

*But, on the other hand, Uncle Abner said that the person that had took a bull by the tail once had learnt sixty or seventy times as much as a person that hadn't, and said a person that started in to carry a cat home by the tail was getting knowledge that was always going to be useful to him, and warn't ever going to grow dim or doubtful.*

— Mark Twain, *Tom Sawyer Abroad* (1894)

## \*11 Tail Inequalities

The simple recursive structure of skip lists made it relatively easy to derive an upper bound on the expected *worst-case* search time, by way of a stronger high-probability upper bound on the worst-case search time. We can prove similar results for treaps, but because of the more complex recursive structure, we need slightly more sophisticated probabilistic tools. These tools are usually called *tail inequalities*; intuitively, they bound the probability that a random variable with a bell-shaped distribution takes a value in the *tails* of the distribution, far away from the mean.

### 11.1 Markov's Inequality

Perhaps the simplest tail inequality was named after the Russian mathematician Andrey Markov; however, in strict accordance with Stigler's Law of Eponymy, it first appeared in the works of Markov's probability teacher, Pafnuty Chebyshev.<sup>1</sup>

**Markov's Inequality.** *Let  $X$  be a non-negative integer random variable. For any  $t > 0$ , we have  $\Pr[X \geq t] \leq E[X]/t$ .*

**Proof:** The inequality follows from the definition of expectation by simple algebraic manipulation.

$$\begin{aligned}
 E[X] &= \sum_{k=0}^{\infty} k \cdot \Pr[X = k] && \text{[definition of } E[X]\text{]} \\
 &= \sum_{k=0}^{\infty} \Pr[X \geq k] && \text{[algebra]} \\
 &\geq \sum_{k=0}^{t-1} \Pr[X \geq k] && \text{[since } t < \infty\text{]} \\
 &\geq \sum_{k=0}^{t-1} \Pr[X \geq t] && \text{[since } k < t\text{]} \\
 &= t \cdot \Pr[X \geq t] && \text{[algebra]} \quad \square
 \end{aligned}$$

Unfortunately, the bounds that Markov's inequality implies (at least directly) are often very weak, even useless. (For example, Markov's inequality implies that with high probability, every node in an  $n$ -node treap has depth  $O(n^2 \log n)$ . Well, *duh!*) To get stronger bounds, we need to exploit some additional structure in our random variables.

<sup>1</sup>The closely related tail bound traditionally called Chebyshev's inequality was actually discovered by the French statistician Irénée-Jules Bienaymé, a friend and colleague of Chebyshev's.

## 11.2 Independence

A set of random variables  $X_1, X_2, \dots, X_n$  are said to be *mutually independent* if and only if

$$\Pr \left[ \bigwedge_{i=1}^n (X_i = x_i) \right] = \prod_{i=1}^n \Pr[X_i = x_i]$$

for all possible values  $x_1, x_2, \dots, x_n$ . For examples, different flips of the same fair coin are mutually independent, but the number of heads and the number of tails in a sequence of  $n$  coin flips are not independent (since they must add to  $n$ ). Mutual independence of the  $X_i$ 's implies that the expectation of the product of the  $X_i$ 's is equal to the product of the expectations:

$$\mathbb{E} \left[ \prod_{i=1}^n X_i \right] = \prod_{i=1}^n \mathbb{E}[X_i].$$

Moreover, if  $X_1, X_2, \dots, X_n$  are independent, then for any function  $f$ , the random variables  $f(X_1), f(X_2), \dots, f(X_n)$  are also mutually independent.

— Discuss limited independence? —  
— Add Chebychev and other moment inequalities? —

## 11.3 Chernoff Bounds

— Replace with Mihai's exponential-moment derivation! —

Suppose  $X = \sum_{i=1}^n X_i$  is the sum of  $n$  mutually independent random *indicator* variables  $X_i$ . For each  $i$ , let  $p_i = \Pr[X_i = 1]$ , and let  $\mu = \mathbb{E}[X] = \sum_i \mathbb{E}[X_i] = \sum_i p_i$ .

**Chernoff Bound (Upper Tail).**  $\Pr[X > (1 + \delta)\mu] < \left( \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^\mu$  for any  $\delta > 0$ .

**Proof:** The proof is fairly long, but it relies on just a few basic components: a clever substitution, Markov's inequality, the independence of the  $X_i$ 's, The World's Most Useful Inequality  $e^x > 1 + x$ , a tiny bit of calculus, and lots of high-school algebra.

