Streaming Random Selection Using the Attenuated Geometric Distribution

Christopher Meek
cameek@uw.edu
University of Washington

Carl Kadie
carlk@msn.com

August 20, 2022

Abstract

The problem of selecting random items from an enumeration is an important and well-studied problem. We define a distribution that arises naturally when considering the problem of selecting a single item at random from an enumeration of items. Due to its similarity with the geometric distribution, we call this distribution the attenuated geometric distribution. The primary difference between the geometric and attenuated geometric distributions is that the bias of the coin is fixed in the case of the geometric distribution and gradually attenuated in the attenuated geometric distribution. We explore properties of the attenuated geometric distribution and give a sampling procedure. Using this sampling procedure, we provide an algorithm for selecting a random items from an enumeration that, as compared to previous published approaches, requires half the number of calls to a random number generator.

1 Introduction

We define a distribution that arises naturally when considering the problem of selecting an item at random from an enumeration of items. The naive streaming algorithm is to select the first item in the enumeration and then to successively replace the selected item with enumerated items with probability $1/n$ for the n^{th} item. This naive streaming algorithm will use $N-1$ calls to a random number generator if there are N items in the enumeration. We can reduce the number of calls required by defining a sampling procedure of a distribution of offsets given a currently selected item. Defining a distribution in this way yields a distribution similar to the geometric distribution. In particular, we can define a distribution over offsets in which the probability of selecting an offset $o = 1$ given the currently selected line is i is $1/(1+i)$ and for $o > 1$ the probability is $p(o|i) = \frac{1}{i+o}\prod_{j=1}^{o-1}(1-\frac{1}{i+j})$. This formula is similar to the formula for the geometric distribution which is typically defined to be the probability of obtaining the first head after n tosses of a biased coin. The primary difference in the case of the attenuated geometric distribution is that the bias of heads on the coin is gradually attenuated. Unlike the naive streaming algorithm, if we can sample from this distribution, we will require fewer calls to a random number generator. Motivated by this scenario, we define a family of distributions that we call *attenuated geometric distribution* and explore properties of the distribution, give a sampling procedure and discuss related work.

2 The Attenuated Geometric Distribution

The attenuated geometric distribution[1] is a distribution over natural numbers with a single scalar parameter $\alpha > 0$. The streaming random selection algorithm described in introduction provides a natural recursive definition for the slightly more general case in which we have a scalar *attenuation parameter*. In particular, the probability of $n = 1$ with attenuation α is $p(n = 1|\alpha) = \frac{1}{1+\alpha}$. For $n > 1$, the natural recurrence is $p(n+1|\alpha) = p(n|\alpha)\frac{n+\alpha}{n+\alpha+1}(1-\frac{1}{n+\alpha}) = p(n|\alpha)\frac{n+\alpha-1}{n+\alpha+1}$. An equivalent non-recursive definition of the *attenuated geometric distribution*

[1]The term modified geometric distribution is another natural name but is used to refer to another distribution.

is

$$p(n|\alpha) = \begin{cases} \frac{1}{1+\alpha} & \text{if } n = 1 \\ \frac{1}{1+\alpha} \prod_{j=0}^{n-2} \frac{j+\alpha}{j+\alpha+2} = \frac{\alpha}{(n+\alpha)(n+\alpha-1)} & \text{if } n > 1 \end{cases}$$

We use $P(n|\alpha)$ to denote the cumulative distribution function for the attenuated geometric distribution.

$$P(n|\alpha) = \sum_{j=1}^{n} p(j|\alpha) = \frac{1}{1+\alpha} + \sum_{j=2}^{n} p(j|\alpha) = 1 - \frac{\alpha}{n+\alpha}$$

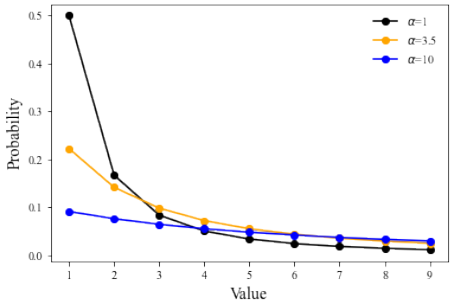

Figure 1: Probability function for different values of α.

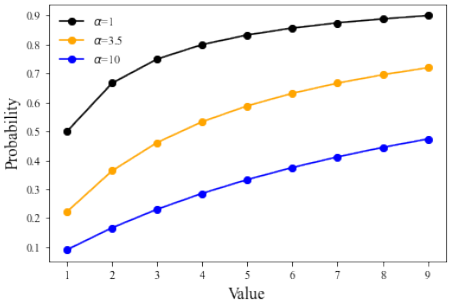

Figure 2: Cumulative distribution function for different values of α.

The mode of the distribution is 1. In addition, as α approaches 0, the probability placed on $n = 1$ approaches 1 (i.e., $\lim_{\alpha \to 0} p(n = 1|\alpha) = 1$).

Figures 1 shows the probability density function for various values of the attenuation parameter. Figure 2 shows the corresponding cumulative density function for these values of the attenuation parameter.

The first moment of the attenuated geometric distribution is ∞ (i.e., $E_p(n|\alpha) = \infty$).

