MOCHA: Federated Multi-Task Learning

Virginia Smith
Stanford / CMU

Chao-Kai Chiang · USC
Maziar Sanjabi · USC
Ameet Talwalkar · CMU
MACHINE LEARNING WORKFLOW

- data & problem
- machine learning model
- optimization algorithm

\[
\min_w \sum_{i=1}^{n} \ell(w, x_i) + g(w)
\]
MACHINE LEARNING WORKFLOW

IN PRACTICE

- data & problem
- machine learning model
- systems setting
- optimization algorithm

\[
\min_w \sum_{i=1}^{n} \ell(w, x_i) + g(w)
\]
how can we perform fast distributed optimization?
BEYOND THE DATACENTER

- Massively Distributed
- Node Heterogeneity
- Unbalanced
- Non-IID
- Underlying Structure
BEYOND THE DATACENTER

- Massively Distributed
- Node Heterogeneity

**Systems Challenges**

- Unbalanced
- Non-IID
- Underlying Structure

**Statistical Challenges**
MACHINE LEARNING WORKFLOW

1. data & problem
2. machine learning model
3. systems setting
4. optimization algorithm

\[ \min_w \sum_{i=1}^{n} \ell(w, x_i) + g(w) \]
MACHINE LEARNING WORKFLOW IN PRACTICE

- data & problem
- systems setting
- machine learning model
- optimization algorithm

\[
\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}, x_i) + g(\mathbf{w})
\]
OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

- Massively Distributed
- Node Heterogeneity

**Statistical Challenges**

**Systems Challenges**
A GLOBAL APPROACH
A LOCAL APPROACH
OUR APPROACH: PERSONALIZED MODELS
OUR APPROACH: PERSONALIZED MODELS
MULTI-TASK LEARNING

\[
\min_{W, \Omega} \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell_t(w_t, x_t^i) + R(W, \Omega)
\]

All tasks related

Outlier tasks

Clusters / groups

Asymmetric relationships

[ZCY, SDM 2012]
FEDERATED DATASETS

Human Activity

Google Glass

Land Mine

Vehicle Sensor
<table>
<thead>
<tr>
<th>Category</th>
<th>Global</th>
<th>Local</th>
<th>MTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Activity</td>
<td>2.23 (0.30)</td>
<td>1.34 (0.21)</td>
<td>0.46 (0.11)</td>
</tr>
<tr>
<td>Google Glass</td>
<td>5.34 (0.26)</td>
<td>4.92 (0.26)</td>
<td>2.02 (0.15)</td>
</tr>
<tr>
<td>Land Mine</td>
<td>27.72 (1.08)</td>
<td>23.43 (0.77)</td>
<td>20.09 (1.04)</td>
</tr>
<tr>
<td>Vehicle Sensor</td>
<td>13.4 (0.26)</td>
<td>7.81 (0.13)</td>
<td>6.59 (0.21)</td>
</tr>
</tbody>
</table>
OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

Systems Challenges
OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

Systems Challenges
GOAL: FEDERATED OPTIMIZATION FOR MULTI-TASK LEARNING

\[ \min_{W, \Omega} \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell_t(w_t^T x_t^i) + R(W, \Omega) \]

- Solve for \( W, \Omega \) in an alternating fashion
  - \( \Omega \) can be updated centrally
  - \( W \) needs to be solved in federated setting

Challenges:

- Communication is expensive
- Statistical & systems heterogeneity
  - Stragglers
  - Fault tolerance
GOAL: FEDERATED OPTIMIZATION FOR MULTI-TASK LEARNING

Idea:
Modify a *communication-efficient* method for the *data center* setting to handle:

- ✔ Multi-task learning
- ✔ Stragglers
- ✔ Fault tolerance
COCOA: COMMUNICATION-EFFICIENT DISTRIBUTED OPTIMIZATION

key idea: control communication

mini-batch methods

one-shot communication
COCOA: PRIMAL-DUAL FRAMEWORK

PRIMAL \geq \text{DUAL}

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(w^T x_i) + \lambda g(w) \quad \max_{\alpha \in \mathbb{R}^n} -\frac{1}{n} \sum_{i=1}^{n} \ell^*(-\alpha_i) - \lambda g^*(X, \alpha)
\]
COCOA: PRIMAL–DUAL FRAMEWORK

challenge #1:
extend to MTL setup
Main assumption:

