Distributed ML: Reducing Communication with ADMM

Venkata Krishna Pillutla*#, SakethaNath J*

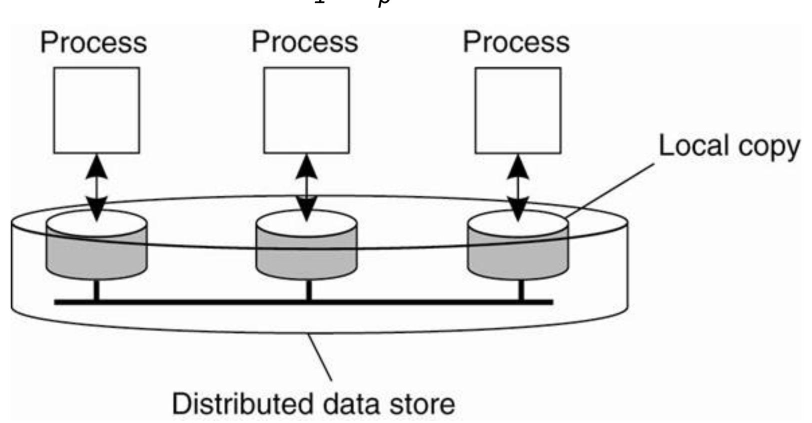
pillutla@cs.cmu.edu, saketh@iitb.ac.in

***Indian Institute of Technology Bombay**

#Carnegie Mellon University

Introduction

- Distributed computing environment with p nodes
- Data is distributed as $D_1 \ldots D_p$



- Global problem is regularized loss minimization:

$$\min_{\mathbf{w}} f(\mathbf{w}) \equiv \min_{\mathbf{w}} \sum_{i=1}^{p} f_i(\mathbf{w})$$

where l is a convex loss function and

$$f_i(\mathbf{w}) = \underbrace{\sum_{j \in \mathcal{D}_i} l(y_j, \mathbf{w}^T \mathbf{x_j})}_{\text{loss over local examples}} + \underbrace{\lambda R(\mathbf{w})}_{\text{regularizer}}$$

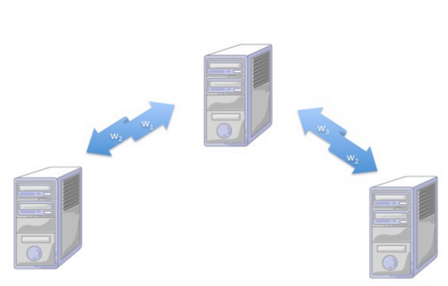
- Examples:
  - Classification: y_j is discrete
  - Regression: y_j is real-valued
- No shared memory: communication is required

Prior Work

- Recast as constrained optimization problem:

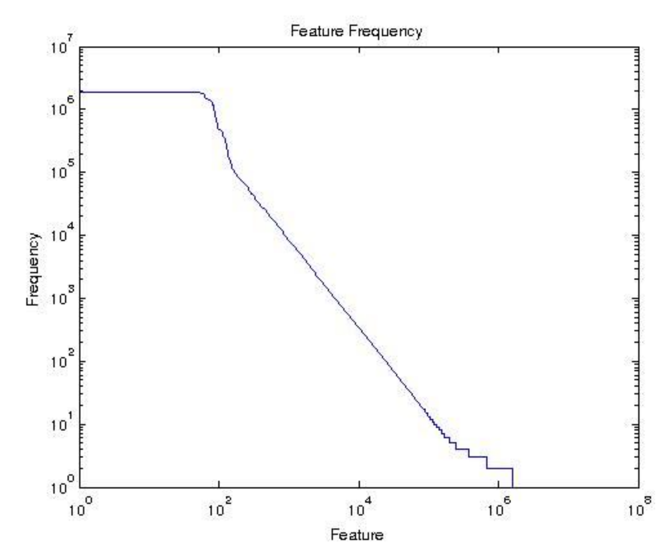
$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_p} \sum_{i=1}^{p} f_i(\mathbf{w}_i)$$
$$\text{s.t. } \mathbf{w}_1 = \mathbf{w}_2 = \ldots = \mathbf{w}_p$$

- Solve using Alternative Direction Method of Multipliers (ADMM)
  - Convex objective and linear constraints
  - Using Lagrange Duality
- Convergence guarantees are known
- Easily parallelized [1,2]
- Each iteration requires *O(d)* communication
  - Not practical for large scale data



