Non-Asymptotic Rates for Communication Efficient Distributed Zeroth Order Strongly Convex Optimization

Anit Kumar Sahu

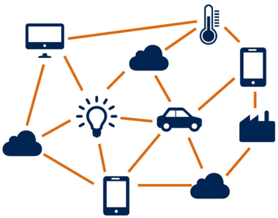
Department of Electrical and Computer Engineering
Carnegie Mellon University

28th November, 2018

Joint work with: Dusan Jakovetic, Dragana Bajovic and Soummya Kar

Motivation: Internet of Things[1]

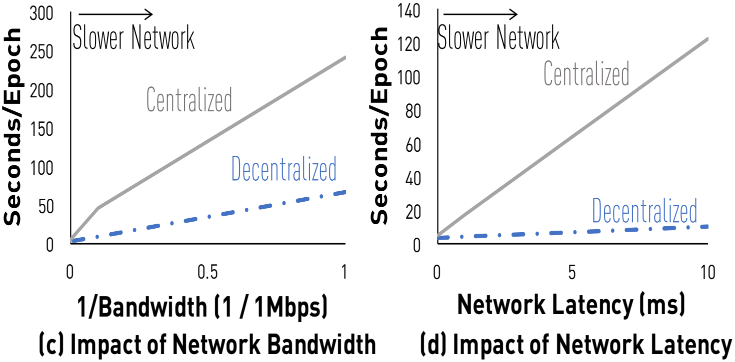
- Myriad of applications: Health monitoring, Smart Home, Smart Campus, Smart Traffic Control
- Needs to be deployed keeping delay sensitivity, scalability, reliability in mind.
- The inherent dynamic nature is a necessary evil.



Internet of Things

[1] https://www.c-mw.net/will-internet-things-impact/

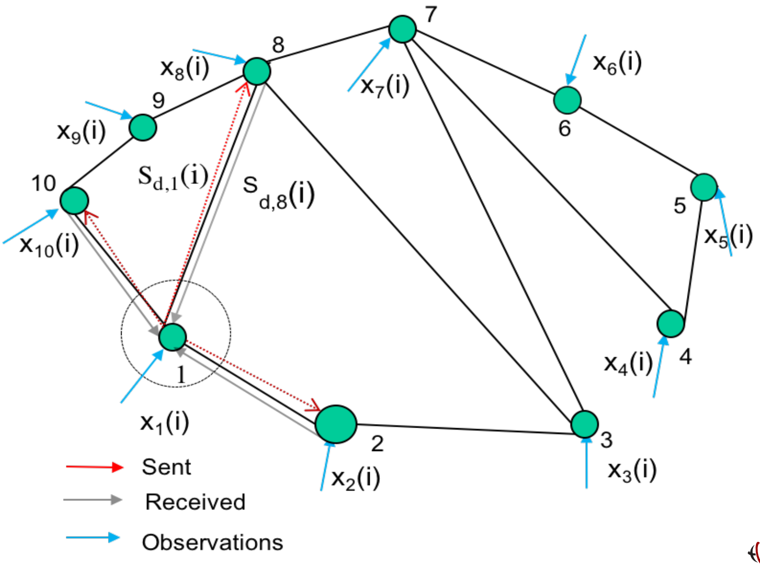
Motivation: Slower Networks



Motivation: Slower Networks[2].

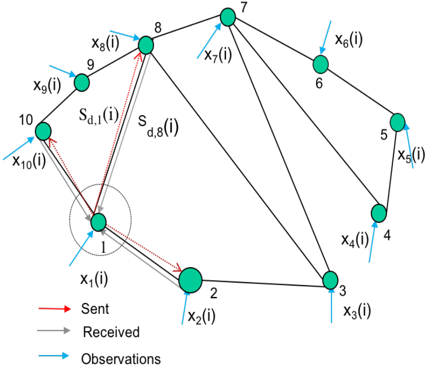
[2] Lian, C. Zhang, H. Zhang, Hsieh, W.Zhang, Liu NIPS '17

Distributed Architecture



Distributed Architecture: Features

- Communication: Raw data never exchanged
- Communication constrained to neighborhood
- Process $S_{d,n}(i)$ is specific to the task.



Distributed Stochastic Optimization

The network of N agents collaboratively aim to solve the following unconstrained problem:

Optimization Problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} \sum_{i=1}^{N} f_i(\mathbf{x}),$$

where $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex function available to node i, $i = 1, ..., N$.

Electrical & Computer ENGINEERING

Assumptions on the cost function

Assumption

For all $i = 1, ..., N$, function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable with Lipschitz continuous gradients. In particular, $\exists L, \mu > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mu \, \mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L\mathbf{I}.$$

Stochastic Zeroth Order Oracle

Stochastic Zeroth Order Oracle (SZO)

A query to the oracle with iterate $\mathbf{x}(k)$ yields, $f(\mathbf{x}(k)) + \mathbf{v}(k)$.

