Robust Adaptive Machine Learning Algorithms for Distributed Signal Processing

Symeon Chouvardas*
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
15784, Ilissia, Athens, Greece

Abstract

Distributed networks comprising a large number of nodes, *e.g.,* Wireless Sensor Networks, Personal Computers (PC's), laptops, smart phones, etc., which cooperate with each other in order to reach to a common goal, constitute a promising technology for several applications. Typical examples include: distributed environmental monitoring, acoustic source localization, power spectrum estimation, etc. Sophisticated cooperation mechanisms can significantly benefit the learning process, through which the nodes achieve their common objective.

In this dissertation, the problem of adaptive learning in distributed networks is studied, focusing on the task of distributed estimation. A set of nodes sense information related to certain parameters and the estimation of these parameters comprises the goal. Towards this direction, nodes exploit locally sensed measurements as well as information springing from interactions with other nodes of the network. Throughout this dissertation, the cooperation among the nodes follows the diffusion optimization rationale and the developed algorithms belong to the APSM algorithmic family.

First, robust APSM–based techniques are proposed. The goal is to "harmonize" the spatial information, received from the neighborhood, with the locally sensed one. This "harmonization" is achieved by projecting the information of the neighborhood onto a convex set, constructed via the locally sensed measurements. Next, the scenario,

*Advisors: S. Theodoridis, Professor, S. Perantonis, Senior Researcher, NCSR "Demokritos", N. Kalouptsidis Professor.

in which a subset of the node set is malfunctioning and produces measurements heavily corrupted with noise, is considered. This problem is attacked by employing the Huber cost function, which is resilient to the presence of outliers. In the sequel, we study the issue of sparsity–aware adaptive distributed learning. The nodes of the network seek for an unknown sparse vector, which consists of a small number of non–zero coefficients. Weighted ℓ_1–norm constraints are embedded, together with sparsity–promoting variable metric projections. Finally, we propose algorithms, which lead to a reduction of the communication demands, by forcing the estimates to lie within lower dimensional Krylov subspaces. The derived schemes serve a good trade-off between complexity/bandwidth demands and achieved performance.

Subject Area: Adaptive Learning, Distributed Signal Processing.
Keywords: Diffusion, Projections, APSM, hyperslabs.

1 Introduction

Distributed networks comprising a number of connected nodes, *e.g.,* Personal Computers (PC's), laptops, smart phones, surveillance cameras and microphones, wireless sensor networks etc., which exchange information in order to reach a common goal, are envisioned to play a central role in many applications. Typical examples of emergent applications involving distributed networks are: distributed environmental monitoring, acoustic source localization, power spectrum estimation, target tracking, surveillance, traffic control, patient monitoring and hospital surveillance, just to name a few [1,3,6,7,11]. All the previously mentioned applications share in common the fact that the nodes are deployed over a geographic region providing spatial diversity to the obtained measurements. Henceforth, the development of algorithms and node cooperation mechanisms, which exploit the information diversity over time and space, so that a common objective to be reached, becomes essential.

In this dissertation, the problem of distributed processing is studied with a focus on the distributed/decentralized estimation task. A number of nodes, which are spread over a geographic region, sense information related to certain parameters; the estimation of these parameters comprises our goal. The main idea behind distributed processing is that the nodes exchange information among them and make decisions/computations in a *collaborative* way instead of working individually, using solely the information that is locally sensed. It is by now well established, that the cooperation among the nodes leads to better results compared to the case where they act as individual learners, see for example [5, 10, 13]. The need to develop node cooperation mechanisms is increased due to the presence of noise in the majority of applications. More specifically, the measurements observed at each node are corrupted by noise, and this fact adds further uncertainty on the obtained estimates of the unknown target parameters. This uncertainty can be reduced via the cooperation of the nodes.

