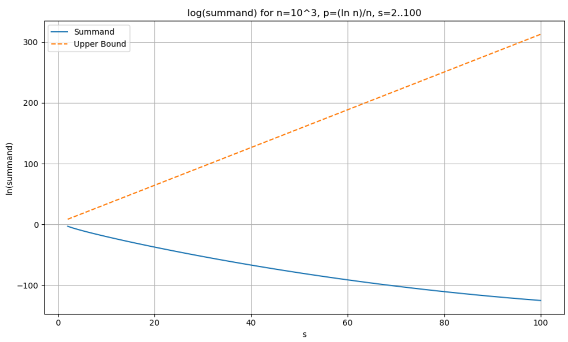
For $|S| \geq 3$:

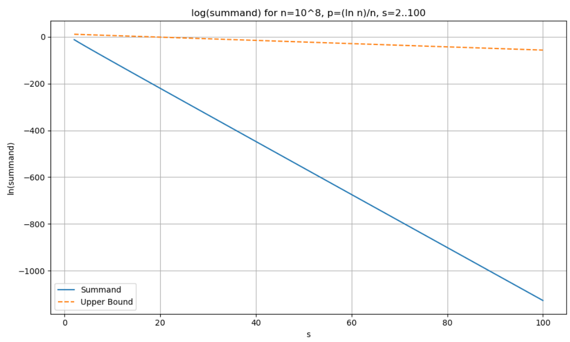
- There are at most $2\binom{n}{s}$ sets S of size s, each of them having $\binom{n}{s-1}$ possible neighbourhoods, and for a given pair S, $N(S)$ we need the probability of it being an obstacle.
- According to (iii), for each of $s-1$ vertices in $N(S)$ we pick two neighbours in S; there are $\binom{s}{2}^{s-1}$ ways to do it and the probability that all the required edges exist is p^{2s-2}.
- None of the $s(n-(s-1))$ edges between S and the complement of $N(S)$ in the other part must exist. (For the remaining edges, we don't care if they are present or not.)

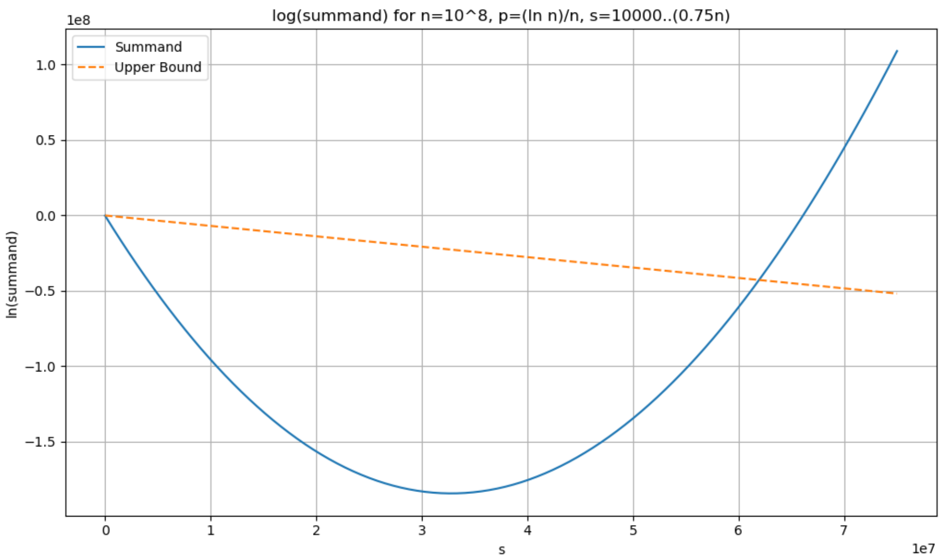
Altogether, the contribution of obstacles of size ≥ 3 is at most

$$V = \sum_{s=3}^{n/2} 2\binom{n}{s}\binom{n}{s-1}\binom{s}{2}^{s-1} p^{2s-2}(1-p)^{s(n-s+1)}.$$

Many sums can be dealt with by guessing which part is small and then using very rough bounds to prove that that part is negligible. A guidance for this can be provided by plotting the summands, so that's what we do now. The charts also include an upper bound we will derive soon.

