# Distributed Training for Large-scale Logistic Models

Siddharth Gopal

Carnegie Mellon Univeristy

21 Aug 2013

[1] Joint work with Yiming Yang presented at ICML'13

# Outline of the Talk

- Logistic Models
- Maximum Likelihood Estimation
- Parallelization
- Experiments

# Logistic Models

Logistic Models model probability of an outcome $Y$ given a predictor $x$.

$$P(Y = y|x; \mathbf{w}) \propto \exp(\mathbf{w}^\top \phi(y, x))$$

Subsumes Multinomial Logistic Regression, Conditional Random fields and Maximum entropy Models.
For example, in Multinomial Logistic Regression

$$P(Y = k|x; \mathbf{w}) = \frac{\exp(w_k^\top x)}{\sum_j \exp(w_j^\top x)}$$

# Focus of the Talk

Train Logistic models on large-scale data.

What is Large-scale ?

- Large number of Training Examples
- High dimensionality
- Large number of Outcomes

# Focus of the Talk

Train Logistic models on large-scale data.

What is Large-scale ?

- Large number of Training Examples
- **High dimensionality**
- **Large number of Outcomes**

# Motivation

Some commonly used data on the web,

| Dataset | #Instances | #Labels | #Features | #Parameters |
|---|---|---|---|---|
| ODP subset | 93,805 | 12,294 | 347,256 | 4,269,165,264 |
| Wikipedia subset | 2,365,436 | 325,056 | 1,617,899 | 525,907,777,344 |
| Image-net | 14,197,122 | 21,841 | - | - |

# Motivation

Some commonly used data on the web,

| Dataset | #Instances | #Labels | #Features | #Parameters |
|---|---|---|---|---|
| ODP subset | 93,805 | 12,294 | 347,256 | 4,269,165,264 |
| Wikipedia subset | 2,365,436 | 325,056 | 1,617,899 | 525,907,777,344 |
| Image-net | 14,197,122 | 21,841 | - | - |

- How can we parallelize the training of such models ?
- How can we optimize different subsets of parameters simultaneously ?

# Maximum Likelihood Estimation (MLE)

Typical MLE estimation

- $N$ training examples, $K$ classes.
- $x_i$ denotes the $i^{th}$ training example.
- Indicator variable $y_{ik}$ denotes whether $x_i$ belongs to class $k$.
- Estimate parameters $\mathbf{w}$ by maximizing the log-likelihood,

$$\max_{\mathbf{w}} \quad \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log P(y_{ik} | x_i ; \mathbf{w}) - \frac{\lambda}{2} \|\mathbf{w}\|^2$$

# Maximum Likelihood Estimation (MLE)

Typical MLE estimation

- $N$ training examples, $K$ classes.
- $x_i$ denotes the $i^{th}$ training example.
- Indicator variable $y_{ik}$ denotes whether $x_i$ belongs to class $k$.
- Estimate parameters $\mathbf{w}$ by maximizing the log-likelihood,

$$\max_{\mathbf{w}} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log P(y_{ik}|x_i; \mathbf{w}) - \frac{\lambda}{2} \|\mathbf{w}\|^2$$

$$\textbf{[OPT1]} \quad \min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log \left( \sum_{k=1}^{K} \exp(w_k^\top x_i) \right)$$

# Parallelization

$$\min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right)$$

# Parallelization

$$\min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K}\exp(w_k^\top x_i)\right)$$

- The log-sum-exp (LSE) function couples all the class-level parameter $w_k$'s together.

# Parallelization

$$\min_{\mathbf{w}} \; \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K}\exp(w_k^\top x_i)\right)$$

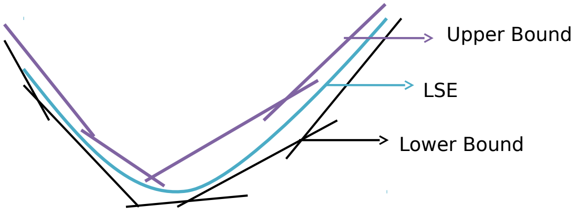
- The log-sum-exp (LSE) function couples all the class-level parameter $w_k$'s together.
- Replace LSE by a parallelizable function
  - This parallelizable function should be an upper-bound
  - It should not make the optimization harder - like introduce non-convexity, non-differentiability etc.

