SCALABLE L1-REGULARIZED LOGISTIC REGRESSION:

IRLS-LARS

Andrea Klein & Dipan Kumar Pal

* We wish to do binary classification on large datasets via LOGISTIC REGRESSION with an L1 norm.

* Recall that logistic regression models the probability that an element x belongs to a class y as a sigmoid function:

$$p(y = 1|x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + exp(-\theta^T x)}$$

* UNDER A LAPLACIAN PRIOR, THE MAP ESTIMATE FOR THE MODEL PARAMETERS (THETA) IS GIVEN BY:

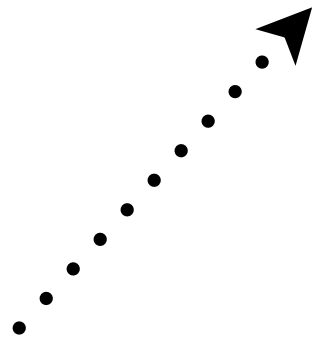$$\min_{\theta} \sum_{i=1}^{M} -\log p(y^{(i)}|\mathbf{x}^{(i)};\theta) + \beta\|\theta\|_1$$

* TECHNICALLY WE SOLVE AN EQUIVALENT ALTERNATIVE PARAMETRIZATION, SUBJECT TO:
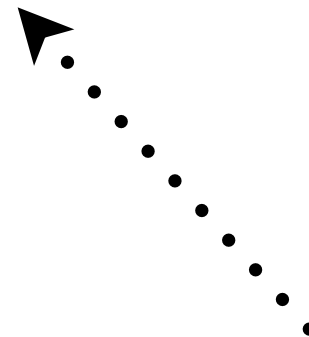
$$\|\theta\|_1 \leq C$$

* REGARDLESS OF HOW YOU FORMULATE THE PROBLEM, IT'S ORDINARILY INEFFICIENT TO FIND THETA FOR LARGE DATASETS.

IRLS-LARS

ITERATIVELY RE-WEIGHTED LEAST SQUARES

LEAST ANGLE REGRESSION

IRLS-LARS

ITERATIVELY RE-WEIGHTED LEAST SQUARES

LEAST ANGLE REGRESSION

* ON EACH STEP OF THE ALGORITHM, WE CHARACTERIZE THE UPDATE DIRECTION (GAMMA) OF THETA AS THE SOLUTION TO A **LEAST-SQUARES PROBLEM**:

$$\gamma^{(k)} = \arg \min_{\gamma} \| (\boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{X}^{\top}) \gamma - \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{z} \|_2^2$$

IRLS-LARS

$$\mathbf{X} = [\mathbf{x}^{(1)}\mathbf{x}^{(2)}\ldots\mathbf{x}^{(M)}]$$

UPDATED ("RE-WEIGHTED") ON EACH ITERATION:

$$\Lambda_{ii} = \sigma(\theta^{(k)\top}\mathbf{x}^{(i)})[1 - \sigma(\theta^{(k)\top}\mathbf{x}^{(i)})],$$

$$z_i = \mathbf{x}^{(i)\top}\theta^{(k)} + \frac{[1 - \sigma(y^{(i)}\theta^{(k)\top}\mathbf{x}^{(i)})]y^{(i)}}{\Lambda_{ii}}.$$

ITERATIVELY   ANGLE
LEAST SQUARES   REGRESSION

★ ON EACH STEP OF THE ALGORITHM, WE CHARACTERIZE THE UPDATE DIRECTION (GAMMA) OF THETA AS THE SOLUTION TO A LEAST-SQUARES PROBLEM:

$$\gamma^{(k)} = \arg\min_{\gamma} \|(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{X}^{\top})\gamma - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z}\|_2^2$$

