Distributed Subgradient Methods for Multi-agent Optimization

Asu Ozdaglar

February 2009

Department of Electrical Engineering & Computer Science

Massachusetts Institute of Technology, USA

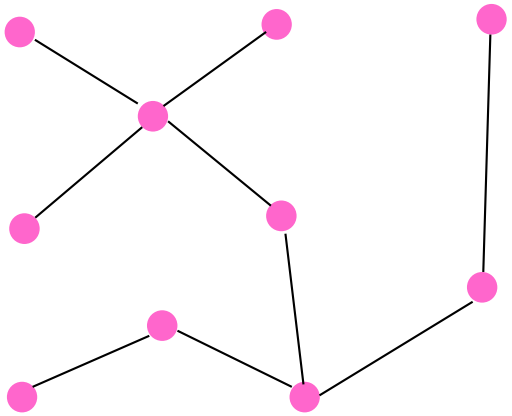
Motivation

- 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
  - Time-varying connectivity
- Control and optimization 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

- 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 Optimization 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 Murray 04, Boyd *et al.* 05
- **Game Theory/Mechanism Design for Distributed Cooperative Control:**
  - Assign each agent a local utility function such that:
    - The equilibrium of the resulting game is the same as (or close to) the global optimum
    - Derive learning algorithms that reach equilibrium
  - **Marden, Arslan, Shamma 07** used this approach for the consensus problem to deal with constraints

This Talk

- Development of a distributed subgradient method for multi-agent optimization **[Nedic, Ozdaglar 08]**
  - Convergence analysis and performance bounds for time-varying topologies under general connectivity assumptions
- Effects of local constraints **[Nedic, Ozdaglar, Parrilo 08]**
- Effects of networked-system constraints: quantization, delay, asynchronism

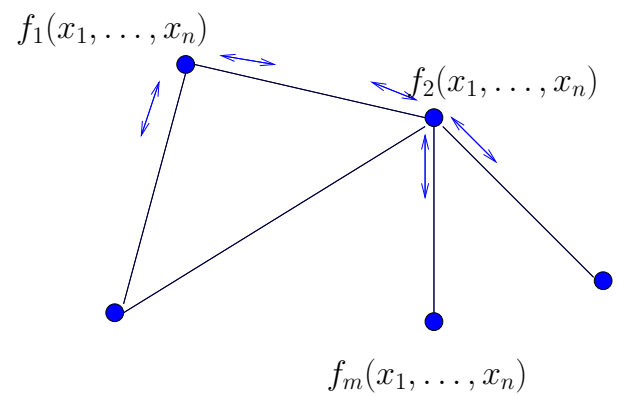
Model

- We consider a network of m agents with node set $V = \{1, \ldots, m\}$

– Agents want to cooperatively solve

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

– 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 $x^i(k) \in \mathbb{R}^n$ to denote the estimate of agent i at time t_k

Agent Update Rule:

- Agent i locally minimizes f_i according to:

$$x^i(k+1) = \sum_{j=1}^{m} a_j^i(k) x^j(k) - \alpha^i(k) d^i(k)$$

$a_j^i(k)$ are weights, $\alpha^i(k)$ is stepsize, $d^i(k)$ is subgradient of f_i at $x^i(k)$

- $a^i(k) = (a_1^i(k), \ldots, a_m^i(k))$ represents i's time-varying neighbors at slot k
- The model includes consensus as a special case ($f_i(x) = 0$ for all i)

Linear Dynamics and Transition Matrices

- We let $A(s)$ denote the matrix whose i^{th} column is the vector $a^i(s)$ and 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 $s \leq k$:

$$x^i(k+1) = \sum_{j=1}^{m} [\Phi(k,s)]^i_j x^j(s) - \sum_{r=s}^{k-1} \left(\sum_{j=1}^{m} [\Phi(k,r+1)]^i_j \alpha^j(r) d^j(r) \right) - \alpha^i(k) d^i(k).$$

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

Assumptions

Assumption (*Weights*) At all times k, we have

(a) The weights $a_j^i(k)$ are nonnegative for all agents i, j.

