On the Convergence of Local SGD on Identical and Heterogeneous Data

Ahmed Khaled

FLOW: Federated Learning One World Seminar
May 13th, 2020
This Talk is Based on

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik
**Tighter Theory for Local SGD on Identical and Heterogeneous Data**
To appear in Artificial Intelligence and Statistics (AISTATS) 2020

Earlier Workshop Papers

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik
**First Analysis of Local GD on Heterogeneous Data**
NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik
**Better Communication Complexity for Local SGD**
NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality
Collaborators

Peter Richtárik
Professor of Computer Science
KAUST

Konstantin Mishchenko
CS PhD candidate
KAUST
Plan

• **Introduction (20 mins)**
  - Problem Definition
  - Mini-batch SGD vs Local SGD
  - Goals and Contributions

• **Theory (20 mins)**
  - Heterogeneous Data
  - Identical Data
Introduction
Federated Learning

• A distributed machine learning setting where data is distributed over many clients with potentially unreliable connections.

• Many applications: mobile text prediction, medical research, and many more!

• Federated Learning poses highly interdisciplinary problems: optimization, privacy, security, information theory, statistics, and many other fields intersect.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, Dave Bacon

Federated Learning: Strategies for Improving Communication Efficiency

NIPS Workshop on Private Multi-Party Machine Learning, 2016
Problem Definition

\[
\min_{x \in \mathbb{R}^d} \left\{ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} [f(x; \xi)] \right\}
\]

- Smooth and \(\mu\)-convex (for \(\mu \geq 0\))
- Model dimension
- We can query stochastic gradients

Example: Finite-sum minimization

\[
f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]

Ubiquitous in machine learning
We desire a scalable, parallel optimization method.

In typical parameter server applications, Mini-batch Stochastic Gradient Descent (Mini-batch SGD) is the popular baseline algorithm.

Communication is the bottleneck.

There are two regimes...
Data Regime 1: Heterogenous Data

- Each client has access to
  - its own optimization objective
  - its own dataset
- Each local objective is also written as a stochastic expectation.
- Arises in Federated Learning applications because the data is inherently distributed, cannot be centralized due to privacy protection.

\[
f_1(x) = \mathbb{E}_{\xi \sim D_1} \{ f_1(x; \xi) \}
\]

\[
f_2(x) = \mathbb{E}_{\xi \sim D_2} \{ f_2(x; \xi) \}
\]
Data Regime 2: Identical Data

- Each client has access to **the same dataset**.
- The clients may draw different samples from the dataset, or **have different sampling distributions**.
- Arises in the parameter server framework.
- Can be **insightful into the usefulness of local steps**.

\[ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} \{ f(x; \xi) \} \]
Mini-batch SGD

If the data is identical:

$$\mathbb{E} \left[ g^m(x_t; \xi^m) \right] = \nabla f(x_t)$$

If the data is heterogeneous:

$$\mathbb{E} \left[ g^m(x_t; \xi^m) \right] = \nabla f_m(x_t)$$

Sample a local stochastic gradient

$g^1(x_t; \xi^1)$

Sample a local stochastic gradient

$g^2(x_t; \xi^2)$

Sample a local stochastic gradient

$g^3(x_t; \xi^3)$

Client

Server

$\uparrow x_t$

$\uparrow x_t$

$\uparrow x_t$
Mini-batch SGD

Sample a local stochastic gradient $g^1(x_t; \xi^1)$

Sample a local stochastic gradient $g^2(x_t; \xi^2)$

Sample a local stochastic gradient $g^3(x_t; \xi^3)$

The server then performs aggregation and averaging:

$$x_{t+1} = x_t - \frac{\gamma}{M} \sum_{m=1}^{M} g^m(x_t; \xi^m)$$

where $\gamma > 0$ is a stepsize
• Equivalently, we can write the parallel mini-batch SGD update as follows:

\[ x_{t+1} = \frac{1}{M} \sum_{m=1}^{M} (x_t - \gamma g^m(x_t; \xi^m)) \]

• Observation: we take one "local step" and follow it by averaging.

• What about multiple local steps?
Local SGD

And then it repeats...