In decentralized networks, the following issues have to be taken into consideration:

- *Performance:* A performance close to the optimal, that is the one associated with the centralized networks, which use all the available data, has to be achieved. In other words, despite the fact that direct communication among some of the nodes cannot be established, sophisticated cooperation mechanisms have to be developed, in order to "push" the performance to be as close as possible to the ideal scenario.

- *Robustness to possible failures:* As it has been already stated, a major drawback of the centralized topology is that if the FC fails then the network collapses. Decentralized networks have to be constructed so as to be robust against possible node failures.

- *Bandwidth and complexity constraints:* The amount of transmitted information has to be as small as possible, in order to keep the bandwidth low. Furthermore, since in decentralized networks a central processing unit with powerful computational capabilities is not present and usually cheap processing units comprise the nodes, low-complexity schemes have to be developed.

- *Adaptivity:* In many applications, such as, source localization, spectrum sensing, etc, the nodes of the network are tasked to estimate non–stationary parameters, *i.e.,* parameters which vary with time. Batch estimation algorithms, which use all the available training data simultaneously, cannot attack such problems. To this end, adaptive techniques have to be developed, where the data are observed sequentially, one per (discrete) time instance and operate in an online fashion for updating and improving the estimates.

The main objective of this dissertation is to develop algorithms in the context of *adaptive estimation* in distributed networks. The diffusion optimization rationale is adopted and the proposed algorithms belong to the Adaptive Projected Subgradient Method (APSM) algorithmic family.

2 Adaptive Robust Algorithms for Distributed Learning

As a first step, distributed algorithms, which follow the diffusion rationale and belong to the family of the Adaptive Projected Subgradient Method, are developed. The proposed algorithms adopt a novel combine–project–adapt cooperation protocol. The intermediate extra projection step of this protocol "harmonizes" the local information, which comprises the input/output measurements, with the information coming from the neighborhood, *i.e.,* the estimates obtained from the neighboring nodes. This is achieved by projecting the vector, occurring by combining the estimates of the neighbourhood, to a convex set, namely a hyperslab, which is constructed by exploiting locally sensed information. The steps of the algorithm can be summarised as follows:

1. **Combination Step:** The estimates from the nodes that belong to the neighbourhood are received and convexly combined with respect to the combination weights.

2. **Projection Step:** The resulting aggregate is first projected onto a properly constructed hyperslab.

3. **Adaptation Step:** The adaptation step is performed.

The following model is adopted. A network of N nodes is considered and each node, k, at time n, has access to the measurements $d_{k,n} \in \mathbb{R}$, $\boldsymbol{u}_{k,n} \in \mathbb{R}^m$ generated by the linear system:

$$d_{k,n} = \boldsymbol{w}_*^T \boldsymbol{u}_{k,n} + v_{k,n}, \tag{1}$$

where $v_{k,n}$ is an additive noise process of zero mean and variance σ_k^2. The goal is the estimation of the $m \times 1$ vector $\boldsymbol{w}_*$.

As we have already mentioned, an APSM–based scheme, which employs projections onto hyperslabs, is developed. The scheme is brought in a distributed fashion by following the diffusion rationale. Moreover, here an extra step is added, that follows the combination stage and precedes the adaptation one. More specifically, the result of the combination step is projected onto the hyperslab $S'_{k,n}$, which is defined as

$$S'_{k,n} = \{\boldsymbol{w} \in \mathbb{R}^m : |d_{k,n} - \boldsymbol{w}^T \boldsymbol{u}_{k,n}| \leq \epsilon'_k\},$$

where $\epsilon'_k > \epsilon_k$ and ϵ_k is the user defined parameter associated with the hyperslabs, that will be used in the adaptation step at node k, *i.e.,*

$$S_{k,n} = \{\boldsymbol{w} \in \mathbb{R}^m : |d_{k,n} - \boldsymbol{w}^T \boldsymbol{u}_{k,n}| \leq \epsilon_k\}.$$

The algorithm comprises the following steps:

1. **Combination Step:** The estimates from the nodes that belong to $\mathcal{N}_k$ are received and convexly combined with respect to the combination weights $a_{k,l}$.

