# 1 Reservoir Sampling

Reservoir Sampling is an algorithm to sample from a potentially infinite stream $\Sigma$. Each element in the sample has the same uniform probability of being picked and the sample has size $M$ always. Getting samples from infinite (or very large) streams has several applications. For example, sampling websites from the net, sampling metro passengers, sampling visitors and many, many more. The context in which i learned about this algorithm was related to sampling edges from a graph with many, many edges. Many more than fit into RAM. To estimate the number of triangles (or other motifs) in the graph, one needs to look at the edges. If one can not look at all of them, at least some are required. Each edge must be samples with equal probabilities. The algorithm goes as follows

1. Initialize

   $t \leftarrow 0$

   $S \leftarrow \emptyset$

For each $x_t \in \Sigma$:

$t \leftarrow t + 1$

If $t \leq M$:

- $S \leftarrow S \cup x_t$
- Else:
  - if $SampleElement(\frac{M}{t})$:
    - pick $x_i$ from $S$ uniformly at random (prob. $\frac{1}{M}$
    - $S \leftarrow S \backslash \{x_i\}$
    - $S \leftarrow S \cup \{x\}$

Proposition:

i) $Pr(x_i \in S) = \frac{M}{t}$ for all $i$ and $t$

By induction:

$$t = M \Rightarrow p(x \in S) = 1$$

Now $t \geq M$

$$Pr(x_i \in S) = Pr(x_i \in S \text{ and } A) + Pr(x_i \in S \text{ and } \bar{A})$$

and

$$Pr(x_i \in S \text{ and } A) = Pr(x_i \in S|A)Pr(A)$$

$$Pr(x_i \in S \text{ and } \bar{A}) = Pr(x_i \in S|A)Pr(\bar{A})$$

Note $Pr(A) = \frac{M}{t}$ is the probability that the element $x_i$ is removed and $x_t$ is added to the samples. $Pr(\bar{A}) = 1 - Pr(A)$.

$$Pr(x_i \in S|A) = \frac{M}{t-1}(1 - \frac{1}{M}),$$

where $\frac{M}{t-1}$ by induction and $1 - \frac{1}{M}$ is the probability that $x_i$ was not removed due to event $A$.

$$Pr(x_i \in S|\bar{A}) = \frac{M}{t-1}.$$

Thus

$$Pr(x_i \in S) = \frac{M}{t-1}(1 - \frac{1}{M})\frac{M}{t} + \frac{M}{t-1}(1 - \frac{M}{t}) = \frac{M}{t}.$$

It is trivial that $|S| = M$ for all $t$.