(b) There exists a scalar $\eta \in (0, 1)$ such that for all agents i,

$$a_i^i(k) \geq \eta \quad \text{and} \quad a_j^i(k) \geq \eta$$

for agents j communicating with agent i at time k.

(c) The agent weight vectors $a^i(k) = [a_1^i(k), \ldots, a_m^i(k)]^T$ are stochastic, i.e., $\sum_{j=1}^m a_j^i(k) = 1$ for all i and k.

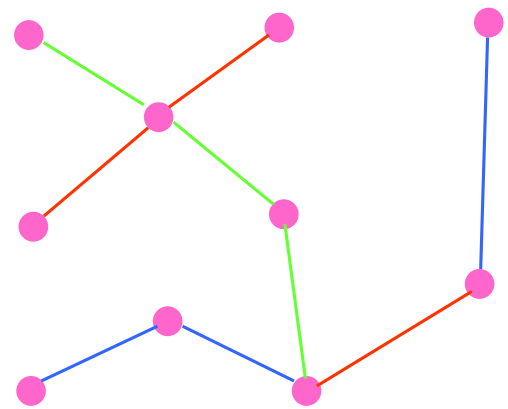
Example: Equal neighbor weights

$$a_j^i(k) = \frac{1}{n_i(k) + 1}$$

where $n_i(k)$ is the number of agents communicating with i at time k

Information Exchange

- Agent i influences any other agent infinitely often - connectivity
- Agent j send his information to neighboring agent i within a bounded time interval - bounded intercommunication interval



At slot k, information exchange may be represented by a directed graph (V, E_k) with $E_k = \{(j,i) \mid a_j^i(k) > 0\}$

Assumption (*Connectivity*) The graph (V, E_∞) is connected, where

$$E_\infty = \{(j,i) \mid (j,i) \in E_k \text{ for infinitely many indices } k\}.$$

Assumption (*Bounded Intercommunication Interval*) There is some $B \geq 1$ s.t.

$$(j,i) \in E_k \cup E_{k+1} \cup \cdots \cup E_{k+B-1} \qquad \text{for all } (j,i) \in E_\infty \text{ and } k \geq 0.$$

Properties of Transition Matrices

Lemma: Let Weights and Information Exchange assumptions hold. We then have

$$[\Phi(s+(m-1)B-1,s)]_j^i \geq \eta^{(m-1)B} \qquad \text{for all } s,\ i, \text{ and } j,$$

where η is the lower bound on weights and B is the intercommunication interval bound.

- We introduce the matrices $D_k(s)$ as follows: for a fixed $s \geq 0$,

  $$D_k(s) = \Phi'\left(s+kB_0-1, s+(k-1)B_0\right) \qquad \text{for } k=1,2,\ldots,$$

  where $B_0 = (m-1)B$.

- By the previous lemma, all entries of $D_k(s)$ are positive.

Convergence of Transition Matrices

Lemma: Let Weights and Information Exchange assumptions hold. For each $s \geq 0$, we have:

(a) The limit $\bar{D}(s) = \lim_{k\to\infty} D_k(s) \cdots D_1(s)$ exists.

(b) The limit $\bar{D}(s)$ has identical rows and the rows are stochastic.

(c) For every j, the entries $[D_k(s) \cdots D_1(s)]_i^j$, $i = 1, \ldots, m$, converge to the same limit $\phi_j(s)$ as $k \to \infty$ with a geometric rate:

$$\left| [D_k(s) \cdots D_1(s)]_i^j - \phi_j(s) \right| \leq 2 \left(1 + \eta^{-B_0}\right) \left(1 - \eta^{B_0}\right)^k$$

where η is the lower bound on weights, B is the intercommunication interval bound, and $B_0 = (m-1)B$.

