Nonconvex Distributed Optimization

Bryan Van Scoy

vanscoy@wisc.edu • vanscoy.github.io

University of Wisconsin–Madison

Distributed optimization

- multiple interacting agents
- agents compute **local** quantities
- agents communicate with **local** neighbors through a network

> Goal is to optimize a **global** performance metric

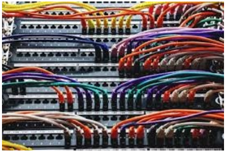
vehicle platoons

drone networks

smart grid

wind farms

load balancing

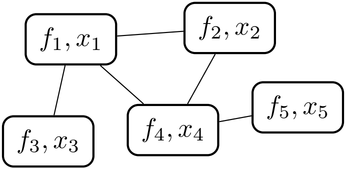


routing and congestion

Problem setup

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

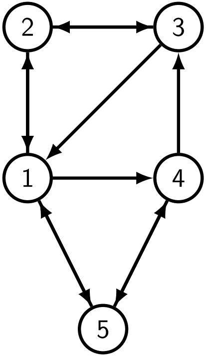
- $f_i : \mathbb{R}^d \to \mathbb{R}$ is the local objective function associated with agent i
- n is the number of agents
- d is the dimension of the problem



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

Communication network

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



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

$$[Wx]_i = \sum_{j=1}^{n} w_{ij}\, x_j$$

$$\sigma \approx 0.7853$$

A first approach

Centralized gradient descent:

$$x_i^{k+1} = x_i^k - \alpha^k \,\mathsf{avg}\big(\{\nabla f_j(x_j)\}_{j=1}^n\big) \qquad\qquad x_i^0 = x^0 \in \mathbb{R}^d$$

- Requires computing an **exact** average at each iteration (costly)
- Linear convergence to optimal solution with constant stepsize

A first approach

Centralized gradient descent:

$$x_i^{k+1} = x_i^k - \alpha^k \,\mathsf{avg}\big(\{\nabla f_j(x_j)\}_{j=1}^n\big) \qquad\qquad x_i^0 = x^0 \in \mathbb{R}^d$$

- Requires computing an **exact** average at each iteration (costly)
- Linear convergence to optimal solution with constant stepsize

Distributed gradient descent:

$$x_i^{k+1} = \sum_{j=1}^n w_{ij}\, x_j^k - \alpha^k \,\nabla f_i(x_i^k) \qquad\qquad x_i^0 \in \mathbb{R}^d$$

- Uses only local communication at each iteration (cheap)
- Linear convergence to **suboptimal** solution with constant stepsize
- **Sublinear** convergence to optimal solution with decaying stepsize

Other distributed algorithms

- Many other distributed algorithms have been proposed recently
- Achieve linear convergence to the optimal solution using two states

$$\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha\nabla f(\mathbf{x}^k) \qquad \textbf{(DGD)}$$

$$\mathbf{x}^{k+1} = 2W\mathbf{x}^k - W^2\mathbf{x}^{k-1} - \alpha\nabla f(\mathbf{x}^k) + \alpha\nabla f(\mathbf{x}^{k-1}) \qquad \textbf{(DIGing)}$$

$$\mathbf{x}^{k+1} = (I+W)\mathbf{x}^k - \tfrac{I+W}{2}\mathbf{x}^{k-1} - \alpha\nabla f(\mathbf{x}^k) + \alpha\nabla f(\mathbf{x}^{k-1}) \qquad \textbf{(EXTRA)}$$

$$\mathbf{x}^{k+1} = (I+W)\mathbf{x}^k - \tfrac{I+W}{2}\big(\mathbf{x}^{k-1} + \alpha\nabla f(\mathbf{x}^k) - \alpha\nabla f(\mathbf{x}^{k-1})\big) \qquad \textbf{(NIDS)}$$

Main result

We construct a novel distributed algorithm with the following properties:

- Worst-case guarantees for **nonconvex** functions
  - convergence rate is the same as centralized gradient descent in terms of number of gradient evaluations, provided we use "enough" communication at each iteration
- Modular approach: communication network can be either
  - directed and time-varying
  - undirected and constant
- Simple convergence proof using a Lyapunov function

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) There exists $\rho \in (0,1)$, called the **contraction factor**, such that

$$\left\|(x - x^\star) - \alpha\left(\nabla f_i(x) - \nabla f_i(x^\star)\right)\right\| \leq \rho \left\|x - x^\star\right\|$$

for all $x \in \mathbb{R}^d$ and all $i \in \{1, \dots, n\}$ where $\alpha > 0$ is the stepsize

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

Nominal algorithm

Parameters: convergence factor $\rho \in (0,1)$, stepsize $\alpha > 0$

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

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

  for agent $i \in \{1,\dots,n\}$ **do**

$$v_i^k = \sum_{j=1}^{n} w_{ij}^k \, x_j^k$$

$$y_i^{k+1} = y_i^k + x_i^k - v_i^k$$

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

  end for

end for

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

Nominal algorithm

Parameters: convergence factor $\rho \in (0,1)$, stepsize $\alpha > 0$

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

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

  for agent $i \in \{1,\dots,n\}$ **do**

$$v_i^k = \sum_{j=1}^{n} w_{ij}^k \, x_j^k$$

$$y_i^{k+1} = y_i^k + x_i^k - v_i^k$$

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

  end for

end for

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

At steady-state,

$$\lim_{k\to\infty} x_i^k = x^\star \qquad \text{and} \qquad \lim_{k\to\infty} y_i^k \propto \nabla f_i(x^\star)$$

Nominal algorithm

Parameters: convergence factor $\rho \in (0,1)$, stepsize $\alpha > 0$

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

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

  for agent $i \in \{1,\dots,n\}$ **do**

$$v_i^k = \sum_{j=1}^{n} w_{ij}^k \, x_j^k$$

$$y_i^{k+1} = y_i^k + x_i^k - v_i^k$$

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

  end for

end for

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

Theorem (Linear convergence)

If $\sigma \le \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$, then the iterate sequence $\{x_i^k\}_{k\ge 0}$ of each agent i converges to the optimal solution $x^\star$ linearly with rate ρ. In other words,

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

Sketch of proof

(1) Write the algorithm in vectorized form

$$\begin{aligned} \mathbf{v}^k &= (W \otimes I_d)\,\mathbf{x}^k \\ \mathbf{u}^k &= \nabla f(\mathbf{v}^k) \\ \mathbf{y}^{k+1} &= \mathbf{y}^k + \mathbf{x}^k - \mathbf{v}^k \\ \mathbf{x}^{k+1} &= \mathbf{v}^k - \alpha\,\mathbf{u}^k - \lambda\,\mathbf{y}^{k+1} \end{aligned}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\nabla f(\mathbf{x}) = \begin{bmatrix} \nabla f_1(x_1) \\ \vdots \\ \nabla f_n(x_n) \end{bmatrix}$, and $\lambda := \sqrt{1-\rho^2}$

Sketch of proof

(1) Write the algorithm in vectorized form

$$\begin{aligned}
\mathbf{v}^k &= (W \otimes I_d)\,\mathbf{x}^k \\
\mathbf{u}^k &= \nabla f(\mathbf{v}^k) \\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \mathbf{x}^k - \mathbf{v}^k \\
\mathbf{x}^{k+1} &= \mathbf{v}^k - \alpha\,\mathbf{u}^k - \lambda\,\mathbf{y}^{k+1}
\end{aligned}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\nabla f(\mathbf{x}) = \begin{bmatrix} \nabla f_1(x_1) \\ \vdots \\ \nabla f_n(x_n) \end{bmatrix}$, and $\lambda := \sqrt{1-\rho^2}$

