# Convergence Rates for Cooperation in Heterogeneous Populations

Andrew Bean, Peter Kairouz, and Andrew Singer
University of Illinois at Urbana Champaign
Urbana, IL 61801, USA
Email: {ajbean, kairouz2, acsinger}@illinois.edu

***Abstract*—We consider the problem of cooperative distributed estimation within a network of heterogeneous agents. In particular, we study the situation where each agent observes an independent stream of Bernoulli random variables, and the goal is for each to determine its own Bernoulli parameter. The agents of this population can be categorized into a small number of subgroups, where within each group the agents all have identical Bernoulli parameters. For a distributed algorithm based on consensus strategies, we examine the rate at which the agent's estimates converge to the correct values. We show that the expected squared error decreases nearly as fast as centralized ML estimation in a homogeneous population. In a heterogeneous population, we derive an approximation to the expected squared error, as a function of the number of observations. Finally, we present simulation results that compare the predicted expected squared error to that observed in the simulations.**

***Index Terms*—gossip algorithms, consensus, diffusion, adaptation, distributed estimation, distributed signal processing**

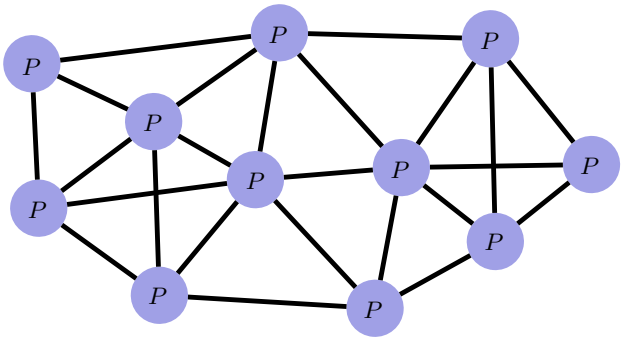


Fig. 1. A homogeneous population of Bernoulli agents.

## I. Introduction

The problem of distributed estimation within a network of agents has been extensively studied. This includes such topics as gossip algorithms [1]–[3], consensus [4]–[6], distributed adaptation and estimation [7]–[10], and others. Related to these is sequential learning or estimation, which includes least mean squares, recursive least squares, kalman filters [11], stochastic approximation [12], etc. In this paper, we contribute to these research areas by considering the problem of distributed estimation within a network of heterogeneous agents. Specifically, we consider populations of agents, each of which is trying to learn the parameters of a model for observed data, but these parameters are only consistent (i.e., the optimal model parameters are the same for all agents) within subpopulations of the whole. We extend the results of [13] by more precisely studying the convergence properties of the algorithm.

We first recall the framework for the problem first given in [13]. We consider a population of $N$ agents, indexed $i \in \{1, ..., N\}$. At each time instant $t \in \{1, 2, ...\}$, agent $i$ makes an observation $x_i(t) \in \{0, 1\}$ drawn according to a Bernoulli distribution with parameter $p_i$. The observations $x_i(t)$ are independent random variables for all $i$ and all $t$. Furthermore, we suppose that there is a partitioning of the population of agents into a number of subpopulations, i.e., $G_1 \cup ... \cup G_K = \{1, ..., N\}$, such that $p_i = P_j$ if and only if $i \in G_j$. We let $G(i)$ denote the subpopulation that agent $i$ belongs to. Lastly, the agents are connected to each other in a network given by adjacency matrix $A$, such that $A_{i,j} = 1$ if nodes $i$ and $j$ are connected, and zero otherwise. Typically, we have $A_{i,i} = 1$ for each agent $i$, and $A = A^T$. From this adjacency matrix, we can also determine the neighborhood $\mathcal{N}_i$ for each agent $i$. Since $A_{i,i} = 1$, we have that $i \in \mathcal{N}_i$.