Proof Outline

- We show that the sequence $\{(D_k \cdots D_1)x\}$ converges for every $x \in \mathbb{R}^m$
- Consider the sequence $\{x_k\}$ with $x_k = D_k \cdots D_1 x$ and write x_k as

$$x_k = z_k + c_k e, \qquad \text{where } c_k = \min_{1 \le i \le m} [x_k]_i$$

- Using the property that each entry of the matrix D_k is positive, we show

$$\|z_k\|_\infty \le \left(1 - \eta^{B_0}\right)^k \|z_0\|_\infty \qquad \text{for all } k.$$

  Hence $z_k \to 0$ with a geometric rate.

- We then show that the sequence $\{c_k\}$ converges to some $\bar{c} \in \mathbb{R}$ and use the contraction constant to establish the rate estimate
- The final relation follows by picking $x = e_j$, the jth unit vector

Convergence of Transition Matrices

Proposition: Let Weights and Information Exchange assumptions hold. For each $s \geq 0$, we have:

(a) The limit $\bar{\Phi}(s) = \lim_{k\to\infty} \Phi(k,s)$ exists.

(b) The limit matrix $\bar{\Phi}(s)$ has identical columns and the columns are stochastic, i.e.,

$$\bar{\Phi}(s) = \phi(s)e',$$

where $\phi(s) \in \mathbb{R}^m$ is a stochastic vector.

(c) For every i, $[\Phi(k,s)]_i^j$, $j = 1, ..., m$, converge to the same limit $\phi_i(s)$ as $k \to \infty$ with a geometric rate, i.e., for all i, j and all $k \geq s$,

$$\left|[\Phi(k,s)]_i^j - \phi_i(s)\right| \leq 2\,\frac{1+\eta^{-B_0}}{1-\eta^{B_0}}\,\left(1-\eta^{B_0}\right)^{\frac{k-s}{B_0}}$$

where η is the lower bound on weights, B is the intercommunication interval bound, and $B_0 = (m-1)B$.

- The rate estimate in part (c) recently improved in **[Nedić, Olshevsky, Ozdaglar, Tsitsiklis 08]**

Convergence Analysis

- Recall the evolution of the estimates (with $\alpha^i(s) = \alpha$):

$$x^i(k+1) = \sum_{j=1}^{m} [\Phi(k,s)]^i_j x^j(s) - \alpha \sum_{r=s}^{k-1} \left(\sum_{j=1}^{m} [\Phi(k,r+1)]^i_j d^j(r) \right) - \alpha d^i(k).$$

- Proof method: Consider a **"stopped process"**, where after time $\bar{k}$, agents stop computing subgradients but keep exchanging their estimates: $d^i(k) = 0$ for all $k \geq \bar{k}$ and all i.

- It can be seen that the stopped process takes the form

$$\bar{x}^i(k+1) = \sum_{j=1}^{m} [\Phi(k,0)]^i_j x^j(0) - \alpha \sum_{r=1}^{\bar{k}} \left(\sum_{j=1}^{m} [\Phi(k,r)]^i_j d^j(r-1) \right)$$

- Using $\lim_{k\to\infty} [\Phi(k,s)]^i_j = \phi_j(s)$ for all i, we see that the limit vector $\lim_{k\to\infty} \bar{x}^i(k)$ exists, independent of i, but depends on $\bar{k}$:

$$\lim_{k\to\infty} \bar{x}^i(k) = y(\bar{k})$$

Behavior of "Stopped Process"

- Stopped process is described by:

$$y(k+1) = y(k) - \alpha \sum_{j=1}^{m} \phi_j(k) d^j(k)$$

  - Subgradients $d^j(k)$ of f_j are computed at $x^j(k)$ instead of $y(k)$

- This would correspond to an **approximate subgradient method** for minimizing $\sum_j f_j(x)$ provided that the values $\phi_j(k)$ are the same for all j

- **Assumption (*Doubly Stochastic Weights*)** The matrices $A(k)$ are doubly stochastic, i.e., $\sum_{i=1}^{m} a_j^i(k) = 1$ for all j and k.
  - Can be ensured when the agents exchange their information simultaneously and coordinate the selection of the weights $a_j^i(k)$