(2) Define the fixed point

$$(\mathbf{v}^\star, \mathbf{u}^\star, \mathbf{y}^\star, \mathbf{x}^\star) = \big(\mathbb{1} \otimes x^\star,\ \nabla f(\mathbb{1} \otimes x^\star),\ -\tfrac{\alpha}{\lambda}\,\nabla f(\mathbb{1} \otimes x^\star),\ \mathbb{1} \otimes x^\star\big)$$

Sketch of proof

(1) Write the algorithm in vectorized form

$$\begin{aligned}
\mathbf{v}^k &= (W \otimes I_d)\,\mathbf{x}^k \\
\mathbf{u}^k &= \nabla f(\mathbf{v}^k) \\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \mathbf{x}^k - \mathbf{v}^k \\
\mathbf{x}^{k+1} &= \mathbf{v}^k - \alpha\,\mathbf{u}^k - \lambda\,\mathbf{y}^{k+1}
\end{aligned}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\nabla f(\mathbf{x}) = \begin{bmatrix} \nabla f_1(x_1) \\ \vdots \\ \nabla f_n(x_n) \end{bmatrix}$, and $\lambda := \sqrt{1-\rho^2}$

(2) Define the fixed point

$$(\mathbf{v}^\star, \mathbf{u}^\star, \mathbf{y}^\star, \mathbf{x}^\star) = \big(\mathbb{1} \otimes x^\star,\ \nabla f(\mathbb{1} \otimes x^\star),\ -\tfrac{\alpha}{\lambda}\nabla f(\mathbb{1} \otimes x^\star),\ \mathbb{1} \otimes x^\star\big)$$

(3) Define the error vectors

$$(\bar{\mathbf{v}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{y}}^k, \bar{\mathbf{x}}^k) = \big(\mathbf{v}^k - \mathbf{v}^\star,\ \mathbf{u}^k - \mathbf{u}^\star,\ \mathbf{y}^k - \mathbf{y}^\star,\ \mathbf{x}^k - \mathbf{x}^\star\big)$$

Sketch of proof

(4) Define the Lyapunov function

$$V(\mathbf{x}, \mathbf{y}) := \|\mathsf{avg}(\bar{\mathbf{x}})\|^2 + \begin{bmatrix} \mathsf{dis}(\bar{\mathbf{x}}) \\ \mathsf{dis}(\bar{\mathbf{y}}) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 1 & \lambda \\ \lambda & \lambda \end{bmatrix} \otimes I_{nd} \right) \begin{bmatrix} \mathsf{dis}(\bar{\mathbf{x}}) \\ \mathsf{dis}(\bar{\mathbf{y}}) \end{bmatrix}$$

where $\mathsf{avg}(\mathbf{x}) = \mathbb{1} \otimes \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\mathsf{dis}(\mathbf{x}) = \mathbf{x} - \mathsf{avg}(\mathbf{x})$

Sketch of proof

(4) Define the Lyapunov function

$$V(\mathbf{x},\mathbf{y}) := \|\mathsf{avg}(\bar{\mathbf{x}})\|^2 + \begin{bmatrix}\mathsf{dis}(\bar{\mathbf{x}})\\ \mathsf{dis}(\bar{\mathbf{y}})\end{bmatrix}^\mathsf{T} \left(\begin{bmatrix}1 & \lambda\\ \lambda & \lambda\end{bmatrix} \otimes I_{nd}\right) \begin{bmatrix}\mathsf{dis}(\bar{\mathbf{x}})\\ \mathsf{dis}(\bar{\mathbf{y}})\end{bmatrix}$$

where $\mathsf{avg}(\mathbf{x}) = \mathbb{1} \otimes \frac{1}{n}\sum_{i=1}^n x_i$ and $\mathsf{dis}(\mathbf{x}) = \mathbf{x} - \mathsf{avg}(\mathbf{x})$

