The Stable Marriage Problem, and The Coupon Collector's Problem

497 - Randomized Algorithms

Sariel Har-Peled

September 17, 2002

1 The Stable Marriage Problem

You have n males and n females. Every male has a ranked list of females he likes, and vice versa. Thus, an instance of his problem might be

$$\begin{array}{llll} A : abcd & B : bacd & C : adcb & D : dcab \\ a : ABCD & b : DCBA & c : ABCD & d : CDAB, \end{array}$$

where capital letters represent females, and small letters represent males. Consider a marriage $M = \{A-a, B-b, C-c, D-d\}$. Note that $C-d$ is a dissatisfied couple, as C prefer to ditch c and marry d, and d would prefer to ditch D and marry C. Thus, this marriage setting is unstable, as $C-d$ would just elope together.

Question 1.1 *Given an full preferences lists, is there a stable marriage?*

The *Proposal Algorithm* works by a man proposing to a woman according to his preference list. A woman either decline, because she is already married to somebody she prefers, or she dispose of her current mate and accept the proposal.

It is easy to argue that the PA terminates in a stable marriage. Indeed, once a woman get married she always stay married. And every time a ditching is performed the ranking of the current mate improves. Thus, a woman can ditch at most n times. As for stability, one can easily argue that it is achieved.

Clearly, this algorithm performers $O(n^2)$ proposals, in the worst case.

Let us change the problem a little: The women fix their preferences lists in advance, and the men (of course) set the preference list in a random uniform fashion. What is the expected number of proposals this algorithm makes? Let T_P denote this number.

Even this, is to hard to analyze. Let us instead define an amnesic algorithm where each male randomly choose which woman to propose to randomly (proposing maybe to the same woman several times). Let T_A denote the number of proposals made by this algorithm.

Remark 1.2 This is an example of *Principle of Deferred Decisions* – we analyze the algorithm like makes its random decision when it really has to, instead of thinking about it as being fixed in advance (i.e., random preference lists precomputed in advanced vs. computing the next woman to propose to randomly).

Definition 1.3 A random variable X *stochastically dominates* a random variable Y, if for any $z \in \mathbb{R}$, we have $\Pr\left[X > z\right] \geq \Pr\left[Y > z\right]$.

Clearly, T_A stochastically dominates T_P.

How do we analyze the behavior of T_A? Well, looking from the outside, it is clear that the algorithm stops once all the women got proposals (as by then, all the women are married, and the situation is stable).

This is a variant of the Coupon Collector problem, described below, and the following bound holds (this follows from Theorem 2.2):

Theorem 1.4 *For any constant $c \in \mathbb{R}$, and $m = n\log n + cn$, we have* $\lim_{n\to\infty} \Pr\left[T_A > m\right] = 1 - \exp\left(-e^{-c}\right)$.

Example 1.5 For $c = 100$, we have

$$\lim_{n\to\infty} \Pr\left[T_A > n\log n + 100n\right] = 1 - \exp\left(-e^{-100}\right) = 1 - \exp\left(-1/e^{100}\right) \leq 1 - \left(1 - 1/e^{100}\right) = \frac{1}{e^{100}},$$

since $e^{-x} = 1 - x + x^2/2! - x^3/3!\ldots$ (Taylor expansion), as such, for $x < 1$, we have $e^{-x} \geq 1 - x$.

This bound is quite small! The distribution of T_A is strongly concentrated around $n\log n$.

2 The Coupon Collector's Problem

There are n types of coupons, and at each trial one coupon is picked in random. How many trials one has to perform before picking all coupons? Let m be the number of trials preformed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

Let $C_i \in \{1, \ldots, n\}$ be the coupon picked in the i-th trial. The j-th trial is a success, if C_j was not picked before in the first $j-1$ trials. Let X_i denote the number of trials from the i-th success, till after the $(i+1)$-th success. Clearly, the number of trials performed is

$$X = \sum_{i=0}^{n-1} X_i.$$

Clearly, the probability of X_i to succeed in a trial is $p_i = \frac{n-i}{n}$, and X_i has geometric distribution with probability p_i. As such $E[X_i] = 1/p_i$, and $\mathrm{var}[X_i] = q/p^2 = (1-p_i)/p_i^2$.

Thus,

$$E\left[X\right] = \sum_{i=0}^{n-1} E\left[X_i\right] = \sum_{i=0}^{n-1} \frac{n}{n-i} = nH_n = n(\ln n + \Theta(1)) = n\ln n + O(n),$$

where $H_n = \sum_{i=1}^{n} \frac{1}{i}$ is the n-th Harmonic number.

As for variance, using the independence of $X_0, \ldots, X_{n-1}$, we have

$$\begin{aligned}
\mathrm{var}\left[X\right] &= \sum_{i=0}^{n-1} \mathrm{var}\left[X_i\right] = \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n-1} \frac{1-(n-i)/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i}{n}\left(\frac{n}{n-i}\right)^2 \\
&= n\sum_{i=0}^{n-1} \frac{i}{(n-i)^2} = n\sum_{i=1}^{n} \frac{n-i}{i^2} = n\left(\sum_{i=1}^{n} \frac{n}{i^2} - \sum_{i=1}^{n} \frac{1}{i}\right) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n.
\end{aligned}$$

Since, $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} = \pi/6$, we have $\lim_{n\to\infty} \frac{\mathrm{var}\left[X\right]}{n^2} = \frac{\pi}{6}$.

