

# Fast Support Vector Machines Using Parallel Adaptive Shrinking on Distributed Systems

Jeyanthi Narasimhan  
School of Electrical  
Engineering and Computer  
Science,  
Washington State University,  
Pullman, WA 99164  
jsalemna@eecs.wsu.edu

Abhinav Vishnu  
Computational Science and  
Mathematics Division,  
Pacific Northwest National  
Laboratory,  
902 Battelle Blvd, Richland,  
WA 99352  
abhinav.vishnu@pnnl.gov

Lawrence Holder  
School of Electrical  
Engineering and Computer  
Science,  
Washington State University,  
Pullman, WA 99164  
holder@wsu.edu

Adolfy Hoisie  
Computational Science and  
Mathematics Division,  
Pacific Northwest National  
Laboratory,  
902 Battelle Blvd, Richland,  
WA 99352  
adolfy.hoisie@pnl.gov

## ABSTRACT

Support Vector Machines (SVM), a popular machine learning technique, has been applied to a wide range of domains such as science, finance, and social networks for supervised learning. Whether it is identifying high-risk patients by health-care professionals, or potential high-school students to enroll in college by school districts, SVMs can play a major role for social good. This paper undertakes the challenge of designing a scalable parallel SVM training algorithm for large scale systems, which includes commodity multi-core machines, tightly connected supercomputers and cloud computing systems. Intuitive techniques for improving the time-space complexity including adaptive elimination of samples for faster convergence and sparse format representation are proposed. Under sample elimination, several heuristics for *earliest possible* to *lazy* elimination of non-contributing samples are proposed. In several cases, where an early sample elimination might result in a false positive, low overhead mechanisms for reconstruction of key data structures are proposed. The algorithm and heuristics are implemented and evaluated on various publicly available datasets. Empirical evaluation shows up to 26x speed improvement on some datasets against the sequential baseline, when evaluated on multiple compute nodes, and an improvement in execution time up to 30-60% is readily observed on a number of other datasets against our parallel baseline.

## 1. INTRODUCTION

Today, simulations and instruments produce exorbitant amounts of data and the rate of data production over the years is expected to grow dramatically [10, 21]. Machine Learning and Data Mining (MLDM) provides algorithms and tools for knowledge extraction from large volumes of data. Several domains such as science, finance and social networks rely on MLDM algorithms for supervised and unsupervised learning [1, 24, 28]. Support Vector Machines (SVM) - a supervised learning algorithm - is ubiquitous due to excellent accuracy and obliviousness to dimensionality. SVM broadly relies on the idea of large margin data classification. It con-

structs a decision surface in the feature space that bisects the two categories and maximizes the margin of separation between classes of points used in the training set. This decision surface is used for classification on the testing set provided by the user. SVM has strong theoretical foundations, and the classification and regression algorithms provide excellent generalization performance [3, 9].

With the increasing data volume and general availability of multi-core machines, several parallel SVM training algorithms are being proposed in the literature. PEGASOS [22] and dual coordinate descent [16] train on extremely large problems, albeit with limitations to linear SVMs. Cao *et al.* have proposed parallel solution extending the previously proposed Sequential Minimal Optimization (SMO) algorithm [18]. However, the empirical evaluation does not show good scalability and the entire dataset is used for training [4]. Other algorithms have been proposed for special architectural aspects such as GPUs [5, 8]. A primary problem with the algorithms proposed above is that they use the complete dataset for margin generation during the entire calculation, even though only a fraction of samples (support vectors) contribute to the hyperplane calculation. *Shrinking* - a technique to eliminate non-contributing samples - has been proposed for sequential SVMs [17] to reduce the time complexity of training. However, no parallel shrinking algorithm for multi-core machines and distributed systems exists in literature.

This paper addresses the limitations of previously proposed approaches and provides a novel parallel SVM training algorithm with adaptive shrinking. We utilize the theoretical framework for shrinking in our parallel solution to improve on the speed of convergence and use a specific format for sample representation in the optimization based on the observation that most of the real world datasets are sparse in nature (see Section 3.1.2). We study the effect of several heuristics (Section 3.3.1) for aggressive to conservative elimination of non-contributing samples during the various stages of execution. The proposed approaches are designed and implemented using state-of-the-art programming models such as Message Passing Interface (MPI) [15] and Global Arrays [19] for design of communication and data storage. These programming models are known to provide optimal performance on multi-core systems,

large scale systems and can be used on cloud computing systems as well. An empirical evaluation of proposed approaches shows up to **3x** speedup in comparison to the original non-elimination algorithms using the same number of processors, and up to **26x** speedup in comparison to libsvm [6].

## 1.1 Contributions

Specifically, this paper makes the following contributions:

1. Design and analysis of parallel algorithms to improve the time complexity of SVM training including adaptive elimination of samples. Several heuristics under the categories of *aggressive*, *average* and *conservative* for elimination of non-contributing samples.
2. Space-efficient SVM training algorithm by using compressed representation of data samples and avoiding the kernel cache. The proposed solution makes it an attractive approach for very large -scale datasets and modern systems.
3. Implementation of our proposed algorithm and evaluation with several datasets on multi-core systems and large-scale tightly-connected supercomputers. The empirical evaluation indicates the efficacy of the proposed approach - 5x-8x speedup on USPS and Mushrooms datasets against the sequential baseline [6] and 20-60% improvement in execution time on several datasets against our parallel no-shrinking baseline algorithm.

The rest of the paper is organized as follows: section 2 provides a background of our work. Section 3 presents a solution space of the algorithms and associated heuristics. Empirical evaluation and analysis is performed in section 4, and section 5 presents the related work. Section 6 presents conclusions and future directions.

## 2. BACKGROUND

Given  $\mathcal{N}$  training data points  $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_{\mathcal{N}}, y_{\mathcal{N}})\}$  where  $\mathbf{x}_i \in \mathbb{R}^d$  and  $y_i \in \{+1, -1\}$ , we solve the standard two-category soft margin non-linear classification problem. Thus the problem of finding a maximal margin separating hyperplane in a high-dimensional space can be formulated as:

$$\min_{\mathbf{w}, \beta} \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C(\sum_i \xi_i)$$

$$\text{subject to } y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) - \beta) \geq 1 - \xi_i \quad i = 1, \dots, \mathcal{N}$$

where  $C$  is a regularization parameter which is a trade-off between the classifier generality and its accuracy on the training set,  $\xi_i$  is a positive slack variable allowing noise in the training set and  $\Phi$  maps the input data to a possibly infinite dimensional space (i.e.  $\Phi: \mathbb{R}^d \mapsto \mathcal{H}$ ).

### 2.1 SVM Training

This is a convex quadratic programming problem [3]. Introducing Lagrange multipliers  $\alpha$  and solving the Lagrangian of the primal to get the Wolfe dual [12], the following formulation is observed:

$$\max_{\alpha} \mathcal{L}_D \equiv \sum_{i=1}^{\mathcal{N}} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\mathcal{N}} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \quad (1)$$

subject to:

$$0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i y_i = 0, \quad \forall i = 1, \dots, \mathcal{N} \quad (2)$$

Minimizing the primal Lagrangian provides the following formulation:

$$\mathbf{w} = \sum_{i=1}^{\mathcal{N}} \alpha_i y_i \Phi(\mathbf{x}_i) \quad (3)$$

SVM training is achieved by a search through the feasible region of the dual problem and maximization of the objective function (1), with the Karush-Kuhn-Tucker (KKT) conditions [3] in identifying the optimal solution. We refer the reader to [3,9] for the full theoretical treatment on the SVMs and training. Samples with  $\alpha_i > 0$  are referred as *support vectors*,  $\zeta$  (Table 1). The support vectors contribute to the definition of the optimal separating hyperplane - other examples can be removed from the dataset. The solution of SV training is given by (3). A new point  $\mathbf{z}$  can be classified with:

$$f(\mathbf{z}) = \text{sgn}(\mathbf{w} \cdot \Phi(\mathbf{z}) - \beta) \quad (4)$$

### 2.2 Sequential Minimal Optimization (SMO)

SVM training by solving the dual problem is typically conducted by splitting a large optimization problem into a series of smaller sub-problems [17]. The SMO algorithm [18,20] uses precisely two samples at each optimization step while solving (1). This facilitates the generation of an analytical solution possible for the quadratic minimization at each step because of the equality constraint in (2). The avoidance of dependencies on numerical optimization packages makes this algorithm a popular choice [6] in SVM training, resulting in simplified design and reduced susceptibility to numerical issues [20].

