Randomized Algorithms

Indicators, linearity of expectation, and the union bound

Alex Gittens
Department of Computer Science, Rensselaer Polytechnic Institute

Fall 2026

Analyzing Quicksort's random work

Last time we saw that randomized Quicksort is always correct, but its running time depends on its random pivots.

Trying to analyze the entire random recursion tree directly is awkward:

- every pivot changes the sizes of all later subproblems;
- the recursion tree is different on every run;
- the comparison count is a complicated function of all those choices.

Instead of tracking the whole execution, we will ask a much smaller question:

> For a fixed pair of keys, are they ever compared?

That change of viewpoint leads naturally to indicator random variables and linearity of expectation.

PART 1

Counting by indicators

1

An event can be turned into a random variable

For an event E, define its indicator

$$\mathbf{1}_E = \begin{cases} 1, & E \text{ occurs,} \\ 0, & E \text{ does not occur.} \end{cases}$$

Then

$$\mathbb{E}[\,\mathbf{1}_E] = 1 \cdot \Pr(E) + 0 \cdot \Pr(E^c) = \Pr(E).$$

This simple identity lets us turn a counting problem into an expectation calculation. If

$$X = \sum_i \mathbf{1}_{E_i},$$

then X counts how many of the events E_i occur.

Warm-up: how many people get their own hat?

n people check their hats, and the hats are returned according to a uniformly random permutation.
For person i, let

$$X_i = \mathbf{1}_{\{\text{person } i \text{ receives their own hat}\}}.$$

Then

$$\Pr(X_i = 1) = \frac{1}{n}, \qquad \mathbb{E}[X_i] = \frac{1}{n}.$$

The total number of people who receive their own hats is

$$X = \sum_{i=1}^{n} X_i.$$

Question

The X_i are dependent. Do we need to understand that dependence in order to compute $\mathbb{E}[X]$?

Linearity without independence

For any random variables $X_1, \ldots, X_m$ with finite expectations,

$$\boxed{\mathbb{E}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mathbb{E}[X_i]}$$

whether or not the variables are independent.
For the hat-check problem,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{n} = 1.$$

Indeed the indicators are dependent: if person 1 gets their own hat, then

$$\Pr(X_2 = 1 \mid X_1 = 1) = \frac{1}{n-1} \neq \frac{1}{n}.$$

But none of that matters for the expectation of the sum.

Why linearity helps

A complicated random quantity is often hard to analyze as a whole. But if we can write it as

$$X = \sum_i X_i$$

where each X_i represents one simple contribution, then

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i].$$

For indicator variables this becomes

$$\boxed{X = \sum_i \mathbf{1}_{E_i} \quad \Longrightarrow \quad \mathbb{E}[X] = \sum_i \Pr(E_i).}$$

So the analysis problem becomes:

> *Identify the elementary events being counted, and compute the probability of each one.*

We now apply exactly this idea to randomized Quicksort.

PART 2

Randomized Quicksort

2

Count comparisons pair by pair

Label the keys by their final sorted ranks $1, \ldots, n$. Assume all keys are distinct.
For each pair $i < j$, define

$$X_{ij} = \mathbf{1}_{\{\text{ranks } i \text{ and } j \text{ are ever compared}\}}.$$

A pair of keys can be compared at most once, so the total number of comparisons is exactly

$$C = \sum_{1 \le i < j \le n} X_{ij}.$$

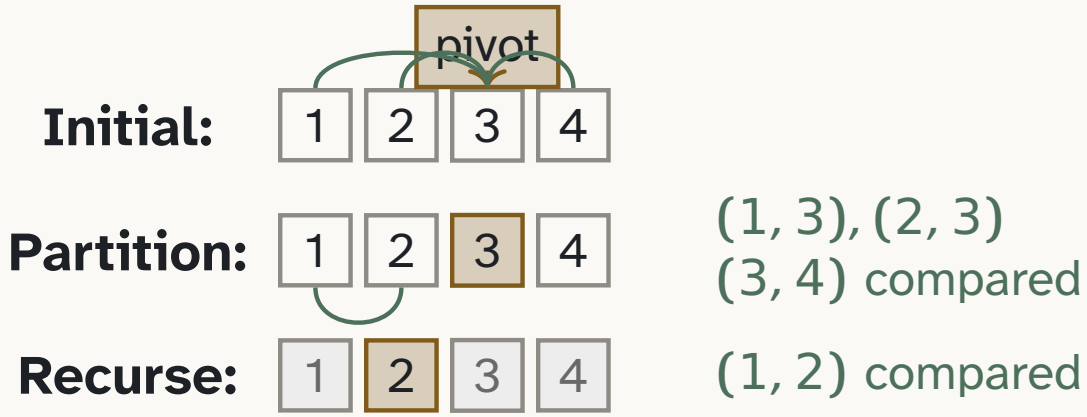
Therefore

$$\mathbb{E}[C] = \sum_{1 \le i < j \le n} \Pr(X_{ij} = 1).$$

The entire analysis has now been reduced to one question:

$$\boxed{\Pr(\text{ranks } i \text{ and } j \text{ are compared}) = ?}$$

Trace one execution before generalizing



Comparison summary

Compared:
$(1, 2), (1, 3), (2, 3), (3, 4)$
Never compared:
$(1, 4), (2, 4)$

When are ranks i and j compared?

