Randomized Algorithms
Lecture 5

Kousha Etessami

Continuing review of Discrete Probability . . . and the Coupon Collector Problem

Recall: "Coupon collecting": we are buying boxes (of cereal), each of which has a uniformly random coupon inside. There are n different types of coupons, and the goal is to collect one of each type, and then stop buying.

Last time we showed the expected number $\mathrm{E}[X]$ of boxes we must buy is $nH(n) = n\ln(n) + \theta(n)$, or more precisely,
$n\ln(n) + \frac{n}{2} \leq nH(n) \leq n\ln(n) + n$.

Today we examine what the probability is that a "run" of the purchasing process is far from that expectation.

Concentration inequalities will be vital:

- Markov Inequality;
- Chebyshev Inequality;
- Chernoff Bound / Hoeffding inequality.

Markov Inequality

Very simple and easy, but very important.

Theorem (3.1, Markov Inequality)

Let X be any random variable that takes only non-negative values. Then for any $a > 0$,

$$\Pr[X \geq a] \;\leq\; \frac{E[X]}{a}.$$

Markov Inequality

Very simple and easy, but very important.

Theorem (3.1, Markov Inequality)

Let X be any random variable that takes only non-negative values. Then for any $a > 0$,

$$\Pr[X \geq a] \leq \frac{E[X]}{a}.$$

Proof.

Define the indicator function $I = I(X)$ by

$$I(x) = \begin{cases} 0 & x < a; \\ 1 & x \geq a. \end{cases};$$

Markov Inequality

Very simple and easy, but very important.

Theorem (3.1, Markov Inequality)

Let X be any random variable that takes only non-negative values. Then for any $a > 0$,

$$\Pr[X \geq a] \;\leq\; \frac{E[X]}{a}.$$

Proof.

Define the indicator function $I = I(X)$ by

$$I(x) \;=\; \begin{cases} 0 & x < a; \\ 1 & x \geq a. \end{cases};$$

Then $X \geq a \cdot I(X)$, and hence $I(X) \leq \frac{X}{a}$.

Markov Inequality

Very simple and easy, but very important.

Theorem (3.1, Markov Inequality)

Let X be any random variable that takes only non-negative values. Then for any $a > 0$,

$$\Pr[X \geq a] \;\leq\; \frac{E[X]}{a}.$$

Proof.

Define the indicator function $I = I(X)$ by

$$I(x) \;=\; \begin{cases} 0 & x < a; \\ 1 & x \geq a. \end{cases};$$

Then $X \geq a \cdot I(X)$, and hence $I(X) \leq \frac{X}{a}$.
Taking expectation of both sides, and using $\mathrm{E}[I] = \Pr[X \geq a]$, we have

$$\mathrm{E}[I] \;=\; \Pr[X \geq a] \leq \frac{\mathrm{E}[X]}{a} \quad .$$

□

Bounding Coupon Collector purchases - Markov

Let X be the number of purchases we have to make in the coupon collector problem until we get all n coupons. Recall, we know $\mathrm{E}[X] = nH_n$, and thus: $n\ln(n) + \frac{n}{2} \leq \mathrm{E}[X] \leq n\ln n + n$.

Suppose we want a lower bound t on the number of boxes we have to buy, such that $\Pr[X \geq t] \leq \frac{1}{2}$.

By Markov's inequality, $\Pr[X \geq t] \leq \frac{\mathrm{E}[X]}{t} \leq \frac{nH_n}{t}$.
Thus, it suffices to let $t = 2nH_n$, to get $\Pr[X \geq 2nH_n] \leq \frac{1}{2}$.
So, $\Pr[X \geq 2(n\ln(n) + n)] \leq \frac{1}{2}$.

However, this bound is way too weak: we can get far smaller probability of failure with $2nH_n$ purchases.

The power of Markov ineq. is that it does not require any other knowledge of the random variable. However for specific problems, we can often do much better.

For example, we can bound the variance.

