A PROXIMAL SPLITTING APPROACH TO REGULARIZED DISTRIBUTED ADAPTIVE ESTIMATION IN DIFFUSION NETWORKS

Wemer M. Wee and Isao Yamada

Department of Communications and Computer Engineering, Tokyo Institute of Technology
2-12-1-S3-60 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan
Email: {wemer, isao}@sp.ce.titech.ac.jp

ABSTRACT

We propose a proximal splitting approach to regularized distributed estimation over networks employing diffusion adaptation strategies. Playing a central role in the proposed framework is the so-called proximity operator, which is a generalization of the convex projection mapping, that enables us to handle convex regularization terms efficiently. The diffusion algorithms developed using the proximal formalism endow networks with new learning abilities and open up possibilities for enhancing performance of the networks by utilizing more general convex penalties. We present performance analysis of the proposed method and provide simulations to demonstrate its feasibility in recovering sparse signals.

Index Terms— Adaptive networks, diffusion strategies, energy conservation, proximity operator, sparsity

1. INTRODUCTION

Distributed adaptive estimation has recently emerged as an attractive and important research area due to its possible applications in wireless sensor networks, dynamic resource allocation and bio-inspired processing [1, 2, 3, 4, 5, 6]. In the typical setting, a set of nodes are allowed to exchange information with each other and perform local computations to improve their own estimates, which are then used to collectively arrive at a solution to the problem. Among the most popular modes of cooperation are the so-called diffusion adaptation strategies [1, 2, 7, 8, 9, 10], which have been proven effective in exploiting the time and spatial diversity of the data, thereby maximizing the learning and tracking abilities of networks.

In this paper, we consider the distributed estimation problem where the parameter of interest is known to satisfy some signal property such as sparsity. To this end, we investigate a variational formulation of the estimation problem with regularization represented by some convex penalty term such as the ℓ_1 norm. Exploiting a priori signal information in this manner leads to improved estimation performance as have been demonstrated in works such as the Lasso [11] and sparsity-aware adaptive filtering [12, 13, 14, 15].

The main objective of this work is to solve the regularized distributed estimation problem by developing a diffusion implementation of the forward-backward splitting method [16, 17, 18, 19]. An important advantage afforded by the proposed design is the use of the proximity operator [20], which enables us to exploit operations that are beyond convex projections [19]. In particular, we will be able to employ the well-known soft-thresholding operation [21], which is an effective tool in a variety of image recovery problems. Although proximal algorithms in general may have limited impact on actual numerical performance, they are considered to be more stable than gradient and subgradient iterations [22]. Moreover, diffusion proximal algorithms open up new possibilities in the consideration of cost functions or constraints in adaptive networks. We also have at our disposal several results that aim to improve the performance of these methods such as acceleration and overrelaxation [23, 24, 25], and using similar techniques may further enhance the performance of the adaptive network. To gain insight into the behavior of the proposed diffusion algorithm, we appeal to the celebrated energy conservation framework [26, 8] to obtain mean and mean-square convergence guarantees and present simulations that demonstrate improved detection of sparsity compared to the standard and sparsity-aware diffusion least-mean-squares (LMS) algorithms [1, 27].

This research was supported in part by the Japan Society for the Promotion of Science Grants-in-Aid (B-21300091). The work of W. M. Wee is also supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

1.1. Relation to Prior Work

Diffusion adaptation strategies were investigated in the seminal work [1] and further discussed and developed in [7, 28, 9, 2, 29]. To improve the performance of diffusion networks in the presence of sparse signals, subgradient-based diffusion LMS filters were proposed in [27, 30]. The present work provides an alternative framework for developing regularized diffusion algorithms using the proximal formalism, which is related to the set-theoretic formulation of diffusion algorithms investigated in [31, 32, 33]. The use of the proximal forward-backward splitting in an adaptive setting has been already considered in [34], where the technique was shown to outperform conventional algorithms in detecting sparsity. This paper, on the other hand, is focused on the design of an adaptive *and* distributed implementation of the algorithm, and we present a mean-square performance analysis of the proposed method based on [7, 27] to determine any relation with existing algorithms.