#### 2.2.1 Gradient updates

Several data structures are maintained during the SMO training [20]. An essential data structure,  $\gamma$ , is described as follows:

$$\gamma_i = \sum_j \alpha_j y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) - y_i \quad (5)$$

The relationship between  $\gamma$  and the gradient of (1) is shown in Table 1. For the rest of the paper,  $\gamma$  and gradient are used interchangeably. In all the algorithms proposed in this work, the key component in the gradient of the dual objective function (1),  $\gamma$ , is maintained for all the samples in the training set/non-shrunk samples and not just the recently optimized samples at a given iteration for reasons explained in Section 3.4.

The update equation is shown below:

$$\begin{aligned} \gamma_i^{new} &= \gamma_i^{old} + \\ & y_{up} * (\alpha_{up}^{new} - \alpha_{up}^{old}) * (\Phi(x_{up}) \cdot \Phi(x_i)) \\ & y_{low} * (\alpha_{low}^{new} - \alpha_{low}^{old}) * (\Phi(x_{low}) \cdot \Phi(x_i)) \end{aligned} \quad (6)$$

where

$$\begin{aligned} i &\in I_0 \cup I_1 \cup I_2 \cup I_3 \cup I_4. \quad I_0 = \{i : 0 < \alpha_i < C\}, \\ I_1 &= \{i : y_i = 1, \alpha_i = 0\}, I_2 = \{i : y_i = -1, \alpha_i = C\}, \\ I_3 &= \{i : y_i = 1, \alpha_i = C\}, I_4 = \{i : y_i = -1, \alpha_i = 0\} \end{aligned} \quad (7)$$

#### 2.2.2 Working Set Selection

The Working set selection describes the selection of samples to be evaluated at each step of the algorithm. Since we work with the first derivative of (1) (refer to (8)), this working set selection is addressed as first-order heuristics. Keerthi *et al.* have proposed multiple possibilities [18]. Algorithm one iterates over all examples in  $I_0$  and the second approach only evaluates the *worst KKT violators*,  $\beta_{up}$  and  $\beta_{low}$ , where at each step, they are calculated as shown in (8). Of these, we have adapted the second modification, and instead of having two loops, we operate only in the innermost loop, avoiding the first costly loop that examines all examples.

We do not compromise on the accuracy of the solution  $w$  (3) because 1) we select the pair of indices based on (8) and not just using  $I_0 \cup \{i_1, i_2\}$  as done in [18] where  $\{i_1, i_2\}$  is a recently optimized pair and 2) of the nature of our  $\gamma$  updates.

$$\begin{aligned}\beta_{up} &= \min\{\gamma_i : i \in I_0 \cup I_1 \cup I_2\} \\ \beta_{low} &= \max\{\gamma_i : i \in I_0 \cup I_3 \cup I_4\}\end{aligned}\quad (8)$$

These values are the two threshold parameters discussed in the optimized version [18]. The optimality condition for termination of the algorithm (considering numerical issues) is

$$\beta_{up} + 2 * \epsilon \geq \beta_{low} \quad (9)$$

where  $\epsilon$  is a user-specified tolerance parameter.

It can be seen from (8), (6) and (7), that the worst violators are gathered by considering all the samples, not just the recently optimized ones and the non-bound samples (i.e.,  $0 < \alpha_k < C$ ) for the next iteration.

### 2.2.3 Adaptive Elimination/Shrinking

Shrinking is a mechanism to expedite the convergence of SVM training phase by eliminating the samples, which would not contribute to the hyperplane [6, 17]. With  $I_1, I_2, I_3$  and  $I_4$  defined as in (7), samples may be eliminated if they satisfy the following decision rule:

$$\begin{aligned}i \in \{I_3 \cup I_4\} \quad \text{and} \quad \gamma_i < \beta_{up} \\ \text{or} \\ i \in \{I_1 \cup I_2\} \quad \text{and} \quad \gamma_i > \beta_{low}\end{aligned}\quad (10)$$

This heuristic is explained in the Figure 1a. The eliminated samples belong to one of the two classes: a) ones that have  $\alpha = 0$  and b) those with  $\alpha = C$ .

## 2.3 Programming Models

This paper uses two programming models - MPI [13, 15] and Global Arrays [19] for designing scalable SMO on distributed systems. Due to space limitations, we provide a brief background of Global Arrays and suggest other literature for MPI [13, 15].

### 2.3.1 Global Arrays

The Global Arrays programming model provides abstractions for distributed arrays, load/store semantics for local partition of the distributed arrays, and one-sided communication to the remote partitions. Global Arrays leverages the communication primitives provided by Communication Runtime for Exascale (ComEx) [27]. Global Arrays programming model has been used for designing many scalable applications in domains such as chemistry [25] and sub-surface modeling [23]. The Global Arrays infrastructure is useful in storing the entire dataset in a compressed row format. The easy access to local and remote portions of distributed arrays facilitates a design of algorithms which would need asynchronous read/write access to the arrays. Global Arrays uses ComEx network communication layer for one-sided communication.

## 3. SOLUTION SPACE

This section begins with a presentation of various steps of the sequential SVM training algorithm 1, which is followed by a discussion of the data structures organization using the parallel programming models. Section 3.2 introduces the parallel training algorithm of the Original algorithm, and presents its time-space complexity. This is followed by a discussion and analysis of multiple parallel shrinking algorithms 3.4

Table 1: Representative notations used and their explanation

Name	Symbol
# of Processors	$p$
# of Training Points	$\mathcal{N}$
Class label	$y_k$
Lagrange multiplier	$\alpha_k$
Set of Support Vectors	$\zeta$
Working set	$\pi$
$\delta L_D / \delta \alpha_k, \gamma_k * y_k$ (5)	$\nabla_k$
Hyperplane threshold	$\beta$
Sample in CSR form	$\tilde{x}$
Indices set $I_{0-4}$ in (7)	$\varsigma$
User-Specified Tolerance	$\epsilon$
Avg $\langle \cdot, \cdot \rangle$ time	$\lambda$
Row-Pointer Array	$\psi$
Average sample length $ x_k $	$m$
Network Latency	$l$
Network Bandwidth	$\frac{1}{G}$

---

### Algorithm 1: Improved SMO - Modification 2 [18]

---

**Input:**  $\mathcal{C}, \sigma, \mathcal{X} \in \mathbb{R}^{\mathcal{N} \times d}, y_i \in \{+1, -1\}, i = 1, 2, \dots, \mathcal{N}$

