Taking snapshots from a stream - preliminary report

Dominik Bojko, Jacek Cichoń

Department of Computer Science
Faculty of Fundamental Problems of Technology
Wrocław University of Science and Technology
Poland

This work is devoted to a certain class of probabilistic snapshots for elements of the observed data stream. We show how one can control their probabilistic properties and we show some potential applications. Our solution can be used to store information from the observed history with limited memory. It can be used for both web server applications and Ad hoc networks and, for example, for automatic taking snapshots from video stream online of unknown size.

The class of algorithms considered in this paper may be treated as a subclass of reservoir sampling algorithms with reservoir of size 1. Our solutions can also be treated as a generalization of the classical Jeffrey Vitter's Algorithm R.

Keywords: big data, streaming, asymptotic distribution

1 Introduction

Suppose that we are observing a long stream $x_1, x_2, \ldots$ of data. Our goal is to keep an element from this stream with a prescribed position. For example, we may want to keep the element x_i with index i close to $\lfloor n/2 \rfloor$ after reading first n elements from the stream. Of course this problem is trivial if we have a direct access to all elements $x_1, \ldots, x_n$ or if we know the number n in advance. But in the case of large amount of data of unknown length keeping all information into memory is expensive or undesirable or impossible. Suppose hence that the number n is unknown and that our memory resources are limited.

In this article we investigate series of randomized procedures which allow us to choose elements located near the required position in the stream of data. All these procedures are based on the same schema. They only differ on a sequence (α_n) of probabilities which are used for control of changes of stored data. In each case we will have $\alpha_1 = 1$. We will use three variables: K, n and $data$. Initially we put $K = 0$, $n = 0$ and we set $data$ as nil value:

```
Initialization: K:= 0; n:= 0; data:= nil;
```

After reading an element x from the stream we call the following update procedure:

```
procedure Update(x)
  n++
  if (random() <= a(n))
```

```
  begin
    K:= 1;
    data: = x
  end else
  begin
    K++
  end
end
```

In the procedure `Update` we used the function `random()` which is a high quality pseudo-random generator of random reals from the interval $[0, 1]$. The function `a(n)` represents the probability sequence (α_n). We call the variable K a **probabilistic snapshot** from data stream.

Connection with reservoir sampling The class of algorithms considered in this paper may be treated as a subclass of reservoir sampling algorithms with reservoir of size 1 (see Tillé [2006], Knuth [1997]). Suppose that we use this procedure with the sequence $\alpha_n = \frac{1}{n}$. In this case we obtain the classical Jeffrey Vitter's Algorithm R published in Vitter [1985]. Let K_n denotes value of the random variable K after reading n items from the stream. It is well known that in this case (i.e. when $\alpha_n = \frac{1}{n}$) we have $\Pr[K_n = i] = \frac{1}{n}$ for each $i \in \{1, \ldots, n\}$, i.e. that the random variable K_n has the uniform distribution on the set $\{1, \ldots, n\}$. Therefore $E[K_n] = (n + 1)/2$.

Applications Here we show some possible applications of methods discussed in this paper (a more detailed discsusion is in Section 4):

- The solution proposed in this paper may by used for storing fixed number of snapshots from an observed movie of unknown length. For example we may want to store short samples of the movie taken at times closed to 0, $\frac{1}{10}T$, $\frac{2}{10}T \ldots \frac{9}{10}T$, T from a movie of length T.
- We may need a sample of data from times close to n, $n - C$, $n - 2 \cdot C$,..., $n - k \cdot C$, where n is an index of current item, C is a fixed distance and k is a reasonably small natural number.
- We may need to observe a sample from stock market in such a way that the snapshots from the past should be less rare than snapshots from times close to the present.

1.1 Mathematical notations and background

We denote by $\mathrm{E}[X]$ the expected value of the random variable X. Let us recall that a discrete random variable X has geometric distribution with parameter $p \in [0, 1]$ ($X \sim \mathrm{Geo}(p)$) if $P[X = k] = (1-p)^{k-1}p$ for $k \geq 1$. If $X \sim \mathrm{Geo}(p)$ then $\mathrm{E}[X] = \frac{1}{p}$. A random variable Y has an exponential distribution with parameter g ($Y \sim \mathrm{Exp}(\lambda)$) if its support is $[0, \infty)$ and $\Pr[Y > x] = e^{-\lambda x}$ for each $x \geq 0$. If $Y \sim \mathrm{Exp}(\lambda)$ then $\mathrm{E}[Y] = \frac{1}{\lambda}$. A random variable Z has the Beta distribution with parameters $a, b > 0$ ($Z \sim \mathrm{B}(a, b)$) if its support is $[0, 1]$ and its density is given by the function $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1 - x)^{b-1}$, where $\Gamma(x)$ is the standard generalization of the factorial function. If $Z \sim \mathrm{B}(a, b)$ then $\mathrm{E}[Z] = \frac{a}{a+b}$.

