

RANDOMIZED INCREMENTAL PROTOCOLS OVER ADAPTIVE NETWORKS

Cassio G. Lopes*

Electrical Engineering, University of Sao Paulo
Escola Politecnica, Sao Paulo, Brazil, 05508-970
Email: cassio@lps.usp.br

Ali H. Sayed†

Dept. of Electrical Engineering, University of California
Los Angeles, CA, 90095.
Email: sayed@ee.ucla.edu

ABSTRACT

We introduce an incremental cooperation mode into the framework of adaptive networks (AN). The method applies to generic topologies and avoids the need to establish a Hamiltonian cycle over the network, generalizing the original incremental mode, while keeping nearly the same mean-square performance, as illustrated by the simulations. We motivate the new mode by relying on an LMS rule at the nodes, and mean-square analysis is provided.

Index Terms— Adaptive filters, distributed estimation, adaptive networks, incremental methods, cooperative systems.

1. INTRODUCTION

An adaptive network (AN) is a collection of N adaptive nodes that observe space-time data and collaborate, according to some cooperation protocol, in order to estimate parameters related to some event of interest.

Several ANs have been proposed based on the *incremental* mode of cooperation [1, 2, 3, 4, 5]. The one main advantage is the ability to reduce the required energy and communication resources to implement distributed algorithms. The one major drawback is the need to establish a Hamiltonian cycle over the network, which is an NP-complete problem, and is not guaranteed to exist in the general case. In this work we relax this requirement by defining a random cooperation walk over the nodes. In the context of ANs, this gives rise to a new probabilistic incremental protocol defined by a Markov chain (MC).

We review the steps necessary to embed the protocol in adaptive networks and revisit and extend available incremental algorithms [1, 2, 5], illustrating that both standard incremental and its probabilistic counterpart yield nearly the same performance in the mean-square sense, suggesting that Hamiltonian cycles may be avoided.

2. RELAXING THE INCREMENTAL PROTOCOL

In order to estimate an unknown vector of parameters w^o , incremental protocols operate sequentially over a pre-established N -steps cooperation cycle that visits all the nodes. Node k observes space-time realizations $\{d_k(i), u_{k,i}\}$ of the random data $\{d_k, \mathbf{u}_k\}$, where d_k is a scalar measurement and \mathbf{u}_k is a $1 \times M$ row regressor vector, and updates an estimate $\psi_{k-1}^{(i)}$ of w^o received from node $k-1$ using its learning rule. Subsequently, the estimate $\psi_k^{(i)}$ at node k is passed to the next node $k+1$, and so on.

*The first author was supported partially by FAPESP.

†This material was based on work supported in part by the National Science Foundation under awards ECS-0601266, ECS-0725441, and CCF-094936.

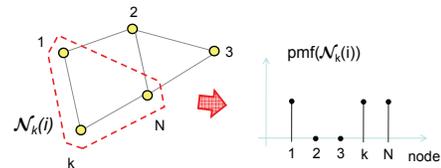


Fig. 1. The probability mass function induced over a neighborhood.

In this work we relax the original cyclic incremental protocol [1, 2] by (a) removing the requirement of a Hamiltonian cycle: nodes decide, locally and on-the-fly, which node comes next in the cooperation effort; and (b) nodes are allowed to be revisited within an N -steps slot, but only one node is communicated with at a time. Cooperation is implemented as a random walk over the network of nodes. For instance, assume the process starts at node k ; the next node ℓ is drawn randomly from a probability mass function (pmf) defined over node k 's neighborhood, $\ell \leftarrow \text{pmf}(\mathcal{N}_k(i))$ (see Fig. 1), and it generates a new estimate ψ_ℓ according to the local learning rule. The cooperation then evolves sequentially and randomly over the available network topology.

The idea of defining random walks on graphs is not new [6]. In fact, the protocol we present here is a distributed instance of randomized algorithms. It has been studied in [7, 8] in the context of distributed deterministic optimization subject to stochastic perturbations. Such protocol is also related to probabilistic diffusion [9] and gossip protocols [10]. Here we motivate a new distributed adaptive algorithm that intertwines *two* stochastic processes: the learning process at the node level, also contaminated by background stochastic noise, and the underlying probabilistic incremental protocol, in charge of cooperation.