(5) The Lyapunov function is decreasing since

$$\begin{aligned} V(\mathbf{x}^{k+1},\mathbf{y}^{k+1}) &= \rho^2\, V(\mathbf{x}^k,\mathbf{y}^k) - \left(\rho^2\,\|\bar{\mathbf{v}}^k\|^2 - \|\bar{\mathbf{v}}^k - \alpha\,\bar{\mathbf{u}}^k\|^2\right) \\ &\quad - 2\rho^2\left(\sigma_0^2\,\|\mathsf{dis}(\bar{\mathbf{x}}^k)\|^2 - \|\mathsf{dis}(\bar{\mathbf{v}}^k)\|^2\right) \\ &\quad - 2\,\sigma_0^2\,\left\|\mathsf{dis}\big(\bar{\mathbf{v}}^k + \lambda\,(\bar{\mathbf{x}}^k + \bar{\mathbf{y}}^k)\big)\right\|^2 \end{aligned}$$

where $\sigma_0 := \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$

Sketch of proof

(4) Define the Lyapunov function

$$V(\mathbf{x},\mathbf{y}) := \|\mathsf{avg}(\bar{\mathbf{x}})\|^2 + \begin{bmatrix}\mathsf{dis}(\bar{\mathbf{x}})\\ \mathsf{dis}(\bar{\mathbf{y}})\end{bmatrix}^\mathsf{T} \left(\begin{bmatrix}1 & \lambda\\ \lambda & \lambda\end{bmatrix} \otimes I_{nd}\right) \begin{bmatrix}\mathsf{dis}(\bar{\mathbf{x}})\\ \mathsf{dis}(\bar{\mathbf{y}})\end{bmatrix}$$

where $\mathsf{avg}(\mathbf{x}) = \mathbb{1} \otimes \frac{1}{n}\sum_{i=1}^n x_i$ and $\mathsf{dis}(\mathbf{x}) = \mathbf{x} - \mathsf{avg}(\mathbf{x})$

(5) The Lyapunov function is decreasing since

$$\begin{aligned} V(\mathbf{x}^{k+1},\mathbf{y}^{k+1}) &= \rho^2\, V(\mathbf{x}^k,\mathbf{y}^k) - \left(\rho^2\,\|\bar{\mathbf{v}}^k\|^2 - \|\bar{\mathbf{v}}^k - \alpha\,\bar{\mathbf{u}}^k\|^2\right) \\ &\quad - 2\rho^2\left(\sigma_0^2\,\|\mathsf{dis}(\bar{\mathbf{x}}^k)\|^2 - \|\mathsf{dis}(\bar{\mathbf{v}}^k)\|^2\right) \\ &\quad - 2\,\sigma_0^2\,\left\|\mathsf{dis}\big(\bar{\mathbf{v}}^k + \lambda\,(\bar{\mathbf{x}}^k + \bar{\mathbf{y}}^k)\big)\right\|^2 \end{aligned}$$

where $\sigma_0 := \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$

(6) Then we have the bound

$$\|x_i^k - x^\star\|^2 \le c\,V(\mathbf{x}^k,\mathbf{y}^k) \le c\,\rho^{2k}\,V(\mathbf{x}^0,\mathbf{y}^0)$$

where $c := \mathsf{cond}\left(\left[\begin{smallmatrix}1 & \lambda\\ \lambda & \lambda\end{smallmatrix}\right]\right)$

Algorithm design (the complicated part)

- Given the algorithm and corresponding Lyapunov function, the convergence proof is quite simple
- The difficult part is finding the algorithm and Lyapunov function
- How we did it:
  - Constructed a small **semidefinite program** that computes the worst-case convergence rate for a given algorithm[1]
  - Constructed a **canonical form** characterizing a large class of distributed algorithms[2]
  - Found the canonical form parameters which **optimize** the worst-case convergence rate

[1] A. Sundararajan, B. Hu, and L. Lessard. Robust convergence analysis of distributed optimization algorithms. Allerton Conference on Communication, Control, and Computing, 2017.
[2] A. Sundararajan, B. Van Scoy, and L. Lessard. A canonical form for first-order distributed optimization algorithms. American Control Conference, 2019 (to appear).

