Probability Definitions and Tools

Sam McCauley
Applied Algorithms, Fall 2024

Goal for Today

Throughout the course we've seen a few definitions for analyzing randomized algorithms, as well as some mathematical tools for analyzing complicated expressions that come up in when performing analysis using these definitions. The goal of this document is to put these definitions and tools in one place so they can be looked up more easily.

Probability Definitions

Two events are independent (intuitively) if they do not affect each other. Independence is crucial because if X and Y are independent, then $\Pr[X \text{ and } Y] = \Pr[X] \cdot \Pr[Y]$.

Definition 1. Events X and Y are independent if $\Pr[X|Y] = \Pr[X]$ or $\Pr[Y|X] = \Pr[Y]$.

Usually we determine independence based on the random process itself rather than the formal definition. For example, if I flip a coin twice in a row, the outcomes of the two flips are independent (since neither affects the other). Similarly, we often assume that hashes are independent: the probability that $h(x) = a$ is independent of the probability that $h(y) = b$, for any x, y, a, b.

Oftentimes, we care about the *average* performance of a random process. This is captured by expectation: in short, expectation is the average value of the result, weighted by the probability of each outcome.

Definition 2. Given a random variable X that takes values in $\{0, \ldots, k\}$, the expectation of X is

$$\mathsf{E}[X] = \sum_{i=0}^{k} i \cdot \Pr[X = i].$$

For some use cases, average does not capture what we want: we want to capture how often a value is out of a certain range. These are often called "concentration bounds." In randomized algorithms, concentration bounds are often given using the following definition.

Definition 3. An event X occurs *with high probability* if $\Pr[X] \geq 1 - 1/n$.

Note that we must have some value n for us to reference when saying that something occurs with high probability. Usually, n is the size of the input to the problem. Let's see one example of how to use this definition.

Example. How many times do I need to flip a coin until I see a heads with high probability?

The probability that I do not see a heads after k coin flips is $1/2^k$. I want the probability that I *do not* see a heads to be $\leq 1/n$ (that way, the probability that I *do* see a heads is $\geq 1 - 1/n$.) Therefore, I want $1/2^k \leq 1/n$. Solving, after $\log_2 n$ coin flips I see a heads with high probability.

When events are not independent, the following loose bound can be very useful for analyzing how they interact.

Definition 4 (Union Bound)**.** For any[a] events X_1 and X_2,

$$\Pr[X_1 \text{ or } X_2] \leq \Pr[X_1] + \Pr[X_2].$$

[a] Even if X_1 and X_2 are not independent!

Logarithms

The definition of a logarithm is that if $b^x = y$, then $\log_b y = x$.

Let's recall some classic log rules:

Theorem 1. For any $a, b, x, y > 0$:

$$\log_b xy = \log_b x + \log_b y \qquad \log_b x/y = \log_b x - \log_b y$$

$$\log_a x = \frac{\log_b x}{\log_b a} \qquad \log_b x^y = y \log_b x$$

Sometimes we see some somewhat complicated expressions involving logarithms, especially when there are logarithms in exponents. My recommendation to simplify them is to use the above log rules to: (1) make all the logs base 2, and (2) put everything into the exponent with 2 as a base.

Let's see some an example of this idea in action.

Example. Let's simplify $n^{1/\log_2 n}$.

Using the above idea, let's make the equation into an exponent with 2 as the base. Substituting $n = 2^{\log n}$, we have that $n^{1/\log_2 n} = (2^{\log n})^{1/\log_2 n}$. From algebra, we know that $(a^b)^c = a^{bc}$, so $(2^{\log_2 n})^{1/\log_2 n} = 2^{(\log_2 n)/\log_2 n} = 2$.

Therefore, $n^{1/\log_2 n} = 2$.

Sometimes, the "wrong" variable is in the base of the number. The following theorem is a black box way to fix it, and its proof uses this same idea of putting everything into an exponent of 2.

Theorem 2. For any $a, b, c > 0$,

$$a^{\log_b c} = c^{\log_b a}.$$

Proof. Let's move everything into the base of 2.

$$a^{\log_b c} = (2^{\log_2 a})^{(\log_2 c)/\log_2 b} = 2^{\log_2 a \cdot \log_2 c/\log_2 b}.$$

Now, a similar sequence of steps get us to the right side of the equation from the theorem.

$$2^{\log_2 a \cdot \log_2 c/\log_2 b} = (2^{\log_2 c})^{\log_2 a/\log_2 b} = c^{\log_b a}.$$

□

Simplifying Probabilistic Expressions

It is very common that probabilistic expressions are like $(1 - 1/x)^y$ or $(1 + 1/x)^y$. Let's discuss a few tools to simplify them.