Considering that the sequence of cooperating nodes is defined randomly, some natural questions are:

1. Will the nodes be regularly visited?
2. What is the expected visit time for a specific node?
3. Does the learning evolution depend on the initial node?
4. How is the network learning impacted in the mean-square sense?

Such questions can be answered by examining how a Markov chain arises from the probabilistic incremental protocol.

3. FROM THE NETWORK TOPOLOGY TO A MARKOV CHAIN

Let $\mathbf{n}(i)$ denote the node indexing (stochastic) process, generated as described in the previous section. For $i \geq 0$, we have that $\mathbf{n}(i) \in \{1, 2, \dots, N\} \triangleq V$, where V is the node set. Any realization $n_i \triangleq \{n(j), 0 \leq j \leq i\}$ defines a cooperation trajectory over the network that is governed by the joint pmf:

$$P(\mathbf{n}(i) = n(i), \dots, \mathbf{n}(1) = n(1), \mathbf{n}(0) = n(0)) \quad (1)$$

or simply $P(n(i), \dots, n(1), n(0)) \triangleq P(n_i)$. We have

$$\begin{aligned} P(n_i) &= P(n(i)|n_{i-1})P(n_{i-1}) \\ &= P(n(i)|n_{i-1})P(n(i-1)|n_{i-2})P(n_{i-2}) \\ &= P(n(i)|n_{i-1}) \cdots P(n(2)|n_1)P(n(1)|n_0)P(n_0) \end{aligned} \quad (2)$$

where $P(n_0) = P(n(0))$. In other words, the probability that we are at a specific node at time i depends on the earlier trajectory followed across the network. If each node assigns its successor independently from previous assignments, then (2) reduces to

$$P(n_i) = P(n(i)|n(i-1)) \cdots P(n(2)|n(1))P(n(0)) \quad (3)$$

and $\mathbf{n}(i)$ becomes a Markov process. A Markov chain (MC) with N states is induced over the network topology, and $P(n(i)|n(i-1))$ is the *transition probability* from node $n(i)$ to node $n(i-1)$, as depicted in Fig.2. An MC is completely described by an initial probability distribution, defined as the $1 \times N$ row vector

$$\pi_0 = [\pi_{0,k}] \triangleq \{P(\mathbf{n}(0) = k); k = 1, \dots, N\} \quad (4)$$

and the $N \times N$ *transition probability matrix* \mathcal{P} , defined as

$$\mathcal{P} = [p_{k\ell}], \quad p_{k\ell} \triangleq P(\mathbf{n}(i) = \ell | \mathbf{n}(i-1) = k) \quad (5)$$

for $k, \ell = 1, \dots, N$. Note that the standard incremental protocol becomes a special case of the novel protocol [6]. As the chain evolves, π_0 is mixed into the chain marginal probability distribution

$$\pi_i = [\pi_{i,k}] \triangleq \{P(\mathbf{n}(i) = k); k = 1, \dots, N\} \quad (6)$$

which represents the probability of the chain being at one of its N states at time i . Its evolution depends on \mathcal{P} as follows. Note that

$$\begin{aligned} \pi_{i,k} &= [\pi_i]_k = \sum_{\ell \in V} P(\mathbf{n}(i-1) = \ell) p_{\ell k} \\ &= \sum_{\ell \in V} [\pi_{i-1}]_\ell [P]_{\ell k} = [\pi_{i-1} \mathcal{P}]_k \end{aligned} \quad (7)$$

for $k = 1, \dots, N$. Relation (7) yields

$$\pi_i = \pi_{i-1} \mathcal{P} = \pi_0 \mathcal{P}^i \quad (8)$$

where \mathcal{P}^i is the i^{th} power of \mathcal{P} and its $k\ell$ entry represents the probability of going from node k to node ℓ in exact i steps. The statistical evolution of a MC depends on the structure of \mathcal{P} , so that different chain classes are possible. An MC induced on a connected graph is *recurrent*, i.e., all states are visited infinitely often [11]. A recurrent chain can be either *periodic*, when the states can be grouped into subclasses periodically visited, or *aperiodic*, when only one entire class exists, with no periodicity (or the period is one). A *recurrent*

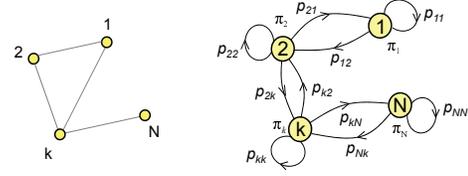


Fig. 2. Network topology (left) and the induced Markov chain (right).