**Data:**  $\alpha \in \mathbb{R}^{\mathcal{N} \times 1}$

**Result:**  $\zeta$

- 1 Initialize  $\gamma_i = -y_i, \alpha_i = 0, \forall i$ ;
  - 2  $i_{low} = \{j \mid y_j = 1, j \in \{1, 2, \dots, \mathcal{N}\}\}$ ;
  - 3  $i_{up} = \{k \mid y_k = -1, k \in \{1, 2, \dots, \mathcal{N}\}\}$ ;
  - 4 **repeat**
  - 5     Update  $\alpha_{i_{low}}$  and  $\alpha_{i_{up}}$  using (11);
  - 6     Assign  $i_{low}$  and  $i_{up}$  to one of  $\varsigma$  using (7);
  - 7      $\forall i$ , Update  $\gamma_i$  using (6);
  - 8     Calculate new  $\beta_{low}$  and  $\beta_{up}$  using (8);
  - 9 **until** (9) *succeeds*;
- 

## 3.1 Preliminaries

Algorithm 1 shows the key steps of our sequential SVM algorithm. This is used as a basis for designing parallel SVM algorithms, with (Algorithms 5 and its variant) and without (Algorithm 3) shrinking. Using Table 1 as reference, at each iteration,  $\forall i, \alpha_i$  is calculated based on (11). In most cases, the objective function is positive definite ( $\rho < 0$ ) 12, which is used as the basis for update. An approach proposed by Platt *et al.* [20] can be used for the update equations, when  $\rho > 0$ .

$$\begin{aligned}\alpha_{i_{low}}^{new} &= \alpha_{i_{low}} - y_{i_{low}} * (\gamma_{i_{up}} - \gamma_{i_{low}}) / \rho \\ \alpha_{i_{up}}^{new} &= \alpha_{i_{up}} + y_{i_{low}} * y_{i_{up}} * (\alpha_{i_{low}} - \alpha_{i_{low}}^{new})\end{aligned}\quad (11)$$

where

$$\begin{aligned}\rho &= 2 * \Phi(\mathbf{x}_{i_{low}}) \cdot \Phi(\mathbf{x}_{i_{up}}) \\ &- \Phi(\mathbf{x}_{i_{up}}) \cdot \Phi(\mathbf{x}_{i_{up}}) - \Phi(\mathbf{x}_{i_{low}}) \cdot \Phi(\mathbf{x}_{i_{low}})\end{aligned}\quad (12)$$

Once (9) is satisfied,  $\beta$  in (4) is calculated as:

$$\beta = \begin{cases} \sum_{i \in I_0} \gamma_i / |I_0| & \text{if } |I_0| \neq 0 \\ (\beta_{low} + \beta_{up}) / 2 & \text{otherwise} \end{cases}$$

### 3.1.1 Distributed Data Structures

There are several data structures required by algorithms 3 and 5, which need to be distributed across different compute nodes. These data structures include the  $\mathcal{X}$  for the input dataset,  $y$  for the sample

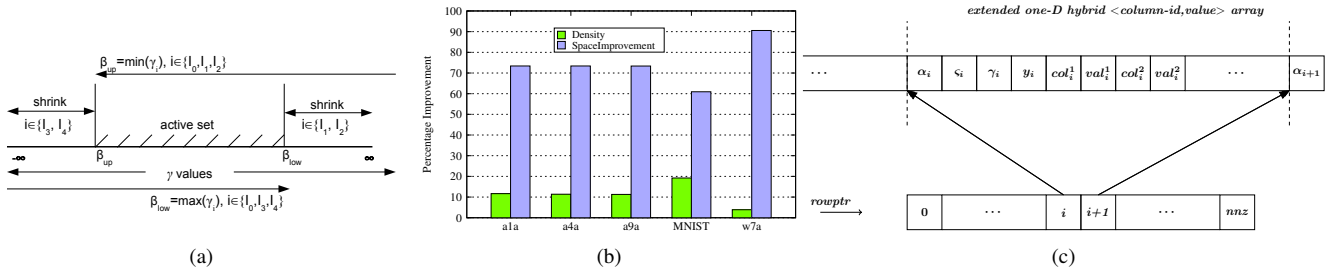


Figure 1: (a): Shrunk ids when (9) is not satisfied (non-optimality). Refer to (10) for the shrinking criterion. (b): Memory Space Conservation with CSR (c):  $\tilde{x}$ : An extended sample prototype in CSR representation. Refer to Table 1 for the meaning of the symbols.

label,  $\alpha$  for the Lagrange multipliers,  $\gamma$  and  $\zeta$ . As presented earlier, the computation can be re-formulated to using a series of kernel calculations. The individual kernel calculations may be stored in a kernel cache, which itself can be distributed among different processes.

However, there are several reasons for avoiding kernel cache for large scale systems. The space complexity of complete kernel cache is  $\Theta(\mathcal{N}^2)$ , which is prohibitive for large inputs - a primary target of this paper. At the same time, the temporal/spatial reuse of individual rows of kernel cache is low as the  $i_{up}$  and  $i_{low}$  typically do not exhibit a temporal/spatial pattern. As the architectural trends exhibit, the available memory per computation unit (such as a core in a multi-core unit) is decreasing rapidly, and it is expected that simple compute units such as Intel Xeon Phi would be commonplace [2], while Graphics Processors are already ubiquitous. At the same time, these compute units provide hardware support for wide-vector instructions, such as fused multiply-add. As a result, the cost of recomputation is expected to be much lower than either caching the complete kernel matrix or conducting off-chip/off-node data movement to get the individual rows of kernel matrix distributed across multiple nodes. Hence, the proposed approaches in this paper avoid the kernel cache altogether.

### 3.1.2 Data Structure Organization

The organization of distributed data structures plays a critical role in reducing time and space complexity of algorithm 1. Most datasets are sparse in nature, with several datasets having less than 20% density. Figure 1b shows the percentage of memory space conserved when using a compressed sparse row (CSR) [11] representation. As shown in Figure 1c, co-locating the algorithmic related data structures and making the column indices as part of representation makes little change in the density and the reduction in space complexity outweighs the additional bookkeeping for the boundaries of various samples.

The core steps of computation requires several kernel calculations and frequent access to the data structures such as the  $y$ ,  $\alpha$ , and  $\gamma$ . Among these,  $y$  is a read-only data structure, while other data structures are read-write. The organization of these data structures with  $\mathcal{X}$  has a significant potential of improving the cache-hit rate of the system, by leveraging the spatial locality. Although a few of these data structures are read-write, the write-back nature of the caches on modern systems make them a better design choice in comparison to the individual data structures distributed across multiple processes in the job. An additional advantage of co-location of these data structures with  $\mathcal{X}$  is that load balancing among processes is feasible which requires contiguous data movement of samples, instead of several individual data structures.

For the proposed approaches, CSR is implemented using Global Arrays [19] programming model. Global Arrays provides seman-

tics for collective creation of a compressed row, facilitating productive use of PGAS models for algorithms 3 and 5. GA provides Remote Memory Access semantics for traditional Ethernet based interconnects and Remote Direct Memory Semantics (RDMA), making it effective for distributed systems such as based on Cloud and tightly-connected supercomputers.

---

#### Algorithm 2: Inner Product using CSR format

---

**Input:** Samples  $\tilde{x}$  and  $\tilde{y}$ ,  $\text{len}(\tilde{x})$ ,  $\text{len}(\tilde{y})$ , where len represents the number of cells in the representation

**Output:**  $\langle \tilde{x}, \tilde{y} \rangle$ , the inner product of  $\tilde{x}$  and  $\tilde{y}$

```

1 /* shift past the padded data */;
2  $\tilde{x} \leftarrow \tilde{x} + 4, \tilde{y} \leftarrow \tilde{y} + 4;$ 
3  $s1 \leftarrow 0, s2 \leftarrow 0, dp \leftarrow 0;$ 
4 while  $s1 < \text{len}(\tilde{x}) - 4$  &  $s2 < \text{len}(\tilde{y}) - 4$  do
5   if  $|\tilde{x}[s1] - \tilde{y}[s2]| < 0$  then
6      $dp \leftarrow dp + \tilde{x}[s1 + 1] * \tilde{y}[s2 + 1];$ 
7      $s1 \leftarrow s1 + 2;$ 
8      $s2 \leftarrow s2 + 2;$ 
9   else if  $\tilde{x}[s1] < \tilde{y}[s2]$  then
10     $s1 \leftarrow s1 + 2;$ 
11   else
12     $s2 \leftarrow s2 + 2;$ 
13 return  $\langle \tilde{x}, \tilde{y} \rangle;$ 

```

---

Algorithm 2 shows the pseudocode for the inner product calculation - the most frequently executed portion in our implementations. While the CSR representation is not conducive for using hardware based vector instructions, we leave the use of such vector units as a future work. The primary objective in this paper is to minimize the space complexity by using CSR representation not used by other papers such as [5]. The inner product is used as a constituent value in  $\gamma$  (line 12 in algorithm 3) calculation by using a simple linear algebra trick.