Nominal algorithm

Parameters: convergence factor $\rho \in (0,1)$, stepsize $\alpha > 0$

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

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

  for agent $i \in \{1,\dots,n\}$ **do**

$$v_i^k = \sum_{j=1}^{n} w_{ij}^k \, x_j^k$$

$$y_i^{k+1} = y_i^k + x_i^k - v_i^k$$

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

  end for

end for

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

Theorem (Linear convergence)

If $\sigma \leq \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$, then the iterate sequence $\{x_i^k\}_{k\geq 0}$ of each agent i converges to the optimal solution $x^\star$ linearly with rate ρ. In other words,

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

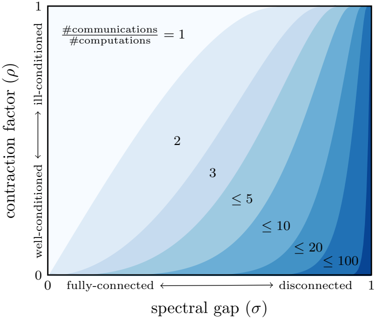
Multi-step gossip

Need graph to be connected enough so that $\sigma \leq \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$

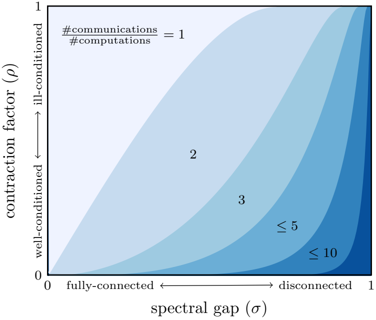
If σ is too large, use multiple gossip steps per iteration

- **standard consensus** if graph is time-varying and/or directed
- **accelerated consensus** if graph is constant and undirected

Standard



Accelerated



Algorithm

Params: convergence factor $\rho \in (0,1)$, spectral gap $\sigma \in [0,1)$, stepsize $\alpha > 0$

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

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

  for agent $i \in \{1,\dots,n\}$ **do**

$$v_i^k = \text{gossip}(\{x_i^k\}, \{w_{ij}^k\}, \rho, \sigma)$$

$$y_i^{k+1} = y_i^k + x_i^k - v_i^k$$

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

  end for

end for

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

gossip function can be:

- **standard consensus** if graph is time-varying and/or directed
- **accelerated consensus** if graph is constant and undirected

Consensus as polynomial filtering

$$\mathbf{v} = p(W)\,\mathbf{x}$$

- Apply a polynomial p of degree m to the gossip matrix W
- m is the number of communication steps required to implement
- Choose p such that $p(1) = 1$ and $|p(w)|$ is small for $w \in [-\sigma, \sigma]$
- Choose m to be the smallest integer such that the spectral gap of $p(W)$ is less than or equal to $\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$

$$p(W) = \begin{cases} W^m & \textbf{standard consensus} \\ \dfrac{T_m(\sigma^{-1}W)}{T_m(\sigma^{-1})} & \textbf{accelerated consensus} \end{cases}$$

T_m is the m^{th} Chebyshev polynomial of the first kind

Standard consensus

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

Initialization: Set $v_i^0 = x_i$ for $i \in \{1, \dots, n\}$, and define the number of rounds of communication

$$m := \left\lceil \frac{\log\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right)}{\log \sigma} \right\rceil$$

for communication round $\ell = 1, \dots, m-1$ **do**

for agent $i \in \{1, \dots, n\}$ **do**

$$v_i^{\ell+1} = \sum_{j=1}^{n} w_{ij}^{\ell} v_i^{\ell}$$

end for

end for

return v_i^m is the estimate of the average of $\{x_i\}$ on agent i

Accelerated consensus

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

Initialization: Set $\gamma^0 = 1$, $\gamma^1 = \sigma^{-1}$, $v_i^0 = x_i$, and $v_i^1 = \sum_{j=1}^n w_{ij} x_j$ for $i \in \{1, \dots, n\}$, 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**

$$\gamma^{\ell+1} = \tfrac{2}{\sigma}\gamma^\ell - \gamma^{\ell-1} \qquad (\gamma^\ell = T_\ell(\sigma^{-1}))$$

