Linear Programming
Problem Set 4 – CS 6515/4540 (Fall 2025)

Solution to problem set 4, question 16

16 Probability Theory

1. A simple example uses two identical Bernoulli random variables. Let X be Bernoulli(p) (i.e. $X \in \{0, 1\}$ and $\Pr[X = 1] = p$), and let $Y = X$ (so X and Y are perfectly correlated). Both X and Y are nonnegative.

   Then
   $$\mathbb{E}[X] = \mathbb{E}[Y] = p, \qquad \mathbb{E}[XY] = \mathbb{E}[X \cdot X] = \mathbb{E}[X] = p.$$

   Hence
   $$\mathbb{E}[XY] = p \neq p^2 = \mathbb{E}[X]\mathbb{E}[Y],$$
   whenever $0 < p < 1$. For instance, with $p = \frac{1}{2}$ we get $\mathbb{E}[XY] = \frac{1}{2}$ while $\mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{4}$.

   This exhibits a pair of nonnegative, correlated random variables with $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$.

2. (a) For $k = 1, 2, \ldots, n$ let X_k be the number of additional draws required to increase the number of distinct coupons seen from $k - 1$ to k. Then
   $$X \;=\; \sum_{k=1}^{n} X_k.$$

   Conditioned on having $k - 1$ distinct coupons, the probability that a single subsequent draw produces a *new* coupon is
   $$p_k = \frac{n - (k - 1)}{n} = \frac{n - k + 1}{n}.$$
   Therefore X_k is a geometric random variable with success probability p_k, so
   $$\mathbb{E}[X_k] = \frac{1}{p_k} = \frac{n}{n - k + 1}.$$

   By linearity of expectation,
   $$\mathbb{E}[X] = \sum_{k=1}^{n} \mathbb{E}[X_k] = n \sum_{k=1}^{n} \frac{1}{n - k + 1} = n \sum_{j=1}^{n} \frac{1}{j} = nH_n,$$
   where H_n is the n-th harmonic number. Using the bound $H_n \leq 1 + \ln n \; + \gamma$; where γ is the Euler-Mascheroni constant
   $$\mathbb{E}[X] \leq n(C + \ln n) = O(n \log n).$$
   for some constant $C > 0$.

   (b) Fix a constant $d > 1$ and set the number of draws as
   $$D = d\, n \ln n.$$

   For a fixed coupon i, the probability it is never drawn in D independent draws is
   $$P[\text{coupon } i \text{ missing after } Ddraws] = \Big(1 - \frac{1}{n}\Big)^{D} \leq \exp\Big(-\frac{D}{n}\Big) = \exp(-d \ln n) = n^{-d},$$

where we used $1 - x \leq e^{-x}$. By the union bound, the probability that some random coupon is missing after D draws is at most

$$\sum_i P(\text{coupon } i \text{ missing after D draws}) = n \cdot n^{-d} = n^{1-d}.$$

Choosing $d = 2$ gives failure probability $n^{1-2} = 1/n$. Hence, with probability at least $1 - 1/n$ every coupon appears within

$$D = 2\, n \ln n = O(n \log n)$$

draws. This proves the desired high-probability bound.