Compressed Counting for Estimating Frequency Moments and Entropy of Data Streams

Ping Li
Department of Statistical Science
Faculty of Computing and Information Science
Cornell University
Ithaca, NY 14853
pingli@cornell.edu

Abstract

Modern massive data are often dynamic and modeled as **data streams**. According to the Turnstile model, the input stream $a_t = (i_t, I_t)$, $i_t \in [1,\ D]$ arriving sequentially describes the underlying signal A, meaning $A_t[i_t] = A_{t-1}[i_t] + I_t$, where the increment I_t can be either positive (insertion) or negative (deletion). The length D could be as large as 2^{64}.

There are numerous interesting and challenging problems in data stream computations. One heavily studied problem is to efficiently compute the α-th frequency moment $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$, using small storage space. A closely related summary statistic is the **Shannon entropy** $H = -\sum_{i=1}^{D} \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}}$, where $F_{(1)}$, the first moment, is the sum of the data stream. H can be estimated using $F_{(\alpha)}$ near $\alpha = 1$, based on the fact that $\frac{1-x^\alpha}{1-\alpha} \to x \log x$, as $\alpha \to 1$.

When $0 < \alpha \leq 2$, well-known algorithms based on **symmetric stable random projections** could estimate moments using only $O\left(1/\epsilon^2\right)$ space, to guarantee that the relative error is within a $1 \pm \epsilon$ factor. Unfortunately, to accurately estimate H, ϵ has to be extremely small, e.g., $\epsilon < 10^{-5}$.

About two years ago, I had the conjecture that, since the first moment $F_{(1)} = \sum_{i=1}^{D} A_t[i] = \sum_{s=1}^{t} I_s$, can be computed exactly using just one counter (to sum the increments and decrements), there might exist an intelligent counting system near $\alpha = 1$ whose complexity will decrease continuously as $\alpha \to 1$. Then I found that using **maximally-skewed stable random projections**, also named as **Compressed Counting**, could possibly achieve this goal. I developed estimators based on *geometric mean* and *harmonic mean* and showed that their variances approached zero at the rate of $O(\Delta)$, where $\Delta = 1 - \alpha$. In addition, I proved that the complexity is $O\left(1/\epsilon\right)$ instead of the previously believed $O\left(1/\epsilon^2\right)$ bound, near $\alpha = 1$. While this is a very significant improvement, unfortunately $O\left(1/\epsilon\right)$ is still too large for entropy estimation if $\epsilon < 10^{-5}$.

Very recently, I found an interesting estimator in the form of $\frac{1}{\Delta^\Delta}\left[\frac{k}{\sum_{j=1}^{k} x_j^{-\alpha/\Delta}}\right]^\Delta$, where x_j's are the projected data and k is the sample size. This estimator has variance proportional to $\Delta^2\left(3 - 2\Delta\right)$, approaching zero extremely fast. We prove its complexity is $O\left(1\right)$ near $\alpha = 1$. This is another very large improvement and leads to highly practical algorithms for entropy estimation. This new estimator is also numerically very stable even for Δ as small as 10^{-10}.