Randomised Algorithms

Example Class 1

Thomas Sauerwald (tms41@cam.ac.uk)

Lent 2022

UNIVERSITY OF CAMBRIDGE

Plan

Schedule:

- Example Class 1 (today)
- Example Class 2 (10 February)
- Demo on Linear/Integer Programming applied to TSP (17 February)
- More Example Classes (3 more slots in February, 3 in March)
- Homework with Feedback?

Structure of Example Classes:

- Model Solution of some questions announced earlier
- Q & A
- (suggestions?)

1st Question

- We consider the coupon collecting problem with n coupons.
  (a) Prove that it takes $n \sum_{k=1}^{n} \frac{1}{k}$ days on expectation to collect all coupons.
  (b) Deduce that the probability it takes more than $n \log n + cn$ days is at most e^{-c}.

1st Question, Part a) (Solution)

Let T be the random variable describing the number of days until a copy from each of the n coupons has been seen. Further, let T_i be the first day after which exactly i different coupons has been seen. Formally:

- Let $Z_1, Z_2, \ldots \in [n]$ be the sequence of drawn coupons
- $T_i := \min\left\{t \geq 0 \colon |\cup_{s=1}^{t} Z_s| = i\right\}$, $(T_0 = 0,\ T_1 = 1$ and $T_n = T)$.

Then, using a telescoping sum and linearity of expectations,

$$\mathbf{E}[\,T\,] = \mathbf{E}[\,T_n - T_0\,] = \mathbf{E}\left[\sum_{k=1}^{n}(T_k - T_{k-1})\right] = \sum_{k=1}^{n}\mathbf{E}[\,T_k - T_{k-1}\,].$$

The random variable $T_k - T_{k-1}$ counts the waiting time between the day having $k-1$ coupons (for the first time) and the day having k coupons (for the first time). This random variable has a geometric distribution with parameter (i.e., success probability) $\frac{n-(k-1)}{n}$, and thus $\mathbf{E}[\,T_k - T_{k-1}\,] = \frac{n}{n-(k-1)}$. Thus,

$$\mathbf{E}[\,T\,] = \sum_{k=1}^{n}\frac{n}{n-(k-1)} = n \cdot \sum_{k=1}^{n}\frac{1}{n-(k-1)} = n \cdot \sum_{k=1}^{n}\frac{1}{k} \approx n \ln n.$$

1st Question, Part b) (Solution)

For the second part of the question, consider any coupon $i \in [n]$ and let $\tau := n \ln n + cn$. Then the waiting time $Z_i := \min\{t \geq 1 \colon Z_t = i\}$ until this coupon is obtained has a geometric distribution with parameter $1/n$.

Therefore,

$$\begin{aligned}
\mathbf{P}\left[\, Y_i > \tau \,\right] &= \left(1 - \frac{1}{n}\right)^{\tau} \\
&= \left(1 - \frac{1}{n}\right)^{n \ln n + cn} \\
&\leq \exp\left(- \ln n - c\right) = \frac{1}{n} \cdot e^{-c},
\end{aligned}$$

where the second inequality used $1 - x \leq e^{-x}$ which holds for any $x \in \mathbb{R}$.

Now by the Union Bound and definition of T and Z_i,

$$\begin{aligned}
\mathbf{P}\left[\, T > \tau \,\right] = \mathbf{P}\left[\bigcup_{i=1}^{n} \{Y_i > \tau\}\right] &\leq \sum_{i=1}^{n} \mathbf{P}\left[\, Z_i > \tau \,\right] \\
&= n \cdot \frac{1}{n} \cdot e^{-c} = e^{-c}.
\end{aligned}$$

1st Question (Additional Remark: Applying Chebyshev) 1/2

- We can also apply Chebyshev to the sum of geometric random variables used in Part a)
- Here we rely on the variance being additive for independent variables:

