

AVOIDING COMMUNICATION IN PRIMAL AND DUAL BLOCK COORDINATE DESCENT METHODS

ADITYA DEVARAKONDA*, KIMON FOUNTOLAKIS†, JAMES DEMMEL‡, AND
MICHAEL W. MAHONEY†

Abstract. Primal and dual block coordinate descent methods are iterative methods for solving regularized and unregularized optimization problems. Distributed-memory parallel implementations of these methods have become popular in analyzing large machine learning datasets. However, existing implementations communicate at every iteration which, on modern data center and super-computing architectures, often dominates the cost of floating-point computation. Recent results on communication-avoiding Krylov subspace methods suggest that large speedups are possible by re-organizing iterative algorithms to avoid communication. We show how applying similar algorithmic transformations can lead to primal and dual block coordinate descent methods that only communicate every s iterations—where s is a tuning parameter—instead of every iteration for the *regularized least-squares problem*. We show that the communication-avoiding variants reduce the number of synchronizations by a factor of s on distributed-memory parallel machines without altering the convergence rate and attains strong scaling speedups of up to $6.1\times$ on a Cray XC30 supercomputer.

Key words. primal and dual methods, communication-avoiding algorithms, block coordinate descent, ridge regression

AMS subject classifications. 15A06; 62J07; 65Y05; 68W10.

1. Introduction. The running time of an algorithm depends on computation, the number of arithmetic operations (F), and communication, the cost of data movement. The communication cost includes the “bandwidth cost”, i.e. the number, W , of words sent either between levels of a memory hierarchy or between processors over a network, and the “latency cost”, i.e. the number, L , of messages sent, where a message either consists of a group of contiguous words being sent, or is used for interprocess synchronization. On modern computer architectures, communicating data often takes much longer than performing a floating-point operation and this gap is continuing to increase. Therefore, it is especially important to design algorithms that minimize communication in order to attain high performance on modern computer architectures. Communication-avoiding algorithms are a new class of algorithms that exhibit large speedups on modern, distributed-memory parallel architectures through careful algorithmic transformations [5]. Much of direct and iterative linear algebra have been re-organized to avoid communication and has led to significant performance improvements over existing state-of-the-art libraries [5, 4, 9, 29, 45, 52]. The results from communication-avoiding Krylov subspace methods [9, 21, 29] are particularly relevant to our work.

The origins of communication-avoiding Krylov subspace methods lie in the s -step Krylov methods work. Van Rosendale’s s -step conjugate gradients method [50], Chronopoulos and Gear’s s -step methods for preconditioned and unpreconditioned symmetric linear systems [15, 16], Chronopoulos and Swanson’s s -step methods for unsymmetric linear systems [17] and Kim and Chronopoulos’s s -step non-symmetric Lanczos method [31] were designed to extract more parallelism than their standard

*EECS Department, University of California, Berkeley, Berkeley, CA 94709 (aditya@eecs.berkeley.edu).

†ICSI and Statistics Department, University of California, Berkeley, Berkeley, CA 94709 (kfount@berkeley.edu, mmahoney@stat.berkeley.edu).

‡Mathematics and EECS Department, University of California, Berkeley, Berkeley, CA 94709 (demmel@berkeley.edu)

Summary of Ops and Memory costs			
Algorithm	Data layout	Ops cost (F)	Memory cost (M)
BCD	1D-column	$O\left(\frac{Hb^2fn}{P} + Hb^3\right)$	$O\left(\frac{fdn+n}{P} + b^2 + d\right)$
CA-BCD		$O\left(\frac{Hb^2sfn}{P} + Hb^3\right)$	$O\left(\frac{fdn+n}{P} + b^2s^2 + d\right)$
BDCD	1D-row	$O\left(\frac{H'b'^2fd}{P} + H'b'^3\right)$	$O\left(\frac{fdn+d}{P} + b'^2 + n\right)$
CA-BDCD		$O\left(\frac{H'b'^2sfd}{P} + H'b'^3\right)$	$O\left(\frac{fdn+d}{P} + b'^2s^2 + n\right)$

Summary of Communication costs			
Algorithm	Data layout	Latency cost (L)	Bandwidth cost (W)
BCD	1D-column	$O(H \log P)$	$O(Hb^2 \log P)$
CA-BCD		$O\left(\frac{H}{s} \log P\right)$	$O(Hb^2s \log P)$
BDCD	1D-row	$O(H' \log P)$	$O(H'b'^2 \log P)$
CA-BDCD		$O\left(\frac{H'}{s} \log P\right)$	$O(H'b'^2s \log P)$

Table 1: Ops (F), Latency (L), Bandwidth (W) and Memory per processor (M) costs comparison along the critical path of classical BCD (Thm. 4.1), BDCD (Thm. 4.2) and communication-avoiding BCD (Thm. 4.6) and BDCD (Thm. 4.7) algorithms for 1D-block column and 1D-block row data partitioning, respectively. H and H' are the number of iterations and b and b' are the block sizes for BCD, and BDCD. We assume that $X \in \mathbb{R}^{d \times n}$ is sparse with fdn non-zeros that are uniformly distributed, $0 < f \leq 1$ is the density of X (i.e. $f = \frac{nnz(X)}{dn}$), P is the number of processors and s is the recurrence unrolling parameter. fbn is the non-zeros of the $b \times n$ matrix with b sampled rows from X at each iteration, and $f'b'd$ is the non-zeros of the $d \times b'$ matrix with b' sampled columns from X at each iteration. We assume that the $b \times b$ and $b' \times b'$ Gram matrices computed at each iteration for BCD and BDCD, respectively, are dense.

counterparts. s -step Krylov methods compute s Krylov basis vectors and perform residual and solution vector updates by using Gram matrix computations and replacing modified Gram-Schmidt orthogonalization with Householder QR [51]. These optimizations enable s -step Krylov methods to use BLAS-3 matrix-matrix operations which attain higher peak hardware performance and have more parallelism than the BLAS-1 vector-vector and BLAS-2 matrix-vector operations used in standard Krylov methods. However, these methods do not avoid communication in the s Krylov basis vector computations. Demmel, Hoemmen, Mohiyuddin, and others [21, 29, 37, 38] introduced the matrix powers kernel optimization which reduces the communication cost of the s Krylov basis vector computations by a factor $O(s)$ for well-partitioned matrices. The combination of the matrix powers kernel along with extensive algorithmic modifications to existing s -step methods and derivation of new s -step methods resulted in what Carson, Demmel, Hoemmen and others call communication-avoiding Krylov subspace methods [9, 21, 29].

We build on existing work by extending those results to machine learning where scalable algorithms are especially important given the enormous amount of data. Block coordinate descent methods are routinely used in machine learning to solve optimization problems [39, 42, 53]. Given a sparse dataset $X \in \mathbb{R}^{d \times n}$ where the rows are features of the data and the columns are data points, the block coordinate descent method can compute the regularized or unregularized least squares solution

Ops and Memory Costs Comparison		
Algorithm	Ops cost (F)	Memory cost (M)
BCD (section 4 Thm. 4.1)	$O\left(\frac{Hb^2fn}{P} + Hb^3\right)$	$O\left(\frac{fdn+n}{P} + b^2 + d\right)$
BDCD (section 4 Thm. 4.2)	$O\left(\frac{H'b'^2fd}{P} + H'b'^3\right)$	$O\left(\frac{fdn+d}{P} + b'^2 + n\right)$
Krylov methods [5]	$O\left(\frac{kfdn}{P}\right)$	$O\left(\frac{fdn}{P} + \min(d, n) + \frac{\max(d, n)}{P}\right)$

Communication Costs Comparison		
Algorithm	Latency cost (L)	Bandwidth cost (W)
BCD (section 4 Thm. 4.1)	$O(H \log P)$	$O(Hb^2 \log P)$
BDCD (section 4 Thm. 4.2)	$O(H' \log P)$	$O(H'b'^2 \log P)$
Krylov methods [5]	$O(k \log P)$	$O(k \min(d, n) \log P)$

Table 2: Computation and communication costs along the critical path of BCD, BDCD, Krylov and TSQR methods. H, H' , and k are the total number of iterations required for BCD, BDCD and Krylov methods, respectively, to converge to a desired accuracy. b and b' are the block sizes for BCD and BDCD, respectively. For Krylov methods we assume a 1D-block row layout if $n < d$ (1D-block column if $n > d$) and replicate the $\min(d, n)$ -dimensional vectors and partition the $\max(d, n)$ -dimensional vectors.

by iteratively solving a subproblem using a block of b rows of X [39, 42, 53]. This process is repeated until the solution converges to a desired accuracy or until the number of iterations has reached a user-defined limit. If X is distributed (in 1D-row or 1D-column layout) across P processors then the algorithm communicates at each iteration in order to solve the subproblem. As a result, the running time for such methods is often dominated by communication cost which increases with P .

There are some frameworks and algorithms that attempt to reduce the communication bottleneck. For example, the CoCoA framework [30] reduces communication by performing coordinate descent on locally stored data points on each processor and intermittently communicating by summing or averaging the local solutions. CoCoA communicates fewer times than coordinate descent – although not provably so – but changes the convergence behavior. HOGWILD! [41] is a lock-free approach to stochastic gradient descent (SGD) where each processor selects a data point, computes a gradient using its data point and updates the solution without synchronization. Due to the lack of synchronization (or locks) processors are allowed to overwrite the solution vector. The main results in HOGWILD! show that if the solution updates are sparse (i.e. each processor only modifies a part of the solution) then running without locks does not affect the final solution with high probability.

In contrast, our results reduce the latency cost in the primal and dual block coordinate descent methods by a factor of s on distributed-memory architectures, for dense and sparse updates without changing the convergence behavior, in exact arithmetic. Hereafter we refer to the primal method as block coordinate descent (BCD) and the dual method as block dual coordinate descent (BDCD). The proofs in this paper assume that X is sparse with fdn non-zeros that are uniformly distributed where $0 < f \leq 1$ is the density of X (i.e. $f = \frac{nnz(X)}{dn}$). Each iteration of BCD samples¹ b rows of X (resp. b' columns of X for BDCD). The resulting $b \times n$ (resp.

¹uniformly, without replacement.

$d \times b'$ for BDCD) sampled matrix contains fbn (resp. $fb'd$ for BDCD) non-zeros. These assumptions simplify our analysis and provide insight into scaling behavior for ideal sparse inputs. We leave extensions of our proofs to general sparse matrices for future work.

The principle behind our communication-avoiding approach is to unroll the BCD and BDCD vector update recurrences by a factor of s , compute Gram-like matrices for the next s iterations, and use linear combinations of the s gradients to update the solution vector. Table 1 summarizes our results for BCD with X stored in a 1D-block column layout and 1D-block row layout for BDCD. Our communication-avoiding variants reduce the latency cost, which is the dominant cost, by a factor of s but increase the bandwidth and flops cost by a factor of s . The algorithms we derive also avoid communication for other data layout schemes, however, we limit our discussion in this paper to the 1D-block column and 1D-block row layouts.

1.1. Contributions. We briefly summarize our contributions:

- We present communication-avoiding algorithms for block coordinate descent and block dual coordinate descent that *provably* reduce the latency cost by a factor of s .
- We analyze the operational, communication and storage costs of the classical and our new communication-avoiding algorithms under two data partitioning schemes and describe their performance tradeoffs.
- We perform numerical experiments to illustrate that the communication-avoiding algorithms are numerically stable for all choices of s tested.
- We show performance results to illustrate that the communication-avoiding algorithms can be up to $6.1\times$ faster than the standard algorithms on up to 1024 nodes of a Cray XC30 supercomputer using MPI.

1.2. Organization. The rest of the paper is organized as follows: Section 2 summarizes existing methods for solving the regularized least squares problem and the communication cost model used to analyze our algorithms. Section 3 presents the communication-avoiding derivations of the BCD and BDCD algorithms. Section 4 analyzes the operational, communication and storage costs of the classical and communication-avoiding algorithms under the 1D-block column and 1D-block row data layouts. Section 5 provides numerical and performance experiments which show that the communication-avoiding algorithms are numerically stable and attain speedups over the standard algorithms. Finally, we conclude in Section 6 and describe directions for future work.

2. Background. We begin by describing the cost model used to analyze the running time of the standard and new, communication-avoiding algorithms. Then we survey existing methods for solving regularized least squares problems. We compare the algorithm costs of these methods and describe their tradeoffs to motivate the need for communication-avoiding block coordinate descent methods.