In [7] and [8], the authors study the problem of distributed parameter estimation using a diffusion protocol for cooperation. In [9], the authors study the problem of distributed parameter estimation for linear state-space models. However, in these works it is assumed that the underlying model parameters $w_i^o$ to be estimated by each agent $i$ are identical, i.e., $w_i^o = w^o$ for all $i$. However, it is conceivable that the population of agents actually consists of a number of subgroups, such that the model parameters to be estimated are the same within a group, but different between different groups. The authors of [13] begin the study of this problem by proposing the heterogeneous framework and presenting a simple algorithm based on consensus strategies. In this paper, we extend the results of [13] by more precisely studying the convergence properties of the algorithm. To this end, we begin in Section II by looking closely at the convergence properties of Bernoulli parameter estimation in a homogeneous population, using a slight variation on the algorithm from [13]. In Section III, we use the results from Section II to approximate the convergence behavior of the algorithm for heterogeneous populations. In Section IV, we present simulation results that compare the predicted expected squared error to that observed in the simulations. Finally, we give some concluding remarks in Section V.

## II. Bernoulli Populations

We begin with the case of a homogeneous population of agents, where each agent observes IID Bernoulli random variables with the same parameter $P$. This situation is depicted in Fig. 1. Here, each agent makes one observation per time instance $t = \{1, 2, ...\}$. This is given by $x_i(t)$ for $i \in \{1, ..., N\}$ for $N$ agents in the network. The vector of observations at time $t$ consisting of the observations of each agent is given by $\mathbf{x}(t)$. The vector $\hat{\mathbf{p}}(t)$ is one consisting of the estimates for each agent, i.e., $[\hat{p}_i(t), ..., \hat{p}_N(t)]^T$. Furthermore, we will assume that the agents cooperate by mixing estimates according to a doubly stochastic, symmetric, irreducible matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ such that $\mathbf{D}\mathbf{1}^{N\times 1} = \mathbf{D}^T\mathbf{1}^{N\times 1} = \mathbf{1}^{N\times 1}$ and $0 \leq \mathbf{D}_{i,j} \leq 1$, i.e., each entry in $\mathbf{D}$ is in the range $[0, 1]$. Finally, since the agents are only able to communicate with each other over the network edges, we have that $\mathbf{D}_{i,j} = 0$ if $A_{i,j} = 0$. Then, we have that the cooperative

algorithm is given by

$$\hat{\mathbf{p}}(t) = \mathbf{D}\left(\frac{t-1}{t}\hat{\mathbf{p}}(t-1) + \frac{1}{t}\mathbf{x}(t)\right), \tag{1}$$

i.e. the update of $\hat{\mathbf{p}}(t)$ involves the incorporation of the new data followed by a diffusion step.

In [13], the authors show that the estimates of all the agents converge to $P$ (in probability). However, there was no discussion of whether the rate of convergence is better than, e.g., noncooperative estimation or how the rate compares to a centralized maximum likelihood estimate. We will now provide results relating to these issues.

First, suppose $E[\hat{\mathbf{p}}(t-1)] = P\mathbf{1}^{N\times 1}$. Then

$$\begin{aligned} E[\hat{\mathbf{p}}(t)] &= E\left[\mathbf{D}\left(\frac{t-1}{t}\hat{\mathbf{p}}(t-1) + \frac{1}{t}\mathbf{x}(t)\right)\right] \\ &= \mathbf{D}\frac{t-1}{t}E[\hat{\mathbf{p}}(t-1)] + \mathbf{D}\frac{1}{t}E[\mathbf{x}(t)] \\ &= P\mathbf{1}^{N\times 1}. \end{aligned}$$

Furthermore, note that

$$\begin{aligned} E[\hat{\mathbf{p}}(1)] &= E[\mathbf{D}\mathbf{x}(1)] \\ &= \mathbf{D}E[\mathbf{x}(1)] \\ &= \mathbf{D}P\mathbf{1}^{N\times 1} \\ &= P\mathbf{1}^{N\times 1}. \end{aligned}$$

Thus, by induction, we have that $E[\hat{p}_i(t)] = P$ for each agent $i$.

We will now consider the variance of $\hat{p}_i(t)$. Since $E[\hat{p}_i(t)] = P$, this variance is equal to the expected squared estimation error. Let $\mathbf{D}$ have an eigenvalue decomposition such that $\mathbf{D} = U\Sigma U^T$, where the columns of $U$ are orthonormal eigenvectors and $\Sigma$ is a diagonal matrix consisting of decreasing eigenvalues $1 = |\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_N|$. (This is possible since we have assumed that $\mathbf{D}$ is a real symmetric doubly stochastic irreducible matrix.)

