Compressed Counting and Random Projections in Data Stream Computations and Entropy Estimation

Ping Li
Department of Statistical Science
Cornell University

Abstract

Many dynamic data, e.g., network traffic data, can be 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_t,

$$A_t[i_t] = A_{t-1}[i_t] + I_t,$$

where the increment I_t can be either positive (insertion) or negative (deletion). Here $D = 2^{64}$ is possible if each $A_t[i]$ corresponds to an IP address. One important task is to measure summary statistics of A_t in real-time (e.g., for detecting anomaly events such as DDoS attacks). Useful summary statistics include the αth frequency moment $F_{(\alpha)}$, and the Shannon entropy H:

$$F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^{\alpha}, \qquad\qquad H = -\sum_{i=1}^{D} \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}}.$$

It is known that H can be approximated by certain functions of $F_{(\alpha)}$ (such as Tsallis entropy or Rényi entropy) by letting $\alpha \rightarrow 1$. Note that computing $F_{(\alpha)}$ exactly requires a counting system with $D = 2^{64}$ counters (which is highly impractical) if $\alpha \neq 1$. However, when $\alpha = 1$, only one counter is needed because $F_{(1)} = \sum_{i=1}^{D} A_t[i] = \sum_{s=1}^{t} I_s$.

Compressed Counting (CC) has been proposed for efficiently and accurately approximating $F_{(\alpha)}$, based on the idea of *maximally-skewed stable random projections*. CC captures the interesting observation that the first moment $F_{(1)}$ is trivial but $F_{(\alpha)}$ is challenging in general. For example, one proposed estimation algorithm of CC exhibits estimation variance (error) proportional to $\Delta = |\alpha - 1|$, which approaches zero as $\alpha \rightarrow 1$ (i.e., $\Delta \rightarrow 0$). Therefore a natural application of CC is to approximate the Shannon entropy using $F_{(\alpha)}$ by letting $\alpha \rightarrow 1$.

In addition, we have also proved that, the sample complexity of CC is $O\left(\frac{1}{\log(1+\epsilon)} + \frac{2\sqrt{\Delta}}{\log^{3/2}(1+\epsilon)} + o\left(\sqrt{\Delta}\right)\right)$, as $\Delta \rightarrow 0$. In other words, in the neighborhood of $\alpha = 1$, the complexity of CC is essentially $O\left(1/\epsilon\right)$ instead of $O\left(1/\epsilon^2\right)$; the latter is the well-known large-deviation complexity bound.