2.1. Modeling Communication. Algorithms have traditionally been analyzed by counting arithmetic (the number of floating-point operations). However, data movement (communication) is another important cost that often dominates arithmetic cost [25, 28]. By combining the arithmetic and communication costs we obtain the following running time model

$$(1) \quad T_{\text{algorithm}} = \underbrace{\gamma F}_{\text{Computation Cost}} + \underbrace{\alpha L + \beta W}_{\text{Communication Cost}}$$

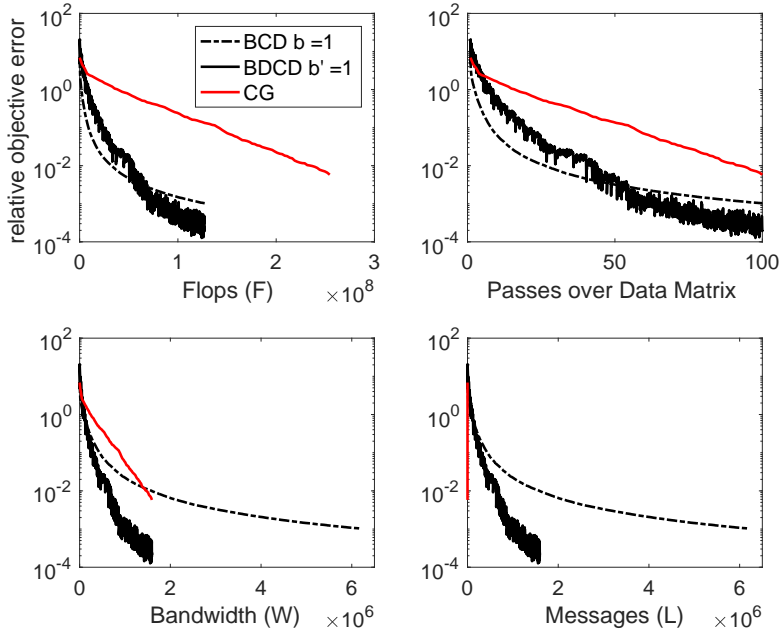


Fig. 1: Comparison of convergence behavior against flops, bandwidth, latency, and data matrix passes of Conjugate Gradients (CG), BCD (with $b = 1$) and BDCD (with $b' = 1$). Convergence is reported in terms of the relative objective error and the experiments are performed on the news20 dataset ($d = 62061$, $n = 15935$, $\text{nnz}(X) = 1272569$) obtained from LIBSVM [14]. We fix the number of CG iterations to $k = 100$, BCD iterations to $H = 100d$ and BDCD iterations to $H' = 100n$ so that each algorithm performs 100 passes over X .

where γ, α , and β are machine-specific parameters that correspond to the time per operation, overhead time per message, and time per word moved, respectively. F, L , and W are algorithm-specific parameters that represent the total number of floating-point operations computed, the number of messages sent and the number of words moved, respectively. Communication models have been well-studied in literature from the LogP [18] and LogGP [3] models to the α - β model (eq. 1). The LogP and LogGP models are refinements of the α - β model, therefore, we use the latter for simplicity. The α - β model applies to both sequential and parallel computations but we focus on the latter in this paper.

2.2. Survey of Regularized Least Squares Methods. The regularized least-squares problem can be written as the following optimization problem:

$$(2) \quad \arg \min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{2n} \|X^T w - y\|_2^2$$

where $X \in \mathbb{R}^{d \times n}$ is the data matrix whose rows are features and columns are data points, $y \in \mathbb{R}^n$ are the labels, $w \in \mathbb{R}^d$ are the weights, and $\lambda > 0$ is a regularization parameter. The unregularized ($\lambda = 0$) and regularized ($\lambda > 0$) least squares problems have been well-studied in literature from directly solving the normal equations to

Relative Objective Error Comparison			
CG iteration	CG error	BDCD error	BCD error
0	6.8735	6.8735	6.8735
1	4.5425	7.8231	1.2826
25	0.5115	0.0441	0.0104
50	0.1326	0.0043	0.0031
75	0.0283	5.0779e-04	0.0016
100	0.0058	1.9346e-04	0.0010

Table 3: Comparison of CG iterations and relative objective error of CG, BDCD ($b' = 1$) and BCD ($b = 1$). We normalize the BDCD and BCD iterations to match each CG iteration reported. If k is the CG iteration, then BCD performs $H = kd$ and BDCD performs $H' = kn$.

other matrix factorization (QR via Gram-Schmidt or Householder, LU, Cholesky, etc.) approaches [20, 6] to Krylov [6, 9, 43] and (primal and dual) block coordinate descent methods [7, 30, 44, 47, 53]. Table 2 summarizes the parallel algorithm costs of the various algorithms just described. Note that we assume that the data matrix, X , is dense for simplicity.

We briefly summarize the difference between the BCD and BDCD algorithms, but defer the derivations to Section 3. The BCD algorithm solves the primal minimization problem (2), whereas, the BDCD algorithm solves the dual minimization problem:

$$(3) \quad \arg \min_{\alpha \in \mathbb{R}^n} \frac{\lambda}{2} \left\| \frac{1}{\lambda n} X \alpha \right\|_2^2 + \frac{1}{2n} \|\alpha + y\|_2^2$$

where $\alpha \in \mathbb{R}^n$ is the dual solution vector. The dual problem [44] can be obtained by deriving the convex conjugate of (2) and has the following primal-dual solution relationship:

$$(4) \quad w = -\frac{1}{\lambda n} X \alpha.$$

Figure 1 illustrates the tradeoff between convergence behavior and theoretical flops, number of passes over X , bandwidth and latency costs of CG, BCD and BDCD based on the costs in Table 2. We plot the sequential flops cost and ignore the $\log P$ factor for latency. We allow each algorithm to make 100 passes over X and plot the relative objective error, $\frac{f(X, w_{opt}, y) - f(X, w_{alg}, y)}{f(X, w_{opt}, y)}$, where $f(X, w, y) = \frac{1}{2n} \|X^T w - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2$. w_{opt} is computed *a priori* from CG with a tolerance of 10^{-15} , and w_{alg} is the solution obtained from each iteration of CG, BCD or BDCD. Since X is not symmetric, CG requires two matrix-vector products at each iteration (one with X and another with X^T). Therefore, the flops cost of CG is twice that of BCD or BDCD. We assume that the two matrix-vector products can be computed with a single pass over X . Ignoring communication, we observe that BCD and BDCD converge faster than CG to 10^{-3} accuracy before stagnating. If low-accuracy suffices, then BCD and BDCD converge faster in terms of flops and passes over X . Since we measure the relative error in the primal objective (rather than the dual objective), the convergence of BDCD does not decrease monotonically. Table 3 summarizes the comparison of objective error progress of CG, BCD and BDCD normalized for CG iterations.

Algorithm 1 Block Coordinate Descent (BCD) Algorithm

- 1: **Input:** $X \in \mathbb{R}^{d \times n}$, $y \in \mathbb{R}^n$, $H > 1$, $w_0 \in \mathbb{R}^d$, $b \in \mathbb{Z}_+$ s.t. $b \leq d$
 - 2: **for** $h = 1, 2, \dots, H$ **do**
 - 3: choose $\{i_m \in [d] \mid m = 1, 2, \dots, b\}$ uniformly at random without replacement
 - 4: $\mathbb{I}_h = [e_{i_1}, e_{i_2}, \dots, e_{i_b}]$
 - 5: $\Gamma_h = \frac{1}{n} \mathbb{I}_h^T X X^T \mathbb{I}_h + \lambda \mathbb{I}_h^T \mathbb{I}_h$
 - 6: $\Delta w_h = \Gamma_h^{-1} \left(-\lambda \mathbb{I}_h^T w_{h-1} - \frac{1}{n} \mathbb{I}_h^T X z_{h-1} + \frac{1}{n} \mathbb{I}_h^T X y \right)$
 - 7: $w_h = w_{h-1} + \mathbb{I}_h \Delta w_h$
 - 8: $z_h = z_{h-1} + X^T \mathbb{I}_h \Delta w_h$
 - 9: **Output** w_H
-

When considering communication, CG is more bandwidth efficient than BCD (but not BDCD) and is *orders of magnitude* more latency efficient than BCD and BDCD. This suggests that reducing the latency cost of BCD and BDCD is an important step in making these algorithms competitive. In this paper, we focus on the design, numerical stability and performance of these communication-avoiding variants and leave the design space exploration of choosing the best algorithm for future work.

3. Communication-Avoiding Primal and Dual Block Coordinate Descent. In this section, we re-derive the block coordinate descent (BCD) (in section 3.1) and block dual coordinate descent (BDCD) (in section 3.2) algorithms starting from the respective minimization problems. The derivation of BCD and BDCD lead to recurrences which can be unrolled to derive communication-avoiding versions of BCD and BDCD, which we will refer to as CA-BCD and CA-BDCD respectively.

3.1. Derivation of Block Coordinate Descent. The minimization problem in (2) can be solved by block coordinate descent with the b -dimensional update

$$(5) \quad w_h = w_{h-1} + \mathbb{I}_h \Delta w_h$$

where $w_h \in \mathbb{R}^d$ and $\mathbb{I}_h = [e_{i_1}, e_{i_2}, \dots, e_{i_b}] \in \mathbb{R}^{d \times b}$, $\Delta w_h \in \mathbb{R}^b$, and $i_k \in [d]$ for $k = 1, 2, \dots, b$. By substitution in (2) we obtain the minimization problem

$$\arg \min_{\Delta w_h \in \mathbb{R}^b} \frac{\lambda}{2} \|w_{h-1} + \mathbb{I}_h \Delta w_h\|_2^2 + \frac{1}{2n} \|X^T w_{h-1} + X^T \mathbb{I}_h \Delta w_h - y\|_2^2$$

with the closed-form solution

$$(6) \quad \Delta w_h = \left(\frac{1}{n} \mathbb{I}_h^T X X^T \mathbb{I}_h + \lambda \mathbb{I}_h^T \mathbb{I}_h \right)^{-1} \left(-\lambda \mathbb{I}_h^T w_{h-1} - \frac{1}{n} \mathbb{I}_h^T X X^T w_{h-1} + \frac{1}{n} \mathbb{I}_h^T X y \right).$$

The closed-form solution requires a matrix-vector multiply using the entire data matrix to compute $\frac{1}{n} \mathbb{I}_h^T X X^T w_{h-1}$. However, this can be avoided by introducing the auxiliary variable:

$$z_h = X^T w_h$$

which, by substituting (5), can be re-arranged into a vector update of the form

$$(7) \quad \begin{aligned} z_h &= X^T w_{h-1} + X^T \mathbb{I}_h \Delta w_h \\ &= z_{h-1} + X^T \mathbb{I}_h \Delta w_h \end{aligned}$$

Algorithm 2 Communication-Avoiding Block Coordinate Descent (CA-BCD) Algorithm

- 1: **Input:** $X \in \mathbb{R}^{d \times n}$, $y \in \mathbb{R}^n$, $H > 1$, $w_0 \in \mathbb{R}^d$, $b \in \mathbb{Z}_+$ s.t. $b \leq d$
 - 2: **for** $k = 0, 1, \dots, \frac{H}{s}$ **do**
 - 3: **for** $j = 1, 2, \dots, s$ **do**
 - 4: choose $\{i_m \in [d] | m = 1, 2, \dots, b\}$ uniformly at random without replacement
 - 5: $\mathbb{I}_{sk+j} = [e_{i_1}, e_{i_2}, \dots, e_{i_b}]$
 - 6: let $Y = [\mathbb{I}_{sk+1}, \mathbb{I}_{sk+2}, \dots, \mathbb{I}_{sk+s}]^T X$.
 - 7: compute the Gram matrix, $G = \frac{1}{n} Y Y^T + \lambda I$.
 - 8: **for** $j = 1, 2, \dots, s$ **do**
 - 9: Γ_{sk+j} are the $b \times b$ diagonal blocks of G .
 - 10:
$$\Delta w_{sk+j} = \Gamma_{sk+j}^{-1} \left(-\lambda \mathbb{I}_{sk+j}^T w_{sk} - \lambda \sum_{t=1}^{j-1} \left(\mathbb{I}_{sk+j}^T \mathbb{I}_{sk+t} \Delta w_{sk+t} \right) - \frac{1}{n} \mathbb{I}_{sk+j}^T X z_{sk} \right. \\ \left. - \frac{1}{n} \sum_{t=1}^{j-1} \left(\mathbb{I}_{sk+j}^T X X^T \mathbb{I}_{sk+t} \Delta w_{sk+t} \right) + \frac{1}{n} \mathbb{I}_{sk+j}^T X y \right)$$
 - 11: $w_{sk+j} = w_{sk+j-1} + \mathbb{I}_{sk+j} \Delta w_{sk+j}$
 - 12: $z_{sk+j} = z_{sk+j-1} + X^T \mathbb{I}_{sk+j} \Delta w_{sk+j}$
 - 13: **Output** w_H
-

and the closed-form solution can be written in terms of z_{h-1} ,

$$(8) \quad \Delta w_h = \left(\frac{1}{n} \mathbb{I}_h^T X X^T \mathbb{I}_h + \lambda \mathbb{I}_h^T \mathbb{I}_h \right)^{-1} \left(-\lambda \mathbb{I}_h^T w_{h-1} - \frac{1}{n} \mathbb{I}_h^T X z_{h-1} + \frac{1}{n} \mathbb{I}_h^T X y \right).$$

In order to make the communication-avoiding BCD derivation easier, let us define

$$\Gamma_h = \frac{1}{n} \mathbb{I}_h^T X X^T \mathbb{I}_h + \lambda \mathbb{I}_h^T \mathbb{I}_h.$$

Then (8) can be re-written as

$$(9) \quad \Delta w_h = \Gamma_h^{-1} \left(-\lambda \mathbb{I}_h^T w_{h-1} - \frac{1}{n} \mathbb{I}_h^T X z_{h-1} + \frac{1}{n} \mathbb{I}_h^T X y \right).$$

This re-arrangement leads to the Block Coordinate Descent (BCD) method shown in Algorithm 1. The recurrence in lines 6, 7, and 8 of Algorithm 1 allow us to unroll the BCD recurrences and avoid communication. We begin by changing the loop index from h to $sk+j$ where k is the outer loop index, s is the recurrence unrolling parameter and j is the inner loop index. Assume that we are at the beginning of iteration $sk+1$ and w_{sk} and z_{sk} were just computed. Then Δw_{sk+1} can be computed by

$$\Delta w_{sk+1} = \Gamma_{sk+1}^{-1} \left(-\lambda \mathbb{I}_{sk+1}^T w_{sk} - \frac{1}{n} \mathbb{I}_{sk+1}^T X z_{sk} + \frac{1}{n} \mathbb{I}_{sk+1}^T X y \right).$$

By unrolling the recurrence for w_{sk+1} and z_{sk+1} we can compute Δw_{sk+2} in terms of w_{sk} and z_{sk}

$$\Delta w_{sk+2} = \Gamma_{sk+2}^{-1} \left(-\lambda \mathbb{I}_{sk+2}^T w_{sk} - \lambda \mathbb{I}_{sk+2}^T \mathbb{I}_{sk+1} \Delta w_{sk+1} \right. \\ \left. - \frac{1}{n} \mathbb{I}_{sk+2}^T X z_{sk} - \frac{1}{n} \mathbb{I}_{sk+2}^T X X^T \mathbb{I}_{sk+1} \Delta w_{sk+1} + \frac{1}{n} \mathbb{I}_{sk+2}^T X y \right).$$