2. **Projection Step:** The resulting aggregate is first projected onto the hyperslab $S'_{k,n}$[1].

3. **Adaptation Step:** The adaptation step is performed.

$$\boldsymbol{\phi}_{k,n} = \sum_{l \in \mathcal{N}_k} a_{k,l} \boldsymbol{w}_{l,n}, \tag{2}$$

$$\boldsymbol{z}_{k,n} = P_{S'_{k,n}} \left(\boldsymbol{\phi}_{k,n} \right), \tag{3}$$

$$\boldsymbol{w}_{k,n+1} = \boldsymbol{z}_{k,n} + \mu_{k,n} \left(\sum_{j \in \mathcal{J}_n} \omega_{k,j} P_{S_{k,j}}(\boldsymbol{z}_{k,n}) - \boldsymbol{z}_{k,n} \right), \tag{4}$$

where $P_{S'_{k,n}}$ and $P_{S_{k,n}}$ are the projection operators onto the respective hyperslabs, $\sum_{j \in \mathcal{J}_n} \omega_{k,j} = 1$ and $\mathcal{J}_n := \overline{\max\{0, n-q+1\}, n}$. As it was experimentally verified, the proposed scheme exhibits an enhanced performance, both in terms of convergence speed as well as steady state error floor, compared to other state of the art algorithms, of similar complexity. Finally, it was proved that the algorithm enjoys a number of nice convergence properties such as monotonicity, strong convergence to a point and consensus.

3 Introducing Robustness to Cope with a Failure of Nodes

Consider a scenario, in which some of the nodes are damaged and the associated observations are very noisy. More specifically, it is assumed that that the noise is additive and white, albeit the standard deviation of the "damaged" nodes becomes larger, compared to the one of the "healthy" nodes. In such cases, the use of loss functions, suggested in the framework of robust statistics, are more appropriate to cope with outliers. A popular cost function of this family is the Huber cost function, *e.g.,* [8, 12].

In the current study we employ a slightly modified version of the Huber cost function, compared to the classical one. The difference is that in our

[1] The projection of a point onto a hyperslab is provided in Chapter 3.

context a 0–th level set is introduced, *i.e.,* a set of points in which the function scores a zero loss. Our goal is to find points, which lie in the previously mentioned 0–th level set. The geometry of the Huber function is illustrated in Fig. In contrast to the hyperslab case, the projection onto the 0–th level set of the Huber cost function, does not admit a closed form. For this reason, projections onto the halfspace, associated to the subgradient of the Huber loss function, take place. We can also include the extra projection step, described in the previous section, by introducing a modified version of the Huber cost function and following a similar rationale as in the hyperslab case. However, instead of projecting the result of the combination step onto an external hyberslab, we project it onto a halfspace that is generated by a properly modified cost function. The proposed algorithm comprises the following steps:

$$\boldsymbol{\phi}_{k,n} = \sum_{l \in \mathcal{N}_k} a_{k,l} \boldsymbol{w}_{l,n}, \tag{5}$$

$$\boldsymbol{z}_{k,n} = P_{H'^{-}_{k,n}} \left(\boldsymbol{\phi}_{k,n} \right), , \tag{6}$$

$$\boldsymbol{w}_{k,n+1} = \boldsymbol{z}_{k,n} + \mu_{k,n} \left(\sum_{j \in \mathcal{J}_n} \omega_{k,j} P_{H^{-}_{k,j}} (\boldsymbol{z}_{k,n}) - \boldsymbol{z}_{k,n} \right), \tag{7}$$

where $P_{H^{-}_{k,j}}$ stands for the projection onto the halfspace associated to the Huber loss function and $P_{H'^{-}_{k,j}}$ is the previously described extra projection step. Under some mild assumptions, the developed algorithm enjoys monotonicity, asymptotic optimality, asymptotic consensus and strong convergence to a point that lies in the consensus subspace. Finally, numerical examples verified that the proposed scheme has an enhanced performance, compared to the other methodologies, in a network with malfunctioning nodes.