1.2. Notation

Let $\mathbb{N}$ be the set of all nonnegative integers and $\mathbb{R}^M$ denote the M-dimensional Euclidean space with norm $\|\cdot\|$. We use capital letters to represent matrices and small letters to denote vectors. Moreover, we use boldface characters to denote random variables and regular font characters for realizations and other deterministic quantities. We use I to denote an identity matrix of appropriate dimensions, $\mathrm{Tr}(\cdot)$ to denote the trace of a matrix, $\mathrm{diag}\{\cdots\}$ as the diagonal matrix consisting of its entries and $\mathrm{col}\{\cdots\}$ to denote a column vector obtained by stacking its entries. The expectation is denoted by $\mathbb{E}[\cdot]$.

2. REGULARIZED DISTRIBUTED ESTIMATION

2.1. Problem Formulation

Consider a network with N nodes in a predefined topology. Suppose that at each time $i \geq 0$, each node k has access to a measurement $d_k(i) \in \mathbb{R}$ of some random process $\boldsymbol{d}_k(i)$ and a regression vector $u_{k,i} \in \mathbb{R}^{1\times M}$ corresponding to a realization of a random process $\boldsymbol{u}_{k,i}$. Furthermore, suppose that $\boldsymbol{u}_{k,i}$ is correlated with $\boldsymbol{d}_k(i)$ and that we have a positive definite covariance matrix $R_{u,k} := \mathbb{E}\boldsymbol{u}_{k,i}^T\boldsymbol{u}_{k,i}$. We assume the data to be related via the model

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

where $w^\circ \in \mathbb{R}^M$ is an unknown vector we wish to estimate and $\boldsymbol{v}_k(i) \in \mathbb{R}$ is a zero mean random variable with variance $\sigma_{v,k}^2$, which is independent of $\boldsymbol{u}_{k,i}$ for all k and i, and independent of $\boldsymbol{v}_l(j)$ for $l \neq k$ or $i \neq j$.

Our main objective in this paper is to develop the proximal forward-backward splitting method for minimizing the following cost in an adaptive and distributed manner:

$$J^{\mathrm{glob}}(w) := \sum_{k=1}^{N} \left\{(d_k(i) - u_{k,i}w)^2 + \gamma R(w)\right\} \tag{2}$$

where $R(w)$ represents some real-valued convex penalty function and $\gamma > 0$ is a regularization parameter. To effectively distribute adaptation among the nodes, we will employ the diffusion strategy as developed in [1, 7, 2, 27] to minimize the cost $J^{\mathrm{glob}}(w)$.

2.2. Proximity Operators and Proximal Splitting

We first recall the forward-backward splitting method before we describe its diffusion implementation. Consider the time varying cost function

$$J_i(w) = F_i(w) + \gamma R_i(w) \quad (\forall i \in \mathbb{N}) \tag{3}$$

where $R_i : \mathbb{R}^M \to (-\infty, +\infty]$ is lower semicontinuous, convex and not identically $+\infty$ while F_i is a differentiable convex function with an L-Lipschitz continuous gradient, that is, $(\forall (x,y) \in \mathbb{R}^M \times \mathbb{R}^M)\ \|\nabla F_i(x) - \nabla F_i(y)\| \leq L\|x-y\|$ for some $L > 0$. Then, under the above hypotheses on the functions F_i and R_i, the cost (3) may be suppressed by the adaptive proximal forward-backward splitting (APFBS) algorithm [34], which is described by the following: Fix $\varepsilon \in (0, \min\{1, 1/L\})$ and $w_{-1} \in \mathbb{R}^M$. Let $\mu \in [\varepsilon, 2/L - \varepsilon]$. For each $i \in \mathbb{N}$, repeat:

$$w_i = \mathrm{prox}_{\mu\gamma R_i}(w_{i-1} - \mu\nabla_w F_i(w_{i-1})) \tag{4}$$

where prox is the so-called proximity operator of index $\kappa \in (0, +\infty)$ defined by

$$\mathrm{prox}_{\kappa R_i} : \mathbb{R}^M \to \mathbb{R}^M : w \mapsto \underset{y\in\mathbb{R}^M}{\arg\min}\left(R_i(y) + \frac{1}{2\kappa}\|w-y\|^2\right). \tag{5}$$

The APFBS algorithm is a time-varying extension of the proximal forward-backward splitting method [19], and is known to satisfy the monotone approximation property [35].