Focus on the ranks

$$i, i+1, \ldots, j.$$

As long as none of these elements has been chosen as a pivot, ranks i and j remain together in the same recursive subproblem.
Now consider the *first* pivot chosen from this interval.

- If it is an interior rank p with $i < p < j$, then p separates i and j into different recursive subproblems. They can never be compared afterward.
- If it is i or j, then that pivot is compared with the other endpoint during the partition step.

Thus

$$\boxed{X_{ij} = 1 \iff \text{the first pivot from } \{i, \ldots, j\} \text{ is } i \text{ or } j.}$$

Why each rank is equally likely to be first

Let $S = \{i, i+1, \ldots, j\}$, and for each $k \in S$ let A_k be the event that k is the first pivot chosen from S.
By relabeling symmetry inside S, all the events A_k have the same probability. Since exactly one of them occurs,

$$\sum_{k \in S} \Pr(A_k) = 1.$$

Therefore each $k \in S$ has probability $1/(j-i+1)$ of being first, and

$$\Pr(X_{ij} = 1) = \Pr(A_i \cup A_j) = \frac{2}{j-i+1}.$$

The first pivot in an interval is uniform

There are $j - i + 1$ ranks in

$$\{i, i+1, \ldots, j\}.$$

Therefore

$$\Pr(X_{ij} = 1) = \frac{2}{j - i + 1}.$$

This probability depends only on the rank distance $j - i$.

Now sum the pairwise probabilities

By linearity of expectation,

$$\mathbb{E}[C] = \sum_{1 \le i < j \le n} \frac{2}{j - i + 1}.$$

Group pairs by their distance

$$d = j - i.$$

For a fixed distance d, there are $n - d$ such pairs. Hence

$$\mathbb{E}[C] = 2 \sum_{d=1}^{n-1} \frac{n - d}{d + 1}.$$

Finish the harmonic-number algebra

Reindexing with $k = d + 1$,

$$\begin{aligned}\mathbb{E}[C] &= 2\sum_{k=2}^{n}\frac{n-k+1}{k} \\ &= 2(n+1)\sum_{k=2}^{n}\frac{1}{k} - 2(n-1) \\ &= 2(n+1)H_n - 4n.\end{aligned}$$

Expected comparisons are logarithmic

Since

$$H_n = \log n + O(1),$$

we obtain

$$\boxed{\mathbb{E}[C] = 2n \log n + O(n) = O(n \log n).}$$

A few things are worth noticing:

- We never analyzed the random recursion-tree shape.
- The indicators X_{ij} are highly dependent; linearity did not care.
- The proof works for every fixed input with distinct keys because only the ranks matter.

The useful decomposition was not “work per recursive call.” It was “one possible comparison per pair of ranks.”

PART 3

Collision events and bounds

3

Collision events and the union bound

Hash k items independently and uniformly into n bins.
For each pair $i < j$, define

$$E_{ij} = \{\text{items } i \text{ and } j \text{ land in the same bin}\}.$$

For one pair,

$$\Pr(E_{ij}) = \frac{1}{n}.$$

The event that *some* collision occurs is

$$\bigcup_{i<j} E_{ij}.$$

Here we do not want the expected number of colliding pairs. We want the probability that at least one bad event occurs.

The union bound

For arbitrary events $E_1, \ldots, E_m$,

$$\boxed{\Pr\left(\bigcup_{r=1}^{m} E_r\right) \leq \sum_{r=1}^{m} \Pr(E_r)}$$

with no independence assumption.
Applying this to pair collisions,

$$\Pr(\text{some collision}) \leq \sum_{i<j} \Pr(E_{ij}) = \binom{k}{2}\frac{1}{n} \leq \frac{k^2}{2n}.$$

So if $k = o(\sqrt{n})$, then $\Pr(\text{collision}) \to 0$.

Consequences of the union bound

The union bound proves a *safe regime*:

$$k \ll \sqrt{n} \quad \Longrightarrow \quad \text{collisions are unlikely.}$$

But when k reaches order $\sqrt{n}$, the upper bound becomes order one. That does *not* prove collisions are likely. It only says this upper bound has stopped being small.

So a natural question remains:

$$\boxed{\text{Is } \sqrt{n} \text{ really the collision threshold?}}$$

To answer the other direction, analyze the complementary event: no collision at all.

Exact no-collision probability

- The first item can go anywhere.
- The second must avoid one occupied bin: probability $1 - 1/n$.
- The third must avoid two occupied bins: probability $1 - 2/n$.
- Continue in the same way.