A sequence $X_1, X_2, \ldots$ of real-valued random variables converge in distribution to a random variable X if $\lim_{n\to\infty} F_n(x) = F(x)$, for every number $x \in \mathbb{R}$ at which the function F is continuous, where F_n and F are the cumulative distribution functions of random variables X_n and X, respectively.

We denote by H_n the n-th harmonic number, i.e. $H_n = \sum_{i=1}^{n} \frac{1}{i}$. We will use the following well known approximation $H_n = \ln(n) + \gamma + \mathrm{O}\left(\frac{1}{n}\right)$, where $\gamma \approx 0.577$ is the Euler - Mascheroni constant. We use the following symbols $[\cdot]$ for the Iverson logical bracket. We will also use the following inequalities (extension of the classical Weierstarass product inequality):

$$1 - \sum_{i=1}^{n} x_i \leq \prod_{i=1}^{n} (1 - x_i) < 1 - \sum_{i=1}^{n} x_i + \sum_{1 \leq i < j \leq n} x_i x_j. \tag{1}$$

which holds for any sequence $x_1, \ldots, x_n$ of real numbers from the interval $[0, 1]$ (see Klamkin and Newman [1970]). We will use several times the following classical inequality, which holds for any non-increasing function $f : [a, b] \to \mathbb{R}$ (where $a \leq b$ are integers):

$$\int_a^b f(x)dx + f(b) \leq \sum_{i=a}^{b} f(i) \leq f(a) + \int_a^b f(x)dx\ . \tag{2}$$

Lemma 1 *For arbitrary sequence $\alpha_1, \ldots, \alpha_n$ we have $\sum_{k=1}^{n} \sum_{i=k}^{n} \alpha_i = \sum_{k=1}^{n} k \cdot \alpha_k$.*

Lemma 2 *For arbitrary sequence $\alpha_1, \ldots, \alpha_n$ of positive real numbers we have*

$$\sum_{k=1}^{n} \sum_{k \leq i < j \leq n} \alpha_i \alpha_j \leq \frac{1}{2} \left(\sum_{k=1}^{n} \sqrt{k} \cdot \alpha_k \right)^2.$$

Proof: Let us observe that if $x, y \geq 0$ then $\min(x, y) \leq \sqrt{x}\sqrt{y}$. Therefore

$$\sum_{k=1}^{n} \sum_{k \leq i < j \leq n} \alpha_i \alpha_j \leq \frac{1}{2} \sum_{k=1}^{n} \sum_{k \leq i,j \leq n} \alpha_i \alpha_j = \frac{1}{2} \sum_{k=1}^{n} \sum_{i,j=1}^{n} \alpha_i \alpha_j \left[(k \leq i) \wedge (k \leq j) \right] =$$
$$\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \sum_{k=1}^{n} [k \leq i] \cdot [k \leq j] = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \min(i, j) \leq$$
$$\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \sqrt{i}\sqrt{j} = \frac{1}{2} \left(\sum_{k=1}^{n} \sqrt{k} \cdot \alpha_k \right)^2.$$

□

2 General Properties

Let us fix a sequence α_n of reals from interval $[0, 1]$ such that $\alpha_1 = 1$. Let us consider the sequence K_n of consecutive integers tracing the value of the variable K during the run of the Algorithm with controlling sequence α_n. Clearly we have $1 \leq K_n \leq n$ for each n. Notice that $\Pr[K_1 = 1] = 1$, $\Pr[K_n = 1] = \alpha_n$ and

$$\Pr[K_{n+1} = k + 1] = (1 - \alpha_{n+1}) \Pr[K_n = k] \tag{3}$$

for $k > 1$. The following formula may be derived from the previous one

$$\Pr[K_n = k] = \alpha_{n-k+1} \prod_{i=n-k+2}^{n} (1 - \alpha_i) . \tag{4}$$

Moreover

$$\Pr[K_n \geq k] = \prod_{i=n-k+2}^{n} (1 - \alpha_i) \tag{5}$$

We will also use the random variable $L_n = n + 1 - K_n$. From the above formula we deduce that

$$\Pr[L_n = k] = \alpha_k \prod_{i=0}^{n-k-1} (1 - \alpha_{n-i}) \tag{6}$$

and

$$\Pr[L_n \geq k] = 1 - \prod_{i=k}^{n} (1 - \alpha_i) \tag{7}$$

The following inequalities follows directly from Lemmas 1 and 2 and the inequalities 1:

Corollary 1 $\sum_{k=1}^{n} k\alpha_k - \frac{1}{2}\left(\sum_{k=1}^{n} \sqrt{k} \cdot \alpha_k\right)^2 \leq \mathrm{E}\,[L_n] \leq \sum_{k=1}^{n} k\alpha_k$.