## 3.2 Parallel SVM training Algorithms

This section lays out the parallel algorithms - with shrinking 5 and without shrinking 3. It also presents the reasoning behind shrinking and the conditions which must be true at the point of shrinking.

### 3.2.1 Parallel algorithm

Algorithm 3 is a parallel no-shrinking variant of the sequential algorithm 1. This algorithm is also referred as *Original* in various sections of this paper. There are several steps in this algorithm. Each process receives  $x_{low}, x_{up}$  from a default process using the MPI broadcast primitive, which is a scalable logarithmic operation

---

**Algorithm 3:** Algorithm 1 parallel version;  $q$ -th CPU perspective.

---

**Data:**  $p$ : # processors,  $P_q$ :  $q$ -th processor,  $0 \leq q < p$ ,  
 $i \in [q * |\mathcal{X}|/p, (q+1) * |\mathcal{X}|/p]$   
**Input:**  $\mathcal{C}, \sigma, \hat{\mathcal{X}} \in \mathbb{R}^{\frac{M}{p} \times d}, y_i \in \{+1, -1\}, \forall i$ ,  
 $\alpha \in \mathbb{R}^{\frac{M}{p} \times 1}$   
**Result:**  $\zeta$

- 1 Initialize  $\gamma_i = -y_i, \alpha_i = 0, \forall i, i_{low}, i_{up}$
- 2 **repeat**
- 3   // MPI Broadcast Operation
- 4   Receive  $\mathbf{x}_{i_{low}}, \mathbf{x}_{i_{up}}$  from proc#0
- 5   Update  $\alpha_{i_{low}}$  and  $\alpha_{i_{up}}$  using (11)
- 6   Constrain  $\alpha$ s as per (2)
- 7   **for**  $\forall i$  **do**
- 8      $li \leftarrow \psi_i$
- 9      $hi \leftarrow \psi_{i+1} - 1$
- 10     $\tilde{x} \leftarrow \text{GAGet}(li, hi)$
- 11     $\tilde{x}[2] := \text{Gradient}$ . Refer to Figure 1c    \*/
- 12    Update  $\tilde{x}[2]$  using (6)
- 13    **if**  $i == i_{up}$  **or**  $i == i_{low}$  **then**
- 14      $\tilde{x}[0] \leftarrow \alpha_{i_{low}}$  **or**  $\alpha_{i_{up}}$
- 15     Update  $\zeta$  using (7)
- 16     // global copy update
- 17     GAPut( $\tilde{x}[0 : 2]$ )                            /\* first 3 cells \*/
- 18     $\beta_{up,local} \leftarrow \min(\beta_{up_i})$
- 19     $\beta_{low,local} \leftarrow \max(\beta_{low_i})$
- 20    // Using MPI Allreduction
- 21     $\beta_{up}, \beta_{low} \leftarrow \text{GlobalMinMax}(\beta_{up,local}, \beta_{low,local}, p)$
- 22 **until** no KKT violators

---

**Algorithm 4:** Parallel update of data structures following shrinking,  $q$ -th CPU perspective.

---

- 1 **for**  $\forall i \in \pi_q$  **do**
- 2    $li \leftarrow \psi[i], hi \leftarrow \psi[i+1] - 1$ ;
- 3    $\tilde{x} \leftarrow \text{GAGet}(li, hi)$ ;
- 4   **if**  $i == i_{up}$  **or**  $i_{low}$  **then**
- 5     update  $\tilde{x}[0], \tilde{x}[1]$ ;                            /\*  $\alpha, \zeta$  \*/
- 6     update  $\tilde{x}[2]$ ;                                    /\*  $\gamma$  \*/
- 7     GAPut( $\tilde{x}[0 : 2]$ );
- 8     **if**  $\neg \text{shrinkitercounter}$  **then**
- 9       apply (10) to  $\tilde{x}$ ;
- 10      update  $\pi_q$ ;
- 11 **if**  $\neg \text{shrinkitercounter}$  **then**
- 12     $\text{shrinkitercounter} \leftarrow \text{MPI\_Allreduce}(|\pi_q|)$ ;
- 13 **else**
- 14     $\text{shrinkitercounter} --$ ;
- 15 update global  $\beta$  values;

---

in the number of processes. Each process independently calculates the new  $\alpha$  corresponding to  $i_{up}$  and  $i_{low}$ . This results in a time complexity of  $O(l + m \cdot G) \cdot \log(p)$  for network communication and three kernel calculations  $3 \cdot \lambda$  (ignoring other integer based calculation).

The *for-loop* over all samples for a process is the predominantly expensive part of the calculation. Each iteration requires the calculation of the *gradient*, which involves several kernel calculations, the  $\zeta$  calculation and update of the global values of  $\alpha_{up}$  or  $\alpha_{low}$ , if

---

**Algorithm 5:** Parallel Shrinking with single call to Algorithm 6.  $q$ -th CPU perspective.

---

**Data:**  $p$ : # processors,  $P_q$ :  $q$ -th processor,  $0 \leq q < p$ ,  $\pi$   
**Input:**  $\mathcal{C}, \sigma, \hat{\mathcal{X}} \in \mathbb{R}^{\frac{M}{p} \times d}, y_i \in \{+1, -1\}, \forall i$ ,  
 $\alpha \in \mathbb{R}^{\frac{M}{p} \times 1}$   
**Result:**  $\zeta$

- 1  $i \in [startindex, endindex)$ , where  
 $startindex \leftarrow q * |\mathcal{X}|/p, endindex \leftarrow (q+1) * |\mathcal{X}|/p$
- 2 Initialize  $\gamma_i = -y_i, \alpha_i = 0, \forall i, i_{low}, i_{up}$
- 3  $\pi_q[0 \dots (endindex - startindex)] \leftarrow startindex - endindex$
- 4 **while** 1 **do**
- 5    $outloop \leftarrow false, tolflag \leftarrow false$
- 6   **if** *shrink* **then**
- 7     **if**  $\beta_{up} < \beta_{low} - 20 \cdot \epsilon$  **then**
- 8       **repeat**
- 9         Receive  $\mathbf{x}_{i_{low}}, \mathbf{x}_{i_{up}}$  from proc#0
- 10        Update  $\alpha_{i_{low}}$  and  $\alpha_{i_{up}}$  using (11)
- 11        Constrain  $\alpha$ s to satisfy (2)
- 12        Perform Algorithm 4
- 13       **until** *failed*
- 14       **else**
- 15          $tolflag \leftarrow true$
- 16     **else**    /\* shrinking done once \*/
- 17       **if**  $\beta_{up} < \beta_{low} - 2 \cdot \epsilon$  **then**
- 18         **repeat**
- 19         Receive  $\mathbf{x}_{i_{low}}, \mathbf{x}_{i_{up}}$  from proc#0
- 20         Update  $\alpha_{i_{low}}$  and  $\alpha_{i_{up}}$  using (11)
- 21         Clip  $\alpha$ s to the box constraint (2)
- 22         Call Algorithm 4
- 23         **until** *failed*
- 24         **else**
- 25          $tolflag \leftarrow true$
- 26     **if** *shrink* & ( $outloop || tolflag$ ) **then**
- 27       gradientreconstruct()                    /\* Algorithm 6 \*/
- 28       **if**  $\beta_{up} < \beta_{low} - 2 \cdot \epsilon$  **then**
- 29          $shrink \leftarrow 0$
- 30          $shrinkitercounter \leftarrow \min\_shrink\_counter$
- 31         Reset  $\pi_q$
- 32       **else**
- 33         break                                    /\* optimality reached \*/
- 34     **else**
- 35       **if**  $outloop || tolflag$  **then**
- 36         break

