Occupancy Problems

Markov's Inequality

Sometimes, it can be hard to say something about the probability that a random variable deviates far from its expectation. One way to avoid making a detailed analysis is by using Markov's inequality. It says that for a random non-negative variable Y, then:

$$P[Y \geq t] \leq \frac{E[Y]}{t}$$

Unfortunately, this is not particularly tight and often not useful. However, if we know the variance of a distribution, we can use Markov's inequality to derive a more useful bound known as Chebyshev's inequality.

Chebyshev's Inequality

The Chebyshev inequality states that, for a random variable X, standard deviation σ and expectation μ:

$$P[|X - \mu| \leq t\sigma] \leq \frac{1}{t^2}$$

We are gonna use this inequality in the analysis of a randomized selection algorithm.

Randomized Selection

A selection algorithm is an algorithm to find the k'th smallest element. The LazySelect algorithm is one such algorithm which uses random sampling. The algorithm:

- So given a k and a set S of size n, we pick $n^{3/4}$ uniformly at random and call that set for R.
- Sort R in $O(n^{3/4} \lg n)$ time.
- We let $x = kn^{-1/4}$, $\ell = \max\{\lfloor x - \sqrt{n} \rfloor, 1\}$ and $h = \min\{\lceil x + \sqrt{n} \rceil, n^{3/4}\}$. Let $a = R_\ell$ and $b = R_h$ and determine their rank by comparing to all elements in S.
- If $k < n^{1/4}$, then P is the set of $y \leq b$ from S. If $k > n^{3/4}$, then P is the set of $y \geq a$ from S. If in between, then P is the set of $a \leq y \leq b$.
- Check if S_k is in P and if $P \leq 4n^{3/4} + 2$. If it is, sort P in $O(P \lg P)$ time and find S_k in P_{k-r_a+1}. Otherwise, try the entire thing again.

So the idea is to have S_k in the set P, which is rather small, so we don't lose too much when sorting the set in the last step.

The claim is that with probability $1 - O(n^{-1/4})$, we find S_k in the first try, yielding $2n + o(n)$ comparisons (Theorem 3.5).

Proof: The number of comparisons is to see. We get $2n$ comparisons determining ranks for a and b. The other steps only perform $o(n)$ comparisons. The algorithm can fail if S_k lands outside the set P. Lets analyze the probability that $S_k < a$. This happens only if fewer than ℓ of the elements in R are $\leq S_k$. Let X_i be an indicator variable for this, then we see that $P[X_i] = k/n$. These variables are actually *Bernouille trials*, which means we can find μ and σ:

$$\mu = E\left[\sum_{i=1}^{n^{3/4}} X_i\right] = \sum_{i=1}^{n^{3/4}} E[X_i] = \frac{kn^{3/4}}{n} = kn^{-1/4}$$

and

$$\sigma^2 = n^{3/4}\left(\frac{k}{n}\right)\left(1 - \frac{k}{n}\right) \leq \frac{n^{3/4}}{4}$$

Or to bound the standard deviation, we have $\sigma \leq n^{3/8}/2$. We can now use Chebyshev's inequality to get:

$$P[|X-\mu| \geq \sqrt{n}] \leq P[|X-\mu| \geq 2n^{1/8}\sigma] = O(n^{-1/4})$$

Following the same argument, this is also the probability that $b < S_k$, meaning the probability it falls outside $[a, b]$ is $O(n^{-1/4})$.
The other way it can fail is if $|P| > 4n^{3/4} + 2$. We note that $|P| = r_b - r_a + 1$, so for the test to fail, we must have $k - r_a > 2n^{3/4} + 1$ or $r_b - k > 2n^{3/4} + 1$. The proof then follows from the above analysis, but where X_i is 1 if the i'th sample has rank less than $k - 2n^{3/4}$ or above $k + 2n^{3/4}$. Once again, we get $O(n^{-1/4}$, which proves the result from Theorem 3.5.

The best known deterministic algorithm uses $3n$ comparisons. The expected running time is also $2n+o(n)$.

Two-Point Sampling

Coupon Collector's Problem

Talk about inequalities → Randomized selection → Prove Theorem 3.5