Information Processing Letters 115 (2015) 923–926

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Weighted sampling without replacement from data streams

Vladimir Braverman [a,1], Rafail Ostrovsky [b,c], Gregory Vorsanger [a,2]

[a] Johns Hopkins University, Department of Computer Science, United States
[b] University of California Los Angeles, Department of Computer Science, United States
[c] University of California Los Angeles, Department of Mathematics, United States

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 5 May 2015
Accepted 18 July 2015
Available online 22 July 2015
Communicated by M. Chrobak

Keywords:
Algorithms
On-line algorithms
Sampling
Streaming algorithms

ABSTRACT

Weighted sampling without replacement has proved to be a very important tool in designing new algorithms. Efraimidis and Spirakis [5] presented an algorithm for weighted sampling without replacement from data streams. Their algorithm works under the assumption of precise computations over the interval $[0, 1]$. Cohen and Kaplan [3] used similar methods for their bottom-k sketches.

Efraimidis and Spirakis ask as an open question whether using finite precision arithmetic impacts the accuracy of their algorithm. In this paper we show a method to avoid this problem by providing a precise reduction from k-sampling without replacement to k-sampling with replacement. We call the resulting method Cascade Sampling.

1. Introduction

Random sampling is a fundamental tool that has many applications in computer science (see e.g., Motwani and Raghavan [12], Knuth [9], Tille [15], and Olken [13]). Random sampling methods are widely used is data stream processing because of their simplicity and efficiency [14, 8,7,6,10,11]. In a stream, the size of the domain and the probability of sampling an element both change constantly; this makes the process of sampling non-trivial. We distinguish between sampling *with replacement*, where all samples are independent (and thus can be repeated), and sampling *without replacement*, where repetitions are prohibited.

In particular, weighted sampling without replacement has proven to be a very important tool. In weighted sampling, each element is given a weight, where the probability of an element being selected is based on its weight. In their work Efraimidis and Spirakis [5] presented an algorithm for weighted sampling without replacement. Cohen and Kaplan [3] use similar methods for their bottom-k sketches. While their preliminary implementation yielded promising results, Efraimidis and Spirakis [5] state, as the main open problem of the paper, "*However, the question if, and to what extent, the finite precision arithmetic affects the algorithms remains an open problem.*"

In this paper we continue this work and provide a new algorithm to avoid the issue of relying on finite precision arithmetic. With this result we show that precision loss is not required in order to sample without replacement. We accomplish this by providing a precise reduction from k-sampling without replacement to k-sampling with replacement, using a special case of k-sampling with replacement, unit sampling (where $k = 1$). Additionally, we believe that in the future our method of expressing differ-

E-mail addresses: vova@cs.jhu.edu (V. Braverman), rafail@cs.ucla.edu (R. Ostrovsky), gregvorsanger@jhu.edu (G. Vorsanger).

[1] This material is based upon work supported in part by the National Science Foundation under Grant No. 1447639, the Google Faculty Award and DARPA grant N660001-1-2-4014. Its contents are solely the responsibility of the authors and do not represent the official view of DARPA or the Department of Defense.

[2] This material is based upon work supported in part by Raytheon BBN Technologies.

http://dx.doi.org/10.1016/j.ipl.2015.07.007
0020-0190/

ent random samples via reduction will provide a tool that allows further translation of other sampling methods into a more effective form for streams.

1.1. Related work

Due to its fundamental nature, the problem of random sampling has received considerable attention in the last few decades.

In 2005, Vitter [16] presented uniform sampling using a reservoir (with and without replacement) over streams. Further, the question of reductions between sampling methods has been addressed before. For instance, Chaudhuri, Motwani and Narasayya [2] briefly discuss reductions for various sampling methods. Cohen and Kaplan [3] use a "mimicking process" in their papers, which is essentially a reduction from sampling without replacement to sampling with replacement.

Chaudhuri, Motwani and Narasayya [2] use the well-known method of "over-sampling", i.e. we sample the set independently until k distinct elements are obtained. Clearly, this schema does not introduce any precision loss, since unit sampling is used as a black-box.