- In this case the stopped process reduces to

$$y(k+1) = y(k) - \frac{\alpha}{m} \sum_{j=1}^{m} d^j(k)$$

Main Convergence Result

Proposition: Let

- Weights, Doubly Stochastic Weights, and Information Exchange assumptions hold
- The subgradients of f_i be uniformly bounded by a constant L, and

$$\max_{1 \le j \le m} \|x^j(0)\| \le \alpha L.$$

Then, for the averages $\hat{x}^i(k)$ of estimates $x^i(0), \ldots, x^i(k-1)$ we have

$$f(\hat{x}^i(k)) \le f^* + \frac{m \; dist^2(y(0), X^*)}{2\alpha k} + \alpha L^2 \left(\frac{C}{2} + 2mC_1 \right)$$

where f^* is the optimal value of $f = \sum_i f_i$, X^* is the optimal set,

$$y(0) = \frac{1}{m} \sum_i x^i(0), \qquad C = 1 + 8mC_1$$

$$C_1 = 1 + \frac{m}{1 - (1 - \eta^{B_0})^{\frac{1}{B_0}}} \; \frac{1 + \eta^{-B_0}}{1 - \eta^{B_0}}, \qquad B_0 = (m-1)B$$

- Estimates are per iteration
- Captures tradeoff between accuracy and computational complexity

Proof Outline

- We analyze the stopped process:

$$y(k+1) = y(k) - \frac{\alpha}{m} \sum_{j=1}^{m} d^j(k)$$

- Establish approximate convergence relation for the running averages

$$\hat{y}(k) = \frac{1}{k} \sum_{h=0}^{k-1} y(h)$$

- Using the convergence rate estimate for the transition matrices $\Phi(k,s)$, we show that $\hat{y}(k)$ are close to the agents' averages $\hat{x}^i(k) = \frac{1}{k}\sum_{h=0}^{k-1} x^i(h)$

$$\|\hat{y}(k) - \hat{x}^i(k)\| \leq 2\alpha L C_1 \qquad \text{for all } i \text{ and } k$$

- Infer the result for the running averages $\hat{x}(k)$

Constrained Consensus Problem

- Estimates of agent i restricted to lie in a **closed convex constraint set X_i**
- We assume that the intersection set $X = \cap_{i=1}^{m} X_i$ is nonempty
- Examples where constraints important:
  - Motion planning and alignment problems, where each agent's position is limited to a certain region or range
  - Distributed constrained multi-agent optimization
- **This talk: Pure consensus problem in the presence of constraints**

Projected Consensus Algorithm

- For the constrained consensus problem, we develop a consensus algorithm based on projections.
- We use $x^i(k) \in X_i$ to denote the estimate of agent i at time t_k
- Given a closed convex set $X \subset \mathbb{R}^n$, and a vector $y \in \mathbb{R}^n$, we define:

$$\mathrm{dist}(y, X) = \min_{x \in X} \|y - x\|, \qquad P_X(y) = \mathrm{argmin}_{x \in X} \|y - x\|$$

Agent Update Rule:

- Agent i updates his estimates subject to his constraint set:

$$x^i(k+1) = P_{X_i}\Big[\sum_{j=1}^{m} a_j^i(k) x^j(k)\Big]$$

where $a^i(k) = (a_1^i(k), \ldots, a_m^i(k))'$ is the weight vector

Connection to Alternating Projections

- Update rule similar to classical alternating projection method

Alternating/Cyclic Projection Methods:

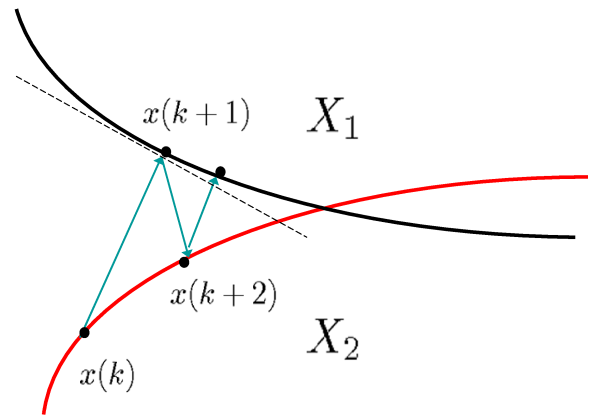
Given closed convex sets $X_1, X_2 \subset \mathbb{R}^n$, find point in $X = X_1 \cap X_2$:

$$x(k+1) = P_{X_1}(x(k))$$
$$x(k+2) = P_{X_2}(x(k+1))$$

Convergence analysis:

X_i affine [Von Neumann; Aronszajn 50],

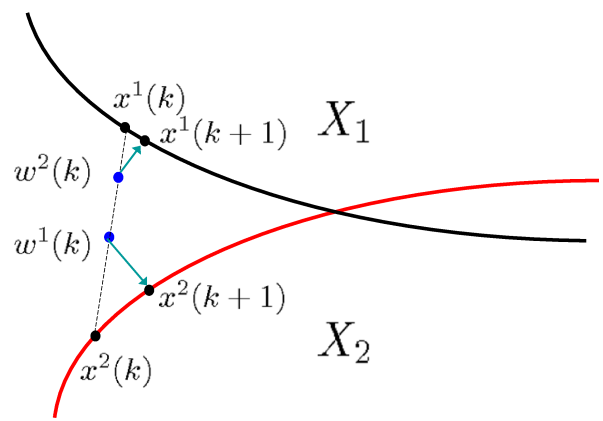
X_i convex [Gubin, Polyak, Raik 67]



Constrained Consensus Algorithm:

Given closed convex sets $X_1, X_2 \subset \mathbb{R}^n$

$$w^1(k) = \sum_j a_j^1(k) x^j(k)$$
$$w^2(k) = \sum_j a_j^2(k) x^j(k)$$
$$x^1(k+1) = P_{X_1}(w^1(k))$$
$$x^2(k+1) = P_{X_2}(w^2(k))$$



Convergence Analysis

- Analysis of the impact of constraints
- Recall the update rule

$$x^i(k+1) = \sum_{j=1}^{m} a_j^i(k) x^j(k) + e^i(k),$$

where the projection error $e^i(k)$ is given by

$$e^i(k) = \mathcal{P}_{X_i}\Big[\sum_{j=1}^{m} a_j^i(k) x^j(k)\Big] - \sum_{j=1}^{m} a_j^i(k) x^j(k) = x^i(k+1) - w^i(k).$$

Proposition: Assume that the intersection $X = \cap_{i=1}^m X_i$ is nonempty. Let Doubly Stochastic Weights assumption hold. Then:

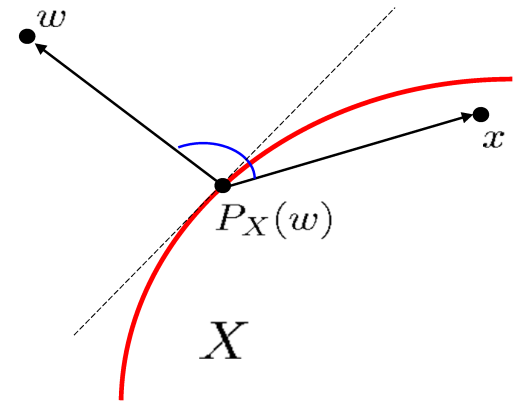
$$\lim_{k\to\infty} e^i(k) = 0 \qquad \text{for all } i.$$

- **Implication:** The analysis translates into unconstrained case
- The proof relies on two lemmas

Convergence of Projection Error

Lemma: Let $w \in \mathbb{R}^n$ and $X \subset \mathbb{R}^n$ closed and convex. For all $x \in X$,

$$\|w - P_X(w)\|^2 \leq \|w - x\|^2 - \|P_X(w) - x\|^2.$$



Lemma: Let Doubly Stochastic Weights assumption hold. Then:

- $\|x^i(k+1) - x\| \leq \|w^i(k) - x\| \ \forall \ x \in X_i$ (nonexpansiveness of projection)
- $\sum_{i=1}^m \|w^i(k) - x\|^2 \leq \sum_{i=1}^m \|x^i(k) - x\|^2 \ \forall \ x$ (by doubly stochasticity)
  - Of independent interest in convergence analysis of doubly stochastic matrices

