ATTAINING OPTIMAL BATCH PERFORMANCE VIA DISTRIBUTED PROCESSING OVER NETWORKS

Xiaochuan Zhao and Ali H. Sayed

Department of Electrical Engineering
University of California, Los Angeles, CA 90095

ABSTRACT

This work shows how the combination weights of diffusion strategies for adaptation and learning over networks can be chosen in order for the network mean-square-error performance to match that of an optimized centralized (or batch) solution. The results show that this is possible regardless of the network topology, however sparse it is, as long as the network is connected without disjoint sub-graphs.

Index Terms— Diffusion adaptation, centralized processing, batch processing, Hastings rule, MRC rule.

1. INTRODUCTION AND RELATION TO PRIOR WORK

Three major classes of distributed strategies for estimation over networks have been proposed in the literature: incremental strategy [1–7], consensus strategy [8–12], and diffusion strategy [13–17]. Incremental strategies require the establishment of a cyclic path that visits all agents in the network sequentially and are therefore prone to link and node failure. On the other hand, diffusion strategies rely on in-network processing and local consultation, and were shown in [18] to have superior mean-square-error (MSE) performance when compared to consensus strategies. The stability of diffusion networks was further shown in [18] to be insensitive to the network topology, while the stability of consensus networks is sensitive to the topology and they can become unstable even when all the individual nodes are stable. For these reasons, we continue our discussion by focusing on the class of diffusion strategies.

In previous work [19], we established that the MSE performance of diffusion strategies for adaptation and learning over networks is dependent on the network topology in an interesting manner. Specifically, we showed that, for sufficiently small step-sizes, the right eigenvector of the combination matrix that corresponds to the eigenvalue at one determines the network performance in lieu of the other eigenvectors. By optimizing over left-stochastic combination matrices, we were then able to show that the MSE performance of adaptive diffusion networks can attain their lowest (i.e., best) value by selecting the combination weights according to the Hastings rule. In this work, we move further to explain how this result implies that diffusion over arbitrary connected networks can be made to achieve the same level of performance as provided by clique networks and by centralized or batch processing. The interesting point to note is that these results will be seen to be *independent* of the network topology. As long as the network is connected so that it does not exhibit disjoint sub-graphs, and as long as the combination weights are adjusted according to the Hastings rule, the diffusion networks can deliver centralized performance no matter how sparse they are.

This work was supported by NSF grant CCF-1011918. Emails: {xzhao, sayed}@ee.ucla.edu.

Notation: We use lowercase letters to denote vectors, uppercase letters for matrices, plain letters for deterministic variables, and boldface letters for random variables. We also use $(\cdot)^*$ to denote conjugate transposition, $(\cdot)^{-1}$ for matrix inversion, $\mathrm{Tr}(\cdot)$ for the trace of a matrix, and $\rho(\cdot)$ for the spectral radius of a matrix. All vectors in our treatment are column vectors, with the exception of the regression vectors, $\boldsymbol{u}_{k,i}$, which are taken to be row vectors for convenience of presentation.

2. DATA MODEL

Consider a network consisting of N agents with an arbitrarily connected topology, as shown in Fig.1a. Each agent k is assumed to sense data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ that satisfy the linear regression model [20]:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w^o + \boldsymbol{v}_k(i), \quad k = 1, \ldots, N \tag{1}$$

where $w^o \in \mathbb{C}^{M\times 1}$ is an $M \times 1$ unknown vector, $\boldsymbol{u}_{k,i} \in \mathbb{C}^{1\times M}$ is a regression vector at time i, and $\boldsymbol{v}_k(i) \in \mathbb{C}$ is a noise signal also at time i. The agents are interested in estimating w^o cooperatively through local in-network processing. We adopt the following assumptions on the statistical properties of the data.