To demonstrate the use of the proximal method for sparse signal recovery, we consider two convex regularization terms as popularized in compressed sensing. The first one is the ℓ_1 norm defined by $R(w_i) := \|w_i\|_1 := \sum_{m=1}^{M} |w_i(m)|$, where $w_i(m)$ represents the m-th entry of the vector w_i. This choice leads to the proximity operator called soft-thresholding which is given by

$$\mathrm{prox}_{\kappa\|\cdot\|_1} w_i = \sum_{m=1}^{M} \mathrm{soft}(w_i(m), \kappa) b_m,$$

where $\mathrm{soft}(x,\kappa) := \max(|x| - \kappa, 0)\mathrm{sgn}(x)$, with sgn representing the signum or sign function, and $\{b_m\}_{m=1}^M$ is the standard orthonormal basis of $\mathbb{R}^M$. The second regularization term we consider is based on the concept of reweighting [36, 15] that seeks to improve the efficiency of the ℓ_1 norm, which gives rise to its weighted version defined by $\|w_i\|_{1,\omega} := \sum_{m=1}^{M} \omega_m^{(i)} |w_i(m)|$, where $\omega_m^{(i)}$ are positive weights that are updated at each iteration as follows: for each $i \in \mathbb{N}$, $\omega_m^{(i)} := (|w_{i-1}(m)| + \varrho)^{-1}$ for $m = 1, 2, \ldots, M$ and some $\varrho > 0$. The associated proximity operator is the soft-thresholding with weighting shown below:

$$\mathrm{prox}_{\kappa\|\cdot\|_{1,\omega}} w_i = \sum_{m=1}^{M} \mathrm{soft}(w_i(m), w_m^{(i)}\kappa) b_m$$

Remark 2.2.1 *Note that other regularization terms such as those for detecting group sparsity [37] and for total variation filtering or denoising [38] can also be considered under this proximal formalism. We provide details regarding these and consider a variation of the forward-backward splitting in an extended version of this paper.*

2.3. Diffusion Adaptation

To promote higher levels of interaction and information exchange among the nodes, we use an approach based on [7]. For this purpose, consider an $N \times N$ matrix C with nonnegative entries $\{c_{lk}\}$ that satisfy $C\mathbb{1} = \mathbb{1}$, $C^T\mathbb{1} = \mathbb{1}$ and $c_{lk} = 0$ if $l \notin \mathcal{N}_k$, where $\mathbb{1}$ denotes the $N \times 1$ vector with unit entries and $\mathcal{N}_k$ denotes the neighborhood of node k (which includes node k itself). Using these coefficients, we may consider the modified local cost function

$$J_k^{\mathrm{loc}}(w) := \sum_{l\in\mathcal{N}_k} c_{lk}(d_l(i) - u_{l,i}w)^2 + \gamma R(w). \tag{6}$$

Now by adding a combination step before or after the adaptation step, one may arrive at different diffusion adaptation strategies. Here we focus only on an adapt-then-combine diffusion method. Together with the combination matrix C, we introduce a matrix A composed of nonnegative weight coefficients $\{a_{l,k}\}$ which satisfy $A^T\mathbb{1} = \mathbb{1}$ and $a_{lk} = 0$ if $l \notin \mathcal{N}_k$. The proximal diffusion algorithm can now be presented as follows. Set $w_{k,-1} = 0 \in \mathbb{R}^M$ for all k. Given the nonnegative coefficients $\{c_{l,k}, a_{l,k}\}$ satisfying the properties above, for each time $i \geq 0$ and for each node k, repeat:

$$\begin{cases} \psi_{k,i} = w_{k,i-1} + \mu_k \sum_{l\in\mathcal{N}_k} c_{l,k} u_{l,i}^T [d_l(i) - u_{l,i} w_{k,i-1}] \\ \chi_{k,i} = \mathrm{prox}_{\mu_k\gamma R}\psi_{k,i} \\ w_{k,i} = \sum_{l\in\mathcal{N}_k} a_{l,k}\chi_{k,i} \end{cases} \tag{7}$$

Observe that (7) is able to implement information exchange in the gradient step and incorporate the regularization parameter in the adaptation step similar to the sparse diffusion LMS developed in [27], in contrast also with the approach in [30].

