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

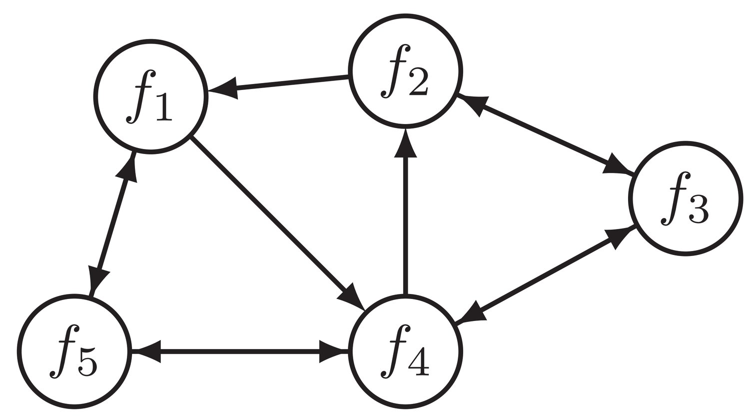
Bryan Van Scoy and Laurent Lessard — University of Wisconsin–Madison

Introduction

Distributed optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \;\; \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

- $f_i : \mathbb{R}^d \to \mathbb{R}$ 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 $\sigma = \|W - \frac{1}{n}\mathbb{1}\mathbb{1}^\mathsf{T}\|$.
- W is **stochastic** if $W\mathbb{1} = \mathbb{1}$ and $\mathbb{1}^\mathsf{T} W = \mathbb{1}^\mathsf{T}$.

For example, a gossip matrix for the above network is

$$W = \begin{bmatrix} 0 & \frac12 & 0 & 0 & \frac12 \\ 0 & 0 & \frac34 & \frac14 & 0 \\ 0 & \frac12 & 0 & \frac12 & 0 \\ \frac14 & 0 & \frac14 & 0 & \frac12 \\ \frac34 & 0 & 0 & \frac14 & 0 \end{bmatrix} \quad \text{with} \quad \sigma \approx 0.7853.$$

Assumptions

(1) Each local function f_i is μ-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 μ, L, and σ.

Algorithm

Notation: Bold indicates concatenated vectors, for example,

$$\mathbf{x}^k := \begin{bmatrix} x_1^k \\ \vdots \\ x_n^k \end{bmatrix} \quad \text{and} \quad \nabla f(\mathbf{v}^k) := \begin{bmatrix} \nabla f_1(v_1^k) \\ \vdots \\ \nabla f_n(v_n^k) \end{bmatrix}.$$

Initialization:

- Set $\mathbf{y}^0 = 0$ and $\mathbf{x}^0$ arbitrary.
- Define the **contraction factor** $\rho = \frac{L-\mu}{L+\mu}$ and
- the number of communications per gradient evaluation

$$m = \underset{r\ge\rho,\, s\ge\sigma}{\text{minimize}} \left\lceil \log_s\!\left(\tfrac{\sqrt{1+r}-\sqrt{1-r}}{2}\right)\right\rceil.$$

for iteration $k = 0, 1, 2, \ldots$ **do**

$\mathbf{v}^k = (W_{k,1}\cdots W_{k,m}\otimes I_d)\,\mathbf{x}^k$ — m communications

$\mathbf{u}^k = \nabla f(\mathbf{v}^k)$ — gradient evaluation

$\mathbf{y}^{k+1} = \mathbf{y}^k + \mathbf{x}^k - \mathbf{v}^k$ — state update

$\mathbf{x}^{k+1} = \mathbf{v}^k - \frac{2}{L+\mu}\,\mathbf{u}^k - \sqrt{1-\rho^2}\,\mathbf{y}^{k+1}$ — state update

end for

Theoretical Results

Theorem (Linear convergence)**.** The iterate sequence $\{x_i^k\}_{k\ge0}$ of each agent $i\in\{1,\ldots,n\}$ converges to the global optimizer $x^\star$ linearly with rate ρ. In other words,

$$\|x_i^k - x^\star\| = \mathcal{O}(\rho^k) \quad \text{for all } i\in\{1,\ldots,n\}.$$

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

Corollary (Time complexity)**.** Suppose it takes

- τ time per communication round and
- unit time for evaluating local gradients.