aperiodic MC is *ergodic* [11], and has the convenient property of converging to a unique stationary chain distribution [12]

$$\pi \triangleq \lim_{i \rightarrow \infty} \pi_i \quad (9)$$

regardless of the initial distribution π_0 . Ultimately it means that, in steady-state, the probability of being at a specific state does not depend on the state the chain was at. By allowing self-loops in the protocol (a node can visit itself)¹, an ergodic MC is induced on the network [6], and a unique stationary distribution π is reached, given by the solution of the equation

$$\pi = \pi \mathcal{P} \quad (10)$$

that is, π is the left eigenvector of \mathcal{P} corresponding to the eigenvalue $\lambda = 1$. In fact, it can be shown that [11]

$$\lim_{i \rightarrow \infty} \mathcal{P}^i = q\pi \quad (11)$$

where $q = \text{col}\{1, \dots, 1\}$ is $N \times 1$. Furthermore, in the long run, the node k 's mean recurrence time t_k , i.e., given that the process is at node k , the expected time to return to it is given by [12]

$$t_k = \frac{1}{[\pi]_k} \quad (12)$$

From (12) we see that it would be interesting to have π close to a uniform distribution, in which case the probabilistic incremental protocol would behave, on average, just like the standard incremental mode: each node is visited once within N iterations. Therefore, we are interested in designing \mathcal{P} so that a convenient π is achieved. Each node has only access to its neighbors, so \mathcal{P} must be specified row-wise, i.e., locally at the nodes, via the pmf's defined over \mathcal{N}_k .

Now, consider two distributions over \mathcal{N}_k : uniform, where $[P]_{k\ell} = \frac{1}{|\mathcal{N}_k|}$, and Metropolis where

$$\begin{cases} p_{k\ell} = 1/\max(\deg_k, \deg_\ell) & \text{if } k \neq \ell \text{ are linked} \\ p_{k\ell} = 0 & \text{if } k \text{ and } \ell \text{ are not linked} \\ p_{kk} = 1 - \sum_{\ell \in \mathcal{N}_k/k} p_{k\ell} & \text{for } k = \ell \end{cases} \quad (13)$$

and \deg_n is node n 's degree. A stationary distribution π must be a solution to (10). Thus, if we can find a vector that is a solution to (10), then this vector is the unique stationary distribution [11]. If we employ the Metropolis rule (13), it is easy to check that the corresponding \mathcal{P} will be doubly stochastic, and q^T (see (11)) is its unique left eigenvector associated with the unit eigenvalue, that is $q^T \mathcal{P} = q^T$; therefore $(\frac{q^T}{N}) \mathcal{P} = 1 \cdot (q^T/N)$ and the chain reaches a uniform distribution $\pi = q^T/N$ over the possible N states. On the

¹If we remove self-loops, the spatial coverage may be accelerated [6], [11]. However, it may give rise to periodicity in the chain evolution.

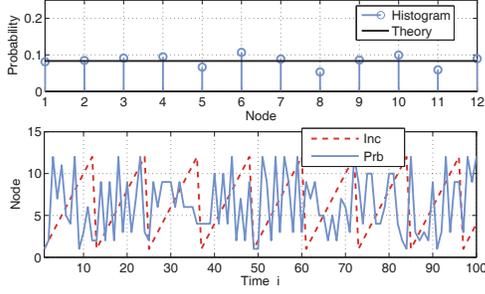


Fig. 3. MC's histogram (top) and node processes (bottom) for Metropolis rule.

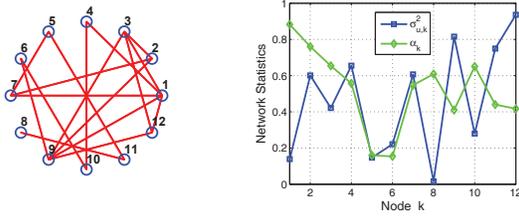


Fig. 4. Network topology and signals parameters.

other hand, if we select the local pmf uniformly distributed over \mathcal{N}_k , the chain stationary distribution can be shown to be [6]

$$[\pi]_k = \frac{\deg_k}{2|E| + N} \quad (14)$$

where $|E|$ is the number of links in the network. In other words, nodes with high degree will be privileged.

