Some Advanced Algorithms Exercises

Nicola Gulmini

1 NP-completeness

Theorem 1.1 (Vertex Cover)**.** $VC <_p IS$

Proof. If for the reductions $CLIQUE <_p IS$ and $CLIQUE <_p VC$ we used the complementary graph, in this case it is not necessary. To prove the theorem, it is sufficient to show that a vertex cover V' of size k induces an independent set in the same graph of size $|V| - k$, simply by taking the nodes of the set V/V'. Formalizing the definition of the reduction function

$$x = \langle G = (V, E), k \rangle \in VC \iff f(x) = \langle G = (V, E), |V| - k \rangle \in IS$$

where the function does not modify the graph but only the required size, so it is polynomial time computable (ptc). Now we have to prove what has been said and we do it through logical equivalences that will allow us to prove both sides of the statement simply by reversing the direction of reading:

$$\begin{aligned} x \in VC &\iff \exists V' \subseteq V, |V'| = k : \forall (u, v) \in E \Rightarrow (u \in V') \vee (u \in V') \\ &\iff \exists V' \subseteq V, |V'| = k : \exists u, v : (u \notin V') \wedge (u \notin V') \Rightarrow (u, v) \notin E \\ &\iff \exists V' \subseteq V, |V'| = k : \exists u, v : (u \in V/V') \wedge (u \in V/V') \Rightarrow (u, v) \notin E \end{aligned}$$

and defining $V'' = V/V'$ yields exactly the definition of independent sets of size $|V''| = |V| - k$, with only logical equivalences. □

Theorem 1.2 (Independent Set)**.** *Given a graph $G = (V, E)$ with $n = |V|$ and $E = |E|$, if $m < \frac{n}{2}$ then an IS with size at least $\frac{n}{2}$ always exists.*

Proof. A graph is connected if and only if all its nodes are connected. That is $\forall u \neq v \in V, \exists (u, v) \in E$, and this implies at worst each node has only one edge connecting it to another, hence $m \geq n - 1$. In this case we have $m < \frac{n}{2}$ which necessarily means having multiple connected components. Between different components the nodes are not connected, so the independent set of the graph will have size equal to the independent sets of the individual

connected components. Each independent set has size ≥ 1 because if we take only one node, there are no edges connecting it to other nodes. It follows that the size of the independent set will be at least the number of connected components. Trivially, if $m = 0$ the size of the independent set will be n, if $m = \frac{n}{2} - 1$ then there will be 2 connected components and the size of the independent set will be *at least* 2. It only remains to show that a graph with $m < \frac{n}{2}$ has at most $\frac{n}{2}$ connected components to conclude. This is simple: assign each edge to only two nodes maximizes the number of connected components, which will be fewer than $\frac{n}{2}$ as the edges. In that case the size of the independent set is equal to the number of connected components. With fewer edges, the number of components increases and thus the size of the independent set increases as well. □

2 Approximation Algorithms

Theorem 2.1 (Vertex Cover)**.** *The set of all nodes forming the edges of a maximal matching are a vertex cover.*

Proof. Let A be a maximal matching, so

$$A = \{\{u, v\} : u, v \in V \wedge \forall e_1 \neq e_2 \in A, e_1 \cup e_2 = \varnothing\} \tag{1}$$

and $\forall e \in E \setminus A$, $A \cup \{e\}$ is not a matching. Let $V' = \{u \in V : \exists e \in A : u \in e\}$. Let's show that each $e \in E$ is covered by V':

1. if $e = \{u, v\} \in A$, by definition $u, v \in V'$, thus e is covered by V';
2. if $e \in E \setminus A$, $A \cup \{e\}$ is not a matching. By definition of matching this means that there must be an edge $e' \in A$ with whom e shares one endpoint, i.e. $e \cap e' = \{w\} \neq \varnothing$. Since $e' \in A$, by the previous point $w \in V'$, so e is covered by V'.

