Incremental Tracking of Multiple Quantiles for Network Monitoring in Cellular Networks

Jin Cao, Li Erran Li, Aiyou Chen and Tian Bu
Bell Labs, Alcatel-Lucent
(cao, erranlli, aychen, tbu)@research.bell-labs.com

ABSTRACT

Network monitoring in cellular networks requires the tracking of quantiles for data distributions of many evolving network measurements (e.g. number of high signaling subscribers per minute). Most quantile estimation algorithms are based on a summary of the empirical data distribution, using either a representative sample or a global approximation of the entire distribution. In contrast, by viewing data as a quantity from a random distribution, the stochastic approximation (SA) for quantile estimation does not keep a global approximation, but rather local approximations at the quantiles of interest, and therefore uses negligible memory even for estimating tail quantiles.

However, the current stochastic approximation algorithm for quantile estimation tracks each quantile separately, and this may lead to a violation of the *monotone property* of quantiles. In this paper, we propose a stochastic approximation technique that enables the simultaneous tracking of multiple quantiles. Our technique maintains the monotone property of different quantiles, and is adaptive to changes in the data distribution. We evaluate its performance using real cellular provider datasets. Our results show that the technique is very efficient.

Categories and Subject Descriptors

C.2.1. [**Computer-Communication Networks**]: Network Operations – Network Monitoring

General Terms

Algorithms, Measurement, Theory

Keywords

Stochastic approximation, cellular networks, incremental quantile estimation

MICNET'09, September 21, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-753-0/09/09 ...$10.00.

1. INTRODUCTION

Cellular network monitoring tracks many event counts per time interval, such as number of mobiles running P2P applications (P2P mobiles), number of mobiles who have generated high number of signaling messages (high signaling subscribers), number of battery attacks. Quantiles summarize the cumulative distributions of these event counts. For example, the 90% quantiles of high signaling subscribers is the number q where 90% of the time, the number of high signaling subscribers is smaller than q. Quantiles need to be tracked continuously over time for monitoring network health.

The main line of work [4, 3, 6, 1] on continuous quantile evaluation keeps a representative sample or summary information (e.g. random subset sum) of the empirical data. The quantile is computed from this summary data. However, to obtain quantile estimates with good accuracy, this summary information tends to be memory expensive, which is especially true for extreme quantiles since the accuracy requirement is much higher. For continuous streams whose underlying data distribution is changing over time, this also may lead to a large bias in quantile estimates since most of the summary information may be out of date.

Rather than maintaining a global approximation of all quantiles, the other line of work of using stochastic approximation (SA) for quantiles is to view the data as a random quantity from a unknown distribution, and incrementally build local approximations of the distribution function only in the neighborhood of the quantiles. As a result, there is no additional memory required except a few summary data (1 or 2 number) at the quantiles. This incremental nature is especially amenable to continuous data updates. It has been shown by Tierney in [8] that, for stationary distribution, the estimated quantile behaves nearly the same as sample quantile (the quantile computed using all data observed so far; the qM sorted observation for q-th quantile with M observed data). Asymptotically, SA estimation and sample quantile are indistinguishable. However, due to the use of derivative information, SA estimation can be sensitive to data order or the particular distribution during intermediate updates. This presents challenges for cellular network monitoring. For example, cellular providers may monitor active flow size quantiles; if the quantiles change abruptly, this may signal network anomaly events. If it is due to inaccuracy of SA at intermediate updates, this can trigger false alarms. False positives may just be annoying to network administrators. However, false negatives can be very serious as active measures are not taken to address the anomalies.

For multiple quantiles, current algorithms [8, 2] derive each quantile estimate in isolation. Due to the multiple local approximations at quantiles of stochastic approximation before convergence, it is quite possible that multiple quantiles estimated independently do not obey the *monotone property*. That is, the value of 90% quantile can be smaller than the value of 80% quantile. As we show in our evaluation, these non-monotone cases do happen in practice, and it can be as high as 20% using our data set. This makes continuous tracking of quantiles difficult. In addition, if the underlying distribution is not stationary, this will pose additional challenges for SA based estimation algorithms.