We start by introducing a variable  $t$ , whose role will become clear shortly.

$$\Pr[X > (1 + \delta)\mu] = \Pr[e^{tX} > e^{t(1+\delta)\mu}]$$

To cut down on the superscripts, I'll usually write  $\exp(x)$  instead of  $e^x$  in the rest of the proof. Now apply Markov's inequality to the right side of this equation:

$$\Pr[X > (1 + \delta)\mu] < \frac{\mathbb{E}[\exp(tX)]}{\exp(t(1 + \delta)\mu)}.$$

We can simplify the expectation on the right using the fact that the terms  $X_i$  are independent.

$$\mathbb{E}[\exp(tX)] = \mathbb{E} \left[ \exp \left( t \sum_i X_i \right) \right] = \mathbb{E} \left[ \prod_i \exp(tX_i) \right] = \prod_i \mathbb{E}[\exp(tX_i)]$$

We can bound the individual expectations  $E[\exp(tX_i)]$  using The World's Most Useful Inequality:

$$E[\exp(tX_i)] = p_i e^t + (1 - p_i) = 1 + (e^t - 1)p_i < \exp((e^t - 1)p_i)$$

This inequality gives us a simple upper bound for  $E[e^{tX}]$ :

$$E[\exp(tX)] < \prod_i \exp((e^t - 1)p_i) < \exp\left(\sum_i (e^t - 1)p_i\right) = \exp((e^t - 1)\mu)$$

Substituting this back into our original fraction from Markov's inequality, we obtain

$$\Pr[X > (1 + \delta)\mu] < \frac{E[\exp(tX)]}{\exp(t(1 + \delta)\mu)} < \frac{\exp((e^t - 1)\mu)}{\exp(t(1 + \delta)\mu)} = (\exp(e^t - 1 - t(1 + \delta)))^\mu$$

Notice that this last inequality holds for *all* possible values of  $t$ . To obtain the final tail bound, we will choose  $t$  to make this bound as small as possible. To minimize  $e^t - 1 - t - t\delta$ , we take its derivative with respect to  $t$  and set it to zero:

$$\frac{d}{dt}(e^t - 1 - t(1 + \delta)) = e^t - 1 - \delta = 0.$$

(And you thought calculus would never be useful!) This equation has just one solution  $t = \ln(1 + \delta)$ . Plugging this back into our bound gives us

$$\Pr[X > (1 + \delta)\mu] < (\exp(\delta - (1 + \delta)\ln(1 + \delta)))^\mu = \left(\frac{e^\delta}{(1 + \delta)^{1 + \delta}}\right)^\mu$$

And we're done! □

This form of the Chernoff bound can be a bit clumsy to use. A more complicated argument gives us the bound

$$\Pr[X > (1 + \delta)\mu] < e^{-\mu\delta^2/3} \text{ for any } 0 < \delta < 1.$$

A similar argument gives us an inequality bounding the probability that  $X$  is significantly *smaller* than its expected value:

**Chernoff Bound (Lower Tail).**  $\Pr[X < (1 - \delta)\mu] < \left(\frac{e^{-\delta}}{(1 - \delta)^{1 - \delta}}\right)^\mu < e^{-\mu\delta^2/2}$  for any  $\delta > 0$ .

### 11.4 Back to Treaps

In our analysis of randomized treaps, we wrote  $i \uparrow k$  to indicate that the node with the  $i$ th smallest key ('node  $i$ ') was a proper ancestor of the node with the  $k$ th smallest key ('node  $k$ '). We argued that

$$\Pr[i \uparrow k] = \frac{[i \neq k]}{|k - i| + 1},$$

and from this we concluded that the expected depth of node  $k$  is

$$E[\text{depth}(k)] = \sum_{i=1}^n \Pr[i \uparrow k] = H_k + H_{n-k} - 2 < 2 \ln n.$$

To prove a worst-case expected bound on the depth of the tree, we need to argue that the *maximum* depth of any node is small. Chernoff bounds make this argument easy, once we establish that the relevant indicator variables are mutually independent.

**Lemma 1.** For any index  $k$ , the  $k-1$  random variables  $[i \uparrow k]$  with  $i < k$  are mutually independent. Similarly, for any index  $k$ , the  $n-k$  random variables  $[i \uparrow k]$  with  $i > k$  are mutually independent.

**Proof:** We explicitly consider only the first half of the lemma when  $k = 1$ , although the argument generalizes easily to other values of  $k$ . To simplify notation, let  $X_i$  denote the indicator variable  $[i \uparrow 1]$ . Fix  $n-1$  arbitrary indicator values  $x_2, x_3, \dots, x_n$ . We prove the lemma by induction on  $n$ , with the vacuous base case  $n = 1$ . The definition of conditional probability gives us

$$\begin{aligned} \Pr \left[ \bigwedge_{i=2}^n (X_i = x_i) \right] &= \Pr \left[ \bigwedge_{i=2}^{n-1} (X_i = x_i) \wedge X_n = x_n \right] \\ &= \Pr \left[ \bigwedge_{i=2}^{n-1} (X_i = x_i) \mid X_n = x_n \right] \cdot \Pr [X_n = x_n] \end{aligned}$$