The sequence $\mathrm{E}\,[K_n]$ satisfies the following simple linear first order difference equation:

Theorem 1 $\mathrm{E}\,[K_{n+1}] = 1 + (1 - \alpha_{n+1})\mathrm{E}\,[K_n]$

3 Special cases

In this section we will consider some examples of sequences (α_n) which may have applications for controlling the history of a massive streams of data. For example, snapshots generated by the sequence $\alpha_n = \frac{1}{n}$ are uniformly distributed in the set $\{1, \ldots, n\}$, hence a collection of such snapshots may be used as random uniformly distributed sample controlling of behavior of a stream.

In a series of six subsections we shall analyze behavior of the random variable K_n for sequences of the form $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ for various fixed parameters $\alpha > 0$ and $g > 0$. This is a summary of our results:

- if $\alpha = 0$ and $g \in (0, 1)$ then the random variable K_n converges in distribution to the geometric distribution $\mathrm{Geo}(g)$. Moreover $\mathrm{E}\,[K_n] \sim \frac{1}{g}$
- if $\alpha \in (0, 1)$ then the random variable $\frac{K_n}{n^\alpha}$ converges in distribution to exponential distribution $\mathrm{Exp}(g)$. Moreover $\mathrm{E}\,[K_n] \sim \frac{n^\alpha}{g}$
- if $\alpha = 1$ then the random variable $\frac{K_n}{n}$ converges in distribution to the beta distribution $\mathrm{Beta}(1, g)$. Moreover $\mathrm{E}\,[K_n] \sim \frac{n}{g+1}$
- if $\alpha \in (1, 2)$ then $\mathrm{E}\,[L_n] \sim \frac{g}{2-\alpha} n^{2-\alpha}$
- if $\alpha = 2$ then $\mathrm{E}\,[L_n] = g \ln(n) + \mathrm{O}\,(1)$
- if $\alpha > 2$ then $\mathrm{E}\,[L_n] = \mathrm{O}\,(1)$.

Notice that for $\alpha > 1$ we do not know asymptotic distributions of sequence of random variables (L_n)

3.1 Fixed value

Let us fix a real number $a > 1$ and let $\alpha_n = \frac{1}{a}$ for each $n > 1$. In this case we have a closed formula for $E[K_n]$, namely, from Eq. (4) we immediately get

$$\Pr[K_n = k] = \begin{cases} \frac{1}{a}(1-\frac{1}{a})^{k-1} & : 1 \le k < n \\ (1-\frac{1}{a})^{n-1} & : k = n \end{cases}$$

From this formula we deduce that

1. The sequence (K_n) of random variables converges in distribution to the geometrical distribution with parameter $\frac{1}{a}$.
2. $E[K_n] = a\left(1 - \left(\frac{a-1}{a}\right)^n\right)$

Therefore the generated snapshot may be used for controlling a behavior of a stream at a fixed position in the past.

3.2 Sublinear Case

In this section we consider the case when $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ for some fixed $g > 0$ and $\alpha \in (0,1)$. We shall show that the normalized random variable $\frac{K_n}{n^\alpha}$ converges in distribution to the exponential $\mathrm{Exp}\,(g)$ distribution.

Theorem 2 *Let $g > 0$ and $\alpha \in (0,1)$. Let $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ and let $x > 0$. Then*

$$\lim_{n\to\infty} \Pr\left[\frac{K_n}{n^\alpha} \le x\right] = 1 - e^{-gx}\,.$$

Proof: Let $k = \lfloor n^\alpha x \rfloor$. Using formula (5) we obtain

$$\Pr\left[\frac{K_n}{n^\alpha} > x\right] = \Pr[K_n > n^\alpha x] = \Pr[K_n > k] = \prod_{i=n-k+1}^{n}\left(1 - \frac{g}{i^\alpha}\right)$$

