Bogo Selection Sort (Piazza)

Abstract

A student in UMD's CMSC132 class decided to make a variant of a BOGO sort that has selection-sort-like properties and asked on our online forum what the runtime might be. He suggested a *median* runtime of $O(n^2 n!)$, so I decided to try and find out what it really is.

```java
public static ArrayList<Integer> bogoSelectionSort(ArrayList<Integer> list) {
    ArrayList<Integer> ans = new ArrayList<>();
    int size = list.size();

    for (int i = 0; i < size; i++) {
        int min = list.get(0);

        for (int j = 0; j < list.size(); j++) {
            if (list.get(j) < min)
                min = list.get(j);
        }
        if (min == list.get(0)) {
            ans.add(min);
            list.remove(0);
        } else {
            Collections.shuffle(list);
            i--;
        }
    }
    return ans;
}
```

1 Best-Case Analysis

Well, obviously the best-case is when we *very, very* luckily always get the minimum to be the first element. Approximately, the outer `for` loop will run n times (if $n =$ `list.size()`), the inner `for` loop will run in time proportional to n, `list.remove(0)` will run in time proportional to n, and `Collections.shuffle(list)` will run in time proportional to n. Overall, best-case is $O(n^2)$.

2 Worst-Case Analysis

We might run forever. $O(\infty)$.

3 Average-Case Analysis

To model the outer `for`-loop, we can describe it as a linear number of operations of the body, conditioned on the index `i`, since the work done in the body is only dependent on the size of the list:

$$T(n) = \sum_{i=0}^{n-1} \text{Inner}(i)$$

For average-case analysis, we care about the *expectation* of our runtime:

$$\begin{aligned} \mathbb{E}\big[T(n)\big] &= \mathbb{E}\left[\sum_{i=0}^{n-1} \text{Inner}(i)\right] & (1) \\ &= \sum_{i=0}^{n-1} \mathbb{E}\big[\text{Inner}(i)\big] \qquad \text{(linearity of expectation)} & (2) \end{aligned}$$

3.1 Expectation of the Inner$(\cdot)$ Function

First, we'll find a closed (closed-ish) form for the inner function, given a particular iteration i. Then, we can find its expectation.

First note that so long as we fail to get the minimum of the list, we do the same amount of work since the size of the list doesn't change. Once we succeed, the size of the list changes and we move on to the next iteration $i+1$. So, our function Inner$(i; k, n)$ (parameterized by k failures and original list size n) has the form

$$\text{Inner}(i, k; n) = k \cdot \text{work on failure} + \text{work on success} \tag{3}$$

Now, for each failure (and the one success), how much work do we do? Regardless of success or failure, we look for the minimum of the remaining elements in the list (taking time $\approx n - i$). If we fail, we shuffle the list, which takes time $\approx n - i$. If we succeed, we add the minimum to the result `ArrayList` and remove it from the *front* of the list, which also takes time $\approx n - i$. So in both cases, we do $\approx 2(n - i)$ work. Our inner function is

$$\begin{aligned} \text{Inner}(i, k; n) &= k \cdot 2(n-i) + 2(n-i) \\ &= 2(n-i)(k+1) \end{aligned}$$

Now, we can start to get the expectation of the Inner$(\cdot)$ function:

$$\begin{aligned} \mathbb{E}\big[\text{Inner}(i, k; n)\big] &= \mathbb{E}\big[2(n-i)(k+1)\big] \\ &= 2(n-i)\mathbb{E}\big[k\big] + 2(n-i) \qquad \text{(linearity of expectation)} \end{aligned}$$

For a given iteration variable i, how many failures can we expect to have before we get a success (minimum is the first element)? How many times we iterate depends on if we happen to get the minimum in the lowest position or not. In terms of success/failure, we want to know how many times we fail to get the minimum before we get the first minimum; from probability, this is modeled using a geometric distribution:

$$P(k;p) = p(1-p)^k$$

with an expected value

$$\mathbb{E}\big[k\big] = \frac{1-p}{p} \tag{4}$$

On iteration i, we have $n-i$ elements in our list, so we have a $\dfrac{1}{n-i}$ probability of getting the minimum as the first element (given a properly implemented shuffle). So specifically, substituting into the equation (4)

