Linear Algebra and its Applications 431 (2009) 760–771

Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

Random projections for the nonnegative least-squares problem

Christos Boutsidis *, Petros Drineas

Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, United States

ARTICLE INFO

Article history:
Received 8 December 2008
Accepted 5 March 2009
Available online 25 April 2009

Submitted by M. Tsatsomeros

Keywords:
Nonnegative least-squares
Sampling
Hadamard transform
Randomized algorithm
Random projections

ABSTRACT

Constrained least-squares regression problems, such as the Nonnegative Least Squares (NNLS) problem, where the variables are restricted to take only nonnegative values, often arise in applications. Motivated by the recent development of the fast Johnson–Lindestrauss transform, we present a fast random projection type approximation algorithm for the NNLS problem. Our algorithm employs a randomized Hadamard transform to construct a much smaller NNLS problem and solves this smaller problem using a standard NNLS solver. We prove that our approach finds a nonnegative solution vector that, with high probability, is close to the optimum nonnegative solution in a relative error approximation sense. We experimentally evaluate our approach on a large collection of term-document data and verify that it does offer considerable speedups without a significant loss in accuracy. Our analysis is based on a novel random projection type result that might be of independent interest. In particular, given a tall and thin matrix $\Phi \in \mathbb{R}^{n\times d}$ $(n \gg d)$ and a vector $y \in \mathbb{R}^d$, we prove that the Euclidean length of Φy can be estimated very accurately by the Euclidean length of $\tilde{\Phi} y$, where $\tilde{\Phi}$ consists of a small subset of (appropriately rescaled) rows of Φ.

* Corresponding author.
E-mail addresses: boutsc@cs.rpi.edu (C. Boutsidis), drinep@cs.rpi.edu (P. Drineas).

0024-3795/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2009.03.026

1. Introduction

The Nonnegative Least Squares (NNLS) problem is a constrained least-squares regression problem where the variables are allowed to take only nonnegative values. More specifically, the NNLS problem is defined as follows:

Definition 1 (*Nonnegative Least Squares (NNLS)*). Given a matrix $A \in \mathbb{R}^{n \times d}$ and a target vector $b \in \mathbb{R}^n$, find a nonnegative vector $x_{opt} \in \mathbb{R}^d$ such that

$$x_{opt} = \arg \min_{x \in \mathbb{R}^d,\, x \geqslant 0} \|Ax - b\|_2^2 . \tag{1}$$

NNLS is a quadratic optimization problem with linear inequality constraints. As such, it is a convex optimization problem and thus it is solvable (up to arbitrary accuracy) in polynomial time [4]. In words, NNLS seeks to find the best nonnegative vector x_{opt} in order to approximately express b as a strictly nonnegative linear combination of the columns of A, i.e., $b \approx Ax_{opt}$.

The motivation for NNLS problems in data mining and machine learning stems from the fact that given least-squares regression problems on nonnegative data such as images, text, etc., it is natural to seek *nonnegative* solution vectors. (Examples of data applications are described in [6].) NNLS is also useful in the computation of the Nonnegative Matrix Factorization [16], which has received considerable attention in the past few years. Finally, NNLS is the core optimization problem and the computational bottleneck in designing a class of Support Vector Machines [22]. Since modern datasets are often massive, there is continuous need for faster, more efficient algorithms for NNLS.

In this paper we discuss the applicability of random projection algorithms for solving constrained regression problems, and in particular NNLS problems. Our goal is to provide fast approximation algorithms as alternatives to the existing exact, albeit expensive, NNLS methods. We focus on input matrices A that are tall and thin, i.e., $n \gg d$, and we present, analyze, and experimentally evaluate a random projection type algorithm for the nonnegative least-squares problem. Our algorithm utilizes a novel random projection type result which might be of independent interest. We argue that the proposed algorithm (described in detail in Section 3), provides relative error approximation guarantees for the NNLS problem. Our work is motivated by recent progress in the design of fast randomized approximation algorithms for unconstrained ℓ_p regression problems [10,7].

The following theorem is the main quality-of-approximation result for our randomized NNLS algorithm.

Theorem 1. *Let $\epsilon \in (0, 1]$. Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be the inputs of the NNLS problem with $n \gg d$. If the input parameter r of the* RandomizedNNLS *algorithm of Section 3 satisfies*

$$\frac{r}{\log r} \geqslant \frac{342 c_o^2 (d+1) \log(n)}{\epsilon^2}, \tag{2}$$

(for a sufficiently large constant c_o)[1] *then the* RandomizedNNLS *algorithm returns a nonnegative vector $\tilde{x}_{opt}$ such that*

$$\|A\tilde{x}_{opt} - b\|_2^2 \leq (1 + \epsilon) \min_{x \in \mathbb{R}^d,\, x \geqslant 0} \|Ax - b\|_2^2 \tag{3}$$

holds with probability at least 0.5.[2] *The running time of the* RandomizedNNLS *algorithm is*

$$O(nd \log(r)) + T_{NNLS}(r, d) . \tag{4}$$

The latter term corresponds to the time required to exactly solve an NNLS problem on an input matrix of dimensions $r \times d$.

[1] c_o is an unspecified constant in [19].

[2] Note that a small number of repetitions of the algorithm suffices to boost its success probability.