One SGD step

At most $H$ local steps: $t \leq H$
Our Contributions
Goal

Can we achieve the same training error as Mini-batch SGD but with less communication?
Our Contributions

• **Heterogeneous data regime:**
  
  • We critically examine data similarity assumptions and show they do not hold for even very simple functions.
  
  • We obtain the first convergence guarantee for Local SGD on arbitrarily heterogeneous local losses. The guarantee shows Local SGD is communication-efficient at least in some settings.

• **Identical data regime:**
  
  • We show that even more dramatic communications savings are possible for convex and strongly-convex objectives.
  
  • In particular, we show that for strongly convex objectives the number of communications can be a constant independent of the total number of iterations!
<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of</td>
<td>Number of Communication Steps</td>
</tr>
<tr>
<td>iterations</td>
<td>$T$</td>
</tr>
<tr>
<td>Synchronization</td>
<td>Interval</td>
</tr>
<tr>
<td>Interval</td>
<td>$H$</td>
</tr>
<tr>
<td>Number of</td>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Nodes</td>
<td>$M$</td>
</tr>
</tbody>
</table>

$$C \geq \frac{T}{H}$$
Theory for Heterogeneous Data
Setting and assumptions

- We assume the existence of at least one minimizer.

\[ x^* \in \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) \right\} \]

- Each function is convex.

- The results can be extended to the strongly convex case.
Related work


However, the last two use the "bounded gradients" assumption...
Related work


Show that communication savings are possible, however they use a *bounded dissimilarity* assumption...
More related work

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang
On the Convergence of FedAvg on Non-IID Data.

Farzin Haddadpour and Mehrdad Mahdavi
On the Convergence of Local Descent Methods in Federated Learning

Consider FedAvg (with sampling)

Obtain results for non-convex objectives under a bounded diversity assumption

More later (and in the paper)...

23
Assumptions on similarity: bounded dissimilarity

- A common assumption to obtain convergence rates is **bounded dissimilarity**:

\[
\frac{1}{M} \sum_{m=1}^{M} \| \nabla f_m(x) - \nabla f(x) \|^2 \leq \sigma^2
\]

for all \( x \in \mathbb{R}^d \)
Assumptions on similarity: bounded dissimilarity

- The bounded dissimilarity condition may not be satisfied for 1-dimensional quadratics:

\[ f_m(x) \overset{\text{def}}{=} \frac{a_m}{2} x^2 \]

\[
\frac{1}{M} \sum_{m=1}^{M} \left\| \nabla f_m(x) - \nabla f(x) \right\|^2
= \left( \frac{1}{M} \sum_{m=1}^{M} \left( a_m - \frac{1}{M} \sum_{j=1}^{M} a_j \right)^2 \right) \cdot x^2
\]

Can be arbitrarily large
Assumptions on similarity: bounded gradients

• The bounded gradients assumption is also in common usage:

\[
\frac{1}{M} \sum_{m=1}^{M} \| \nabla f_m(x) \|^2 \leq G^2 \quad \text{for all } x \in \mathbb{R}^d
\]

• Problem 1: special case of bounded dissimilarity without the benefit of characterizing similarity..

\[
\frac{1}{M} \sum_{m=1}^{M} \| \nabla f_m(x) - \nabla f(x) \|^2 = \frac{1}{M} \sum_{m=1}^{M} \| \nabla f_m(x) \|^2 - \| \nabla f(x) \|^2 \leq G^2
\]
Assumptions on similarity: bounded gradients

- The bounded gradients assumption is also in common usage:

\[
\frac{1}{M} \sum_{m=1}^{M} \| \nabla f_m(x) \|^2 \leq G^2
\]

- Problem 2: **contradicts** global strong convexity.

Assumptions on similarity: bounded gradients

• The bounded gradients assumption is also in common usage:

\[ \frac{1}{M} \sum_{m=1}^{M} \| \nabla f_m(x) \|^2 \leq G^2 \]

• Problem 3: questionable applicability to practice.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. 
Reducing Noise in GAN Training with Variance Reduced Extragradient. 