Each subproblem is solved to accuracy $\Theta$

$$\Theta \in [0, 1) \approx$$

amount of local computation vs. communication

exactly solve

inexactly solve
COCOA: COMMUNICATION PARAMETER

Main assumption:

Challenge #2: make communication more flexible

- exactly solve
- inexactlly solve
MOCHA: COMMUNICATION-EFFICIENT FEDERATED OPTIMIZATION

\[
\min_{\mathbf{W}, \Omega} \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell_t(\mathbf{w}_t^T \mathbf{x}_t^i) + \mathcal{R}(\mathbf{W}, \Omega)
\]

- Solve for \( \mathbf{W}, \Omega \) in an alternating fashion
- Modify CoCoA to solve \( \mathbf{W} \) in federated setting

\[
\min_{\mathbf{\alpha}} \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell^*(-\alpha_t^i) + \mathcal{R}^*(X\mathbf{\alpha})
\]

\[
\min_{\Delta \mathbf{\alpha}_t} \sum_{i=1}^{n_t} \ell^*(-\alpha_t^i - \Delta \alpha_t^i) + \langle \mathbf{w}_t(\mathbf{\alpha}), \mathbf{X}_t \Delta \mathbf{\alpha}_t \rangle + \frac{\sigma'}{2} \| \mathbf{X}_t \Delta \mathbf{\alpha}_t \|_{M_t}^2
\]
MOCHA: PER-DEVICE, PER-ITERATION APPROXIMATIONS

New assumption: each subproblem is solved to accuracy $\theta^h_t \in [0, 1]$

Stragglers (Statistical heterogeneity)
- Difficulty of solving subproblem
- Size of local dataset

Stragglers (Systems heterogeneity)
- Hardware (CPU, memory)
- Network connection (3G, LTE, …)
- Power (battery level)

Fault tolerance
- Devices going offline
New assumption: each subproblem is solved to accuracy $\theta^h_t$

and assume: $\mathbb{P}[\theta^h_t := 1] < 1$

**Theorem 1.** Let $\ell_t$ be $L$-Lipschitz, then

$$T \geq \frac{1}{(1 - \Theta)} \left( \frac{8L^2n^2}{\epsilon} + \tilde{c} \right)$$

$1/\epsilon$ rate

**Theorem 2.** Let $\ell_t$ be $(1/\mu)$-smooth, then

$$T \geq \frac{1}{(1 - \Theta)} \frac{\mu + n}{\mu} \log \frac{n}{\epsilon}$$

linear rate
Algorithm 1 MOCHA: Federated Multi-Task Learning Framework

1: **Input:** Data $X_t$ stored on $t = 1, \ldots, m$ devices
2: Initialize $\alpha^{(0)} := 0$, $v^{(0)} := 0$
3: for iterations $i = 0, 1, \ldots$ do
4:   for iterations $h = 0, 1, \cdots, H_i$ do
5:     for devices $t \in \{1, 2, \ldots, m\}$ in parallel do
6:       call local solver, returning $\theta^h_t$-approximate solution $\Delta \alpha_t$
7:       update local variables $\alpha_t \leftarrow \alpha_t + \Delta \alpha_t$
8:     reduce: $v \leftarrow v + \sum_t X_t \Delta \alpha_t$
9:     Update $\Omega$ centrally using $w(v) := \nabla R^*(v)$
10: Compute $w(v) := \nabla R^*(v)$
11: return: $W := [w_1, \ldots, w_m]$
MOCHA IS ROBUST TO STATISTICAL HETEROGENEITY

MOCHA & COCOA PERFORM PARTICULARLY WELL IN HIGH-COMMUNICATION SETTINGS
MOCHA SIGNIFICANTLY OUTPERFORMS ALL COMPETITORS

[BY 2 ORDERS OF MAGNITUDE]
### FAULT TOLERANCE

**W-Step**

![Graph showing W-Step results]

**Full Method**

![Graph showing Full Method results]

**MOCHA IS ROBUST TO DROPPED NODES**
OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

- Massively Distributed
- Node Heterogeneity

Statistical Challenges

Systems Challenges