In this paper, we make the following contributions. First, we proposes a stochastic approximation scheme that estimates multiple quantiles incrementally over time. At any given time, the monotone property of quantiles are maintained. Second, we have validated our algorithm using real cellular provider data. Our scheme only needs to keep track of quantiles of interest and has no additional memory requirement. In contrast, non-SA based algorithm's space requirement depends on which quantile is estimated (more samples are needed for extreme quantiles). Our scheme is extremely light weight. To the best of our knowledge, our algorithm is the first stochastic approximation algorithm that maintains the monotone property during incremental updates.

The rest of this paper is organized as follows. In Section 2, we briefly review stochastic approximation for a single quantile. We present our *monotoneSA* algorithm in Section 3. In Section 4, we evaluate our algorithms using real network data.

2. STOCHASTIC APPROXIMATION: A PRELIMINARY

Let $\{x_t\}$ be an incoming data stream with a distribution $\mathcal{F}_t$. Suppose $\mathcal{F}_t(\cdot)$ is a continuous distribution with positive derivatives on its domain. In this section, we give some preliminaries on incremental quantile estimation of $\{x_t\}$ using stochastic approximation. We present the algorithm for a single quantile, and discuss issues arise for multiple quantile estimations.

2.1 A Single Quantile

A stochastic approximation (SA) for online incremental quantile estimate is proposed by [7, 8], under the assumption that $\mathcal{F}_t = \mathcal{F}$, that is, the data distribution does not change over time. Let p be a probability whose quantile is of interest, and let θ_t be the true quantile of $\mathcal{F}_t$ w.r.t. p. The SA method for estimating θ is as follows. Let S_{t-1} be the quantile estimate up to time $t-1$. With the arrival of the tth observation x_t, the SA quantile estimate is updated by

$$S_t = S_{t-1} + a_t(p - I(x_t \leq S_{t-1})), \qquad (1)$$

where $a_t > 0$ is a pre-defined sequence of positive numbers, and $I(\cdot)$ is the indicator function. Let $f_t = \mathcal{F}_t'(\theta) > 0$ be the derivative of $\mathcal{F}_t$ (density) at the true quantile θ_t. We can write a_t in the form of

$$a_t = f_t^{-1} w_t, \qquad (2)$$

where w_t is refereed as the weight associated with data x_t.

When the data distribution is stationary, i.e., $\mathcal{F}_t = \mathcal{F}$ and hence the density $f_t = f$, the following two lemmas give the property of the SA quantile estimate in (1) [7, 8].

LEMMA 1. *If $\sum_t w_t = \infty$, and $\sum_t w_t^2 < \infty$, the SA estimate will converge with probability 1 to θ ([7]).*

LEMMA 2. *When $w_t = \alpha t^{-1}$ (thus satisfy the convergence condition in Lemma 1), then $\sqrt{t}(S_t - \theta)$ will converge to a normal distribution with mean zero and a fixed variance. In addition, the variance will be minimized when*

$$w_t = t^{-1}, \quad (t^{-1} \textit{ weights}), \qquad (3)$$

with a variance σ^2/f^2, where σ is the variance of the stationary distribution.

Of course, in practice, since the derivative f is not known exactly, it is estimated from data [8]. However, if the true value f is close to 0 such as at the tails, the estimate may become unstable.

When the data distribution $\mathcal{F}_t$ changes over time, the diminishing weight t^{-1} is no longer appropriate. In this case, to track the true quantile θ_t w.r.t p, Chen et al.[2] suggested setting w_t in Eq. 2 by

$$w_t = w, \quad \text{(constant weights)}, \qquad (4)$$

where $w > 0$ is a fixed constant. At the same time, the derivative f_t is estimated from data using an exponentially weighted average with the same weight w.

In fact, the use of constant weights (4) has been strongly suggested by Chen et al. [2] even for stationary data, as their simulation results suggest that it gives a good estimate and is less prone to bad initial values. This is supported by the following weak convergence result of the quantile estimate S_t for stationary data [5].

LEMMA 3. *$S_t - \theta$ converges in distribution to a random variable with mean 0 and fixed variance, as $t \to \infty$.*

In fact, one can further reduce the variability and hence improve the accuracy of the quantile estimates by averaging S_t ([5]).