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. 
Revisiting Stochastic Extragradient. 
To appear in the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.
There are no results that apply to arbitrarily heterogeneous data
The alternative

• Our theory is built upon the variance at the optimum

\[ \sigma_{\text{dif}}^2 \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\xi \sim \mathcal{D}_m} \left[ \| \nabla f_m (x_*; \xi) \|^2 \right] \]

• Naturally relates the difference between the functions at a single point.

• Zhang and Li show that when this quantity is zero, we get linear convergence for strongly convex objectives with any \( H \).

Chi Zhang and Qianxiao Li.
Distributed Optimization for Over-Parameterized Learning.
Main Theorem (Heterogeneous Data)

For any sufficiently small step size

\[
E[f(\bar{x}_T) - f(x_*)] \leq \frac{4\|r_0\|^2}{\gamma T} + \frac{20\gamma \sigma^2_{\text{dif}}}{M} + 16\gamma^2 L(H - 1)^2 \sigma^2_{\text{dif}}.
\]

\[
\gamma \leq \min \left\{ \frac{1}{4L'}, \frac{1}{8L(H - 1)} \right\}
\]
Main Theorem (Heterogeneous Data)

For any sufficiently small step size

\[
E \left[ f(\bar{x}_T) - f(x_*) \right] \leq \frac{4\|r_0\|^2}{\gamma T} + \frac{20\gamma\sigma^2_{\text{dif}}}{M} + 16\gamma^2 L(H - 1)^2 \sigma^2_{\text{dif}}.
\]

\[
\gamma \leq \min \left\{ \frac{1}{4L'}, \frac{1}{8L(H - 1)} \right\}
\]

Same as Mini-batch SGD (up to constants)

An error term controlled by the synchronization interval $H$
Communication Complexity

• If we properly chose the stepsize...

• **Communication Complexity**: iterations to guarantee

\[
\mathbb{E} [f(\bar{x}_T) - f(x_*)] \leq \varepsilon
\]

\[
C = \Omega \left( \frac{\|r_0\|^2}{\varepsilon} \max \left\{ L, \frac{\sigma^2_{\text{dif}}}{HM \varepsilon}, \frac{\sqrt{L}\sigma_{\text{dif}}}{\sqrt{\varepsilon}} \right\} \right)
\]

- Initial distance to the optimum
- **Smoothness constant**
- **Synchronization Interval**
- **# clients**

Desired accuracy
Communication Complexity

- For a small enough desired accuracy:

\[
C = \Omega \left( \frac{\| r_0 \|^2 \sqrt{L\sigma_{\text{dif}}}}{\varepsilon^{3/2}} \right) \quad C = \Omega \left( \frac{\| r_0 \|^2 \sigma_{\text{dif}}^2}{\varepsilon^2 M} \right)
\]

We get a reduction in the number of communications as a function of the accuracy even for arbitrarily heterogeneous data!
Optimal Synchronization Interval

• We show that the optimal $H$ for attaining the same rate as Minibatch SGD is

$$H = 1 + \left\lfloor T^{1/4} M^{-3/4} \right\rfloor$$

• And the corresponding communication complexity then is

$$C = \Omega \left( \min\{T, T^{3/4} M^{-3/4}\} \right)$$
Experimental Results

![Graph showing experimental results with different local steps](image)
Theory for Identical Data
Setting

- We assume that each $f(x; \xi)$ is almost surely convex and smooth. All clients share the same objective. Will present results for $\mu$-strongly convex as well.

- The measure of variance is also the variance at the optimum:

$$\sigma^2_{\text{opt}} \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\xi \sim D_m} \left[ \| \nabla f(x_\star; \xi) \|^2 \right]$$
Background 1

• Stich (2019) analyzes Local SGD with identical data.

• For strongly convex objectives, the communication complexity to reach the same error as Minibatch SGD is:

$$C = \Omega \left( \sqrt{\kappa MT} \right)$$

The condition number

$$\kappa \overset{\text{def}}{=} \frac{L}{\mu}$$

# clients
Background 2

- One-shot averaging is running SGD on each node and communicating only once at the end.

