WAKE FOREST UNIVERSITY

Department of Computer Science

S-Step Dual Coordinate Descent for Dual Support Vector Machines

Zishan Shao[1] , Aditya Devarakonda[1]

Department of Computer Science, Wake Forest University

Introduction

Support Vector Machine (SVM) (Boser et al., 1992) are supervised learning models used for classification tasks by finding the hyperplane (or set of hyperplanes) in a high-dimensional space. The primal problem could be reconstructed by introducing two lagrangian multipliers: α, μ. Therefore, the SVM problem could be reformed to:

L1-SVM with kernel

$$\mathcal{L} = \arg\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathcal{K}(a_{i,:}a_{j,:}^T) - \sum_{i=1}^{m}\alpha_i \quad (1)$$

L2-SVM with kernel

$$\mathcal{L} = \arg\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathcal{K}(a_{i,:}a_{j,:}^T) - \sum_{i=1}^{m}\alpha_i + \frac{1}{4C}\sum_{i=1}^{n}{\alpha_i}^2 \quad (2)$$

Preliminary results indicate substantial improvements in running time without sacrificing the quality of solution, thereby making s-step DCD a promising method for large-scale sparse machine-learning tasks.

Methodology

Considering the matrix $A \in \mathcal{R}^{m*n}$

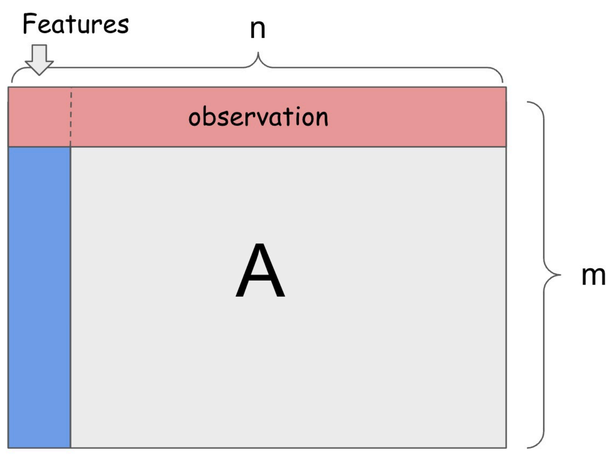


Figure 1: Definition of Trainset Matrix A

We adapt the dual coordinate descent algorithm in parallel computation with OpenMPI in C program. The matrix was partitioned with features, which allows the program to perform local computations of kernel and communicate until all local computation are finished.

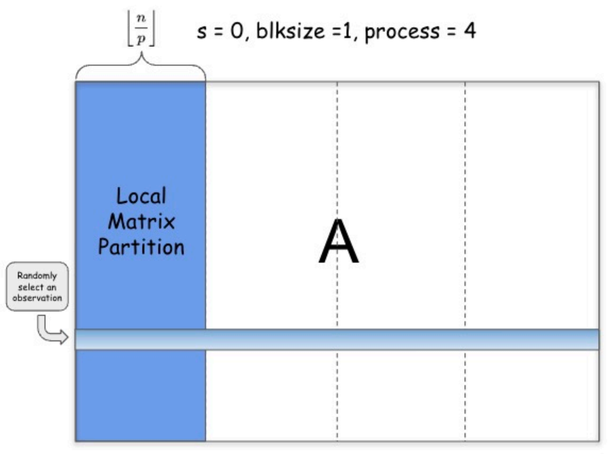


Figure 2: Partition the dataset by features

The gram matrix was then computed by AA^T

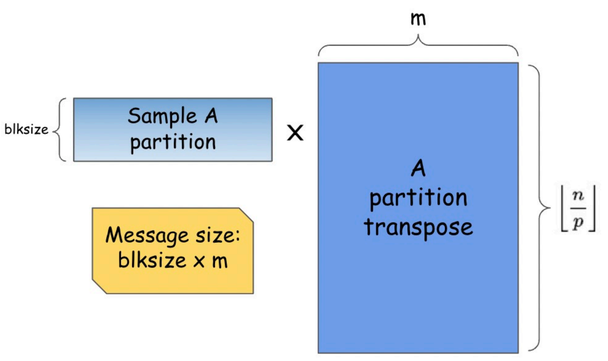


Figure 3: Gram Matrix Computation, DCD

S-step method works by selecting s block of observations with replacement, conduct local gram matrix computations and communicate s times more information, thereby reduce the latency time with a sacrifice in bandwidth cost.