4 Sparsity–Aware Adaptive Distributed Learning

As a next step, an APSM–based sparsity–promoting adaptive algorithm for distributed learning in ad–hoc networks is developed. At each time instance and at each node of the network, a hyperslab is constructed based on the received measurements; this defines the region in which the solution is searched for. Sparsity encouraging variable metric projections onto these sets have been adopted. In addition, sparsity is also imposed by employing variable metric projections onto weighted ℓ_1 balls. A combine adapt cooperation strategy is followed.

Let us introduce, here, the sparsity promoting *variable metric projection*, onto the respective hyperslabs, with respect to the matrix $\boldsymbol{G}_n$, defined as [14]:

$$\forall \boldsymbol{w} \in \mathbb{R}^m, \quad P_{S_n}^{(\boldsymbol{G}_n)}(\boldsymbol{w}) := \boldsymbol{w} + \beta_n \boldsymbol{G}_n^{-1} \boldsymbol{u}_n, \tag{8}$$

where

$$\beta_n = \begin{cases} \dfrac{d_n - \boldsymbol{u}_n^T \boldsymbol{w} + \epsilon}{\|\boldsymbol{u}_n\|^2_{\boldsymbol{G}_n^{-1}}}, & \text{if } d_n - \boldsymbol{u}_n^T \boldsymbol{w} < -\epsilon, \\ 0, & \text{if } |d_n - \boldsymbol{u}_n^T \boldsymbol{w}| \le \epsilon, \\ \dfrac{d_n - \boldsymbol{u}_n^T \boldsymbol{w} - \epsilon}{\|\boldsymbol{u}_n\|^2_{\boldsymbol{G}_n^{-1}}}, & \text{if } d_n - \boldsymbol{u}_n^T \boldsymbol{w} > \epsilon, \end{cases}$$

and $\|\boldsymbol{u}_n\|^2_{\boldsymbol{G}_n^{-1}}$ denotes the weighted norm, with definition $\|\boldsymbol{u}_n\|^2_{\boldsymbol{G}_n^{-1}} := \boldsymbol{u}_n^T \boldsymbol{G}_n^{-1} \boldsymbol{u}_n$ (see Appendix C). Note that if $\boldsymbol{G}_n = \boldsymbol{I}_m$, then (8) is the Euclidean projection onto a hyperslab. The positive definite diagonal matrix $\boldsymbol{G}_n^{-1}$ is constructed following similar philosophy as in [2, 15]. The i-th coefficient of its diagonal equals to $g_{i,n}^{-1} = \frac{1-\overline{\alpha}}{m} + \overline{\alpha} \frac{|w_i^{(n)}|}{\|\boldsymbol{w}_n\|_1}$, where $\overline{\alpha} \in [0, 1)$ is a parameter, that determines the extend to which the sparsity level of the unknown vector will be taken into consideration, and $w_i^{(n)}$ denotes the i-th component of $\boldsymbol{w}_n$. In order to grasp the reasoning of the variable metric projections, consider the ideal situation, in which $\boldsymbol{G}_n^{-1}$ is generated by the unknown vector $\boldsymbol{w}_*$. It is easy to verify that $g_{i,n}^{-1} > g_{i',n}^{-1}$, if $i \in \operatorname{supp}(\boldsymbol{w}_*)$, and $i' \notin \operatorname{supp}(\boldsymbol{w}_*)$, where $\operatorname{supp}(\cdot)$ stands for the support set of a vector, *i.e.,* the set of the non–zero coefficients. Hence, employing the variable metric projection, the amplitude of each coefficient of the vector used to construct $\boldsymbol{G}_n^{-1}$ determines the weight that will be assigned to the corresponding coefficient of the second term of the right hand side in (8). That is, components with smaller magnitude are multiplied with small coefficients of $\boldsymbol{G}_n^{-1}$. Loosely speaking, the variable metric projections accelerate the convergence speed when tracking a sparse vector, since by assigning different weights pushes the coefficients of the estimates with small amplitude to diminish faster. The geometric implication of it is that the projection is made to "lean" towards the direction of the more significant components of the currently available estimate.