- This communication complexity tells us that one-shot averaging is not convergent. **But it should be. Why?**

\[ C = \Omega \left( \sqrt{\kappa MT} \right) \]

- There are no results for minimizing convex (but not strongly convex) objectives.

Grows with the total number of iterations
Theorem (Identical Data, Strong Convexity)

- With an appropriately chosen constant stepsize:

\[ \mathbb{E} \left[ \left\| x_T - x_\star \right\|^2 \right] = \tilde{O} \left( \frac{\|r_0\|^2}{T^2} + \frac{\sigma_{opt}^2}{\mu^2 MT} + \frac{\sigma_{opt}^2 \kappa (H - 1)}{\mu^2 T^2} \right) \]

- The condition number
- Strong convexity constant
- Error Term
- Same as Minibatch SGD
Interpreting the Result

\[ \mathbb{E} \left[ \| x_T - x_\ast \|^2 \right] = \tilde{O} \left( \frac{\| r_0 \|^2}{T^2} + \frac{\sigma^2_{\text{opt}}}{\mu^2 M T} + \frac{\sigma^2_{\text{opt}} \kappa (H - 1)}{\mu^2 T^2} \right) \]

- **Optimal synchronization interval**
  \[ H = 1 + \left\lceil \frac{T}{(\kappa M)} \right\rceil \]

- Reaches the same convergence error as Mini-batch SGD (up to absolute constants) but with a communication complexity of \( \tilde{\Omega}(\min(T, \kappa M)) \)

Number of communications can be constant!
Interpreting the Result

\[ E \left[ \| x_T - x_* \| ^2 \right] = \tilde{O} \left( \frac{\| x_0 - x_* \| ^2}{T^2} + \frac{\sigma^2}{\mu^2 MT} + \frac{\kappa \sigma^2 (H - 1)}{\mu^2 T^2} \right) \]

- **One-shot averaging**
  - Put \( H = T + 1 \), then we obtain a convergence rate of

\[ \tilde{O} \left( \frac{\sigma^2 \kappa}{\mu^2 T} \right) \]

- An improvement, but applying Jensen's inequality yields

\[ \tilde{O} \left( \frac{\sigma^2}{\mu^2 T} \right) \]

There is room for improvement!
Theorem (Identical Data, Convexity)

• For the (non-strongly) convex case, we get a similar result

\[ \mathbb{E} [f(\bar{x}_T) - f(x_*)] \leq \frac{2}{\gamma T} \| x_0 - x_* \|^2 + \frac{2\gamma \sigma^2}{M} + 4\gamma^2 L\sigma^2 (H - 1) \]

• Same guarantee as the heterogenous case, but with a linear instead of quadratic dependence on the synchronization interval.

• Translates to more communications savings
Concurrent Work

• Similar results for identical data were obtained in concurrent work of Stich and Karimireddy who use a different proof technique.

Sebastian U. Stich and Sai Praneeth Karimireddy

• More discussion is given in the paper.
Experimental Results
Open Questions 1

• Can we get **better convergence results** for Local SGD or Federated Averaging compared to Minibatch SGD?

• For general convex objectives and identical data:
  
  Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro.  
  **Is Local SGD Better than Minibatch SGD?**  

• For heterogeneous data, client sampling and using two stepsizes:
  
  **SCAFFOLD: Stochastic Controlled Averaging for Federated Learning.**  
Open Questions 2

• Do local methods give benefits other than optimization?

• Meta Learning point of view:

Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan
*Improving Federated Learning Personalization via Model Agnostic Meta Learning*

• Another perspective on personalization:

Filip Hanzely and Peter Richtárik
*Federated Learning of a Mixture of Global and Local Models*
Questions?

Thank you!
On Non-convex Objectives

- Even for the single-machine finite-sum optimization problem, convergence bounds in the non-convex setting often rely on restrictive assumptions.

- Often results rely on bounded variance or gradient dissimilarity assumptions.

- The relation of these assumptions to each other is not clear.

- We consider this and obtain a more general result in our new paper:

Ahmed Khaled and Peter Richtárik.
Better Theory for SGD in the Nonconvex World
References


References


References


References


 Acknowledgments

• PDF icon made by Pixel Buddha on www.flaticon.com.