2.1.1 An Alternative Interpretation

Before we move on, we first give an alternative interpretation of the stochastic algorithm that will be used later to motivate us to improve some of its deficiencies. To simplify the explanation, let's assume that $f_t = \mathcal{F}_t'(\theta_t)$ is known.

Given observations up to $t-1$, S_{t-1} is the approximated quantile for probability p, i.e. $P(x \leq S_{t-1}) \approx p$. Now with the observation x_t and its associated weight w_t, the probability $P(x \leq S_{t-1})$ can be updated by

$$P(x \leq S_{t-1}) \approx (1 - w_t)p + w_t I(x_t \leq S_{t-1}). \qquad (5)$$

Now with the distribution derivative f_t, we can approximate $\mathcal{F}_t$ locally at $(S_{t-1}, (1-w_t)p + w_t I(x_t \leq S_{t-1}))$ using a linear function with slope f_t, i.e.,

$$\hat{\mathcal{F}}_t(x) \approx (1 - w_t)p + w_t I(x_t \leq S_{t-1})) + (x - S_{t-1})f_t.$$

Now setting this equals to p, we obtained the SA quantile estimate in (1), i.e.

$$S_t = S_{t-1} + w_t/f_t(p - I(x_t \leq S_{t-1})),$$

What we have shown above is that the SA quantile estimate is essentially derived from a local approximation of the $\mathcal{F}_t$ at the quantile point S_t. This local approximation is extremely simple (a linear function), and is incrementally updated with every new arrival. Such a local approximation is very different from many of the proposed approaches that try to build a global approximation using data summaries. Because of its simplicity, there is essentially no memory requirement even for a tail quantile.

2.2 Issues with Multiple Quantiles

To obtain online quantiles estimates for more than one probabilities, a naive method is to run the SA algorithm (1) for each of the probabilities in isolation. However, this simple method can lead to a violation of the monotone property of the quantile estimates. That is, if $p(1) < p(2)$, there is no guarantee that the estimated quantile for $p(1)$ is strictly less than $p(2)$. We illustrate using the case of two quantiles.

Let $p(1) < p(2)$ be the two probabilities whose quantiles are of interest to us. Let the corresponding SA quantile estimates given data up to $t-1$ be $S_{t-1}(1)$ and $S_{t-1}(2)$. And suppose the densities of distribution $\mathcal{F}_t$ are known at the true quantiles. Denote them by $f_t(1), f_t(2)$ for $p(1), p(2)$ respectively. Now with the new observation x_t, we update the quantile estimates by

$$S_t(1) = S_{t-1}(1) + w_t/f_t(1)(p(1) - I(x_t \leq S_{t-1}(1))), \quad (6)$$

$$S_t(2) = S_{t-1}(2) + w_t/f_t(2)(p(2) - I(x_t \leq S_{t-1}(2))), \quad (7)$$

If $S_{t-1}(1) < S_{t-1}(2)$, given the relations above, we cannot guarantee that $S_t(1) < S_t(2)$. Figure 1 gives an example of this violation of the monotone property, using the interpretation of the SA algorithm that is given in Section 2.1.1.

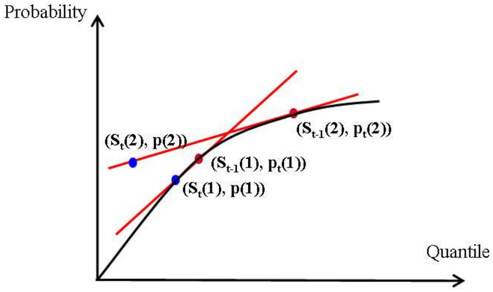

Figure 1: An illustration of the out-of-orderness problem of SA algorithm for multiple quantiles.

In Figure 1, the two red points represent the quantile estimates at time $t-1$, whose cumulative probabilities, denoted by $p_t(1)$ and $p_t(2)$, are already updated upon arrival of the new data x_t using (5). The two red lines represent the local approximation of the distribution at quantile points $(S_{t-1}(1), p_t(1))$ and $(S_{t-1}(2), p_t(2))$. However, in this case, the slope at quantile point at $S_{t-1}(2)$ is significantly less than the slope at $S_{t-1}(1)$. Now the SA updates on the two linear lines result in a violation of the monotone property.