$$\begin{aligned}
\mathbb{E}\big[k\big] &= \frac{1-p}{p}\bigg|_{p=\frac{1}{n-i}} \\
&= \frac{1-\dfrac{1}{n-i}}{\dfrac{1}{n-i}} \\
&= n-i-1
\end{aligned}$$

we need an average of $n-i-1$ *failures* before getting our first success. Substituting this result into the expectation of our Inner$(\cdot)$ function,

$$\begin{aligned}
\mathbb{E}\big[\text{Inner}(i;n)\big] &= 2(n-i)(n-i-1) + 2(n-i) \\
&= 2(n-i)^2
\end{aligned}$$

3.2 Putting It All Together

Coming back to our overall average runtime,

$$\begin{aligned}
\mathbb{E}\big[T(n)\big] &= \sum_{i=0}^{n-1} \mathbb{E}\big[\text{Inner}(i;n)\big] \\
&= 2\sum_{i=0}^{n-1}(n-i)^2 \\
&= 2\sum_{i=1}^{n} i^2 && \text{(change of variables)} \\
&= \frac{n(n+1)(2n+1)}{3} && \text{(sum of squares formula)} \qquad \in O(n^3)
\end{aligned}$$

4 Median Analysis

Suppose we used the median instead of the average/mean/expectation. We'd like to just reuse the analysis from above but just substitute the median of the geometric distribution: $\left\lceil \dfrac{-1}{\lg(1-p)} \right\rceil - 1$.

However, there are a few issues that we will need to resolve:

1. In general, there is no linearity of medians.
2. There's no consistent notation for medians, so we'll use $\mu_{1/2}\big[\,\cdot\,\big]$ as a sort of median operator.
3. Medians not continuous in the same way that expectations are. (See that the median of the geometric distribution involves a ceiling, which introduces steps/non-linearities.) We'll need to bound the expression.

4.1 Linearity of Medians (Special Case)

Although there is no linearity of medians in general, we can show that linearity does hold for the geometric distribution. Specifically: monotonically decreasing probability distributions are closed under (positive) scalar multiplication and addition.

1. Under scalar multiplication, $\mu_{1/2}\big[aX\big] = a\mu_{1/2}\big[X\big], \forall a > 0$. This makes sense since the order of the elements doesn't change if we scale all numbers up or down by the same amount.
2. Under addition, adding Y to X doesn't change the order of X (and vice versa). If X_i ¡= X_j, then when we add Y_i and Y_j, where $Y_i <= Y_j$, it can't be the case that X_i was made to be bigger than $X_j + Y_j$, since even if $X_i = X_j$, we can only add at most $Y_i <= Y_j$.

4.2 Median of the Inner($\cdot$) Function

Now we're ready to apply the results from above to our Inner($\cdot$) function. We can start with the same Inner($\cdot$) function from our average case analysis:

$$
\begin{aligned}
\text{Inner}(i, k; n) &= 2(n-i)(k+1) \\
\mu_{1/2}\big[\text{Inner}(i, k; n)\big] &= \mu_{1/2}\big[2(n-i)(k+1)\big] \\
&= 2(n-i)\mu_{1/2}\big[k+1\big] \qquad \text{(applying linearity)} \\
&= 2(n-i)\Big(\mu_{1/2}\big[k\big]+1\Big) \qquad \text{(applying linearity again)} \\
&= 2(n-i)\left\lceil \frac{-1}{\lg(1-p)} \right\rceil \qquad \text{(median of the geometric distribution)} \\
&< 2(n-i)\left(\frac{-1}{\lg(1-p)}+1\right) \qquad \text{(upper bound on ceil}(\cdot)) \\
&= 2(n-i)\left(\frac{-1}{\lg\left(1-\frac{1}{n-i}\right)}+1\right)
\end{aligned}
$$

However, note that this function is ill-defined for $i = n-1$ since we get $\lg(0)$. So we need to evaluate what happens when $i = n-1$ separately: this occurs when there are $n-(n-1) = 1$ element left in our list. Since that one element is guaranteed to be the minimum, we do $2(n-(n-1)) = 2$ amount of work. To be precise, the median of $\mathrm{Inner}(i;n)$ is bounded

$$\mu_{1/2}\big[\mathrm{Inner}(i;n)\big] < \begin{cases} 2(n-i)\left(\dfrac{-1}{\lg\left(1-\dfrac{1}{n-i}\right)}+1\right), & i < n-1 \\ 2, & i = n-1 \end{cases} \tag{5}$$

