Nonconvex Distributed Optimization

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

Introduction

Distributed optimization problem:

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

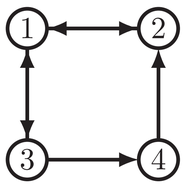
- $f_i : \mathbb{R}^d \to \mathbb{R}$ is the local objective function on agent i
- n is the number of agents
- d is the dimension of the problem

Goal: Each agent must compute the global optimizer by communicating with local neighbors and performing local computations

Communication Network

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 $\mathbb{1}^\mathsf{T}W = \mathbb{1}^\mathsf{T}$



$$W = \begin{bmatrix} 0 & \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \qquad \sigma = \frac{2}{3}$$

Assumptions

(1) There exists a **stationary point** $x^\star \in \mathbb{R}^d$ such that

$$\sum_{i=1}^{n} \nabla f_i(x^\star) = 0.$$

(2) The gradient descent operator $T_i := I - \alpha\nabla f_i$ with stepsize $\alpha > 0$ is a contraction with respect to $x^\star$ with **contraction factor** $\rho \in (0,1)$. In other words,

$$\|T_i(x) - T_i(x^\star)\| \le \rho\,\|x - x^\star\|$$

for all $x \in \mathbb{R}^d$ and all $i \in \{1,\dots,n\}$.

(3) Each agent $i \in \{1,\dots,n\}$ has access to the i^{th} row of a stochastic gossip matrix with **spectral gap** $\sigma \in [0,1)$.

Algorithm

Initialization: Set $y_i^0 = 0 \in \mathbb{R}^d$ and $x_i^0 \in \mathbb{R}^d$ arbitrary
for iteration $k = 0, 1, 2, \dots$ **do**
  for agent $i \in \{1,\dots,n\}$ **do**

$$v_i^k = \mathsf{gossip}(\{x_i^k\}, \{w_{ij}\}, \rho, \sigma)$$
$$y_i^{k+1} = y_i^k + x_i^k - v_i^k$$
$$x_i^{k+1} = v_i^k - \alpha\,\nabla f_i(v_i^k) - \sqrt{1-\rho^2}\,y_i^{k+1}$$

  end for
end for
return $x_i^k \in \mathbb{R}^d$ is the estimate of $x^\star$ on agent i at iteration k

---

Function: $\mathsf{gossip}(\{x_i\}, \{w_{ij}\}, \rho, \sigma)$

Initialization: Set $\gamma^0 = 1$, $\gamma^1 = \frac{1}{\sigma}$, $v_i^0 = x_i$, $v_i^1 = \sum_{j=1}^n w_{ij}x_j$, and define the number of rounds of communication

$$m := \left\lceil \frac{\cosh^{-1}\left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{\rho}\right)}{\cosh^{-1}\left(\frac{1}{\sigma}\right)} \right\rceil$$

for communication round $\ell = 1, \dots, m-1$ **do**
  for agent $i \in \{1,\dots,n\}$ **do**

$$\gamma^{\ell+1} = \tfrac{2}{\sigma}\gamma^\ell - \gamma^{\ell-1}$$
$$v_i^{\ell+1} = \tfrac{2}{\sigma}\tfrac{\gamma^\ell}{\gamma^{\ell+1}}\sum_{j=1}^n w_{ij}v_j^\ell - \tfrac{\gamma^{\ell-1}}{\gamma^{\ell+1}}v_i^{\ell-1}$$

  end for
end for
return $v_i^m \in \mathbb{R}^d$ is the estimate of the average of $\{x_i\}$ on agent i

Theoretical Results

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

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

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

- τ time for communication (multiplying by W), and
- unit time for computation (evaluating $\{\nabla f_i\}$).

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

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

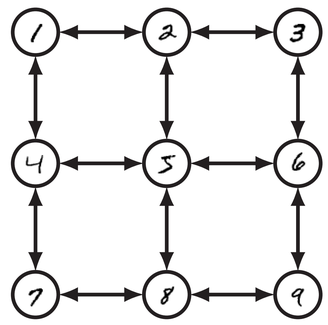
as $\kappa \to \infty$ and $\sigma \to 1$ where $\rho = \frac{\kappa-1}{\kappa+1}$.

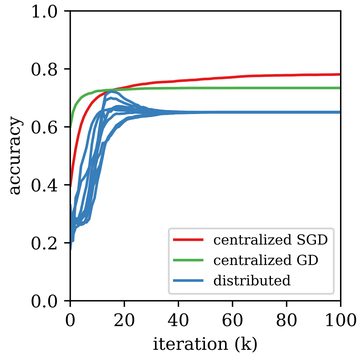
Distributed Machine Learning

- MNIST dataset of handwritten numbers
- Each agent has 1000 samples of the **same** digit
- Data is **private** to each agent (not communicated)
- Agents communicate model with **local** neighbors
- The model on agent i is

$$\text{prediction} = \text{softmax}(A_i \cdot \text{image} + b_i)$$

- Agents learn a **global** model for predicting **all** digits





Target Localization

- Target located at position $x^\star = (p^\star, q^\star) \in \mathbb{R}^2$
- Agent i knows its position $(p_i, q_i) \in \mathbb{R}^2$ and distance to the target

$$r_i = \sqrt{(p_i - p^\star)^2 + (q_i - q^\star)^2}$$

- The objective function $f_i : \mathbb{R}^2 \to \mathbb{R}$ on agent i is

$$f_i(p,q) = \tfrac{1}{2}\left(\sqrt{(p_i - p)^2 + (q_i - q)^2} - r_i\right)^2$$

- Each agent learns the position of the target