---

they are locally owned by the process. The GAPut operation (line 17) updates only the indices, which were updated during the calculation, reducing the overall communication cost. The computation cost of this step is  $\Theta(\lambda \cdot \frac{|\mathcal{X}|}{p})$ . For a sufficiently large  $\frac{|\mathcal{X}|}{p}$ , the  $\zeta$  calculation and the communication cost to update the global copy of  $\alpha$  can be ignored. The last step of the algorithm is to obtain the globally maximum and minimum of  $\beta_{low}$ , and  $\beta_{up}$ , respectively. This is designed using MPI Allreduction operation which has a time-complexity of  $\Theta(l \cdot \log(p))$  (The bandwidth term can be ignored, since this step involves a communication of only two scalars).

### 3.3 Shrinking Algorithms

Joachims *et al.* [17] and Lin *et al.* [6] have previously demonstrated the impact of adaptive elimination of samples - shrinking. This technique is a heuristic, since the sufficient conditions to identify the samples to be eliminated are unknown [17]. For the eliminated samples, the Lagrange multipliers are kept fixed and they are not considered during the working set selection and the check for optimality. This results in time-complexity reduction, since the gradient for eliminated samples is not computed. The primary intuition behind shrinking is that only a small subset of samples contributes towards hyperplane definition:

$$\mathcal{A} = \{k \mid \gamma_k < \beta_{low} \text{ or } \gamma_k > \beta_{up}, 0 < \alpha_k\} \quad \text{and} \quad (13)$$

$$|\mathcal{A}| \ll |\mathcal{X}|$$

It is expected that when the optimization is at the early stage, some of the bound samples ( $\alpha_k = 0$ ,  $\alpha_k = C$ ) stabilize [17].

$$\text{At non-optimality after sufficient iterations:} \quad (14)$$

$$\hat{\mathcal{A}} = \{k \mid \beta_{low} \geq \gamma_k \geq \beta_{up}\}$$

where  $\hat{\mathcal{A}}$  is the set of violators from where working set variables are chosen and one or more samples from the set  $\mathcal{X} - \hat{\mathcal{A}}$  can be eliminated without changing the current solution. Specifically, (10) presents a variant of the condition proposed previously by Lin *et al.* for shrinking. The overhead of calculating which samples to shrink is expected to be  $\Theta(1)$ , since the computation only involves a few conditions.

However, there are several problems with this assumption. It is possible that samples with  $\alpha \in \{0, C\}$  - which were previously eliminated - eventually stabilize to a value between 0 and  $C$ . A premature elimination of these samples may result in the incorrect definition of hyperplane. A *conservative* approach to decide on the execution of this condition may not be beneficial, since much of the calculation would likely have completed. In essence, it is very difficult to predict the point at which to execute this condition. Lin *et al.* have proposed to use  $\min(|\hat{\mathcal{A}}|, 1000)$  iterations as the point to perform shrinking. However, there is no intuitive reasoning behind selecting a value to begin or executed shrinking. A discussion on spectrum of heuristics for shrinking is presented in the next section.

#### 3.3.1 Shrinking Heuristics

The heuristics for shrinking considered in this paper are to address the concerns of early elimination of the samples, while still reducing the overall time for SVM convergence. In general,  $|\zeta| \ll |\mathcal{X}|$ . Using this *intuition*, we propose several heuristics for shrinking, which are based on the  $|\mathcal{X}|$ . An *aggressive* shrinking heuristic would use a  $n \cdot |\mathcal{X}|$  as the iteration count for initial shrinking, where  $n \ll 1$ . A *conservative* shrinking heuristic could use a larger value of  $n$ . This method is referred to as *numsamples* based approach in Table 3. An alternative technique is to use initial shrinking iteration counter to be a *random* value, similar to the approach proposed by Lin *et al.* For each of these heuristics, the subsequent calculation of shrinking iteration is a minimum of  $|\mathcal{X}|$  or  $|\hat{\mathcal{A}}|$  (depending on the algorithms) and the value of shrinking iteration calculated using the proposed heuristics. See section 4.3 for further discussion on this topic.

### 3.4 Gradient Reconstruction

Gradient reconstruction is an important step in ensuring that the previously eliminated samples are not *false positives* and that they are on the correct side of the hyperplane in the final solution. Algorithm 6 shows the key steps involved in updating  $\gamma$  values during

the gradient reconstruction step of the algorithm 5. The algorithm 5 corresponds to shrinking with single gradient-reconstruction. An algorithm, which corresponds to multiple gradient reconstruction (Refer to Table 3) can be derived from this. However, due to lack of space, it is not presented explicitly.

Algorithm 6 finds the  $\gamma$  values of all the eliminated samples from the previous gradient reconstruction. To achieve this, it needs  $\mathcal{X} - \hat{\mathcal{A}}$ , which results in the communication of samples owned by each process. The time complexity of this step is  $l + |\mathcal{X} - \hat{\mathcal{A}}| \cdot G \approx |\mathcal{X} - \hat{\mathcal{A}}| \cdot G$ . The communication cost may be non-negligible for distributed systems, hence it is necessary to consider heuristics which limit the execution of gradient synchronization step. Also evident from the loop structure is the fact that the outer loop considers all eliminated samples of the  $q$ -th CPU and updates their gradient values. This is a computationally expensive operation since line 9 involves kernel calculations ((5) from section 2.2), so this algorithm is called only when global violators are within a specific threshold (e.g., lines 7 and 17 in Algorithm 5). Since  $\gamma$  plays an important role in both the  $\alpha$  updates (11) and working set selection (Section 2.2.2), we maintain it for all the active samples throughout the program execution.

Considering a less-noisy dataset,  $|\zeta| \ll \mathcal{N}$  and on an average,  $\pi_q = \frac{\zeta}{p}$ . Then,  $|\omega \cap \zeta|$  is small if not 0 and the computational time complexity expected for  $q$ -th CPU for Algorithm 6 is  $|\omega_q| \cdot |\zeta| \cdot \lambda = \left\lfloor \frac{\mathcal{X} - \zeta}{p} \right\rfloor \cdot |\zeta| \cdot \lambda$ . The tradeoff between  $|\omega|$  and  $|\zeta|$  is clear making this essential algorithm a bottleneck in achieving the overall speedup in convergence. As a result, we have considered single and multi heuristics for  $\gamma$ -reconstruction as shown in Figure 3.

