### 2.4.1. *Example: Coupon Collector's Problem*

The coupon collector's problem arises from the following scenario. Suppose that each box of cereal contains one of $n$ different coupons. Once you obtain one of every type of coupon, you can send in for a prize. Assuming that the coupon in each box is chosen independently and uniformly at random from the $n$ possibilities and that you do not collaborate with others to collect coupons, how many boxes of cereal must you buy before you obtain at least one of every type of coupon? This simple problem arises in many different scenarios and will reappear in several places in the book.

Let $X$ be the number of boxes bought until at least one of every type of coupon is obtained. We now determine $\mathbf{E}[X]$. If $X_i$ is the number of boxes bought while you had exactly $i-1$ different coupons, then clearly $X = \sum_{i=1}^{n} X_i$.

The advantage of breaking the random variable $X$ into a sum of $n$ random variables $X_i$, $i = 1, \ldots, n$, is that each $X_i$ is a geometric random variable. When exactly $i-1$ coupons have been found, the probability of obtaining a new coupon is

$$p_i = 1 - \frac{i-1}{n}.$$

Hence, $X_i$ is a geometric random variable with parameter $p_i$, and

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}.$$

Using the linearity of expectations, we have that

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbf{E}[X_i] \\ &= \sum_{i=1}^{n} \frac{n}{n-i+1} \\ &= n \sum_{i=1}^{n} \frac{1}{i}.\end{aligned}$$

The summation $\sum_{i=1}^{n} 1/i$ is known as the *harmonic number* $H(n)$, and as we show next, $H(n) = \ln n + \Theta(1)$. Thus, for the coupon collector's problem, the expected number of random coupons required to obtain all $n$ coupons is $n \ln n + \Theta(n)$.

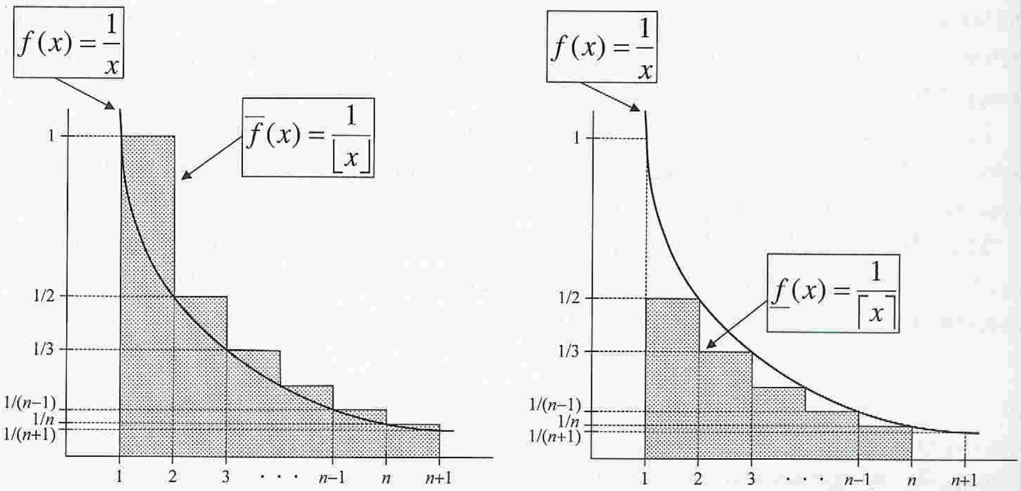
**Lemma 2.10:** *The harmonic number* $H(n) = \sum_{i=1}^{n} 1/i$ *satisfies* $H(n) = \ln n + \Theta(1)$.

***Proof:*** Since $1/x$ is monotonically decreasing, we can write

$$\ln n = \int_{x=1}^{n} \frac{1}{x}\,dx \le \sum_{k=1}^{n} \frac{1}{k}$$

and

$$\sum_{k=2}^{n} \frac{1}{k} \le \int_{x=1}^{n} \frac{1}{x}\,dx = \ln n.$$



(a) Approximating $1/x$ from above.

(b) Approximating $1/x$ from below.

**Figure 2.1:** Approximating the area below $f(x) = 1/x$.

This is clarified in Figure 2.1, where the area below the curve $f(x) = 1/x$ corresponds to the integral and the areas of the shaded regions correspond to the summations $\sum_{k=1}^{n} 1/k$ and $\sum_{k=2}^{n} 1/k$.

Hence $\ln n \le H(n) \le \ln n + 1$, proving the claim. ■

As a simple application of the coupon collector's problem, suppose that packets are sent in a stream from a source host to a destination host along a fixed path of routers. The host at the destination would like to know which routers the stream of packets has passed through, in case it finds later that some router damaged packets that it processed. If there is enough room in the packet header, each router can append its identification number to the header, giving the path. Unfortunately, there may not be that much room available in the packet header.

Suppose instead that each packet header has space for exactly one router identification number, and this space is used to store the identification of a router chosen uniformly at random from all of the routers on the path. This can actually be accomplished easily; we consider how in Exercise 2.18. Then, from the point of view of the destination host, determining all the routers on the path is like a coupon collector's problem. If there are $n$ routers along the path, then the expected number of packets in the stream that must arrive before the destination host knows all of the routers on the path is $nH(n) = n \ln n + \Theta(n)$.