*Popular tradition may be used to defend what seems irrational, and you can also say that sometimes it is not irrational, for it is likely that unlikely things should happen.*

— Aristotle, *Poetics* (c.335 BCE)

*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 gitting 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.[1]

**Markov's Inequality.** *Let $Z$ be a non-negative integer random variable. For any real number $z > 0$, we have* $\Pr[Z \geq z] \leq \mathrm{E}[Z]/z$.

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

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

□

[1]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. Just to be extra confusing, some sources refer to what we're calling Markov's inequality as "Chebyshev's inequality" or "Bienaymé's inequality".

In particular, Markov's inequality implies the following bound on the probability that any random variable $X$ is significantly larger than its expectation:

$$\Pr[X \geq (1+\delta)\,\mathrm{E}[X]] \leq \frac{1}{1+\delta}.$$

Unfortunately, the bounds that Markov's inequality (directly) implies are generally too weak to be useful. (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.

## 11.2 Independence

Two random variables are ***independent*** if knowing the value of one variable gives us no additional knowledge about the distribution of the other. More formally, $X$ and $Y$ are and only if

$$\Pr[X = x \ \wedge\ Y = y] = \Pr[X = x] \cdot \Pr[Y = y],$$

or equivalently,

$$\Pr[X = x \mid Y = y] = \Pr[X = x]$$

for all possible values $x$ and $y$. For example, two flips of the same (ideal) fair coin are independent, but the number of heads and the number of tails in a sequence of $n$ coin flips are not independent (because they must add to $n$). Variables that are not independent are called ***dependent***. Independence of $X$ and $Y$ has two important consequences.

- The expectation of the product of two independent random variables is equal to the product of their expectations $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.
- If $X$ and $Y$ are independent, then for any function $f$, the random variables $f(X)$ and $f(Y)$ are also independent.

Neither of these properties hold for dependent random variables.

More generally, a collection of random variables $X_1, X_2, \ldots, X_n$ are said to be ***mutually independent*** (or ***fully 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 $x_1, x_2, \ldots, x_n$. 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:

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

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

In some contexts, especially in the analysis of hashing, we can only realistically assume a limited form of independence. A set of random variables is ***pairwise independent*** if every pair of variables in the set is independent. More generally, for any positive integer $k$, we say that a set of random variables is ***$k$-wise independent*** if every subset of size $k$ is mutually independent. A set of variables is fully independent if and only if it is $k$-wise independent for *all* $k$.

Every $k$-wise independent set of random variables is also $j$-wise independent for all $j < k$, but the converse is not true. For example, let $X$, $Y$, and $Z$ be independent random bits, and let $W = (X + Y + Z) \bmod 2$. The set of random variables $\{W, X, Y, Z\}$ is 3-wise independent but not 4-wise (or fully) independent.

## 11.3 Chebyshev's Inequality

Most of our analysis of randomized algorithms and data structures boils down to analyzing sums of indicator variables.[2] So far we have only been interested in the expected value of these sums, but we can prove stronger conditions under various assumptions about independence.

Consider a collection $X_1, X_2, \ldots, X_n$ of random bits, and for each index $i$, let $p_i = \mathrm{E}[X_i] = \Pr[X_i = 1]$. Let $X$ denote the sum $\sum_i X_i$ and let $\mu = \mathrm{E}[X]$; linearity of expectation immediately implies that $\mu = \sum_i p_i$.

**Chebyshev's Inequality.** *If the indicator variables $X_1, X_2, \ldots, X_n$ are pairwise independent, then $\Pr[(X-\mu)^2 \geq z] < \mu/z$ for all $z > 0$.*

**Proof:** For each index $i$, let $Y_i := X_i - p_i$, and let $Y := \sum_i Y_i = X - \mu$.

$$\begin{aligned}
\mathrm{E}[Y^2] &= \mathrm{E}\left[\sum_{i,j} Y_i Y_j\right] && \text{[definition of } Y\text{]} \\
&= \sum_{i,j} \mathrm{E}[Y_i Y_j] && \text{[linearity]} \\
&= \sum_i \mathrm{E}[Y_i^2] + \sum_{i \neq j} \mathrm{E}[Y_i Y_j] \\
&= \sum_i \mathrm{E}[Y_i^2] + \sum_{i \neq j} \mathrm{E}[Y_i] \cdot \mathrm{E}[Y_j] && \text{[pairwise independence]} \\
&= \sum_i \mathrm{E}[Y_i^2] + 0 && \text{[}\mathrm{E}[Y_i] = 0\text{, by linearity]} \\
&= \sum_i \big((p_i(1-p_i)^2 + (1-p_i)(0-p_i)^2\big) && \text{[definition of E]} \\
&= \sum_i p_i(1-p_i) \;<\; \sum_i p_i \;=\; \mu
\end{aligned}$$

Because the random variable $Y^2$ is non-negative, Markov's inequality now directly implies the bound $\Pr[Y^2 \geq z] \leq \mathrm{E}[Y^2]/z < \mu/z$, which completes the proof. □