Variance and Covariance

Recall, the *variance* of a random variable is

$$\mathrm{Var}[X] := \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

Definition (3.3)

The *covariance* of two random variables X and Y is defined as

$$\mathrm{Cov}[X, Y] \doteq \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] = \mathrm{E}[XY] - E[X]E[Y]\,.$$

Theorem (3.2)

For any two random variables X, Y, we have

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y].$$

Variance and Covariance

Recall, the *variance* of a random variable is

$$\mathrm{Var}[X] \;:=\; \mathrm{E}[(X-\mathrm{E}[X])^2] \;=\; \mathrm{E}[X^2]-\mathrm{E}[X]^2.$$

Definition (3.3)

The *covariance* of two random variables X and Y is defined as

$$\mathrm{Cov}[X,Y] \;\doteq\; \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right] \;=\; \mathrm{E}[XY]-E[X]E[Y]\;.$$

Theorem (3.2)

For any two random variables X, Y, we have

$$Var[X+Y] \;=\; Var[X]+Var[Y]+2Cov[X,Y].$$

Proof.

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathrm{E}[(X+Y)^2]-\mathrm{E}[X+Y]^2\\ &= \mathrm{E}[X^2]+\mathrm{E}[Y^2]+2\mathrm{E}[XY]-\mathrm{E}[X]^2-\mathrm{E}[Y]^2-2\mathrm{E}[X]\mathrm{E}[Y]\\ &= \mathrm{Var}[X]+\mathrm{Var}[Y]+2(\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]).\end{aligned}$$

□

(pairwise) Independent Random Variables

Theorem (3.3)

If X, Y are a pair of independent random variables, then

$$E[XY] = E[X] \cdot E[Y].$$

(pairwise) Independent Random Variables

Theorem (3.3)

If X, Y are a pair of independent random variables, then

$$E[XY] = E[X] \cdot E[Y].$$

Corollary (3.4)

If X, Y are a pair of independent random variables, then

$$Cov[X, Y] = 0$$

and

$$Var[X + Y] = Var[X] + Var[Y].$$

Chebyshev Inequality

Theorem (3.2, Chebyshev Inequality)

For every $a > 0$,

$$\Pr[|X - E[X]| \geq a] \;\leq\; \frac{Var[X]}{a^2}.$$

Chebyshev Inequality

Theorem (3.2, Chebyshev Inequality)

For every $a > 0$,

$$\Pr[|X - E[X]| \geq a] \ \leq \ \frac{Var[X]}{a^2}.$$

Proof.

First note that for any $a > 0$,

$$|X - \mathrm{E}[X]| \geq a \iff (X - \mathrm{E}[X])^2 \geq a^2$$

Chebyshev Inequality

Theorem (3.2, Chebyshev Inequality)

For every $a > 0$,

$$\Pr[|X - E[X]| \geq a] \leq \frac{Var[X]}{a^2}.$$

Proof.

First note that for any $a > 0$,

$$|X - \mathrm{E}[X]| \geq a \iff (X - \mathrm{E}[X])^2 \geq a^2$$

Apply Markov's Inequality to the random variable $(X - \mathrm{E}[X])^2$. We get:

$$\Pr[|X - \mathrm{E}[X]| \geq a] = \Pr[(X - \mathrm{E}[X])^2 \geq a^2] \leq \frac{\mathrm{E}[(X - \mathrm{E}[X])^2]}{a^2} = \frac{\mathrm{Var}[X]}{a^2}.$$

□

Bounding Coupon Collector purchases - Markov

Recall that X is the number of purchases of the coupon collector problem and $\mathrm{E}[X] = nH_n \leq n \ln n + n$.

Using Markov's inequality, we can get that $\Pr[X \geq n^2 H_n] \leq \frac{1}{n}$.

We can do better with Chebyshev's inequality ...

Bounding Coupon Collector purchases - Chebyshev

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2}.$$

Bounding Coupon Collector purchases - Chebyshev

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2}.$$

- Need to evaluate $\mathrm{Var}[X]$, which is $\mathrm{Var}[X_1 + \ldots + X_n]$.
  Recall that X_i is the *number of boxes* bought to get the i-th new card.
- Corollary 3.4: for independent Y, Z, $\mathrm{Var}[Y + Z] = \mathrm{Var}[Y] + \mathrm{Var}[Z]$.
- Are these X_i's independent? Yes.

Bounding Coupon Collector purchases - Chebyshev

$$\Pr[|X - \mathrm{E}[X]| \geq a] \leq \frac{\mathrm{Var}[X]}{a^2}.$$

- Need to evaluate $\mathrm{Var}[X]$, which is $\mathrm{Var}[X_1 + \ldots + X_n]$.