Algorithm 3 Block Dual Coordinate Descent (BDCD) Algorithm

- 1: **Input:** $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$, $y \in \mathbb{R}^n$, $H' > 1$, $\alpha_0 \in \mathbb{R}^n$, $b' \in \mathbb{Z}_+$ s.t. $b' \leq n$
 - 2: **Initialize:** $w_0 \leftarrow \frac{-1}{\lambda n} X \alpha_0$
 - 3: **for** $h = 1, 2, \dots, H'$ **do**
 - 4: choose $\{i_m \in [n] \mid m = 1, 2, \dots, b'\}$ uniformly at random without replacement
 - 5: $\mathbb{I}_h = [e_{i_1}, e_{i_2}, \dots, e_{i_{b'}}]$
 - 6: $\Theta_h = \frac{1}{\lambda n^2} \mathbb{I}_h^T X^T X \mathbb{I}_h + \frac{1}{n} \mathbb{I}_h^T \mathbb{I}_h$
 - 7: $\Delta \alpha_h = -\frac{1}{n} \Theta_h^{-1} (-\mathbb{I}_h^T X^T w_{h-1} + \mathbb{I}_h^T \alpha_{h-1} + \mathbb{I}_h^T y)$
 - 8: $\alpha_h = \alpha_{h-1} + \mathbb{I}_h \Delta \alpha_h$
 - 9: $w_h = w_{h-1} - \frac{1}{\lambda n} X \mathbb{I}_h \Delta \alpha_h$
 - 10: **Output** α'_H and w'_H
-

By induction we can show that Δw_{sk+j} can be computed using w_{sk} and z_{sk}

$$(10) \quad \Delta w_{sk+j} = \Gamma_{sk+j}^{-1} \left(-\lambda \mathbb{I}_{sk+j}^T w_{sk} - \lambda \sum_{t=1}^{j-1} (\mathbb{I}_{sk+j}^T \mathbb{I}_{sk+t} \Delta w_{sk+t}) \right. \\ \left. - \frac{1}{n} \mathbb{I}_{sk+j}^T X z_{sk} - \frac{1}{n} \sum_{t=1}^{j-1} (\mathbb{I}_{sk+j}^T X X^T \mathbb{I}_{sk+t} \Delta w_{sk+t}) + \frac{1}{n} \mathbb{I}_{sk+j}^T X y \right).$$

for $j = 1, 2, \dots, s$. Due to the recurrence unrolling we can defer the updates to w_{sk} and z_{sk} for s steps. Notice that the first summation in (10) computes the intersection between the coordinates chosen at iteration $sk+j$ and $sk+t$ for $t = 1, \dots, j-1$ via the product $\mathbb{I}_{sk+j}^T \mathbb{I}_{sk+t}$. Communication can be avoided in this term by initializing all processors to the same seed for the random number generator. The second summation in (10) computes the Gram-like matrices $\mathbb{I}_{sk+j}^T X X^T \mathbb{I}_{sk+t}$ for $t = 1, \dots, j-1$. Communication can be avoided in this computation by computing the $sb \times sb$ Gram matrix $G = \left(\frac{1}{n} [\mathbb{I}_{sk+1}, \mathbb{I}_{sk+2}, \dots, \mathbb{I}_{sk+s}]^T X X^T [+ \mathbb{I}_{sk+1}, \mathbb{I}_{sk+2}, \dots, \mathbb{I}_{sk+s}] + \lambda I \right)$ once before the inner loop and redundantly storing it on all processors. Finally, at the end of the s inner loop iterations we can perform the vector updates

$$(11) \quad w_{sk+s} = w_{sk} + \sum_{t=1}^s (\mathbb{I}_{sk+t} \Delta w_{sk+t}),$$

$$(12) \quad z_{sk+s} = z_{sk} + X^T \sum_{t=1}^s (\mathbb{I}_{sk+t} \Delta w_{sk+t}).$$

The resulting communication-avoiding BCD (CA-BCD) algorithm is shown in Algorithm 2.

3.2. Derivation of Block Dual Coordinate Descent. The solution to the primal problem (2) can also be obtained by solving the dual minimization problem shown in (3) with the primal-dual relationship shown in (4). The dual problem (3) can be solved using block coordinate descent which iteratively solves a subproblem in $\mathbb{R}^{b'}$, where $1 \leq b' \leq n$ is a tunable block-size parameter. Let us first define the dual vector update for $\alpha_h \in \mathbb{R}^n$

$$(13) \quad \alpha_h = \alpha_{h-1} + \mathbb{I}_h \Delta \alpha_h.$$

Where h is the iteration index, $\mathbb{I}_h = [e_{i_1}, e_{i_2}, \dots, e_{i_{b'}}] \in \mathbb{R}^{n \times b'}$, $i_k \in [n]$ for $k = 1, 2, \dots, b'$ and $\Delta\alpha_h \in \mathbb{R}^{b'}$. By substitution in (3), $\Delta\alpha_h$ is the solution to a minimization problem in $\mathbb{R}^{b'}$ as desired:

$$(14) \quad \arg \min_{\Delta\alpha_h \in \mathbb{R}^{b'}} \frac{1}{2\lambda n^2} \|X\alpha_{h-1} + X\mathbb{I}_h\Delta\alpha_h\|_2^2 + \frac{1}{2n} \|\alpha_{h-1} + \mathbb{I}_h\Delta\alpha_h + y\|_2^2.$$

Finally, due to (4) we obtain the primal vector update for $w_h \in \mathbb{R}^d$

$$(15) \quad w_h = w_{h-1} - \frac{1}{\lambda n} X\mathbb{I}_h\Delta\alpha_h.$$

From (13), (14), and (15) we obtain a block coordinate descent algorithm which solves the dual minimization problem. Henceforth, we refer to this algorithm as block dual coordinate descent (BDCD). Note that by setting $b' = 1$ we obtain the SDCA algorithm [44] with the least-squares loss function.

The optimization problem (14) which computes the solution along the chosen coordinates has the closed-form

$$(16) \quad \Delta\alpha_h = - \left(\frac{1}{\lambda n^2} \mathbb{I}_h^T X^T X \mathbb{I}_h + \frac{1}{n} \mathbb{I}_h^T \mathbb{I}_h \right)^{-1} \left(\frac{1}{\lambda n^2} \mathbb{I}_h^T X^T X \alpha_{h-1} + \frac{1}{n} \mathbb{I}_h^T \alpha_{h-1} + \frac{1}{n} \mathbb{I}_h^T y \right).$$

Let us define $\Theta_h \in \mathbb{R}^{b' \times b'}$ such that

$$\Theta_h = \left(\frac{1}{\lambda n^2} \mathbb{I}_h^T X^T X \mathbb{I}_h + \frac{1}{n} \mathbb{I}_h^T \mathbb{I}_h \right).$$

From this we have that at iteration h , we compute the solution along the b' coordinates of the linear system

$$(17) \quad \Delta\alpha_h = -\frac{1}{n} \Theta_h^{-1} (-\mathbb{I}_h^T X^T w_{h-1} + \mathbb{I}_h^T \alpha_{h-1} + \mathbb{I}_h^T y)$$

and obtain the BDCD algorithm shown in Algorithm 3. The recurrence in lines 7, 8, and 9 of Algorithm 3 allow us to unroll the BDCD recurrences and avoid communication. We begin by changing the loop index from h to $sk + j$ where k is the outer loop index, s is the recurrence unrolling parameter and j is the inner loop index. Assume that we are at the beginning of iteration $sk + 1$ and w_{sk} and α_{sk} were just computed. Then $\Delta\alpha_{sk+1}$ can be computed by

$$\Delta\alpha_{sk+1} = -\frac{1}{n} \Theta_{sk+1}^{-1} (-\mathbb{I}_{sk+1}^T X^T w_{sk} + \mathbb{I}_{sk+1}^T \alpha_{sk} + \mathbb{I}_{sk+1}^T y).$$

Furthermore, by unrolling the recurrences for w_{sk+1} and α_{sk+1} we can analogously to (10) show by induction that

$$(18) \quad \Delta\alpha_{sk+j} = -\frac{1}{n} \Theta_{sk+j}^{-1} \left(-\mathbb{I}_{sk+j}^T X^T w_{sk} + \frac{1}{\lambda n} \sum_{t=1}^{j-1} (\mathbb{I}_{sk+j}^T X^T X \mathbb{I}_{sk+t} \Delta\alpha_{sk+t}) \right. \\ \left. + \mathbb{I}_{sk+j}^T \alpha_{sk} + \sum_{t=1}^{j-1} (\mathbb{I}_{sk+j}^T \mathbb{I}_{sk+t} \Delta\alpha_{sk+t}) + \mathbb{I}_{sk+j}^T y \right)$$

Algorithm 4 Communication-Avoiding Block Dual Coordinate Descent (CA-BDCD) Algorithm

- 1: **Input:** $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$, $y \in \mathbb{R}^n$, $H' > 1$, $\alpha_0 \in \mathbb{R}^n$, $b' \in \mathbb{Z}_+$ s.t. $b' \leq n$
 - 2: **Initialize:** $w_0 \leftarrow \frac{1}{\lambda n} X \alpha_0$
 - 3: **for** $k = 0, 1, \dots, \frac{H'}{s}$ **do**
 - 4: **for** $j = 1, 2, \dots, s$ **do**
 - 5: choose $\{i_m \in [n] | m = 1, 2, \dots, b'\}$ uniformly at random without replacement
 - 6: $\mathbb{I}_{sk+j} = [e_{i_1}, e_{i_2}, \dots, e_{i_{b'}}]$
 - 7: let $Y = X [\mathbb{I}_{sk+1}, \mathbb{I}_{sk+2}, \dots, \mathbb{I}_{sk+s}]$.
 - 8: compute the Gram matrix, $G' = \frac{1}{\lambda n^2} Y^T Y + \frac{1}{n} I$.
 - 9: **for** $j = 1, 2, \dots, s$ **do**
 - 10: Θ_{sk+j} are the $b' \times b'$ diagonal blocks of G' .
 - 11:
$$\Delta \alpha_{sk+j} = -\frac{1}{n} \Theta_{sk+j}^{-1} \left(-\mathbb{I}_{sk+j}^T X^T w_{sk} + \frac{1}{\lambda n} \sum_{t=1}^{j-1} \left(\mathbb{I}_{sk+j}^T X^T X \mathbb{I}_{sk+t} \Delta \alpha_{sk+t} \right) \right. \\ \left. + \mathbb{I}_{sk+j}^T \alpha_{sk} + \sum_{t=1}^{j-1} \left(\mathbb{I}_{sk+j}^T \mathbb{I}_{sk+t} \Delta \alpha_{sk+t} \right) + \mathbb{I}_{sk+j}^T y \right)$$
 - 12: $\alpha_{sk+j} = \alpha_{sk+j-1} + \mathbb{I}_{sk+j} \Delta \alpha_{sk+j}$
 - 13: $w_{sk+j} = w_{sk+j-1} - \frac{1}{\lambda n} X \mathbb{I}_{sk+j} \Delta \alpha_{sk+j}$
 - 14: **Output** $\alpha_{H'}$ and $w_{H'}$
-

for $j = 1, 2, \dots, s$. Note that due to unrolling the recurrence we can compute $\Delta \alpha_{sk+j}$ from w_{sk} and α_{sk} which are the primal and dual solution vectors from the previous outer iteration. Since the solution vector updates require communication, the recurrence unrolling allows us to defer those updates for s iterations at the expense of additional computation. The solution vectors can be updated at the end of the inner iterations by

$$(19) \quad w_{sk+s} = w_{sk} - \frac{1}{\lambda n} X \sum_{t=1}^s (\mathbb{I}_{sk+t} \Delta \alpha_{sk+t}),$$

$$(20) \quad \alpha_{sk+s} = \alpha_{sk} + \sum_{t=1}^s (\mathbb{I}_{sk+t} \Delta \alpha_{sk+t}).$$

The resulting communication-avoiding BDCD (CA-BDCD) algorithm is shown in Algorithm 4.

4. Analysis of Algorithms. From the derivations in Section 3, we can observe that the primal and dual block coordinate descent algorithms perform computations on XX^T and $X^T X$, respectively. This implies that, along with the convergence rates, the shape of X is a key factor in choosing between the two methods. Furthermore, the data partitioning scheme used to distribute X between processors may cause one method to have a lower communication cost than the other. In this section we analyze the cost of BCD and BDCD under two data partitioning schemes: 1D-block row (feature partitioning) and 1D-block column (data point partitioning). In both cases, we derive the associated computation, storage, and communication costs in order to compare the classical algorithms to our communication-avoiding variants.

We describe the tradeoffs between the choice of data partitioning scheme and its effect on the communication cost of the BCD and BDCD algorithms. We assume that the matrix $X \in \mathbb{R}^{d \times n}$ is sparse with fdn uniformly distributed non-zeros where $0 < f \leq 1$ is the density of X . We assume that the computed Gram matrices and residual and solution vectors are dense and that vectors in \mathbb{R}^n are partitioned and vectors in \mathbb{R}^d are replicated for 1D-block column. The reverse holds if X is stored in a 1D-block row layout. Since X is sparse the analysis of the computational cost includes passes over the sparse data structure instead of just the floating-point operations associated with the sparse matrix - sparse matrix multiplication (i.e. Gram matrix computation). Therefore, our analysis gives bounds on the local operations for each processor. We begin in Section 4.1 with the analysis of the BCD and BDCD algorithms and then analyze our new, communication-avoiding variants in Section 4.2.

4.1. Classical Algorithms. We begin with the analysis of the BCD algorithm with X stored in a 1D-block column layout and show how to extend this proof to BDCD with X in a 1D-block row layout.