Implication:

$\lim_{k\to\infty} \sum_{i=1}^m \|w^i(k) - \bar{x}\|^2 - \|x^i(k+1) - \bar{x}\|^2 = 0$, for all $\bar{x} \in X$

Convergence of the Estimates

- Recall that the transition matrices are defined as follows:

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

- Using the transition matrices and the decomposition of the estimate evolution,

$$x^i(k+1) = \sum_{j=1}^{m} [\Phi(k,s)]_j^i \, x^j(s) + \sum_{r=s+1}^{k} \left(\sum_{j=1}^{m} [\Phi(k,r)]_j^i \, e^j(r-1) \right) + e^i(k).$$

- Use a two-time scale analysis, where we define a similar **"stopped process"**

$$y(k) = \frac{1}{m} \sum_{j=1}^{m} x^j(s) + \frac{1}{m} \sum_{r=s+1}^{k} \left(\sum_{j=1}^{m} e^j(r-1) \right)$$

- It can be seen that the agent estimates reach a **consensus**:

$$\lim_{k\to\infty} \|x^i(k) - y(k)\| = 0 \qquad \text{for all } i$$

Proposition: (*Convergence*) Let Weights, Doubly Stochastic Weights, and Information Exchange assumptions hold. We have for some $\tilde{x} \in X$,

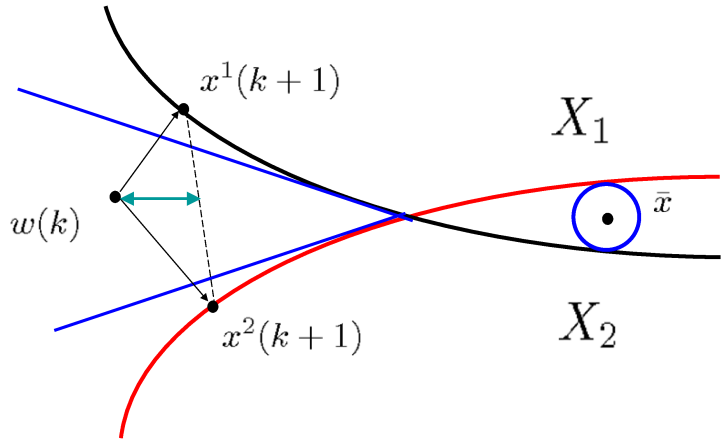
$$\lim_{k\to\infty} \|x^i(k) - \tilde{x}\| = 0 \qquad \text{for all } i.$$

Rate Analysis

Assumption (*Interior Point*): There exists a vector $\bar{x}$ such that

$$\bar{x} \in \text{int}(X) = \text{int}\Big(\cap_{i=1}^{m} X_i \Big),$$

i.e., there exists some scalar $\delta > 0$ such that $\{z \mid \|z - \bar{x}\| \leq \delta\} \subset X$.



Proposition: Let Interior Point Assumption hold. Let the weight vectors $a^i(k)$ be given by $a^i(k) = (1/m, \ldots, 1/m)'$ for all i and k. Then,

$$\sum_{i=1}^{m} \|x^i(k) - \tilde{x}\|^2 \leq \left(1 - \frac{1}{4R^2}\right)^k \sum_{i=1}^{m} \|x^i(0) - \tilde{x}\|^2 \qquad \text{for all } k \geq 0,$$

where $\tilde{x} \in X$ is the limit of the sequence $\{x^i(k)\}$, and $R = \frac{1}{\delta} \sum_{i=1}^{m} \|x^i(0) - \bar{x}\|$, i.e., **the convergence rate is linear**.