It is then possible to show that for some positive constant $C$, $\mathbf{D}^t_{i,j} \leq \frac{1}{N} + C|\lambda_2|^t$ for all $i$ and $j$, where the notation $\mathbf{D}^t_{i,j}$ indicates the element of matrix $\mathbf{D}^t$ at row $i$ and column $j$.

To consider $\text{var}[\hat{p}_i(t)]$, we note that

$$\hat{\mathbf{p}}(t) = \frac{1}{t}\sum_{j=1}^{t}\mathbf{D}^{t-j+1}\mathbf{x}(j). \tag{2}$$

Therefore, we have that

$$\hat{p}_i(t) = \frac{1}{t}\sum_{j=1}^{t}\mathbf{d}_i^{(t-j+1)}\mathbf{x}(j), \tag{3}$$

where $\mathbf{d}_i^{(t-j_1)}$ is the $i^{\text{th}}$ row of the matrix $\mathbf{D}^{t-j+1}$. We can then conclude that

$$\begin{aligned} &\text{var}[\hat{p}_i(t)] \\ &= \text{var}\left[\frac{1}{t}\sum_{j=1}^{t}\mathbf{d}_i^{(t-j+1)}\mathbf{x}(j)\right] \\ &= \frac{1}{t^2}\text{var}\left[\sum_{j=1}^{t}\mathbf{d}_i^{(t-j+1)}\mathbf{x}(j)\right] \\ &= \frac{1}{t^2}\sum_{j=1}^{t}\text{var}\left[\mathbf{d}_i^{(t-j+1)}\mathbf{x}(j)\right] \\ &= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{k=1}^{N}\text{var}\left[\left[\mathbf{d}_i^{(t-j+1)}\right]_k x_k(j)\right] \\ &= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{k=1}^{N}\left[\mathbf{d}_i^{(t-j+1)}\right]_k^2 \text{var}[x_k(j)] \\ &= \frac{P(1-P)}{t^2}\sum_{j=1}^{t}\sum_{k=1}^{N}\left[\mathbf{d}_i^{(t-j+1)}\right]_k^2 \\ &\leq \frac{P(1-P)}{t^2}\sum_{j=1}^{t}\sum_{k=1}^{N}\left(\frac{1}{N} + C|\lambda_2|^{t-j+1}\right)^2 \\ &= \frac{P(1-P)}{Nt} + \frac{P(1-P)}{t^2}\sum_{j=1}^{t}\left(2C|\lambda_2|^j + NC^2|\lambda_2|^{2j}\right) \\ &\leq \frac{P(1-P)}{Nt} + \frac{P(1-P)}{t^2}\sum_{j=1}^{\infty}\left(2C|\lambda_2|^j + NC^2|\lambda_2|^{2j}\right) \\ &= \frac{P(1-P)}{Nt} + \frac{P(1-P)}{t^2}\left(\frac{2C|\lambda_2|}{(1-|\lambda_2|)} + \frac{NC^2|\lambda_2|^2}{(1-|\lambda_2|^2)}\right). \end{aligned}$$

In particular, we note that the rate is dominated by $\frac{P(1-P)}{Nt}$. This is important because $\frac{P(1-P)}{Nt}$ is the expected squared error for a centralized maximum likelihood estimate of the Bernoulli parameter, based on all of the observations from the $N$ agents over $t$ time instants. In other words, the distributed cooperative estimation suffers a small, asymptotically negligible regret with respect to centralized estimation.

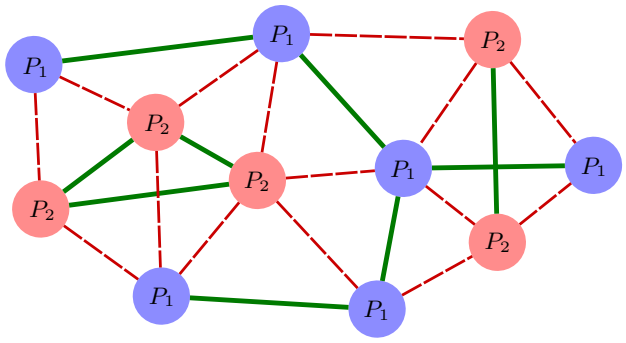


Fig. 2. A heterogeneous population of Bernoulli agents. Cooperation over the solid green edges is helpful. Cooperation over the dashed red edges is detrimental.