IRLS-LARS
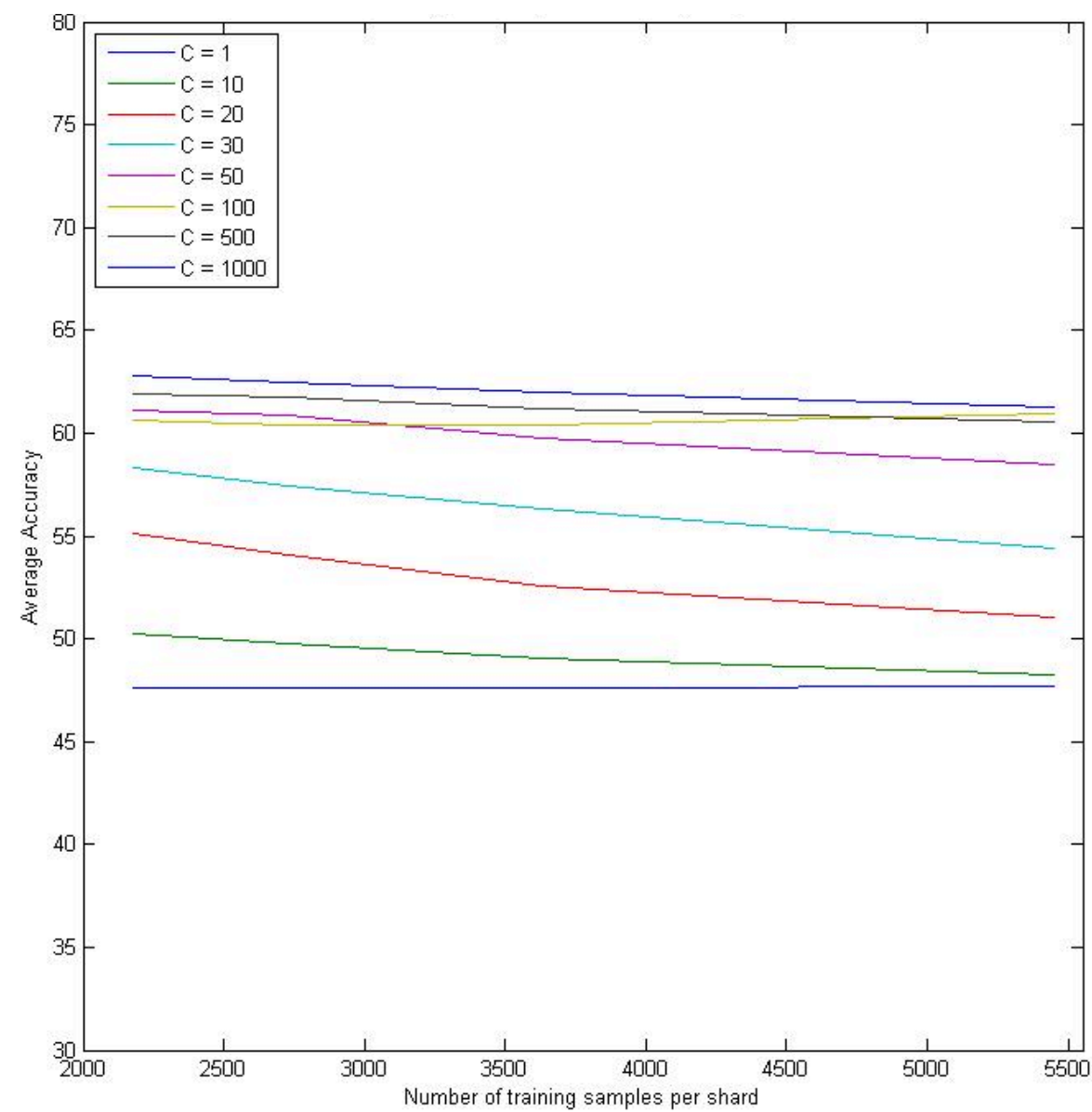
ITERATIVELY RE-WEIGHTED LEAST SQUARES