THEOREM 4.1. *H iterations of the Block Coordinate Descent (BCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block column partitions with a block size b , on P processors along the critical path costs*

$$F = O\left(\frac{Hb^2fn}{P} + Hb^3\right) \text{ ops, } M = O\left(\frac{fdn+n}{P} + b^2 + d\right) \text{ words of memory.}$$

Communication costs

$$W = O(Hb^2 \log P) \text{ words moved, } L = O(H \log P) \text{ messages.}$$

Proof. The BCD algorithm computes a $b \times b$ Gram matrix, Γ_h , solves a $b \times b$ linear system to obtain Δw_h , and updates the vectors w_h and z_h . Computing the Gram matrix requires that each processor locally compute a $b \times b$ block of inner-products and then perform an all-reduce (a reduction and broadcast) to sum the partial blocks. Since the $b \times n$ sub-matrix $\mathbb{I}_h^T X$ has bfn non-zeros, the parallel Gram matrix computation ($\mathbb{I}_h^T X X^T \mathbb{I}_h$) requires $O(\frac{b^2fn}{P})$ operations (there are b^2 elements of the Gram matrix each of which depend on fn non-zeros) and communicates $O(b^2 \log P)$ words, with $O(\log P)$ messages. In order to solve the subproblem redundantly on all processors, a local copy of the residual is required. Computing the residual requires $O(\frac{bfn}{P})$ operations, and communicates $O(b \log P)$ words, in $O(\log P)$ messages. Once the residual is computed the subproblem can be solved redundantly on each processor in $O(b^3)$ flops. Finally, the vector updates to w_h and z_h can be computed without any communication in $O(b + \frac{bfn}{P})$ flops on each processor. The critical path costs of H iterations of this algorithm are $O(\frac{Hb^2fn}{P} + Hb^3)$ flops, $O(Hb^2 \log P)$ words, and $O(H \log P)$ messages. Each processor requires enough memory to store w_h , Γ_h , Δw , \mathbb{I}_h and $\frac{1}{P}$ -th of X , z_h , and y . Therefore the memory cost of each processor is $d + b^2 + 2b + \frac{fdn+2n}{P} = O(\frac{fdn+n}{P} + b^2 + d)$ words per processor. \square

Note that if $\frac{fn}{P} > b$, then the Gram matrix computation cost dominates the cost of solving the subproblem. Furthermore, the (distributed) storage cost of X dominates the cost of storing the $b \times b$ Gram matrix.

THEOREM 4.2. *H' iterations of the Block Dual Coordinate Descent (BDCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block row partitions with a block size b' , on P processors along the critical path costs*

$$F = O\left(\frac{H'b'^2fd}{P} + H'b'^3\right) \text{ ops, } M = O\left(\frac{fdn + d}{P} + b'^2 + n\right) \text{ words of memory.}$$

Communication costs

$$W = O\left(H'b'^2 \log P\right) \text{ words moved, } L = O(H' \log P) \text{ messages.}$$

Proof. The BDCD algorithm computes a $b' \times b'$ Gram matrix, Θ_h , solves a $b' \times b'$ linear system to obtain $\Delta\alpha_h$, and updates the vectors α_h and w_h . Since Θ_h requires inner-products between columns of X , a 1D-block row partitioning scheme ensures that all processors contribute to each entry of Θ_h . A similar cost analysis to the one used in Theorem 4.1 proves this theorem. \square

Note that if $\frac{fd}{P} > b$, then the Gram matrix computation cost dominates the cost of solving the subproblem. Furthermore, the (distributed) storage cost of X dominates the cost of storing the $b \times b$ Gram matrix.

If X is stored in a 1D-block row layout, then each processor stores a disjoint subset of the features of X . Since BCD selects b features at each iteration, 1D-block row partitioning could lead to load imbalance. In order to avoid load imbalance we re-partition the chosen b features into 1D-block column layout and proceed by using the 1D-block column BCD algorithm. Re-partitioning the b features requires communication, so we begin by bounding the maximum number of features assigned to a single processor². These bounds only holds with high probability since the features are chosen uniformly at random. To attain bounds on the bandwidth cost we assume that each sampled row of X has fn non-zeros.

LEMMA 4.3. *Given a matrix $X \in \mathbb{R}^{d \times n}$ and P processors such that each processor stores $\Theta\left(\lfloor \frac{d}{P} \rfloor\right)$ features, if b features are chosen uniformly at random, then the worst case maximum number of features, $\eta(b, P)$, assigned to a single processor w.h.p. is:*

$$\eta(b, P) = \begin{cases} O\left(\frac{b}{P} + \sqrt{\frac{b \log P}{P}}\right) & \text{if } b > P \log P, \\ O\left(\frac{\log b}{\log \log b}\right) & \text{if } b = P, \\ O\left(\frac{\log P}{\log \frac{P}{b}}\right) & \text{if } b < \frac{P}{\log P}. \end{cases}$$

Proof. This is the well-known generalization of the balls and bins problem introduced by Gonnet [27] and later extended by Mitzenmacher [36] and Raab et. al. [40].

Note that a similar result holds for the BDCD algorithm with X stored in a 1D-block column layout.

THEOREM 4.4. *H iterations of the Block Coordinate Descent (BCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block row partitions with a block size b , on P processors along the critical path costs*

$$F = O\left(\frac{Hb^2fn}{P} + Hb^3\right) \text{ ops, } M = O\left(\frac{fdn + n}{P} + b^2 + d\right) \text{ words of memory.}$$

²the bandwidth cost of re-partitioning is bounded by the processor with maximum load (i.e. maximum number of features).

For small messages, communication costs w.h.p.

$$W = O\left((b^2 + \eta(b, P)fn) H \log P\right) \text{ words moved, } L = O(H \log P) \text{ messages.}$$

For large messages, communication costs w.h.p.

$$W = O(Hb^2 \log P + H\eta(b, P)fn) \text{ words moved, } L = O(HP) \text{ messages.}$$

Proof. The 1D-block row partitioning scheme implies that the $b \times b$ Gram matrix, Γ_h , computation may be load imbalanced. Since we randomly select b rows, some processors may hold multiple rows while others hold none. In order to balance the computational load we perform an all-to-all to convert the $b \times n$ sampled matrix into the 1D-block column layout. The amount of data moved is bounded by the max-loaded processor, which from Lemma 4.3, stores $O(\eta(b, P))$ rows w.h.p. in the worst-case. This requires $W = O(\eta(b, P)fn \log P)$ and $L = O(\log P)$ for small messages or $W = O(\eta(b, P)fn)$ and $L = O(HP)$ for large messages. The all-to-all requires additional storage on each processor of $M = O\left(\frac{bfn}{P}\right)$ words. Once the sampled matrix is converted, the BCD algorithm proceeds as in Theorem 4.1. By combining the cost of the all-to-all over H iterations and the costs from Theorem 4.1, we obtain the costs for the BCD algorithm with X stored in a 1D-block row layout. \square

Note that the additional storage for the all-to-all does not dominate since $b < d$ by definition.

THEOREM 4.5. *H' iterations of the Block Dual Coordinate Descent (BDCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block column partitions with a block size b' , on P processors along the critical path costs w.h.p.*

$$F = O\left(\frac{H'b'^2fd}{P} + H'b'^3\right) \text{ ops, } M = O\left(\frac{fdn + d}{P} + b'^2 + n\right) \text{ words of memory.}$$

For small messages, communication costs w.h.p.

$$W = O\left(\left(b'^2 + \eta(b', P)fd\right) H' \log P\right) \text{ words moved, } L = O(H' \log P) \text{ messages.}$$

For large messages, communication costs w.h.p.

$$W = O\left(H'b'^2 \log P + H'\eta(b', P)fd\right) \text{ words moved, } L = O(H'P) \text{ messages.}$$

Proof. The BDCD algorithm computes a $b' \times b'$ Gram matrix, Θ_h . A 1D-block column partitioning scheme implies that the Gram matrix computation will be load imbalanced and, therefore, requires an all-to-all to convert the sampled matrix into a 1D-block row layout. A similar cost analysis to the one used in Theorem 4.4 proves this theorem. \square

4.2. Communication-Avoiding Algorithms. In this section, we derive the computation, storage, and communication costs of our communication-avoiding BCD and BDCD algorithm under the 1D-block row and 1D-block column data layouts. In both cases we show that our algorithm reduces the latency costs by a factor of s over the classical algorithms. We begin with the CA-BCD algorithm in 1D-block column layout and, then show how this proof extends to CA-BDCD in 1D-block row layout.

THEOREM 4.6. *H iterations of the Communication-Avoiding Block Coordinate Descent (CA-BCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block column partitions with a block size b , on P processors along the critical path costs*

$$F = O\left(\frac{Hb^2sfn}{P} + Hb^3\right) \text{ ops, } M = O\left(\frac{fdn+n}{P} + b^2s^2 + d\right) \text{ words of memory.}$$

Communication costs

$$W = O(Hb^2s \log P) \text{ words moved, } L = O\left(\frac{H}{s} \log P\right) \text{ messages.}$$

Proof. The CA-BCD algorithm computes the $sb \times sb$ Gram matrix, $G = \frac{1}{n}YY^T + \lambda I$, where $Y = [\mathbb{I}_{sk+1}, \mathbb{I}_{sk+2}, \dots, \mathbb{I}_{sk+s}]^T X$, solves s ($b \times b$) linear systems to compute Δw_{sk+j} and updates the vectors w_{sk+s} and z_{sk+s} . Computing the Gram matrix requires that each processor locally compute a $sb \times sb$ block of inner-products and then perform an all-reduce (a reduction and broadcast) to sum the partial blocks. This operation requires $O\left(\frac{b^2s^2fn}{P}\right)$ operations (there are s^2b^2 elements of the Gram matrix each of which depends on fn non-zeros), communicates $O(s^2b^2 \log P)$ words, and requires $O(\log P)$ messages. In order to solve the subproblem redundantly on all processors, a local copy of the residual is required. Computing the residual requires $O\left(\frac{bsfn}{P}\right)$ flops, and communicates $O(sb \log P)$ words, in $O(\log P)$ messages. Once the residual is computed the subproblem can be solved redundantly on each processor in $O(b^3s + b^2s^2)$ flops. Finally, the vector updates to w_{sk+s} and z_{sk+s} can be computed without any communication in $O\left(bs + \frac{bsfn}{P}\right)$ flops on each processor. Since the critical path occurs every $\frac{H}{s}$ iterations (every outer iteration), the algorithm costs $O\left(\frac{Hb^2sfn}{P} + Hb^3\right)$ flops, $O(Hb^2s \log P)$ words, and $O\left(\frac{H}{s} \log P\right)$ messages. Each processor requires enough memory to store w_{sk+j} , G , Δw_{sk+j} , \mathbb{I}_{sk+j} and $\frac{1}{P}$ -th of X , z_{sk+j} , and y . Therefore the memory cost of each processor is $d + s^2b^2 + 2sb + \frac{fdn+2n}{P} = O\left(\frac{fdn+n}{P} + b^2s^2 + d\right)$ words per processor. \square

THEOREM 4.7. *H' iterations of the Communication-Avoiding Block Dual Coordinate Descent (CA-BDCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block row partitions with a block size b' , on P processors along the critical path costs*

$$F = O\left(\frac{H'b'^2sfd}{P} + H'b'^3\right) \text{ ops, } M = O\left(\frac{fdn+d}{P} + b'^2s^2 + n\right) \text{ words of mem.}$$

Communication costs

$$W = O(H'b'^2s \log P) \text{ words moved, } L = O\left(\frac{H'}{s} \log P\right) \text{ messages.}$$

Proof. The CA-BDCD algorithm computes the $sb' \times sb'$ Gram matrix, $G' = \frac{1}{\lambda n^2}Y^TY + \frac{1}{n}I$, where $Y = X [\mathbb{I}_{sk+1}, \mathbb{I}_{sk+2}, \dots, \mathbb{I}_{sk+s}]$. The 1D-block column partitioning layout ensures that each processor computes a partial $sb' \times sb'$ block of the Gram matrix. A similar cost analysis to Theorem 4.6 proves this theorem. \square

THEOREM 4.8. *H iterations of the Communication-Avoiding Block Coordinate Descent (CA-BCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block row*

partitions with a block size b , on P processors along the critical path costs

$$F = O\left(\frac{Hb^2sfn}{P} + Hb^3\right) \text{ ops, } M = O\left(\frac{(d+bs)fn+n}{P} + b^2s^2 + d\right) \text{ words.}$$

For small messages, communication costs w.h.p.

$$W = O((b^2s + \eta(sb, P)fn)H \log P) \text{ words moved, } L = O\left(\frac{H}{s} \log P\right) \text{ messages.}$$

For large messages, communication costs w.h.p.

$$W = O(Hb^2s \log P + H\eta(sb, P)fn) \text{ words moved, } L = O\left(\frac{H}{s}P\right) \text{ messages.}$$

Proof. The 1D-block row partitioning scheme implies that the $sb \times sb$ Gram matrix computation may be load imbalanced. Since we randomly select sb rows, some processors may hold multiple chosen rows while some hold none. In order to balance the computational load we perform an all-to-all to convert the $sb \times n$ sampled matrix into the 1D-block column layout. The amount of data moved is bounded by the max-loaded processor, which from Lemma 4.3, stores $O(\eta(sb, P))$ rows w.h.p. in the worst-case. This requires $W = O(\eta(sb, P)fn \log P)$ and $L = O(\log P)$ for small messages or $W = O(\eta(sb, P)fn)$ and $L = O(HP)$ for large messages. The all-to-all requires additional storage on each processor of $M = O\left(\frac{bsfn}{P}\right)$ words. Once the sampled matrix is converted, the BCD algorithm proceeds as in Theorem 4.6. By combining the cost of the all-to-all over H iterations and the costs from Theorem 4.6, we obtain the costs for the CA-BCD algorithm with X stored in a 1D-block row layout. \square

Note that the additional storage for the all-to-all may dominate if $d < bs$. Therefore, b and s must be chosen carefully.