---

#### Algorithm 6: Gradient Reconstruction; $q$ -th CPU perspective.

---

**Data:**  $p$ : # processors,  $P_q$ :  $q$ -th processor,  $0 \leq q < p$ ,  $\pi$

**Input:**  $\sigma$ ,  $\hat{\mathcal{X}} \in \mathbb{R}^{\frac{\mathcal{N}}{p} \times d}$ ,  $y_i \in \{+1, -1\}, \forall i$

```

1 // Gather eliminated samples of this process;
2  $li \leftarrow \psi_0$ ;
3  $hi \leftarrow \psi_{|\mathcal{X}|}$ ;
4 GAget( $li, hi$ );
5  $\omega_q = \hat{\mathcal{X}} - \pi_q$ ;
6 for  $\forall \check{a} \in \omega_q$  do
7      $my\gamma \leftarrow 0$ ;
8     for  $\{\forall \check{b} \in \mathcal{X} \mid \check{b}[0] > 0\}$  do
9          $my\gamma += \check{b}[0] * \check{b}[3] * (\Phi(\mathbf{a}) \cdot \Phi(\mathbf{b}))$ ;
10         $\check{a}[2] = my\gamma - \check{a}[3]$ ;
11        GAPut( $\check{a}[2]$ );
12 // MPI All reduction;
13 Update global  $\beta_{low}$  and  $\beta_{up}$ ;
```

---

## 4. EMPIRICAL EVALUATION

This section provides an empirical evaluation of the proposed approaches in the previous section. The empirical evaluation is conducted across multiple dimensions: datasets, number of processes, shrinking/no-shrinking, heuristics for selection of shrinking steps. The performance evaluation uses up to 512 processes (32 compute nodes), and several datasets use between 1 and 32 compute nodes. As a result, the proposed approaches can be used on multi-core machines such as a desktop, supercomputers or cloud computing systems. For each dataset, we compare our results with LIBSVM [6], version 3.17, with *shrinking enabled*.

The upcoming sections provide a brief description of the datasets,

Table 2: Dataset Characteristics and hyperparameter settings

Name	Training Set Size	Testing Set Size	C	$\sigma^2$
MNIST	60000	10000	10	25
Adult-7 (a7a)	16100	16461	32	64
Adult-9 (a9a)	32561	16281	32	64
USPS	7291	2007	8	16
Mushrooms	8124	N/A	8	64
Web (w7a)	24692	25057	32	64
IJCNN	49990	91701	0.5	1

experimental testbed and followed by empirical results. Due to accessibility limitations, the performance evaluation is conducted on a tightly connected supercomputer, although the generality of our proposed solution makes it effective for cloud computing systems as well.

## 4.1 Datasets

Table 2 provides a description of the datasets used for performance evaluation in this paper. The MNIST<sup>1</sup> dataset represents images of handwritten digits. The dimensions are formed by flattening the 28x28 pixel box into one-dimensional array of floating point values between 0 and 1, with 0 representing black and 1 white. The 10-class dataset is converted into a two-class one by representing even digits as class -1 and odd digits as +1. The sparse binary Adult dataset represents the collected census data for income prediction. Web dataset is used to categorize web pages based on their text [20]. USPS represents a collection of handwritten text recognition, collected by United States Postal Service. The mushrooms data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. The IJCNN dataset represents the first problem of International Joint Conference on Neural Nets challenge 2001. Hyperparameter settings for the datasets have been selected after doing multi-fold cross-validation [6]. These are shown in Table 2. The hyperparameter  $C$  is described in section 2.1 and  $\sigma^2$  is the kernel width in the Gaussian kernel:  $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2)$ . It is straightforward to use other kernels in this work.

## 4.2 Experimental Testbed

All our experiments were run on PNNL Institutional Computing (PIC) cluster<sup>2</sup>. PIC Cluster consists of 692 dual-socket nodes with 16 cores per socket AMD Interlagos processors, running at 2.1 GHz with 64 GB of 1600 MHz memory per node (2 GB/socket). The nodes are connected using InfiniBand QDR interconnection network. The empirical evaluation consists of a mix of results on single node (multi-core) and multiple nodes (distributed system). While the performance evaluation is on a tightly-connected system, most modern Cloud providers provide programming models such as MPI and Global Arrays, hence this solution is deployable on them as well.

## 4.3 Heuristics: An overview

Table 3 provides a list of heuristics, which are used for evaluation on datasets presented in the previous section. Specific values to *aggressive*, *conservative* and *average* methods for shrinking are provided. To emulate the heuristics evaluated by Lin *et al.*, we also compare several values of *random* sample elimination as suggested in the Table. Line 2 in the table is to be interpreted as shrinking

<sup>1</sup>deeplearning.net/data

<sup>2</sup>pic.pnnl.gov/resources.stm

Table 3: Heuristics. Description and classification. \*: Aggressive shrinking class, •: Conservative, ◊: Average

#	Shrinking Type	$\gamma$ -Recon.	Name	Class
1)	None	N/A	Original	N/A
2)	random: 2	Single	Single2	*
3)	random: 500	Single	Single500	*
4)	random: 1000	Single	Single1000	◊
5)	numsamples: 5%	Single	Single5pc	*
6)	numsamples: 10%	Single	Single10pc	◊
7)	numsamples: 50%	Single	Single50pc	•
8)	random: 2	Multi	Multi2	*
9)	random: 500	Multi	Multi500	*
10)	random: 1000	Multi	Multi1000	◊
11)	numsamples: 5%	Multi	Multi5pc	*
12)	numsamples: 10%	Multi	Multi10pc	◊
13)	numsamples: 50%	Multi	Multi50pc	•
14)	Default	Default	LIBSVM	N/A

every 2 iterations(*aggressive*), with a single call to gradient reconstruction. Optimization proceeds without shrinking after this call. Similarly, line 13 can be read as shrinking whenever the number of iterations reach half the number of samples (*conservative*) with multiple calls to  $\gamma$  reconstruction as deemed fit and optimization proceeds with shrinking throughout until convergence.

## 4.4 Results and Analysis

Figures 2 and 3 show the results for Adult-7 and Adult-9 datasets, respectively. A speedup of  $\approx 2x$  is observed on Adult-7 dataset using the Multi500 and Multi5pc heuristics in comparison to Original algorithm, and 3-3.5x in comparison to LIBSVM. Among all the approaches, Multi2 has the highest time in  $\gamma$  Reconstruction (referred as Recon-Time in the figures), largely because it eliminates samples prematurely, while other heuristics allow the  $\alpha$  values to stabilize before elimination. It is worthwhile noting that each of the Multi\* heuristics are better than Single shrinking for these datasets. For each of the adult datasets,  $|\zeta| \ll X$ , which is a suitable condition for shrinking. Since the proposed heuristics are precise, the accuracy and time for classification for each of these datasets is similar, and only a representative information is shown in Table 9. It is also worthwhile noting that the implementation of the original algorithm is near optimal, as it scales well with increasing number of processes.

The results for USPS dataset, as shown in Figure 5 show the efficacy of highly aggressive Multi2 heuristic, with Multi5pc being the second best. These results validate our premise, that  $|\zeta|$  is typically small, and a multiple 5% heuristic such as Multi5pc can provide significant elimination of computation for SVM, resulting in faster convergence. As discussed previously, the first  $\gamma$  reconstruction is executed at  $20 \cdot \epsilon$ , while others are executed  $2 \cdot \epsilon$ . However, with Multi\* heuristics, the number of times the gradient is reconstructed at the terminating condition can be predicted *a priori*. As shown in the USPS results, each of the Multi\* shrinking heuristics, although spend significantly more time in gradient reconstruction(6), still reduce the overall execution time. For USPS dataset, an overall speedup of  $\approx 1.7x$  is observed in comparison to Original implementation, and  $5x$  in comparison to LIBSVM.