In the algorithm which is presented here, we go one step further, as far as sparsity is concerned. In a second stage, additional sparsity-related constraints, which are built around the weighted ℓ_1 ball, are employed, [4]. A sparsity promoting adaptive scheme, based on set-theoretic estimation arguments, in which the constraints are weighted ℓ_1 balls, was presented in [9]. Given a vector of weights $\boldsymbol{\psi}_n = [h_1^{(n)}, \dots, h_m^{(n)}]^T$, where $h_i^{(n)} > 0, \forall i = 1, \dots, m$, and a positive radius, δ, the weighted ℓ_1 ball is defined as: $B_{\ell_1}[\boldsymbol{\psi}_n, \delta] := \{\boldsymbol{w} \in \mathbb{R}^m : \sum_{i=1}^m h_i^{(n)} |w_i| \le \delta\}$. The projection onto $B_{\ell_1}[\boldsymbol{\psi}_n, \delta]$,

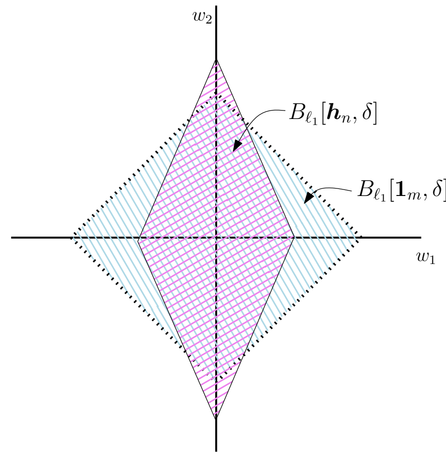


Figure 1: Illustration of a weighted ℓ_1 ball (solid line magenta) and an unweighted ℓ_1 ball (dashed line blue).

is given in [9, Theorem 1], and the geometry of these sets is illustrated in Fig. 1.

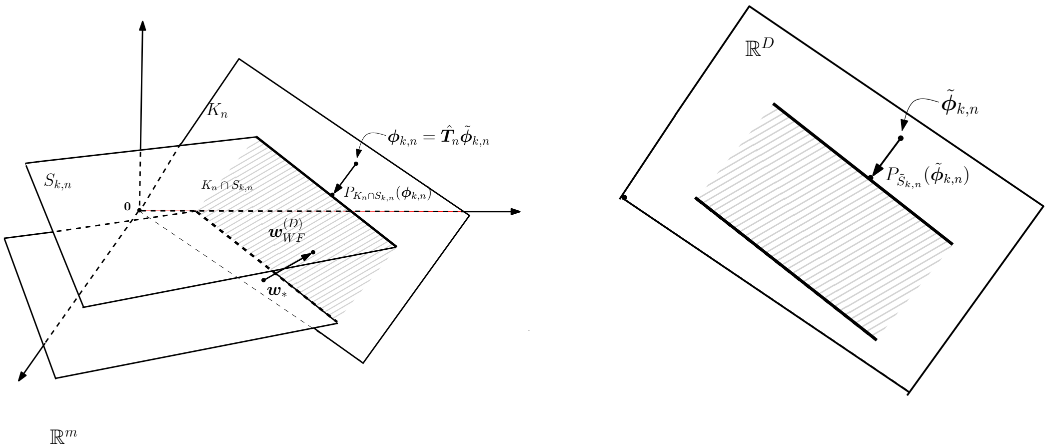
The steps of the algorithm are summarized in the sequel:

$$\boldsymbol{w}_{k,n+1} = P^{(\boldsymbol{G}_n)}_{B_{\ell_1}[\boldsymbol{\psi}_n,\delta]} \left(\boldsymbol{\phi}_{k,n} + \mu_{k,n} \left(\sum_{j\in\mathcal{J}} \omega_{k,j} P^{(\boldsymbol{G}_n)}_{S_{k,j}}(\boldsymbol{\phi}_{k,n}) - \boldsymbol{\phi}_{k,n} \right)\right), \quad (9)$$