4.3 Putting It All Together (Again)

Coming back to our original problem: given our definition of $T(n)$, we want to find its median:

$$\begin{aligned} \mu_{1/2}\big[T(n)\big] &= \mu_{1/2}\left[\sum_{i=0}^{n-1}\mathrm{Inner}(i;n)\right] && (6)\\ &= \sum_{i=0}^{n-1}\mu_{1/2}\big[\mathrm{Inner}(i;n)\big] \qquad \text{(monotonic linearity of medians)} && (7) \end{aligned}$$

Now, we'll try to simplify as far as we can:

$$\begin{aligned} \mu_{1/2}\big[T(n)\big] &= 2+\sum_{i=0}^{n-2}\mu_{1/2}\big[\mathrm{Inner}(i;n)\big] \\ &< 2+2\sum_{i=0}^{n-2}(n-i)\left(\frac{-1}{\lg\left(1-\dfrac{1}{n-i}\right)}+1\right) \\ &= 2+2\sum_{i=2}^{n}i\left(\frac{-1}{\lg\left(1-\dfrac{1}{i}\right)}+1\right) \qquad \text{(change of variables)} \\ &= 2+2\sum_{i=2}^{n}i+2\sum_{i=2}^{n}\frac{-1}{\lg\left(1-\dfrac{1}{i}\right)} \\ &= n(n+1)+2\sum_{i=2}^{n}\frac{-i}{\lg\left(1-\dfrac{1}{i}\right)} \qquad \text{(Gauss's sum)} \end{aligned}$$

There's not an easy way to get a function like $\dfrac{-i}{\lg\left(1-\dfrac{1}{i}\right)}$ to an explicit form, so we'll try bounding it with a polynomial. If we guess i^2,

$$\begin{aligned}
\lim_{i\to\infty} \frac{i^2}{\dfrac{-i}{\lg\left(1-\dfrac{1}{i}\right)}} &= \lim_{i\to\infty}\left(-i\lg\left(1-\frac{1}{i}\right)\right) \\
&= \lim_{i\to\infty} \frac{-\lg\left(1-\dfrac{1}{i}\right)}{1/i} \\
&= \lim_{i\to\infty} \frac{-\left(1-\dfrac{1}{i}\right)^{-1}/i^2}{-1/i^2} && \text{(l'Hôpital's Rule)} \\
&= \lim_{i\to\infty} \frac{1}{1-\dfrac{1}{i}} \\
&= 1
\end{aligned}$$

Since 1 is finite and $\neq 0$, our "bad" function from above: $\dfrac{-i}{\lg\left(1-\dfrac{1}{i}\right)}$ can be bounded by a constant multiple of i^2, so long as we choose a big enough multiple. It turns out though, that i^2 is already big enough: assuming that $i > 0$, if we try to find the intersection points between the 2 functions:

$$\begin{aligned}
i^2 &= \frac{-i}{\lg\left(1-\dfrac{1}{i}\right)}, \forall i > 0 \\
\frac{-1}{i} &= \lg\left(1-\frac{1}{i}\right) \\
x &= \lg(1+x), \forall x < 0 && \text{(letting } x = \tfrac{-1}{i}\text{)} \\
2^x &= x+1, \forall x < 0
\end{aligned}$$

For negative x, there are no intersection points between 2^x and $x+1$. Going back to our median,

$$\begin{aligned}
\mu_{1/2}\big[T(n)\big] &= n(n+1) + 2\sum_{i=2}^{n} \frac{-i}{\lg\left(1-\dfrac{1}{i}\right)} \\
&< n(n+1) + 2\sum_{i=2}^{n} i^2 \\
&= n(n+1) + \frac{n(n+1)(2n+1)}{3} && \text{(sum of squares formula)} \\
&= \frac{2}{3}n^3 + 2n^2 + \frac{4}{3}n \\
&\in O(n^3)
\end{aligned}$$