3. PERFORMANCE ANALYSIS

In this section, we consider the estimates $w_{k,i}$ as realizations of a random process $\boldsymbol{w}_{k,i}$ and determine the performance of algorithm (7) in terms of its mean-square behavior. To proceed, we introduce the error vectors $\widetilde{\boldsymbol{w}}_{k,i} = w^{\circ} - \boldsymbol{w}_{k,i}$, $\widetilde{\boldsymbol{\psi}}_{k,i} = w^{\circ} - \boldsymbol{\psi}_{k,i}$, $\widetilde{\boldsymbol{\chi}}_{k,i} = w^{\circ} - \boldsymbol{\chi}_{k,i}$ and the global quantities

$$\boldsymbol{w}_i = \mathrm{col}\{\boldsymbol{w}_{1,i}, \cdots, \boldsymbol{w}_{N,i}\}, \quad \boldsymbol{\psi}_i = \mathrm{col}\{\boldsymbol{\psi}_{1,i}, \cdots, \boldsymbol{\psi}_{N,i}\},$$

$$\widetilde{\boldsymbol{w}}_i = \begin{bmatrix} \widetilde{\boldsymbol{w}}_{1,i} \\ \vdots \\ \widetilde{\boldsymbol{w}}_{N,i} \end{bmatrix}, \quad \widetilde{\boldsymbol{\psi}}_i = \begin{bmatrix} \widetilde{\boldsymbol{\psi}}_{1,i} \\ \vdots \\ \widetilde{\boldsymbol{\psi}}_{N,i} \end{bmatrix}, \quad \widetilde{\boldsymbol{\chi}}_i = \begin{bmatrix} \widetilde{\boldsymbol{\chi}}_{1,i} \\ \vdots \\ \widetilde{\boldsymbol{\chi}}_{N,i} \end{bmatrix}.$$

Furthermore, we consider the block diagonal matrix

$$\mathcal{M} = \mathrm{diag}\{\mu_1 I_M, \ldots, \mu_N I_M\}$$

and the extended block weighting matrices $\mathcal{C} = C \otimes I_M$ and $\mathcal{A} = A \otimes I_M$, where $\otimes$ denotes the Kronecker product operation. We also introduce the quantities

$$\boldsymbol{D}_i = \mathrm{diag}\Big\{ \sum_{l \in \mathcal{N}_1} c_{l1} \boldsymbol{u}_{l,i}^T \boldsymbol{u}_{l,i}, \ldots, \sum_{l \in \mathcal{N}_N} c_{lN} \boldsymbol{u}_{l,i}^T \boldsymbol{u}_{l,i} \Big\}$$

$$\boldsymbol{g}_i = \mathcal{C}^T \mathrm{col}\{\boldsymbol{u}_{1,i}^T \boldsymbol{v}_1(i), \ldots, \boldsymbol{u}_{N,i}^T \boldsymbol{v}_N(i)\}$$

We then conclude that the error vectors satisfy the recursion

$$\widetilde{\boldsymbol{w}}_i = \mathcal{A}^T [I - \mathcal{M}\boldsymbol{D}_i] \widetilde{\boldsymbol{w}}_{i-1} - \mathcal{A}^T \mathcal{M} \boldsymbol{g}_i + \gamma \mathcal{A}^T \mathcal{M} \mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i), \quad (8)$$

where $\mathcal{P}_\kappa$ is a clipping or limiter function defined by $\mathcal{P}_\kappa(x) := (|x+\kappa| - |x-\kappa|)/(2\kappa)$ that is applied component-wise, i.e., $\mathcal{P}_\kappa(\boldsymbol{x}_i) := \mathrm{col}\{\mathcal{P}_\kappa(\boldsymbol{x}_{1,i}), \ldots, \mathcal{P}_\kappa(\boldsymbol{x}_{N,i})\}$. This results from the fact that the soft-thresholding and the limiter function satisfy the following property: $\mathrm{soft}(x,\kappa) = x - \kappa\mathcal{P}_\kappa(x)$. In fact, operator $\mathcal{P}_\kappa$ is the projection onto the set $B_\infty(0;1) := \{x \in \mathbb{R}^M \,|\, \|x\|_\infty \le 1\}$, where $\|\cdot\|_\infty$ is the ℓ_∞ norm. We have thus shown that (7) can be expressed in a form similar to those of the subgradient-based diffusion LMS algorithms. Observe further that compared to the zero-attracting (ZA) diffusion LMS in [27], the update based on $\mathcal{P}_{\mu\gamma}$ exerts a linear attraction in the range $(-\mu\gamma, \mu\gamma)$, and that the sign function may be seen as the limiting behavior of $\mathcal{P}_{\mu\gamma}$ as $\mu\gamma$ becomes very small.