LEAST ANGLE REGRESSION

* IN PRACTICE, WE USE AN ALGORITHM CALLED LARS TO EFFICIENTLY SOLVE FOR THE STEP DIRECTION WITHIN THE IRLS FORMULATION. FINALLY, ONCE WE KNOW THE DIRECTION OF THE UPDATE, WE USE A BACKTRACKING LINE SEARCH TO DETERMINE HOW BIG A STEP TO TAKE.
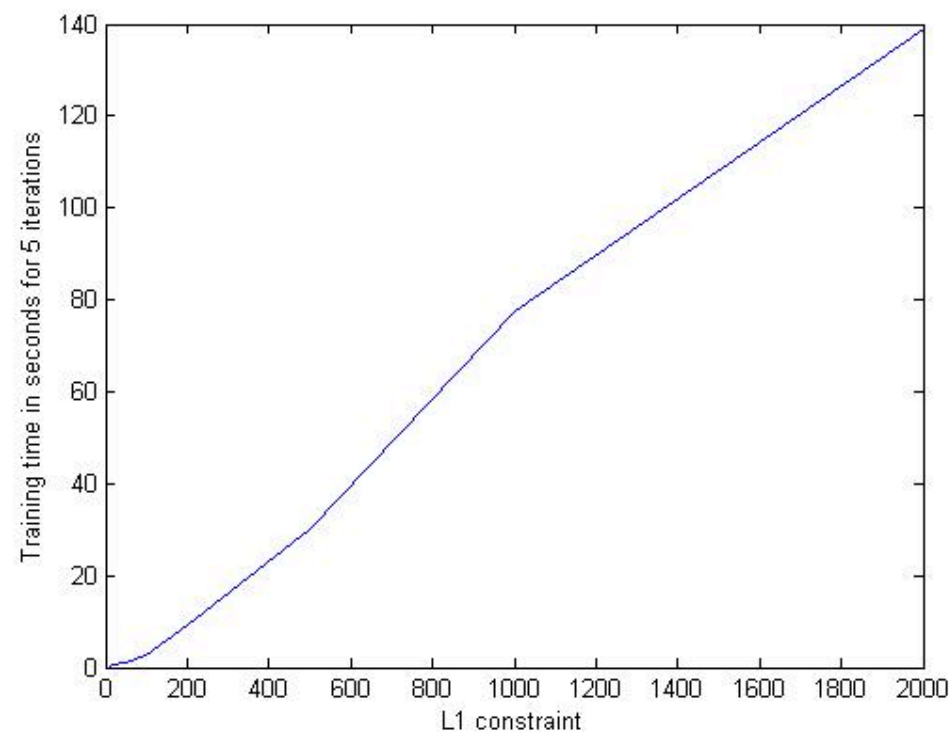
PERFORMANCE

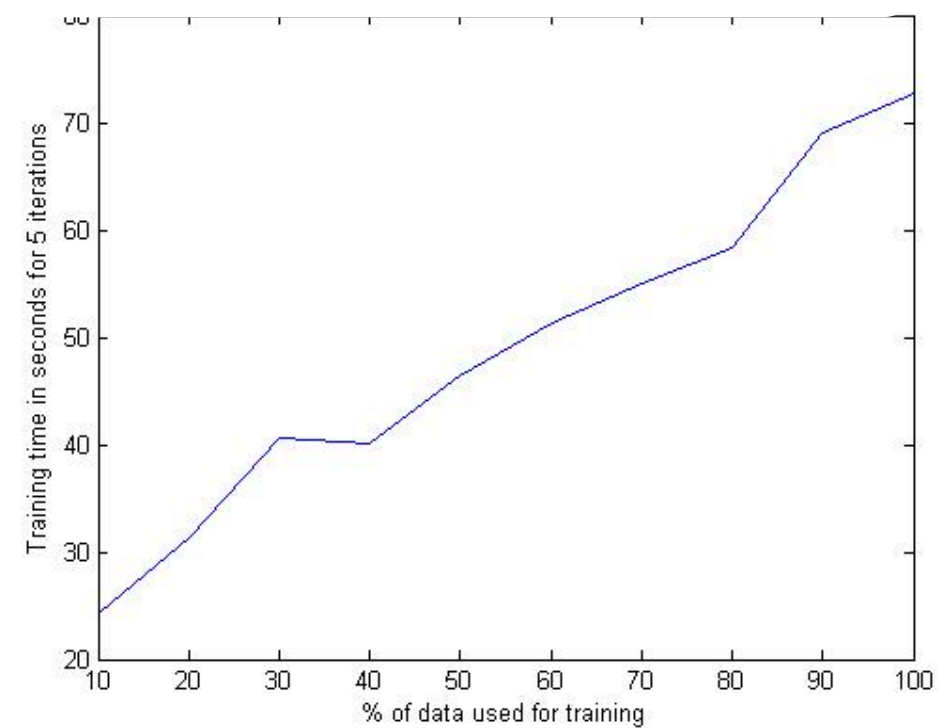## Average accuracy vs. # of training samples

