Sparse System Identification by Exponentially Weighted Adaptive Parallel Projection and Generalized Soft-Thresholding

Masao Yamagishi∗, Masahiro Yokawa†, and Isao Yamada∗

∗Tokyo Institute of Technology, 2-12-1-S3-60 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
E-mail: {myamagi, isao}@sp.ss.titech.ac.jp Tel: +81-3-5734-2503
†Niigata University, 2-8050 Ikarashi, Nishi-ku, Niigata, 950-2181, Japan
E-mail: yokawa@eng.niigata-u.ac.jp Tel: +81-25-262-7549

Abstract—In this paper, we propose a novel online scheme named the adaptive proximal forward-backward splitting method to suppress the sum of ‘smooth’ and ‘nonsmooth’ convex functions, both of which are assumed time-varying. We derive a powerful algorithm for sparse system identification by defining each function as a certain average squared distance (‘smooth’) and a weighted ℓ1-norm (‘nonsmooth’). The smooth term brings an exponentially weighted average of the metric projections of the current estimate onto linear varieties, of which the number grows as new measurements arrive. The presented recursive formula realizes an efficient computation of the average (the exponentially weighted adaptive parallel projection), which contributes the fast and stable convergence. The nonsmooth term, on the other hand, brings the weighted soft-thresholding, contributing the enhancement of the filter sparsity. The weights are adaptively controlled according to the filter coefficients so that the soft-thresholding gives significant impacts solely to inactive coefficients (coefficients close to zero). The numerical example demonstrates the efficacy of the proposed algorithm.

I. INTRODUCTION

Recently, there are increasing interests in developing adaptive filtering algorithms which exploit the sparsity of an unknown system to be estimated. The so-called variable-metric projection family of algorithms [1], including the proportionate NLMS/APA [2], [3], have been developed to promote effectively the sparsity of the estimate. On the other hand, the soft-thresholding technique has been proposed for a denoising problem in [4], and it is also known to enhance the sparsity of the estimate. Indeed, the soft-thresholding can be seen as an example of the so-called proximity operator of the ℓ1-norm function.

In [5], the authors have proposed a natural strategy bridging the above two techniques (i.e., variable-metric projection and soft-thresholding). Its idea is the use of proximal forward-backward splitting [6], [7] to suppress the sum of two functions: (i) a certain average of the squared distances to a closed convex set (‘smooth’) and (ii) a weighted ℓ1-norm function (‘nonsmooth’). The time-varying inner product used extensively in the variable metric projection is employed in the derivation of the algorithm. It has been shown that the simple weighted soft-thresholding techniques derived naturally by the proximal forward-backward splitting can significantly improve the convergence of the variable metric projection type schemes. There is still room for further improvement by taking a more reasonable way to average the squared distances.

In this paper, we propose an online scheme named the adaptive proximal forward-backward splitting method to suppress the sum of ‘smooth’ and ‘nonsmooth’ convex functions, both of which are assumed time-varying. We derive a powerful algorithm for sparse system identification by defining each function as an exponentially-weighted average of the squared distances (‘smooth’) and a weighted ℓ1-norm (‘nonsmooth’). The smooth term brings the exponentially weighted adaptive parallel projection, an exponentially-weighted average of the metric projections of the current estimate onto linear varieties, of which the number grows as new measurements arrive. Since every linear variety defined at each iteration is exploited over the following iteration steps (with a forgetting factor), the exponentially weighted adaptive parallel projection is expected to be effective particularly when the unknown system is changing slowly. For efficient computation of the exponentially-weighted average of a growing number of projections, we fix the metric to the Euclidean and present a recursive formula: the computational complexity at each iteration keeps constant although the number of projections to be computed is growing. (The idea of the recursive formula for hyperplanes, i.e., a special case of the current study, was originally introduced in [8], [9] in the framework of the adaptive projected subgradient method [12].) The nonsmooth term, on the other hand, brings the weighted soft-thresholding, contributing the enhancement of the filter sparsity. The key idea is the adaptive control of the weights according to the filter coefficients so that the soft-thresholding gives significant impacts solely to inactive coefficients (coefficients close to zero). Along this idea, we present three examples of the adaptive weight controlling, two of which are motivated by the ℓ0 norm constraint LMS (ℓ0-LMS) [14] and the reweighted zero-attracting (RZA) LMS [15], respectively. The numerical example shows the efficacy of the proposed algorithm in the sparse system identification problem.