$\mathcal{F}_k$ is the σ-algebra generated by the collection of random variables $\{\,\mathbf{L}(s), \mathbf{v}_i(s)\}$, $i = 1, ..., N$, $s = 0, ..., k-1$.

Related Work: Zeroth Order Stochastic Optimization

- Centralized Case: [KWSA Kiefer '52], [SPSA Spall '92], [Randomized Smoothing based RDSA Nesterov '11], [Duchi '12](Mirror Descent based)
- Non-smooth Case: For LASSO, [Wang '18] with a sparse Hessian assumption show $\mathrm{O}(s^3 \log d)$ dependence, in terms of sparsity s and dimension d. However, rate is $O(k^{-1/3})$.
- Distributed zeroth order optimization with a static network for non-convex losses [Garcia, Hong '17].
- Zeroth order optimization for ADMM, attacks on neural networks [Liu, Chen et.al. '17]
- Best known dimension dependence for one sampled random direction; Zeroth Order Stochastic Frank Wolfe $\mathrm{O}(d^{1/3})$[Sahu, Zaheer, Kar '18]

Electrical & Computer ENGINEERING

Communication Cost

Define the communication cost $\mathcal{C}_t$ to be the expected per-node number of transmissions up to iteration k, i.e.,

$$\mathcal{C}_k = \mathbb{E}\left[\sum_{s=0}^{k-1} \mathbb{I}_{\{node\ C\ \ transmits\ at\ \ s\}}\right],$$

where $\mathbb{I}_A$ represents the indicator of event A.

Electrical & Computer ENGINEERING

Gradient Approximation: KWSA

For dimension $j \in \{1, \cdots, d\}$ agent i queries for $f_i(\mathbf{x}_i(k) + c_k \mathbf{e}_j)$ and $f_i(\mathbf{x}_i(k) - c_k \mathbf{e}_j)$ at time k.

Keifer Wolfowitz Stochastic Approximation

$$\mathbf{e}_j^\top \mathbf{g}_i(\mathbf{x}_i(k)) = \frac{f_i\left(\mathbf{x}_i(k) + c_k \mathbf{e}_j\right)}{2c_k} - \frac{f_i\left(\mathbf{x}_i(k) - c_k \mathbf{e}_j\right)}{2c_k} + \frac{\hat{v}_{i,j}^{+}(k) - \hat{v}_{i,j}^{-}(k)}{2c_k},$$

Randomized Gradient Approximation

KWSA uses $2d$ function evaluations at each iteration.

The following scheme by [Nesterov '11] uses 2 function evaluations at each iteration.

Randomized Gradient Approximation

$$\widetilde{\mathbf{g}}_i(\mathbf{x}_i(k)) = \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k}\right) - \widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k} \mathbf{z}_{i,k},$$

where $\widehat{f}_i(\mathbf{x}_i(k)) = f_i(\mathbf{x}_i(k)) + \widehat{v}_i(k; \mathbf{x}_i(k))$, $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$ and $\mathbf{z}_{i,k} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)$.

Improvised Randomized Gradient Approximation

Zeroth order schemes are plagued by biased gradient estimates.

De-biasing akin to the scheme used in kernel density estimation then yields

Randomized Gradient Approximation

$$
\begin{aligned}
\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) &\doteq 2\widetilde{\mathbf{g}}_i\left(\mathbf{x}_i(k), \frac{c_k}{2}\right) - \widetilde{\mathbf{g}}_i\left(\mathbf{x}_i(k), c_k\right) \\
&= \frac{4\widehat{f}_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right) - 4\widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k} \\
&- \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right) - \widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k}
\end{aligned}
$$

where $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$ and $\mathbf{z}_{i,k} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)$.

Improvised Randomized Gradient Approximation

Assumption

The $z_{i,k}$'s are drawn from a distribution P such that $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$, $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right]$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right]$ are finite.

- $\mathbf{z}_{i,k} \sim \mathcal{N}(0, \mathbf{I}_d)$, then $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d(d+2)$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right] = d(d+2)(d+4)$.
- $\mathbf{z}_{i,k}$'s are drawn uniformly from the l_2-ball of radius $\sqrt{d}$, then we have, $\|\mathbf{z}_{i,k}\| = \sqrt{d}$, $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d^2$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d^3$.

Communication Efficient 0th order Optimization

While ensuring $MSE = O(1/k^{1/2})$, can MSE-communication rate be improved? Yes!

Assumption

For each $i = 1, \ldots, N$, the sequence of measurement noises $\{v_i(k; \mathbf{x}_i(k))\}$ satisfies for all $k = 0, 1, \ldots$:

$$\mathbb{E}[\, v_i(k; \mathbf{x}_i(k)) \,|\, \mathcal{F}_k, \mathbf{z}_{i,k}] = 0, \text{ almost surely (a.s.)}$$
$$\mathbb{E}[\, v_i(k; \mathbf{x}_i(k))^2 \,|\, \mathcal{F}_k, \mathbf{z}_{i,k}] \leq c_v \|\mathbf{x}_i(k)\|^2 + \sigma_v^2, \text{ a.s.,}$$

where c_v and σ_v^2 are nonnegative constants.