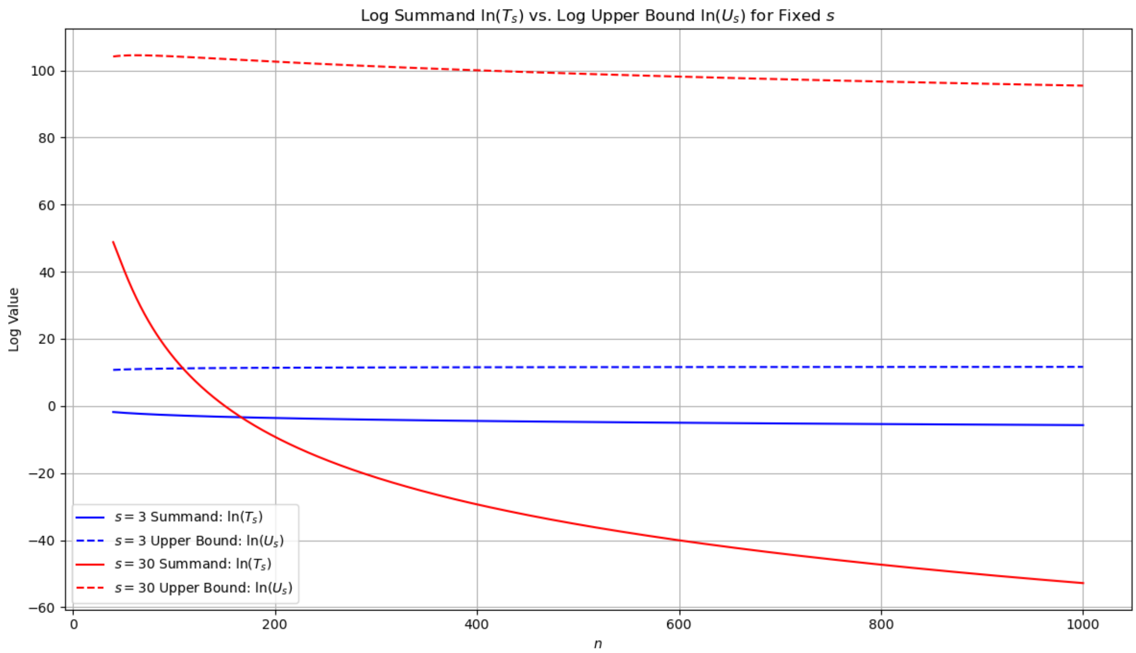


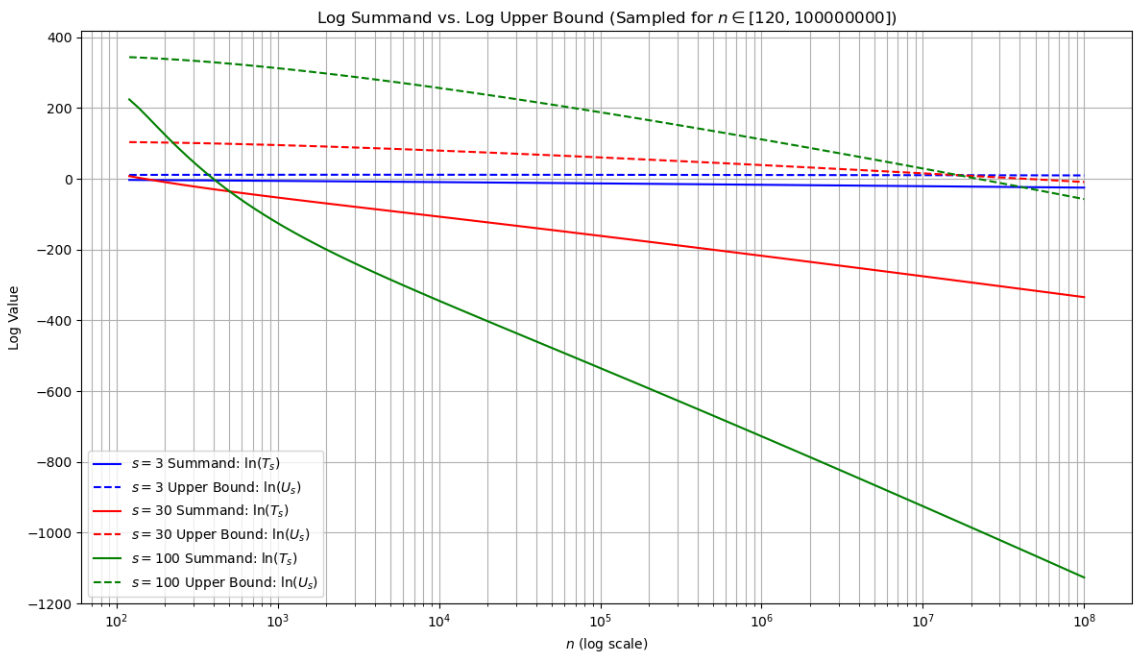




It seems that at some point, the summands start growing quickly with s (superexponential growth suggested by the log plot), so we do need an upper bound like (ii) on s. Visualizations reveal this trend only for sufficiently large values of n and s. Given that the summands seem smallest in the middle but large at both extremes of the $s \in [3, n/2]$ range, we cannot simplify the sum by cutting off the tails or focusing on a narrow central portion.