(for sufficiently large n). Therefore

$$\Pr\left[\frac{K_n}{n^\alpha} > x\right] < \left(1 - \frac{g}{n^\alpha}\right)^k \le \left(1 - \frac{g}{n^\alpha}\right)^{n^\alpha x - 1} \tag{8}$$

and

$$\Pr\left[\frac{K_n}{n^\alpha} > x\right] > \left(1 - \frac{g}{(n-k+1)^\alpha}\right)^k > \left(1 - \frac{g}{(n-\lfloor n^\alpha x\rfloor)^\alpha}\right)^{n^\alpha x} = \left(1 - \frac{g/(1-\lfloor n^\alpha x\rfloor/n)}{n^\alpha}\right)^{n^\alpha x},$$

so we see that both bounds on $\Pr\left[\frac{K_n}{n^\alpha} > x\right]$ converges to the same limit e^{-gx} when n tends to infinity. □

Theorem 3 *If $g > 0$, $\alpha \in (0,1)$ and $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ then $\lim_{n\to\infty} \mathrm{E}\left[\frac{K_n}{n^\alpha}\right] = \frac{1}{g}$.*

Proof: From Theorem 2 and Fatou's Lemma we get $\frac{1}{g} \leq \liminf_{n\to\infty} \mathrm{E}\left[\frac{K_n}{n^\alpha}\right]$. On the other hand, inequality (8) applied for $n > g^{1/\alpha}$ implies that

$$\mathrm{E}\left[\frac{K_n}{n^\alpha}\right] = \int_0^\infty \Pr\left[\frac{K_n}{n^\alpha} > t\right] dt \leq \int_0^\infty \left(1 - \frac{g}{n^\alpha}\right)^{n^\alpha t - 1} dt = -\frac{1}{(n^\alpha - g)\ln(1 - \frac{g}{n^\alpha})} \stackrel{n\to\infty}{\longrightarrow} \frac{1}{g}\,.$$

□

Corollary 2 *If $g > 0$ and $\alpha_n = \min\{1, \frac{g}{n}\}$ then $\mathrm{E}[K_n] = \frac{n^\alpha}{g} + \mathrm{o}(n^\alpha)$.*

Remark. We know much more precise results in some special cases. For example, if $\alpha_n = \frac{1}{\sqrt{n}}$ then $\mathrm{E}[K_n] = \sqrt{n} - \frac{1}{2} + \frac{1}{2\sqrt{n}} + \frac{1}{8n} + \mathrm{O}\left(\frac{1}{n^{3/2}}\right)$.

3.3 Linear Case

In this section we consider the case when $\alpha_n = \min\{1, \frac{g}{n}\}$ for some fixed $g > 0$. We shall show that the normalized random variable $\frac{K_n}{n}$ converges in distribution to the Beta(1,g) distribution.

Theorem 4 *Let $g > 0$. Let $\alpha_n = \min\{1, \frac{g}{n}\}$ and let $x \in (0,1)$. Then*

$$\lim_{n\to\infty} \Pr\left[\frac{K_n}{n} \leq x\right] = 1 - (1-x)^g\,.$$

Proof: Let $k = \lfloor nx \rfloor$. Then we have

$$\Pr\left[\frac{K_n}{n} > x\right] = \Pr[K_n > nx] = \Pr[K_n > k] = \prod_{i=n-k+1}^{n} (1 - \alpha_i)\,.$$

Therefore, for sufficiently large n, we have

$$\Pr\left[\frac{K_n}{n} > x\right] = \prod_{i=n-k+1}^{n} \left(1 - \frac{g}{i}\right)\,.$$

Hence

$$\ln\left(\Pr\left[\frac{K_n}{n} > x\right]\right) = -\sum_{i=n-k+1}^{n} \ln\left(\frac{1}{1-\frac{g}{i}}\right) = -\sum_{i=n-k+1}^{n} \sum_{a\geq 1} \frac{1}{a}\left(\frac{g}{i}\right)^a =$$
$$-g\sum_{i=n-k+1}^{n} \frac{1}{i} - \sum_{a\geq 2} \frac{g^a}{a} \sum_{i=n-k+1}^{n} \frac{1}{i^a}\,.$$

Notice that $\sum_{i=n-k+1}^{n} \frac{1}{i} = H_n - H_{n-k}$, and

$$-g(H_n - H_{n-k}) = -g\left(\ln(n) - \ln(n-k) + \mathrm{O}\left(\frac{1}{n}\right) + \mathrm{O}\left(\frac{1}{n - \lfloor nx \rfloor}\right)\right) = \ln\left(\frac{n - \lfloor nx \rfloor}{n}\right)^g + O\left(\frac{1}{n}\right)$$