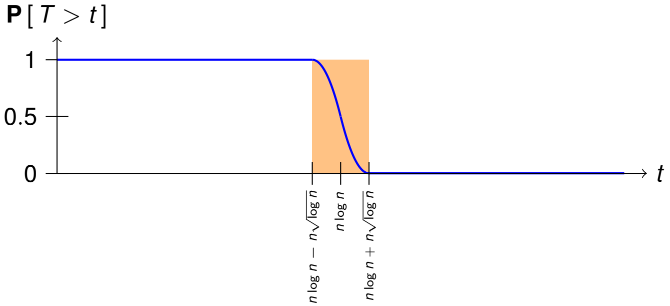
$$\begin{aligned}
\mathbf{V}[\,T\,] &= \mathbf{V}\left[\sum_{k=1}^{n} T_k - T_{k-1}\right] \\
&= \sum_{k=1}^{n} \mathbf{V}[\,T_k - T_{k-1}\,] \\
&= \sum_{k=1}^{n} \frac{1 - \frac{n-(k-1)}{n}}{(\frac{n-(k-1)}{n})^2} \\
&\leq n^2 \cdot \sum_{k=1}^{n} \frac{1}{n-(k-1)^2} \\
&\leq n^2 \cdot \sum_{k=1}^{\infty} \frac{1}{k^2} \\
&\leq n^2 \cdot \frac{\pi^2}{6}.
\end{aligned}$$

1st Question (Additional Remark: Applying Chebyshev) 2/2

- We derived $\mathbf{V}[\,T\,] \leq n^2 \cdot \frac{\pi^2}{6}$.
- We also computed $\mathbf{E}[\,T\,] = n \cdot \sum_{k=1}^{n} \frac{1}{k} \approx n \log n$.
- Applying Chebyshev with $\lambda = n\sqrt{\log n}$ yields:

$$\mathbf{P}\left[\,|T - \mathbf{E}[\,T\,]| \geq n\sqrt{\log n}\,\right] \leq \frac{\mathbf{V}[\,T\,]}{(n\sqrt{\log n})^2} \leq \frac{\pi^2}{6\log n} \stackrel{n\to\infty}{\to} 0.$$

- This implies concentration of T; the distribution of the upper tail drops sharply from 1 to 0:



2nd Question

Let $X_1, X_2, \ldots, X_n$ be n independent geometric random variables, each with parameter p (so $\mathbf{E}[\,X_i\,] = 1/p$ for each $i = 1, 2, \ldots, n$). Derive a Chernoff bound for $X := \sum_{i=1}^{n} X_i$.

2nd Question (Solution)

- First Approach: Use recipe for Chernoff Bounds by bounding $\mathbf{E}\left[\, e^{tX_i} \,\right]$ (a bit technical, since the random variable X_i has unbounded range)
- Second Approach: Relate sum of geometric random variables to a sum of Bernoulli random variables and apply one of the (nicer) Chernoff Bounds

- Let $X := X_1 + \cdots + X_n$ be the sum of n independent geometric random variables with $\mathbf{E}[\, X \,] = n/p$.
- We wish to upper bound, for any $\delta > 0$,
$$\mathbf{P}[\, X > (1+\delta)\mathbf{E}[\, X \,]\,].$$
- How can we express this event in terms of a sum of Bernoulli variables?
Hint: Imagine writing out all the outcomes of the n geometric variables as a single binary string ($1 = $ success, $0 = $ fail)
- $Y_1, Y_2, \ldots, Y_k$, with $k := (1+\delta)n/p$ are Bernoulli random variables (coin flips), and $Y := \sum_{i=1}^{k} Y_i$ has less than n successes:
$$\begin{aligned}
\mathbf{P}[\, X > (1+\delta)\mathbf{E}[\, X \,]\,] &= \mathbf{P}[\, Y < n \,] \\
&= \mathbf{P}[\, Y < kp - (kp - n) \,] \\
&= \mathbf{P}\left[\, Y < (1 - \frac{kp-n}{kp}) \cdot \mathbf{E}[\, Y \,] \,\right] \\
&\leq \exp\left(-\frac{1}{2} \left(\frac{kp-n}{kp} \right)^2 kp \right) \leq \exp\left(-\frac{1}{2} \frac{\delta^2 n}{(1+\delta)} \right).
\end{aligned}$$