Now recall that  $X_n = 1$  (which means  $1 \uparrow n$ ) if and only if node  $n$  has the smallest priority of all nodes. The other  $n-2$  indicator variables  $X_i$  depend only on the order of the priorities of nodes 1 through  $n-1$ . There are exactly  $(n-1)!$  permutations of the  $n$  priorities in which the  $n$ th priority is smallest, and each of these permutations is equally likely. Thus,

$$\Pr \left[ \bigwedge_{i=2}^{n-1} (X_i = x_i) \mid X_n = x_n \right] = \Pr \left[ \bigwedge_{i=2}^{n-1} (X_i = x_i) \right]$$

The inductive hypothesis implies that the variables  $X_2, \dots, X_{n-1}$  are mutually independent, so

$$\Pr \left[ \bigwedge_{i=2}^{n-1} (X_i = x_i) \right] = \prod_{i=2}^{n-1} \Pr [X_i = x_i].$$

We conclude that

$$\Pr \left[ \bigwedge_{i=2}^n (X_i = x_i) \right] = \Pr [X_n = x_n] \cdot \prod_{i=2}^{n-1} \Pr [X_i = x_i] = \prod_{i=1}^{n-1} \Pr [X_i = x_i],$$

or in other words, that the indicator variables are mutually independent. □

**Theorem 2.** The depth of a randomized treap with  $n$  nodes is  $O(\log n)$  with high probability.

**Proof:** First let's bound the probability that the depth of node  $k$  is at most  $8 \ln n$ . There's nothing special about the constant 8 here; I'm being generous to make the analysis easier.

The depth is a sum of  $n$  indicator variables  $A_k^i$ , as  $i$  ranges from 1 to  $n$ . Our Observation allows us to partition these variables into two mutually independent subsets. Let  $d_{<}(k) = \sum_{i < k} [i \uparrow k]$  and  $d_{>}(k) = \sum_{i > k} [i \uparrow k]$ , so that  $\text{depth}(k) = d_{<}(k) + d_{>}(k)$ . If  $\text{depth}(k) > 8 \ln n$ , then either  $d_{<}(k) > 4 \ln n$  or  $d_{>}(k) > 4 \ln n$ .

Chernoff's inequality, with  $\mu = \mathbb{E}[d_{<}(k)] = H_k - 1 < \ln n$  and  $\delta = 3$ , bounds the probability that  $d_{<}(k) > 4 \ln n$  as follows.

$$\Pr[d_{<}(k) > 4 \ln n] < \Pr[d_{<}(k) > 4\mu] < \left(\frac{e^3}{4^4}\right)^\mu < \left(\frac{e^3}{4^4}\right)^{\ln n} = n^{\ln(e^3/4^4)} = n^{3-4 \ln 4} < \frac{1}{n^2}.$$

(The last step uses the fact that  $4 \ln 4 \approx 5.54518 > 5$ .) The same analysis implies that  $\Pr[d_{>}(k) > 4 \ln n] < 1/n^2$ . These inequalities imply the crude bound  $\Pr[\text{depth}(k) > 4 \ln n] < 2/n^2$ .

Now consider the probability that the treap has depth greater than  $10 \ln n$ . Even though the distributions of different nodes' depths are *not* independent, we can conservatively bound the probability of failure as follows:

$$\Pr \left[ \max_k \text{depth}(k) > 8 \ln n \right] = \Pr \left[ \bigwedge_{k=1}^n (\text{depth}(k) > 8 \ln n) \right] \leq \sum_{k=1}^n \Pr[\text{depth}(k) > 8 \ln n] < \frac{2}{n}.$$

This argument implies more generally that for any constant  $c$ , the depth of the treap is greater than  $c \ln n$  with probability at most  $2/n^{c \ln c - c}$ . We can make the failure probability an arbitrarily small polynomial by choosing  $c$  appropriately.  $\square$

This lemma implies that any search, insertion, deletion, or merge operation on an  $n$ -node treap requires  $O(\log n)$  time with high probability. In particular, the expected *worst-case* time for each of these operations is  $O(\log n)$ .

## Exercises

1. Prove that for any integer  $k$  such that  $1 < k < n$ , the  $n - 1$  indicator variables  $[i \uparrow k]$  with  $i \neq k$  are *not* mutually independent. [Hint: Consider the case  $n = 3$ .]
2. Recall from Exercise 1 in the previous note that the expected number of descendants of any node in a treap is  $O(\log n)$ . Why doesn't the Chernoff-bound argument for depth imply that, with high probability, *every* node in a treap has  $O(\log n)$  descendants? The conclusion is clearly bogus—Every treap has a node with  $n$  descendants!—but what's the hole in the argument?
3. Recall from the previous lecture note that a *heater* is a sort of anti-treap, in which the priorities of the nodes are given, but their search keys are generated independently and uniformly from the unit interval  $[0, 1]$ .

Prove that an  $n$ -node heater has depth  $O(\log n)$  with high probability.