May 2012

MADALGO seminars by Hossein Jowhari, Aarhus University

Near-optimal space bounds for L_p samplers

Abstract:

In this talk, I will present near-optimal space bounds for L_p-samplers in data streams. Given a stream of additions and subtractions to the coordinates of an underlying vector x, an L_p sampler is a streaming algorithm that reads the updates once and takes a sample coordinate with probability proportional to the L_p distribution of x. More precisely, the i-th coordinate is picked with the chance corresponding to the weight |x_i|^p. Here I will present an eps-relative error L_p sampler requiring roughly O(eps^{-p}log^2n) space for p in (0,2). This result improves the previous bounds by Monemizadeh and Woodruff (SODA 2010) and Andoni, Krauthgamer and Onak (FOCS 2011).

As an application of these samplers, an upper bound will be demonstrated for finding duplicates in data streams using L_1 samplers.
In case the length of the stream is long enough, our L_1 sampler leads to a O(log^2 n) space algorithm for this problem, thus improving the previous bound due to Gopalan and Radhakrishnan.

If time permits, I also show an Omega(log^2 n) lower bound for sampling from 0,+-1 vectors. This matches the space of our sampling algorithms for constant eps>0. These bounds are obtained using reductions from the communication complexity problem of Augmented Indexing.

Joint work with Mert Saglam and Gabor Tardos in PODS 2011.