THEOREM 4.9. *H iterations of the Communication-Avoiding Block Dual Coordinate Descent (CA-BDCD) algorithm with the matrix $X \in \mathbb{R}^{d \times n}$ stored in 1D-block column partitions with a block size b' , on P processors along the critical path costs*

$$F = O\left(\frac{H'b'^2sfd}{P} + H'b'^3\right) \text{ ops, } M = O\left(\frac{(n+b's)fd+d}{P} + b'^2s^2 + n\right) \text{ words.}$$

For small messages, communication costs w.h.p.

$$W = O\left((b'^2s + \eta(sb', P)fd)H' \log P\right) \text{ words moved, } L = O\left(\frac{H'}{s} \log P\right) \text{ msgs.}$$

For large messages, communication costs w.h.p.

$$W = O\left(H'b'^2s \log P + H'\eta(sb', P)fd\right) \text{ words moved, } L = O\left(\frac{H'}{s}P\right) \text{ messages.}$$

Proof. The CA-BDCD algorithm computes a $sb' \times sb'$ Gram matrix, G . A 1D-block column partitioning scheme implies that the Gram matrix computation will be load imbalanced and, therefore, requires an all-to-all to convert the sampled matrix into a 1D-block row layout. A similar cost analysis to the one used in Theorem 4.8 proves this theorem. \square

Summary of datasets						
Name	Features (d)	Data Points (n)	NNZ%	σ_{min}	σ_{max}	Source
news20	62,061	15,935	0.13	$1.7e-6$	$6.0e+5$	LIBSVM [32]
a9a	123	32,561	11	$4.9e-6$	$2.0e+5$	UCI [33]
real-sim	20,958	72,309	0.24	$1.1e-3$	$9.2e+2$	LIBSVM [35]

Table 4: Properties of the LIBSVM datasets used in our experiments. We report the largest and smallest singular values (same as the eigenvalues) of $X^T X$.

The communication-avoiding variants that we have derived require a factor of s fewer messages than their classical counterparts, at the cost of more computation, bandwidth and memory. This suggests that s must be chosen carefully to balance the additional computation, bandwidth and memory usage with the reduction in the latency cost. This suggests that if latency is the dominant cost then our communication-avoiding variants can attain a s -fold speedup.

5. Experimental Evaluation. We proved in Section 4 that the CA-BCD and CA-BDCD algorithms reduce latency (the dominant cost) at the expense of additional bandwidth and computation. The recurrence unrolling we propose may also affect the numerical stability of CA-BCD and CA-BDCD since the sequence of computations and vector updates are different. In Section 5.1 we experimentally show that the communication-avoiding variants are numerically stable (in contrast to some CA-Krylov methods [9, 10, 11, 12, 13, 29]) and, in Section 5.2, we show that the communication-avoiding variants can lead to large speedups on a Cray XC30 supercomputer using MPI.

5.1. Numerical Experiments. The algorithm transformations derived in Section 3 require that the CA-BCD and CA-BDCD operate on Gram matrices of size $sb \times sb$ instead of size $b \times b$ every outer iteration. Due to the larger dimensions, the condition number of the Gram matrix increases and may have an adverse affect on the convergence behavior. We explore this tradeoff between convergence behavior, flops, communication and the choices of b and s for the standard and communication-avoiding algorithms. All numerical stability experiments were performed in MATLAB version R2016b on a 2.3 GHz Intel i7 machine with 8GB of RAM with datasets obtained from the LIBSVM repository [14]. Datasets were chosen so that all algorithms were tested on a range of shapes, sizes, and condition numbers. Table 4 summarizes the important properties of the datasets tested. For all experiments, we set the regularization parameter to $\lambda = 1000\sigma_{min}$. The regularization parameter reduces the condition numbers of the datasets and allows the BCD and BDCD algorithms to converge faster. In practice, λ should be chosen based on metrics like prediction accuracy on the test data (or hold-out data). Smaller values of λ would slow the convergence rate and require more iterations, therefore we choose λ so that our experiments have reasonable running times. We do not explore tradeoffs among λ values, convergence rate and running times in this paper. In order to measure convergence behavior, we plot the relative solution error, $\frac{\|w_{opt} - w_h\|_2}{\|w_{opt}\|_2}$, where w_h is the solution obtained from the coordinate descent algorithms at iteration h and w_{opt} is obtained from conjugate gradients with $tol = 1e-15$. We also plot the relative objective error, $\frac{f(X, w_{opt}, y) - f(X, w_h, y)}{f(X, w_{opt}, y)}$, where $f(X, w, y) = \frac{1}{2n} \|X^T w - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2$, the primal objective. We use the primal objective to show convergence behavior for BCD, BDCD and their communication-avoiding variants. We explore the tradeoff between the block

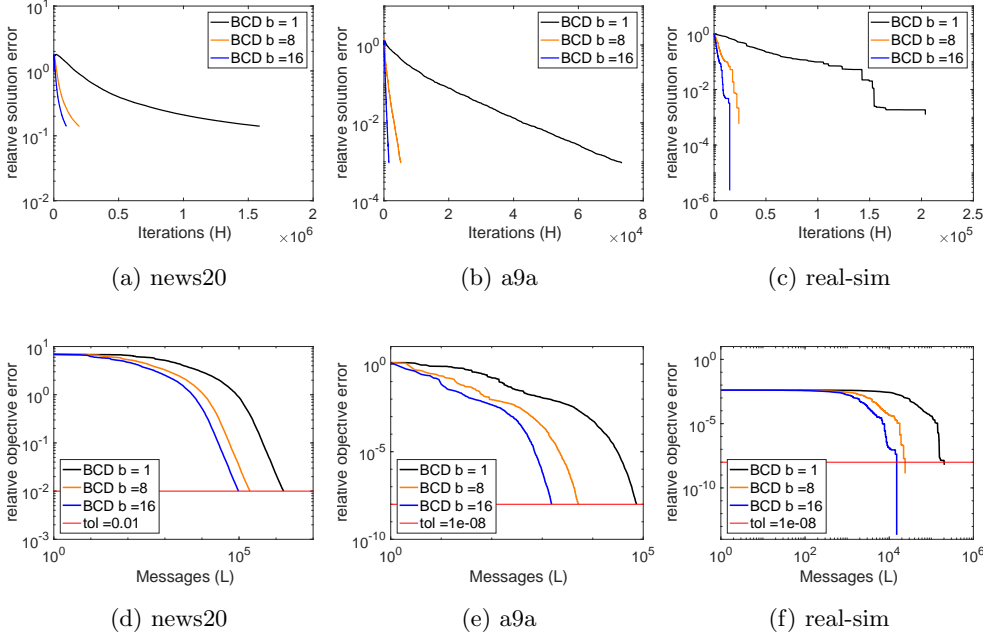


Fig. 2: We compare the convergence behavior of BCD for several block sizes, b , such that $1 \leq b < d$ on several machine learning datasets. We show relative solution error (top row, Figs. 2a-2c) and objective error (bottom row, Fig. 2d-2f) convergence plots with $\lambda = 1000\sigma_{min}$. We fix the objective error tolerance for news20 to $1e-2$ and $1e-8$ for a9a and real-sim. The x-axis for Figures 2d-2f show the number of messages required on a \log_{10} scale. Since BCD communicates at every iteration, the x-axis is also equivalent to the number of iterations (modulo \log_{10} scale).

sizes, b and b' , and convergence behavior to test BCD and BCD stability due to the choice of block sizes. Then, we fix the block sizes and explore the tradeoff between s , the recurrence unrolling parameter, and convergence behavior to study the stability of the communication-avoiding variants. Finally, for both sets of experiments we also plot the algorithm costs against convergence behavior to illustrate the theoretical performance tradeoffs due to choice of block sizes and choice of s . For the latter experiments we assume that the datasets are partitioned in 1D-block column for BCD and 1D-block row for BCD. We plot the sequential flops cost for all algorithms, ignore the $\log P$ factor for the number of messages and ignore constants. We obtain the Gram matrix computation cost from the SuiteSparse [19] routine `ssmultsym`³.

5.1.1. Block Coordinate Descent. Recall that the BCD algorithm computes a $b \times b$ Gram matrix and solves a b -dimensional subproblem at each iteration. Therefore, one should expect that as b increases the algorithm converges faster but requires more flops and bandwidth per iteration. So we begin by exploring the block size vs. convergence behavior tradeoff for BCD with $1 \leq b < d$.

³Symbolically executes the sparse matrix - sparse matrix multiplication and reports an estimate of the flops cost (counting multiplications and additions).

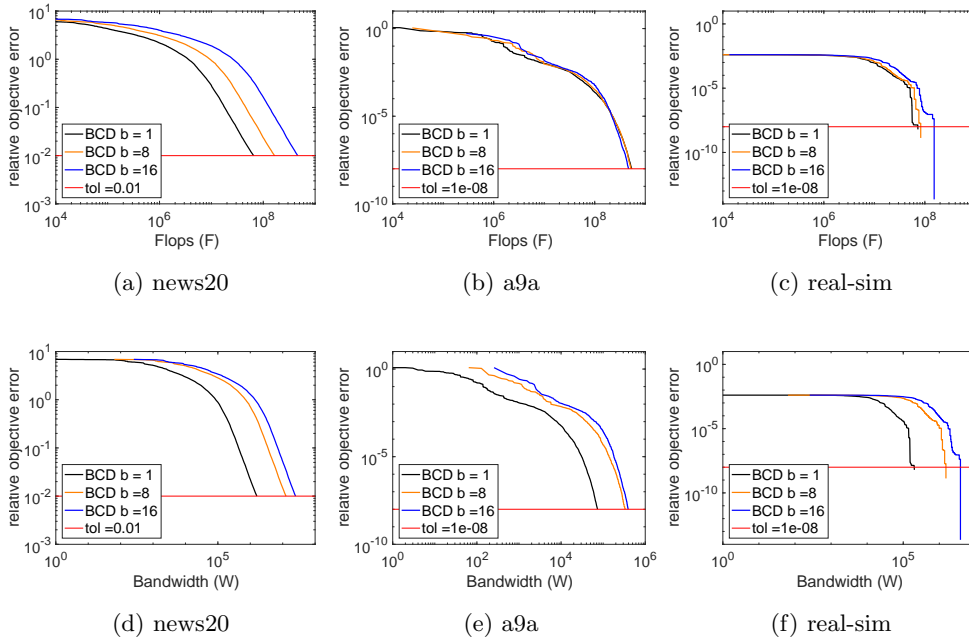


Fig. 3: We compare the convergence behavior of BCD for several block sizes, b , such that $1 \leq b < d$ on several machine learning datasets. Flops cost (top row, Figs. 3a-3c) and bandwidth cost (middle row, Figs. 3d-3f) versus convergence with $\lambda = 1000\sigma_{min}$.

Figure 2 shows the convergence behavior of the datasets in Table 4 in terms of the relative solution error (Figs. 2a-2c) and relative objective error (Figs. 2d-2f). The x-axis for the latter figures are on \log_{10} scale. Note that the number of messages is equivalent to the number of iterations, since BCD communicates every iteration. We observe that the convergence rates for all datasets improve as the block sizes increase.

Figure 3 shows the convergence behavior (in terms of the objective error) vs. flops and bandwidth costs for each dataset. From these results, we observe that BCD with $b = 1$ is more flops and bandwidth efficient, whereas $b > 1$ is more latency efficient (from Figs. 2d-2f). This indicates the existence of a tradeoff between BCD convergence rate (which depends on the block size) and hardware-specific parameters (like flops rate, memory/network bandwidth and latency).

5.1.2. Communication-Avoiding Block Coordinate Descent. Our derivation of the CA-BCD algorithm showed that by unrolling the vector update recurrences we can reduce the latency cost of the BCD algorithm by a factor of s . However, this comes at the cost of computing a larger $sb \times sb$ Gram matrix whose condition number is larger than the $b \times b$ Gram matrix computed in the BCD algorithm. The larger condition number implies that the CA-BCD algorithm may not be stable for $s > 1$ due to round-off error. We begin by experimentally showing the convergence behavior of the CA-BCD algorithm on the datasets in Table 4 with fixed block sizes of $b = 16$ for news20, a9a, and real-sim, respectively.

Figure 4 compares the convergence behavior of BCD and CA-BCD for $s > 1$.

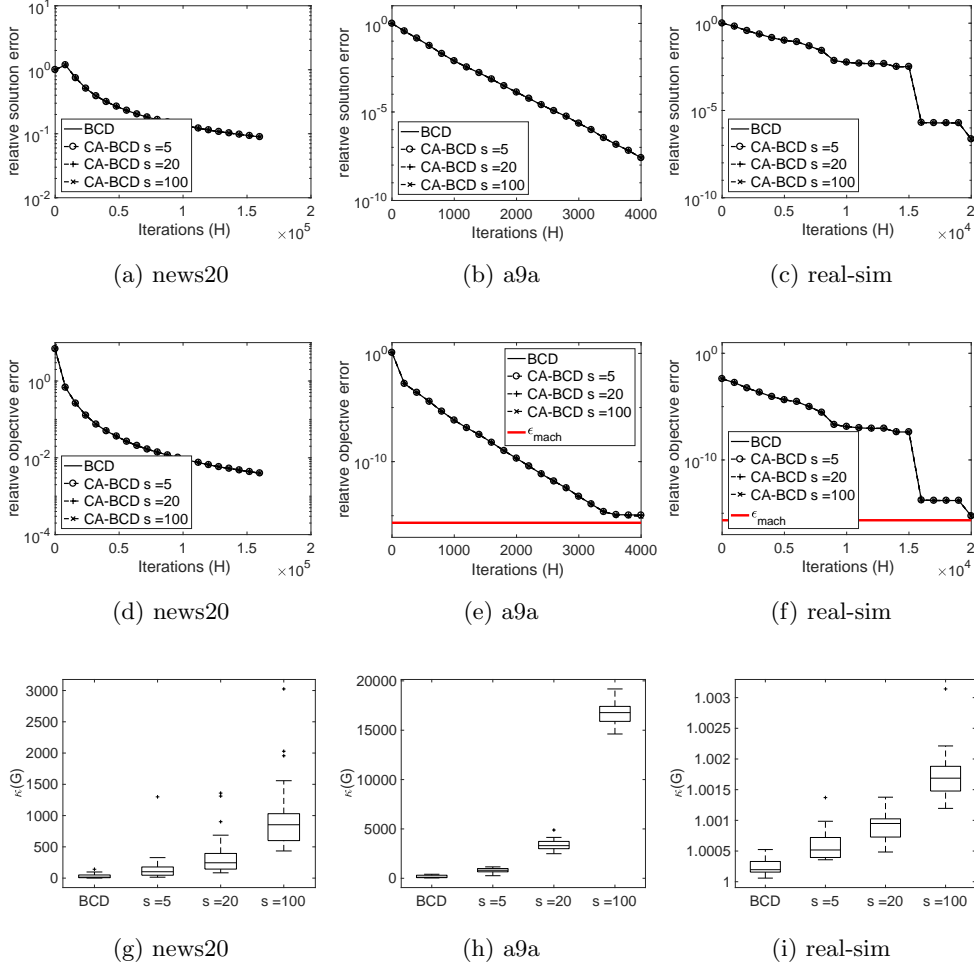


Fig. 4: We compare the convergence behavior of BCD and CA-BCD with several values of s . Relative solution error (top row, Figs. 4a-4c), relative objective error (middle row, Figs. 4d-4f), and statistics of the Gram matrix condition numbers (bottom row, Figs. 4g-4i) versus convergence. The block size for each dataset is set to $b = 16$. The boxplots (Figs. 4g - 4i) use standard MATLAB convention [1].