Figure 4 shows the performance of various approaches on MNIST dataset using 256 and 512 processes - equivalent of 16 and 32 compute nodes. There are several take away messages - the original implementation scales well providing about 1.2x speedup or  $\approx 90\%$  efficiency. Several Multi\* heuristics perform very well, with lit-

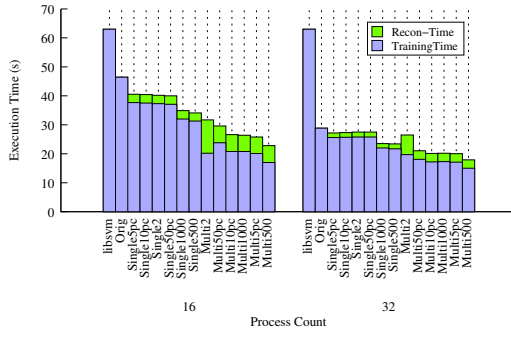


Figure 2: Adult (a7a) Dataset Performance

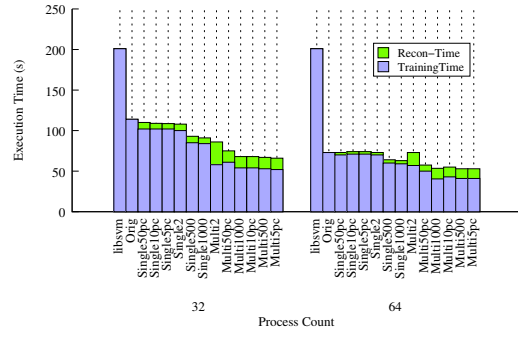


Figure 3: Adult (a9a) Dataset Performance

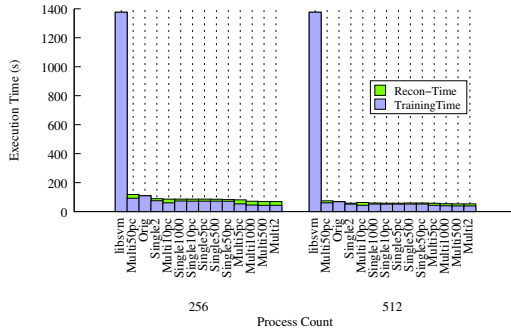


Figure 4: MNIST Dataset Performance

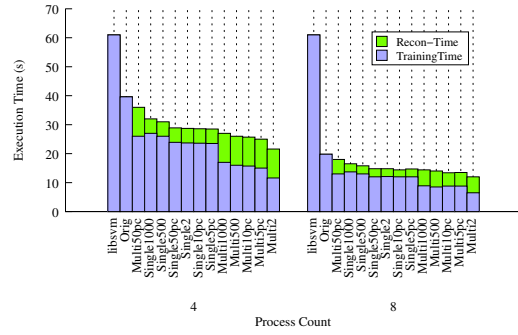


Figure 5: USPS Dataset Performance

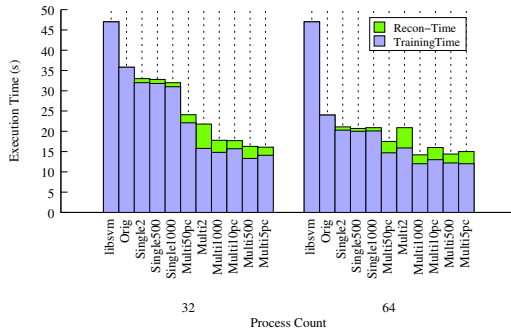


Figure 6: w7a Dataset Performance

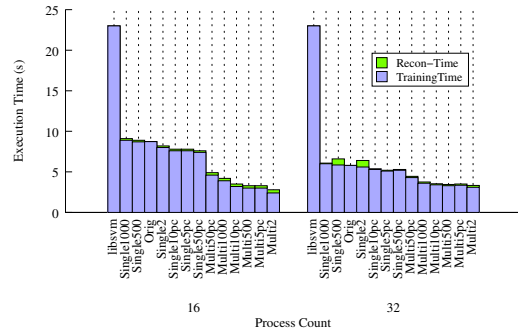


Figure 7: Mushroom Dataset Performance

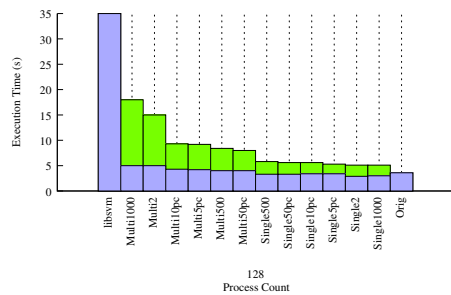


Figure 8: IJCNN Dataset Performance.

Name	Test Acc. - Ours(%)	Test Acc.-LIBSVM(%)	Speedup(Original)	Speedup (LIBSVM)
Adult-7	83.75	84.81	1.6x	3.5x
Adult-9	84.68	83.12	1.4x	3.7x
USPS	97.6	97.75	1.7x	5x
Web	98.86	98.8	1.6x	3.3x
MNIST	96.63	98.62	1.2x	26x
Mushrooms	N/A	N/A	3x	8.2x
IJCNN	90.3	96.58	N/A	9x (Orig)

Figure 9: Summary of Results (Testing Accuracy and Relative Speedups between our best performing heuristic with the Original Implementation 3 and LIBSVM)



tle difference in execution time among them. For 256 processes, a speedup of more than 1.3x is achieved using Multi1000 approach in comparison to the original approach and 26x over LIBSVM. The fact that more time is spent in  $\gamma$  reconstruction is outweighed by the overall reduction in the training time. Similar trends are observed with the w7a dataset shown in Figure 6, where 2.3x speedup is observed on 32 processes and 1.6x speedup is observed on 64 processes, with up to 3.3x speedup in comparison to LIBSVM.

Figure 7 shows the performance on the Mushrooms dataset. In comparison to other datasets Mushrooms dataset requires significantly more relative time for training due to the higher values of  $C$  and  $\sigma^2$ . As a result, the reconstruction time is relatively small.  $\frac{276}{8124}$  are support vectors, which is less than 5% of the overall training set. Here as well, Multi5pc provides near-optimal performance resulting in  $\approx 3x$  speedup, while Multi2 is slightly better than that. Again, it is fair to conclude that Multi5pc is a good heuristic in extracting the benefits of shrinking. Up to 8.2x improvement is observed in comparison to LIBSVM.

Figure 8 shows the performance of IJCNN data set. We have used this as an example to indicate that shrinking is not beneficial for all datasets and different setting of hyperparameters. For several datasets, we have observed that higher values of hyperparameters results in faster elimination of samples, potentially providing benefits of shrinking. This opens up a new avenue for research where shrinking is integrated in the cross-validation step to get parameters suitable for both shrinking and better generalization. As shown in the figure, the original implementation is the best, while each of the other approaches result in significant degradation due to shrinking. However, in comparison to LIBSVM, a speedup of up to 9x is observed with the Original implementation.

## 5. RELATED WORK

We discuss SVM training algorithms in literature under two major branches of study: 1) the sample selection methods and 2) parallel algorithms.

### 5.1 Sample Selection

Multiple researchers have proposed algorithms for selection of samples, which can be used for faster convergence. Active set methods solve the dual optimization problem by considering a part of the dataset in a given iteration until global convergence [4, 5, 8, 14, 17, 18, 20]. The primary approach is to decompose large Quadratic Programming tasks into small ones. Other approaches include the reformulation of the optimization problem, which does not require the decomposition [31]. The seminal SMO [20] and SVM<sup>light</sup> [17] are active set sequential methods and SVM-GPU [5], and PSMO [4] are examples of parallel decomposition methods whereas Woodsend *et al.* [29] is an example of a parallel non-decomposition solution. A primary problem with the working set methods is the inability to address noisy, non-separable datasets [30]. However, the simplicity, ease of implementation and strong convergence properties make them an attractive choice for solving large-scale classification problems. Other researchers have considered different values ( $> 2$ ) of the working set [8, 17].

### 5.2 Parallel Algorithms

With the advent of multi-core systems and cluster computing, several parallel and distributed algorithms have been proposed in literature. This section provides a brief overview of these algorithms.