## III. Heterogeneous Bernoulli Populations

We will now consider heterogeneous populations, i.e., the situation where there are various subgroups observing different types of sources, as shown in Fig. 2. The algorithm used in this setting involves having each agent compute an estimate of the parameter $p_i$ based on only it's own observations. This will be written as $p_i^\ell(t)$ and we will call it the *private estimate*. In particular, this is taken to be

$$p_i^\ell(t) = \begin{cases} \frac{1}{2} & \text{if } t = 0 \\ \frac{1}{t}\sum_{\tau=1}^{t} x_i(\tau) & \text{if } t > 0. \end{cases}$$

We will then choose the elements of $\mathbf{D}(t)$ as follows: First, each agent will decide which neighboring agents it will take messages from by comparing its own private estimate to those of its neighbors. In particular, agent $i$ will take a message from neighbor $j \in \mathcal{N}_i$ if $|p_i^\ell - p_j^\ell| \geq \gamma_t$, where $\gamma_t$ is a threshold for cooperation between agents. Hence, agent $i$ will take messages from $|\tilde{\mathcal{N}}_i(t)|$ neighbors, where $\tilde{\mathcal{N}}_i(t)$ is the subset of neighbors that agent $i$ will cooperate with during time $t$. Once this has been determined, the diffusion

weights can be determined as follows:

$$\mathbf{D}_{i,j}(t) = \begin{cases} \frac{1}{\max\{|\hat{\mathcal{N}}_i(t)|,|\hat{\mathcal{N}}_j(t)|\}} & \text{if } \begin{cases} |p_i^\ell - p_j^\ell| < \gamma_t \\ i \neq j \\ \text{and } A_{i,j} = 1 \end{cases} \\ 0 & \text{if } \begin{cases} |p_i^\ell - p_j^\ell| \geq \gamma_t \\ \text{or } A_{i,j} = 0 \end{cases} \\ 1 - \sum_{k \neq i} \mathbf{D}_{i,k}(t) & \text{if } i = j, \end{cases}$$

Therefore, $\mathbf{D}(t)$ is essentially a time varying Metropolis weight matrix, as in [14]. It it possible to show that if we choose $\gamma_t = Ct^\delta$ for some positive constant $C$ and $-\frac{1}{2} < \delta < 0$, the subpopulations will be correctly differentiated and each agent's estimate $\hat{p}_i(t)$ will converge to the true parameter of the model of its observations.

We will now study the convergence properties of this distributed algorithm. To this end, we will attempt to approximate the expected squared error. In particular, suppose that we choose $\gamma_t = Ct^\delta$ such that $\gamma_1$ is large, causing all of the agents to initially collaborate with their neighbors. What will happen is that the estimates of all of the agents will converge to the neighborhood of the global mean parameter, and the expected squared error will remain approximately constant for some time. At some point, the subpopulations will disconnect from each other, as a result of the private estimates improving and the collaboration radius $\gamma_t$ becoming more selective (smaller). We will call this the *time to disconnect* and represent it by $t^*$. After the time $t^*$, the subpopulations quickly disconnect from each other, and the expected squared error gradually converges to that of the centralized maximum likelihood estimate within connected subsets of the subpopulations.

To approximate the time to disconnect, we will use basic methods from large deviations theory. In particular, we would like to approximate the time when an edge between agents of different subpopulations (a "bad link") has a low probability of being used for collaboration. Consider a scenario with two subpopulations, with parameters $P_1$ and $P_2 > P_1$. The probability of collaboration on the bad edge between connected agents $i$ with $P_1$ and $j$ with $P_2$ is $P[|p_i^\ell - p_j^\ell| < \gamma_t]$, which can be approximated using large deviations. Specifically, we note that $p_i^\ell - p_j^\ell = \frac{1}{t}\sum_\tau (x_i(\tau) - x_j(\tau))$. The result is that

$$P[\text{bad connection}] \approx e^{-tI(-\gamma_t)},$$

where $I(p)$ is the large deviations rate function, given by

$$I(p) = p\theta(p) - \ln\left(ae^{-\theta(p)} + b + ce^{\theta(p)}\right).$$

For shorthand, we will define $I(t)$ as $I(-\gamma_t)$. Here, we have that $a = (1-P_1)P_2$, $b = P_1P_2 + (1-P_1)(1-P_2)$, $c = P_1(1-P_2)$, and