□

Exercise 2.1 (Subset Sum)**.** *Write a 2-approximation algorithm for Subset Sum, linear in the instance size $|\langle S, t\rangle|$.*

Answer. Search for the maximum (it is done in linear time in the size of S) element of S and place it first, then the algorithm is the same as the one seen in class, with the exact same approximation factor analysis, except that it is not necessary to sort S, so the complexity is no longer $O(|\langle S, t\rangle|)$ but only $O(|\langle S, t\rangle|)$. □

Exercise 2.2 (Resource Selection)**.** *Extend to F subsets for each task and obtain a F-approximation algorithm.*

Answer. The problem is

$$\begin{cases} \mathcal{I} : \langle R, \{Y_t^j, t \in [1,m], j \in [1,F]\}, k \rangle \\ \mathcal{Q} : \exists Y_t^{s_t} \ \forall t \in [1,m] : |\bigcup_{t=1}^m Y_t^{s_t}| \leq k? \end{cases} . \tag{2}$$

Suppose that F_RS is NPH. An ILP formulation of the problem could be:

$$\begin{cases} \min \sum_{i=1}^n x_i & \\ \sum_{k=1}^F y_j^k \geq 1 & \forall j \in [1,m] \\ x_i \geq y_j^k & \text{if } r_i \in Y_j^k \\ x_i, y_j^k \in \{0,1\} & \forall i \in [1,n], j \in [1,m], k \in [1,F]. \end{cases} \tag{3}$$

Consider the optimal solution $\mathbf{x}^*, \mathbf{y}^*$ of the continuous relaxation and the function

$$\hat{x}_i = \text{round}(x_i^*) = \begin{cases} 1 & \text{if } x_i^* \geq 1/F \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and same for $\mathbf{y}^*$. Note that the rounding solution is feasible: in $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ at least one element has to be $\geq 1/F$ because

$$\sum_{k=1}^F y_j^{k*} = y_j^{1*} + \cdots + y_j^{F*} \geq 1 \ \forall j \in [1,m] \Rightarrow \exists k_j \in [1,F] : y_j^{k_j*} \geq 1/F \tag{5}$$

so $x_i^* \geq y_j^{k_j*} \geq 1/F \Rightarrow \text{round}(x_i^*) = 1$. Now observe that $c(\mathbf{x}^*, \mathbf{y}^*) \leq F \cdot c(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ so

$$\rho = \frac{c(\mathbf{x}^*, \mathbf{y}^*)}{c(\hat{\mathbf{x}}, \hat{\mathbf{y}})} \leq \frac{F \cdot c(\hat{\mathbf{x}}, \hat{\mathbf{y}})}{c(\hat{\mathbf{x}}, \hat{\mathbf{y}})} = F. \tag{6}$$

□

3 Algorithms for number-theory problems

Property 3.1. $\forall d \in \mathbb{Z} \setminus \{0\}$, $d|0$.

Proof. $d|0 \iff \exists k \in \mathbb{Z} : 0 = kd$, and it holds for $k = 0$, $\forall d \in \mathbb{Z} \setminus \{0\}$. □

Property 3.2. $(a \neq 0) \wedge (d|a) \Rightarrow |d| \leq |a|$.

Proof. $d|a \iff a = kd \iff |a| = |kd|$, so $|d| \leq |kd| = |a|$. The hypothesis $(a \neq 0)$ is necessary because if $a = 0$ then $d|a$ $\forall d \in \mathbb{Z} \setminus \{0\}$ and $|d| > |a| = 0$. □

Property 3.3. $(a|b) \wedge (b|a) \Rightarrow |a| = |b|$.

Proof. $a|b \iff \exists k_1 \in \mathbb{Z} : b = k_1 a$ and $b|a \iff \exists k_2 \in \mathbb{Z} : a = k_2 b = k_2 k_1 a \iff k_1 k_2 = 1$. There are two cases:

1. $k_1 = k_2 = 1 \Rightarrow a = b$;
2. $k_1 = k_2 = -1 \Rightarrow a = -b$

so $|a| = |b|$. □

Property 3.4. $(d|a) \wedge (d|b) \Rightarrow d|(a \pm b)$.

Proof. $d|a \iff \exists k_1 \in \mathbb{Z} : a = k_1 d$ and $d|b \iff \exists k_2 \in \mathbb{Z} : b = k_2 d$. Summing the two equations we get $a + b = (k_1 + k_2)d \iff d|(a + b)$, and subtracting we get $(a - b) = (k_1 - k_2)d \iff d|(a - b)$. □

Property 3.5. $\forall x, y \in \mathbb{Z}, (d|a) \wedge (d|b) \Rightarrow d|(ax + by)$.