Figure 3 compares the probabilistic incremental protocol and its standard counterpart in terms of spatial coverage for a Metropolis rule². The top plot shows the statistics (theory and histogram) after 3000 iterations over the network topology depicted in the left plot of Fig. 4. The bottom plot of Fig. 3 shows a single cooperation trajectory.

4. INCREMENTAL LMS REVISITED

We embed an LMS learning rule at the nodes, and consider the original distributed incremental LMS algorithm [1], [4] except that the node assignment is now governed by the node indexing process $\{n(i)\}$. We skip the swift derivations and present just the algorithm's recursion

$$\psi_{n(i)} = \psi_{n(i-1)} + \mu_{n(i)} \mathbf{u}_{n(i)}^* (d_{n(i)} - \mathbf{u}_{n(i)} \psi_{n(i-1)}) \quad (15)$$

Note that the goal is to remove the Hamiltonian cycle while maintaining, if possible, mean-square performance. We use for evaluation the mean-square deviation (MSD), and the mean-square error (MSE), defined as follows

$$\text{MSD}(i) = E \|\mathbf{w}^\circ - \psi_{n(i-1)}\|^2 \quad (16)$$

$$\text{MSE}(i) = E \|\mathbf{d}_{n(i)} - \mathbf{u}_{n(i)} \psi_{n(i-1)}\|^2 \quad (17)$$

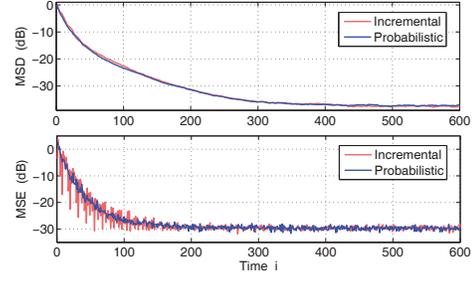


Fig. 5. Top: MSD. Bottom: MSE.

where the expectations are carried out over the learning process and the cooperation trajectory.

Figure 4 shows the settings for Example 1, where signals are Gaussian with regressor signal power and background noise power given by $\sigma_{u,k}^2$ and $\sigma_{v,k}^2 = 10^{-3}$, respectively, and correlation index given by α_k [1]. Both instances, incremental and probabilistic, have been simulated, and their mean-square convergence is nearly the same, as shown in Fig. 5. Note also the smoothing effect arising in the mean-square error (MSE) from the natural shuffling effect of the Markov sampling across the nodes. The probabilistic incremental protocol can be also extended to the aggregated gradient case [13].

5. STEADY-STATE ANALYSIS

Mean-square analysis for the proposed protocol is challenging. Here we provide a simplified version; it is an extension of the work in [1, 4, 5], with the added intricacy of the random cooperation protocol. We assume the data \mathbf{u}_k arise from a circular Gaussian distribution and define the following error quantities: $\tilde{\psi}_{n(i)} = \mathbf{w}^\circ - \psi_{n(i)}$ (weight error vector at time i), $\mathbf{e}_{a,n(i)} = \mathbf{u}_{n(i)} \tilde{\psi}_{n(i-1)}$ (a priori error), $\mathbf{e}_{p,n(i)} = \mathbf{u}_{n(i)} \tilde{\psi}_{n(i)}$ (a posteriori error), as well as the corresponding weighted errors $\mathbf{e}_{a,n(i)}^\Sigma = \mathbf{u}_{n(i)} \tilde{\psi}_{n(i-1)}^\Sigma$ and $\mathbf{e}_{p,n(i)}^\Sigma = \mathbf{u}_{n(i)} \tilde{\psi}_{n(i)}^\Sigma$, for some Hermitian matrix $\Sigma > 0$. By adopting the usual space-time data model $\mathbf{d}_k(i) = \mathbf{u}_{k,i} \mathbf{w}^\circ + \mathbf{v}_k(i)$ and manipulating the algorithm's rule (15), we arrive at the following space-time weighted energy conservation relation [1, 2, 14]:

$$\|\tilde{\psi}_{n(i)}\|_\Sigma^2 + \frac{|\mathbf{e}_{a,n(i)}^\Sigma|^2}{\|\mathbf{u}_{n(i)}\|_\Sigma^2} = \|\tilde{\psi}_{n(i-1)}\|_\Sigma^2 + \frac{|\mathbf{e}_{p,n(i)}^\Sigma|^2}{\|\mathbf{u}_{n(i)}\|_\Sigma^2} \quad (18)$$