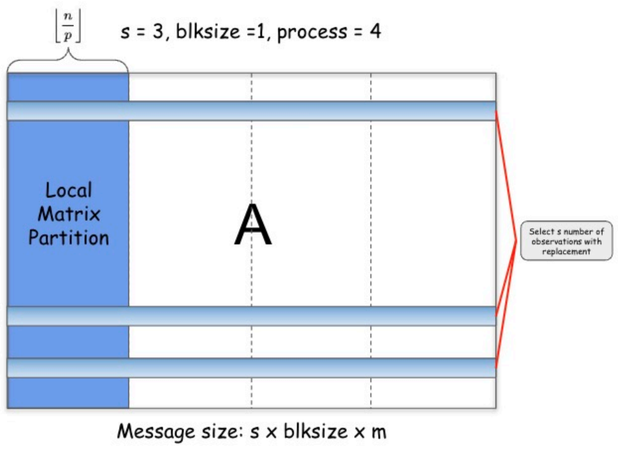


Figure 4: Partition the dataset by features

The gram matrix of s-step method was computed by:

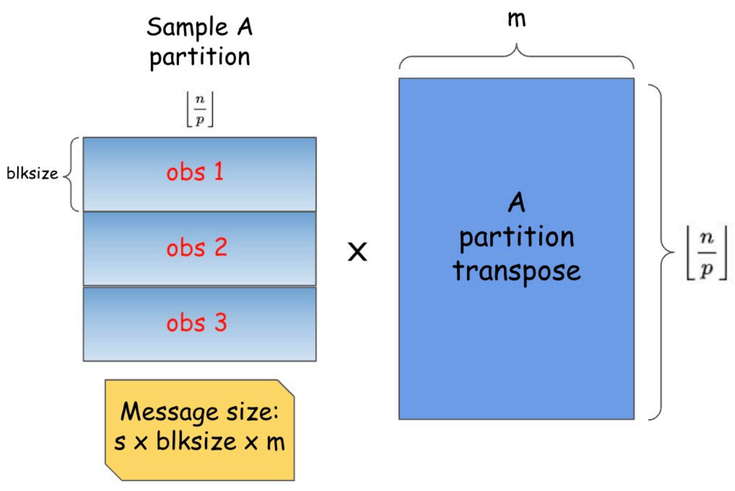


Figure 5: Gram Matrix Computation, DCD

Result

Convergence Experiment:

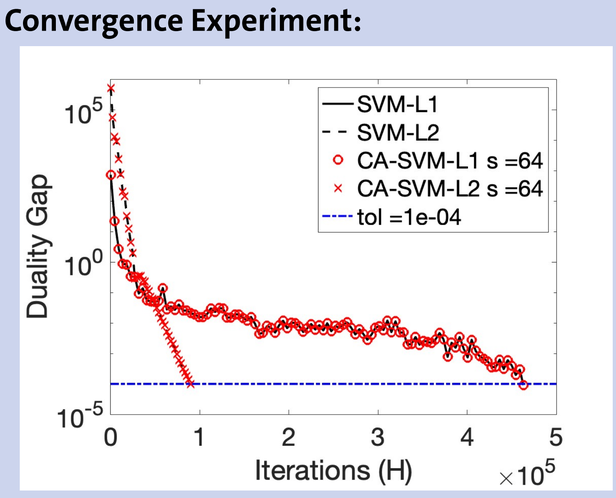
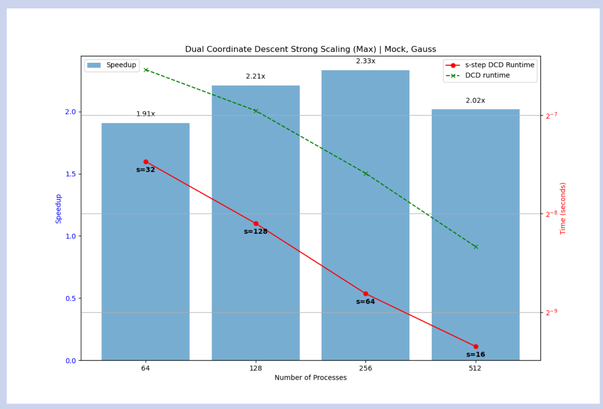


Figure 6: Convergence behavior, DCD, diabete

Strong scaling Experiment:



Runtime Breakdown:

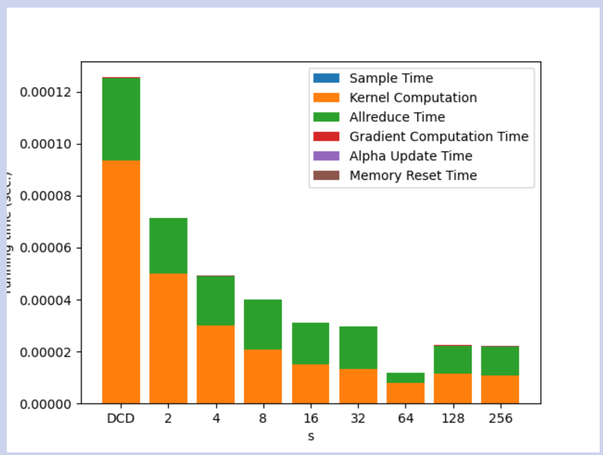


Figure 7: Duke Breast-Cancer, gauss

References

[1] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery.

[2] Aditya Devarakonda. Avoiding communication in first order methods for optimization. 2018.

[3] Cho-Jui Hsieh et al. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 2008.