Reservoir Sampling

Craig Scott

July 19, 2019

1 Introduction

Suppose there exists a large set T_i with elements $\{t_0, t_1, ...t_n\}$. If we wanted to develop an algorithm to randomly choose an element within that set, then we would have to count the set where length(T) = n. Suppose that we choose a random number m s.t. $0 \leq m \leq n$. We would then have to iterate over that set to arrive at the index T_m. The computation time would be of the order O(nm), with worst case O(n^2) and best case O(n).

This is feasible approach, but when the size of the set approaches magnitudes of billions, then a more efficient approach must be taken.

2 Paraphrased Implementation

With the knowledge that the probability of choosing any random single item out of the set is $\frac{1}{i}$ where i is the i^{th} step in the iteration. Generate a random number between [0,i). If rand == 0, ie. with the probability being $\frac{1}{i}$, then the storage is replaced with t_i. There is, and most likely will be multiple values replaced within the storage buffer. The result will give a number randomly chosen with a probability $\frac{1}{n}$.

3 Proof : Single Element

Suppose there exists a set T_i with elements $\{t_0, t_1, ...t_n\}$. Suppose the probability for choosing a random value in the set T_i is $\frac{1}{n}$. This means that the first element chosen has the probability of being chosen $\frac{1}{i}$, where i is the i^{th} step in the iteration. Since it is the first iteration, the probability of being chosen is 1. This should have a time complexity of O(n) and space complexity of O(1).

We conject that the probability of an item being chosen at the i^{th} step is $\frac{1}{i}$. Then what is the probability of the $i + 1$ element being accepted? This should follow P(i accepted) $\cdot$ P(i not replaced by i+1)

$$P(i + 1) = \frac{1}{i} \cdot (1 - \frac{1}{1 + i}) \tag{1}$$

after some elementary algebra, this is just equal to $\frac{1}{i+1}$. After n iterations, what is the probability of the final element to be chosen? Expanding this out we obtain a telescoping series-product

$$P(n) = \frac{1}{i}(1 - \frac{1}{i+1})(1 - \frac{1}{i+2})(1 - \frac{1}{i+3})...(1 - \frac{1}{n}) = \frac{1}{n} \tag{2}$$