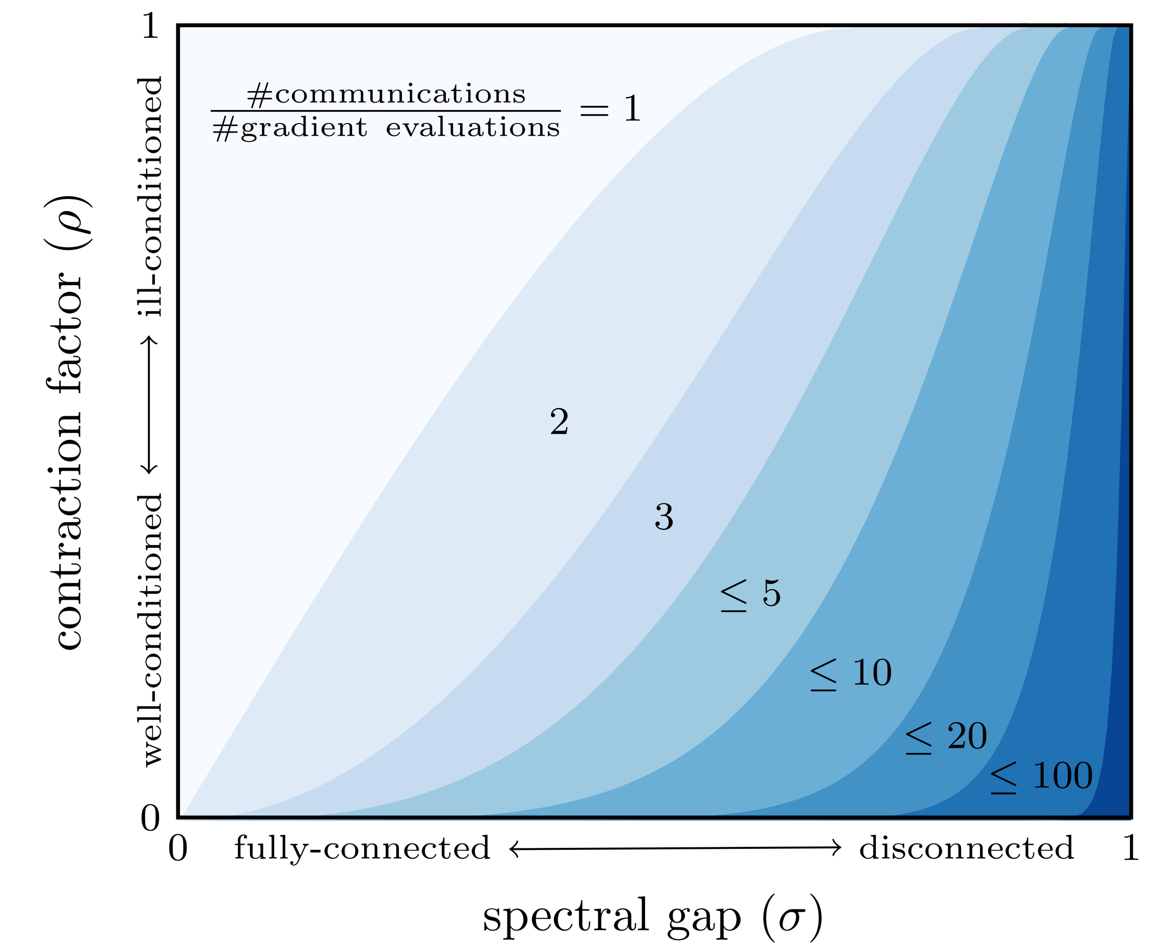
Then the time to obtain a solution with precision $\epsilon > 0$ is

$$\mathcal{O}\!\left(\kappa\left(1+\tfrac{\tau}{\sqrt{1-\sigma}}\right)\ln\!\left(\tfrac{1}{\epsilon}\right)\right)$$

as $\kappa\to\infty$ and $\sigma\to1$, where $\kappa = L/\mu$.