The theoretical properties of the algorithm are studied and it is shown that under some mild assumptions, the scheme enjoys monotonicity, asymptotic optimality and strong convergence to a point that lies in the consensus subspace. Finally, numerical examples verify the enhanced performance obtained by the proposed scheme compared to other algorithms, which have been developed in the context of sparsity–aware adaptive learning.

5 Dimensionality Reduction in Distributed Adaptive Learning via Krylov Subspaces

In this section, the problem of dimensionality reduction in adaptive distributed learning is studied. As in the previous sections, the algorithm, to be presented here, is based on the APSM algorithmic family. At each time instant and at each node of the network, a hyperslab is constructed based on the received measurements and this defines the region in which the solution is searched for. Moreover, in order to reduce the number of transmitted coefficients, which is dictated by the dimension of the unknown vector, we seek



(a) (b)

Figure 2: (a) Geometrical illustration of the algorithm for $q = 1$. The aggregate $\boldsymbol{\phi}_{k,n}$, which belongs in the subspace, is projected onto the intersection of the subspace and the hyperslab, generated by the measurement data. (b) The algorithmic scheme in the reduced dimension space, i.e., $\mathbb{R}^D$.

for possible solutions in a subspace of lower dimensionality; the technique will be developed around the Krylov subspace rationale. Our goal is to find a point that belongs to the intersection of this infinite number of hyperslabs and the respective Krylov subspaces. This is achieved via a sequence of projections onto the property sets as well as the Krylov subspaces. The proposed schemes are brought in a decentralized form by adopting the combine-adapt cooperation strategy among the nodes.

The steps of the algorithm can be encoded in the following formula:

$$\tilde{\boldsymbol{w}}_{k,n+1} = \tilde{\boldsymbol{\phi}}_{k,n} + \tilde{\mu}_{k,n} \left(\sum_{j \in \mathcal{J}} \omega_{k,j} P_{\tilde{S}_{k,j}}(\tilde{\boldsymbol{\phi}}_{k,n}) - \tilde{\boldsymbol{\phi}}_{k,n} \right), \tag{10}$$

where the vectors $tilde\boldsymbol{w}_{k,n+1}$, $\tilde{\boldsymbol{\phi}}_{k,n+1}$ belong to the reduced dimension space and the reduced dimension hyperslab is given by: $\tilde{S}_{k,n} := \{\tilde{\boldsymbol{w}} \in \mathbb{R}^D : |d_{k,n} - \boldsymbol{u}_{k,n}^T \hat{\boldsymbol{T}}_n \tilde{\boldsymbol{w}}| \leq \epsilon_k, \}$ where $\hat{\boldsymbol{T}}_n$ is a matrix, the columns of which, span the Krylov subspace. The geometrical interpretation of the algorithm is illustrated in Fig. 2

As in the previously derived schemes, the theoretical properties of the algorithm are studied and it is shown that the scheme enjoys monotonicity, asymptotic optimality and strong convergence to a point that lies in the intersection of the consensus subspace with the Krylov Subspace. Finally, numerical examples verify that the proposed scheme provides a good trade-off between the number of transmitted coefficients and the respective performance.

6 List of Publications

Journal Publications

- S. Chouvardas, K. Slavakis, and S. Theodoridis. Adaptive robust distributed learning in diffusion sensor networks. *IEEE Transactions on Signal Processing*, 59(10):4692–4707, 2011.

- S. Chouvardas, K. Slavakis, Y. Kopsinis, and S. Theodoridis. A sparsity promoting adaptive algorithm for distributed learning. *IEEE Transactions on Signal Processing*, 60(10):5412 –5425, Oct. 2012.