II. SPARSE ADAPTIVE FILTERING PROBLEM

Let ℝ and ℕ denote the sets of all real numbers and nonnegative integers, respectively. Denote the set ℕ\{0} by ℕ* and transposition of a matrix or a vector by (·)T. Suppose that we observe the output sequence (d(k))k∈ℕ ⊂ ℝ (i.e., d(k) ∈ ℝ, ∀k ∈ ℕ) that obeys the following model (see Fig. 1):

d(k) = u∗(k)h∗ + nk,

where k ∈ ℕ denotes the time index, N ∈ ℕ* the tap length, u∗ := [u∗(0), ..., u∗(N − 1)]T ∈ ℝN a known vector defined with the input sequence (u(k))k∈ℕ ∋ ℝ, h∗ ∈ ℝN the unknown system to be estimated (e.g., echo impulse response), and nk ∈ ℝ the noise process. Throughout this paper we consider h∗ ∈ ℝN to be sparse, i.e., few coefficients are significantly different from zero (active coefficients) and many coefficients are zero or near-zero (inactive coefficients).
Algorithm 1: For an arbitrarily chosen \( \mathbf{h}_0 \in \mathbb{R}^N \), generate a sequence \((\mathbf{h}_k)_{k \in \mathbb{N}} \subset \mathbb{R}^N \) by

\[
\mathbf{h}_{k+1} := \text{prox}_{\frac{\mu_k}{\psi_k}}^{\frac{1}{\mu_k} \psi_k} \left( \mathbf{h}_k - \frac{\mu_k}{L_k} \nabla (\psi_k \varphi_k) (\mathbf{h}_k) \right) \tag{2}
\]

where \((\mu_k)_{k \in \mathbb{N}} \subset (0, 2)\) is the step-size and

\[
\text{prox}_{\frac{\mu_k}{\psi_k}}^{\frac{1}{\mu_k} \psi_k} (x) := \arg \min_{y \in \mathbb{R}^N} \left( \psi_k(y) + \frac{L_k}{2\mu_k} \| x - y \|_2^2 \right), \forall x \in \mathbb{R}^N
\]

is called the proximity operator of \( \psi_k \) of index \( \mu_k/L_k > 0 \) (with respect to the norm \( \| \cdot \|_{Q_k} \)).

Note that this algorithm is a time-variation extension of the proximal-forward-backward splitting method [6], [7] and Algorithm 1 is identical to the original method if \( \varphi_k, \psi_k \) and \( Q_k \) are time-invariant. Algorithm 1 satisfies the (strictly) monotone approximation property:

\[
\| \mathbf{h}_{k+1} - \mathbf{h}_k \|_{Q_k} < \| \mathbf{h}_k - \mathbf{h}_0 \|_{Q_k}, \tag{3}
\]

for every \( \mathbf{h}_0 \in \Omega := \arg \min_{\mathbf{h} \in \mathbb{R}^N} \varphi_k(\mathbf{h}), \) if \( \Omega \neq \emptyset \) and \( \mathbf{h}_k \in \Omega \).

For better understanding of Algorithm 1, we present some examples of \( \varphi_k \) and \( \psi_k \) which have been studied in [5].

Example 1 (Weighted squared distance for \( \varphi_k \)): Define the smooth part \( \varphi_k(\mathbf{h}) \) of the objective function \( \Theta_k \) as follows:

\[
\varphi_k(\mathbf{h}) := \frac{1}{2} \sum_{i \in I_k} w_i^{(k)} \Delta Q_k^2 (\mathbf{h}, S_i). \tag{4}
\]

Here, \((S_k)_{k \in \mathbb{N}} \) is a sequence of closed convex sets, each of which has its elements consistent with the data available at time instant \( k, \) \( I_k := \{0, 1, \ldots, q - 1, k \} \) is the index of the closed convex sets, \( w_i^{(k)} \in (0, 1], i \in I_k, \) the weights satisfying \( \sum_{i \in I_k} w_i^{(k)} = 1, \) and \( d_{Q_k}(x; C) := \min_{y \in C} \| x - y \|_{Q_k} \) is the distance between an arbitrary point \( x \in \mathbb{R}^N \) and a closed convex set \( C \subset \mathbb{R}^N \). In this case, the Lipschitz constant of the gradient of \( \varphi_k \) is one. The update equation (2) then becomes the following:

