Frugal Coordinate Descent for Large-Scale NNLS

Vamsi K. Potluru
Department of Computer Science
University of New Mexico
ismav@cs.unm.edu

Abstract

The Nonnegative Least Squares (NNLS) formulation arises in many important regression problems. We present a novel coordinate descent method which differs from previous approaches in that we do not explicitly maintain complete gradient information. Empirical evidence shows that our approach outperforms a state-of-the-art NNLS solver in computation time for calculating radiation dosage for cancer treatment problems.

Introduction

Nonnegativity is a natural constraint in modeling many real-life scenarios, such as when dealing with chemical concentrations, brain activations and color intensities. In the case of linear regression, this leads to the Nonnegative Least Squares (NNLS) problem. Given inputs corresponding to vector $\boldsymbol{x}$ and matrix $\mathbf{W}$ of size $m \times n$, the NNLS problem is defined as follows:

$$\min_{\boldsymbol{h} \geq \mathbf{0}} f(\boldsymbol{h}) = \frac{1}{2}\|\boldsymbol{x} - \mathbf{W}\boldsymbol{h}\|_2^2 \tag{1}$$

One of the first algorithms to solve this problem was proposed by Lawson and Hanson (Lawson and Hanson 1974), and over the last couple of decades other algorithms have been developed (Kim, Sra, and Dhillon 2006; Bro and De Jong 1997).

We propose a coordinate descent scheme to solve NNLS. Our method is similar to the successful approach by Hsieh et al. (Hsieh et al. 2008) for solving linear SVM, which has been recently generalized to Nonnegative Quadratic Programming (NQP) by Nesterov (Nesterov 2010). Earlier, Franc, Hlavac and Navara (Franc, Hlavac, and Navara 2005) proposed a coordinate descent algorithm for NNLS; however, their approach of applying coordinate descent for solving NNLS is not optimized for large datasets. In particular, they compute $\mathbf{W}^\top \mathbf{W}$ which can be expensive. Experiments indicate that we converge quickly to a usable solution.

Coordinate Descent for NNLS

We optimize one coordinate at a time similar to the previous coordinate descent approach (Franc, Hlavac, and Navara 2005). However, our method avoids the expensive computation of the matrix product $\mathbf{W}^\top \mathbf{W}$. (Since we are updating only one coordinate at a time, computing the full gradient information is unnecessary.) The plain version of our **F**rugal **C**oordinate **D**escent algorithm (**FCD**) is presented in Algorithm 1.

Algorithm 1 FCD($\boldsymbol{x}, \mathbf{W}, \mathbf{h}$)

(If $\mathbf{h}$ is not specified, let $\mathbf{h} = 0$.)
Let $\boldsymbol{z} = \sum_i \mathbf{W}_i h_i$.
repeat
  for $i = 1, \ldots, n$ **do**
    $G = \langle \mathbf{W}_i, \boldsymbol{x} - \boldsymbol{z} \rangle$
    if $h_i = 0$ **then**
      $G \leftarrow min(G, 0)$
    end if
    if $G \neq 0$ **then**
      $\boldsymbol{z} \leftarrow \boldsymbol{z} + (\max(h_i - \frac{G}{\|\mathbf{W}_i\|^2}, 0) - h_i)\mathbf{W}_i$
      $h_i \leftarrow \max(h_i - \frac{G}{\|\mathbf{W}_i\|^2}, 0)$
    end if
  end for
until convergence
Output: Vector $\boldsymbol{h}$.

The convergence condition of the algorithm can be specified in a couple of different ways. One of them is to specify the stopping threshold of relative change in the norm of the current solution or objective value across outer iterations of the algorithm. Another is to explicitly set the number of outer loops or total computation time. Finally, one could use an approximate satisfiability of KKT conditions of the NNLS problem depending on the required precision of the solution. The proof of convergence and its rate have been previously discussed (Nesterov 2010).

There are two important cases for NNLS corresponding to "tall and thin" ($m \gg n$) and "short and fat" ($m \ll n$). Some of the algorithms compute the matrix product $\mathbf{W}^\top \mathbf{W}$ ($O(mn^2)$) while others work with $\mathbf{W}$ directly. Our algorithm is especially suitable when the matrix $\mathbf{W}$ is not thin.

We suggest three modifications that could potentially further speed up our algorithm. They are random permutations (Nesterov 2010), shrinking (Joachims 1998), and random projections (Boutsidis and Drineas 2009).

