Distributed Optimization over Random Networks

Ilan Lobel and Asu Ozdaglar

Allerton Conference
September 2008

Operations Research Center and Electrical Engineering & Computer Science

Massachusetts Institute of Technology, USA

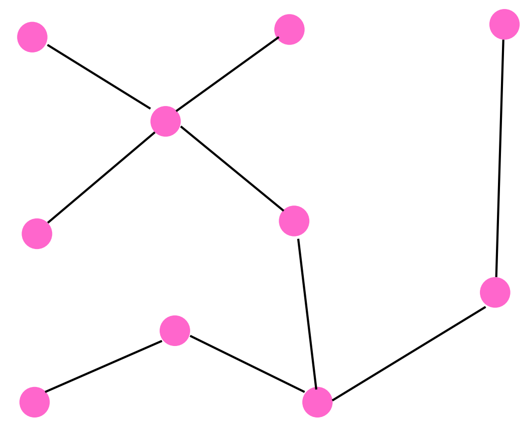
Motivation

- Increasing interest in **distributed control and coordination** of networks consisting of multiple autonomous agents
- Motivated by many emerging networking applications, such as ad hoc wireless communication networks and sensor networks, characterized by:
  - Lack of centralized control and access to information
  - Randomly varying connectivity
- Control algorithms for such networks should be:
  - Completely distributed relying on local information
  - Robust against changes in the network topology

Multi-Agent Optimization Problem

Goal: Develop a general computational model for cooperatively optimizing a global system objective through local interactions and computations in a **multi-agent system with randomly varying connectivity**

- Global objective is a combination of individual agent performance measures



Examples:

- *Consensus problems*: Alignment of estimates maintained by different agents
  - Control of moving vehicles (UAVs), computing averages of initial values
- *Parameter estimation in distributed sensor networks*:
  - Regression-based estimates using local sensor measurements
- *Congestion control in data networks with heterogeneous users*

Related Literature

- **Parallel and Distributed Algorithms:**
  - General computational model for distributed asynchronous optimization
    - Tsitsiklis 84, Bertsekas and Tsitsiklis 95
- **Consensus and Cooperative Control:**
  - Analysis of group behavior (flocking) in dynamical-biological systems
    - Vicsek 95, Reynolds 87, Toner and Tu 98
  - Mathematical models of consensus and averaging
    - Jadbabaie *et al.* 03, Olfati-Saber and Murray 04, Olshevsky, Tsitsiklis 07
- **Distributed Multi-agent Optimization:**
  - Distributed subgradient methods with local information and network effects
    - Nedić and Ozdaglar 07, Nedić, Olshevsky, Ozdaglar, Tsitsiklis 07, Nedić, Ozdaglar, Parrilo 08
- **Existing Work:**
  - Focus on deterministic models of network connectivity
  - Worst-case assumptions about connectivity of agents (e.g. bounded communication intervals between nodes)

Random Network Models and Our Contribution

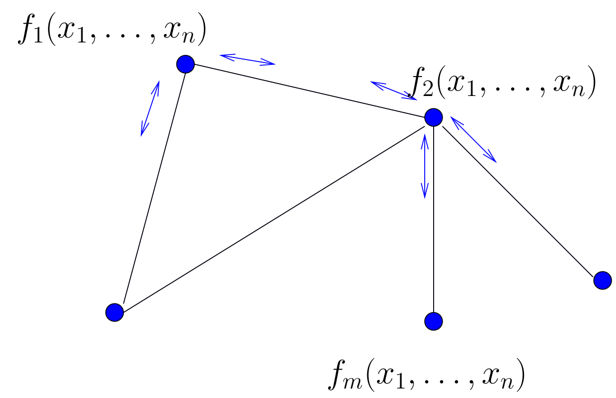
- Recent interest in random algorithms
  - **Randomized gossip algorithms** for averaging or computation of separable functions
    - Boyd *et al.* 05, Mosk-Aoyama and Shah 07
  - **Consensus/agreement over random networks**
    - Availability of links between agents modeled probabilistically
    - Hatano and Mesbahi 05, Wu 06, Tahbaz-Salehi and Jadbabaie 08
- **Our Contributions:**
  - General random network model
  - Development of a distributed subgradient method for multi-agent optimization over a random network
  - Convergence analysis and performance bounds