Therefore $\lim_{n\to\infty}(-g(H_n - H_{n-\lfloor nx \rfloor})) = \ln((1-x)^g)$. Let $A_{n,k} = \sum_{a\geq 2} \frac{g^a}{a} \sum_{i=n-k+1}^{n} \frac{1}{i^a}$. Then

$$0 < A_{n,k} < \sum_{a\geq 2} g^a \sum_{i=n-k+1}^{\infty} \frac{1}{i^a} = \sum_{i=n-k+1}^{\infty} \sum_{a\geq 2} \frac{g^a}{i^a} = \sum_{i=n-k+1}^{\infty} \frac{g^2}{i^2} \frac{1}{1-\frac{g}{i}} < 2g^2 \sum_{i=n-k+1}^{\infty} \frac{1}{i^2}$$

(the last inequality holds for $n > \frac{2g}{1-x}$), hence

$$A_{n,k} < 2g^2 \sum_{i=n-k+1}^{\infty} \frac{1}{i(i-1)} = \frac{2g^2}{n-k} = \frac{2g^2}{n - \lfloor nx \rfloor} = \mathrm{O}\left(\frac{1}{n}\right) .$$

Thus

$$\ln\left(\Pr\left[\frac{K_n}{n} > x\right]\right) = \ln\left(1 - \frac{\lfloor nx \rfloor}{n}\right)^g + \mathrm{O}\left(\frac{1}{n}\right).$$

Therefore

$$\Pr\left[\frac{K_n}{n} > x\right] = \left(1 - \frac{\lfloor nx \rfloor}{n}\right)^g e^{\mathrm{O}\left(\frac{1}{n}\right)} = \left(1 - \frac{\lfloor nx \rfloor}{n}\right)^g + \mathrm{O}\left(\frac{1}{n}\right) ,$$

so $\lim_{n\to\infty} \Pr\left[\frac{K_n}{n} > x\right] = (1-x)^g$. □

Corollary 3 *If $g > 0$ and $\alpha_n = \min\{1, \frac{g}{n}\}$ then $\lim_{n\to\infty} \mathrm{E}\left[\frac{K_n}{n}\right] = \frac{1}{g+1}$.*

Proof: Notice that $\frac{K_n}{n} \leq 1$, hence the sequence $(\frac{K_n}{n})_{n\in\mathbb{N}}$ is bounded, therefore the convergence in distribution of the sequence $(\frac{K_n}{n})_{n\in\mathbb{N}}$ to a random variable Y with Beta(1,g) distribution implies the convergence of moments (see e.g. Billingsley [2012]), hence

$$\lim_n \mathrm{E}\left[\frac{K_n}{n}\right] = \int_0^1 x \frac{d}{dx}\left(1 - (1-x)^g\right) dx = \frac{1}{1+g} .$$

□

Corollary 4 *If $g > 0$ and $\alpha_n = \min\{1, \frac{g}{n}\}$ then $\mathrm{E}[K_n] = \frac{n}{g+1} + \mathrm{o}(n)$.*

Remark 1. A slightly more complicated calculus shows that in the case considered in this section we have $\mathrm{E}[K_n] = \frac{n+1}{g+1} + \mathrm{O}\left(\frac{1}{n^g}\right)$.

Remark 2. If $\alpha_n = \frac{1}{n}$ then the snapshot K_n is uniformly distributed in $\{1, \ldots, n\}$.

3.4 Subquadratic Case

In this section we consider the case when $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ for some fixed $g > 0$ and $\alpha \in (1, 2)$. We investigate the random variable $L_n = n + 1 - K_n$. Let us observe that if $\alpha > 1$ then $3 - 2\alpha < 2 - \alpha$.

Theorem 5 *If $g > 0$, $\alpha \in (1, 2)$ and $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ then*

$$\mathrm{E}\left[L_n\right] = \frac{g}{2-\alpha} n^{2-\alpha} + \begin{cases} \mathrm{O}\left(n^{3-2\alpha}\right) & : \alpha < \frac{3}{2} \\ \mathrm{O}\left(\ln^2(n)\right) & : \alpha = \frac{3}{2} \\ \mathrm{O}\left(1\right) & : \alpha > \frac{3}{2} \end{cases}$$

Proof: Suppose that $h \geq 1$. Using the standard interpretation of finite sums as Riemann's integral sums (see inequality 2) we easily deduce that

$$\sum_{k=h}^{n} \frac{k}{k^\alpha} = \frac{n^{2-\alpha}}{2-\alpha} + \mathrm{O}\left(1\right) \ .$$

and

$$\sum_{k=h}^{n} \frac{\sqrt{k}}{k^\alpha} = \begin{cases} \frac{n^{3/2-\alpha}}{3/2-\alpha} + \mathrm{O}\left(1\right) & : \alpha < \frac{3}{2} \\ \ln(n) + \mathrm{O}\left(1\right) & : \alpha = \frac{3}{2} \\ \mathrm{O}\left(1\right) & : \alpha > \frac{3}{2} \end{cases} \ .$$

