A Distributed Optimization Algorithm over Time-Varying Graphs with Efficient Gradient Evaluations

BRYAN VAN SCOY AND LAURENT LESSARD UNIVERSITY OF WISCONSIN–MADISON

Introduction

Distributed optimization problem:

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

- \( f_i : \mathbb{R}^d \rightarrow \mathbb{R} \) is the local objective function on agent \( i \)
- \( n \) is the number of agents
- \( d \) is the dimension of the objective function

Goal: agents compute the global optimizer by communicating with local neighbors and performing local computations.

Communication Network

We model the communication network using a gossip matrix.

- A matrix \( W = [w_{ij}] \in \mathbb{R}^{n \times n} \) is a gossip matrix if \( w_{ij} = 0 \) whenever agent \( i \) does not receive information from agent \( j \).
- The spectral gap is \( \sigma = \|W - \frac{1}{n} \mathbf{1} \mathbf{1}^T\| \).
- \( W \) is stochastic if \( W \mathbf{1} = \mathbf{1} \) and \( \mathbf{1}^T W = \mathbf{1}^T \).

For example, a gossip matrix for the above network is

\[
W = \begin{bmatrix}
0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\
0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 \\
0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 \\
\frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{2} \\
\frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 \\
\end{bmatrix}
\]

with \( \sigma \approx 0.7853 \).

Algorithm

Notation: Bold indicates concatenated vectors, for example,

\[
x^k := \begin{bmatrix} x_{k1}^T \\ \vdots \\ x_{kn}^T \end{bmatrix}
\]

\[
\nabla f(v^k) := \begin{bmatrix} \nabla f_1(v_{1k}) \\ \vdots \\ \nabla f_n(v_{nk}) \end{bmatrix}
\]

Initialization:
- Set \( y^0 = 0 \) and \( x^0 \) arbitrary.
- Define the contraction factor \( \rho = \frac{L - \mu}{L + \mu} \) and
- the number of communications per gradient evaluation \( m \) as a function of the spectral gap \( \sigma \).

for iteration \( k = 0, 1, 2, \ldots \) do

\[
v^k = (W_{k,1} \cdots W_{k,m} \otimes I_d) x^k
\]

\[
u^k = \nabla f(v^k)
\]

\[
y^{k+1} = y^k + x^k - v^k
\]

\[
x^{k+1} = x^k - \frac{2}{L + \mu} u^k - \sqrt{1 - \rho^2} y^{k+1}
\]

end for

Theoretical Results

Theorem (Linear convergence). The iterate sequence \( \{x^k\}_{k \geq 0} \) of each agent \( i \in \{1, \ldots, n\} \) converges to the global optimizer \( x^* \) linearly with rate \( \rho \). In other words,

\[
\|x^k - x^*\| = \mathcal{O}(\rho^k)
\]

for all \( i \in \{1, \ldots, n\} \).

Our decentralized algorithm has the same worst-case convergence rate as centralized gradient descent in terms of the number of gradient evaluations.

Assumptions

1. Each local function \( f_i \) is \( \mu \)-smooth and \( L \)-strongly convex.
2. At each round of communication, each agent \( i \) has access to the \( i \)th row of a stochastic gossip matrix with spectral gap \( \sigma \in (0, 1) \).
3. Each agent knows the global parameters \( \mu, L, \) and \( \sigma \).

Corollary (Time complexity). Suppose it takes

- \( \tau \) time per communication round and
- unit time for evaluating local gradients.

Then the time to obtain a solution with precision \( \epsilon > 0 \) is

\[
\mathcal{O}\left( \kappa \left( 1 + \frac{1}{\sqrt{\sigma}} \right) \ln \left( \frac{1}{\epsilon} \right) \right)
\]

as \( \kappa \to \infty \) and \( \sigma \to 1 \), where \( \kappa = L/\mu \).

Convergence Rate

Our algorithm uses an optimized ratio between the number of communication rounds and gradient evaluations.

To account for the extra communication, define a step as one round of communication and at most one gradient evaluation.

The convergence rate of our algorithm per step is \( \rho^{1/m} \) as shown below, where dashed lines indicate the optimal algorithm.

**Theorem** (Linear convergence). The iterate sequence \( \{x^k\}_{k \geq 0} \) of each agent \( i \in \{1, \ldots, n\} \) converges to the global optimizer \( x^* \) linearly with rate \( \rho \). In other words,
\[
\|x^k - x^*\| = \mathcal{O}(\rho^k)
\]
for all \( i \in \{1, \ldots, n\} \).

Our decentralized algorithm has the same worst-case convergence rate as centralized gradient descent in terms of the number of gradient evaluations.

#communications
#gradient evaluations = 1
2
3
\leq 5
\leq 10
\leq 20
\leq 100
well-conditioned
ill-conditioned
fully-connected
disconnected

\|x^k - x^*\| = \mathcal{O}(\rho^k)
\]
for all \( i \in \{1, \ldots, n\} \).

Our decentralized algorithm has the same worst-case convergence rate as centralized gradient descent in terms of the number of gradient evaluations.

Assumptions

1. Each local function \( f_i \) is \( \mu \)-smooth and \( L \)-strongly convex.
2. At each round of communication, each agent \( i \) has access to the \( i \)th row of a stochastic gossip matrix with spectral gap \( \sigma \in (0, 1) \).
3. Each agent knows the global parameters \( \mu, L, \) and \( \sigma \).

Corollary (Time complexity). Suppose it takes

- \( \tau \) time per communication round and
- unit time for evaluating local gradients.

Then the time to obtain a solution with precision \( \epsilon > 0 \) is

\[
\mathcal{O}\left( \kappa \left( 1 + \frac{1}{\sqrt{\sigma}} \right) \ln \left( \frac{1}{\epsilon} \right) \right)
\]

as \( \kappa \to \infty \) and \( \sigma \to 1 \), where \( \kappa = L/\mu \).

Convergence Rate

Our algorithm uses an optimized ratio between the number of communication rounds and gradient evaluations.

To account for the extra communication, define a step as one round of communication and at most one gradient evaluation.

The convergence rate of our algorithm per step is \( \rho^{1/m} \) as shown below, where dashed lines indicate the optimal algorithm.

Our algorithm has near-optimal convergence rate in terms of the number of steps.