Assumption 1 (Statistical properties of data)

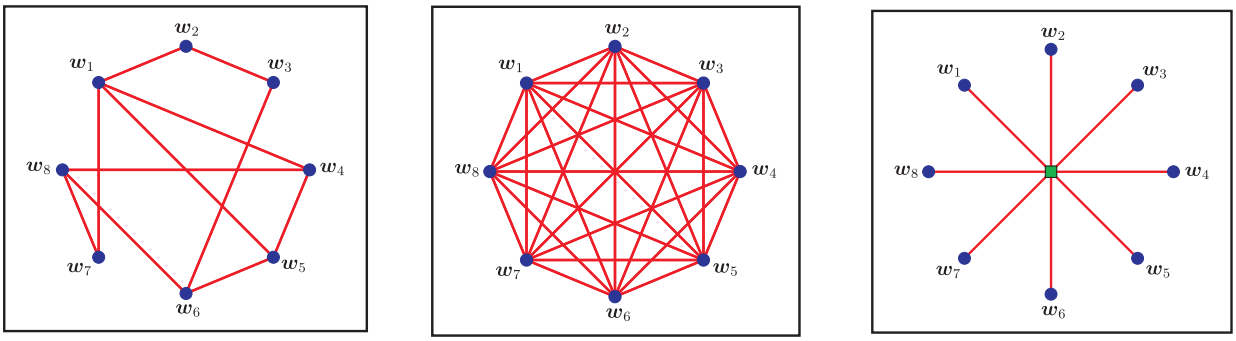
1. *The regression data $\boldsymbol{u}_{k,i}$ are temporally white and spatially independent random variables with zero mean and uniform covariance matrix $R_u \triangleq \mathbb{E}\boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i} > 0$.*
2. *The noise signals $\boldsymbol{v}_k(i)$ are temporally white and spatially independent random variables with zero mean and variances $\sigma_{v,k}^2$.*
3. *The regressors $\boldsymbol{u}_{k,i}$ and the noise signals $\boldsymbol{v}_\ell(j)$ are mutually-independent for all k and ℓ, i and j.* ■

We also adopt the following assumption, which is common in the literature of stochastic approximation algorithms (see [19]).

Assumption 2 (Small step-sizes) *The step-size used in the adaptation strategies is sufficiently small, i.e., $\mu \ll 1$, such that terms depending on higher-order powers of μ can be ignored and the strategies are mean-square stable whenever necessary.* ■

3. DIFFUSION STRATEGIES

Diffusion strategies constitute a powerful class of distributed mechanisms for estimating w^o cooperatively [14, 15, 17]. For each agent, an adaptation step is performed to incorporate local data and a consultation step is used to aggregate information form the neighbors.



(a) An arbitrarily connected topology. (b) A clique topology. (c) A star topology.

Fig. 1. Three different topologies for adaptive networks.

According to the order of these two steps, two variations are possible [15]: adapt-then-combine (ATC) and combine-then-adapt (CTA). It has been shown that ATC in general has superior MSE performance than CTA [15, 19], so we shall focus on the ATC diffusion strategy in this work even though the conclusions apply equally well to CTA.

For ATC diffusion, the kth agent performs the following operations at each time instant i:

$$\begin{cases} \boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} + \mu \boldsymbol{u}_{k,i}^{*}(\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k}\boldsymbol{\psi}_{\ell,i} \end{cases} \tag{2}$$

where $\mu > 0$ denotes the step-size parameter, $\mathcal{N}_k$ denotes the neighborhood of agent k, and $a_{\ell k}$ is the nonnegative combination weight that agent k assigns to the data from agent $\ell \in \mathcal{N}_k$. A sum constraint is imposed on the combination weights $\{a_{\ell k}\}$:

$$\sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \qquad k = 1, \ldots, N \tag{3}$$

All weights $\{a_{\ell k}\}$ are collected into a matrix $A \in \mathbb{R}^{N \times N}$ such that the (ℓ, k)th entry of A is $a_{\ell k}$. Moreover, $a_{\ell k} = 0$ wherever $\ell \notin \mathcal{N}_k$. It follows from (3) that A is a left-stochastic matrix.

The choice of A affects the performance of the distributed strategy. Different combination rules, such as uniform, Laplacian, maximum degree, Metropolis, relative degree, relative degree-variance, and relative variance have been proposed in the literature [15, 17]. The following analysis explains how the choice of A influences performance.