Let us now put $h = \left\lceil g^{\frac{1}{\alpha}} \right\rceil - 1$. From the above observations and Corollary 1 we get

$$\mathrm{E}\left[L_n\right] \leq \sum_{k=1}^{h-1} k + \sum_{k=h}^{n} \frac{kg}{k^\alpha} = \frac{g}{2-\alpha} n^{2-\alpha} + \mathrm{O}\left(1\right)$$

and

$$\mathrm{E}\left[L_n\right] \geq \sum_{k=1}^{h-1} k + \sum_{k=h}^{n} \frac{kg}{k^\alpha} - \frac{1}{2}\left(\sum_{k=1}^{h-1} \sqrt{k} + g\sum_{k=h}^{n} \frac{\sqrt{k}}{k^\alpha}\right)^2 =$$
$$\frac{g}{2-\alpha} n^{2-\alpha} + \mathrm{O}\left(1\right) + \frac{1}{2}\left(\sum_{k=1}^{h-1} \sqrt{k} + g\sum_{k=h}^{n} \frac{\sqrt{k}}{k^\alpha}\right)^2$$

If $\alpha < \frac{3}{2}$ then we get

$$\mathrm{E}\left[L_n\right] \geq \frac{g}{2-\alpha} n^{2-\alpha} + \mathrm{O}\left(1\right) + \left(\mathrm{O}\left(n^{3/2-\alpha}\right)\right)^2 = \frac{g}{2-\alpha} n^{2-\alpha} + \mathrm{O}\left(n^{3-2\alpha}\right) \ .$$

The cases $\alpha = \frac{3}{2}$ and $\alpha \in (\frac{3}{2}, 2)$ are similar. □

3.5 Quadratic Case

In this section we consider the case when $\alpha_n = \min\{1, \frac{g}{n^2}\}$ for some fixed $g > 0$ and we analyze random variable $L_n = n + 1 - K_n$.

Theorem 6 *Suppose that $g > 0$ and $\alpha_n = \min\{1, \frac{g}{n^2}\}$. Then $\mathrm{E}[L_n] = g\ln(n) + \mathrm{O}(1)$.*

Proof: Let $h = \lceil\sqrt{g}\rceil - 1$. Directly from Corollary 1 we have

$$\mathrm{E}[L_n] \leq \sum_{k=1}^{n} k\alpha_k \leq \sum_{k=1}^{h} k + g\sum_{k=h+1}^{n} \frac{1}{k} = g\ln(n) + \mathrm{O}(1) \ .$$

So we have

$$\mathrm{E}[L_n] \geq \sum_{k=1}^{n} k\alpha_k - \frac{1}{2}\left(\sum_{k=1}^{n} \sqrt{k}\cdot\alpha_k\right)^2 \geq$$

$$\frac{h(h+1)}{2} + gH_n - gH_h - \frac{1}{2}\left(\sum_{k=1}^{h}\sqrt{k} + g\sum_{k=h+1}^{n}\frac{1}{k^{3/2}}\right)^2 = g\ln(n) + \mathrm{O}(1) \ .$$

□

Remark For some special cases there are a much more precise formulas for $\mathrm{E}[L_n]$. For example, if $\alpha_n = \frac{1}{n^2}$ then we have $\mathrm{E}[L_n] = H_{n+1} - 1$ and if $\alpha_n = \min\{1, \frac{4}{n^2}\}$ then $\mathrm{E}[L_n] = 4H_{n+1} + \frac{12}{n+2} - 10$.

3.6 Superquadratic Case

Finally, in this section we consider the case when $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$ for some fixed $g > 0$ and $\alpha > 2$. This case is the least interesting for our purpose, but we concisely consider $a > 2$ for completeness. Again, we apply random variable $L_n = n + 1 - K_n$.

Theorem 7 *Suppose that $g > 0$, $\alpha > 2$ and $\alpha_n = \min\{1, \frac{g}{n^\alpha}\}$. Then $\mathrm{E}[L_n] = \mathrm{O}(1)$.*

Proof: Notice that $\alpha - 1 > 1$. Let $h = \lfloor g^{1/\alpha}\rfloor - 1$. Directly from Corollary 1, we have

$$\mathrm{E}[L_n] \leq \sum_{k=1}^{h} k + g\sum_{k=h+1}^{n}\frac{k}{k^\alpha} = \frac{h(h+1)}{2} + g\sum_{k=h+1}^{n}\frac{1}{k^{\alpha-1}} < \infty$$

□

Remark L_n is a number of update of algorithm which saved the last snapshot. We see that in the superquadratic case algorithm save only data from the very beginning of the stream. Therefore the practical value of the snapshots constructed in this way is negligible.