2nd Question (Solution based on First Approach 1/2)

- First note that if X_i is geometric with parameter p, then

$$\mathbf{E}\left[e^{tX_i}\right] = \sum_{k=1}^{\infty} e^{tk} p(1-p)^{k-1} = pe^t \sum_{k=1}^{\infty} e^{t(k-1)}(1-p)^{k-1} = \frac{pe^t}{1-e^t(1-p)} = \frac{p}{e^{-t}-1+p},$$

assuming t is chosen so that $e^t(1-p) < 1$ (later, we will choose a t satisfying $t < p$ which implies this inequality)

- Using $e^{-t} \geq -t+1$,

$$\mathbf{E}\left[e^{tX_i}\right] \leq \frac{p}{p-t} = \left(1-\frac{t}{p}\right)^{-1}.$$

- Now returning to the recipe of deriving Chernoff bounds,

$$\begin{aligned}
\mathbf{P}\left[X \geq (1+\delta)\mu\right] \leq \mathbf{P}\left[e^{tX} \geq e^{t(1+\delta)\mu}\right] &= \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t(1+\delta)\mu}} \\
&= \frac{\left(1-\frac{t}{p}\right)^{-n}}{e^{t(1+\delta)n/p}} \\
&= \exp\left(-t(1+\delta)n/p + n\cdot(-\ln(1-\frac{t}{p}))\right),
\end{aligned}$$

and now choosing $t = (1-\frac{1}{1+\delta})p$ yields

$$\mathbf{P}\left[X \geq (1+\delta)\mu\right] \leq \exp\left(-n\cdot(\delta - \ln(1+\delta))\right).$$

This is slightly better than the previous bound, at least for large values of δ!

2nd Question (Solution based on First Approach 2/2)

- For the lower bound, one can derive similarly for $t > 0$ sufficiently small,

$$\mathbf{E}\left[e^{-tX}\right] \leq \left(1 + \frac{t}{p_i}\right)^{-1}.$$

- Then following the recipe of the Chernoff bound,

$$\begin{aligned}
\mathbf{P}[X \leq (1-\delta)\mu] \leq \mathbf{P}\left[e^{-tX} \geq e^{-t(1+\delta)\mu}\right] &= \frac{\mathbf{E}\left[e^{-tX}\right]}{e^{-t(1+\delta)\mu}} \\
&= \frac{\left(1+\frac{t}{p}\right)^{-n}}{e^{-t(1+\delta)n/p}} \\
&= \exp\left(t(1+\delta)n/p + n \cdot (-\ln(1+\frac{t}{p}))\right),
\end{aligned}$$

and now choosing $t = (\frac{1}{1-\delta} - 1)p$ yields

$$\mathbf{P}[X \leq (1-\delta)\mu] \leq \exp\left(-n \cdot (\delta - \ln(1-\delta))\right).$$

3rd Question

Using the concentration result for QuickSort in class, prove that it implies a bound of $O(n \log n)$ for the expected number of comparisons.

Recall: We proved for the number of comparisons $H := \sum_{i=1}^{n} H_i$,

$$\mathbf{P}[\, H \leq 24n \log n \,] \geq 1 - n^{-1}.$$

3rd Question (Solution)