We assume that the network topology is *standard* (also called strongly-connected), meaning that it is connected and contains at least one self-loop (i.e., $a_{kk} > 0$ some k). Let us introduce the weight error vector:

$$\widetilde{\boldsymbol{w}}_{k,i} \triangleq w^{o} - \boldsymbol{w}_{k,i} \tag{4}$$

and the noise variance matrix profile R_v:

$$R_v \triangleq \mathrm{diag}\{\sigma_{v,1}^2, \ldots, \sigma_{v,N}^2\} \tag{5}$$

In [19], we showed that under Assumptions 1 and 2 and for standard networks, the network mean-square-deviation (MSD) of the diffusion strategy (2) is given by

$$\begin{aligned} \mathrm{MSD}_{\mathrm{diff}}^{\mathrm{network}} &\triangleq \lim_{i\to\infty} \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \\ &= \frac{\mu M}{2} p^{\mathsf{T}} R_v p + O(\mu^2) \end{aligned} \tag{6}$$

and the network excess mean-square-error (EMSE) is given by

$$\begin{aligned} \mathrm{EMSE}_{\mathrm{diff}}^{\mathrm{network}} &\triangleq \lim_{i\to\infty} \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|_{R_u}^2 \\ &= \frac{\mu \mathrm{Tr}(R_u)}{2} p^{\mathsf{T}} R_v p + O(\mu^2) \end{aligned} \tag{7}$$

where p is the right eigenvector of A corresponding to the eigenvalue at one and whose entries $\{p_k\}$ are normalized to add up to one:

$$Ap = p, \qquad p^{\mathsf{T}}\mathbb{1} = 1, \qquad p_k > 0 \tag{8}$$

where $\mathbb{1}$ is the $N \times 1$ vector with all entries equal to one. We established in [19] the following result, which provides one choice for the combination matrix A to minimize (6) and (7).

Theorem 1 (Optimal mean-square performance for diffusion) *Under Assumptions 1 and 2, one combination rule that minimizes the MSD and EMSE performance for standard networks is given by the Hastings rule:*

$$a_{\ell k} = \begin{cases} \dfrac{\sigma_{v,k}^2}{\max\{|\mathcal{N}_k|\sigma_{v,k}^2, |\mathcal{N}_\ell|\sigma_{v,\ell}^2\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}, & \ell = k \end{cases} \tag{9}$$

where $|\mathcal{N}_k|$ denotes the cardinality of the neighborhood $\mathcal{N}_k$. Moreover, the minimum MSD and EMSE performance levels attained by the network are given by

$$\mathit{MSD}_{\mathit{diff,opt}}^{\mathit{network}} = \frac{\mu M}{2} \frac{1}{\mathrm{Tr}(R_v^{-1})} + O(\mu^2) \tag{10}$$

$$\mathit{EMSE}_{\mathit{diff,opt}}^{\mathit{network}} = \frac{\mu \mathrm{Tr}(R_u)}{2} \frac{1}{\mathrm{Tr}(R_v^{-1})} + O(\mu^2) \tag{11}$$

and the dominant mode of mean-square convergence is

$$\mathit{mode}_{\mathit{diff}} = 1 - 2\mu\lambda_{\min}(R_u) + O(\mu^2) \tag{12}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of its positive-definite argument.

Proof See [19]. ■

An interesting conclusion that follows readily from this theorem is that the optimal attainable MSE performance and the corresponding convergence rate for *standard* networks are both *independent* of the network topology.

Corollary 1 (Insensitivity to topology) *Regardless of the topology linking the N agents, as long as the network is standard, the optimal MSE performance, e.g., MSD and EMSE, and convergence rate that it can achieve are given by the same values* (10)–(12) *up to variations of the order of μ^2.*

Although the optimal mean-square performance does not change with the topology, the Hastings rule (9), which defines the combination rule, does depend on the topology and will change accordingly.

3.1. Clique Networks

Another useful conclusion follows from Theorem 1.

Corollary 2 (Comparison to clique networks) *Using the Hastings rule* (9)*, diffusion strategies over arbitrary N-agent standard networks can achieve the same MSE performance and convergence rate as N-agent fully-connected networks, i.e., clique networks (see Fig.1b). The differences in performance are of the order of μ^2.*

Clique networks can also achieve the same optimal performance levels (10) – (11) by using simpler combination rules than the Hastings rule (9). For instance, one possible choice of A for clique networks is the rank-one matrix:

$$A_{\text{clique}} \triangleq p^o \mathbb{1}^{\mathsf{T}} \tag{13}$$

where

$$p^o \triangleq \frac{R_v^{-1}\mathbb{1}}{\mathbb{1}^{\mathsf{T}} R_v^{-1}\mathbb{1}} \tag{14}$$

