TOPICS IN MATRIX SAMPLING ALGORITHMS

By

Christos Boutsidis

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Petros Drineas, Thesis Adviser
Kristin P. Bennett, Member
Sanmay Das, Member
Malik Magdon-Ismail, Member
Michael W. Mahoney, Member
Mark Tygert, Member

Rensselaer Polytechnic Institute
Troy, New York

May 2011
(For Graduation May 2011)

ABSTRACT

We study three fundamental problems of Linear Algebra, lying in the heart of various Machine Learning applications, namely: (i) Low-rank Column-based Matrix Approximation, (ii) Coreset Construction in Least-Squares Regression, and (iii) Feature Selection in k-means Clustering. A high level description of these problems is as follows: given a matrix A and an integer r, what are the r most "important" columns (or rows) in A? A more detailed description is given momentarily.

1. Low-rank Column-based Matrix Approximation. We are given a matrix A and a target rank k. The goal is to select a subset of columns of A and, by using only these columns, compute a rank k approximation to A that is as good as the rank k approximation that would have been obtained by using all the columns.

2. Coreset Construction in Least-Squares Regression. We are given a matrix A and a vector $\mathbf{b}$. Consider the (over-constrained) least-squares problem of minimizing $\|\mathrm{A}\mathbf{x} - \mathbf{b}\|_2$, over all vectors $\mathbf{x} \in \mathcal{D}$. The domain $\mathcal{D}$ represents the constraints on the solution and can be arbitrary. The goal is to select a subset of the rows of A and $\mathbf{b}$ and, by using only these rows, find a solution vector that is as good as the solution vector that would have been obtained by using all the rows.

3. Feature Selection in K-means Clustering. We are given a set of points described with respect to a large number of features. The goal is to select a subset of the features and, by using only this subset, obtain a k-partition of the points that is as good as the partition that would have been obtained by using all the features.

We present novel algorithms for all three problems mentioned above. Our results can be viewed as follow-up research to a line of work known as "Matrix Sampling Algorithms". Frieze et al [59] presented the first such algorithm for the Low-rank Matrix Approximation problem. Since then, such algorithms have been developed for several other problems, e.g. Regression [47], Graph Sparsification [131], and Linear Equation Solving [128]. Our contributions to this line of research are: (i) improved algorithms for Low-rank Matrix Approximation and Regression (ii) algorithms for a new problem domain (K-means Clustering).