Proof:

$$E(n|\alpha) = \sum_{n=1}^{\infty} np(n|\alpha) = \frac{1}{\alpha+1} + \alpha \sum_{n=2}^{\infty} \frac{n}{(n+\alpha)(n+\alpha-1)}$$

We focus on the infinite sum, simplifying the expression while ensuring that the original sum is larger than the simplified expressions. First, we drop summands from the infinite sum. Let $m = 3$ if $\alpha < 3$, otherwise $m = \lceil \alpha \rceil$.

$$\sum_{n=2}^{\infty} \frac{n}{(n+\alpha)(n+\alpha-1)} > \sum_{n=m}^{\infty} \frac{n}{(n+\alpha)(n+\alpha-1)}$$

Second, we use the fact that $n/(n+\alpha) \geq 1/2$ for each of the remaining summands.

$$\sum_{n=m}^{\infty} \frac{n}{(n+\alpha)(n+\alpha-1)} \geq \frac{1}{2} \sum_{n=m}^{\infty} \frac{1}{n+\alpha-1}$$

The infinite sum $\sum_{n=1}^{\infty} 1/n$ is divergent and the sum of any finite number of terms in this sequence is finite. Because the final infinite sum is equal to the divergent sequence with a finite set of initial summands removed, it is also divergent.

3 Application to Streaming Sampling Algorithms

In this section, we define the inverse transform sampling procedure for the attenuated geometric distribution and describe an algorithm using this sampling procedure to sample a random item from an enumeration. Our algorithm improves on existing approaches by using half the number of calls to a random number generator.

The formula for the cumulative distribution function makes defining an inverse transform sampling procedure easy. If we are given a sampling procedure that can produce a random

sample r in $(0, 1)$ then we simply solve for n setting the cumulative distribution function to r. Thus, if $r \sim U(0, 1)$ then $\lceil \frac{r\alpha}{1-r} \rceil \sim p(n|\alpha)$.[2]

Algorithm 1 An efficient algorithm for sampling a random item

Input: An *enumeration* of items where 1 is index of first item.
Output: None if *enumeration* is empty or a random item.

```
index ← 1
item ← None
while TRUE do
    next_item ← enumeration.nth(index)
    if next_item is None then
        return item
    else
        item = next_item
        r = random()
        index = index + max(1, ⌈rα/(1−r)⌉)
```

Algorithm 1 samples a random item from an enumeration using samples from the attenuated geometric distribution. We assume that we are given an *enumeration* and that one can retrieve the n^{th} item in the enumeration by calling *enumeration.nth*(n). We prove that this algorithm samples a random item by induction on the number n of items visited in the enumeration. For the base case, when one item is visited, we have selected the one item with probability one and thus all items visited have the same probability. Now we assume the claim is true for visiting n items and show it to be true for visiting $n + 1$ items. Before visiting the $n+1^{th}$ item, each of the items $1, \ldots n$ are selected with probability $1/n$. There are n paths for selecting the $n + 1$ item each having probability $1/(n(n + 1))$ and, thus, the item is selected with probability $1/(n + 1)$. The previously selected item will remain the selected item with probability $n/(n + 1)$ and thus, each of the items $1, \ldots n$ will remain the selected item with probability $n/(n(n + 1))$, which is $1/(n + 1)$ which concludes the proof.

Next, we consider the expected number of calls to the random number generator when we use random samples from attenuated geometric distribution to skip items in the enumeration as described in Algorithm 1. We can see that the median of the attenuated geometric distribution given α is α. Note that $\alpha \geq 1$ throughout its use in Algorithm 1. Thus, in expectation, for half of the sampled offsets we will more than double the index of the selected item. One call to random is required for each for each update of the selected item, thus, in expectation, $O(\log(N))$ calls to the random number generator are required if there are N items in the enumeration. This is the expected and optimal speedup [Vit85].

Previous approaches take an alternative approach in which, a scalar score is sampled from a fixed distribution and the score is associated with each item. The item with the lowest score is kept. In order to not require a random sample for each item in the enumeration, a geometric distribution is used where the bias is computed using the current lowest score and the cumulative distribution function of the fixed distribution. The geometric distribution is used to compute an offset and that item is selected and given a score from a random sample from the fixed distribution conditioned so that the sample is less than the previous lowest score. It is natural to use a uniform distribution to simplify sampling. This scoring approach requires two calls to the random number generator per item selected from the enumeration whereas, Algorithm 1 demonstrates that only one call is require when sampling offsets from the attenuated geometric distribution.

The more general streaming problem of sampling k items from an enumeration is a natural generalization. While one can define a natural generalization of the attenuated geometric distribution to solve this problem, the resulting distribution is less tractable. The scoring approach described above, however, generalizes nicely.[3] One simply needs to keep the k lowest scoring items (e.g., in a priority queue). This approach can be further generalized to handled weighted items. In the case of equally weights items, the need for storing scores for the k items can be eliminated as well as the need for a data structure to store them (e.g., the priority queue). By selecting the fixed score distribution to be the uniform distribution

[2]Correct implementation depends on the implementation the procedure for sampling a uniform random number (whether it can return 0, or 1) and properly treating potential overflow of the ratio. If the random number generator can generate 0 then the formula should be $\max(1, \lceil \frac{r\alpha}{1-r} \rceil)$.

[3]These solutions are sometimes called reservoir sampling algorithms.

over $(0, 1)$ we need only maintain a current minimum score and a new minimum score can be computed using a random sample and computing the k^{th} order statistic. In this case, a selected item randomly replaces one of the current k items [Li94]. A similar outcome can be achieved with other continuous univariate distributions (e.g., the exponential distribution) but procedures for sampling from a uniform random distribution are typically provided by programming languages or random sampling packages. The cost of eliminating the k scores is an additional call to a random number generator. In particular, this approach requires three calls to a random number generator for each item selected from the enumeration.

References

[Li94] Kim-Hung Li. Reservoir-Sampling Algorithms of Time Complexity o(n(1+log(n/n))). *ACM Transactions on Mathematical Software*, 20(4):481–493, 1994.

[Vit85] Jeffrey S. Vitter. Random Sampling with a Reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985.