It can be verified that (13) minimizes (6) and achieves the same MSD performance (10). The choice (13) and (14) has a useful interpretation. It means that each agent uses a maximal ratio combination (MRC) rule [20] since for any $\ell \in \{1, 2, \ldots, N\}$,

$$a_{\ell 1} = a_{\ell 2} = \cdots = a_{\ell N} = \frac{\sigma_{v,\ell}^{-2}}{\sum_{k=1}^{N} \sigma_{v,k}^{-2}} \triangleq a_\ell \tag{15}$$

Result (15) means that the information emanating from every agent ℓ is weighted by the same amount a_ℓ before it reaches any of the other agents in the clique network. The weight is inversely proportional to the noise variance at the source agent ℓ, meaning that data emanating from noisy agents are weighted less heavily. Substituting (15) into the combination step in the diffusion strategy (2) leads to

$$\boldsymbol{w}_{1,i} = \boldsymbol{w}_{2,i} = \cdots = \boldsymbol{w}_{N,i} = \sum_{\ell=1}^{N} a_\ell \boldsymbol{\psi}_{\ell,i} \triangleq \boldsymbol{w}_i \tag{16}$$

Substituting (16) into (2) reduces the diffusion strategy to the following batch algorithm:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu \sum_{\ell=1}^{N} a_\ell \boldsymbol{u}_{\ell,i}^{*} (\boldsymbol{d}_\ell(i) - \boldsymbol{u}_{\ell,i}\boldsymbol{w}_{i-1}) \tag{17}$$

Expression (17) has the form of a centralized solution, as expected. In a clique network, every agent has access to the data from across the entire network and therefore every agent can run the batch solution (17). By examining (17), we find that every agent computes a weighted combination of the instantaneous approximate gradients from across the network (the terms appearing inside the sum and weighted by a_ℓ). This implementation is conceptually equivalent to a data fusion center sitting at the hub of a network with a star topology as shown in Fig.1c. We refer to the adaptation rule (17) as a *weighted centralized LMS* solution.

4. CENTRALIZED (BATCH) LMS

The argument that led to the centralized LMS implementation (17) shows that diffusion strategies over arbitrary standard networks using (9) achieve the same optimal mean-square performance as (17) using the optimal coefficients (15). In other words, the distributed diffusion strategy (2) can match the best performance that one can extract from the batch solution (17); this best performance occurs when the MRC coefficients (15) are used.

More generally, it is instructive to evaluate the performance of the batch LMS solution (17) for arbitrary nonnegative weights $\{a_\ell\}$ that are not necessarily chosen according to (15) but are still required to satisfy

$$\sum_{\ell=1}^{N} a_\ell = 1, \qquad \ell = 1, 2, \ldots, N \tag{18}$$

Define the steady-state MSD and EMSE for the centralized (or batch) LMS solution (17) as

$$\text{MSD}_{\text{batch}} \triangleq \lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \tag{19}$$

$$\text{EMSE}_{\text{batch}} \triangleq \lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{R_u}^2 \tag{20}$$

and denote the jth eigenvalue of R_u by λ_j.

Theorem 2 (Mean-square performance for batch LMS) *Under Assumptions 1 and 2, the MSD and EMSE performance of the centralized (or batch) LMS implementation* (17) *for arbitrary nonnegative coefficients $\{a_\ell\}$ satisfying* (18) *are given by*

$$\textit{MSD}_{\textit{batch}} = \kappa_1 \cdot a^{\mathsf{T}} R_v a \tag{21}$$

$$\textit{EMSE}_{\textit{batch}} = \kappa_2 \cdot a^{\mathsf{T}} R_v a \tag{22}$$

where $a \triangleq \text{col}\{a_1, \ldots, a_N\}$ and

$$\kappa_1 \triangleq \frac{\sum_{j=1}^{M} \frac{\mu}{2-\mu\lambda_j}}{1 - \|a\|^2 \sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j}} \tag{23}$$

$$\kappa_2 \triangleq \frac{\sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j}}{1 - \|a\|^2 \sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j}} \tag{24}$$

The dominant mode of mean-square convergence is given by

$$\textit{mode}_{\textit{batch}} \triangleq \rho\left((I - \mu\Lambda)^2 + \mu^2\|a\|^2 \lambda\lambda^{\mathsf{T}}\right) \tag{25}$$

where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_M\}$ and $\lambda = \text{col}\{\lambda_1, \ldots, \lambda_M\}$.