Properties used

- LSE is a convex-function
- Convex function can be approximated to any precision by piecewise linear functions.

$$\max_{j}\{a_j^\top \gamma + b_j\} \leq \log\left(\sum_{k=1}^{K} \exp(\gamma_k)\right) \leq \max_{j'}\{c_{j'}^\top \gamma + d_{j'}\}$$

$$a, c \in \mathcal{R}^K \quad b, d \in \mathcal{R}$$



# Bound 1 - Piecewise Linear Bound (Hsiung et al)

$$\max_j \{a_j^\top \gamma + b_j\} \leq \log\left(\sum_{k=1}^{K} \exp(\gamma_k)\right) \leq \max_{j'} \{c_{j'}^\top \gamma + d_{j'}\}$$

$$a, c \in \mathcal{R}^K \quad b, d \in \mathcal{R}$$

Advantages

- The bound can be made arbitrarily accurate by increasing the number of pieces.

Disadvantages

- Max-function makes the objective non-differentiable.
- The number of variational parameters grows with the approximation level.
- Optimizing the variational parameter is hard.

# Bound 2 - Double Majorization (Bouchard 2007)

The LSE is bound by,

$$\log\left(\sum_{k=1}^{K}\exp(w_k^\top x_i)\right) \leq a_i + \sum_{k=1}^{K}\log(1+\exp(w_k^\top x_i - a_i)) \quad , \; a_i \in \mathcal{R}$$

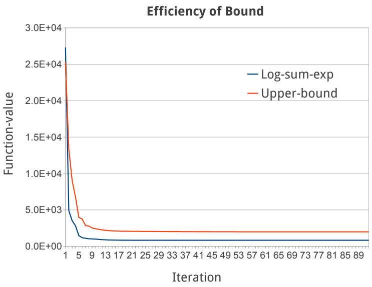
Advantages

- The bound is parallelizable.
- It is an upper bound.
- It is differentiable and **convex**.

# Bound 2 - Double Majorization (Bouchard 2007)

Disadvantage

- The bound is not tight enough.

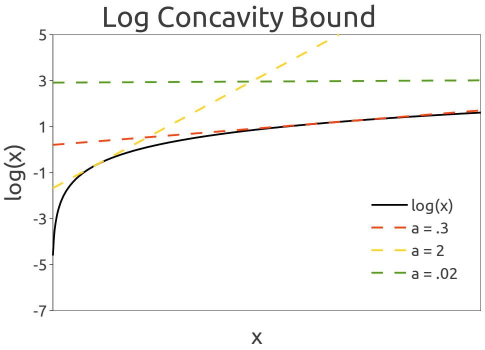



The gap between true objective and upper-bounded objective on the 20-newsgroup dataset.

# Bound 3 - Log Concavity

A relatively famous bound using the concavity of the log-function

$$\log(x) \leq ax - \log(a) - 1 \quad \forall\ x, a > 0$$



# Bound 3 - Log Concavity

Applying to the LSE function,

$$\log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right) \leq a_i \sum_{k=1}^{K} \exp(w_k^\top x_i) - \log(a_i) - 1$$

Advantages

- The bound is parallelizable.
- It is differentiable.
- Optimizing the variational parameter $a_i$ is easy.
- The upper bound is exact at $a_i = \frac{1}{\sum\limits_{k=1}^{K} \exp(w_k^\top x_i)}$.

Disadvantage

- The combined objective is non-convex.

# Reaching Optimality

**MLE estimation** $\min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right)$

**Log-concavity Bound** $\log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right) \leq a_i \sum_{k=1}^{K} \exp(w_k^\top x_i) - \log(a_i) - 1$

# Reaching Optimality

**MLE estimation** $\min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right)$

**Log-concavity Bound** $\log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right) \leq a_i \sum_{k=1}^{K} \exp(w_k^\top x_i) - \log(a_i) - 1$

Combined Objective

$$F(W, A) = \frac{\lambda}{2}\sum_{k=1}^{K}\|w_k\|^2 + \sum_{i=1}^{N}\left[-\sum_{k=1}^{K} y_{ik} w_k^\top x_i + a_i \sum_{k=1}^{K} \exp(w_k^\top x_i) - \log(a_i) - 1\right]$$

# Reaching Optimality

**MLE estimation** $\min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} w_k^\top x_i + \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right)$

**Log-concavity Bound** $\log\left(\sum_{k=1}^{K} \exp(w_k^\top x_i)\right) \leq a_i \sum_{k=1}^{K} \exp(w_k^\top x_i) - \log(a_i) - 1$

Combined Objective

$$F(W, A) = \frac{\lambda}{2}\sum_{k=1}^{K} \|w_k\|^2 + \sum_{i=1}^{N}\left[-\sum_{k=1}^{K} y_{ik} w_k^\top x_i + a_i \sum_{k=1}^{K} \exp(w_k^\top x_i) - \log(a_i) - 1\right]$$

Despite the non-convexity, we can show that

- The combined objective has a unique minima.
- This minimum coincides with the optimal MLE solution.