Chebyshev's inequality immediately gives us significantly tighter bounds than Markov's inequality when the component indicator variables $X_i$ are pairwise independent. The following bounds hold for all positive real numbers $\Delta$ and $\delta$:

$$\boxed{\begin{array}{ll}
\Pr[X \geq \mu + \Delta] < \dfrac{\mu}{\Delta^2} & \Pr[X \geq (1+\delta)\mu] < \dfrac{1}{\delta^2 \mu} \\[2ex]
\Pr[X \leq \mu - \Delta] < \dfrac{\mu}{\Delta^2} & \Pr[X \leq (1-\delta)\mu] < \dfrac{1}{\delta^2 \mu}
\end{array}}$$

[2] This focus on sums of indicator variables is a feature (or if you prefer, a handicap) of these lecture notes, not of randomized algorithm analysis more broadly!

The inequalities on the left are called *additive* tail bounds; the inequalities on the right are called *multiplicative* tail bounds. The inequalities in the top row bounds the *upper tail* of $X$'s probability distribution; the inequalities in the bottom row bound the *lower tail*.

An important consequence of Chebyshev's inequality is the so-called ***Law of Large Numbers***, which states that if we repeat the same experiment many times, the statistical average of the outcomes tends toward the expected value of a single experiment with overwhelming probability. More formally, let $X_1, X_2, X_3, \ldots$ be independent random bits, with $\Pr[X_i = 1] = p$ for all $i$, and for any index $n$, let $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ denote the mean of the first $n$ bits. The ***weak*** law of large numbers states that

$$\lim_{n\to\infty} \Pr\left[|\overline{X}_n - p| > \varepsilon\right] = 0$$

for any real $\varepsilon > 0$; this law follows almost immediately from Chebyshev's inequality. The ***strong*** law of large numbers states that

$$\Pr\left[\lim_{n\to\infty} \overline{X}_n = p\right] = 1;$$

the proof of this law is considerably more involved.

## 11.4 Higher Moment Inequalities

Chebyshev's inequality and its corollaries are special cases of a more general family of so-called ***moment inequalities***, which consider the distribution of higher even powers of the random variable $X$ under stronger assumptions about the independence of the components $X_i$. Here I'll state the most general result without proof; the derivation follows the same general outline as the proof of Chebyshev's inequality above, but with more, you know, math.

**General Moment Inequality.** *For any fixed integer $k > 0$, if the variables $X_1, X_2, \ldots, X_n$ are $2k$-wise independent, then $\Pr[(X-\mu)^k \geq z] = O(\mu^k/z)$ for all $z > 0$. The hidden constant in the $O(\cdot)$ notation depends on $k$.*

This inequality immediately implies the following asymptotic tail bounds for all positive real numbers $\Delta$ and $\delta$; again, the hidden $O(\cdot)$ constants depend on the fixed independence parameter $k$.

$$\boxed{\begin{array}{ll} \Pr[X \geq \mu + \Delta] = O\left(\left(\frac{\mu}{\Delta^2}\right)^k\right) & \Pr[X \geq (1+\delta)\mu] = O\left(\left(\frac{1}{\delta^2\mu}\right)^k\right) \\ \Pr[X \leq \mu - \Delta] = O\left(\left(\frac{\mu}{\Delta^2}\right)^k\right) & \Pr[X \geq (1-\delta)\mu] = O\left(\left(\frac{1}{\delta^2\mu}\right)^k\right) \end{array}}$$

In short, the more independence we can assume about the indicator variables $X_i$, the higher the degree of the polynomial tails in the distribution of $X$, and the more likely that $X$ stays close to its expected value.

## 11.5 Chernoff Bounds

For *fully* independent random variables, even tighter tail bounds were developed in the early 1950s by Herman Chernoff, who was then a mathematics professor at the University of Illinois. However, in strict accordance with Stigler's Law, the first published version of "Chernoff bounds",

which appeared in a 1952 paper by Chernoff that gave the bounds their name, were actually due to his colleague Herman Rubin.[3]

**Exponential moment inequality.** *If the indicator variables $X_1, X_2, \ldots, X_n$ are fully independent, then $E[\alpha^X] \le e^{(\alpha-1)\mu}$ for any $\alpha \ge 1$.*

**Proof:** The definition of expectation immediately implies that

$$E[\alpha^{X_i}] \;=\; p_i\alpha^1 + (1-p_i)\alpha^0 \;=\; (\alpha-1)p_i + 1.$$

Thus, by The World's Most Useful Inequality $1 + t \le e^t$, we have

$$E[\alpha^{X_i}] \;\le\; e^{(\alpha-1)p_i}.$$

Full independence of the $X_i$'s now immediately implies

$$E[\alpha^X] \;=\; \prod_i E[\alpha^{X_i}] \;\le\; \prod_i e^{(\alpha-1)p_i} \;=\; e^{(\alpha-1)\mu}$$

and we're done. □

**Chernoff bound (upper tail).** *If the indicator variables $X_1, X_2, \ldots, X_n$ are fully independent, then $\Pr[X \ge x] \le e^{x-\mu}(\mu/x)^x$ for all $x \ge \mu$.*