The fundamental reason for the out-of-order phenomenon is because we are using multiple local approximations to the distribution, not a global one. Although the local approximation is good enough for the neighborhood, it is not a good global approximation. Obviously, such a situation would not occur when the quantile estimates are very close to the actual values. For stationary data where the biggest adjustment occur at the beginning this means the out-of-orderness maybe restricted to early iterations and thus rare. However, this may not be the case when the the data is non-stationary, and hence there is constant adjustment of the quantile estimates.

Obviously, a quick fix to the monotonicity problem is the following. Suppose that the SA estimates for the K quantiles w.r.t. $p = (p(1), \ldots, p(K))$ are $S_t = (S_t(1), \ldots, S_t(K))$. Then to obtain monotone estimates using S_t, for $1 \leq i \leq K$, we can set

$$\tilde{S}_t(i) = \max_{j \leq i} S_t(j), \quad \text{or} \quad \tilde{S}_t(i) = \min_{j \geq i} S_t(j).$$

It is easy to see that $\tilde{S}_t(i)$ is monotone using either approaches. However, a drawback of this method is when the monotonicity fails for S_t (which is the problem that we are trying to fix), then the new estimates $\tilde{S}_t$ will have ties which does not seem desirable.

3. THE MONOTONESA ALGORITHM

Let $\{x_t\}$ be an incoming data stream with a distribution $\mathcal{F}_t$, where $\mathcal{F}_t(\cdot)$ is strictly monotone, and has positive derivatives on its domain. We are interested in an online algorithm that is able to track the the quantiles of $\mathcal{F}_t$ w.r.t. a set of K probabilities $p = (p(1), p(2), \ldots, p(K))$. Denote that the quantile estimates are $S_t = (S_t(1), S_t(2), \ldots, S_t(K))$. We require that the quantile estimates are strictly monotone, i.e. $S_t(1) < S_t(2) < \ldots < S_t(K)$.

In this section, we present our *monotoneSA* algorithm for quantile estimation. Our algorithm relies on an incremental approximation $\hat{\mathcal{F}}_t$, to the distributional function $\mathcal{F}_t$, upon new data arrivals. The quantile estimates are just the quantiles of the approximation $\hat{\mathcal{F}}_t$ w.r.t. probabilities p. Our quantile estimates are not out of order since the approximation is a globally increasing function constructed from local approximations at each of the estimated quantiles.

3.1 The Algorithm

The algorithm goes as follows. Suppose that from some initial samples, we obtain an initial estimate of the distribution function, denoted by $\hat{\mathcal{F}}_0$, From it, we obtain the initial quantile estimates S_0 w.r.t. probabilities $\{p(i), i = 1, \ldots, K\}$, and their respective derivatives $f_0 = (f_0(1), \ldots, f_0(K))$ of $\hat{\mathcal{F}}_0$ at the quantiles. Now at each time t, with the new arrival x_t, we shall update the distribution approximation by

$$\hat{\mathcal{F}}_t(x) = (1 - w_t)\hat{\mathcal{F}}_{t-1}(x) + w_t I(x \geq x_t), \quad (8)$$

where w_t is the weight associated with x_t, and $I(\cdot)$ is the indicator function. Evaluating the above at the estimated quantile points at time $t-1$, and using the fact that for $1 \leq i \leq K$, $\hat{\mathcal{F}}_{t-1}(S_{t-1}(i)) \approx p(i)$, we have

$$\hat{\mathcal{F}}_t(S_{t-1}(i)) \approx (1 - w_t)p(i) + w_t I(S_{t-1}(i) \geq x_t).$$

Denote the updated probability by $p_t(i)$, i.e.,

$$p_t(i) = (1 - w_t)p(i) + w_t I(S_{t-1}(i) \geq x_t), \quad \text{(probability update)}, \quad (9)$$

and let $p_t = (p_t(1), p_t(2), \ldots, p_t(K))$. With the updated quantile probabilities S_{t-1} and derivative estimates f_{t-1}, we use linear interpolation to construct a globally increasing function as an approximation of $\hat{\mathcal{F}}_t$ such that at in the neighborhood of each $S_{t-1}(i)$ it is a linear function with the slope specified by $f_{t-1}(i)$, and these linear segments around the quantile points are extended as much as possible under the constraints of monotonicity of the interpolation function.

