Discussion 14A

1. **How Many Coupons?**

Consider the coupon collecting problem covered in note 17. There are n distinct types of coupons that we wish to collect. Every time we buy a box, there is one coupon in it, with equal likelihood of being any one of the types of coupons. We want to figure out how many boxes we need to buy in order to get one of each coupon. For this problem, we want to bound the probability that we have to buy lots of boxes — say substantially more than $n \log n$ boxes.

a) We represent X, the number of boxes we have to buy, as a sum of other random variables. Let X_i represent the number of boxes you buy to go from $i-1$ to i *distinct* coupons in your hand. Write X as a sum of X_i's. Argue that each X_i is an independent random variable with a geometric distribution.

b) We wish to use Chebyshev's inequality to bound the probability we have to buy substantially more than $n \log n$ boxes. In order to do this, we need to compute the variance of a geometric random variable. There are multiple ways of doing this, including a recursive trick like that used in Lily's Lottery.

Another approach is to use series techniques. We use the following lemma in our proof:

$$\sum_{k=1}^{\infty} k(k+1)(1-p)^{k-1} = \frac{2}{p^3}.$$

Prove this lemma. (Hint: what is the sum of the geometric series $\sum_{k=0}^{\infty}(1-p)^k$? Take the derivative of both sides - what happens?)

c) If $X_i \sim \text{Geo}(p)$, show that $\mathbb{E}[X_i^2] = \sum_{k=1}^{\infty} k(k+1)p(1-p)^{k-1} - \sum_{k=1}^{\infty} kp(1-p)^{k-1}$

d) Use your lemma and the fact that $\mathbb{E}[X_i] = 1/p$ to simplify part c) to show $\mathbb{E}[X_i^2] = \frac{2}{p^2} - \frac{1}{p}$

e) Show that variance of a geometric variable with parameter p is $\frac{1-p}{p^2}$. We will later use the simpler upper bound, $Var[X_i] < \frac{1}{p^2}$.

f) Make use of the fact that $\sum_{i=1}^{\infty} \frac{1}{i^2}$ is a positive constant $\frac{\pi^2}{6} \leq 2$ to show that the $Var[X] \leq 2n^2$.

g) This means that the standard deviation for X scales like n and not like the expectation $n \ln n$.
Use Chebyshev's inequality to show that $Pr[X \geq \alpha n \ln n]$ tends to zero for any $\alpha > 1$ as $n \to \infty$. (Hint: Recall that we estimated in the note that $\mathbb{E}[X] \approx n(\ln n + \gamma) \approx n \ln n$.)

2. **How many errors?**

We want to send a length-n message through a noisy channel with 50% redundancy added (so there are $\frac{3}{2}n$ symbols total in the codeword). We use a Reed-Solomon code and you can assume that the underlying finite field is $GF(q)$ with $q > 2n$. In order to ensure an overall probability of 0.9 that the receiver can decode the entire message correctly, how noisy can the channel be? i.e. what is the highest probability of symbol error that we can tolerate?