  Recall that X_i is the *number of boxes* bought to get the i-th new card.
- Corollary 3.4: for independent Y, Z, $\mathrm{Var}[Y + Z] = \mathrm{Var}[Y] + \mathrm{Var}[Z]$.
- Are these X_i's independent? Yes. X_i is a geometrically distributed r.v. that only depends on the values n and i (and not on what cards we have collected or how long it took to collect them).

Bounding Coupon Collector purchases - Chebyshev

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2}.$$

- Need to evaluate $\mathrm{Var}[X]$, which is $\mathrm{Var}[X_1 + \ldots + X_n]$.

  Recall that X_i is the *number of boxes* bought to get the i-th new card.
- Corollary 3.4: for independent Y, Z, $\mathrm{Var}[Y + Z] = \mathrm{Var}[Y] + \mathrm{Var}[Z]$.
- Are these X_i's independent? Yes. X_i is a geometrically distributed r.v. that only depends on the values n and i (and not on what cards we have collected or how long it took to collect them).
- Hence the random variables $X_1, \ldots, X_n$ are all mutually independent, and

$$\mathrm{Var}[X] \;=\; \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_n].$$

Bounding Coupon Collector purchases - Chebyshev

Each X_i is a geometric random variable with parameter $p = \frac{n-(i-1)}{n}$.

Bounding Coupon Collector purchases - Chebyshev

Each X_i is a geometric random variable with parameter $p = \frac{n-(i-1)}{n}$.

Fact (3.8)

For any geometric random variable X with parameter p,
$E[X] = p^{-1}$, *and* $Var[X] = \frac{1-p}{p^2}$.

(These facts are well known. See chapter 3 of book for a proof.)

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2} \;=\; \frac{\sum_{j=1}^{n} \mathrm{Var}[X_j]}{a^2}.$$

Bounding Coupon Collector purchases - Chebyshev

Each X_i is a geometric random variable with parameter $p = \frac{n-(i-1)}{n}$.

Fact (3.8)

For any geometric random variable X with parameter p,

$$E[X] = p^{-1}, \quad \textit{and} \quad Var[X] = \frac{1-p}{p^2}.$$

(These facts are well known. See chapter 3 of book for a proof.)

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2} \;=\; \frac{\sum_{j=1}^{n} \mathrm{Var}[X_j]}{a^2}.$$

Each individual X_j is geometric with parameter $\frac{n-(j-1)}{n}$, so each X_j has

$$\mathrm{Var}[X_j] \;=\; \frac{j-1}{n}\left(\frac{n}{(n+1-j)}\right)^2 \;\leq\; \left(\frac{n}{n+1-j}\right)^2.$$

Bounding Coupon Collector purchases - Chebyshev

Each X_i is a geometric random variable with parameter $p = \frac{n-(i-1)}{n}$.

Fact (3.8)

For any geometric random variable X with parameter p,

$E[X] = p^{-1}$, *and* $Var[X] = \frac{1-p}{p^2}$.

(These facts are well known. See chapter 3 of book for a proof.)

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2} \;=\; \frac{\sum_{j=1}^{n} \mathrm{Var}[X_j]}{a^2}.$$

Each individual X_j is geometric with parameter $\frac{n-(j-1)}{n}$, so each X_j has

$$\mathrm{Var}[X_j] \;=\; \frac{j-1}{n}\left(\frac{n}{(n+1-j)}\right)^2 \;\leq\; \left(\frac{n}{n+1-j}\right)^2.$$

$$\mathrm{Var}[X] \;\leq\; n^2 \sum_{j=1}^{n} \left(\frac{1}{n+1-j}\right)^2 = n^2 \sum_{j=1}^{n} \left(\frac{1}{j}\right)^2$$

Bounding Coupon Collector purchases - Chebyshev

Theorem *[Euler,1741]*

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$$

Hence,

$$\mathrm{Var}[X] \leq n^2 \sum_{j=1}^{n} \frac{1}{j^2} \leq \frac{n^2\pi^2}{6}.$$