Proof. Same reasoning as the previous property. □

Property 3.6. $\forall a, b \in \mathbb{Z} \setminus \{0\} \Rightarrow \gcd(a, b) \in [1, \min\{|a|, |b|\}]$.

Proof. We have to show that:

1. $\gcd(a, b) \geq 1$;
2. $\gcd(a, b) \leq \min\{|a|, |b|\}$.

Recalling the definition of greatest common divisor

$$\gcd(a, b) = \begin{cases} \max\{d > 0 : (d|a) \wedge (d|b)\} & \forall a, b \in \mathbb{Z} : |a| + |b| > 0 \\ 0 & a = b = 0 \end{cases} \tag{7}$$

we know that $d \geq 1$ because of the hypothesis $a, b \neq 0$, so $\gcd(a, b) \geq 1$.

From the property 3.2,

$$((d|a) \wedge (d|b)) \Rightarrow ((|d| \leq |a|) \wedge (|d| \leq |b|)) \tag{8}$$

and $(d > 0) \Rightarrow (|d| = d)$, so $d \leq \min\{|a|, |b|\}$ and finally $\gcd(a, b) \leq \min\{|a|, |b|\}$. □

Property 3.7. $\gcd(a, 0) = |a|$.

Proof.

$$\begin{aligned}\gcd(a,0) &= \max\{d>0 : (d|a) \wedge (d|0)\} \\ &= \max\{d>0 : d|a\} \\ &= \max\{d>0 : |d| \leq |a|\} = |a|. \end{aligned} \tag{9}$$

□

Property 3.8. $\gcd(a,b) = \gcd(b,a) = \gcd(|a|,|b|)$.

Proof. Omitted. □

Property 3.9. $\gcd(a,ka) = |a| \ \forall a,k \in \mathbb{Z}$.

Proof. Omitted. □

Lemma 3.1. $(n|ab) \wedge (\gcd(n,a)=1) \Rightarrow n|b$.

Proof. $n|ab \iff ab = kn \iff ab \mod n = 0 \iff a \mod nb \mod n = 0$ but $\gcd(a,n) = 1 \Rightarrow a \mod n \neq 0$ so $b \mod n = 0 \iff n|b$. □

Theorem 3.2. *a^{-1} is unique in $\mathbb{Z}_n^*$*

Proof. It is sufficient to show that $\nexists x \neq y \in \mathbb{Z}_n^* : (xa \mod n = 1) \wedge (ya \mod n = 1)$. By contraddiction, suppose that $\exists x \neq y \in \mathbb{Z}_n^*$: $(xa \mod n = 1) \wedge (ya \mod n = 1)$.

$$\begin{aligned} xa \mod n = ya \mod n &\iff xa \mod n - ya \mod n = 0 \\ &\iff (ax - ay) \mod n = 0 \\ &\iff ax - ay = kn \text{ for some } k \in \mathbb{Z} \\ &\iff a(x-y) = kn \end{aligned} \tag{10}$$

and $a \in \mathbb{Z}_n^* \Rightarrow \gcd(a,n) = 1$, so for the previous lemma $(x-y)|n$.

$$\begin{aligned} (x-y)|n &\iff (x-y) \mod n = 0 \\ &\iff x \mod n - y \mod n = 0 \\ &\iff x \mod n = y \mod n \end{aligned} \tag{11}$$

where the last quation holds because $x \mod n = x \in \mathbb{Z}_n^*$ and $y \mod n = y \in \mathbb{Z}_n^*$. We have obtained $x = y$, that contradicts the hypothesis. □

4 Randomized techniques and their applications

Exercise 4.1. *Implement* $RANDOM(\{0,1,2\})$ *using the* $BIAS()$ *primitive.*

Answer. Just make 3 calls at $BIAS()$, one for each iteration and make the following association: $001 \to 0$, $010 \to 1$ and $100 \to 2$ which all have probability $p(1-p)^2$. The number of iterations is a geometric random variable with parameter $3p(1-p)^2$, so the expected number of calls is $1/(p(1-p)^2)$. □