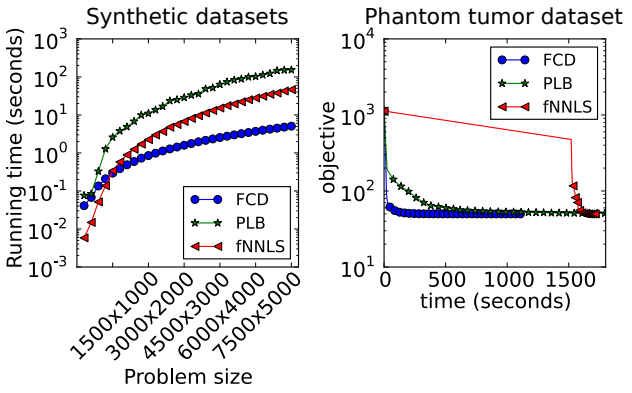

Figure 1: (Left) Mean running times for each problem size where the elements of the matrix $\boldsymbol{W}$ and vector $\boldsymbol{x}$ are drawn uniformly at random from $[0, 1]$. The running times for the solvers should be taken with a grain of salt because of the different stopping criterion used. (Right) Running times versus objective values for our (FCD) algorithm and the competing FNNLS and PLB algorithms on the phantom tumor dataset.

Experiments

In this section, we compare our algorithm with two NNLS solvers called PLB (Kim, Sra, and Dhillon 2006) and FNNLS (Bro and De Jong 1997). First, we applied our algorithm FCD and the competing solvers on various synthetic datasets ranging in size from 300×200 to 9000×6000. Next, we consider a large dataset obtained from a phantom commonly used for benchmarking radiosurgery treatment planning systems by Luan et al. (Luan et al. 2009). The size of the input matrix W is 42875×20268. Also, we consider a skull base tumor case that was treated with carbon ion therapy which was obtained from the German Cancer Research Center (DKFZ), of Heidelberg, Germany. The size of the input matrix W is 227920×6505. Clinically, each column of the matrix W represents the radiation energy distribution deposited by a "shot" of radiation in Gamma Knife radiosurgery. The matrix x represents the ideal radiation energy deposition as prescribed by the physician. The sought variable h denotes the beam-on time need for each shot (i.e., a column of W) to create a radiation dose distribution that is as close to the ideal as possible. The results of running times for the synthetic and the phantom datasets are shown in Figure 1. Similarly, the running times versus objective values for the real tumor dataset is shown in Figure 2.

Our algorithm was implemented in MATLAB (`http://www.mathworks.com`) similar to the PLB algorithm. We used the default settings for the competing algorithm as given by the implementation. All of our experiments were run on a 3.2 Ghz Intel machine with 24GB of RAM and the number of threads set to one.

We note that our algorithm converges rapidly to within 1% of final value very fast. This accuracy is good enough in practice for radiation dosage calculations.

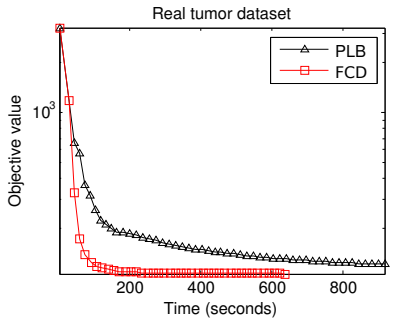

Figure 2: Running times verses objective values for FCD and PLB are shown for the real tumor dataset .

Conclusions and Future Work

We have presented a coordinate-descent algorithm to solve the NNLS problem. The new algorithm is simple to implement and its rate of convergence is at least linear. We have shown its application to two examples of dose calculation in radiation therapy. Our algorithm has the potential to be parallelized (Bradley et al. 2011).

Acknowledgement

The author acknowledges the support from NIBIB grants 1 R01 EB 000840 and 1 R01 EB 005846, and also the fruitful discussions with Sergey Plis, Thomas Hayes, Vince Calhoun, Shuang Luan and Barak Pearlmutter.

References

Boutsidis, C., and Drineas, P. 2009. Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications* 431(5-7):760–771.

Bradley, J. K.; Kyrola, A.; Bickson, D.; and Guestrin, C. 2011. Parallel coordinate descent for l1-regularized loss minimization. In *ICML*, 321–328.

Bro, R., and De Jong, S. 1997. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics* 11(5):393–401.

Franc, V.; Hlavac, V.; and Navara, M. 2005. Sequential coordinate-wise algorithm for the non-negative least squares problem. In *Computer Analysis of Images and Patterns*, 407.

Hsieh, C.-J.; Chang, K.-W.; Lin, C.-J.; Keerthi, S. S.; and Sundararajan, S. 2008. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, 408–415. New York, NY, USA: ACM.

Joachims, T. 1998. Making large-scale SVM learning practical. LS8-Report 24, Universität Dortmund, LS VIII-Report.

Kim, D.; Sra, S.; and Dhillon, I. 2006. *A new projected quasi-newton approach for the nonnegative least squares problem*. Citeseer.

Lawson, C., and Hanson, R. 1974. Solving least squares problems, 340 pp.

Luan, S.; Swanson, N.; Chen, Z.; and Ma, L. 2009. Dynamic gamma knife radiosurgery. *Physics in Medicine and Biology* 54:1579.

Nesterov, Y. 2010. Efficiency of coordinate descent methods on huge-scale optimization problems. *CORE Discussion Papers*.