For each $1 \leq i \leq K-1$, denote

$$\text{right}_t(i) = (S_{t-1}(i) + \Delta_t(i), p_t(i) + f_{t-1}(i)\Delta_t(i)), \qquad (10)$$

which is the point right to the quantile point $(S_{t-1}(i), p_t(i))$ with a slope $f_{t-1}(i)$, and

$$\text{left}_t(i+1) = (S_{t-1}(i+1) - \Delta_t(i), p_t(i+1) - f_{t-1}(i+1)\Delta_t(i)), \qquad (11)$$

which is the point left to the quantile point $(S_{t-1}(i+1), p_t(i+1))$ with a slope $f_{t-1}(i+1)$. Then we obtain a value $\Delta_t(i) > 0$ such that these two data points are non-decreasing in their coordinates, i.e.,

$$S_{t-1}(i) + \Delta_t(i) \leq S_{t-1}(i+1) - \Delta_t(i),$$

and

$$p_t(i) + f_{t-1}(i)\Delta_t(i) \leq p_t(i+1) - f_{t-1}(i+1)\Delta_t(i).$$

It is obvious that these indicate that

$$\Delta_t(i) \leq \min\left(\frac{S_{t-1}(i+1) - S_{t-1}(i)}{2}, \frac{p_t(i+1) - p_t(i))}{f_{t-1}(i) + f_{t-1}(i+1)}\right). \qquad (12)$$

In our algorithm, we choose $\Delta_t(i)$ to the be maximum possible value on the right hand side.

Now the distribution approximation $\hat{\mathcal{F}}_t$ is obtained by connecting the quantile point $(S_{t-1}(1), p_t(1))$, the points $right_t(i)$, $left_t(i)$, $i = 1, \ldots, K-1$, and the quantile point $(S_{t-1}(K), p_t(K))$, and finally extending the piece-wise function beyond the two boundary points so that it reaches the extreme y-values 0 and 1. The new quantile estimates S_t is just the quantiles that corresponds to the updated distribution function, i.e., for each $i, 1 \leq K$,

$$\hat{\mathcal{F}}_t(S_t(i)) = p_i. \quad \text{(quantile update)}. \qquad (13)$$

To derive the derivative estimate f_t at estimated quantiles $S_t(i)$, we use a similar approach as in [8, 2], i.e.,

$$f_t(i) = (1 - w_t)f_{t-1}(i) + w_t(2c)^{-1}I(|x_t - S_t(i)| \leq c), \qquad (14)$$

where $c > 0$ is a tunable parameter representing the window size around $S_t(i)$ used to estimate the derivatives. In our implementation, we use a value of c that is a fraction of the estimated inter-quantile range, and the window size c is the same for all quantiles. It may be advantageous to choose window sizes that are not uniform across all quantiles, but we have not explored that.

Our algorithm can be summarized Figure 2.

At the ith iteration, the line segment around the ith quantile point, $(S_{t-1}(i), p_t(i))$, determined by the neighborhood points $left_t(i)$ and $right_t(i)$ is a line with slope $1/f_{t-1}(i)$. This is the same approximation that the simple SA algorithm uses. (see Section 3.1.1). Therefore, if the solution to the quantile updating equation, $\hat{\mathcal{F}}_t(S_t(i)) = p_i$ (Equation 13) occurs in this line segment, then we would obtain the same SA quantile estimates as in (1). Obviously, such an situation occurs when the adjustment from p to p_t upon the new data arrival x_t is sufficiently small.