Now, for some node k , let $R_{u,k} = E \mathbf{u}_{u,k} \mathbf{u}_k^* = T_k \Lambda_k T_k^*$, where T_k is unitary and Λ_k is diagonal and contains the eigenvalues of $R_{u,k}$. Defining the transformed error quantities $\bar{\psi}_{n(i)} = T_{n(i)}^* \tilde{\psi}_{n(i)}$, $\bar{\psi}_{n(i-1)} = T_{n(i)}^* \tilde{\psi}_{n(i-1)}$, $\bar{\mathbf{u}}_{n(i)} = \mathbf{u}_{n(i)} T_{n(i)}$, $\bar{\Sigma}_{n(i)} = T_{n(i)}^* \Sigma T_{n(i)}$, and via the diagonal notation [1, 14], we arrive at the following relation

$$E \|\bar{\psi}_{n(i)}\|_{\bar{\Sigma}_{n(i)}}^2 = E \|\bar{\psi}_{n(i-1)}\|_{\bar{\Sigma}_{n(i)}}^2 + g_{n(i)} \bar{\sigma}_{n(i)} \quad (19)$$

where $\bar{F}_{n(i)} = I - 2\mu_{n(i)} \Lambda_{n(i)} + \gamma \mu_{n(i)}^2 \Lambda_{n(i)}^2 + \mu_{n(i)}^2 \lambda_{n(i)} \lambda_{n(i)}^T$, with $\gamma = 1$ for complex data and $\gamma = 2$ for real data, $g_{n(i)} = \mu_{n(i)}^2 \sigma_{v,n(i)}^2 \lambda_{n(i)}^T$ is a row vector, with $\lambda_{n(i)} = \text{diag}\{\Lambda_{n(i)}\}$, and

²For comparison, we assume it is possible, e.g., via multihop, to establish a Hamiltonian cycle.

$\bar{\sigma}_{n(i)} = \text{diag}\{\bar{\Sigma}_{n(i)}\}$ [1, 14]. In steady-state, the chain reaches a stationary distribution π , and the indexing process approaches a random process ℓ , i.e., $n(i) \rightarrow \ell$ with $\ell \sim \pi$ (i.e., ℓ is distributed according to π). Due to the shuffling and smoothing effect of the protocol, we assume that, in steady-state, the estimates ψ_ℓ delivered by the nodes are statistically similar, so that the transformed weight error vector becomes $\bar{\psi}_\ell \rightarrow \bar{\psi}$. Taking the expectation over $\ell \in V$ in (19) results in:

$$\sum_{\ell} E\|\bar{\psi}\|_{\bar{\sigma}_\ell}^2[\pi]_\ell = \sum_{\ell} E\|\bar{\psi}\|_{\bar{F}_\ell \bar{\sigma}_\ell}^2[\pi]_\ell + \sum_{\ell} g_\ell \bar{\sigma}_\ell[\pi]_\ell \quad (20)$$

Using properties of weighted norms and diagonal notation [1, 14], and with $\bar{\sigma}_\ell \rightarrow \bar{\sigma}$, we arrive at

$$E\|\bar{\psi}\|_{(I - \sum_{\ell} \bar{F}_\ell[\pi]_\ell) \bar{\sigma}}^2 = \left(\sum_{\ell} g_\ell[\pi]_\ell \right) \bar{\sigma} \quad (21)$$

The MSD in steady-state can be calculated as $E\|\bar{\psi}\|_q^2$, where now $q = \text{col}\{1, \dots, 1\}$ is $M \times 1$ [1]. Solving for $\bar{\sigma}$ in $(I - \sum_{\ell} \bar{F}_\ell[\pi]_\ell) \bar{\sigma} = q$ and plugging the result back in (21), yields the network mean-square deviation in steady-state

$$\text{MSD} \approx \left(\sum_{\ell \in V} g_\ell[\pi]_\ell \right) \left(I - \sum_{\ell \in V} \bar{F}_\ell[\pi]_\ell \right)^{-1} q \quad (22)$$

The EMSE and MSE arise from similar procedures. Note that the resulting mean-square performance is a *weighted sum* of the data statistics in terms of the chain stationary distribution. If π is uniform, then the nodes' statistics are equally accounted for.

Figures 6 and 7 show the MSD across the nodes for the network settings of Fig. 4 and different stepsizes. Note that the adopted assumption is quite reasonable: different nodes achieve quite similar mean-square (deviation) performance for a range of stepsize values.