An iterative and **parallel** block coordinate descent algorithm to converge to the unique minimum.

**Algorithm 1** A parallel block coordinate descent

**Initialize** : $t \leftarrow 0, \mathbf{A}^0 \leftarrow \frac{1}{K}, \mathbf{W}^0 \leftarrow 0.$

**While :** Not converged

*In parallel* : $\mathbf{W}^{t+1} \leftarrow \arg\min_W F(W, \mathbf{A}^t)$

$\mathbf{A}^{t+1} \leftarrow \arg\min_A F(\mathbf{W}^{t+1}, A)$

$t \leftarrow t + 1$

# Experimental Comparison

## Datasets

| Dataset | # instances | #Leaf-labels | #Features | #Parameters | Parameter Size (approx) |
|---|---|---|---|---|---|
| **CLEF** | 10,000 | 63 | 80 | 5,040 | 40KB |
| **NEWS20** | 11,260 | 20 | 53,975 | 1,079,500 | 4MB |
| **LSHTC-small** | 4,463 | 1,139 | 51,033 | 227,760,279 | 911MB |
| **LSHTC-large** | 93,805 | 12,294 | 347,256 | 4,269,165,264 | 17GB |

## Optimization Methods

- Double Majorization Bound (DM)
- Log concavity Bound (LC)
- Limited Memory BFGS (LBFGS) - the most widely used method.
- Alternating Direction Method of Multipliers (ADMM)

# Time Complexity

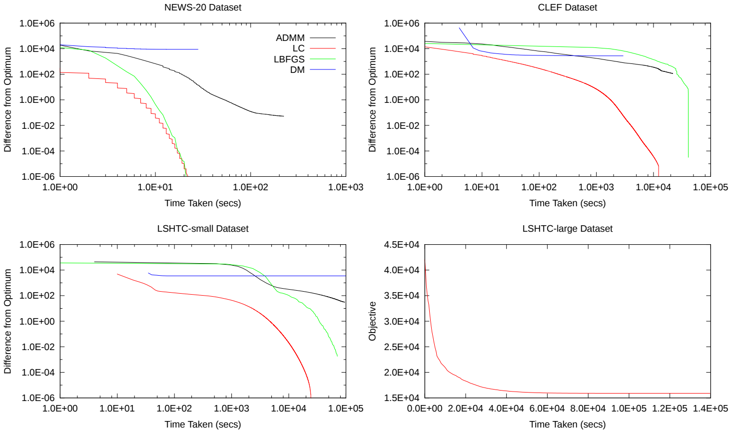
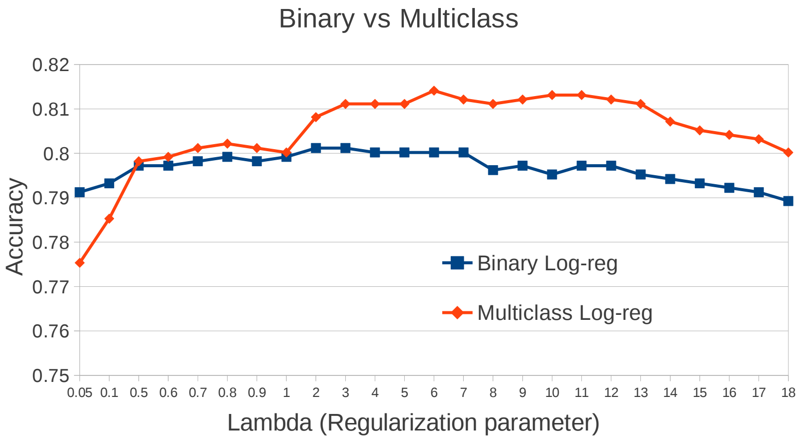



Figure : The difference from the true optimum vs time

# Conclusion

- Discussed multiple ways to perform distributed training of large-scale Logistic Models.
- The LC method seem to offer the best trade-off between accuracy and time.
- Several open questions,
    - Effect of the regularization parameter $\lambda$.
    - Effect of the correlation between the parameters.


Binary vs Multiclass
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
Accuracy
0.05 0.1 0.5 0.6 0.7 0.8 0.9 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Lambda (Regularization parameter)
Binary Log-reg
Multiclass Log-reg