Proof Omitted due to space limitations. ■

It is easy to verify that $0 < \|a\|^2 \leq 1$. Under Assumption 2, the scaling factors $\{\kappa_1, \kappa_2\}$ in (23) and (24) can be bounded as

$$\sum_{j=1}^{M} \frac{\mu}{2-\mu\lambda_j} \leq \kappa_1 \leq \frac{\sum_{j=1}^{M} \frac{\mu}{2-\mu\lambda_j}}{1-\sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j}} \tag{26}$$

$$\sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j} \leq \kappa_2 \leq \frac{\sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j}}{1-\sum_{j=1}^{M} \frac{\mu\lambda_j}{2-\mu\lambda_j}} \tag{27}$$

Both upper and lower bounds in (26) and (27) are simultaneously minimized by the same choice of $\{a_\ell\}$ as in (15). In addition, it can be verified that

$$\kappa_1 = \frac{\mu M}{2} + O(\mu^2) \tag{28}$$

$$\kappa_2 = \frac{\mu \mathrm{Tr}(R_u)}{2} + O(\mu^2) \tag{29}$$

Then, we arrive at the following conclusion, which confirms our earlier analysis.

Corollary 3 (Optimal mean-square performance for batch LMS) *Under Assumptions 1 and 2, the MSD and EMSE performance of the centralized (or batch) LMS algorithm* (17) *are minimized by using* (15) *and the optimal MSD and EMSE performance are given by*

$$MSD_{batch}^{opt} = \frac{\mu M}{2} \frac{1}{\mathrm{Tr}(R_v^{-1})} + O(\mu^2) \tag{30}$$

$$EMSE_{batch}^{opt} = \frac{\mu \mathrm{Tr}(R_u)}{2} \frac{1}{\mathrm{Tr}(R_v^{-1})} + O(\mu^2) \tag{31}$$

respectively. Moreover, the dominant mode of mean-square convergence is

$$mode_{batch} = 1 - 2\mu\lambda_{\min}(R_u) + O(\mu^2) \tag{32}$$

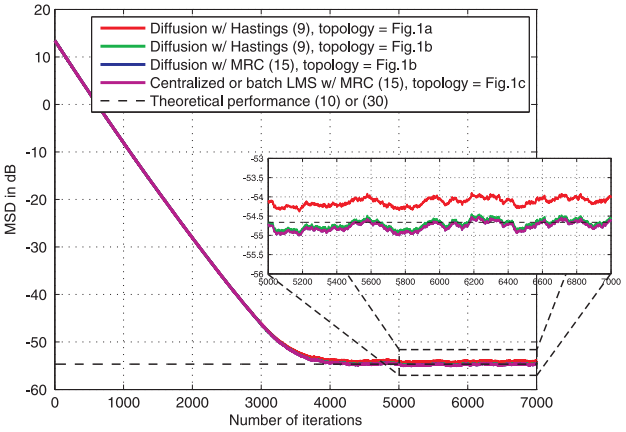
5. SIMULATION RESULTS AND CONCLUSIONS

We illustrate the results via simulations. We compare four different cases: (i) ATC diffusion (2) with the Hastings weights (9) over the graph in Fig.1a, (ii) ATC diffusion (2) with the Hastings weights (9) over the graph in Fig.1b, (iii) ATC diffusion (2) with the MRC weights (15) over the graph in Fig.1b, (iv) batch LMS (17) with the MRC weights (15) over the graph in Fig.1c. The step-size is $\mu = 0.002$. The noise variance profile across all agents is shown in Fig.2(c) and the spectrum of R_u is shown in Fig.2(d). The unknown parameter w^o is randomly generated. The MSD and EMSE learning curves are obtained by averaging over 100 experiments and are plotted in Figs.2(a) and 2(b), respectively. The theoretical results (10) and (11), or (30) and (31), are also plotted in Figs.2(a) and 2(b), respectively. It can be seen that theory matches simulations rather well (where the learning curves for ATC diffusion with MRC weights (15) overlap with the one with the Hastings weights (9) for the same graph in Fig.1b).