\[
\mathbf{h}_{k+1} := \text{prox}_{\frac{\mu_k}{\psi_k}}^{\frac{1}{\mu_k} \psi_k} \left( \mathbf{h}_k + \mu_k \sum_{i \in I_k} w_i^{(k)} \left( P_{S_i}^{Q_k} (\mathbf{h}_k) - \mathbf{h}_k \right) \right) \tag{5}
\]

where \((\mu_k)_{k \in \mathbb{N}} \subset (0, 2)\) is the step-size and the metric projection of \( x \in \mathbb{R}^N \) onto a closed convex set \( C \subset \mathbb{R}^N \) is defined by \( P_{S_i}^{Q_k}(x) := \arg \min_{y \in C} \| x - y \|_{Q_k} \). The algorithm (5) covers many existing algorithms as its special cases (see Example 2 below and [5]).

Example 2 (Squared distance to linear variety for \( \varphi_k \)): Let \( I_k := \{0\}, \) \( S_k := \arg \min_{\mathbf{h} \in \mathbb{R}^N} \| \Delta \mathbf{h} \|_r \) in (5), where we use \( \| \cdot \|_r \) for the standard Euclidean norm on \( \mathbb{R}^r \). Moreover, we restrict \( Q_k \) to a diagonal matrix, i.e., \( Q_k := \text{diag}(q_{1}^{(k)}, q_{2}^{(k)}, \ldots, q_{n}^{(k)}) \) (Note that many of the variable metric projection type methods, such as the proportionate NLMS/APA [2], [3], utilize diagonal matrices (see [1])). Then we have

\[
\mathbf{h}_{k+1} := \text{prox}_{\frac{\mu_k}{\psi_k}}^{\frac{1}{\mu_k} \psi_k} \left( \mathbf{h}_k - \mu_k Q_k^{-1} \mathbf{U}_k \Gamma_k^{-1} \Delta \mathbf{h}_k \right) \tag{6}
\]

A set \( C \subset \mathbb{R}^N \) is said to be convex if \( \alpha x + (1 - \alpha) y \in C \) for \( x, y \in C \) and \( \alpha \in (0, 1). \)
where $\Gamma_k := U_k^T \Sigma_k^{-1} U_k + \delta I$. The regularization parameter $\delta \geq 0$ is introduced for numerical stability. The algorithm (6) significantly extends the standard NLMS/APA [10], [11] and the proportionate NLMS/APA [2], [3]. The APA, for instance, is reproduced by setting $Q_k := I$ and $\psi_k := 0$, where $I$ is the identity matrix.

Example 3 (Weighted $\ell_1$-norm for $\psi_k$): In order to exploit the sparsity of the unknown system, define $\psi_k$ as below:

$$
\psi_k(h) := \lambda \sum_{i=1}^{N} \omega_i^{(k)} |h_i|, \quad h := [h_1, h_2, \ldots, h_N]^T \in \mathbb{R}^N,
$$

where $\lambda > 0$ is the regularization parameter, and $\omega_i^{(k)} > 0$, $i \in \{1, 2, \ldots, N\}$, the weights of the $\ell_1$ norm defined with available knowledge. We restrict $Q_k$ to a diagonal matrix, i.e., $Q_k := \text{diag} \{q_1^{(k)}, q_2^{(k)}, \ldots, q_N^{(k)}\}$. Then the proximity operator of $\psi_k$ becomes

$$
\text{prox}_{\mu \psi_k}(h) = \sum_{i=1}^{N} \text{sgn}(h_i) \max \left\{ |h_i| - \frac{\mu \lambda \omega_i^{(k)}}{q_i^{(k)}}, 0 \right\} e_i,
$$

where $\text{sgn}(\cdot)$ is the signum function defined by $\text{sgn}(x) := x/|x|$ if $x \neq 0$, $\text{sgn}(x) := 0$ otherwise, for all $x \in \mathbb{R}$, and $\{e_i\}_{i=1}^{N}$ is the standard orthonormal basis of $\mathbb{R}^N$, $i \in \{1, 2, \ldots, N\}$, with the value 1 assigned to its $i$th position. We call the operator in (8) Adaptively Weighted Soft-Thresholding (AWST). Intuitively, AWST cuts off the components having smaller absolute values than the threshold $\mu \lambda \omega_i^{(k)}/q_i^{(k)}$.