Model

- We consider a network of n agents with node set $\mathcal{N} = \{1, \ldots, n\}$
  - Agents want to cooperatively solve

$$\min_{x \in \mathbb{R}^m} \sum_{i=1}^{n} f_i(x)$$

  - Function $f_i(x) : \mathbb{R}^n \to \mathbb{R}$ is a convex objective function known only by node i



- Agents update and send their information at discrete times $t_0, t_1, t_2 \ldots$
- We use $x_i(k) \in \mathbb{R}^m$ to denote the estimate of agent i at time t_k

Agent Update Rule:

- Agent i updates his estimate according to:

$$x^i(k+1) = \sum_{j=1}^{n} a_{ij}(k) x_j(k) - \alpha(k) d_i(k)$$

- $a_{ij}(k)$: weight, $\alpha(k)$: stepsize, $d_i(k)$: subgradient of $f_i(x)$ at $x = x_i(k)$
- The model includes consensus as a special case ($f_i(x) = 0$ for all i)

Random Network Model

- Let $x^l(k)$ denote the vector of the l^{th} component of all agent estimates at time k, i.e., $x^l(k) = (x_1^l(k), \ldots, x_n^l(k))$ for all $l = 1, \ldots, m$

- Agent update rule implies that the component vectors of agent estimates evolve according to

$$x^l(k+1) = A(k)x^l(k) - \alpha(k)d^l(k),$$

  - $d^l(k) = (d_1^l(k), \ldots, d_n^l(k))$ is a vector of the l^{th} component of the subgradient vector of each agent
  - $A(k)$ is a matrix with components $A(k) = [a_{ij}(k)]_{i,j \in \mathcal{N}}$

- We assume that $A(k)$ is a **random matrix** that describes the time-varying connectivity of the network

Assumptions

Assumption (*Weights*) Let $\mathcal{F} = (\Omega, \mathcal{B}, \mu)$ be a probability space such that Ω is the set of all $n \times n$ stochastic matrices, $\mathcal{B}$ is the Borel σ-algebra on Ω and μ is a probability measure on $\mathcal{B}$.

(a) There exists a scalar γ with $0 < \gamma < 1$ such that $A_{ii} \geq \gamma$ for all i and all $A \in \Omega$.

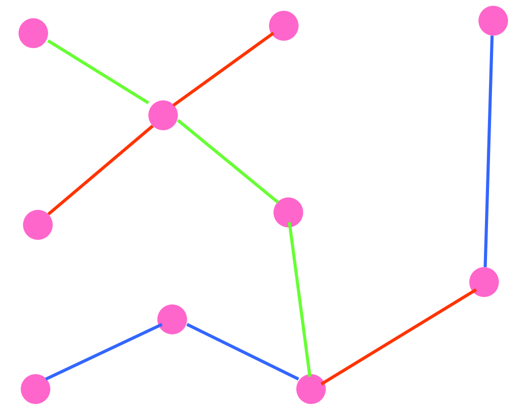
(b) For all $k \geq 0$, the matrix $A(k)$ is drawn independently from probability space $\mathcal{F}$.

Implications:

- Each agent takes convex combination of the information from his neighbors
- The induced graph, i.e., the graph $(\mathcal{N}, \mathcal{E}_+(k))$, $\mathcal{E}_+(k) = \{(j,i) \mid a_{ij}(k) > 0\}$, is a random graph that is independent and identically distributed over time k
  - This allows edges at any time k to be correlated
- Formally, we define a product probability space $(\Omega^\infty, \mathcal{B}^\infty, \mu^\infty) = \prod_{k=1}^{\infty}(\Omega, \mathcal{B}, \mu)$.
- The sequence $A^\infty = \{A(k)\}$ is drawn from this product probability space.