Exercise 4.2. *Implement* $RANDOM(\{0,1,\ldots,n-1\})$ *using the primitive* $RANDOM(\{0,1\})$

Answer. The $RANDOM(\{0,1\})$ primitive is like $BIAS()$, but with p known and equal to $1/2$. So we can use the previous exercise to conclude that the expected number of total calls to implement $RANDOM(\{0,1,\ldots,n-1\})$ is $\mathbb{E}(s(p)) = \frac{1}{1/2(1-1/2)^{n-1}} = 2^n$. □

Exercise 4.3. *Use the Monte Carlo method to estimate the value of* π.

Answer. Let us consider the $[-1,1]\times[-1,1] \subset \mathbb{R}^2$ square and its inscribed circumference with $r = 1$. Then, generating k points $(x_i, y_i) \leftarrow \text{RANDOM}([-1,1]\times[-1,1])$ we can define the random variable $Z_i = 1 \iff \sqrt{x_i^2 + y_i^2} \leq 1$ which means that the extracted point is inside the circumference. Since the area of the square is $\|[-1,1]\|^2 = 4$ and the area of the circumference is $\pi r^2 = \pi$, the probability that an extracted point (x_i, y_i) lays inside the circumference is

$$\mathbb{P}((x_i, y_i) \text{ inside the circumference}) = \frac{\text{area of crf}}{\text{area of the square}} = \frac{\pi}{4}$$

At the end of the calls, we sum up all the Z_i's, each with probability $p_i = \pi/4$, actually defining a binomial random variable $Z = \sum_{i=1}^{k} Z_i$ with expected value $\mathbb{E}(Z) = k\pi/4$.

Finally, giving the Monte Carlo Counting Theorem considering a known $\alpha = \pi/4$ and so $k_{\min} = \frac{3\cdot 4}{\pi\varepsilon^2}\ln(2/\delta) \approx \frac{4}{\varepsilon^2}\ln(2/\delta)$ we have

$$\frac{4}{\pi}\left|\frac{1}{k}\mathbb{E}(Z) - \frac{\pi}{4}\right| > \varepsilon$$

with probability δ. To give some numbers, with minimum absolute error of $\varepsilon = 0.01$ and a confidence of 95% (which is $\delta = 0.05$) we have to perform about $4\ln(2/0.05)/0.0001 \approx 150000$ calls. □

4.1 Balls into bins, occupancy problems

Exercise 4.4. *Suppose that we toss balls into b bins until some bin contains two balls. Each toss is independent, and each ball is equally likely to end up in any bin. What is the expected number of ball tosses?*

Answer. From the definition of the expected value, let X be the desired number of ball tosses, its expectation is

$$\mathbb{E}(X) = \sum_{x=2}^{b+1} x\mathbb{P}(x)$$

where we start from $x = 2$ because for $x = 1$ the probability of collision is null, and we stop at $x = b + 1$ because then the probability is 1. $\mathbb{P}(x)$ is the probability to obtain a collision at the x-th toss, after $x - 1$ tosses without collisions. For this reason, we have to take into account also the probability to have no collisions in $x - 1$ tosses, that is

$$1 \cdot \left(1 - \frac{1}{b}\right) \cdot \dots \cdot \left(1 - \frac{x-2}{b}\right) = \prod_{j=1}^{x-2} \left(1 - \frac{j}{b}\right).$$

The probability to have a collision only at the x-th ball toss, after $x-1$ tosses without any collisions, is $(x-1)/b$ and the expected number becomes

$$\mathbb{E}(X) = \sum_{x=2}^{b+1} x \cdot \frac{(x-1)}{b} \prod_{j=1}^{x-2} \left(1 - \frac{j}{b}\right).$$

If we want to give a bound at this formula, using $(1 - j/b) < e^{-j/b}$, so it becomes

$$\mathbb{E}(X) < \frac{1}{b} \sum_{x=2}^{b+1} x(x-1) \prod_{j=1}^{x-2} e^{-j/b} = \frac{1}{b} \sum_{x=2}^{b+1} x(x-1) e^{-\frac{(x-2)(x-3)}{2b}}.$$

Another bound (that we call *bound 2*) that is not so informative, is given by $(1 - j/b) \leq (1 - 1/b)$ for $j \geq 1$. The formula becomes

$$\mathbb{E}(X) \leq \frac{1}{b} \sum_{x=2}^{b+1} x(x-1)(1-1/b)^{x-2} < \frac{1}{b} \sum_{x=2}^{b+1} x(x-1) e^{(2-x)/b}$$

where in the last statement we used $(1-1/b)^{x-2} < e^{-(x-2)/b}$. In Fig. 1 there is a simulation with $b \in [0, 200]$ bins and 100 trials to estimate the expected value.