PERFORMANCE

Training Time vs. Constraint Size

Training Time vs. Training Size

Algorithm 2 ParallelSGD($\{c^1, \ldots c^m\}, T, \eta, w_0, k$)

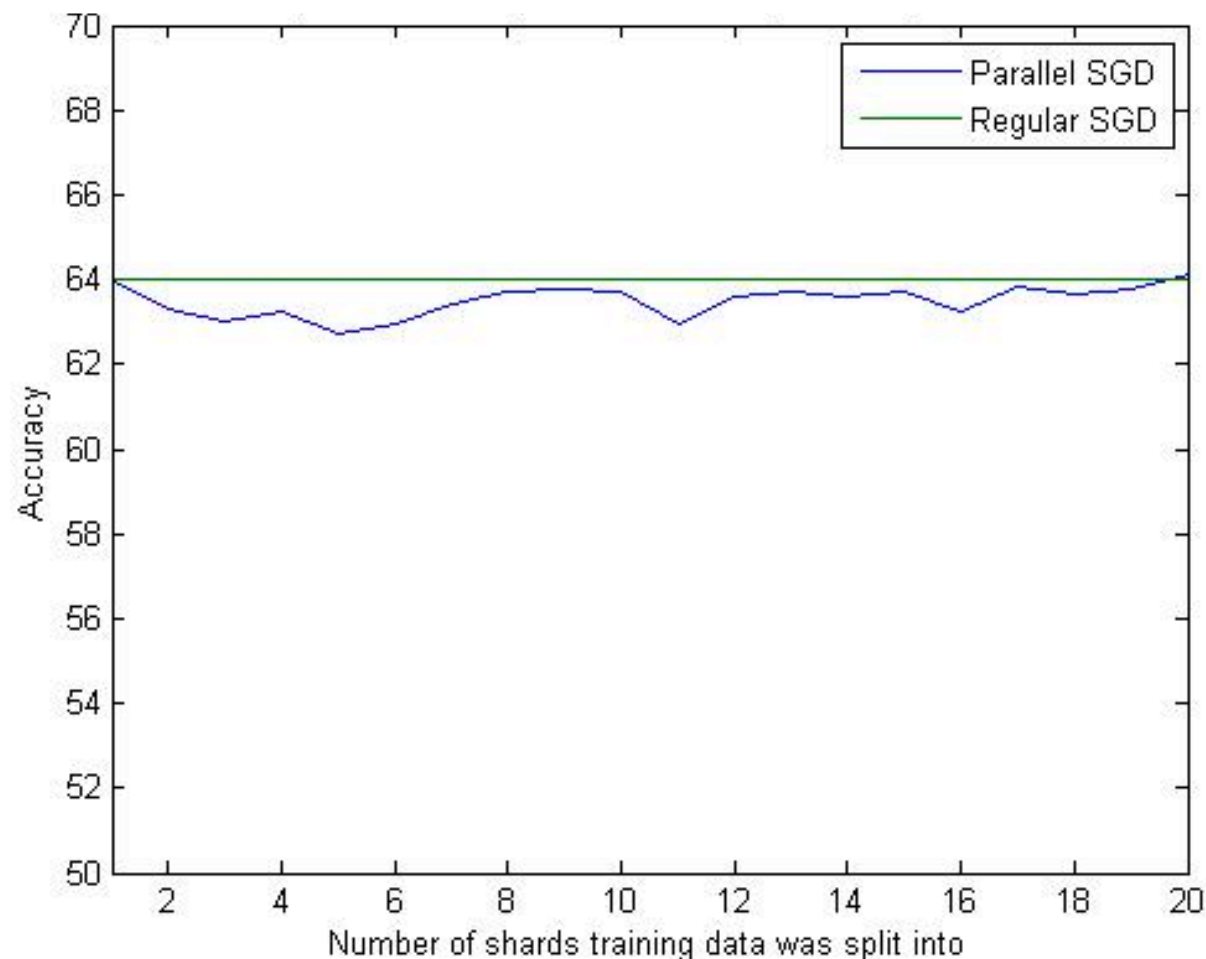**for all** $i \in \{1, \ldots k\}$ **parallel do**
  $v_i = \text{SGD}(\{c^1, \ldots c^m\}, T, \eta, w_0)$ on client
end for
Aggregate from all computers $v = \frac{1}{k} \sum_{i=1}^{k} v_i$ and **return** v