Connectivity

- Consider the expected value of the random matrices $A(k)$,
$$\tilde{A} = E[A(k)] \qquad \text{for all } k \geq 0.$$
- Define the edge set induced by the positive elements of the matrix $\tilde{A}$,
$$\tilde{\mathcal{E}} = \{(j,i) \mid \tilde{A}_{ij} > 0\}$$



Assumption (*Connectivity*) The mean connectivity graph $(\mathcal{N}, \tilde{\mathcal{E}})$ is strongly connected.

- This assumption ensures that in expectation, the information of an agent i reaches every other agent j through a directed path.
- Assume without loss of generality that the scalar $\gamma > 0$ of part (a) of the Weights Assumption satisfies
$$\min_{(j,i)\in\tilde{\mathcal{E}}} \frac{\tilde{A}_{ij}}{2} \geq \gamma.$$

Linear Dynamics and Transition Matrices

- We introduce the **transition matrices**

$$\Phi(k,s) = A(s)A(s+1)\cdots A(k-1)A(k) \qquad \text{for all } k \geq s$$

- We use these matrices to relate $x^i(k+1)$ to $x^j(s)$ at time $s \leq k$:

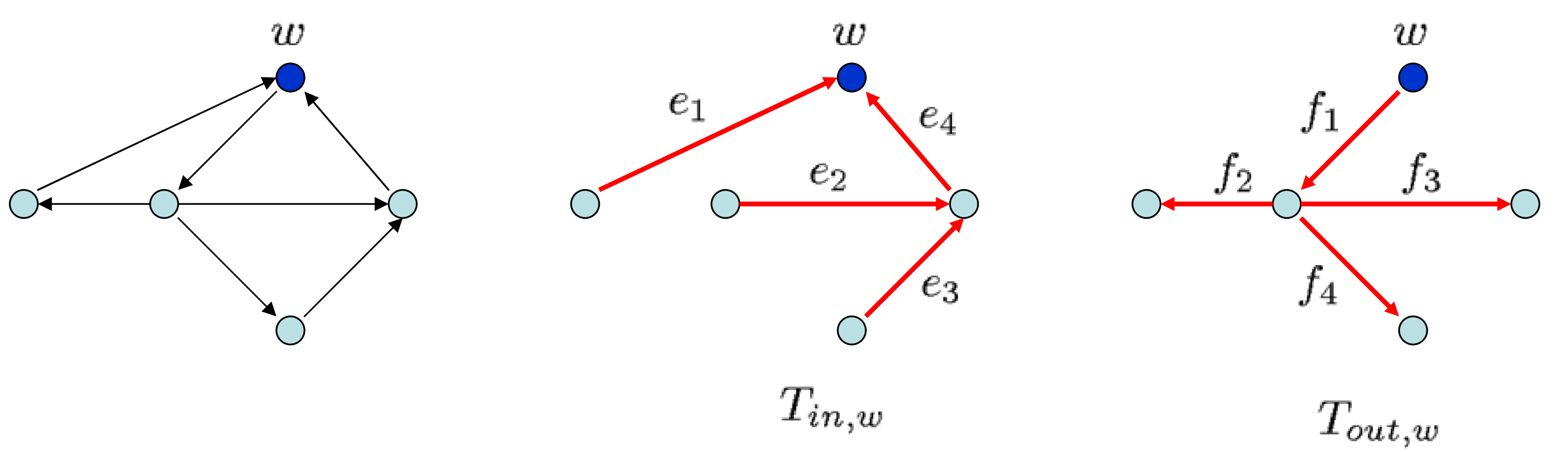
$$x_i(k+1) = \sum_{j=1}^{n}[\Phi(k,s)]_{ij}x_j(s) - \sum_{r=s+1}^{k}\left(\sum_{j=1}^{m}[\Phi(k,r)]_{ij}\alpha(r-1)d_j(r-1)\right) - \alpha(k)d_i(k).$$

- We analyze convergence properties of the distributed method by establishing:
  - Convergence of (random) transition matrices $\Phi(k,s)$ (consensus part)
  - Convergence of an approximate subgradient method (effect of optimization)