Unfortunately, the amount of resources required to determine this information is a function of the weight distribution for the data set, and thus can be arbitrarily large.

In particular, consider the case when there is an element with weight that is overwhelmingly larger than the rest of the population. In this case, the number of repetitions found while sampling with replacement is significantly larger then k.

Probably the first effective non-streaming solution for the weighted sampling without replacement problem was the algorithm of Wong and Easton [17]. It is used by many other algorithms (see Olken [13] for the discussion). For data streams, Efraimidis and Spirakis [5] proposed an algorithm that is based on the "exponent method". The algorithm requires precise computations of random keys $r^{1/w(p)}$, where $r \sim U[0,1]$. The sample generated is composed of the k elements with maximal keys. Cohen and Kaplan [3] used similar methods as a building block for their bottom-k sketches. The bottom-k sketch is an effective construction that has been extensively used for various applications including approximations of aggregative queries over data streams. As Cohen and Kaplan [3] show, these methods are very effective in practical applications and are superior to the sketches that are based on sampling with replacement.

While effective in practice, the algorithms of Efraimidis and Spirakis and Cohen and Kaplan introduce a loss of accuracy, since their techniques require additional floating point arithmetic operations.

1.2. Results

In this paper we show that the tradeoff between precision and performance is not a necessary property of sampling without replacement from data streams and construct a precise streaming reduction from k-sampling without replacement to k-sampling with replacement. This result provides a practical improvement to the algorithms of Efraimidis and Spirakis in cases where high accuracy is required.

Our method is yields a surprisingly simple algorithm, given the importance of sampling without replacement and the existence of many previous methods. We call this algorithm Cascade Sampling. In particular, when used with the algorithm from [2] Cascade Sampling requires $O(k)$ memory, constant time per element and the same precision as in [2].

Algorithm 1 Black-Box WR2: Algorithm for Weighted Unit Sampling.

1. $W \leftarrow 0$.
2. Initialize reservoir with length $r = 1$, λ_0.
3. For each tuple t in stream:
   (a) Get next tuple t with weight $w(t)$
   (b) $W \leftarrow W + w(t)$
   (c) Set $\lambda_0 = t$ with prob. $\frac{w(t)}{W}$
4. Return λ_0

1.3. Intuition

Let Λ be *any* algorithm that maintains a unit weighted sample from stream D. Similarly to the over-sampling method, we maintain instances of Λ. Namely, we maintain k instances $\Lambda_1, \ldots, \Lambda_k$. However, we introduce the idea of *stream modification*. That is, instead of applying Λ independently and symmetrically on D, we apply Λ_i on the modified stream D_i that does not contain samples of Λ_j for $j < i$. In particular, Λ_i may process its input elements in an order *different from the order of their arrival in* D. This simple but novel idea is sufficient to solve the problem. In particular, we can claim that the input of Λ_i is a random set that precisely matches the definition of weighted sampling without replacement. Since we use Λ as a black box with only a constant number of auxiliary variables, specifically pointers, the resulting schema is a precise reduction.

2. Definitions

An important building block of our algorithm is the concept of a unit sample, that is, the ability to sample a single element from a set.

Definition 1. Let S be a finite set of elements and let w be a non-negative function $w : S \to R$. A random element X_S with values from S is a **unit weighted random sample** if, for any $a \in S$, $P(X_S = a) = \frac{w(a)}{w(S)}$. Here $w(S) = \sum_{a \in S} w(a)$.

For an algorithm instantiating weighted unit sampling we provide Black-Box WR2 from [2]. Black-Box WR2 is a unit sample when $r = 1$ (Algorithm 1).

Definition 2. A **data stream** is an ordered, set of elements, $p_1, p_2, \ldots, p_n$, that can be observed only once. An algorithm A is a streaming sampling algorithm if A outputs a sample using a single pass over the data set.

Definition 3. A set $X = \{X_1, \ldots, X_k\}$ is called a k-**sample with replacement** from S if $X_1, \ldots, X_k$ are independent random unit samples from S.