

April 2009

**MADALGO seminar by Jelani Nelson, Massachusetts Institute of Technology (MIT)**

## **Revisiting Norm Estimation in Data Streams**

Abstract:

The problem of estimating the  $p$ th moment  $F_p$  ( $p$  nonnegative and real) in data streams is as follows. There is a vector  $x$  which starts at 0, and many updates of the form  $x_i \leftarrow x_i + v$  come sequentially in a stream. The algorithm also receives an error parameter  $0 < \epsilon < 1$ . The goal is then to output an approximation with relative error at most  $\epsilon$  to  $F_p = \|x\|_p^p$ .

Previously, it was known that polylogarithmic space (in the vector length  $n$ ) was achievable if and only if  $p \leq 2$ . We make several new contributions in this regime, including:

- (\*) An optimal space algorithm for  $0 < p < 2$ , which, unlike previous algorithms which had optimal dependence on  $1/\epsilon$  but sub-optimal dependence on  $n$ , does not rely on Nisan's PRG.
- (\*) A near-optimal space algorithm for  $p = 0$  with optimal update and query time.
- (\*) A near-optimal space algorithm for the "distinct elements" problem ( $p = 0$  and all updates have  $v = 1$ ) with optimal update and query time.
- (\*) Improved  $L_2 \rightarrow L_2$  dimensionality reduction in a stream.
- (\*) New 1-pass lower bounds to show optimality and near-optimality of our algorithms, as well as of some previous algorithms (the "AMS sketch" for  $p = 2$ , and the  $L_1$ -difference algorithm of Feigenbaum *et al.*).

As corollaries of our work, we also obtain a few separations in the complexity of moment estimation problems:  $F_0$  in 1 pass vs. 2 passes,  $p = 0$  vs.  $p > 0$ , and  $F_0$  with strictly positive updates vs. arbitrary updates.

Joint work with:

Daniel Kane, Harvard University  
David Woodruff, IBM Almaden.