$\mathcal{F}_k$ is given by the σ-algebra generated by the collection of random variables $\{\, \mathbf{L}(s), \mathbf{v}(k; \mathbf{x}(k)), \mathbf{z}_{i,s}\}$, $i = 1, \ldots, N$, $s = 0, \ldots, k-1$.

Electrical & Computer ENGINEERING

- For each node i, at every time k, we introduce a binary random variable $\psi_{i,k}$, where

$$\psi_{i,k} = \begin{cases} \rho_k & \textit{with probability } \zeta_k \\ 0 & \textit{otherwise}, \end{cases}$$

- $\psi_{i,k}$'s are independent both across time and nodes.

Communication Efficiency: Parameters

We specifically take ρ_k and ζ_k of the form

$$\rho_k = \frac{\rho_0}{(k+1)^{\epsilon/2}}, \zeta_k = \frac{\zeta_0}{(k+1)^{(\tau_1/2-\epsilon/2)}},$$

where $0 < \epsilon < \tau_1$ and $0 < \tau_1 \leq 1$.

$$\beta_k = (\rho_k \zeta_k)^2 = \frac{\beta_0}{(k+1)^{\tau_1}}, \alpha_k = \frac{a}{k+1}.$$

Communication Efficiency: Graph Sequence

Define the random time-varying Laplacian $\mathbf{L}(k)$, where $\mathbf{L}(k) \in \mathbb{R}^{N\times N}$ as follows:

$$\mathbf{L}_{i,j}(k) = \begin{cases} -\psi_{i,k}\psi_{j,k} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ \sum_{l\neq i} \psi_{i,k}\psi_{l,k} & i = j. \end{cases}$$

Assumption

The inter-agent communication graph is connected on average, i.e., $\lambda_2(\overline{\mathbf{L}}) > 0$.

Communication Efficient 0th order Optimization: Update

For arbitrary initializations $\mathbf{x}_i(0) \in \mathbb{R}^d$, $i = 1, ..., N$, the update rule at node i is given as follows:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \sum_{j \in \Omega_i(k)} \psi_{i,k}\psi_{j,k}\left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) - \alpha_k \widehat{\mathbf{g}}_i(\mathbf{x}_i(k)),$$

The weight sequence $\{\alpha_k\}$ is given by

$$\alpha_k = \alpha_0/(k+1).$$

Communication Efficient Zeroth Order Optimization: Convergence

For each node i's solution estimate $\mathbf{x}_i(k)$, there holds:

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O(1/k^{1/2}).$$

The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k} \zeta_t\right] = O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right),$$

leading to the following MSE-communication rate:

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\mathcal{C}_k^{-\frac{2}{3}+\zeta}\right),$$

where ζ can be arbitrarily small.

Electrical & Computer ENGINEERING

Convergence Under 2nd order smoothness

Assumption

For all $i = 1, ..., N$, the functions $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ have their Hessian to be M-Lipschitz, i.e.,

$$\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|, \forall i = 1, \cdots, N.$$

For each node i's solution estimate $\mathbf{x}_i(k)$, there holds:

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O(1/k^{2/3}).$$

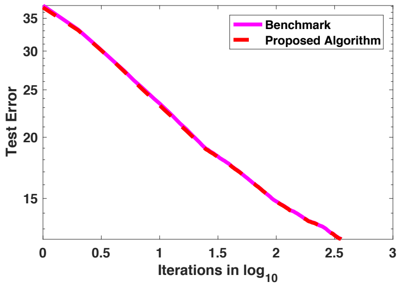
The MSE-communication rate:

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\mathcal{C}_k^{-\frac{8}{9}+\zeta}\right),$$

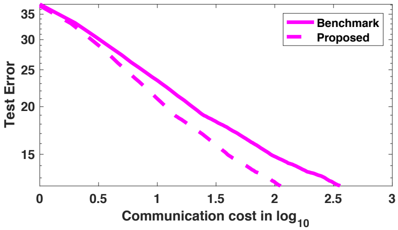
where ζ can be arbitrarily small.

Simulation Experiments: Abalone Dataset

10 agent network, 4177 data points, 8 features, Reported MSE on test set of 577 data points.



Communication Efficient 0th order Optimization: Test Error vs Iteration



Communication Efficient 0th order Optimization: Test Error vs Communication Cost

Summary and Future Work

- Proposed a communication efficient zeroth order optimization algorithm.
- Established non-asymptotic MSE-communication rates.
- Future work: Extension to non-convex functions and improving dimension dependence.

Thank you!

Questions?