6. REMARKS AND FUTURE WORK

In the new distributed adaptive algorithm proposed, learning is governed by two intertwined stochastic processes: the adaptive algorithms run by observations captured at the node level, and the Markov process $n(i)$ conducting the cooperation.

Simulations corroborate that, in the mean-square sense, the Hamiltonian cycle may be relaxed. The new algorithm can be also viewed as a Hidden-Markov process, and energy conservation methods [14] were employed for mean-square deviation analysis. A complete mean-square performance analysis will be pursued in future work.

Future extensions include randomized incremental RLS, and the use of multiple tokens [6], where several random incremental processes take place concurrently. Moreover, not only nodes can learn, but also the cooperation process may be entitled with learning capabilities, which may be captured by the transition probabilities kept at each node. By doing so, some nodes may be probabilistically preferred due to their performance, which may improve the overall network performance (see (22)).

7. REFERENCES

- [1] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 4064-4077, August 2007.
- [2] A. H. Sayed and C. G. Lopes, "Adaptive processing over distributed networks," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E90-A, no. 8, pp. 1504-1510, 2007.

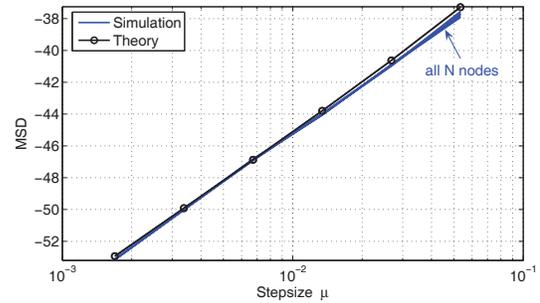


Fig. 6. MSD as a function of the stepsize for all nodes.

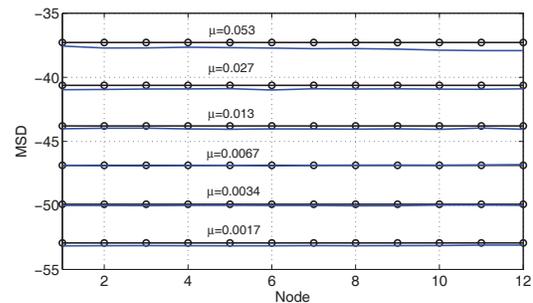


Fig. 7. MSD as a function of the nodes for decreasing stepsizes.

- [3] L. Li, J. Chambers, C. G. Lopes and A. H. Sayed, "Distributed estimation over an adaptive incremental network based on the affine projection algorithm," to appear in *IEEE Trans. Signal Processing*, 2009.
- [4] C. G. Lopes and A. H. Sayed, "Distributed adaptive incremental strategies: formulation and performance analysis," *Proc. ICASSP*, pp. 584-587, vol. 3, Toulouse, France, May 2006.
- [5] A. H. Sayed and C. G. Lopes, "Distributed recursive least-squares strategies over adaptive networks," *Proc. Asilomar Conference on Signals, Systems and Computers*, Monterey, CA, Oct. 2006, pp. 233-237.
- [6] L. Lovász "Random walks on graphs: A survey" *Combinatorics, Paul Erdos is Eighty*, Vol. 2, Keszthely, Hungary, 1993, pp. 1-46.
- [7] B. Johansson, M. Rabi and M. Johansson "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM J. Optim.* Vol. 20, Issue 3, pp. 1157-1170, 2009
- [8] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," to appear in *SIAM Journal on Optimization*, 2009.
- [9] C. G. Lopes and A. H. Sayed, "Diffusion least-mean-squares over adaptive networks," *Proc. ICASSP*, Honolulu, Hawaii, vol. 3, pp. 917-920, April 2007.
- [10] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, Vol. 52, No. 6, June 2006, pp. 2508-2530.
- [11] R. G. Gallager. *Discrete Stochastic Processes*, Kluwer Academic Publishers, Boston, 1996.
- [12] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Texts in Applied Mathematics, Springer-Verlag, New York, 1999.
- [13] D. Blatt, A. O. Hero and H. Gauchman "A convergent incremental gradient algorithm with a constant stepsize," *SIAM Journal on Optimization*, Vol. 18, No. 1, pp. 29-51, Feb. 2007.
- [14] A. H. Sayed. *Adaptive Filters*, Wiley, NJ, 2008.