□

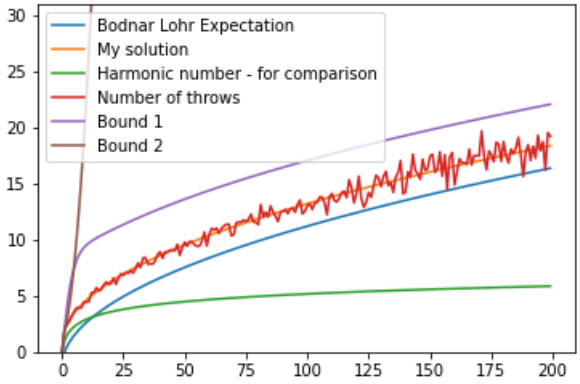

Figure 1: The bound 2 is useless. The harmonic number is used for comparison, and the *Bodnar Lohr Expectation* is taken from `https://sites.math.rutgers.edu/~ajl213/CLRS/Ch5.pdf`. The script to generate this plot is in `https://github.com/nicolagulmini/balls_and_bins`.

4.2 Generalized Coupon Collecting

Suppose that there is a collection of n uniformly distributed cards and for each packet there are $m \leq n$ (usually $m \ll n$) different cards. Analyze the number of packets to buy to complete the collection in expectation and also in high probability.

4.2.1 Expectation

Suppose that j cards of the collection are already found. What is the probability to find at least a new card with the next packet?

$$\mathbb{P}(\text{at least a new card between the } m \text{ cards of the pack} | j \text{ found}) = \\ 1 - \mathbb{P}(\text{only already found cards} | j \text{ found}) = 1 - (j/n)^m.$$

Let Z_j be the number of packets to buy before the one that contains at least a new card, given j cards found. Here we have to distinguish the probability to find exactly one new card, or to find *at least* a new card. To follow the same approach as the case $m = 1$, we will sum up all the Z_j but taking into account that the results will be overestimated, because in some packets there could be more new cards, and in those cases we jump from j found cards to $j' \geq j + 2$, skipping the case in which $j + 1$ cards are found. Instead, we will

count all the j's. The number of packets Z_j is a geometric random variable

$$Z_j \sim \mathcal{G}\left(1 - \left(\frac{j}{n}\right)^m\right) \Rightarrow \mathbb{E}(Z_j) = \frac{1}{1 - \left(\frac{j}{n}\right)^m}$$

and the expected number of packets to complete the collection is

$$\mathbb{E}(Z) \leq \mathbb{E}\left(\sum_{j=0}^{n-1} Z_j\right) = \sum_{j=0}^{n-1} \mathbb{E}(Z_j) = \sum_{j=0}^{n-1} \frac{1}{1 - \left(\frac{j}{n}\right)^m}.$$

4.2.2 High probability

Since the Z_j's are independent but not identically distributed, and Z is a sum of geometric (not bernoulli) random variables, we cannot use the Chernoff's Bounds.

Let's fix a card i and the event $E_i =$ "first r packs do not contain i". We have

$$\mathbb{P}(E_i) = \mathbb{P}(i \text{ is not in a packet})^r = (1 - \mathbb{P}(\text{a packet contains } i))^r$$

and the probability that a packet contains i is

$$\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-m+1} = \sum_{k=0}^{m-1} \frac{1}{n-k} = \sum_{t=n-m+1}^{n} \frac{1}{t} = H(n) - H(n-m)$$

so

$$\mathbb{P}(E_i) = \left(1 - \sum_{t=n-m+1}^{n} \frac{1}{t}\right)^r < e^{-r\sum_{t=n-m+1}^{n} \frac{1}{t}}.$$