Now note that due to the nonlinearity of (4), the function $\mathcal{P}_\kappa$ in (8) is dependent on the noise sequence. We need some simplifying assumptions to make the analysis more tractable. Appealing to the nonexpansiveness of the proximity (or projection) operator [19, 39], i.e., $\|\mathrm{prox}_{\kappa R} x - \mathrm{prox}_{\kappa R} y\| \le \|x - y\|$, we see that $\|\mathcal{P}_{\mu\gamma}(\psi_{k,i}) - \mathcal{P}_{\mu\gamma}(w_{k,i-1})\| \le \mu_k \sum_{l \in \mathcal{N}_k} c_{l,k} \|u_{l,i}^T [d_l(i) - u_{l,i} w_{k,i-1}]\|$. Assuming that the step-sizes are sufficiently small and that at steady-state, $d_l(i) - u_{l,i} w_{k,i-1}$ is small for each $l \in \mathcal{N}_k$, we employ $\mathcal{P}_{\mu\gamma}(w_{k,i-1})$ instead of $\mathcal{P}_{\mu\gamma}(\psi_{k,i})$. We will make use of this simplifying assumption to obtain a more manageable steady-state analysis.

3.1. Mean Convergence

By taking the expectation of both sides of (8) and assuming the spatial and temporal independence of the regressors, we conclude that the mean-error vector evolves according to the following dynamics:

$$\mathbb{E}\widetilde{\boldsymbol{w}}_i = \mathcal{A}^T (I - \mathcal{M}\mathcal{D}) \mathbb{E}\widetilde{\boldsymbol{w}}_{i-1} + \gamma \mathcal{A}^T \mathcal{M} \mathbb{E}\mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i)$$

where $\mathcal{D} := \mathbb{E}\boldsymbol{D}_i$. Recall that a square matrix $\mathcal{X}$ is called stable if $\mathcal{X}^i \to 0$ as $i \to +\infty$. Since $\mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i)$ has bounded entries, algorithm (8) converges in the mean if $\mathcal{A}^T(I - \mathcal{M}\mathcal{D})$ is a stable matrix, and this is guaranteed if $I - \mathcal{M}\mathcal{D}$ is stable, as the entries on the columns of $\mathcal{A}^T$ add up to one and $\mathcal{M}$ is diagonal. Hence we obtain:

Theorem 3.1 *The diffusion algorithm* (7) *asymptotically converges in the mean for any initial condition if the step-sizes satisfy:*

$$0 < \mu_k < \frac{2}{\lambda_{\max}(\sum_{l=1}^N c_{lk} R_{u,l})}, \qquad k = 1, \ldots, N \quad (9)$$

where $\lambda_{\max}(X)$ denotes the maximum eigenvalue of a symmetric positive definite matrix X. Moreover, we have the following bias:

$$\mathrm{bias} := \lim_{i \to +\infty} \mathbb{E}\widetilde{\boldsymbol{w}}_i = \gamma \mathcal{B} \mathcal{A}^T \mathcal{M} \lim_{i \to +\infty} \mathbb{E}\mathcal{P}_{\mu\gamma}(\boldsymbol{w}_{i-1}) \quad (10)$$

where $\mathcal{B} := [I - \mathcal{A}^T(I - \mathcal{M}\mathcal{D})]^{-1}$. ∎

As one may expect, the regularization parameter γ and the step-sizes $\{\mu_k\}$ affect the bias of the estimates as shown above. Thus, we need to set a small γ to minimize the bias. These properties were also observed for the sparse diffusion LMS developed in [27].

3.2. Mean-Square Convergence

Here we investigate the behavior of the steady-state mean-square deviation (MSD) at each node k. Under the energy conservation framework [7, 27] and using the independence assumptions, we establish the following variance relation for (7):

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\Sigma^2 = \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 + \mathbb{E}[\boldsymbol{g}_i^T \mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{g}_i] + \Delta_{\Sigma,i}(\gamma) \quad (11)$$

where Σ is any symmetric positive definite matrix that we are free to choose, with $\Sigma' := \mathbb{E}(I - \boldsymbol{D}_i\mathcal{M})\mathcal{A}\Sigma\mathcal{A}^T(I - \mathcal{M}\boldsymbol{D}_i)$ and $\Delta_{\Sigma,i}(\gamma) := \gamma^2 \eta_{\Sigma,i} - \gamma\zeta_{\Sigma,i}$ for some terms $\zeta_{\Sigma,i}$ and $\eta_{\Sigma,i}$ given by

$$\begin{aligned} \eta_{\Sigma,i} &:= \mathbb{E}\|\mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i)\|_{\mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}}^2 \ge 0 \\ \zeta_{\Sigma,i} &:= 2\mathcal{M}\mathbb{E}\mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i)\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{g}_i \\ &\quad - 2\mathbb{E}\mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i)^T\mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T[I - \mathcal{M}\boldsymbol{D}_i]\widetilde{\boldsymbol{w}}_{i-1} \end{aligned}$$