4 Applications

In this chapter we will discuss two examples of applications discussed in the previous chapters of probabilistic snapshot.

4.1 Linear sampling

Let us assume that we are observing a data stream and that after reading the n-th item we would like to have access to elements laying near the points $\{\frac{k}{10} \cdot n : k = 0, \ldots, 10\}$. Of course, there is no problem with the element laying near 0 - it is sufficient to store the first element from the stream. As an element laying near n we may take the current item. So we must propose some mechanism for dealing with remaining 9 points.

Let us consider a series $K_n^1, \ldots K_n^M$ of independent random variables generated by the sequence $\alpha_n = \frac{1}{n}$. We know (see 3.3) that this is a sequence of independent random variables uniformly distributed in the set $\{1, \ldots, n\}$. Let $X_n = \{K_n^1, \ldots K_n^M\}$. Let us fix some $0 < \varepsilon < \frac{1}{10}$ and let $I_k = \{a \in \mathbb{N} : |\frac{a}{n} - \frac{k}{10}| < \varepsilon\}$. Let us observe that $\Pr[I_k \cap X_n = \emptyset] \approx (1 - 2\varepsilon)^M$. This approximation is accurate for large n. So, for simplicity we shall assume that we have an equality. Therefore

$$\Pr[\bigvee_{k=1}^{9} (I_k \cap X_n = \emptyset)] \leq 9 \cdot (1 - 2\varepsilon)^M \ .$$

The solution of inequality $9 \cdot (1 - 2\varepsilon)^M \leq \eta$ is given by $M \geq \frac{\log(\frac{\eta}{9})}{\log(1-2\epsilon)}$. By putting into this formula $\varepsilon = \frac{1}{100}$ and $\eta = 10^{-10}$ we get $M \geq 1248.5$. Therefore, if we take $M = 1250$ snapshots then

$$\Pr[\bigwedge_{k=1}^{9} (I_k \cap X_n \neq \emptyset)] > 1 - \frac{1}{10^{10}} \ .$$

Hence, with a very high probability, for each $k \in \{1, \ldots, 10\}$ we are able to choose a point from the set X_n which approximates $\frac{kn}{10}$ with precision 1%.

We performed numerical experiments with a collection of 1250 independent probabilistic snapshots $\mathcal{K}_n$= $(K_n^1, \ldots K_n^{1250})$ evolving independently according to the sequence $\alpha_n = \frac{1}{n}$. In the experiment whose results are shown in Fig. 1 after each call to the procedure Update we calculated the quality of set of snapshots defined as

$$Q(\mathcal{K}_n) = \frac{1}{n} \max \left\{ \min \left\{ \left| \frac{kn}{10} - K_n^j \right| : j = 1, \ldots, 1250 \right\} : k = 1, \ldots, 9 \right\} \ .$$

We may observe in this figure several regularities which are connected with the specific method of generation of sequences $\mathcal{K}_n$ (for example, it is clear that sequences $\mathcal{K}_n$ and $\mathcal{K}_{n+1}$ are not independent of each other). We can see that the predicted 1% precision has been achieved.

4.2 Bitcoin capitalization

We used bitcoin cap data from Quandl [2017] to test probabilistic snapshots concentrated at the end of the data stream. The data stream consists of 1439 records from 2013-09-30 to 2017-12-31 (one record for one day). Since the length of this stream of data is relatively short, we decided to use only 100 probabilistic snapshots. We used the probability sequence $\alpha_n = \frac{0.1}{\sqrt{n}}$. From Sec. 3.2 we know that in this case the expected value of snapshots is close to $10\sqrt{1439} \approx 380$.