- S. Chouvardas, K. Slavakis, and S. Theodoridis. Trading off complexity with communication costs in distributed adaptive learning via Krylov subspaces for dimensionality reduction. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):257–273, 2013.

- Symeon Chouvardas, Konstantinos Slavakis, Sergios Theodoridis, and Isao Yamada. Stochastic analysis of hyperslab–based adaptive projected subgradient method under bounded noise. *To appear in IEEE Signal Processing Letters*, 2013.

Conference Publications

- S. Chouvardas, K. Slavakis, and S. Theodoridis. A novel adaptive algorithm for diffusion networks using projections onto hyperslabs. In *CIP*, pages 393–398. IEEE, 2010 **(best student paper award)**.

- Symeon Chouvardas, Konstantinos Slavakis, and Sergios Theodoridis. Trading off communications bandwidth with accuracy in adaptive diffusion networks. In *ICASSP*, pages 2048–2051. IEEE, 2011.

- Symeon Chouvardas, Konstantinos Slavakis, Yannis Kopsinis, and Sergios Theodoridis. Sparsity-promoting adaptive algorithm for distributed learning in diffusion networks. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1084–1088. IEEE, 2012.

- Symeon Chouvardas, Gerasimos Mileounis, Nicholaos Kalouptsidis, and Sergios Theodoridis. A greedy sparsity–promoting LMS for distributed adaptive learning in diffusion networks. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013.

- Symeon Chouvardas, Gerasimos Mileounis, Nicholaos Kalouptsidis, and Sergios Theodoridis. Training-Based and Blind Algorithms for Sparsity-Aware Distributed Learning. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2013.

References

[1] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102–114, 2002.

[2] Jacob Benesty and Steven L. Gay. An improved pnlms algorithm. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1881–1884, may 2002.

[3] Doron Blatt and Alfred O Hero. Energy-based sensor network source localization via projection onto convex sets. *Signal Processing, IEEE Transactions on*, 54(9):3614–3619, 2006.

[4] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[5] F.S. Cattivelli and A.H. Sayed. Diffusion lms strategies for distributed estimation. *IEEE Transactions on Signal Processing*, 58(3):1035–1048, 2010.

[6] Chee-Yee Chong and Srikanta P Kumar. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE*, 91(8):1247–1256, 2003.

[7] Deborah Estrin, Lewis Girod, Greg Pottie, and Mani Srivastava. Instrumenting the world with wireless sensor networks. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 4, pages 2033–2036. IEEE, 2001.

[8] PJ Huber and EM Ronchetti. *Robust Statistics*. New York: Willey, 2009.

[9] Y. Kopsinis, K. Slavakis, and S. Theodoridis. Online sparse system identification and signal reconstruction using projections onto weighted ℓ_1 balls. *IEEE Transactions on Signal Processing*, 59(3):936 –952, 2011.

[10] Jinchao Li and Ali H Sayed. Modeling bee swarming behavior through diffusion adaptation with asymmetric information sharing. *EURASIP Journal on Advances in Signal Processing*, 18, 2012.

[11] Joel B Predd, SB Kulkarni, and H Vincent Poor. Distributed learning in wireless sensor networks. *Signal Processing Magazine, IEEE*, 23(4):56–69, 2006.

[12] M. Rabbat and R. Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27. ACM New York, NY, USA, 2004.

[13] Sheng-Yuan Tu and Ali H Sayed. Foraging behavior of fish schools via diffusion adaptation. In *CIP*, pages 63–68. IEEE, 2010.

[14] M. Yukawa and I. Yamada. Set-theoretic adaptive filtering based on data-driven sparsification. *International Journal of Adaptive Control and Signal Processing*, 25:707–722.

[15] M. Yukawa and I. Yamada. A unified view of adaptive variable-metric projection algorithms. *EURASIP Journal on Advances in Signal Processing*, 2009:2–2, 2009.