Note that we have used an assumption that $\boldsymbol{v}_k(i)$ is independent of $\boldsymbol{\psi}_{k,i}$ for each k, which is quite reasonable at steady-state. We may also consider using our simplifying approximation on $\mathcal{P}_{\mu\gamma}(\boldsymbol{\psi}_i)$ to obtain an estimate for the steady-state behavior. Now setting $\mathcal{G} = \mathbb{E}[\boldsymbol{g}_i\boldsymbol{g}_i^T]$, we can rewrite the variance relation (11) in the form

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\Sigma^2 = \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 + \mathrm{Tr}[\Sigma\mathcal{A}^T\mathcal{M}\mathcal{G}\mathcal{M}\mathcal{A}] + \Delta_{\Sigma,i}(\gamma) \quad (12)$$

where $\mathrm{Tr}(\cdot)$ denotes the trace operator. We turn to the vectorized form to simplify (11). Let $\sigma = \mathrm{vec}(\Sigma)$ and $\sigma' = \mathrm{vec}(\Sigma')$, where the vec operator stacks the columns of Σ on top of each other. We now use the notations $\|w\|_\sigma^2$ and $\|w\|_\Sigma^2$ to denote the same quantity. Using the Kronecker product property $\mathrm{vec}(U\Sigma V) = (V^T \otimes U)\mathrm{vec}(\Sigma)$, we can vectorize Σ' and replace it by the simpler linear vector relation $\sigma' = \mathcal{F}\sigma$ where the matrix $\mathcal{F}$ is given by

$$\begin{aligned} \mathcal{F} = (I \otimes I)\{&I - I \otimes (\mathcal{D}\mathcal{M}) - (\mathcal{D}^T\mathcal{M}) \otimes I \\ &+ \mathbb{E}[(\boldsymbol{D}_i^T\mathcal{M}) \otimes (\boldsymbol{D}_i\mathcal{M})]\}(\mathcal{A} \otimes \mathcal{A}). \end{aligned} \quad (13)$$

Using the property $\mathrm{Tr}(\Sigma X) = \mathrm{vec}(X^T)^T\sigma$, we can rewrite (12) as:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\sigma^2 = \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\mathcal{F}\sigma}^2 + [\mathrm{vec}(\mathcal{A}^T\mathcal{M}\mathcal{G}^T\mathcal{M}\mathcal{A})]^T\sigma + \Delta_{\Sigma,i}(\gamma) \quad (14)$$

Similar to [27], we have the following mean-square stability guarantee for the diffusion forward-backward algorithm:

Theorem 3.2 *For any initial condition, the diffusion algorithm* (7) *asymptotically converges in the mean-square sense if the step-sizes are chosen such that they satisfy* (9) *and that the matrix* $\mathcal{F}$ *given by* (13) *is stable.* ∎

Now assuming the matrix $I - \mathcal{F}$ is invertible, we obtain the following from (14) in steady-state:

$$\lim_{i\to+\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2_{(I-\mathcal{F})\sigma} = [\mathrm{vec}(\mathcal{A}^T\mathcal{M}\mathcal{G}^T\mathcal{M}\mathcal{A})]^T\sigma + \Delta_\Sigma \qquad (15)$$

where $\Delta_\Sigma := \lim_{i\to+\infty} \Delta_{\Sigma,i}(\gamma)$. The steady-state expression (15) allows us to derive the MSD through the proper selection of the free weighting parameter σ or Σ [7, 27]. For instance, if we define the steady-state MSD at node k by $\mathrm{MSD}_k := \lim_{i\to+\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, we can compute it by considering a weighting using a block matrix Q_k that has an identity matrix at block (k,k) and zeros elsewhere, with vectorized form $q_k = \mathrm{vec}(\mathrm{diag}(e_k) \otimes I_M)$ where e_k being the column vector with unit entry at position k and zeros elsewhere. Thus, choosing $\sigma_k = (I - \mathcal{F})^{-1}q_k$, we obtain

$$\mathrm{MSD}_k = [\mathrm{vec}(\mathcal{A}^T\mathcal{M}\mathcal{G}^T\mathcal{M}\mathcal{A})]^T(I - \mathcal{F})^{-1}q_k + \Delta_{\Sigma_k} \qquad (16)$$

Note that the expressions above revert to those obtained for the diffusion LMS when $\gamma = 0$. Now by treating $\mathcal{P}_{\mu\gamma}$ as the proximity operator of the indicator function on the set $B_\infty(0;1)$, we may express it in terms of a subgradient [22] and by using similar arguments in [27], we can show that by selecting some appropriate value for $\mu\gamma$, the term Δ_{Σ_k} will be negative if w° is sparse. Hence, the proximal diffusion algorithm will have better MSD compared to diffusion LMS. However, this will not hold in general if w° is not sparse, and therefore in such cases, the proximal diffusion algorithm is expected to perform worse compared to the standard diffusion algorithm.