Properties of Transition Matrices

- We consider the edge set $\mathcal{E}(k) = \{(j,i) \mid [A(k)]_{ij} \geq \gamma\}$
- We construct a probabilistic event in which the edges of the graphs $\mathcal{E}(k)$ are activated over time k in such a way that information propagates from every agent to every other agent in the network.
- We fix a node $w \in \mathcal{N}$ and consider two directed spanning trees in the mean connectivity graph $(\mathcal{N}, \tilde{\mathcal{E}})$: an in-tree rooted at w, $T_{in,w}$, and an out-tree rooted at w, $T_{out,w}$
- We consider a specific ordering of the edges of these spanning trees:

$$T_{in,w} = \{e_1, e_2, \ldots, e_{n-1}\}, \qquad T_{out,w} = \{f_1, f_2, \ldots, f_{n-1}\}$$



Properties of Transition Matrices (Continued)

Given any time $s \geq 0$, we define the events:

$$C_l(s) = \{A^\infty \in \Omega^\infty \mid A_{e_l}(s+l) \geq \gamma\} \qquad \text{for all } l = 1, \ldots, n-1,$$

$$D_l(s) = \{A^\infty \in \Omega^\infty \mid A_{f_l}(s+(n-1)+l) \geq \gamma\} \qquad \text{for all } l = 1, \ldots, n-1,$$

$$G(s) = \bigcap_{l=1,\ldots,n-1} \Big(C_l(s) \cap D_l(s) \Big).$$

- $G(s)$ denotes the event in which each edge in the spanning trees $T_{in,w}$ and $T_{out,w}$ are activated sequentially following time s in the given order

Lemma: Let Weights and Connectivity Assumptions hold. For any $s \geq 0$, let $A^\infty \in G(s)$. Then,

$$[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1} \qquad \text{for all } i, j, \text{and all } k \geq s + 2(n-1) - 1.$$

Lemma: For any $s \geq 0$, the following hold:

(a) The events $C_l(s)$ and $D_l(s)$ for all $l = 0, \ldots, n-1$ are mutually independent and

$$P(C_l(s)) \geq \gamma, \quad \text{and} \quad P(D_l(s)) \geq \gamma \qquad \text{for all } l = 0, \ldots, n-1.$$

(b) $P(G(s)) \geq \gamma^{2(n-1)}$.

Geometric Decay

Assumption (*Doubly Stochastic Weights*) The matrices $A(k)$ are doubly stochastic with probability 1.

- We introduce the metric

$$b(k,s) = \max_{(i,j)\in\{1,\ldots,n\}^2} \left| [\Phi(k,s)]_i^j - \frac{1}{n} \right| \quad \text{for all } k \geq s \geq 0$$

Lemma: Let Connectivity and Doubly Stochastic Weights assumptions hold. Then,

$$E[b(k,s)] \leq C\beta^{k-s} \quad \text{for all } k \geq s,$$

where β and C are given by

$$C = \left(3 + \frac{2}{\gamma^{2(n-1)}}\right) \exp\left\{-\frac{\gamma^{4(n-1)}}{2}\right\}, \quad \beta = \exp\left\{-\frac{\gamma^{4(n-1)}}{4(n-1)}\right\}$$

Analysis of the Subgradient Method

- Recall the evolution of the estimates:

$$x_i(k+1) = \sum_{j=1}^{n}[\Phi(k,s)]_{ij}x_j(s) - \sum_{r=s+1}^{k}\left(\sum_{j=1}^{m}[\Phi(k,r)]_{ij}\alpha(r-1)d_j(r-1)\right) - \alpha(k)d_i(k).$$

- We introduce a related sequence: Let $y(0)$ be given by $y(0) = \frac{1}{n}\sum_{j=1}^{n} x_j(0)$ and

$$y(k+1) = y(k) - \frac{\alpha(k)}{n}\sum_{j=1}^{n} d_j(k),$$

or equivalently

$$y(k) = \frac{1}{n}\sum_{j=1}^{n} x_j(0) - \frac{1}{n}\sum_{s=1}^{k}\alpha(s)\sum_{j=1}^{n} d_j(s-1).$$

- This iteration can be viewed as an approximate subgradient method for minimizing $f(x) = \sum_j f_j(x)$, in which a subgradient at $x = x_j(k)$ is used instead of a subgradient at $x = y(k)$.

