Yuancheng Luo

Updated QR Decompositions for Efficient NNLS and its GPU Parallelization

3368 A V Williams Building
University of Maryland
College Park
MD 20742-3255
yluo1@umd.edu
Ramani Duraiswami

In many signal processing applications, the non-negative least squares problem (NNLS) of "moderate" size (in a few hundred to a thousand variables) arises. Efficient solutions of these problems would enable online applications, in which the estimation can be performed as data is acquired. We parallelize a version of the *active-set* iterative algorithm derived from the original algorithm of Lawson and Hanson (1974) on a graphics processor. This algorithm requires the solution of an unconstrained least squares problem in every step of the iteration for a matrix composed of the "passive columns" of the original system matrix. To achieve improved performance we use parallelizable procedures to efficiently update and "downdate" the QR factorization of the matrix at the current iteration to account for inserted and removed columns, and efficient data structures that account for GPU memory access patterns.

The NNLS problem has roots in data-modelling where we optimize a set of underlying parameters that is used to describe observed data. The underlying parameters denote a set m variables in a $m \times 1$ vector $x = \{x_1, x_2, \cdots, x_m\}^T$. The observed data is composed of n observations in a $n \times 1$ vector $b = \{b_1, b_2, \cdots, b_n\}^T$. Suppose that the observed data are linear functions of the underlying parameters in the model, then the linear functions may be expressed as a $n \times m$ matrix A where $Ax = b$ describes a linear mapping from the parameters in x to the observations in b.

In the general case where $n \geq m$, the dense overdetermined system of linear equations may be solved via a least squares approach by decomposing matrix $A = QR$ where Q is an orthogonal $n \times m$ matrix and R is an upper-triangular $m \times m$ matrix. The resulting matrix equation may be rearranged as $Rx = Q^T b$ and solved via back-substitution.

Sometimes, the underlying parameters are constrained to be non-negative in order to reflect real-world prior information. When the data is corrupted by noise, the estimated parameters may not satisfy these constraints, producing answers which are not usable. In these cases, it is necessary to explicitly enforce

the non-negativity constraints and so we solve for

$$\min_x f(x) = \tfrac{1}{2}\|Ax - b\|^2, \qquad x_i \geq 0.$$

The seminal work of Lawson and Hanson in ref. [3] provided the first widely used method for solving this NNLS problem. This algorithm, later referred to as the active-set method, partitions the set of parameters or variables into the active and passive-sets. The active-set contains the variables with value zero and those that violate the constraints in the problem. The passive-set contains the variables that do not violate the constraint. By iteratively updating a feasibility vector with components from the passive-set, each iteration is reduced to an unconstrained linear least squares sub-problem that is solvable via QR.

We denote the unconstrained sub-problem as the linear system $A^P y = b$ where matrix A^P contains the column vectors in matrix A that correspond to variables in the passive-set. Observe that any changes between the active and passive-sets at each iteration are generally limited to the exchange of a single variable; usually one column vector is added or removed from A^P at each iteration. We make an important distinction that exchanged variables that have remained in the same set throughout several iterations have a lower propensity for future exchanges. This leads to an efficient algorithm that does not recompute the entire $A^P = QR$ decomposition at each step but rather modifies previous Q and R matrices with regards to two cases:

1. A new variable added to set P expands or updates matrix A^P by a single column.
2. The removal of a variable from set P shrinks or downdates matrix A^P by a single column.

Furthermore, we maintain a separate ordering for the columns of A^P by the relative time of insertions and deletions over iterations. This stack-like ordering ensures that variables more recently added to set P are placed near the top and computationally less expensive to update. Similarly, variables to be removed from set P are likely located near the top and are computationally less expensive to downdate.

Our update procedure is based on the modified Gram-Schmidt algorithm for orthogonalizing the inserted column with respect to all the existing columns in Q. The time-complexity of the update step is $O(nm)$. The parallel time-complexity is $O(m \log n)$. The downdate procedure involves a series of Given's rotations that introduces zeros to a single row of R. The time-complexity of the downdate step is $O(nm)$. The parallel time-complexity is $O(m)$.

We implement our algorithm on NVIDIA's Compute Unified Device Architecture. For a comparison, Matlab's built-in lsqnonneg routine implements a

version of the Lawson and Hanson active-sets algorithm that solves the sub-problem via a full QR decomposition based on Intel's optimized Math Kernel Library code-base. Other active-set variants in literature include the Fast NNLS (FNNLS) algorithm in ref. [1] and the Projective Quasi-Newton NNLS (PQN-NNLS) algorithm in ref. [2]. For experiments results, we apply the listed algorithms to a deconvolution problem with data obtained from terrain laser imaging. We show that our algorithm achieves a moderate speed-up over the lsqnonneg routine and a substantial speed-up over the FNNLS and PQN-NNLS algorithms for our data-set.

References

[1] R. Bro, S. D. Jong, A fast non-negativity-constrained least squares algorithm, Journal of Chemometrics, Vol. 11, No. 5, pp. 393-401, 1997.

[2] D. Kim, S. Sra, and I. S. Dhillon, A New Projected Quasi-Newton Approach for the Non-negative Least Squares Problem. Technical Report TR-06-54, Computer Sciences, The Univ. of Texas at Austin, 2006.

[3] C. L. Lawson and R. J. Hanson, Solving Least Squares Problems, PrenticeHall, 1987.