4. SIMULATION RESULTS

We now present simulation results to demonstrate the advantage of the designed diffusion forward-backward method compared to its subgradient-based sparse diffusion LMS formulation in [27]. We consider the proximal diffusion algorithm based on the ℓ_1 norm and its counterpart diffusion ZA-LMS, and we also use their reweighted versions, which for the subgradient-based algorithm leads to what is called the diffusion RZA-LMS algorithm. We compare these algorithms by examining their learning curves at steady-state.

We consider a connected network with $N = 20$ nodes. The regressors are of length $M = 100$, zero-mean Gaussian, and spatially and temporally independent. The background white noise power is randomly set to $\sigma^2_{v,k} \in (0.01, 0.1)$ for each node k. The unknown sparse vector has two nonzero entries that are chosen randomly, with values between 0 and 1. We use a uniform step-size of $\mu = 0.1$ for all simulations. On the other hand, the regularization γ is set as follows: for the ℓ_1 norm-based, $\gamma = 0.1$ and for the reweighted version, $\gamma = 0.001$ with $\varrho = 0.01$. We also use these parameters for the corresponding sparse diffusion LMS algorithms. Furthermore, we consider the relative degree rule [7] as our combination strategy and we do not consider any measurement exchange, i.e., $C = I$. The expectation is calculated by averaging 200 independent experiments.

Figure 1 shows the learning behavior of each algorithm in terms of the network MSD, which is defined as the average MSD across all nodes in the network. Observe that the proposed algorithms perform better than their diffusion LMS counterparts at steady-state, and this outcome is true throughout the network as seen in Figure 2. Notice that due to reweighting, the diffusion RZA-LMS reaches a lower steady-state value compared to the diffusion forward-backward algorithm, but using the same reweighting technique for the diffusion forward-backward algorithm results to a better effect.

We also considered the case where the system is completely non-sparse. As shown in Figure 3, the standard diffusion LMS outperforms the diffusion forward-backward and ZA-LMS, but the reweighted versions were able to perform comparably to the diffusion LMS. Moreover, in both cases, we see that the proximal diffusion algorithms outperform their subgradient-based counterparts. Thus, as in [27], we have shown that the reweighted algorithm yields a slight performance loss only on the case where the system is completely non-sparse.

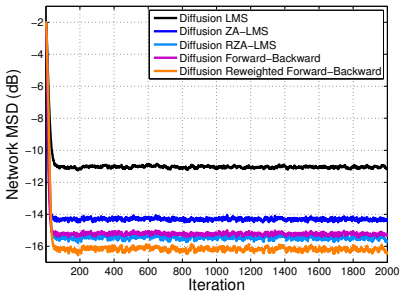

Fig. 1. Network MSD

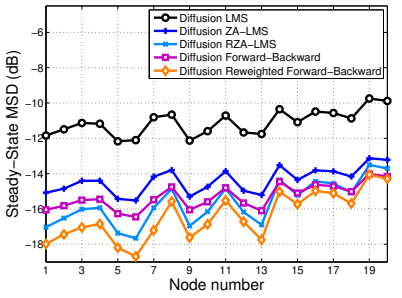

Fig. 2. Node MSD

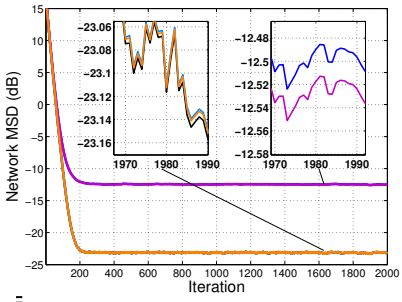

Fig. 3. Network MSD of non-sparse system

5. CONCLUSIONS

We proposed a proximal splitting approach to regularized distributed estimation over adaptive networks. A diffusion implementation of the proximal forward-backward splitting method was developed, and by exploiting properties of the proximity operator, we showed that the mean-square performance analysis can be handled in a way similar to that of subgradient-based algorithms. Simulations demonstrated the advantage of using the proposed method compared to other diffusion LMS algorithms in detecting sparsity.