Therefore

$$\Pr(\text{no collision}) = \prod_{j=0}^{k-1} \left(1 - \frac{j}{n}\right).$$

Using $1 - x \le e^{-x}$,

$$\begin{aligned}\Pr(\text{no collision}) &\le \exp\left(-\sum_{j=0}^{k-1} \frac{j}{n}\right) \\ &= \exp\left(-\frac{k(k-1)}{2n}\right).\end{aligned}$$

The collision threshold

The two bounds now give complementary information.
If

$$k \ll \sqrt{n},$$

then the union bound gives

$$\Pr(\text{collision}) \to 0.$$

If

$$k \gg \sqrt{n},$$

then

$$\Pr(\text{no collision}) \leq \exp\left(-\frac{k(k-1)}{2n}\right) \to 0,$$

so

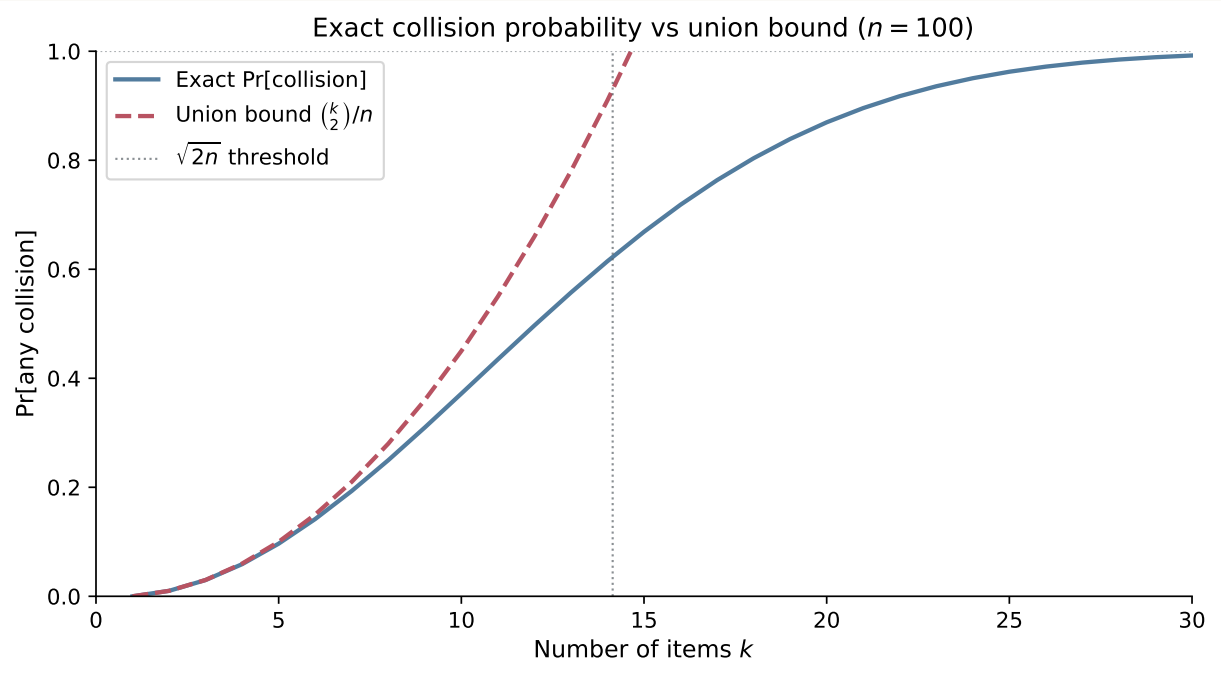
$$\Pr(\text{collision}) \to 1.$$

Thus

$$\boxed{k = \Theta(\sqrt{n})}$$

is the transition scale between the sparse regime $k \ll \sqrt{n}$, where collisions are unlikely, and the dense regime $k \gg \sqrt{n}$, where collisions are likely.

Bounds versus exact probability



For $k = c\sqrt{n}$, the no-collision probability approaches approximately

$$e^{-c^2/2},$$

so the probability of at least one collision changes from near 0 to near 1 on precisely this scale.

The union bound finds the correct scale without needing the exact formula.

Two decompositions

Today we used the same elementary events in two different ways.

To count how many events occur on average:

$$X = \sum_i \mathbf{1}_{E_i} \quad \Longrightarrow \quad \boxed{\mathbb{E}[X] = \sum_i \Pr(E_i)}.$$

To bound the probability that at least one event occurs:

$$E = \bigcup_i E_i \quad \Longrightarrow \quad \boxed{\Pr(E) \le \sum_i \Pr(E_i)}.$$

Neither identity requires independence.
The modeling decision comes first: are we asking for an *expected count*, or for the probability of *any bad event*?

Expectation is not enough

For randomized Quicksort we proved

$$\mathbb{E}[C] = O(n \log n).$$

But an expectation alone does not answer questions such as:

- How often is the runtime much larger than its mean?
- Can we say that a good runtime occurs with high probability?
- What additional information about a random variable gives stronger control of deviations?

These questions lead from expectation to tail bounds.
Next we begin with the weakest possible information—the mean—and derive Markov's inequality, then ask what variance lets us improve.

Questions?