We plot the relative solution error, relative objective error and statistics of the Gram matrix condition numbers. The convergence plots indicate that CA-BCD shows almost no deviation from the BCD convergence. While the Gram matrix condition numbers increase with s for CA-BCD, those condition numbers are not so large as to significantly alter the numerical stability. Figures 4e and 4f show that the objective error converges very close to ϵ_{mach} . The well-conditioning of the real-sim dataset in addition to the regularization and small block size (relative to d) makes the Gram matrices almost perfectly conditioned. Based on these results, it is likely that the factor of s increase in flops and bandwidth will be the primary bottleneck.

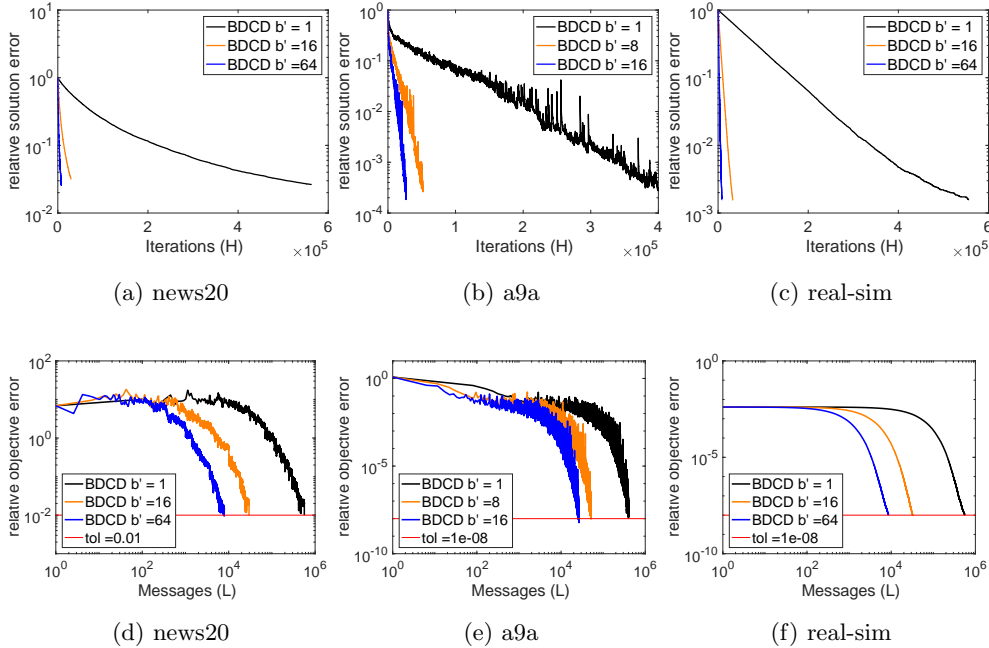


Fig. 5: We compare the convergence behavior of BDCD for several block sizes, b' , such that $1 \leq b' < n$ on several machine learning datasets. We show relative solution error (top row, Figs. 5a-5c) and objective error (bottom row, Fig. 5d-5f) convergence plots with $\lambda = 1000\sigma_{min}$. The x-axis for Figures 5d-5f show the number of messages required on a \log_{10} scale. Since BDCD communicates at every iteration, the x-axis is also equivalent to the number of iterations (modulo \log_{10} scale).

5.1.3. Block Dual Coordinate Descent. The BDCD algorithm solves the dual of the regularized least-squares problem by computing a $b' \times b'$ Gram matrix obtained from the columns of X (instead of the rows of X for BCD) and solves a b' -dimensional subproblem at each iteration. Similar to BCD, we expect that as b' increases, the BDCD algorithm converges faster at the cost of more flops and bandwidth. We explore this tradeoff space by comparing the convergence behavior (solution error and objective error) and algorithm costs for BDCD with $1 \leq b' < n$.

Figure 5 shows the convergence behavior on the datasets in Table 4 for various block sizes and measures the relative solution error (Figs. 5a-5c) and relative objective error (Figs. 5d-5f). Similar to BCD, as the block sizes increase the convergence rates of each dataset improves. However, unlike BCD, the objective error does not immediately decrease for some datasets (news20 and a9a). This is expected behavior since BDCD minimizes the dual objective (see Section 3.2) and obtains the primal solution vector, w_h , by taking linear combinations of b' columns of X and w_{h-1} . This also accounts for the non-monotonic decrease in the primal objective and primal solution errors.

Figure 6 shows the convergence behavior (in terms of the objective error) vs. flops and bandwidth costs of BDCD for the datasets and block sizes tested in Figure 5. We see that small block sizes are more flops and bandwidth efficient while large block

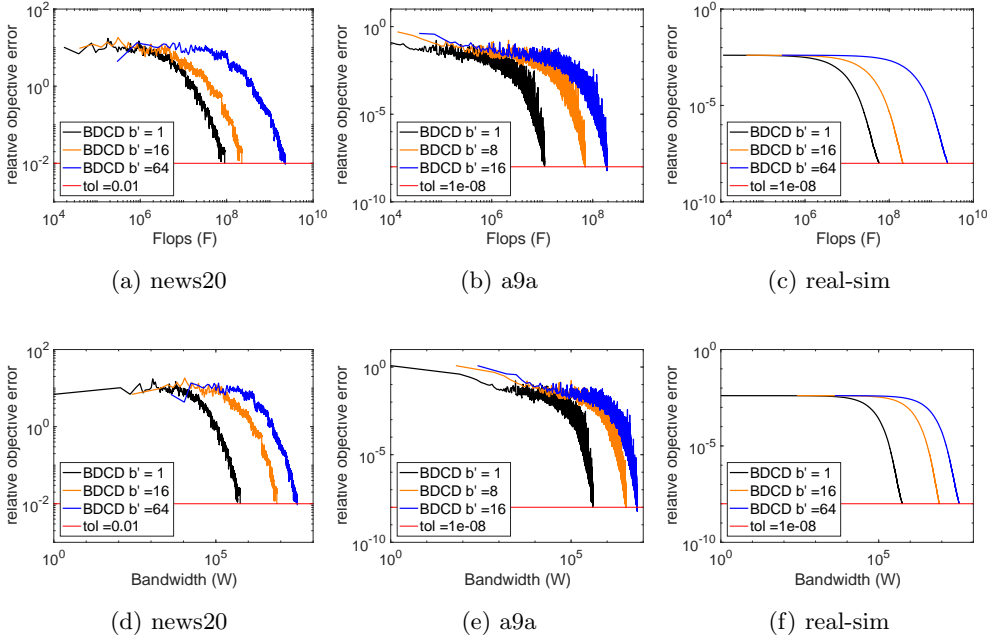


Fig. 6: We compare the convergence behavior of BDCD for several block sizes, b' , such that $1 \leq b' < n$ on several machine learning datasets. Flops cost (top row, Figs. 6a-6c), and bandwidth cost (middle row, Figs. 6d-6f) versus convergence with $\lambda = 1000\sigma_{min}$.

sizes are latency efficient (from Figs. 5d-5f). Due to this tradeoff it important to select block sizes that balance these costs based on machine-specific parameters.

5.1.4. Communication-Avoiding Block Dual Coordinate Descent. The CA-BDCD algorithm avoids communication in the dual problem by unrolling the vector update recurrences by a factor of s . This allows us to reduce the latency cost by computing a larger $sb' \times sb'$ Gram matrix instead of a $b' \times b'$ Gram matrix in the BDCD algorithm. The larger condition number implies that the CA-BDCD algorithm may not be stable, so we begin by experimentally showing the convergence behavior of the CA-BCD algorithm on the datasets in Table 4.

Figure 7 compares the convergence behavior of BDCD and CA-BDCD for $s > 1$ with block sizes of $b' = 64, 16$, and 64 for the news20, a9a and real-sim datasets, respectively. The results indicate that CA-BDCD is numerically stable for all tested values of s on all datasets. While the condition numbers of the Gram matrices increase with s , the numerical stability is not significantly affected. The well-conditioning of the real-sim dataset in addition to the regularization and small block size (relative to n) make the Gram matrices almost perfectly conditioned.

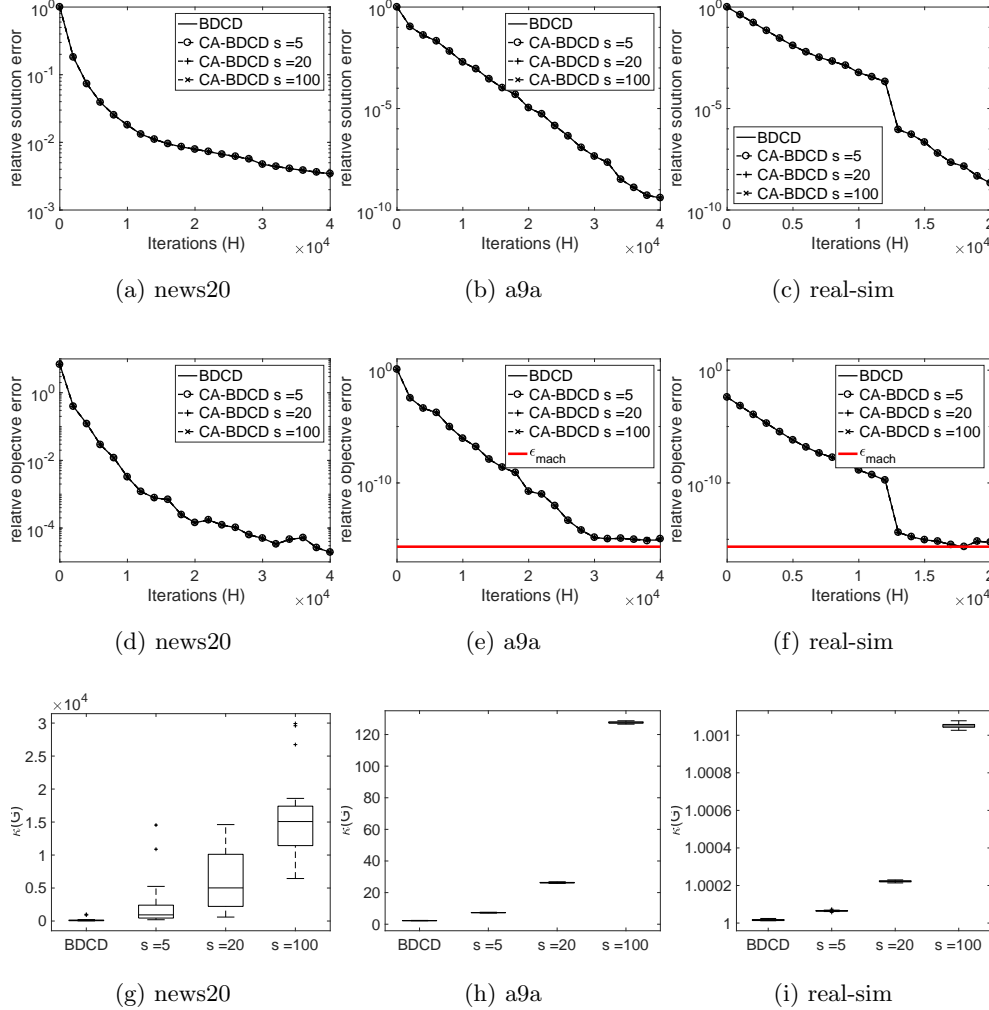


Fig. 7: We compare the convergence behavior of BDCD and CA-BDCD with several values of s . Relative solution error (top row, Figs. 7a-7c), relative objective error (middle row, Figs. 7d-7f), and statistics of the Gram matrix condition numbers (bottom row, Figs. 7g-7i) versus convergence. The block sizes for each dataset are: news20 with $b' = 64$, a9a with $b' = 16$, and real-sim with $b' = 64$.