As was done in [2], we can also update the quantile estimation for every batch of M new data arrival. In this case, the only change here is that we do not perform the steps in 2-4 in the algorithm (Figure 2) until we have M data arrivals.

monotoneSA Algorithm

```
0. Initialization:
     Let the initial quantile estimate be $S_0$ and
     density estimate be $f_0$
     Let $x_t$ be an incoming data at time $t \geq 1$

1. Probability update for quantile $S_{t-1}$:
     $p_t(i) = (1 - w_t)p(i) + w_t I(S_{t-1}(i) \geq x_t), i = 1, \ldots, K$

2. Compute $\Delta_t$ using (12)

3. Interpolate to construct an approximation to $\mathcal{F}_t$, $\hat{\mathcal{F}}_t$:
     connecting the first quantile point $(S_{t-1}(1), p_t(1))$,
     the points $right_t(i)$ and $left_t(i)$, $i = 1, \ldots, K-1$
     and the last quantile point $(S_{t-1}(K), p_t(K))$.
     Extending this piece-wise function beyond the two
     boundary points to reach extreme y-values 0 and 1.

4. Update the quantile estimate $S_t$:
     $1 \leq i \leq K$, $\hat{\mathcal{F}}_t(S_t(i)) = p_i$

5. Update the density estimate $f_t$:
     $f_t(i) = (1 - w_t)f_{t-1}(i) + w_t(2c)^{-1}I(|x_t - S_t(i)| \leq c)$
```

Figure 2: Algorithm for incremental tracking of multiple quantiles

3.1.1 An Illustration

Figure 3 give an illustration of our algorithm for finding two probability quantiles. The black smooth curve represent the hypothetical smooth approximation of data distribution $\mathcal{F}_t$, and the red curve is our piece-wise linear approximation using the estimated quantile points $(S_{t-1}(1), p_t(1))$ and $(S_{t-1}(2), p_t(2))$ and their respective derivatives. The two figures correspond to different scenarios for the value of $\Delta_t(1)$ defined in (12): the top figure illustrates the case when $\Delta_t(1)$ takes the second value in the equation, and bottom figure illustrates the case when $\Delta_t(1)$ takes the first value in the equation.

3.2 The Choice of Weights

We consider two types of data: data with stationary distributions and data with non-stationary distributions. We also consider two types of weights: $w_t = 1/t$ or constant $w_t = w$.

When $\mathcal{F}_t$ is stationary, it has been shown that the simple stochastic approximation (SA) algorithm will lead to convergence for both kinds of weights [5]. For $1/t$, the convergence is to the true quantile in probability 1, and for constant step sizes, the convergence is in distribution to a random variable with mean of the true quantile. We extend the convergence result for our modified SA algorithm with $1/t$ weights.

THEOREM 1. *As $t \to \infty$, our Algorithm (Figure 2) with $1/t$ weights will converge almost surely to true quantiles. Furthermore, results in Lemma 2 hold true for our quantile estimates.*

The convergence can be intuitively understood as follows. At each iteration, our algorithm will result in a quantile update similar to that of (1), but now with a more complicated $\{a_t\}$ sequence for each quantile. In fact it can be shown that our $\{a_t\}$ sequence is asymptotically the same that of

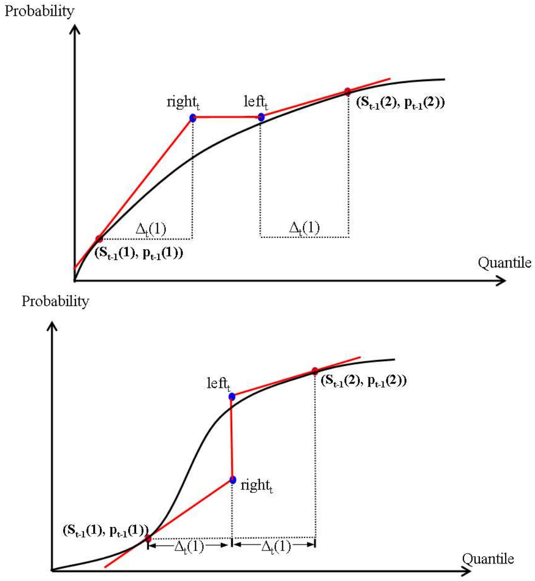

Figure 3: Illustrations of the algorithm for two quantile points with probability $p(1)$ and $p(2)$. The two figures correspond to different cases of $\Delta_t(1)$ values in 12.

the simple SA algorithm. Therefore, with probability 1, our estimates will be the same as the conventional ones.