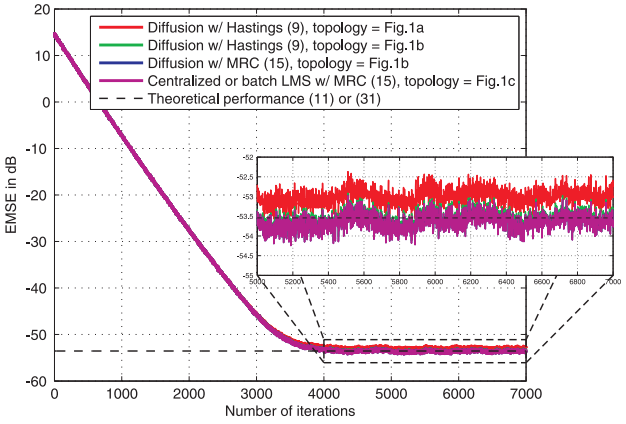
From Theorem 1 and Corollary 3, we conclude that by adopting the Hastings weights (9) and using small step-sizes, diffusion strategies can match the performance of optimal centralized (or batch) LMS solutions. Besides, from (9) and (15), we see that knowledge of the noise variances is necessary for computing the combination weights. In real applications where this knowledge is unavailable, each agent can learn it by the following recursive estimation [19]:

$$\hat{\boldsymbol{\sigma}}_{v,k}^2(i) = (1-\nu)\hat{\boldsymbol{\sigma}}_{v,k}^2(i-1) + \nu|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}|^2 \tag{33}$$

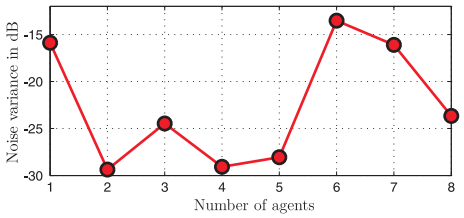
where $0 < \nu \ll 1$ is the forgetting factor.



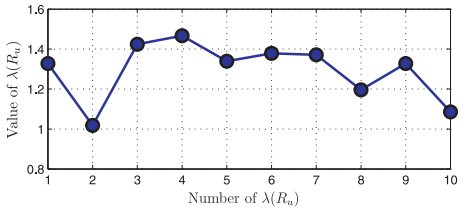
(a) MSD learning curves for various algorithms.



(b) EMSE learning curves for various algorithms.



(c) The noise variance profile.



(d) The spectrum of R_u.

Fig. 2. Simulation results.

6. REFERENCES

[1] B. T. Polyak and Y. Z. Tsypkin, "Pseudogradient adaptation and training algorithms," *Automat. Remote Contr.*, vol. 34, pp. 377–397, 1973.

[2] D. P. Bertsekas, "A new class of incremental gradient methods

for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, 1997.

[4] A. Nedic and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," in *Stochastic Optimization: Algorithms and Applications*, S. Uryasev and P. M. Pardalos, Eds., pp. 263–304. Kluwer Academic Publishers, 2000.

[5] M. G. Rabbat and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 798–808, Apr. 2005.

[6] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 223–229, Aug. 2007.

[7] D. Blatt, A. Hero, and H. Gauchman, "A convergent incremental gradient method with constant step size," *SIAM J. Optim.*, vol. 18, no. 1, pp. 29 – 51, Feb. 2007.

[8] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2365–2382, June 2009.

[9] A. Nedic and A. Ozdaglar, "Cooperative distributed multi-agent optimization," in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. C. Eldar, Eds., chapter 10, pp. 340 – 386. Cambridge Univ. Press, 2009.

[10] S. Sardellitti, M. Giona, and S. Barbarossa, "Fast distributed average consensus algorithms based on advection-diffusion processes," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 826 – 842, Feb. 2010.

[11] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[12] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.

[13] C. G. Lopes and A. H. Sayed, "Distributed processing over adaptive networks," in *Proc. Adapt. Sensor Array Process. Workshop (ASAP)*, MIT Lincoln Laboratory, MA, June 2006, pp. 1–5.

[14] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[15] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[16] S. Chouvardas, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4692–4707, Oct. 2011.

[17] A. H. Sayed, "Diffusion adaptation over netowrks," in *E-Reference Signal Processing*, R. Chellapa and S. Theodoridis, Eds. Elsevier, 2013, Also available online at http://arxiv.org/abs/1205.4220 [cs.MA], May 2012.

[18] S-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive netowrks," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6217–6234, Dec. 2012.

[19] X. Zhao and A. H. Sayed, "Performance limits for distributed estimation over LMS adaptive netowrks," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5107–5124, Oct. 2012.

[20] A. H. Sayed, *Adaptive Filters*, Wiley, NJ, 2008.