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

end for

end for

return v_i^m is the estimate of the average of $\{x_i\}$ on agent i

Complexity

At each iteration, agents must:

- perform m steps of communication with local neighbors
- compute their local gradient

Suppose it takes τ time for communication and unit time for computation

Complexity

At each iteration, agents must:

- perform m steps of communication with local neighbors
- compute their local gradient

Suppose it takes τ time for communication and unit time for computation

Corollary (Time complexity)

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

$$\mathcal{O}\Big(\kappa\left(1+\tfrac{\tau}{1-\sigma}\right)\ \ln\Big(\tfrac{1}{\epsilon}\Big)\Big) \qquad \text{(standard consensus)}$$

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

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

Complexity

At each iteration, agents must:

- perform m steps of communication with local neighbors
- compute their local gradient

Suppose it takes τ time for communication and unit time for computation

Corollary (Time complexity)

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

$$\mathcal{O}\left(\kappa \left(1 + \frac{\tau}{1-\sigma}\right) \ln\left(\frac{1}{\epsilon}\right)\right) \qquad \text{(standard consensus)}$$

$$\mathcal{O}\left(\kappa \left(1 + \frac{\tau}{\sqrt{1-\sigma}}\right) \ln\left(\frac{1}{\epsilon}\right)\right) \qquad \text{(accelerated consensus)}$$

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

If each f_i is **smooth strongly convex** with condition ratio κ, then a lower bound using accelerated consensus is

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

K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. ICML, 2017.

Target localization

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

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

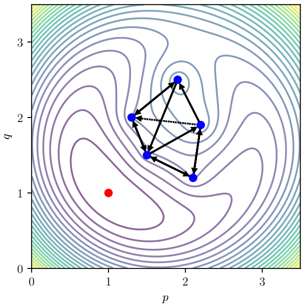
- The objective function $f_i : \mathbb{R}^2 \to \mathbb{R}$ associated to agent i is

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

- To locate the target, agents solve

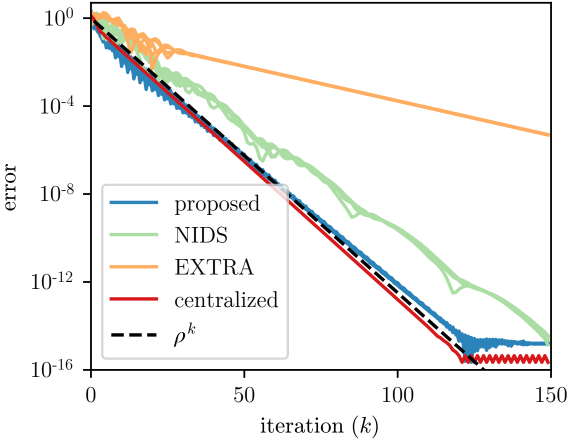
$$\underset{p,q \in \mathbb{R}}{\text{minimize}} \;\; \frac{1}{n}\sum_{i=1}^{n} f_i(p,q)$$

- The optimal stepsize is $\alpha = 2$



Target localization

- Plot of the error $\|x_i^k - x^\star\|$ for each of the $n = 5$ agents
- Our algorithm does one computation and $m = 6$ communications per iteration



Summary

- Worst-case guarantees for **nonconvex** functions
  - convergence rate is the same as centralized gradient descent in terms of number of gradient evaluations, provided we use "enough" communication at each iteration
- Modular approach: communication network can be either
  - directed and time-varying
  - undirected and constant
- Simple convergence proof using a Lyapunov function
- Particularly useful when gradient evaluations are expensive

Paper available at: `https://arxiv.org/abs/1905.11982`