5.1.5. Stopping Criterion. At iteration h of BCD, we solve a b -dimensional subproblem

$$\Delta w_h = \left(\frac{1}{n} \mathbb{I}_h^T X X^T \mathbb{I}_h + \lambda \mathbb{I}_h^T \mathbb{I}_h \right)^{-1} \left(-\lambda \mathbb{I}_h^T w_{h-1} - \frac{1}{n} \mathbb{I}_h^T X z_{h-1} + \frac{1}{n} \mathbb{I}_h^T X y \right).$$

Note that the b -dimensional vector, $(-\lambda \mathbb{I}_h^T w_{h-1} - \frac{1}{n} \mathbb{I}_h^T X z_{h-1} + \frac{1}{n} \mathbb{I}_h^T X y)$, is the subsampled primal residual vector and is explicitly computed at every iteration. Therefore, a natural stopping criteria is to occasionally compute the full-dimensional resid-

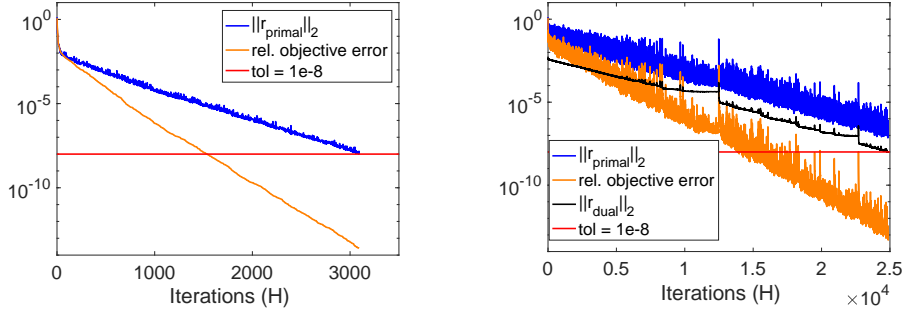
(a) BCD on a9a dataset ($b = 16$).(b) BDCD on a9a dataset ($b' = 16$).

Fig. 8: We plot the relative objective error, norm of the primal residual (for BCD Figure 8a), and norm of the dual residual (for BDCD Figure 8b) for the a9a dataset with block sizes $b = b' = 16$.

ual to check for convergence. Figure 8a illustrates the convergence of the residual in comparison to the relative objective error (the optimal objective value is obtained with Conjugate Gradients) for the a9a dataset with $b = 16$. Since the optimal objective value is, in general, unknown the residual can be used as an upper bound on the objective error.

At iteration h of BDCD, we solve the b' -dimensional subproblem

$$\Delta\alpha_h = -\frac{1}{n} \left(\frac{1}{\lambda n^2} \mathbb{I}_h^T X^T X \mathbb{I}_h + \frac{1}{n} \mathbb{I}_h^T \mathbb{I}_h \right)^{-1} \left(-\mathbb{I}_h^T X^T w_{h-1} + \mathbb{I}_h^T \alpha_{h-1} + \mathbb{I}_h^T y \right)$$

This b' -dimensional vector, $(-\mathbb{I}_h^T X^T w_{h-1} + \mathbb{I}_h^T \alpha_{h-1} + \mathbb{I}_h^T y)$, is the sub-sampled dual residual vector. One can similarly compute the full-dimensional dual residual occasionally to check for convergence. Figure 8b illustrates the convergence of the dual residual in comparison to the relative objective error, and the primal residual. We can observe that the dual residual is a lower bound on the primal residual, therefore, the dual problem should be solved to higher accuracy. In subsequent performance experiments we occasionally compute the primal residual for (CA)-BCD and the dual residual for (CA)-BDCD to test for convergence.

5.2. Performance Experiments. In Section 5.1 we showed tradeoffs between convergence behavior and algorithm costs for several datasets. In this section, we explore the performance tradeoffs of standard vs. CA variants on datasets obtained from LIBSVM [14]. We implemented these algorithms in C/C++ using Intel MKL for (sparse and dense) BLAS routines and MPI [22] for parallel processing. While Sections 4 and 5.1 assumed dense data for the theoretical analysis and numerical experiments, our parallel implementation stores the data in CSR (Compressed Sparse Row) format. We used a Cray XC30 supercomputer (“Edison”) at NERSC [2] to run

Algorithm	Name	Features (d)	Data Points (n)	NNZ%	residual tolerance (tol)
BCD	a9a	123	32561	11	1e-2
	covtype	54	581012	22	1e-1
	mnist8m	784	8100000	25	1e-1
BDCD	news20	62061	15935	0.13	1e-2
	e2006	150360	3308	0.93	1e-2
	rcv1	47236	3000	0.17	1e-3

Table 5: LIBSVM datasets used in our performance experiments.

our experiments on the datasets shown in Table 5. We used a 1D-column layout for datasets with $n > d$ and a 1D-row layout for $n < d$. We ensured that the parallel file I/O was load-balanced (i.e. each processor read roughly equal bytes) and found that the non-zero entries were reasonably well-balanced⁴. We constrain the running time of (CA-)BCD and (CA-)BDCD by fixing the residual tolerance for each dataset to the values described in Table 5. We ran many of these datasets for smaller tolerances of $1e - 8$ and found that our conclusions did not significantly change.

Section 5.2.1 compares the strong scaling behavior of the standard BCD and BDCD algorithms against their CA variants, Section 5.2.2 shows the running time breakdown to illustrate the flops vs. communication tradeoff, and Section 5.2.3 compares the speedups attained as a function of the number of processors, block size and recurrence unrolling parameter, s .

5.2.1. Strong Scaling. All strong scaling experiments were conducted with one MPI process per processor (flat-MPI) with one warm-up run and three timed runs. Each data point in Figure 9 represents the maximum running time over all processors averaged over the three timed runs. For each dataset in Figure 9 we plot the BCD running times, the fastest CA-BCD running times for $s \in \{2, 4, 8, 16, 32\}$, and the ideal scaling behavior. We show the scaling behavior of all datasets for $b \in \{1, 8\}$ to illustrate how the CA-BCD speedups are affected by the choice of block size, b . When the BCD algorithm is entirely latency dominated (i.e. Figure 9a), CA-BCD attains speedups between $3.6\times$ to $6.1\times$. When the BCD algorithm is flops and bandwidth dominated (i.e. Figure 9b), CA-BCD attains modest speedups between $1.2\times$ to $1.9\times$. The strong scaling behavior of the BDCD and CA-BDCD algorithms is shown in Figures 9c and 9d. CA-BDCD attains speedups between $1.6\times$ to $2.9\times$ when latency dominates and $1.1\times$ to $3.4\times$ when flops and bandwidth dominated.

While we did not experiment with weak scaling, we can observe from our analysis (in Section 4) that the BCD and BDCD algorithms achieve perfect weak scaling (in theory). It is likely that the CA-BCD and CA-BDCD algorithms would attain weak-scaling speedups by reducing the latency cost by a factor of s , if latency dominates.

5.2.2. Running Time Breakdown. Figure 10 shows the running time breakdown of BCD and CA-BCD for $s \in \{2, 4, 8, 16, 32\}$ on the mnist8m dataset. We plot the breakdowns for $b \in \{1, 8\}$ at scales of 64 nodes and 1024 nodes to illustrate CA-BCD tradeoffs for different flops vs. communication ratios. Figures 10a and 10b show the running time breakdown at 64 nodes for $b = 1$ and $b = 8$, respectively. In both cases flops dominate communication and most of the speedup for CA-BCD is due to faster flops. Since BCD with $b = 1$ is memory-bandwidth bound, CA-BCD with

⁴For datasets with highly irregular sparsity structure, additional load balancing is likely required but we leave this for future work.

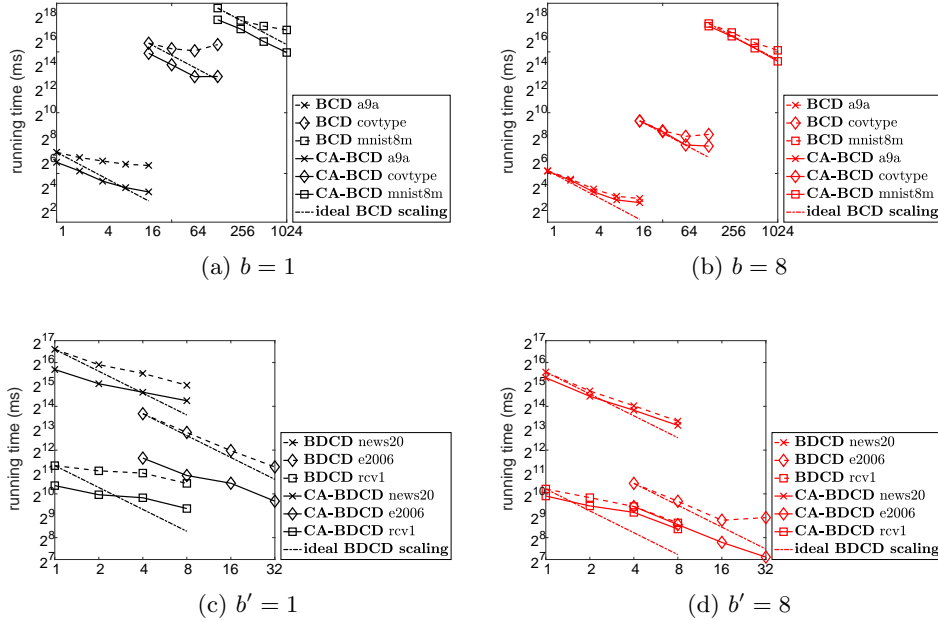


Fig. 9: Strong scaling results for BCD/CA-BCD (top row, Figs. 9a-9b) and BDCD/CA-BDCD (bottom row, Figs 9c-9d). We report the ideal strong scaling behavior for BCD and BDCD to illustrate the performance improvements gained from the communication-avoiding variants.

$s > 1$ increases the computational complexity and allows each processor to achieve higher flops performance through the use of BLAS-3 GEMM operations instead of BLAS-1 dot product operations. For $s \geq 8$ CA-BCD begins to saturate memory-bandwidth, therefore, speedup for $s > 8$ is due to reduction in communication time. For $b = 8$, memory-bandwidth is saturated at $s < 8$. The flops running time improves for $s < 8$ since the BLAS-3 calls can use larger, more cache-efficient tile sizes. For $s \geq 8$ CA-BCD becomes CPU-bound and does not attain any speedup over BCD. Furthermore, in the $b = 8$ setting, communication is more bandwidth dominated and less communication speedup is expected. On 1024 nodes (Figures 10c and 10d), where communication and latency costs are more dominant, CA-BCD attains larger communication and overall speedups. These experiments suggest that the CA-BCD and CA-BDCD algorithms, for appropriately chosen values of s , can attain large speedups when latency is the dominant cost.

5.2.3. Speedup Comparison. Figure 11 summarized the speedups attainable on the mnist8m dataset at 64 nodes and 1024 nodes for several combinations of block sizes (b) and recurrence unrolling values (s). We normalize the speedups to BCD with $b = 1$. At small scale (Figure 11a) we see speedups of $1.95\times$ to $2.91\times$ since flops and bandwidth are the dominant costs. The speedup for larger block sizes is due to faster convergence (i.e. fewer iterations and messages) and due to the use of BLAS-3 matrix-matrix operations. Even at small scale we see that CA-BCD is fastest for all block sizes tested. At large scale, when latency dominates, (Figure 11b) we observe

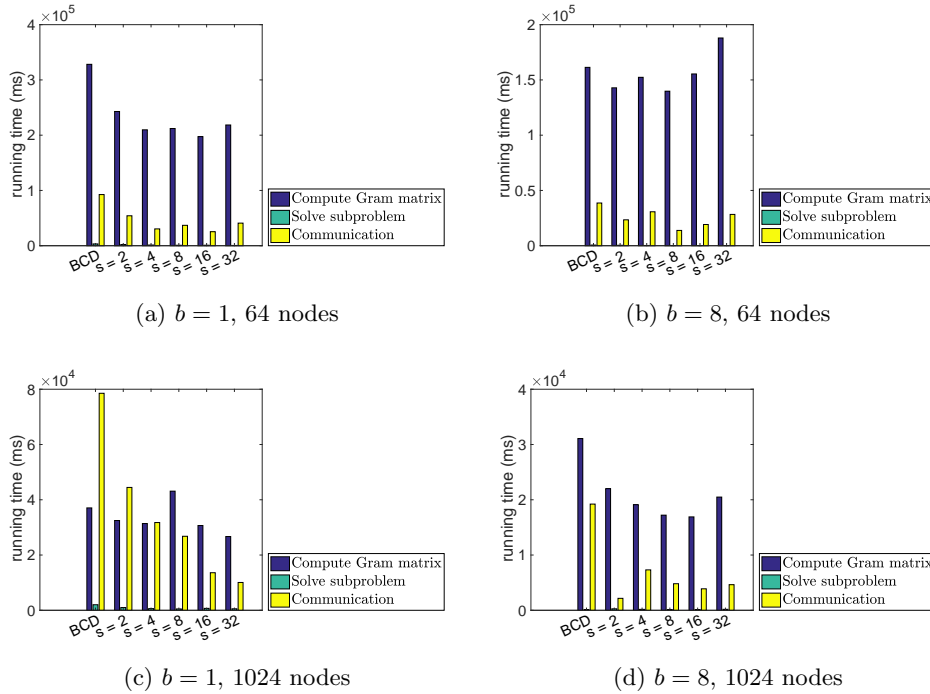


Fig. 10: Running time breakdown for the mnist8m dataset for $b = 1$ (left column, Figs. 10a-10c) and $b = 8$ (right column, Figs. 10b-10d). We report the breakdown for 64 nodes (top row, Figs. 10a-10b) and for 1024 nodes (bottom row, Figs. 10c-10d) using the fastest timed run for each algorithm and setting.

greater speedups of $3.62\times$ to $5.98\times$. Once again, we see that CA-BCD is fastest for all block sizes tested. From Figure 9b, we see that BCD and CA-BCD for mnist8m with $b = 8$ would likely scale beyond 1024 nodes. Therefore, we can expect greater speedups for $b = 8$, when latency becomes the dominant cost.

6. Conclusion and Future Work. In this paper, we have shown how to extend the communication-avoiding technique of CA-Krylov subspace methods to block coordinate descent and block dual coordinate descent algorithms in machine learning. We showed that in some settings, BCD and BDCD methods may converge faster than traditional Krylov methods – especially when the solution does not require high-accuracy. We analyzed the computation, communication and storage costs of the classical and communication-avoiding variants under two partitioning schemes. Our experiments showed that CA-BCD and CA-BDCD are numerically stable algorithms for all values of s tested, experimentally showed the tradeoff between algorithm parameters and convergence. Finally, we showed that the communication-avoiding variants can attain large speedups of up to $6.1\times$ on a Cray XC30 supercomputer using MPI.

