Distributed Machine Learning
Algorithms and Open Source Implementations

Martin Jaggi
ETH Zurich
monthly

Zürich Machine Learning and Data Science

[ Link to Website ]
Machine Learning?
AI?

(Prediction)
Classification & Regression
Classification

Training data
Classification

Training data
Classification

Training data
Classification

Training data
Classification

Training data
The Training Algorithm

Training data
The Training Algorithm

Perceptron
(Rosenblatt 1957)

Support-Vektor-Maschine
(Cortes & Vapnik 1995)
The Training Algorithm

Perceptron (Rosenblatt 1957)

Support-Vektor-Maschine (Cortes & Vapnik 1995)
The Training Algorithm

Perceptron  
(Rosenblatt 1957)

Support-Vektor-Maschine  
(Cortes & Vapnik 1995)
The Training Algorithm

Perceptron
(Rosenblatt 1957)

Support-Vektor-Maschine
(Cortes & Vapnik 1995)
The Training Algorithm

Perceptron
(Rosenblatt 1957)

Support-Vektor-Maschine
(Cortes & Vapnik 1995)
The Training Algorithm

Perceptron
(Rosenblatt 1957)

Support-Vektor-Maschine
(Cortes & Vapnik 1995)

\[ \mathbf{w} \leftarrow \mathbf{w} + \lambda \cdot \mathbf{x} \]

(Stochastic Gradient Descent)
What has changed?

1950s: $10^3$ FLOPS

2010s: $10^{15}$ FLOPS
What has changed?

1950s: $10^3$ FLOPS

2010s: $10^{15}$ FLOPS

“the embryo of an electronic computer that ... will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.” 1958
What has changed?

1950s: $10^3$ FLOPS

2010s: $10^{15}$ FLOPS

"the embryo of an electronic computer that ... will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

1958
Machine Learning?

Some Applications w/ Big Data

Classification & Regression
Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars
Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars
Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars
Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

[how-old.net](http://how-old.net)
Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

how-old.net
Text Data

- Spam Detection
- User Content
- Medical Data

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Type</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative</td>
<td>neutral</td>
<td>But i wanna wear my Concords tomorrow though but i don't</td>
</tr>
<tr>
<td>positive</td>
<td>neutral</td>
<td>Gonna watch Grey's Anatomy all day today and tomorrow(:(</td>
</tr>
<tr>
<td>negative</td>
<td>neutral</td>
<td>@CoachVac heey do you know anything about UVA's fallll</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>@DustyEf when that sun is high in that Texas sky, I'll be buuuuu</td>
</tr>
<tr>
<td>neutral</td>
<td>positive</td>
<td>Up 20 points in my money league with Vernon Davis and L.</td>
</tr>
<tr>
<td>neutral</td>
<td>positive</td>
<td>DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actually</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
<td>The Rick Santorum signing that was scheduled for tomorrow</td>
</tr>
<tr>
<td>positive</td>
<td>neutral</td>
<td>@dreami9 lol yep looks like it! Was after El Clasico on Sundaa</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>Back in Stoke on Trent for the 2nd time today!</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>First Girls Varsity Basketball Game tomorrow at 6:00 pm Th</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>#UFC lightweights @Young___Assassin VS @jamievarner set</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>@OOOOO_WEEEE slide thru sometime this weekend ill have</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
<td>@DannyB618 Sure absolutely-- I meant out of the Bachman</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
<td>@RichardGordon48 re Levein discussion on Wed. Can't keep</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>Today In History November 02, 1958 Elvis gave a party at h</td>
</tr>
<tr>
<td>neutral</td>
<td>positive</td>
<td>Hustle cause you got to then kick back n party everyday like</td>
</tr>
<tr>
<td>positive</td>
<td>positive</td>
<td>I can't sleep. Way too exited about Vancouver tomorrow! I'm</td>
</tr>
</tbody>
</table>
Text Data

- Spam Detection
- User Content
- Medical Data
Medical: Genetic Data
Audio Data

- Hearing aids
- Voice Recognition
- Automatic Translation
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics
Internet Data

- Advertizing
- Recommender Systems
Internet Data

- Advertizing
- Recommender Systems
Internet Data

- Advertisizing
- Recommender Systems
Insurance & Finance

- Business-Analytics
- Targeted Advertizing
- Fraud
- Risk
- Customer Relations
- Marketing
Your turn!

Getting Started with Machine Learning

- scikit learn (python)

- real data applications: kaggle.com
Distributed Machine Learning

What if the data does not fit onto one computer anymore?
Distributed Machine Learning

What if the data does not fit onto one computer anymore?
Distributed Machine Learning

What if the data does not fit onto one computer anymore?