We finally mention that a different way of using a weighted $\ell_1$-norm for sparse system identification was recently proposed [13], which is based on the use of the projection onto a weighted $\ell_1$-norm constraint-set in the frame of the adaptive projected subgradient method [12].

IV. EXponentially WIGHTED ADAPTIVE PARALLEL PROJECTION WITH AWST

We present the proposed algorithm in the frame of the adaptive proximal forward-backward splitting method. We derive the algorithm by employing exponentially decaying weights for $w_i^{(k)}$ in (5). To realize an efficient recursive formula to compute the exponentially weighted average of a growing number of projections, we fix the metric to the standard Euclidean distance, i.e., $Q_k := I$, $\forall k \in \mathbb{N}$. The proposed algorithm is the composition of AWST and the exponentially weighted adaptive parallel projection as shown below.

Algorithm 2: In (5), let $g_k := \{0, 1, \ldots, k\}$, $S_k := \text{arg min}_{h \in \mathbb{R}^N} \|h_k(h)\|_F$, and the weights of the time-varying smooth convex term $\psi_k$ are selected as $w_i^{(k)} := \frac{1}{\lambda} e_i^{(k)}$, where $\mathcal{I}_k := \{i = 0, \ldots, k\}$ for all $i \in \mathcal{I}_k$, with $v \in [0, 1]$, $(0^0 = 1$ is applied for special selection $v = 0$). Then we obtain the following algorithm:

$$
h_{k+1} := \text{prox}_{\mu \psi_k}(h_k + \mu \sum_{i=0}^{k} \Gamma_k^{-1} U_i e_i(h_k)),
$$

where $\Gamma_k := U_k^T U_k + \delta I$. The regularization parameter $\delta \geq 0$ is introduced for numerical stability. This algorithm has the recursive form as shown below.

1) Initialization: select arbitrary $h_0 \in \mathbb{R}^N$ and let $\mathcal{Y}_0 = 0 \in \mathbb{R}$, $f_0 = 0 \in \mathbb{R}^N$ and $D_0 = \mathcal{O} \in \mathbb{R}^{N \times N}$.

2) At each time $k \in \mathbb{N}$,

$$
\begin{align*}
\mathcal{Y}_k &:= v \mathcal{Y}_{k-1} + 1, \\
f_k &:= v f_{k-1} + u_k \Gamma_k^{-1} d_k, \\
D_k &:= v D_{k-1} + U_k \Gamma_k^{-1} U_k^T,
\end{align*}
$$

$$
h_{k+1} := \text{prox}_{\mu \psi_k}(h_k + \mu \sum_{i=0}^{k} (f_i - D_i h_k)).
$$

It should be mentioned that the basic idea of the recursive formula for the special case of $r = 1$ was presented in [8], [9].

Now we turn our attention to the designs of $\omega_i^{(k)}$ in the proximity operator in (8). The threshold $\mu \lambda \omega_i^{(k)}/q_i^{(k)}$ is desired to be small for active coefficients since these coefficients should never shrink to zero. Hence the weight $\omega_i^{(k)}$ for each active coefficient should be small, leading to the idea of controlling the weight $\omega_i^{(k)}$ adaptively as a function of $h_i$ (the $i$th component of $h_k$). We thus let $\omega_i^{(k)} := \nu(h_i)$ with $\nu : \mathbb{R} \to (0, \infty)$ and present below three examples of the function $\nu$ (see Fig. 2).

Example 4:

1) (Weight design in [5]) Define

$$
\nu(x) := \begin{cases} 
\epsilon, & \text{if } |x| > \tau, \forall x \in \mathbb{R}, \\
1, & \text{otherwise},
\end{cases}
$$

where $\epsilon \approx 0$ is a small positive constant, and $\tau > 0$ is the thresholding parameter for the selection of active coefficients.

2) (Weight design motivated by $\ell_0$-LMS$^5$ [14]) Define

$$
\nu_{\ell_0}(x) := e^{-\beta_{\ell_0} |x|}, \forall x \in \mathbb{R},
$$

where $\beta_{\ell_0} > 0$ is a large positive constant.

3) (Weight design motivated by RZA-LMS$^5$ [15]) Define

$$
\nu_{\text{RZA}}(x) := \frac{1}{1 + \beta_{\text{RZA}} |x|}, \forall x \in \mathbb{R}
$$

where $\beta_{\text{RZA}} > 0$ is a large positive constant.