We added to generated probabilistic snapshots to points: the first one, and the last one. The results of this experiment is shown at Fig. 2. Let us remark that the most left points at these diagrams corresponds

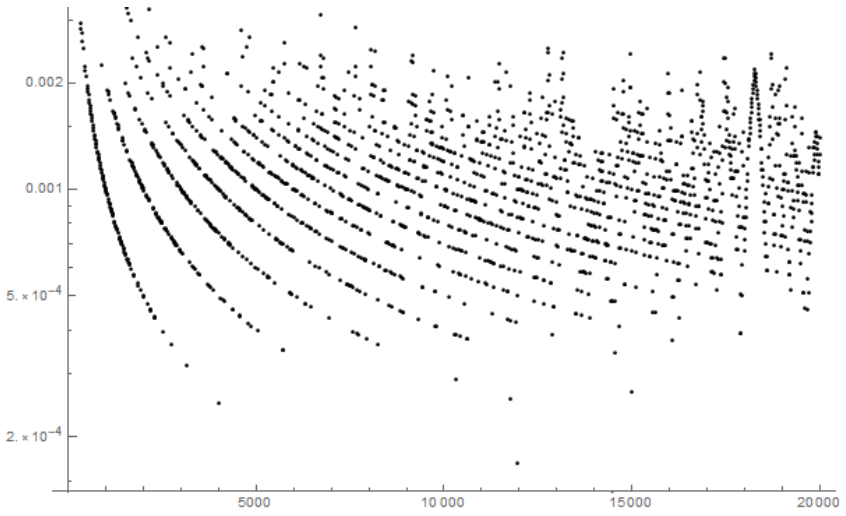

Fig. 1: Quality of approximation of points $\frac{kn}{10}$, $k = 1, \ldots, 9$, by a collection of 1250 snapshots.

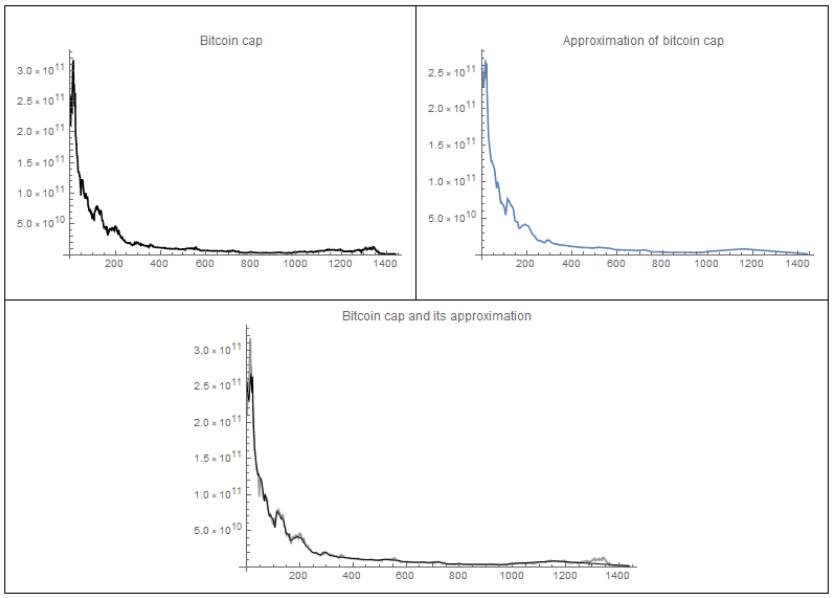

Fig. 2: Precision of bitcon cap approximation by 100 snapshots.

to data from the day 2017-12-31 and the most right point to 2013-09-30 (so we reversed the typical order of such kind of diagrams). Note that despite the large fluctuations in the bitcoin market at the end of 2017, using only 100 snapshots makes it possible to reproduce the main trend of this parameter of the bitcoins

market quite faithfully.

Acknowledgment

The authors would like to thank Zbigniew Gołębiewski and Jakub Lemiesz for usefull hints and comments on the proofs presented in this paper. This research was supported by the Polish National Science Center Grant 2013/09/B/ST6/02258.

References

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, anniversary edition, 2012. ISBN 1118122372, 978-1118122372.

M. S. Klamkin and D. J. Newman. Extensions of the Weierstrass Product Inequalities. *Mathematics Magazine*, 43(3):137–141, 1970. ISSN 0025570X, 19300980. URL `http://www.jstor.org/stable/2688388`.

D. E. Knuth. *The Art of Computer Programming, Volume 2 (3rd Ed.): Seminumerical Algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1997. ISBN 0-201-89684-2.

Quandl. Bitcoin Mining Statistics. `https://www.quandl.com/data/BITCOINWATCH/MINING`, 2017.

Y. Tillé. *Sampling Algorithms*. Springer Series in Statistics. Springer, 2006. ISBN 9780387308142. URL `https://books.google.pl/books?id=2auW1rVAwGMC`.

J. S. Vitter. Random Sampling with a Reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, Mar. 1985. ISSN 0098-3500. doi: 10.1145/3147.3165. URL `http://doi.acm.org/10.1145/3147.3165`.