For fixed weights, we do not have a formal proof but it has been observed empirically that the same weak convergence result holds as in the case of the simple SA algorithm (Lemma 4). For non-stationary data, the diminishing weight $1/t$ is no longer valid. In this case, we have to choose a fixed weight to be able to track the quantiles over time.

3.3 Further Discussions

We have several comments here. First, our quantile updating scheme is based on an incremental distribution approximation by interpolating at the updated quantile points. Local to the quantile points, our approximation is the same linear function as in the case of straightforward SA (see Section 3.1.1). Yet globally, it is an increasing function. This interpolation scheme is designed to be simple so that the updated quantiles can be easily computed, and at the same time lead to good approximations as in the case of straightforward SA (Theorem 1). Our approach here opens up the possibility of using other more elaborate interpolation (or approximation) schemes given the local approximations at the quantiles. For extreme tails, we can also use an asymptotic model to overcome some of the instabilities of SA (due to a very small derivative, see Lemma 2). However, we should be careful in choosing a such interpolation or approximation model since it may lead to biases in estimates. For example, our experimental studies have shown that if we use the linear interpolation by connecting the quantile points directly (without using the local derivatives), the quantile estimates converges for stationary data but with a bias.

Second, on computational cost. Since the distribution approximation is piecewise linear, it is trivial to find the quantile points of the function for updating (see Equation 13). All we need to do is to first find out which line segment each $p(i)$ falls into, and then solve $p(i)$ for that line segment.

Third, the derivative f_t is a vector of estimated derivatives (density). It is not crucial to obtain exact values of the derivatives. In fact, if we replace f_t by a vector of fixed positive constant, that the quantile estimates derived from the algorithm still give good approximations. However, it is more efficient to use a value of f_t that is close to the actual derivatives of the distribution function since the quantile estimates will stabilize faster around the true value.

Last, we have assumed that $\mathcal{F}_t(\cdot)$ is a strictly increasing continuous function. For a discrete distribution, we can not apply the algorithm unmodified since the derivative estimate may become infinite. However, a simple fix is to add a small random noise to the data, which can be chosen in a data dependent fashion. Obviously, by doing so, we may introduce a small bias for the estimated quantiles.

4. EXPERIMENTAL RESULTS

In this section, we demonstrate our proposed algorithm using data collected from one 3G wireless network of a US national-wide provider using a proprietary measurement product. For privacy consideration, the data shown here are sanitized by a multiplication with a unspecified constant.

The data are collected for wireless network security purposes, and many reflect different kinds of faults (such as battery attacks, worm outbreaks, port scans, number of flooded mobiles etc). There are 18 event counts over a period of a week under our study, and the counts are collected every minute. The behavior of these event counts varies: some show a clear daily trend but some show a subtle trend, and some have a large variability but others vary little.

Since these are network counts, and they are discrete. To break the ties, we add a small random noise to the data. The random noise is generated from a uniform distribution on the interval [0,1]. We compare three kinds of quantile estimates: *SimpleSA* (simple stochastic approximation method that treats each quantile separately), $MonotoneSA$ (Algorithm in Figure 2), and *movingQuantile* (empirical quantile estimates using data in the left neighborhood). The probabilities under consideration is $p = 0.5, 0.7, 0.9, 0.95, 0.99$. Since the data is non-stationary, we choose a fixed weight $w = 0.05$ for both $SimpleSA$ and $MonotoneSA$, and use a window size of 100 for the moving quantiles. The window size is chosen to be compatible with the weight 0.05 given to the data[1]; it is probably also the maximum possible value since we are also interested in the 99% quantile.

On average, the out-of-orderness of *SimpleSA* occurs for 5% of the time. However, there is a wide variation between different count series. For example, for two time series, this occurs more than 20% of the time, but for many others, it occurs about 1% of time. Close inspection shows that most of the out-of-orderness occur at the tails for $p = 0.95$ and $p = 0.99$.