There are many ways to design the parameters $\tau, \beta_{\ell_0}$ and $\beta_{\text{RZA}}$: for example, we may design the parameters based on noise statistics such as the variance. Examples 4.2 and 4.3 are motivated by the following fact. A subgradient of the weighted $\ell_1$-norm $\psi_k$ in (7) is given by $\lambda \sum_{i=1}^{N} \omega_i^{(k)} \text{sgn}(h_i) e_i \in \partial \psi_k(h) := \{v \in \mathbb{R}^N | \langle v, h - h \rangle_I + \psi_k(h) \leq \psi_k(h), \forall h \in \mathbb{R}^N\}$. The third term of the $\ell_0$-LMS (or RZA-LMS) algorithm

$^5\ell_0$-LMS is described by the following equation:

$$
h_{k+1} := h_k + \mu u_k e_k(h_k) - \lambda \sum_{i=1}^{N} \text{sgn}(h_i^{(k)}) e_i,
$$

where $\mu > 0$ is the step-size and $\beta_{\ell_0}$ is a positive constant.

$^6$RZA-LMS is described by the following equation:

$$
h_{k+1} := h_k + \mu u_k e_k(h_k) - \lambda \sum_{i=1}^{N} \frac{\text{sgn}(h_i^{(k)}) e_i}{1 + \beta_{\text{RZA}} |h_i^{(k)}|},
$$

where $\mu > 0$ is the step-size and $\beta_{\text{RZA}}$ is a positive constant.
is obtained by substituting $\omega^{(k)}_{i} = \nu_{h_{0}}(h_{i})$ (or $\omega^{(k)}_{i} = \nu_{RZA}(h_{i})$) into the subgradient.

The use of these three designs robustifies active coefficients against the effect of AWST. In particular, even though $h^{*}$ is dispersive, AWST does not adversely affect the algorithm performance because most weights $\omega^{(k)}_{i}$ become zero approximately. Note that the computational complexity of AWST is relatively small, because it requires $O(N)$ multiplications at most.

V. NUMERICAL EXAMPLE

We examine the efficacy of Algorithm 2 with AWST in the context of an echo cancellation problem. We use the sparse echo impulse response $h^{*}$ of length $N = 512$ initialized according to ITU-T G.168 [16] (in this case, $h^{*}$ has only 64 non-zero components). The input signal $u_{k}$ is generated according to a first order autoregressive (AR(1)) process: $u_{k} = \rho u_{k-1} + w_{k}$, where $w_{k} \sim \mathcal{N}(0, 1)$ and $\rho = 0.8$. The noise $w_{k}$ is zero mean white Gaussian and signal-to-noise ratio (SNR) = 30 dB, where $\text{SNR} := 10 \log_{10}(E[z_{k}^{2}]/E[w_{k}^{2}])$ with $z_{k} := u_{k}^{2}h^{*}$ ($E[.]$ denotes expectation). For all the algorithms, we set the initial vector to $h_{0} := 0 \in \mathbb{R}^{N}$. Figure 3 depicts a comparison of the algorithms in the sense of system-mismatch $\eta(h_{k}) := 10 \log_{10}(\|h^{*} - h_{k}\|^{2}_{2})$ averaged over 300 runs. The step-size for each algorithm is chosen in such a way that the convergence speed of all algorithms is the same. The regularization parameter is chosen to obtain the best results in our experiments. Comparing Proposed with EW-APP, we see that AWST notably improves the performance. Moreover, comparing Proposed with APFBS, we see that the exponentially weighted adaptive parallel projection achieves significant improvement.

Note that we observe better performance of Proposed than the other algorithms in other settings: both in the case where SNR is 15dB or 20dB and in the case where $h^{*}$ has only a few big coefficients and all other coefficients are very small (e.g., all components are 0.0313 except the first component $h^{*}_{1} := 0.707$).

VI. CONCLUSION

We have proposed a novel online scheme named the adaptive proximal forward-backward splitting method. In the frame of this scheme, we have produced the composition of the exponentially weighted adaptive parallel projection and the Adaptively Weighted Soft-Thresholding (AWST) with the adaptive control of the weight of AWST. This composition has achieved excellent performance in convergence as well as in accuracy.

ACKNOWLEDGEMENTS

This work was partially supported by SCAT (Support Center for Advanced Telecommunications) and JSPS Grants-in-Aid (B-21300091, 20760252, 0909539).

REFERENCES