Does More Data Help?
\[ \Delta w^{(1)} := \gamma x_i \]

\[ \Delta w^{(5)} := \gamma x_i \]
Distributed Machine Learning

\[ \Delta w^{(1)} := \gamma x_i \]

\[ \Delta w^{(5)} := \gamma x_i \]

\[ w := w + \sum_k \Delta w^{(k)} \]
Distributed Machine Learning

\[ \Delta \mathbf{w}^{(1)} := \gamma \mathbf{x}_i \]

\[ \Delta \mathbf{w}^{(5)} := \gamma \mathbf{x}_i \]

\[ \mathbf{w} := \mathbf{w} + \sum_k \Delta \mathbf{w}^{(k)} \]
Problem 1

The Cost of Communication

- Reading $\nu$ from Memory (RAM)
  $100 \text{ ns}$

- Sending $\nu$ to another Machine
  $500'000 \text{ ns}$

- One Typical Map-Reduce Iteration (Hadoop)
  $10'000'000'000 \text{ ns}$
Problem 2

* Parallel Algorithms are Hard
* Single Machine Solvers are Fast

- no reusability of good single machine algorithms
Model Averaging Does Not Work
Model Averaging Does Not Work
Communication Efficient
Distributed Dual Coordinate Ascent

machine 1

machine 2

machine 3

machine 4

machine 5
Communication Efficient Distributed Dual Coordinate Ascent

\[ \Delta w^{(1)} \]

\[ \Delta w^{(5)} \]

\[ w := w + \sum_k \Delta w^{(k)} \]

CoCoA
Communication Efficient Distributed Dual Coordinate Ascent

$$\mathbf{w} := \mathbf{w} + \sum_k \Delta \mathbf{w}^{(k)}$$
Communication Efficient Distributed Dual Coordinate Ascent

CoCoA

\[ \Delta \mathbf{w}^{(1)} \]

\[ \Delta \mathbf{w}^{(5)} \]

\[ \mathbf{w} := \mathbf{w} + \sum_k \Delta \mathbf{w}^{(k)} \]
“Big Data Analytics” Applications

Classification
- Support Vector Machine (SVM)
- Logistic Regression
- Structured Prediction

Regression
- Ridge Regression
- Sparse Least Squares variants
- Lasso, Elastic-Net (Feature Selection, Compressed Sensing)

\[
\min_{w \in \mathbb{R}^d} \left[ \frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \ell_i(w^T x_i) \right]
\]
Experiments vs Spark MLlib

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training $n$</th>
<th>Features $d$</th>
<th>Sparsity</th>
<th>$\lambda$</th>
<th>Workers $K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cov</td>
<td>522,911</td>
<td>54</td>
<td>22.22%</td>
<td>1e-6</td>
<td>4</td>
</tr>
<tr>
<td>rcv1</td>
<td>677,399</td>
<td>47,236</td>
<td>0.16%</td>
<td>1e-6</td>
<td>8</td>
</tr>
<tr>
<td>imagenet</td>
<td>32,751</td>
<td>160,000</td>
<td>100%</td>
<td>1e-5</td>
<td>32</td>
</tr>
</tbody>
</table>

In Figure 3 we explore the effect of increasing $H$. As described above, increasing $H$ may cause convergence to become less of an impediment. In Figure 4, we attempt to scale the averaging step of each algorithm by using various $H$. While noting that it is a benefit to avoid having to tune this data-dependent parameter.

Our results indicate that COCOA (H=1e6) and mini-batch-CD (H=100) converge more quickly than all other algorithms, even when accounting for different batch sizes. For these datasets, they are 25x faster than the best competitor. This is a testament to the algorithm's ability to achieve a good balance between convergence speed and communication.

In comparing each algorithm and dataset, we analyze the progress in primal objective value as a function of both time (Figure 1) and communication (Figure 2). For all competing methods, we summarize the results.

To avoid communication while still making significant global progress, we use COCOA for the cov dataset on a single machine. Implementing mini-batch SDCA (denoted mini-batch-CD) as described in [3, 1], we see clearly that COCOA is able to converge to a more accurate solution in all cases:

- For the cov dataset, COCOA achieves a .001-accurate solution in less than 200 seconds, while all other methods require significantly more time.
- For the rcv1 dataset, COCOA converges much faster than the other methods, with an average speedup of 25x.
- For the imagenet dataset, COCOA achieves the best performance in terms of reduction in log primal suboptimality, reaching a .001-accurate solution in approximately 300 seconds.

These results demonstrate the effectiveness of COCOA in handling large-scale optimization problems efficiently, even in distributed settings.
Thanks

Email: m.jaggi@gmail.com

Web: da.inf.ethz.ch
    spinningbytes.com

all our Spark code is available on github

joint work with Virginia Smith, Martin Takáč, Chenxin Ma, Simone Forte, Tribhuvanesh Orekondy, Jonathan Terhorst, Sanjay Krishnan, Aurelien Lucchi, Peter Richtarik, Thomas Hofmann, Michael I. Jordan