This implies a weak bound on the concentration of X, using Chebyshev inequality, but this is going to be quite weaker than what we implied we can do.

A stronger bound, follows from the following observation. Let Z_i^r denote the event that the i-th coupon was not picked in the first r trials. Clearly,

$$\Pr\left[Z_i^r\right] = \left(1 - \frac{1}{n}\right)^r \le e^{-r/n}.$$

Thus, for $r = \beta n \log n$, we have $\Pr\left[Z_i^r\right] \le e^{-(\beta n \log n)/n = n^{-\beta}}$. Thus,

$$\Pr\left[X > \beta n \log n\right] \le \Pr\left[\bigcup_i Z_i^{\beta n \log n}\right] \le n \cdot \Pr\left[Z_1\right] \le n^{-\beta+1}.$$

This is quite strong, but still not as strong as we can do.

We need the following:

For any $y \ge 1$, and $|x| \le 1$, we have

$$\left(1 - x^2 y\right) e^{xy} \le (1+x)^y \le e^{xy}$$

Lemma 2.1 *Let $c > 0$ be a constant, $m = n \ln n + cn$ for a positive integer n. Then for any constant k, we have*

$$\lim_{n\to\infty} \binom{n}{k} \left(1 - \frac{k}{n}\right)^m = \frac{\exp(-ck)}{k!}.$$

Proof: By the above formula, we have

$$\left(1 - \frac{k^2 m}{n^2}\right) \exp\left(-\frac{km}{n}\right) \le \left(1 - \frac{k}{n}\right)^m \le \exp\left(-\frac{km}{n}\right).$$

Observe also that $\lim_{n\to\infty} \left(1 - \frac{k^2 m}{n}\right) = 1$, and $\exp\left(-km/n\right) = n^{-k} \exp\left(-ck\right)$. Also,

$$\lim_{n\to\infty} \binom{n}{k} \frac{k!}{n^k} = \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} = 1.$$

Thus,

$$\lim_{n\to\infty} \binom{n}{k} \left(1 - \frac{k}{n}\right)^m = \lim_{n\to\infty} \frac{n^k}{k!} \exp\left(-\frac{km}{n}\right) = \lim_{n\to\infty} \frac{n^k}{k!} n^{-k} \exp\left(-ck\right) = \frac{\exp(-ck)}{k!}. \qquad \blacksquare$$

Theorem 2.2 *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then, for any constant $c \in \mathbb{R}$, and $m = n \ln n + cn$, we have*

$$\lim_{n\to\infty} \Pr\left[X > m\right] = 1 - \exp\left(-e^{-c}\right).$$

Proof: We have $\Pr\left[X > m\right] = \Pr\left[\cup_i Z_i^m\right]$. By inclusion-exclusion, we have

$$\Pr\left[\bigcup_i Z_i^m\right] = \sum_{i=1}^{n} (-1)^{i+1} P_i^n,$$

where

$$P_j^n = \sum_{1\le i_1 < i_2 < \ldots < i_j \le n} \Pr\left[\bigcap_{v=1}^{j} Z_{i_v}^m\right].$$

Let $S_k^n = \sum_{i=1}^{k} (-1)^{i+1} P_i^n$. We know that $S_{2k}^n \le \Pr\left[\bigcup_i Z_i^m\right] \le S_{2k+1}^n$.

By symmetry,

$$P_k^n = \binom{n}{k} \Pr\left[\bigcap_{v=1}^{k} Z_v^m\right] = \binom{n}{k}\left(1 - \frac{k}{n}\right)^m,$$

Thus, $P_k = \lim_{n\to\infty} P_k^n = \exp\left(-ck\right)/k!$, by Lemma 2.1.

Let

$$S_k = \sum_{j=1}^{k} (-1)^{j+1} P_j = \sum_{j=1}^{k} (-1)^{j+1} \frac{\exp(-cj)}{j!}$$

Clearly, $\lim_{k\to} S_k = 1 - \exp(-e^{-c})$ by the Taylor expansion of $\exp(x)$ for $x = -e^{-c}$. Indeed,

$$\exp(x) = \sum_{j=0}^{\infty} \frac{x^j}{j!} = \sum_{j=0}^{\infty} \frac{(-e^{-c})^j}{j!} = 1 + \sum_{j=0}^{\infty} \frac{(-1)^j e^{-cj}}{j!}$$

Clearly, $\lim_{n\to\infty} S_k^n = S_k$ and $\lim_{k\to\infty} S_k = 1 - \exp\left(-e^{-c}\right)$. Thus, (using fluffy math), we have

$$\lim_{n\to\infty} \Pr\left[X > m\right] = \lim_{n\to\infty} \Pr\left[\cup_{i=1}^{n} Z_i^m\right] = \lim_{n\to\infty} \lim_{k\to\infty} S_k^n = \lim_{k\to\infty} S_k = 1 - \exp\left(-e^{-c}\right)$$

■