$$e^{\theta(p)} = \frac{bp + \sqrt{b^2p^2 + 4ac(1-p^2)}}{2c(1-p)}$$

Since the natural scales for studying features of this convergence are logarithmic in time and magnitude, we will convert $e^{-tI(-\gamma_t)}$ to such a scale. This gives us $\tilde{p}(\tilde{t}) \approx -e^{\tilde{t}}I(e^{\tilde{t}})$, where $\tilde{p}(\tilde{t}) = \ln(P[\text{bad connection}])$ and $\tilde{t} = \ln(t)$. The time when $\tilde{p}(\tilde{t})$ begins to rapidly decrease is approximately when $\frac{\mathrm{d}}{\mathrm{d}\tilde{t}}\tilde{p}(\tilde{t}) = -1$. This occurs approximately when $tI(t) = 1$, and this can be found numerically. Hence, we choose $t^*$ such that $t^*I(\gamma_{t^*}) = 1$ and $\gamma_{t^*} < |P_1 - P_2|$. At this point, we will simply note that this does not take into account the number of edges that connect agents of different groups. Many edges should increase the time to disconnect, so the estimate of $t^*$ presented here should be somewhat too early.

To approximate the convergence behavior after the disconnect time, we assume that the estimates converged to the global average of the parameters. We will also assume that the subgroups have completely disconnected from each other, the agents within subgroups are connected by some path, and that the mixing time is instantaneous. Then the expected squared error within the group associated with $P_1$ after $t^*$ is approximately given by

$$E\left[\left(\frac{t^*N(P_1+d) + \sum_{\tau=t^*+1}^{t}\sum_{i=1}^{N} x_i(\tau)}{Nt} - P_1\right)^2\right],$$

where $d = \frac{P_1+P_2}{2} - P_1$ is the estimated error right before the subgroups disconnect and the agents $1,...,N$ belong to subgroup $P_1$. It can be shown that this leads us to

$$E[(\hat{p}_i(t) - P_1)^2] \approx \frac{P_1(1-P_1)}{Nt} + \frac{d^2t^{*2} - \frac{1}{N}t^*P_1(1-P_1)}{t^2}.$$

Again, as in the homogeneous case, we see that the convergence is dominated by $\frac{P_1(1-P_1)}{Nt}$, and therefore the convergence rate is nearly as good as centralized maximum likelihood within the subpopulation.

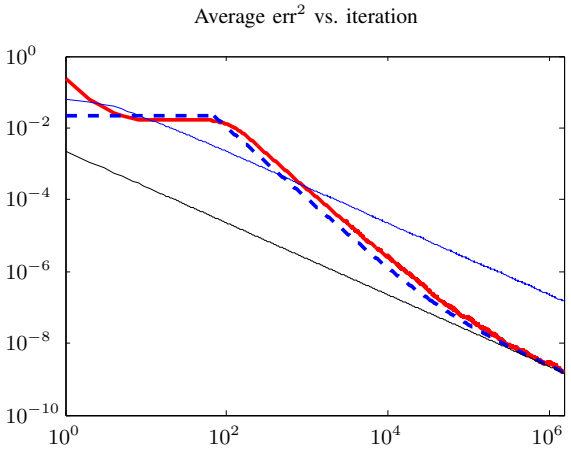


Fig. 3. Simulation with two Bernoulli subpopulations. The upper thin line shows the simulated noncooperative average squared error. The lower thin line shows the predicted centralized maximum likelihood squared error. The thick solid line shows the simulated cooperative average squared error in a heterogeneous population. The thick dashed line shows the estimated average squared error for the estimation algorithm for heterogeneous populations.