- Linear convergence rate extends to time-varying weights with $X_i = X$
- Convergence rate for time-varying weights and different local constraints open

Summary

- We presented a distributed subgradient method for multi-agent optimization
- The method can operate over networks with time-varying connectivity
- We proposed a constrained consensus policy for the case when agents have local constraints on their estimates
- This policy has connections to the classical alternating projection method
- We analyzed the convergence and rate of convergence of the algorithms

Putting Things Together: Constrained Distributed Optimization

- Each agent has a convex closed local constraint set X_i **[Nedić, Ozdaglar, Parrilo 08]**
- Agent i updates his estimate by

$$v^i(k) = \sum_{j=1}^{m} a_j^i(k) x^j(k)$$

$$x^i(k+1) = P_{X_i}\left[v^i(k) - \alpha(k) d^i(k)\right],$$

$\alpha(k) > 0$ is a diminishing stepsize sequence and $d^i(k)$ is a subgradient of $f_i(x)$ at $x = v^i(k)$.

Results:

- Agent estimates generated by this algorithm converge to the same optimal solution for the cases when the weights are constant and equal, and when the weights are time-varying but $X_i = X$ for all i.
- Convergence analysis in the general case open!

Optimization over Random Networks

- Existing work focuses on deterministic models of network connectivity (i.e., worst-case assumptions about intercommunication intervals)
- Time-varying connectivity modeled probabilistically **[Lobel, Ozdaglar 08]**
- Each component l of agent estimates evolve according to

$$x_l(k+1) = A(k)x_l(k) - \alpha(k)d_l(k)$$

- We assume that the matrix $A(k)$ is a random matrix drawn independently over time from a probability space of stochastic matrices.
- This allows edges at any time k to be correlated.

Results:

- We establish properties of random transition matrices by constructing positive probability events in which information propagates from every node to every other node in the network
- We provide convergence analysis of a stochastic subgradient method under different stepsize rules

Limits on Communication

Quantization Effects:

- Agents have access to quantized estimates due to storage and communication bandwidth constraints **[Nedić, Olshevsky, Ozdaglar, Tsitsiklis 07]**
- Agent i updates his estimate by

$$x^i(k+1) = \sum_{j=1}^{m} a_j^i(k) x_Q^j(k) - \alpha d^i(k)$$

$$x_Q^i(k+1) = \lfloor x^i(k+1) \rfloor$$

$\lfloor \cdot \rfloor$ represents rounding down to the nearest integer multiple of $1/Q$

Delay Effects:

- Agents have access to outdated estimates due to communication delays **[Bliman, Nedić, Ozdaglar 08]**
- In the presence of delays, agent i updates his estimate by

$$x^i(k+1) = \sum_{j=1}^{m} a_j^i(k) x^j(k - t_j^i(k)) - \alpha d^i(k)$$

$t_j^i(k)$ is the delay in passing information from j to i

Applications to Social Networks

- Growing interest in dynamics in a social network of communicating agents
- A specific example is learning and information aggregation over networks
- Consensus policies can be used to develop and analyze myopic/quasi-myopic learning models **[Golub, Jackson 07]**, **[Acemoglu, Ozdaglar, ParandehGheibi 08]**
- In the context of social networks, these rules may be too myopic:
  - **Alternative approach:** Bayesian learning over social networks **[Acemoglu, Dahleh, Lobel, Ozdaglar 08]**

References

- A. Nedić and A. Ozdaglar, "Distributed Subgradient Methods for Multi-agent Optimization," IEEE Transactions on Automatic Control, 2008.
- A. Ozdaglar, "Constrained Consensus and Alternating Projections," Proc. of Allerton Conference, 2007.
- A. Nedić, A. Olshevsky, A. Ozdaglar, and J.N. Tsitsiklis, "On Distributed Averaging Algorithms and Quantization Effects," IEEE Transactions on Automatic Control, 2008.
- A. Nedić, A. Ozdaglar, and P.A. Parrilo, "Constrained Consensus and Optimization in Multi-agent Networks," submitted for publication, 2008.
- I. Lobel and A. Ozdaglar, "Distributed Subgradient Methods over Random Networks," Proc. of Allerton Conference, 2008.
- D. Acemoglu, M. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian Learning in Social Networks," submitted for publication, 2008.
- D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi "Spread of (Mis)Information over Social Networks," working paper, 2008.

All papers can be downloaded from web.mit.edu/asuman/www