While CA-BCD and CA-BDCD appear to be stable, numerical analysis of these methods and proofs of stability would be interesting directions for future work. Extending the communication-avoiding technique to other algorithms (SGD, L-BFGS, Newton’s method, etc.), regularization (LASSO, Elastic-net, etc.) and loss functions

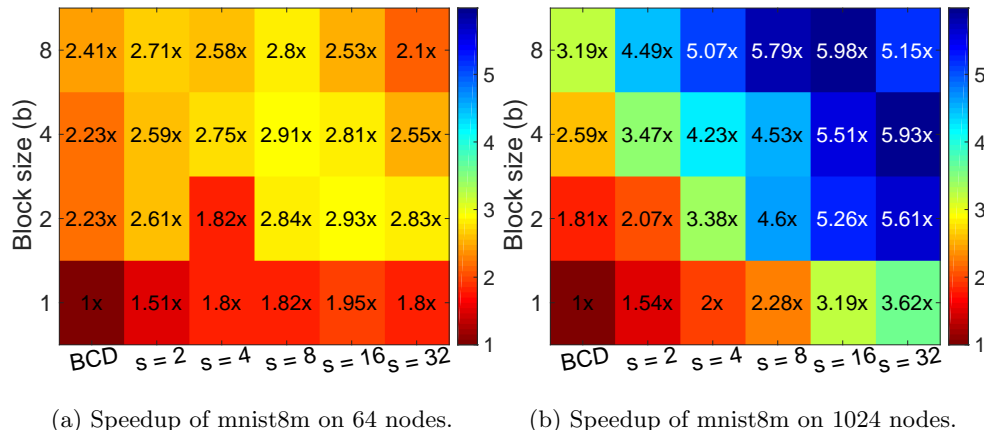


Fig. 11: Heatmaps of the speedups achieved for CA-BCD on the mnist8m dataset for various settings of b and s . On the left (Fig. 11a) we show speedups for 64 nodes and on the right (Fig. 11b) we show speedups for 1024 nodes.

(SVM, logistic, etc.) would be particularly interesting.

Acknowledgements. AD is supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE 1106400. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract Nos. DE-AC02-05CH11231, DE-SC0010200, and DE-SC0008700. This work is supported by Cray, Inc. under Grant No. 47277 and the Defense Advanced Research Projects Agency XDATA program. Research partially funded by ASPIRE Lab industrial sponsors and affiliates Intel, Google, Hewlett-Packard, Huawei, LGE, NVIDIA, Oracle, and Samsung.

REFERENCES

- [1] *Box plots*. <https://www.mathworks.com/help/stats/box-plots.html>.
- [2] *NERSC Edison configuration*. <http://www.nersc.gov/users/computational-systems/edison/configuration/>.
- [3] A. ALEXANDROV, M. F. IONESCU, K. E. SCHAUSER, AND C. SCHEIMAN, *LogGP: Incorporating long messages into the LogP model for parallel computation*, Journal of parallel and distributed computing, 44 (1997), pp. 71–79.
- [4] G. BALLARD, *Avoiding Communication in Dense Linear Algebra*, PhD thesis, EECS Department, University of California, Berkeley, Aug 2013.
- [5] G. BALLARD, E. CARSON, J. DEMMEL, M. HOEMMEN, N. KNIGHT, AND O. SCHWARTZ, *Communication lower bounds and optimal algorithms for numerical linear algebra*, Acta Numerica, 23 (2014), pp. 1–155.
- [6] Å. BJÖRCK, *Numerical Methods for Least Squares Problems*, Society for Industrial and Applied Mathematics, 1996.
- [7] L. BOTTOU, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of Computation Statistics, Springer, 2010, pp. 177–186.
- [8] J. BRUCK, C.-T. HO, S. KIPNIS, E. UPFAL, AND D. WEATHERSBY, *Efficient algorithms for all-to-all communications in multipoint message-passing systems*, IEEE Transactions on Parallel and Distributed Systems, 8 (1997), pp. 1143–1156.

- [9] E. CARSON, *Communication-Avoiding Krylov Subspace Methods in Theory and Practice*, PhD thesis, EECS Department, University of California, Berkeley, Aug 2015.
- [10] E. CARSON AND J. DEMMEL, *A residual replacement strategy for improving the maximum attainable accuracy of s-step krylov subspace methods*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 22–43.
- [11] E. CARSON AND J. W. DEMMEL, *Accuracy of the s-step lanczos method for the symmetric eigenproblem in finite precision*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 793–819.
- [12] E. CARSON, N. KNIGHT, AND J. DEMMEL, *Avoiding communication in nonsymmetric lanczos-based krylov subspace methods*, SIAM Journal on Scientific Computing, 35 (2013), pp. S42–S61.
- [13] E. CARSON, N. KNIGHT, AND J. DEMMEL, *An efficient deflation technique for the communication-avoiding conjugate gradient method*, Electronic Transactions on Numerical Analysis, 43 (2014), pp. 125–141.
- [14] C.-C. CHANG AND C.-J. LIN, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2 (2011), pp. 1–27.
- [15] A. CHRONOPOULOS AND C. GEAR, *On the efficient implementation of preconditioned s-step conjugate gradient methods on multiprocessors with memory hierarchy*, Parallel Computing, 11 (1989), pp. 37 – 53.
- [16] A. CHRONOPOULOS AND C. GEAR, *s-step iterative methods for symmetric linear systems*, Journal of Computational and Applied Mathematics, 25 (1989), pp. 153 – 168.
- [17] A. T. CHRONOPOULOS AND C. D. SWANSON, *Parallel iterative s-step methods for unsymmetric linear systems*, Parallel Computing, 22 (1996), pp. 623–641.
- [18] D. CULLER, R. KARP, D. PATTERSON, A. SAHAY, K. E. SCHAUSER, E. SANTOS, AND T. SUBRAMONIAN, R. AND VON EICKEN, *LogP: Towards a realistic model of parallel computation*, vol. 28, ACM, 1993.
- [19] T. A. DAVIS, S. RAJAMANICKAM, AND W. M. SID-LAKHDAR, *A survey of direct methods for sparse linear systems*, Acta Numerica, 25 (2016), p. 383566, <https://doi.org/10.1017/S0962492916000076>.
- [20] J. DEMMEL, L. GRIGORI, M. HOEMMEN, AND J. LANGOU, *Communication-avoiding parallel and sequential QR and LU factorizations*, SIAM Journal of Scientific Computing, (2008).
- [21] J. DEMMEL, M. HOEMMEN, M. MOHIYUDDIN, AND K. YELICK, *Avoiding communication in computing Krylov subspaces*, Tech. Report UCB/EECS-2007-123, EECS Department, University of California, Berkeley, Oct 2007.
- [22] M. P. I. FORUM, *MPI: A message-passing interface standard*, 1994.
- [23] K. FOUNTOLAKIS AND J. GONDZIO, *Performance of First- and Second-Order Methods for L1-Regularized Least Squares Problems*, ArXiv e-prints, (2015), <https://arxiv.org/abs/1503.03520>.
- [24] K. FOUNTOLAKIS AND R. TAPPENDEN, *Robust Block Coordinate Descent*, ArXiv e-prints, (2014), <https://arxiv.org/abs/1407.7573>.
- [25] S. H. FULLER AND L. I. MILLETT, *Computing performance: Game over or next level?*, Computer, (2011), pp. 31–38.
- [26] A. GITTESS, A. DEVARAKONDA, E. RACAH, M. RINGENBURG, L. GERHARDT, J. KOTTALAM, J. LIU, K. MASCHHOFF, S. CANON, J. CHHUGANI, P. SHARMA, J. YANG, J. DEMMEL, J. HARRELL, V. KRISHNAMURTHY, M. W. MAHONEY, AND PRABHAT, *Matrix factorizations at scale: A comparison of scientific data analytics in spark and c+mpi using three case studies*, in 2016 IEEE International Conference on Big Data (Big Data), Dec 2016, pp. 204–213.
- [27] G. H. GONNET, *Expected length of the longest probe sequence in hash code searching*, Journal of the ACM (JACM), 28 (1981), pp. 289–304.
- [28] S. L. GRAHAM, M. SNIR, AND C. A. PATTERSON, *Getting up to speed : the future of supercomputing*, National Academies Press, Washington, DC, 2005.
- [29] M. HOEMMEN, *Communication-avoiding Krylov subspace methods*, PhD thesis, University of California, Berkeley, 2010.
- [30] M. JAGGI, V. SMITH, M. TAKÁČ, J. TERHORST, S. KRISHNAN, T. HOFMANN, AND M. I. JORDAN, *Communication-efficient distributed dual coordinate ascent*, in Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS’14, Cambridge, MA, USA, 2014, MIT Press, pp. 3068–3076.
- [31] S. KIM AND A. CHRONOPOULOS, *An efficient nonsymmetric Lanczos method on parallel vector computers*, Journal of Computational and Applied Mathematics, 42 (1992), pp. 357 – 374.
- [32] K. LANG, *Newsweeder: Learning to filter netnews*, in Proceedings of the 12th International Machine Learning Conference, 1995.

- [33] M. LICHMAN, *UCI machine learning repository*, 2013, <http://archive.ics.uci.edu/ml>.
- [34] J. MAREČEK, P. RICHÁRIK, AND M. TAKÁČ, *Distributed block coordinate descent for minimizing partially separable functions*, in Numerical Analysis and Optimization, Springer, 2015, pp. 261–288.
- [35] A. MCCALLUM, *SRAA: Simulated/real/aviation/auto usenet data*. <https://people.cs.umass.edu/~mccallum/data.html>.
- [36] M. D. MITZENMACHER, *The Power of Two Choices in Randomized Load Balancing*, PhD thesis, EECS Department, University of California, Berkeley, 1996.
- [37] M. MOHIYUDDIN, *Tuning Hardware and Software for Multiprocessors*, PhD thesis, EECS Department, University of California, Berkeley, May 2012.
- [38] M. MOHIYUDDIN, M. HOEMMEN, J. DEMMEL, AND K. YELICK, *Minimizing communication in sparse matrix solvers*, in Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09, New York, NY, USA, 2009, ACM, pp. 36:1–36:12.
- [39] Y. NESTEROV, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 22 (2012), pp. 341–362.
- [40] M. RAAB AND A. STEGER, “balls into bins” – a simple and tight analysis, in Randomization and Approximation Techniques in Computer Science, Springer, 1998, pp. 159–170.
- [41] B. RECHT, C. RÉ, S. WRIGHT, AND F. NIU, *Hogwild: A lock-free approach to parallelizing stochastic gradient descent*, in Advances in Neural Information Processing Systems, 2011, pp. 693–701.
- [42] P. RICHÁRIK AND M. TAKÁČ, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 144 (2014), pp. 1–38.
- [43] Y. SAAD, *Iterative methods for sparse linear systems*, SIAM, 2003.
- [44] S. SHALEV-SHWARTZ AND T. ZHANG, *Stochastic dual coordinate ascent methods for regularized loss*, The Journal of Machine Learning Research, 14 (2013), pp. 567–599.
- [45] E. SOLOMONIK, *Provably efficient algorithms for numerical tensor algebra*, PhD thesis, EECS Department, University of California, Berkeley, Aug 2014.
- [46] J. STAMPER, A. NICULESCU-MIZIL, S. RITTER, G. GORDON, AND K. KOEDINGER, *Algebra 2008-2009 from challenge data set*, in KDD Cup 2010 Educational Data Mining Challenge, 2010.
- [47] M. TAKÁČ, P. RICHÁRIK, AND N. SREBRO, *Distributed mini-batch SDCA*, CoRR, abs/1507.08322 (2015).
- [48] R. THAKUR AND W. D. GROPP, *Improving the performance of MPI collective communication on switched networks*, (2002).
- [49] R. THAKUR AND W. D. GROPP, *Improving the performance of collective operations in mpich*, in Recent Advances in Parallel Virtual Machine and Message Passing Interface, Springer, 2003, pp. 257–267.
- [50] J. VAN ROSENDALE, *Minimizing inner product data dependencies in conjugate gradient iteration*, IEEE Computer Society Press, Silver Spring, MD, Jan 1983.
- [51] H. F. WALKER, *Implementation of the GMRES method using Householder transformations*, SIAM Journal on Scientific and Statistical Computing, 9 (1988), pp. 152–163.
- [52] S. WILLIAMS, M. LIJEWSKI, A. ALMGREN, B. VAN STRAALLEN, E. CARSON, N. KNIGHT, AND J. DEMMEL, *s-step Krylov subspace methods as bottom solvers for geometric multigrid*, in Parallel and Distributed Processing Symposium, 2014 IEEE 28th International, IEEE, 2014, pp. 1149–1158.
- [53] S. J. WRIGHT, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34.
- [54] H.-F. YU, H.-Y. LO, H.-P. HSIEH, J.-K. LOU, T. G. MCKENZIE, J.-W. CHOU, P.-H. CHUNG, C.-H. HO, C.-F. CHANG, J.-Y. WENG, E.-S. YAN, C.-W. CHANG, T.-T. KUO, P. T. CHANG, C. PO, C.-Y. WANG, Y.-H. HUANG, Y.-X. RUAN, Y.-S. LIN, S.-D. LIN, H.-T. LIN, AND C.-J. LIN, *Feature engineering and classifier ensemble for kdd cup 2010*, in JMLR Workshop and Conference Proceedings, 2011.
- [55] Y. ZHANG, M. J. WAINWRIGHT, AND J. C. DUCHI, *Communication-efficient algorithms for statistical optimization*, in Advances in Neural Information Processing Systems, 2012, pp. 1502–1510.
- [56] M. ZINKEVICH, M. WEIMER, L. LI, AND A. J. SMOLA, *Parallelized stochastic gradient descent*, in Advances in Neural Information Processing Systems, 2010, pp. 2595–2603.