Analysis of the Subgradient Method (Continued)

- We assume that the subgradients of f_i are uniformly bounded by a constant L, and $\max_{1 \leq j \leq n} \|x_j(0)\| \leq L$.
- We consider (weighted) averaged-vectors $\tilde{x}_i(k)$ defined for all $k \geq 0$ as

$$\tilde{x}_i(k) = \frac{1}{\sum_{s=1}^{k} \alpha(s)} \sum_{t=1}^{k} \alpha(t) x_i(t)$$

Proposition: An upper bound on the objective value $f(\tilde{x}_i(k))$ for each i is given by

$$\begin{aligned} f(\tilde{x}_i(k)) \quad &\leq \quad f^* + \frac{n}{2\sum_{r=1}^{k} \alpha(r)} dist^2(y(0), X^*) + \frac{nL^2}{2\sum_{r=1}^{k} \alpha(r)} \sum_{t=1}^{k} \alpha^2(t) \\ &\quad + \frac{3nL^2}{\sum_{r=1}^{k} \alpha(r)} \sum_{t=1}^{k} \alpha(t) \left[n \sum_{s=0}^{t-1} \alpha(s-1) b(t-1, s) + 2\alpha(t-1) \right] \end{aligned}$$

Main Convergence Result for Diminishing Stepsize Rule

We consider a diminishing stepsize rule: The stepsize α_k is decreasing and is s.t.

$$\sum_{k=0}^{\infty} \alpha(k) = \infty \qquad \text{and} \qquad \sum_{k=0}^{\infty} \alpha(k)^2 = A < \infty.$$

Proposition: Let Connectivity-Doubly Stochastic Weights assumptions hold. For all $i \in \mathcal{N}$, we have

$$\lim_{k\to\infty} E[f(\tilde{x}_i(k))] = f^* \text{ and}$$

$$\liminf_{k\to\infty} f(\tilde{x}_i(k)) = f^* \quad \text{with probability } 1.$$

Follows from the lemma:

Lemma: Let Connectivity-Doubly Stochastic Weights assumptions hold. We have

$$\lim_{k\to\infty} E\left[\frac{1}{\sum_{r=1}^{k} \alpha(r)} \sum_{t=1}^{k} \sum_{s=0}^{t-1} \alpha(t)\alpha(s-1)b(t-1,s)\right] = 0 \qquad \text{and}$$

$$\liminf_{k\to\infty} \frac{1}{\sum_{r=1}^{k} \alpha(r)} \sum_{t=1}^{k} \sum_{s=0}^{t-1} \alpha(t)\alpha(s-1)b(t-1,s) = 0 \qquad \text{with probability } 1.$$

Main Convergence Result for Constant Stepsize Rule

We consider a constant stepsize rule: The stepsize α_k satisfies $\alpha_k = \alpha$ for all k and for some scalar $\alpha > 0$

Proposition: Let Connectivity-Doubly Stochastic Weights assumptions hold. For all $i \in \mathcal{N}$, we have

$$\limsup_{k\to\infty} E[f(\tilde{x}_i(k))] \leq f^* + \alpha n L^2 \left(\frac{(1+2n)C}{1-\beta} + \frac{1}{2} \right),$$

where C and β are the constant governing the geometric decay, i.e., $E[b(k,s)] \leq C\beta^{k-s}$ for all $k \geq s$.

Conjecture: For all $i \in \mathcal{N}$, we have

$$\limsup_{k\to\infty} f(\tilde{x}_i(k)) \leq f^* + \alpha n L^2 \left(\frac{(1+2n)C}{1-\beta} + \frac{1}{2} \right) \quad \text{with probability 1.}$$

Conclusions

- We presented a distributed subgradient method for multi-agent optimization with a random network model
- We analyzed the convergence of the algorithm under different stepsize rules
- **Extensions:**
  - Optimization algorithms for the case when probability of link failures depends on the current information state of each node – state-dependent consensus
  - Asynchronous optimization algorithms: Analysis of delay effects on performance
  - Nonconvex local objectives
  - Effects of constraints: distributed primal-dual approaches
  - Second-order distributed optimization algorithms