Relative Performance of Parallel vs. Regular SGD

* More strenuous (parallel) tests on larger datasets

* If time, we'd like to explore the algorithm's performance on more diverse datasets

* We may also incorporate parallel SCD (sequential coordinate descent), parallelizing over features rather than samples:

Algorithm 2 Shotgun: Parallel SCD

Choose number of parallel updates $P \geq 1$.
Set $\mathbf{x} = \mathbf{0} \in \mathbb{R}_+^{2d}$
while not converged **do**
    Choose random subset of P weights in $\{1, \ldots, 2d\}$.
    In parallel on P processors
        Get assigned weight j.
        Set $\delta x_j \longleftarrow \max\{-x_j, -(\nabla F(\mathbf{x}))_j/\beta\}$.
        Update $x_j \longleftarrow x_j + \delta x_j$.
end while

THE END!

REFERENCES:

* EFFICIENT L1 REGULARIZED LOGISTIC REGRESSION (LEE ET AL)

* PARALLELIZED STOCHASTIC GRADIENT DESCENT (ZINKEVICH ET AL)

* PARALLEL COORDINATE DESCENT FOR L1 REGULARIZED LOSS MINIMIZATION (BRADLEY ET AL)

* LEAST ANGLE REGRESSION (EFRON ET AL)

$$\mathbf{X} = [\mathbf{x}^{(1)}\mathbf{x}^{(2)}\ldots\mathbf{x}^{(M)}].$$

Let the diagonal matrix $\mathbf{\Lambda}$ and the vector $\mathbf{z}$ be defined as follows: for all $i = 1, 2, \ldots, M$:

$$\Lambda_{ii} = \sigma(\theta^{(k)\top}\mathbf{x}^{(i)})[1 - \sigma(\theta^{(k)\top}\mathbf{x}^{(i)})], \tag{7}$$

$$z_i = \mathbf{x}^{(i)\top}\theta^{(k)} + \frac{[1 - \sigma(y^{(i)}\theta^{(k)\top}\mathbf{x}^{(i)})]y^{(i)}}{\Lambda_{ii}}. \tag{8}$$

Then, we have that $\mathbf{H}(\theta^{(k)}) = -\mathbf{X}\mathbf{\Lambda}\mathbf{X}^\top$ and $\mathbf{g}(\theta^{(k)}) = \mathbf{X}\mathbf{\Lambda}(\mathbf{z} - \mathbf{X}^\top\theta^{(k)})$, and thus Equation (5) can be rewritten as:

$$\gamma^{(k)} = (\mathbf{X}\mathbf{\Lambda}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{\Lambda}\mathbf{z}. \tag{9}$$

Thus $\gamma^{(k)}$ is the solution to the following weighted least squares problem:

$$\gamma^{(k)} = \arg\min_{\gamma} \|(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{X}^\top)\gamma - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z}\|_2^2. \tag{10}$$