**Proof:** Consider any fixed value $x \ge \mu$. The function $t \mapsto (x/\mu)^t$ is monotonically increasing, so

$$\Pr[X \ge x] = \Pr\left[\left(\frac{x}{\mu}\right)^X \ge \left(\frac{x}{\mu}\right)^x\right].$$

Markov's inequality implies that

$$\Pr\left[\left(\frac{x}{\mu}\right)^X \ge \left(\frac{x}{\mu}\right)^x\right] \;\le\; \frac{E[(x/\mu)^X]}{(x/\mu)^x}.$$

Finally, applying the exponential moment inequality at $\alpha = x/\mu$ gives us

$$E[(x/\mu)^X] \le e^{((x/\mu)-1)\mu} = e^{x-\mu},$$

which completes the proof. □

A nearly identical argument implies a similar bound on the lower tail:

**Chernoff bound (lower tail).** *If the indicator variables $X_1, X_2, \ldots, X_n$ are fully independent, then $\Pr[X \le x] \le e^{x-\mu}(\mu/x)^x$ for all $x \le \mu$.*

**Proof:** For any $x \le \mu$, we immediately have $\Pr[X \le x] = \Pr[(x/\mu)^X \ge (x/\mu)^x]$. (The direction of the inequality changes because $x/\mu \le 1$.) The remainder of the proof is unchanged. □

[3] "Since that seemed to be a minor lemma in the ensuing paper I published (Chernoff, 1952), I neglected to give [Rubin] credit. I now consider it a serious error in judgment, especially because his result is stronger, for the upper bound, than the asymptotic result I had derived."

The particular value $x/\mu$ in these results may seem arbitrary, but in fact it's chosen vary carefully. The same arguments imply that

$$\Pr[X \geq x] \leq \frac{e^{(\alpha-1)\mu}}{\alpha^x} \text{ for } \textit{all } \alpha > 1 \qquad \text{and} \qquad \Pr[X \leq x] \leq \frac{e^{(\alpha-1)\mu}}{\alpha^x} \text{ for } \textit{all } \alpha < 1.$$

A bit of calculus implies that the right sides of these inequalities are minimized when $\alpha = x/\mu$.

Direct substitution now implies the following more traditional forms of Chernoff bounds, for any positive reals $\Delta$ and $\delta$. Unlike the polynomial moment bounds we derived earlier, Chernoff bounds for the upper and lower tails of $X$ are asymmetric.

$$\boxed{\begin{array}{ll} \Pr[X \geq \mu + \Delta] \leq e^{\Delta}\left(\frac{\mu}{\mu+\Delta}\right)^{\mu+\Delta} & \Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \\ \Pr[X \leq \mu - \Delta] \leq e^{-\Delta}\left(\frac{\mu}{\mu-\Delta}\right)^{\mu-\Delta} & \Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \end{array}}$$

## 11.6 Back to Treaps

In our analysis of randomized treaps, we wrote $\boldsymbol{i \uparrow k}$ 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

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

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, and 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 $x_2, x_3, \ldots, x_n$. We prove the lemma by induction on $n$, with the vacuous base case $n = 1$. The definition of conditional probability gives us

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

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,

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

The inductive hypothesis implies that the variables $X_2, \ldots, X_{n-1}$ 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 deliberately 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 $n$. Lemma 1 allows us to partition these variables into two mutually independent subsets. We define

$$\mathit{depth}_<(k) := \sum_{i<k}[i \uparrow k] \qquad \text{and} \qquad \mathit{depth}_>(k) := \sum_{i>k}[i \uparrow k],$$

so that $\mathit{depth}(k) = \mathit{depth}_<(k) + \mathit{depth}_>(k)$. If $\mathit{depth}(k) > 8\ln n$, then either $\mathit{depth}_<(k) > 4\ln n$ or $\mathit{depth}_>(k) > 4\ln n$.

To bound the probability that $\mathit{depth}_<(k) > 4\ln n$, we apply Chernoff's inequality

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

with $\mu = \mathrm{E}[\mathit{depth}_<(k)] = H_k - 1 < \ln n$ and $\delta = 3$, as follows.

$$\begin{aligned}
\Pr[\mathit{depth}_<(k) > 4\ln n] &< \Pr[\mathit{depth}_<(k) > 4\mu] \\
&< \left(\frac{e^3}{4^4}\right)^\mu \\
&< \left(\frac{e^3}{4^4}\right)^{\ln n} = n^{\ln(e^3/4^4)} = n^{3-4\ln 4} < \frac{1}{n^2}.
\end{aligned}$$

(The last step uses the fact that $4\ln 4 \approx 5.54518 > 5$.) The same analysis implies that $\Pr[\mathit{depth}_>(k) > 4\ln n] < 1/n^2$. These inequalities imply the crude bound $\Pr[\mathit{depth}(k) > 8\ln n] < 2/n^2$.

Now consider the probability that the treap has maximum depth greater than $8\ln n$. Even though the distributions of different nodes' depths are *not* independent, we can conservatively bound the probability of failure using the union bound, as follows:

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

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. □

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.