Motivation

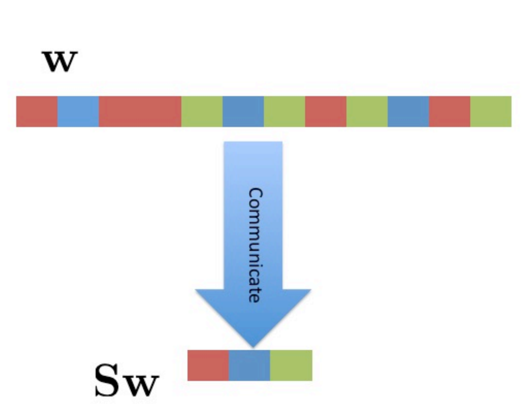
- Not all features are equally important
- Most real, big datasets are skewed
  - Feature occurrence frequencies follow approx power laws



- Features are strongly correlated
  - Examples: NLP (bag of words): synonyms
  - Examples: Vision: Nearby pixels
- Intuition: Giving equal importance is wasteful
- Figure: Feature occurrence

Approach

- Project features onto a lower *d'*-dimensional subspace *for communication*
- In other words, ***relax constraints***
- For instance:
  - Subsample features randomly
  - Averages of some groups of features
  - Features with highest variance
- Formally, define the projection via $\mathbf{S} \in \mathbb{R}^{d' \times d}$



- The new problem is now:

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_p} \sum_{i=1}^{p} f_i(\mathbf{w}_i)$$
$$\text{s.t. } \mathbf{Sw}_1 = \mathbf{Sw}_2 = \ldots = \mathbf{Sw}_p$$

Algorithm

- Solve using ADMM
- Dual variables: $\boldsymbol{\lambda} \in \mathbb{R}^{d'}$
- Augmented Lagrangian Parameter: ρ
- Iterations:

$$\mathbf{c}^{(t+1)} = \boldsymbol{\lambda}_i^{(t)} - \boldsymbol{\lambda}_{i-1}^{(t)} - \rho\left(\mathbf{Sw}_i^{(t)} + \frac{\mathbf{Sw}_{i-1}^{(t)} + \mathbf{Sw}_{i+1}^{(t)}}{2}\right)$$
$$\mathbf{w}_i^{(t+1)} = \operatorname{argmin}_{\mathbf{w}} \{f_i(\mathbf{w}) + \rho\|\mathbf{Sw}\|_2^2 + \mathbf{w}^T\mathbf{S}^T\mathbf{c}^{(t+1)}\}$$
$$\boldsymbol{\lambda}_i^{(t+1)} = \boldsymbol{\lambda}_i^{(t)} + \frac{\rho}{2}(\mathbf{Sw}_i^{(t+1)} - \mathbf{Sw}_{i+1}^{(t+1)})$$

- Communication required: *O(d')* per iteration

Theoretical Result

- Ridge Regression
- Assume: $|y_i| \le \beta, \|\mathbf{x}\|_\infty \le R$
- Notation:
  - Relaxed optimum: $\mathbf{w_S}$
  - Original optimum: $\mathbf{w}^*$
- *Theorem*:

$$\|\mathbf{w_S} - \mathbf{w}^*\| \le \sqrt{d - d'}\sqrt{p}(p-1)(c_1 + \sqrt{d}\|\mathbf{w}^*\|c_2)$$

- Proof by Linear Algebra
- Always true
- Can get a better high probability result
- Additionally, assume gradient of *f* exists and is Lipschitz
- *Theorem*:

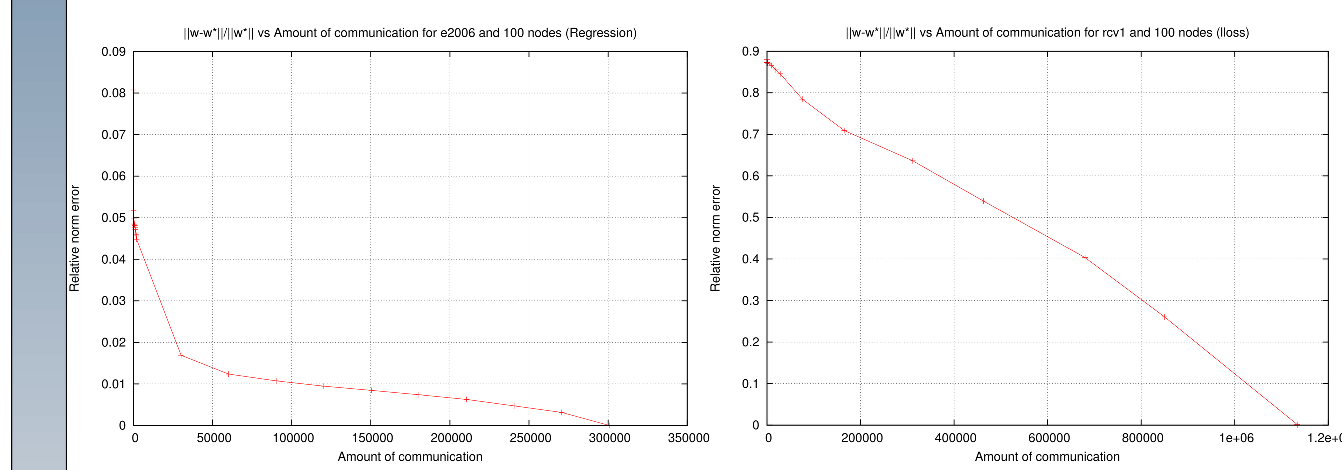
$$f(\mathbf{w_S}) - f(\mathbf{w}^*) \le c_3(d - d')(p-1)^2$$