Convergence Rate

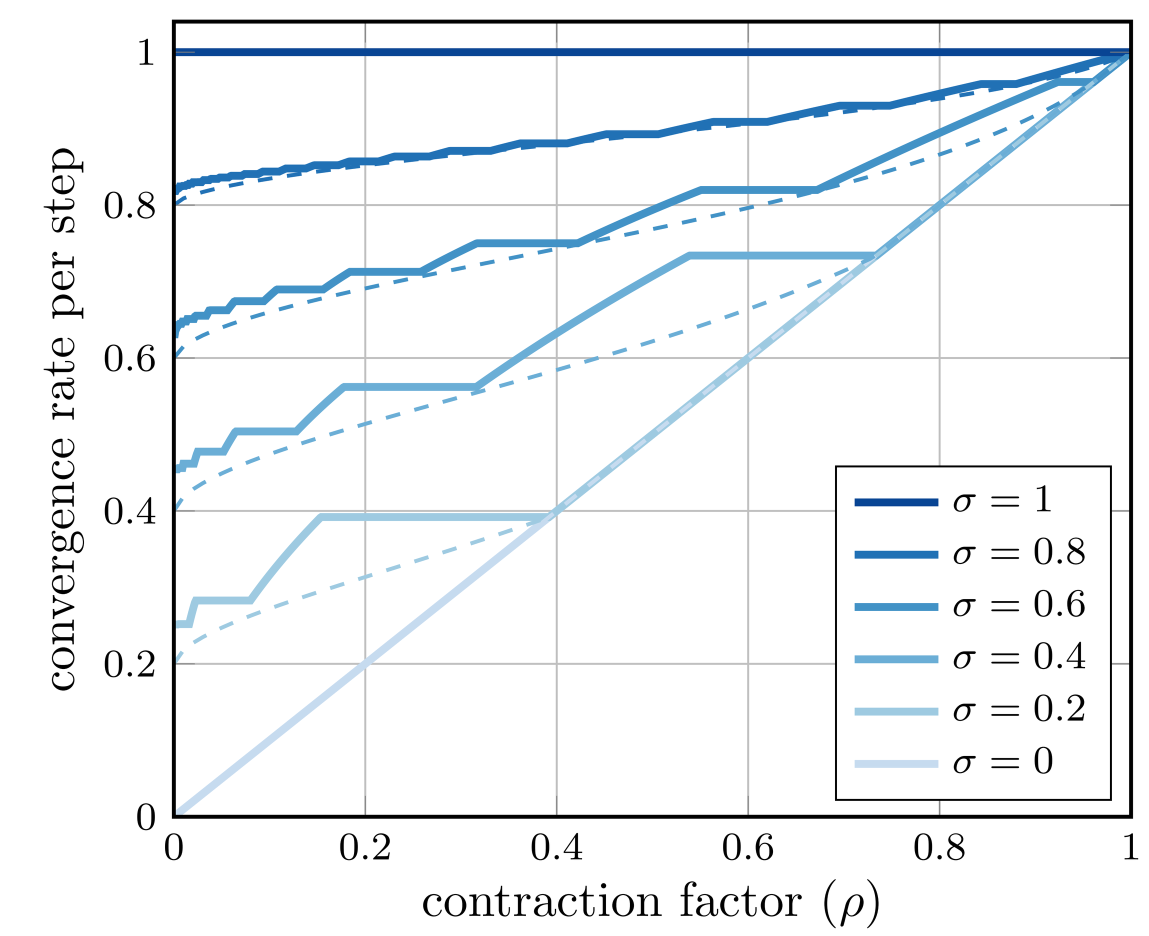
Our algorithm uses an optimized ratio between the number of communication rounds and gradient evaluations.



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

iteration	1			2			3		
step	1	2	3	4	5	6	7	8	9
communication	✓	✓	✓	✓	✓	✓	✓	✓	✓
gradient evaluation			✓			✓			✓

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



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

[a] A. Sundararajan, B. Van Scoy, and L. Lessard. Analysis and design of first-order distributed optimization algorithms over time-varying graphs. arXiv:1907.05448, 2019