There are two key equations we will use here.

Theorem 3. For any $x > 2$,

$$2 \leq \left(1 + \frac{1}{x}\right)^x \leq e \quad \text{and} \quad \frac{1}{4} \leq \left(1 - \frac{1}{x}\right)^x \leq \frac{1}{e}.$$

In fact, $(1 + 1/x)^x$ and $(1 - 1/x)^x$ are increasing, so the right inequalities get closer and closer to equalities as x gets larger.

Let's look at an example probability problem that uses this inequality.

Example. Let's say we store n items in a hash table with chaining; the hash table has exactly n buckets. Let's say our hash function is perfectly random: it maps each item to a given bucket with probability $1/n$, independent of the bucket of any other item. What is the probability that a given bucket (say bucket 1) is empty?

Bucket 1 is empty if and only if no item hashes to it. The probability that the first item *does* hash to bucket 1 is $1/n$—so the probability that the first item *does not* hash to bucket 1 is $1 - 1/n$. Similarly, the probability that the second item does not hash to bucket 1 is $1 - 1/n$, and so on. Each item's probability of hashing to bucket 1 is independent, so we can multiply them to find the probability of all the events occurring. Therefore, bucket 1 is empty with probability

$$(1 - 1/n)^n \leq 1/e.$$

As mentioned above, $(1 - 1/n)^n \approx 1/e$ (i.e. $1/e$ is in fact the correct answer) if n is large. We can also say that so long as $n \geq 2$, bucket 1 is empty with probability at least $1/4$.

Oftentimes, we want to deal with expressions like $(1 - 1/x)^y$—that is to say, the exponent does not exactly match up with the value in the parentheses. Fortunately, an extra step of algebra solves this problem. Long story short, we can do the following:

Theorem 4. For any $x, y \geq 2$,

$$\left(\frac{1}{2}\right)^{y/x} \leq \left(1 - \frac{1}{x}\right)^y \leq \left(\frac{1}{e}\right)^{y/x}$$

Proof. First, the right side inequality:

$$\left(1 - \frac{1}{x}\right)^y = \left(\left(1 - \frac{1}{x}\right)^x\right)^{y/x} \leq \left(\frac{1}{e}\right)^{y/x}.$$

Now, the left:

$$\left(1 - \frac{1}{x}\right)^y = \left(\left(1 - \frac{1}{x}\right)^x\right)^{y/x} \geq \left(\frac{1}{2}\right)^{y/x}.$$ □

Let's see two examples of how to use this. First, we'll look at a slightly bigger hash table; then, a significantly smaller one.

Example. Let's say we store n items in a hash table with chaining; the hash table has exactly $2n$ buckets. What is the probability that a given bucket (say bucket 1) is empty?

Now, the probability that a given item does not hash to bucket 1 is $1 - 1/(2n)$. As before, bucket 1 is empty with probability

$$\left(1 - \frac{1}{2n}\right)^n = \left(\left(1 - \frac{1}{2n}\right)^2 n\right)^{1/2} \leq \frac{1}{e^{1/2}}.$$

So when we double the size of our hash table, the expected fraction of empty buckets goes from (assuming n is large) $\approx 1/e \approx .37$ to $\approx 1/\sqrt{e} \approx .61$.

Now, let's look at a smaller table. We'll see that a given bucket is full with high probability.

Example. Let's say we store n items in a hash table with chaining; the hash table has exactly[a] $n/\log_e n$ buckets. What is the probability that a given bucket (say bucket 1) is empty?

Now, the probability that a given item does not hash to bucket 1 is $1 - (\log_e n)/n$. As before, bucket 1 is empty with probability

$$\left(1 - \frac{\log n}{n}\right)^n = \left(\left(1 - \frac{\log n}{n}\right)^{n/\log n}\right)^{\log n} \leq \frac{1}{e^{\log_e n}} = 1/n.$$

Therefore, we can say that if there are $n/\log n$ buckets, each of them is nonempty with high probability.

[a] Log base e is usually called the natural log, and written $\ln n$. I'll keep the base e to be explicit.