CS5330: Randomized Algorithms Spring 2020

Review Session Proceedings

Prof. Arnab Bhattacharyya

The format of the review session was open, with people raising their hands and asking questions. In the following, I will try to summarize the main items of discussion.

Total distance moved by random permutation

Exercise 2.1 of MU: Let $a_1, \ldots, a_n$ be a random permutation of $\{1, \ldots, n\}$. Find $\mathbb{E}\left[\sum_{i=1}^n |a_i - i|\right]$.

Define the indicator variable X_{ij} for the event that $a_i = j$. Note that $|a_i - i| = \sum_{j=1}^n X_{ij}|j - i|$. We know that $\mathbb{E}[X_{ij}] = 1/n$ for any i and j. Hence, by linearity of expectation:

$$\mathbb{E}\left[\sum_{i=1}^n |a_i - i|\right] = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[X_{ij} \cdot |j - i|] = \frac{1}{n} \sum_{i,j \in [n]} |j - i| = \Theta(n^2).$$

The last equality is a routine calculation that we didn't do fully on board. You can see it in the following way. Every i is either $\leq n/2$ or $> n/2$. First, assume it's $\leq n/2$. In that case, $\sum_{j=i+1}^n (j - i) = \Theta(n^2)$. Similarly, if $i > n/2$, $\sum_{j=1}^{i-1}(i - j) = \Theta(n^2)$.

2018 Midterm

In Problem 2b, the variance should be $\frac{1}{5}(1^2 + 2^2 + 3^2 + 4^2 + 5^2) - 3^2 = 2$. Problem 6 is good practice in computing probabilities, expectations and variances. In Problem 6d, there are some incorrect constant factors in the solution.

Number of inversions in a random permutation

What is the expected number of inversions in a random permutation (i.e., $i < j$ but $\pi(i) > \pi(j)$)? For each $i < j$, define X_{ij} to indicate that i, j is an inversion. Note that $\mathbb{E}[X_{ij}] = 1/2$ because $\pi(i) < \pi(j)$ and $\pi(i) > \pi(j)$ are equally likely. So, $\mathbb{E}[\sum_{i<j} X_{ij}] = \frac{1}{2}\binom{n}{2}$.

What is the variance of the number of inversions? We use the formula $\mathsf{Var}\left[\sum_{i<j} X_{ij}\right] = \sum_{i<j} \mathsf{Var}[X_{ij}] + \sum_{i<j,k<\ell,\{i,j\}\neq\{k,\ell\}} \mathsf{Cov}(X_{ij}, X_{k\ell})$. The variance of each X_{ij} is $1/4$. Note that X_{ij} and $X_{k\ell}$ are independent if i, j, k, ℓ are all distinct. Thus, we can bound:

$$\mathsf{Var}\left[\sum_{i<j} X_{ij}\right] \leq \frac{1}{4}\binom{n}{2} + \sum_{i<j,i<\ell,j\neq\ell} \mathbb{E}[X_{ij}X_{i\ell}] + \sum_{i<j,k<i} \mathbb{E}[X_{ij}X_{ki}] + \sum_{i<j,j<\ell} \mathbb{E}[X_{ij}X_{j\ell}] + \sum_{i<j,k<j} \mathbb{E}[X_{ij}X_{kj}]$$

Each of the above expectations can be explicitly computed (e.g., $\mathbb{E}[X_{ij}X_{i\ell}] = 1/3$ and $\mathbb{E}[X_{ij}X_{ki}] = 1/6$), but in any case, they are all constants. Each sum goes over $O(n^3)$ indices. So, the variance is bounded by $O(n^3)$.

Randomized Rounding for Hitting Set LP

This is Problem 7 of the second optional problem set. In the randomized rounding step, you take a solution t^* of the LP and obtain a set T by putting each $j \in [m]$ into T with probability t^*_j independently.

Fix a set S_i. The probability that T does not hit S_i is $\prod_{j \in S_i}(1-t^*_j) \leq \prod_{j \in S_i} e^{-t^*_j} = e^{-\sum_{j \in S_i} t^*_j} \leq e^{-1}$ where the last inequality holds because of the LP constraint. Thus, in expectation, $\leq n/e$ sets are not hit.

In order to hit all the sets, we can repeat the above process $r = O(\log n)$ times independently to get sets $T_1, \ldots, T_r$. The probability that some S_i is hit by none of these sets is $\leq e^{-r} < \frac{1}{n^2}$. Thus, with high probability, each of the sets $S_1, \ldots, S_n$ is hit by $\hat{T} = T_1 \cup \cdots \cup T_r$. The expected size of $\hat{T}$ is $O(\log n)$ times the LP optimal value (which is at most the optimal hitting set size). The desired conclusion follows by Markov's inequality.

Longest Common Subsequence

This is problem 3 of the second optional problem set. The probability that the LCS is of length $\geq k$ is, by the union bound, at most $\leq \binom{n}{k}^2 \frac{1}{n^k} \leq \left(\frac{e^2 n}{k^2}\right)^k$ where we used the inequality $\binom{n}{k} \leq (en/k)^k$.

Now, we use the fact that:

$$\begin{aligned}
\mathbb{E}[|\mathsf{LCS}(x,y)|] &= \sum_{k \geq 0} \Pr[|\mathsf{LCS}(x,y)| \geq k] \\
&= \sum_{0 \leq k \leq 3\sqrt{n}} \Pr[|\mathsf{LCS}(x,y)| \geq k] + \sum_{k > 3\sqrt{n}} \Pr[|\mathsf{LCS}(x,y)| \geq k] \\
&\leq 3\sqrt{n} + \sum_{k > 3\sqrt{n}} \left(\frac{e^2 n}{k^2}\right)^k \\
&\leq 3\sqrt{n} + \sum_{k > 3\sqrt{n}} \left(\frac{e^2 n}{9n}\right)^k \\
&\leq 3\sqrt{n} + 10
\end{aligned}$$

where the last line follows by summing the geometric sequence with ratio $e^2/9 < 1$.

Other comments

With high probability typically means that the probability is $1 - 1/n^c$ where c is a constant. Typically, c can be made arbitrarily large by tuning the constants in the algorithm. I am not a big fan of this notation, so I will usually be more precise in specifying the probability bound.

For Question 6 in the second optional problem set: you first need to take the average of $O(r^2/\epsilon^2)$ copies of X to reduce its variance, so that the average will be in the interval $[(1-\epsilon)\mu, (1+\epsilon)\mu]$ with probability $2/3$. Finally, you can take the median of $O(\log 1/\delta)$ copies of these averages to boost the success probability to $1 - \delta$.