6. REFERENCES

[1] 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, Jul. 2008.

[2] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4795–4810, Sep. 2010.

[3] J. Li and A. H. Sayed, "Modeling bee swarming behavior through diffusion adaptation with asymmetric information sharing," *EURASIP J. Adv. Signal Process.*, no. 1, p. 18, Jan. 2012.

[4] S.-Y. Tu and A. H. Sayed, "Foraging behavior of fish schools via diffusion adaptation," *Proc. 2nd Int. Workshop on Cognitive Inform. Process.*, pp. 63–68, Jun. 2010.

[5] J. Chen, X. Zhao, and A. H. Sayed, "Bacterial motility via diffusion adaptation," in *Proc. 44th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2010, pp. 1930–1934.

[6] F. S. Cattivelli and A. H. Sayed, "Modeling bird flight formations using diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2038–2051, May 2011.

[7] ——, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[8] A. H. Sayed, "Diffusion Adaptation over Networks," in *E-Reference Signal Processing*, R. Chellappa and S. Theodoridis, Eds. New York: Elsevier, 2013. [Online]. Available: http://arxiv.org/abs/1205.4220

[9] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

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

[11] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Royal Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[12] Y. Chen, Y. Gu, and A. O. Hero, "Sparse LMS for system identification," in *Proc. IEEE ICASSP*, 2009, pp. 3125–3128.

[13] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online adaptive estimation of sparse signals: where RLS meets the ℓ_1-norm," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3436–3447, Jul. 2010.

[14] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, Aug. 2010.

[15] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online sparse system identification and signal reconstruction using projections onto weighted ℓ_1 balls," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 936–952, Mar. 2011.

[16] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space," *J. Math. Anal. Appl.*, vol. 72, pp. 383–390, 1979.

[17] D. Gabay, "Applications of the method of multipliers to variational inequalities," in *Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems*, North Holland, Amsterdam, 1983.

[18] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control Optim.*, vol. 29, pp. 119–138, 1991.

[19] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[20] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci. Paris Sér. A Math.*, vol. 255, pp. 2897–2899, 1962.

[21] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[22] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program.*, vol. 129, no. 2, pp. 163–195, Jun. 2011.

[23] A. Beck, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[24] M. Yamagishi, M. Yukawa, and I. Yamada, "Acceleration of adaptive proximal forward-backward splitting method and its application to sparse system identification," in *Proc. IEEE ICASSP*, May 2011, pp. 4296–4299.

[25] M. Yamagishi and I. Yamada, "Over-relaxation of the fast iterative shrinkage-thresholding algorithm with variable stepsize," *Inverse Problems*, vol. 27, no. 10, pp. 1–15, Oct. 2011.

[26] A. H. Sayed, *Fundamentals of Adaptive Filtering*. New York: Wiley, 2003.

[27] P. Di Lorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, Mar. 2013.

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

[29] N. Takahashi and I. Yamada, "Link probability control for probabilistic diffusion least-mean squares over resource-constrained networks," in *Proc. IEEE ICASSP*, 2010, pp. 3518–3521.

[30] Y. Liu, C. Li, and Z. Zhang, "Diffusion sparse least-mean squares over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4480–4485, Aug. 2012.

[31] R. L. G. Cavalcante, I. Yamada, and B. Mulgrew, "An adaptive projected subgradient approach to learning in diffusion networks," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2762–2774, Jul. 2009.

[32] 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.

[33] S. Chouvardas, K. Slavakis, Y. Kopsinis, and S. Theodoridis, "A sparsity promoting adaptive algorithm for distributed learning," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5412–5425, Oct. 2012.

[34] Y. Murakami, M. Yamagishi, M. Yukawa, and I. Yamada, "A sparse adaptive filtering using time-varying soft-thresholding techniques," in *Proc. IEEE ICASSP*, 2010, pp. 3734–3737.

[35] I. Yamada and N. Ogura, "Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions," *Numer. Funct. Anal. Optim.*, vol. 25, no. 7-8, pp. 593–617, 2004.

[36] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 877–905, Oct. 2008.

[37] Y. Chen, Y. Gu, and A. O. Hero, "Regularized least-mean-square algorithms," 2010. [Online]. Available: http://arxiv.org/abs/1012.5066

[38] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

[39] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, 1979.