We derive upper bounds for different components of the summands with the goal of isolating terms dependent on s within the summation and moving the remaining terms before the sum. It turns out that we can replace some occurences of s with $n/2$ in such a way that we get $\sum z^s$ with z depending only on n, not s, and then we can sum it as a geometric series. This can only work if the summand for any fixed s in our range is $o(1)$ as $n \to \infty$. Computer-generated plots suggest it is true.





For estimating large binomial coefficients, we use

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \le \frac{n^k}{k!} \le \left(\frac{en}{k}\right)^k$$

which is true because $e^k = \sum_{i=0}^{\infty} \frac{k^i}{i!} > \frac{k^k}{k!}$, thus $\frac{1}{k!} \le \frac{e^k}{k^k}$.

Next,

$$s(n-s+1) = sn - s^2 + s \ge sn - s\frac{n}{2} + s = \frac{1}{2}sn \qquad \text{and} \qquad \forall x \in \mathbb{R} \quad 1 - x \le e^{-x},$$

hence

$$(1-p)^{s(n-s+1)} \le (1-p)^{\frac{1}{2}sn} \le e^{-p\left(\frac{1}{2}sn\right)}.$$

Put together,

$$\begin{aligned} V &= \sum_{s=3}^{n/2} 2\binom{n}{s}\binom{n}{s-1}\binom{s}{2}^{s-1} p^{2s-2}(1-p)^{s(n-s+1)} \\ &\le 2\sum_{s=3}^{n/2}\left(\frac{en}{s}\right)^s\left(\frac{en}{s-1}\right)^{s-1}\frac{s^{s-1}(s-1)^{s-1}}{2^{s-1}}p^{2s-2}e^{-\frac{1}{2}psn} \\ &\le \frac{4}{enp^2}\sum_{s=3}^{n/2}\frac{e^{2s}(np)^{2s}}{2^s s}e^{-\frac{1}{2}s(np)}. \end{aligned}$$

The property "G has a PM" is monotone, i.e. adding edges to a graph that already has a PM does not destroy the PM, and thus it is sufficient to provide a proof for some value of p and we know the result will also hold for any larger value of p. We are also assuming $f(n) \to \infty$, so $np = \ln n + f(n)$ can be assumed to be between $\ln n$ and $2\ln n$ for a sufficiently large n. We use these bounds in estimating V; for instance, $\exp(-\frac{1}{2}(np)) \le 1/\sqrt{n}$. We also use $1/s \le 1/3$.

$$\begin{aligned} V &\le \frac{4n}{3e(\ln n)^2}\sum_{s=3}^{n/2}\left(\frac{e^2(np)^2}{2}e^{-\frac{1}{2}(np)}\right)^s \\ &\le \frac{4n}{3e(\ln n)^2}\sum_{s=3}^{n/2}\left(\frac{2e^2(\ln n)^2}{\sqrt{n}}\right)^s \\ &\le \frac{4n}{3e(\ln n)^2}\left(\frac{2e^2(\ln n)^2}{\sqrt{n}}\right)^3\cdot\sum_{s=0}^{\infty}\left(\frac{2e^2(\ln n)^2}{\sqrt{n}}\right)^s \\ &= O\left(\frac{n}{\ln^2 n}\cdot\frac{\ln^6 n}{n^{3/2}}\right) = O\left(n^{-1/2}\cdot\ln^4 n\right) = o(1). \end{aligned}$$

Remark: Notice that none of the bounds we used is asymptotic, i.e. all of them hold everywhere we used them (if this was not true we might run into trouble with our estimations since s depends on n too, thus we cannot carelessly manipulate the base [1]). The base of the s-th powers in the summands is $o(1)$, so for any large enough n it is below a constant smaller than 1, and for such n our sum can be bounded from above by a convergent geometric series (which we can extend to infinity since we don't care what the actual sum of the series is). The convergence of the base to $o(1)$ is slow, though: it only gets below 1 for $n > 17 \cdot 10^6$.

[1] An example of this kind of error: $e = \left(1 + \frac{1}{n}\right)^n \sim (1+0)^n = 1$.

If we did not analyze cherries separately, we would get $O(\ln^2 n)$ as the upper bound — too big.

If we did not impose (ii), we could not estimate $s(n-s+1)$; without the $\exp(-\frac{1}{2}(np))$, some individual summands would not be $o(1)$, especially not those for s close to n.