Figure 4 and 5 shows the result for two probability quantiles 0.5 and 0.95, for two of these network event counts: number of observed P2P mobiles (subscribers who run P2P applications), and number of high signaling subscribers (subscribers who generate high number of control messages such as call setup). It is interesting to see that for both datasets, the two estimates are almost indistinguishable.

[1] The equivalent window size for an exponentially weighted moving average scheme with weights w is $2/w$.

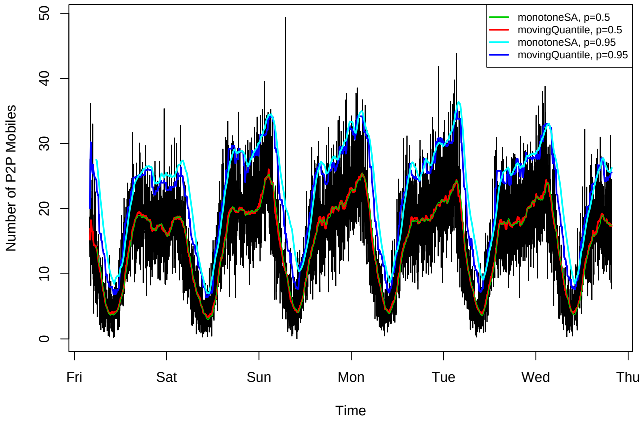

Figure 4: Quantile Estimates for the time series of number of observed P2P mobiles observed in a wireless network. The $MonotoneSA$ estimates shown are the averaged estimates using the left neighborhood of size 100.

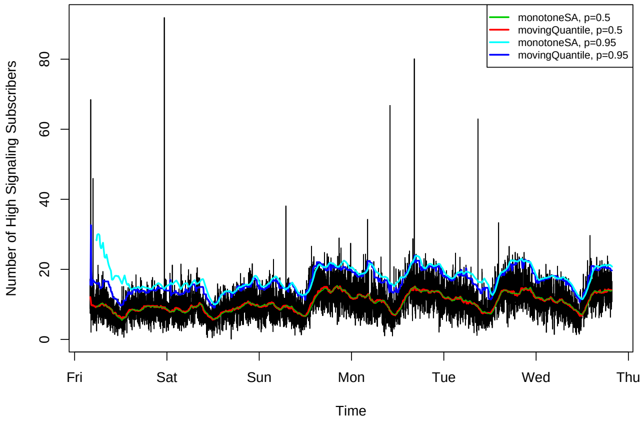

Figure 5: Quantile Estimates for the time series of number of high signaling subscribers observed in a wireless network. The $MonotoneSA$ estimates shown are the averaged estimates using the left neighborhood of size 100.

5. CONCLUSION AND FUTURE WORK

Tracking quantiles for evolving continuous streams is very important for real-time monitoring of cellular networks. We present a stochastic approximation technique that enables accurate tracking of multiple quantiles at the same time. Our space requirement is constant. Our technique enables incremental updates. This avoids the delay and computational overhead of periodic computation. Our algorithm also maintains the important monotone property.

We evaluate our algorithm using real cellular provider network data. Our evaluation shows that our algorithm converges very fast and can accurately track changing quantiles over time.

For future work, we are interested in extending our algorithm to the distributed setting.

6. REFERENCES

[1] A. Arasu and G. S. Manku. Approximate counts and quantiles over sliding windows. In *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 286–296, New York, NY, USA, 2004. ACM.

[2] F. Chen, D. Lambert, and J. C. Pinheiro. Incremental quantile estimation for massive tracking. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000.

[3] A. C. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 389–398, 2002.

[4] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. How to summarize the universe: dynamic maintenance of quantiles. In *VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases*, pages 454–465. VLDB Endowment, 2002.

[5] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 1997.

[6] X. Lin, H. Lu, J. Xu, and J. X. Yu. Continuously maintaining quantile summaries of the most recent n elements over a data stream. In *ICDE '04: Proceedings of the 20th International Conference on Data Engineering*, page 362, Washington, DC, USA, 2004. IEEE Computer Society.

[7] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[8] L. Tierney. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal of Scientific and Statistical Computing*, 4(4), 1983.