Now we want to find a r that ensures, with high probability, to complete the collection:

$$\begin{aligned}
&\mathbb{P}(\text{complete the collection with } r \text{ packets}) = \\
&= 1 - \mathbb{P}(\text{at least one card not found in } r \text{ packets}) = \\
&= 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \geq 1 - \sum_{i=1}^{n} \mathbb{P}(E_i) > 1 - ne^{-r\sum_{t=n-m+1}^{n} \frac{1}{t}}
\end{aligned}$$

and, in order to obtain the high probability, r has to satisfy:

$$1 - ne^{-r\sum_{t=n-m+1}^{n} \frac{1}{t}} = 1 - \frac{1}{n^d} \iff e^{-r\sum_{t=n-m+1}^{n} \frac{1}{t}} = \frac{1}{n^{d-1}}.$$

We obtain $r = \frac{(d-1)\ln n}{H(n)-H(n-m)}$. Note that with $m = 1$, since $H(n) - H(n-1) = 1/n$, the result is $r|_{m=1} = n(d-1)\ln n$. In Fig. 2 there are two simulations with increasing n and increasing m, fixing $d = 2$ and estimating the expectations with 100 trials.

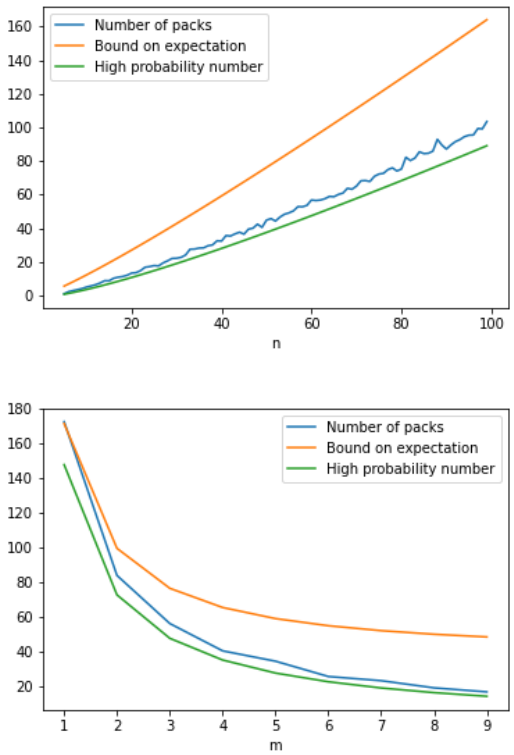

Figure 2: The script to generate these plots are in `https://github.com/nicolagulmini/generalized_coupon_collecting`.

Exercise 4.5. *Given a positive integer k, consider the square $[0,1] \times [0,1] \subset \mathbb{R}^2$ and its natural partitioning in k^2 subsquares with $1/k$ length sides. Determine a number of calls at the RANDOM($[0,1] \times [0,1]$) procedure that guarantees with probability at least $1 - 1/k^2$ that at least one point is extracted from each subsquare.*

Answer. The event E_m ="at least one point in each subsquare in m calls" is the complementary of $\bar{E}_m$ ="there are not points in some subsquare in m calls":

$$\mathbb{P}(E_m) = 1 - \mathbb{P}(\bar{E}_m).$$

The event $\bar{E}_m$ is the union of the events e_i ="no points in the i-th subsquare in m calls" for each $i \in [1, k^2]$:

$$\mathbb{P}(\bar{E}_m) = \mathbb{P}\left(\bigcup_{i=1}^{k^2} e_i\right) \leq \sum_{i=1}^{k^2} \mathbb{P}(e_i)$$

where the last statement holds for the union bound. The probability of the event e_i is $(1 - 1/k^2)^m$, so

$$\mathbb{P}(\bar{E}_m) \leq \sum_{i=1}^{k^2} \left(1 - \frac{1}{k^2}\right)^m = k^2 \left(1 - \frac{1}{k^2}\right)^m.$$

