Local SGD: Unified Theory and New Efficient Methods

Eduard Gorbunov1,2,3, Filip Hanzely4, Peter Richtárik3
1MIPT (Russia), 2Yandex (Russia), 3KAUST (Saudi Arabia), 4TTIC (United States)

1. The Problem

\[
\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]

Options for \( f_i(x) \):

- \( f_i(x) = \mathbb{E}_{\xi_i} [f_i(x)] \)
- \( f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x) \)

Problem: Distributed optimization / training, where \( n \) workers (devices/clients) jointly solve a problem by communicating with a central server.

Assumptions: Smoothness and quasi-strong convexity of local loss functions:

- \( \|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\| \)
- \( f_i(x) \geq f_i(x^*) + \langle \nabla f_i(x^*), x^* - x \rangle + \frac{\mu}{2} \|x - x^*\|^2 \)

2. Local-SGD

- General theoretical framework for local first-order methods covering
- local shifts
- importance sampling
- variance reduction
- variable number of local steps

New efficient methods

- S-Local-SVRG – the first linearly converging stochastic method with local updates in heterogeneous case

3. Our Contributions

- Our framework recovers tight rates for known optimizers
- fills missing gaps for known methods
- extends the established optimizers

4. The Framework

\[
x_{k+1} = \begin{cases} 
  x_k - \gamma_k g_k, & \text{if } c_{k+1} = 0 \\
  x_k - \gamma_k \bar{g}_k, & \text{if } c_{k+1} = 1 
\end{cases}
\]

\( c_k = \begin{cases} 
  0, & \text{if } k \mod \tau \neq 0 \\
  1, & \text{if } k \mod \tau = 0 
\end{cases} \quad \text{encodes the time of communication}

Examples of \( \{c_k\}_{k \geq 0} \):

- Constant local loop
- Random local loop

The assumption below covers a very broad class of methods.

- \( \frac{1}{n} \sum f_i(x_1) + \cdots + f_i(x_n) - \frac{1}{n} \sum f_i(x^*) \leq \frac{1}{2} \sum \|x_i - x_i^*\|^2 \)
- \( E \left( \sum_{i=1}^{n} \sigma_i^2 \right) \leq 2 \left( 2E \|f(x) - f(x^*)\| + E \sigma_i^2 + E W_i \right) 
  + \left( E \left( \sum_{i=1}^{n} g_i^2 \right) \right) \leq 2L E \|x - x^*\| + \frac{\mu}{2} \|x - x^*\|^2 
  + E \left( \sum_{i=1}^{n} E \|W_i\| \right) 
  + \left( E \left( \sum_{i=1}^{n} g_i^2 \right) \right) \leq 2L E \|x - x^*\| + \frac{\mu}{2} \|x - x^*\|^2 
  + E \left( \sum_{i=1}^{n} E \|W_i\| \right) 
  + \left( E \left( \sum_{i=1}^{n} g_i^2 \right) \right) \leq 2L E \|x - x^*\| + \frac{\mu}{2} \|x - x^*\|^2 
  + E \left( \sum_{i=1}^{n} E \|W_i\| \right) 
  + \left( E \left( \sum_{i=1}^{n} g_i^2 \right) \right) \leq 2L E \|x - x^*\| + \frac{\mu}{2} \|x - x^*\|^2 
  + E \left( \sum_{i=1}^{n} E \|W_i\| \right) 
\]

5. Main Theorem: Simplified Version

- \( x^K = \frac{1}{n} \sum_{k=0}^{K-1} y_k \)
- \( E \left[ f(x^K) - f(x^*) \right] \leq \left( 1 - \min \{ \gamma, \frac{1}{L} \} \right)^{K} E \left[ f(x^{(0)}) - f(x^*) \right] + \frac{\mu}{2} \|x^{(0)} - x^*\|^2 \)

6. New Method: Shifted Local-SVRG

- Finite-sum case: \( f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x) \)
- \( x_{k+1} = x_k - \gamma_k g_k; \quad g_k = \nabla f_k(x_k) - \nabla f_k(x^{\bar{k}}) + \nabla f_k(x^{\bar{k}}) \)
- \( y_{k+1} = x_{k+1} \quad \text{with prob. } \frac{q}{m} \quad y_{k+1} = y_k \quad \text{with prob. } \frac{1-q}{m} \quad q = \frac{1}{m} \)

- Iteration complexity:

- The first linearly converging local method for heterogeneous data!

It is just an example. In fact, our approach covers a lot of different setups, methods and even the algorithms without local updates.

7. Numerical Experiments

We conducted several numerical experiments on logistic regression problem with \( L \)-regularization.

\[
f_i(x) = \frac{1}{m} \sum_{j=1}^{m} \log \left( 1 + \exp \left( \langle w_{0,i} + x_i, x \rangle \cdot b_{0,i} \right) \right) + \frac{\beta}{2} \|x\|^2
\]

- Weights: \( w_k = \frac{1}{1 + \exp \left( \langle w_{0,i}, x \rangle \right)} \)
- Weights: \( W_k = \frac{1}{1 + \exp \left( \langle w_{0,i}, x \rangle \right)} \)

- Reflects smoothness properties of the problem and noises introduced by stochastic gradients and functions dissimilarity
- Describes the process of local shifts’ learning and variance reduction
- Bounds the workers iterates’ discrepancy