Suppose we are willing to make $2\mathrm{E}[X] = 2nH_n$ purchases. The probability we fail to get all cards is

$$\begin{aligned} \Pr[X > 2\mathrm{E}[X]] &= \Pr[X - \mathrm{E}[X] > \mathrm{E}[X]] \\ &= \Pr[|X - \mathrm{E}[X]| > \mathrm{E}[X]]. \qquad \text{(as } X \geq 0\text{)} \end{aligned}$$

Bounding Coupon Collector purchases - Chebyshev

Theorem *[Euler,1741]*

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$$

Hence,

$$\mathrm{Var}[X] \leq n^2 \sum_{j=1}^{n} \frac{1}{j^2} \leq \frac{n^2\pi^2}{6}.$$

Suppose we are willing to make $2\mathrm{E}[X] = 2nH_n$ purchases. The probability we fail to get all cards is

$$\begin{aligned} \Pr[X > 2\mathrm{E}[X]] &= \Pr[X - \mathrm{E}[X] > \mathrm{E}[X]] \\ &= \Pr[|X - \mathrm{E}[X]| > \mathrm{E}[X]]. \qquad \text{(as } X \geq 0\text{)} \end{aligned}$$

Using Chebyshev Inequality with $a = \mathrm{E}[X]$:

$$\begin{aligned} \Pr[|X - \mathrm{E}[X]| \geq \mathrm{E}[X]] &\leq \frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} \leq \frac{\pi^2 n^2}{6n^2 H_n^2} \\ &= \frac{\pi^2}{6H_n^2} \leq \frac{2}{\ln^2 n}. \end{aligned}$$

Note: This improves over $\frac{1}{2}$, which is what Markov gives us.

Bounding Coupon Collector purchases - Union bound

Theorem (1.2, Union bound)

Let $E_1, E_2, \ldots$ be a finite or countably infinite sequence of events. Then

$$\Pr\left[\bigcup_{i \geq 1} E_i\right] \leq \sum_{i \geq 1} \Pr[E_i].$$

Similar to Markov ineq., there is almost no requirement to the union bound!

Bounding Coupon Collector purchases - Union bound

Let E_i be the "bad" event where card i is still missing at time T.

$$\Pr\left[E_i\right] \leq \left(1 - \frac{1}{n}\right)^T.$$

Thus, by a union bound,

$$\Pr[X \geq T] = \Pr\left[\cup_{i \geq 1}^{n} E_i\right] \leq n\left(1 - \frac{1}{n}\right)^T.$$

Bounding Coupon Collector purchases - Union bound

Once again we use $(1-1/n)^n \leq 1/e$. If $T = (1+\varepsilon)n \ln n$,

$$n\left(1-\frac{1}{n}\right)^T \leq n\left(\left(1-\frac{1}{n}\right)^n\right)^{(1+\varepsilon)\ln n}$$
$$\leq n(e^{-1})^{(1+\varepsilon)\ln n} = n^{-\varepsilon}.$$

Thus, for example if $\varepsilon = 1$,

$$\Pr[X \geq 2n \ln n] \leq n^{-1}.$$

As $\mathrm{E}[X] \geq n \ln n$,

$$\Pr[X \geq 2\mathrm{E}[X]] \leq \Pr[X \geq 2n \ln n] \leq n^{-1}.$$

Coupon collector bounds

$$\Pr[X \geq 2\mathrm{E}[X]] \leq \frac{1}{2} \qquad \text{(Markov)}$$

$$\Pr[X \geq 2\mathrm{E}[X]] \leq \frac{2}{\ln(n)^2} \qquad \text{(Chebyshev)}$$

$$\Pr[X \geq 2\mathrm{E}[X]] \leq \frac{1}{n} \qquad \text{(Union bound)}$$

Using "Chernoff bounds" for "negatively correlated" r.v.'s, one can also show

$$\Pr[X \leq (1-\varepsilon)(n-1)\ln n] \leq e^{-n^{\varepsilon}}.$$

However, we will not establish this result in this course.

Wrapping up today

Next week we will continue the theme of “bounding deviation from the mean” by introducing some very important concentration inequalities, which apply first and foremost to sums of independent random variables, called

Chernoff bounds / Hoeffding’s inequality.

First, in the next lecture we give a simple randomized algorithm to approximate the **Maximum** Cut in a graph, and show how to *derandomize* it using conditional expectation.