18.317 Combinatorics, Probability, and Computations on Groups 1 October 2001

Lecture 10

Lecturer: Igor Pak *Scribe: Igor Pavlovsky*

Coupon Collector's Problem

Recall that the problem concerns a prudent shopper who tries, in several attempts, to collect a complete set of N different coupons. Each attempt provides the collector with a coupon randomly chosen from N known kinds, and there is an unlimited supply of coupons of each kind. It is easy to estimate the expected waiting time to collect all N coupons:

$$\mathrm{E}[\text{time to collect } N \text{ coupons}] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + N = N(\log N + \gamma + O(\frac{1}{N}))$$

Similarly,

$$\begin{aligned}\mathrm{E}[\text{time to collect } \tfrac{N}{2} \text{ coupons}] &= 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{\frac{N}{2}+1} \\ &= N(\log N + \gamma + O(\frac{1}{N}) - \log\frac{N}{2} - \gamma - O(\frac{1}{N})) \\ &= N\log 2 + O(1)\end{aligned}$$

Let's explore the problem further. Suppose our collector attempts to obtain coupons $T = \alpha N$ times. We'd like to know how many different kinds of coupons he can be expected to have, and also how many of these are not repeated. For such estimates, we need the following tool:

Theorem 1 (Chernoff bound) *Let X_i, $i = 1 \ldots n$, be independent Poisson trials, with outcomes 1 and 0 with probabilities p_i and $1 - p_i$ respectively. Set $X = \sum_{i=1}^n X_i$, $\mu = \mathrm{E}[X] = \sum p_i$. Then for every $\delta > 0$, the following bounds hold:*

$$\begin{aligned}\Pr[X > (1+\delta)\mu] &< \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \\ \Pr[X < (1-\delta)\mu] &< e^{-\mu\delta^2/2}\end{aligned}$$

■

In our situation, at every step

$$\begin{aligned}\Pr[\text{got a coupon we already have}] &< \alpha \\ \Pr[\text{got a new coupon}] &> 1 - \alpha\end{aligned}$$

Consider the worst-case *independent*-trials processes defined by the above inequalities transformed into equalities. Bounds on these processes will give us the desired bounds for the coupon collector's problem. Clearly,

$$\begin{aligned}\mathrm{E}[\text{number of repeated coupons}] &< \alpha T \\ \mathrm{E}[\text{number of different coupons}] &> (1-\alpha)T\end{aligned}$$

Therefore, by Chernoff bounds,

$$\Pr[\#\text{ rep coupons} > (1+\delta)\alpha T] < \epsilon_1^T$$
$$\Pr[\#\text{ diff coupons} < (1-\delta)(1-\alpha)T] < \epsilon_2^T$$

where ϵ_1, ϵ_2 depend on α and δ. Now, the event [# non-repeated coupons $< A - B$] is contained in the union of the events [# diff coupons$< A$] $\cup$ [# rep coupons $> B$]. Hence,

$$\Pr[\#\text{ non-rep coupons} < (1-\delta)(1-\alpha)T - (1+\delta)\alpha T] < \epsilon_1^T + \epsilon_2^T < \epsilon^T$$

for large T and some ϵ dependent on α and δ. Recalling that $T = \alpha N$, conclude

Theorem 2 *After αN steps, the number of non-repeated coupons accumulated by the collector is greater than cN with probability $> 1 - \epsilon^N$ for large N. Here $\alpha, c, \epsilon > 0$, one of α or c may be chosen arbitrarily, and the other two parameters are produced by the theorem.* ∎

Mixing Time of Random Walks

We have finally arrived at our goal

Theorem 3 *Let X_t be a lazy random walk on G starting at 1, with a random generating set of size k. Denote by $\operatorname{sep}(T)$ the separation distance after T steps. Then exists a constant $c > 0$ such that $\operatorname{sep}(c \log |G|) < \epsilon$ with probability $> 1 - \delta - \gamma$, provided that $k = 2\log|G| + 2\log\frac{1}{\epsilon} + 2\log\frac{1}{\delta} + \log\frac{1}{\gamma}$.*

Proof: Combine the Erdős–Rényi machine and coupon collector's problem to ensure enough non-repeated generators. Check parameters, done. ∎

Our next aim is to prove that random Cayley graphs are expanders. We begin with a few observations on matrices and random walks. Let A be the transition matrix for the random walk $X_{t+1} = X_t \cdot g_i$ on G with generators $g_1, \ldots, g_k$. If at each step the generator g_i is chosen uniformly, then the entries of A are numbers 0 and $\frac{1}{k}$. Let $v = e_1$ be the initial state vector (begin at 1). The vector $A^t v$ gives the probability distribution on G after t steps, and can be written as

$$A^t v = u + \lambda_1^t w_1 + \lambda_2^t w_2 + \ldots$$

where $\lambda_1 \geq \lambda_2 \geq \ldots$ are the eigenvalues of A. If these eigenvalues are real, they clearly must lie between -1 and 1 for the distribution to converge to the uniform probability distribution U.