Now we want a value for m such that the probability of the event E_i is at least $1 - 1/k^2$, so

$$\mathbb{P}(E_m) = 1 - \mathbb{P}(\bar{E}_m) \geq 1 - k^2 \left(1 - \frac{1}{k^2}\right)^m \equiv 1 - \frac{1}{k^2}$$

so we want to find m that satisfies

$$k^2 \left(1 - \frac{1}{k^2}\right)^m = \frac{1}{k^2} \iff m = \log_{\left(1-\frac{1}{k^2}\right)}(1/k^4) = -4 \log_{\left(1-\frac{1}{k^2}\right)} k$$
$$= \frac{-4 \ln k}{\ln(1 - 1/k^2)}$$

and, using $(1 - 1/k^2) < e^{-1/k^2}$, we can obtain a lower bound for m

$$m = \frac{-4 \ln k}{\ln(1 - 1/k^2)} > \frac{-4 \ln k}{\ln(e^{-1/k^2})} = 4k^2 \ln k.$$

Note that the same result can be obtained in the following way

$$k^2 \left(1 - \frac{1}{k^2}\right)^m < k^2 e^{-m/k^2} \equiv \frac{1}{k^2} \iff -\frac{m}{k^2} < -4 \ln k \iff m > 4k^2 \ln k.$$

□

4.3 Occupancy problems

We have m identical balls and n bins/boxes. We throw the balls randomly and mutually independently. At the end of these $m \geq n$ tosses, we want to analyze the maximum number of balls in any one bin.

First, let x_b be the random variable that models the number of balls in the bin $b \in [1, n]$ at the end of the tosses. We have

$$\mathbb{P}(x_b = i) = \binom{m}{i} \frac{1}{n^i} \left(1 - \frac{1}{n}\right)^{m-i}.$$

We want to study, for some fixed $k \in [1, m]$, the probability that there are *at least* k balls in the box b. We have

$$\mathbb{P}(x_b \geq k) = \sum_{i=k}^{m} \binom{m}{i} \frac{1}{n^i} \left(1 - \frac{1}{n}\right)^{m-i}.$$

We now bound to get a simpler formula to handle. The first simplification we make is on the binomial coefficient via the Stirling approximation

$$\binom{m}{i} \leq \left(\frac{em}{i}\right)^i, \forall m \geq 1, \forall i \in (0, m]$$

by which we obtain the following

$$\begin{aligned}
\mathbb{P}(x_b \geq k) &\leq \sum_{i=k}^{m} \left(\frac{em}{i}\right)^i \frac{1}{n^i} \left(1 - \frac{1}{n}\right)^{m-i} \\
&= \sum_{i=k}^{m} \left(\frac{em}{ni}\right)^i \left(1 - \frac{1}{n}\right)^{m-i} \\
&\leq \sum_{i=k}^{m} \left(\frac{em}{ni}\right)^i \qquad (12)
\end{aligned}$$

where in the last step we used $(1 - 1/n)^{m-i} \leq 1$. Now we can extend the

sum from m to $+\infty$ and replace i in the denominator with k:

$$\begin{aligned}\mathbb{P}(x_b \geq k) &\leq \sum_{i=k}^{+\infty} \left(\frac{em}{nk}\right)^i \\ &= \sum_{t=0}^{+\infty} \left(\frac{em}{nk}\right)^{t+k} \\ &= \sum_{t=0}^{+\infty} \left(\frac{em}{nk}\right)^t \left(\frac{em}{nk}\right)^k \\ &= \left(\frac{em}{nk}\right)^k \sum_{t=0}^{+\infty} \left(\frac{em}{nk}\right)^t. \end{aligned} \tag{13}$$

In this way we have made the base independent from the index, obtaining a geometric series. Fixing k so that the argument is $< 1/2$

$$\frac{em}{nk} < \frac{1}{2} \iff k > \frac{2em}{n}$$

we can conclude that the result of the geometric series is less than 2:

$$\sum_{t=0}^{+\infty} \left(\frac{em}{nk}\right)^t = \frac{1}{1 - \frac{em}{nk}} < 2$$

and so $\mathbb{P}(x_b \geq k) < 2\left(\frac{em}{nk}\right)^{\frac{2em}{n}} < 2\left(\frac{em}{nk}\right)^k$. To talk about high probability it is necessary to find a $k^\star$ that satisfies

$$\begin{aligned}\mathbb{P}(\text{in some bin there are } \geq k \text{ balls}) &= \mathbb{P}\left(\bigcup_{b=1}^{n} [x_b \geq k^\star]\right) \\ &\leq \sum_{b=1}^{n} \mathbb{P}(x_b \geq k^\star) \leq n \cdot \frac{1}{n^{d+1}} = \frac{1}{n^d}.\end{aligned}$$

