Stochastic Data Streams

S. Muthukrishnan

Google Inc.

The classical data stream problems are now greatly understood [1]. This talk will focus on problems with stochastic (rather than deterministic) streams where the underlying physical phenomenon generates probabilistic data or data distributions. While basic streaming problems are now getting reexamined with stochastic streams [2,3], we will focus on a few novel problems.

1. Estimating tails. Each item in the data stream random variable $X_i, 1 \leq i \leq n$, $X_i\{0,1\}$, identically distributed such that $E(X_i) = 0$, $E(X_i^2) = \sigma^2 > 0$ but bounded. The query is to estimate $\Pr[\sum X_i \leq c]$. This is a special case of more general probabilistic streams where each X_i may be drawn from a different distribution, and we wish to answer this query using sublinear space. We provide an algorithm using the well known result below (joint work with Krzysztof Onak).

Theorem 1. (Berry-Esseen Theorem*) Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with $E(X_i) = 0$, $E(X_i^2) = \sigma^2 > 0$, and $E(|X_i|^3) = \rho < \infty$. Also, let $Y_n = \sum_i X_i/n$ with F_n the cdf of $Y_n\sqrt{n}/\sigma$, and Φ the cdf of the standard normal distribution. Then there exists a positive constant C such that for all x and n, $|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$.*

2. Max Triggering. The data stream is a sequence of independent random variables $x_1, x_2 \ldots$. The problem is to determine a point τ and trigger an alarm based on x_τ. The goal is to trigger the alarm when x_τ is the maximum possible over a length n. Define $M = E(max_{i=1}^{i=n} x_i)$, the expected value of the maximum. This is target for any streaming algorithm. We will be able to show simply that there exists a streaming algorithm that finds x_τ such that $E(x_\tau) \geq M/2$. Our algorithm will be a suitable implementation of the stopping rule that gives the well known Prophet inequalities [4].

References

1. Muthukrishnan, S.: Data Streams: Algorithms and Applications. In: Foundations and Trends in Theoretical Computer Science. NOW publishers (2005); Also Barbados Lectures (2009), `http://www.cs.mcgill.ca/~denis/notes09.pdf`
2. Jayram, T.S., McGregor, A., Muthukrishnan, S., Vee, E.: Estimating statistical aggregates on probabilistic data streams. ACM Trans. Database Syst. 33(4) (2008)
3. Cormode, G., Garofalakis, M.: Sketching probabilistic data streams. SIGMOD, 281–292 (2007)
4. Samuel-Cahn, E.: Comparisons of optimal stopping values and prophet inequalities for independent non-negative random variables. Ann. Prob. 12, 1213–1216 (1984)

R. Královič and D. Niwiński (Eds.): MFCS 2009, LNCS 5734, p. 55, 2009.