## IV. Simulations

To evaluate the quality of our approximation to the expected squared error of the heterogeneous cooperative algorithm, we randomly placed 200 agents within a 1 unit by 1 unit square. We formed a network connection if two agents were within 0.25 unit of each other. We used two subpopulations: one with a Bernoulli parameter $P_1 = 0.35$, and the other with parameter $P_2 = 0.65$. The cooperation radius is given by $Ct^\delta = t^{-0.4}$. Figure 3 shows results from this simulation. In this plot, the upper thin line shows the simulated noncooperative average squared error. As expected, this decreases like $\frac{0.35\times0.65}{t}$. The lower thin line shows the predicted centralized maximum likelihood squared error, which decreases like $\frac{0.35\times0.65}{100t}$, since there are 100 agents in each subpopulation. It should not be possible to do better than this lower thin line. The thick solid line shows the simulated cooperative average squared error in the heterogeneous population. It can be observed that the estimates indeed converge to a particular squared error and stay here until a certain point. After this point, the average squared error begins to decrease, eventually coming very close to the predicted centralized maximum likelihood squared error curve. The thick dashed line shows

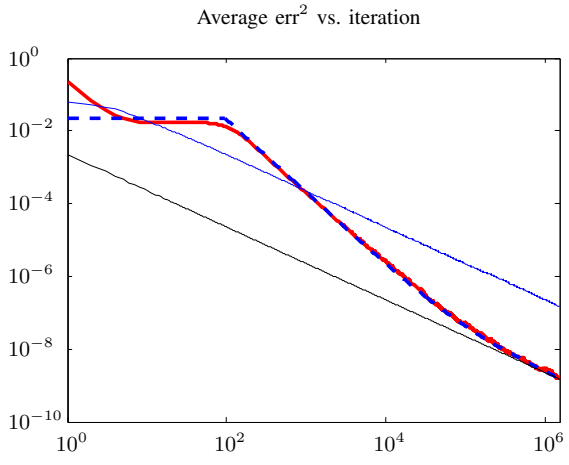


Fig. 4. Simulation with two Bernoulli subpopulations. The thin curves and the thick solid curve are the same as in Fig. 3. The thick dashed curve uses the value of $t^*$ that fits the downward trend of the average squared error.

the estimated average squared error for the estimation algorithm. We can see that the estimated disconnect time in this instance was $t^* \approx 69$. As suggested earlier, the approximation method for $t^*$ gives a value that is a bit early compared to what the simulations indicate. Determining a better method for approximating $t^*$ could be a point for future study. In this case, fitting the downward portion of the simulated curve to our predicted trend indicates that $t^* \approx 96$. This fit is shown in Fig. 4.

## V. Conclusion

In this paper, we considered convergence rates for the problem of cooperative distributed estimation within a network of heterogeneous agents. First, we studied homogeneous populations of Bernoulli agents, and demonstrated that such a population can achieve a convergence rate that is nearly as good as centralized maximum likelihood parameter estimation. We then considered the case of heterogeneous populations, and derived an approximation to the expected squared error. Finally, we presented simulation results that compared the approximated expected squared error to that observed in the simulations.

There are many directions that could be looked at from here. For example, we could consider the consequence of knowing the number of subpopulations or knowing the minimum distance between the underlying optimal subpopulation parameter values. For a more adaptive algorithm, rather than asymptotic, we could consider an algorithm with fixed, rather than decreasing, step size, in order to accomodate time varying underlying model parameters. It would also be interesting to consider the types of messages that would be sent over communication links. In our case, we assume that both cooperative and noncooperative infinite precision estimates are sent to neighbors over the links, but restricting this communication to only the cooperative estimate could be considered, or we may even consider sending quantized messages, such as resampled symbols as is done in [15].

## References


[1] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, November 2010.

[2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006.

[3] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, vol. 5, December 2003, pp. 4997–5002.

[4] M. H. Degroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, March 1974.

[5] J. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Institute of Technology, November 1984.

[6] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. AC-31, no. 9, pp. 803–812, September 1986.

[7] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3166, July 2008.

[8] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[9] S. Ram, V. V. Veeravalli, and A. Nedic, "Distributed and recursive parameter estimation in parametrized linear state-space models," *IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 488–492, February 2010.

[10] S. Kirti and A. Scaglione, "Scalable distributed kalman filtering through consensus," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2008, pp. 2725–2728.

[11] A. H. Sayed, *Fundamentals of Adaptive Filtering*. Hoboken, NJ: Wiley-Interscience, 2003.

[12] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[13] A. Bean and A. Singer, "Cooperative estimation in heterogeneous populations," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, November 2011, pp. 696–699.

[14] L. Xiao, S. Boyd, and S. Lall, "Distributed average consensus with time-varying metropolis weights," unpublished.

[15] A. D. Sarwate and T. Javidi, "Opinion dynamics and distributed learning of distributions," in *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computation*, September 2011.