Architecture specific solutions such as GPUs [5, 8] have been proposed, and other approaches require a special cluster setup [14]. Graf *et al.* have proposed Cascade SVM [14], which provides a

parallel solution to the dual optimization problem. The primary approach is to divide the original problem in completely independent sub-problems, and recursively combine the independent solutions to obtain the final set of support vectors. However, this approach suffers from load imbalance, since many processes may finish their individual sub-problem before others. As a result, this approach does not scale well for very large scale processes - a primary target of our approach.

The advent of SIMD architectures such as GPUs has resulted in research conducted for Support Vector Machines on GPUs [5]. Under this approach, a thread is created for each data point in the training set and the MapReduce paradigm is used for compute-intensive steps. The primary approach proposed in this paper is suitable for large scale systems, and not restricted to GPUs.

Several researchers have proposed alternative mechanisms for solving QP problems. An example of variable projection method is proposed by Zanghirati and Zanni [32]. They use an iterative solver for QP problems leveraging the decomposition strategy of SVM<sup>light</sup> [17]. Chang *et al.* [7] have also considered more than 2 active set size and solves the problem using Incomplete Cholesky Factorization and Interior Point method (IPM). Woodsend *et al.* [29] have proposed parallelization of linear SVM using IPM and a combination of MPI and OpenMP. However, their approach is not an active set method, as it does not decompose a large problem into smaller ones. There are approaches like [22] that solve the primal problem for linear SVMs for very large problems, but the primary objective of this paper is to scale the most popular 2-working set based methods due to their ubiquity.

As evident from the literature study above, none of the previously proposed approaches use adaptive elimination of samples on large scale systems, which has a significant potential in reducing the execution time for several datasets.

## 6. CONCLUSIONS AND FUTURE WORK

This paper has endeavored to address the limitations of previously proposed approaches and provided a novel parallel Support Vector Machine algorithm with adaptive shrinking. It explored various design aspects of the algorithm and the associated implementation, such as space complexity reduction by using sparse data structures, intuitive heuristics for adaptive shrinking of samples, and adaptive reconstitution of the data structures. We have used state-of-the-art programming models such as Message Passing Interface (MPI) [15] and Global Arrays [26] for the design of communication and data storage in the implementations. Empirical evaluation has demonstrated the efficacy of our proposed algorithm and the heuristics.

The future work involves shrinking with second order heuristics for working set selection, with deeper evaluation of heuristics and, considering other algorithms and working set sizes for faster elimination of samples. It will also be interesting to study shrinking under different architectures like GPUs. Though the proposed approach does complete elimination of kernel cache, it is possible to use deep memory hierarchy for keeping active portions of the kernel cache. The future work would also involve optimizations on upcoming architectures such as Intel MIC architecture, and AMD Fusion APU architecture.

## 7. REFERENCES

- [1] P. Balaprakash, Y. Alexeev, S. A. Mickelson, S. Leyffer, R. L. Jacob, and A. P. Craig. Machine learning based load-balancing for the cesm climate modeling package. 2013.
- [2] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Karp,

- S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snively, T. Sterling, R. S. Williams, K. Yelick, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Keckler, D. Klein, P. Kogge, R. S. Williams, and K. Yelick. Exascale computing study: Technology challenges in achieving exascale systems peter kogge, editor and study lead, 2008.
- [3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2:121–167, June 1998.
- [4] L. J. Cao, S. S. Keerthi, C.-J. Ong, J. Q. Zhang, U. Periyathamby, X. J. Fu, and H. P. Lee. Parallel sequential minimal optimization for the training of support vector machines. *IEEE Transactions on Neural Networks*, 17(4):1039–1049, July 2006.
- [5] B. Catanzaro, N. Sundaram, and K. Keutzer. Fast support vector machine training and classification on graphics processors. In *Proceedings of the 25th international conference on Machine Learning, ICML '08*, pages 104–111. ACM, 2008.
- [6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [7] E. Y. Chang, K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu, and H. Cui. Psvm: Parallelizing support vector machines on distributed computers. In *NIPS, 2007*. Software available at <http://code.google.com/p/psvm>.
- [8] A. Cotter, N. Srebro, and J. Keshet. A GPU-tailored approach for training kernelized SVMs. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11*, pages 805–813, 2011.
- [9] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge University Press, 2000.
- [10] DOE ASCAC Subcommittee. Synergistic Challenges in Data-Intensive Science and Exascale Computing, 2013.
- [11] J. Dongarra. Sparse matrix storage formats. In Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, 2000.
- [12] R. Fletcher. *Practical methods of optimization; (2nd ed.)*. Wiley-Interscience, New York, NY, USA, 1987.
- [13] A. Geist, W. Gropp, S. Huss-Lederman, A. Lumsdaine, E. L. Lusk, W. Saphir, T. Skjellum, and M. Snir. MPI-2: Extending the message-passing interface. In *Euro-Par, Vol. 1*, pages 128–135, 1996.
- [14] H. P. Graf, E. Cosatto, L. Bottou, I. Durdanovic, and V. Vapnik. Parallel support vector machines: The cascade svm. In *Advances in Neural Information Processing Systems*, pages 521–528. MIT Press, 2005.
- [15] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A High-Performance, Portable Implementation of the MPI Message Passing Interface Standard. *Parallel Computing*, 22(6):789–828, 1996.
- [16] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, pages 408–415, New York, NY, USA, 2008. ACM.
- [17] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in kernel methods*, pages 169–184. MIT Press, 1999.
- [18] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K.R.K.Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649, 2001.
- [19] J. Nieplocha, R. J. Harrison, and R. J. Littlefield. Global Arrays: A Nonuniform Memory Access Programming Model for High-Performance Computers. *Journal of Supercomputing*, 10(2):169–189, 1996.
- [20] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*, pages 185–208, Cambridge, MA, USA, 1999. MIT Press.
- [21] Report from the DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization. Scientific Discovery at the Exascale, 2011.
- [22] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, pages 807–814, New York, NY, USA, 2007. ACM.
- [23] Subsurface Transport over Multiple Phases. STOMP. <http://stomp.pnl.gov/>.
- [24] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. DrĂĉghici. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, 06 2007.
- [25] M. Valiev, E. Bylaska, N. Govind, K. Kowalski, T. Straatsma, H. V. Dam, D. Wang, J. Nieplocha, E. Apra, T. Windus, and W. de Jong. Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, 181(9):1477 – 1489, 2010.
- [26] A. Vishnu, J. Daily, and B. Palmer. Scalable PGAS Communication Subsystem on Cray Gemini Interconnect. Pune, India, 2012. HiPC.
- [27] A. Vishnu, D. J. Kerbyson, K. Barker, and H. J. J. V. Dam. Designing scalable pgas communication subsystems on blue gene/q. Boston, 2013. 3rd Workshop on Communication Architecture for Scalable Systems.
- [28] A. Vossen. Support vector machines in high-energy physics. 2008.
- [29] K. Woodsend and J. Gondzio. Hybrid mpi/openmp parallel linear support vector machine training. *J. Mach. Learn. Res.*, 10:1937–1953, Dec. 2009.
- [30] E. Yom-tov. A parallel training algorithm for large scale support vector machines. *Neural Information Processing Systems Workshop on Large Scale Kernel Machines*, 2004.
- [31] L. Yuh-jye and O. L. Mangasarian. RSVM: Reduced support vector machines. Technical Report 00–07, Data Mining Institute, Computer Sciences Department, University of Wisconsin, 2001.
- [32] G. Zanghirati and L. Zanni. A parallel solver for large quadratic programs in training support vector machines. *Parallel Computing*, 29(4):535–551, 2003.