- Let H be the number of comparisons performed by Quicksort.
- In the lectures, we proved that $\mathbf{P}[\,H > 24n\log n\,] \leq n^{-1}$
- From Part IA Algorithms, we know the fact that $H \leq n^2$.
- Let us now bound $\mathbf{E}[\,H\,]$:

$$
\begin{aligned}
\mathbf{E}[\,H\,] &= \sum_{x=1}^{n^2} \mathbf{P}[\,H = x\,]\cdot x \\
&\leq \sum_{x=1}^{24n\log n} \mathbf{P}[\,H = x\,]\cdot x + \sum_{x=24n\log n+1}^{n^2} \mathbf{P}[\,H = x\,]\cdot x \\
&\leq (24n\log n)\cdot \sum_{x=1}^{24n\log n} \mathbf{P}[\,H = x\,] + n^2 \sum_{x=24n\log n+1}^{n^2} \mathbf{P}[\,H = x\,] \\
&= (24n\log n)\cdot \mathbf{P}[\,X \leq 24n\log n\,] + n^2\cdot \mathbf{P}[\,H > 24n\log n\,] \\
&\leq (24n\log n)\cdot 1 + n^2\cdot n^{-1} \\
&\leq 24n\log n + n \leq 25n\log n.
\end{aligned}
$$

4th Question

Design a randomised algorithm for the following problem. The input consists of an $n \times n$ matrix A with entries in $\{0, 1\}$ and a vector x of length n with entries in the real interval $[0, 1]$. The goal is to return a vector y of length n with entries in $\{0, 1\}$ such that

$$\max_{i=1,\ldots,n} |(Ax)_i - (Ay)_i| \leq 2\sqrt{n \log n}$$

with probability at least $1 - 2 \cdot n^{-2}$.
Hint: Your algorithm should have the property that for any $1 \leq i, j \leq n$, $\mathbf{E}[\, A_{i,j} \cdot y_j \,] = A_{i,j} x_j$.

4th Question (Example)

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 0.5 \\ 0.25 \end{pmatrix}$$

$$A \cdot x = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.8 \\ 0.5 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1.3 \\ 0.25 \end{pmatrix}$$

Now take an integral vector:

$$y = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad \Rightarrow \quad A \cdot y = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$$

The largest gap between any coordinate in $A \cdot x$ and $A \cdot y$ is $|1.3 - 2| = 0.7$.

4th Question (Solution)

- For any $1 \leq j \leq n$, let Y_j be a Bernoulli distribution with parameter $x_j \in [0,1]$. Note $\mathbf{E}[\,Y_i\,] = x_i$, and thus $\mathbf{E}\left[\,A_{ij}Y_j\,\right] = A_{i,j}x_j$. Further, for any row i define

$$Z = Z(i) := (AY)_i - (AX)_i = \sum_{j=1}^{n} A_{ij}(Y_j - x_j).$$

- We will check that $|Z| > 2\sqrt{n\log n}$ with sufficiently small probability. First

$$\mathbf{P}\left[\,Z > 2\sqrt{n\log n}\,\right] = \mathbf{P}\left[\sum_{j=1}^{n} A_{ij}Y_j \geq \sum_{j=1}^{n} A_{ij}x_j + 2\sqrt{n\log n}\right]$$

  and note that $\sum_{j=1}^{n} A_{ij}Y_j$ is the sum of $m = \sum_{j=1}^{n} A_{ij}$ independent Bernoulli's.

- Using the nice version of Chernoff Bounds (additive form), we have

$$\mathbf{P}\left[\sum_{j=1}^{n} A_{ij}Y_j \geq \sum_{j=1}^{n} A_{ij}x_j + 2\sqrt{n\log n}\right] \leq \exp\left(-8\frac{n\log n}{m}\right) \leq \exp(-8\log n) = \frac{1}{n^8}.$$

  That is $\mathbf{P}\left[\,Z > \sqrt{n\log n}\,\right] \leq n^{-8}$.

- Applying the same argument we get $\mathbf{P}\left[\,Z < -\sqrt{n\log n}\,\right] \leq n^{-8}$ and thus $\mathbf{P}\left[\,|Z| > \sqrt{n\log n}\,\right] < 2n^{-8}$ by the Union Bound.
- Finally, applying Union Bound over all $i = 1,\ldots,n$ yields

$$\mathbf{P}\left[\max_{i=1,\ldots,n} |(AY)_i - (AX)_i| > \sqrt{n\log n}\right] \leq n \cdot 2n^{-8} < n^{-2}.$$