Experiments

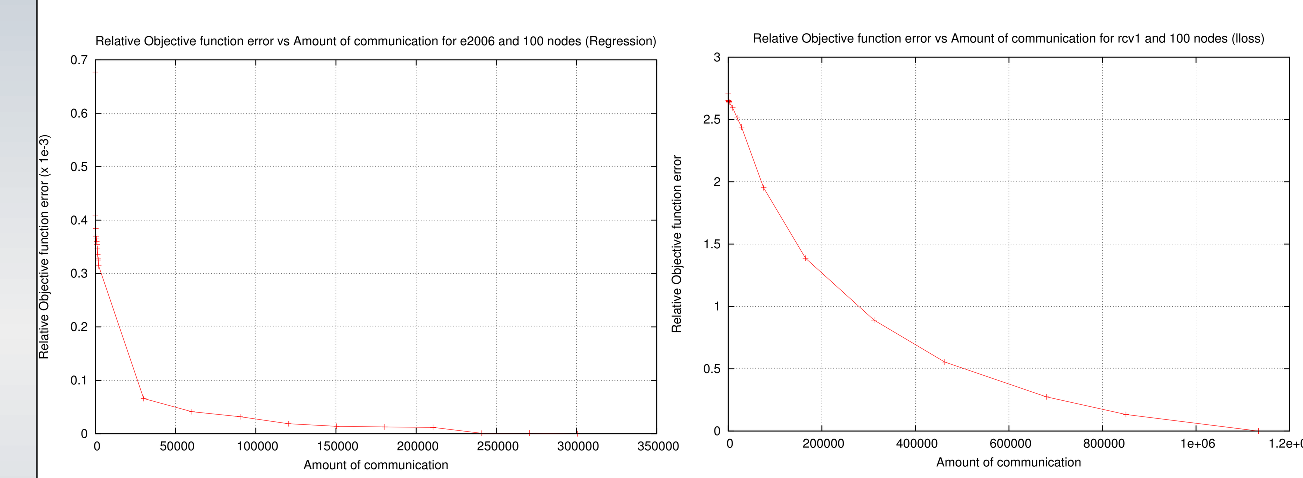
- Predicting financial volatility from financial reports [4](Regression)
  - E2006-tfidf
  - 27k documents with 10k words in each document
- Text categorization, Reuters Corpus Volume 1 [3](Classification)
  - 800k newswire stories

- Inner optimization used is gradient descent
- Loss function
  - Squared loss for regression
  - Logistic Regression for classification
- Best values of parameters chosen by cross-validation
- Stopping condition: Primal and Dual Residuals are small

$$\frac{\|\mathbf{w} - \mathbf{w}^*\|}{\|\mathbf{w}^*\|} \text{ vs communication}$$



$$\frac{f(\mathbf{w}) - f(\mathbf{w}^*)}{f(\mathbf{w}^*)} \text{ vs. communication}$$



Conclusions

- Can reduce communication cost up to 90% and still do well!
- Performance guarantees
- Works on real datasets

Bibliography

1. D.P. Bertsekas and J.N. Tsitsiklis. Parallel and Distributed Computation: Numerical Methods. 1997.
2. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. 2011.
3. M.-R. Amini, N. Usunier, C. Goutte. Learning from Multiple Partially Observed Views- an Application to Multilingual Text Categorization. *Advances in Neural Information Processing Systems* 22, p. 28-36, 2009.
4. S. Kogan, D. Levin, B.R. Routledge, J.S. Sagi, N.A. Smith. Predicting risk from financial reports with regression. In *In Proceedings of the North American Association of Computational Linguistics Human Language Technologies Conference*, pages 272-280, 2009.