# Convergence Analysis of Local Update Methods

**Don Kurian Dennis**
Carnegie Mellon University
dondennis@cmu.edu

**Shuhua Yu**
Carnegie Mellon University
shuhuay@andrew.cmu.edu

## 1 Introduction

One of the go to optimization procedure used in many large scale optimization tasks is mini-batch stochastic gradient descent (SGD). Conceptually speaking, in minibatch-SGD when optimizing over $\theta \in \Theta$, at $t = 0$, we start of at some $\theta^{(0)}$ and apply the following update rule at each step $t$,

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_t g_t.$$

Here $g_t$ is a stochastic gradient, with the property that $\mathbb{E}g_t = \nabla f(x_t)$. Compared to traditional gradient descent, replacing the full gradient computation in with the stochastic gradient decreases the memory and compute requirement for each step. Moreover, the stochastic gradient computation step can be easily parallelized and enjoys a linear speedup with the number of workers. Such parallelization is beneficial in scenarios where the computation is the bottleneck and communication is cheap — for instance in data local clusters. However, as more and more machine learning moves towards edge computing nodes (ex, mobile phones and fitness trackers), new settings arise where communication cost is a significant bottle neck. A prime example being Federated Learning (FL).

For such settings, a very natural change one can introduce to SGD to reduce communication costs is performing *local updates*. That is, each worker $k$ starts of with the same parameters $w_k^{(0)} = w_0$, and perform say $M$ local updates,

$$w_k^{(t+1)} = w_k^{(t)} - \eta_t g_t,$$

before the parameters on each of the workers is finally combined and re-synchronized.

Many methods loosely based on local updates have been proposed and have seen success in practice (ex. FedAvg), but the convergence analysis for these methods even for convex case is only being understood now. We wish to go deeper into the relevant literature to get a better picture of questions in this space.

## 2 Reading List

We plan to go over two main results.

1. Local SGD Converges Fast and Communicates Little, *Sebastian U. Stich*.
   This paper shows precise convergence analysis for local SGC on convex problem. This paper though, does not show that there is an advantage in termsof number of evaluated gradients when compared to mini-batch SGD. One of our goals would be to get some intuition regarding the difficulty here.
2. Sub-Sampled Newton Methods: Globally Convergent Algorithms, *Farbod Roosta-Khorasani, Michael W. Mahoney*
   This paper analyse second-order sub-sampled methods. They show global convergent rates of certain sub-sampled second order methods on convex problems. Or main goal for this paper would be a summary of the main arguments of the proofs used. As an additional goal, we wish to use the second-order sampling method used herein to see if it can be used to get an efficient yet faster algorithm for communication bottlenecked distributed optimization settings.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.