Let us now analyze the number of empty boxes, in the case $n = m$ (because it is interesting: it shows an counterintuitive imbalance). We have $\mathbb{P}(x_b = 0) = (1 - 1/n)^n$ for each $b \in [1, n]$. Let y_b be the indicator variable $y_b = 1 \iff x_b = 0$. We want to know the number of empty boxes, so $y = \sum_{b=1}^{n} y_b$. Note that the y_b *not* are independent, but by the linearity of the expected value

$$\mathbb{E}(y) = \mathbb{E}\left(\sum_{b=1}^{n} y_b\right) = \sum_{b=1}^{n} \mathbb{E}(y_b) = \sum_{b=1}^{n} \left(1 - \frac{1}{n}\right)^n = n\left(1 - \frac{1}{n}\right)^n < ne^{-1} = \frac{n}{e}$$

and for $n \geq 2$ we have $n/4 \leq n\left(1 - \frac{1}{n}\right)^n$ which shows a rather unbalanced allocation. In the general case $m \geq n \geq 1$ we have $\mathbb{E}(y) = n(1 - 1/n)^m < ne^{-m/n}$.

4.4 Bingo

A bingo card has 15 numbers to be extracted. Let us suppose that those numbers are sampled in an uniform manner from the set $[1, 90]$ (actually it is not true, because in each column there must be a number and in each row there must be five numbers, so a card with, for instance, 1, 2, 3, 4, 5, is not valid...). Let us focus on one player with one card. Let X_ℓ be the number of extractions to get the ℓ-th match with a card number, with $\ell \in [1, 14]$ and $X_\ell \in [\ell, 89]$. Then, inspired by the geometric random variable, the probability to get the new match in a fixed $k \in [2, 90 - X_\ell]$ extractions is

$$\begin{aligned}
\mathbb{P}(X_{\ell+1} = X_\ell + k | X_\ell) &= \prod_{j=0}^{k-2} \left(1 - \frac{15-\ell}{90 - X_\ell - j}\right) \frac{15-\ell}{90 - X_\ell - k + 1} \\
&< \prod_{j=0}^{k-2} e^{-\frac{15-\ell}{90-X_\ell-j}} \frac{15-\ell}{90 - X_\ell - k + 1} \\
&= e^{-\sum_{j=0}^{k-2} \frac{15-\ell}{90-X_\ell-j}} \frac{15-\ell}{90 - X_\ell - k + 1} \\
&= e^{-(15-\ell)\sum_{t=90-X_\ell-k+2}^{90-X_\ell} \frac{1}{t}} \frac{15-\ell}{90 - X_\ell - k + 1} \\
&= e^{(15-\ell)(H(90-X_\ell-k+1) - H(90-X_\ell))} \frac{15-\ell}{90 - X_\ell - k + 1}.
\end{aligned}$$

Note that for $k = 2$ the formula becomes $\left(1 - \frac{15-\ell}{90-X_\ell}\right)\frac{15-\ell}{90-X_\ell-1}$ and for $k = 1$ we have $\mathbb{P}(\text{next extraction is a card number} | X_\ell) = \frac{15-\ell}{90-X_\ell}$.

The expectation is, by definition,

$$\begin{aligned}
\mathbb{E}(X_{\ell+1} | X_\ell) &= (X_\ell + 1)\frac{15-\ell}{90 - X_\ell} + \sum_{k=2}^{90-X_\ell} (X_\ell + k)\mathbb{P}(X_{\ell+1} = X_\ell + k | X_\ell) \\
&< (X_\ell + 1)\frac{15-\ell}{90 - X_\ell} + \\
&\quad \sum_{k=2}^{90-X_\ell} (X_\ell + k) e^{(15-\ell)(H(90-X_\ell-k+1) - H(90-X_\ell))} \frac{15-\ell}{90 - X_\ell - k + 1}.
\end{aligned}$$