A SPARSE ADAPTIVE FILTERING USING
TIME-VARYING SOFT-THRESHOLDING TECHNIQUES

Yukihiro Murakami, Masao Yamagishi, Masahiro Yukawa, and Isao Yamada

1Dept. of Communications and Integrated Systems, Tokyo Institute of Technology, Japan.
2Mathematical Neuroscience Lab., Brain Science Institute, RIKEN, Japan.
E-mail: {murakami,myamagi,isao}@comm.ss.titech.ac.jp, myukawa@riken.jp

ABSTRACT

In this paper, we propose a novel adaptive filtering algorithm based on an iterative use of (i) the proximity operator and (ii) the parallel variable-metric projection. Our time-varying cost function is a weighted sum of squared distances (in a variable-metric sense) plus a possibly nonsmooth penalty term, and the proposed algorithm is derived along the idea of proximal forward-backward splitting in convex analysis. For application to sparse-system identification problems, we employ the (weighted) $\ell_1$ norm as the penalty term, leading to a time-varying soft-thresholding operator. As the simple example of the proposed algorithm, we present the variable-metric affine projection algorithm composed with the time-varying soft-thresholding operator. Numerical examples demonstrate that the proposed algorithms notably outperform their counterparts without soft-thresholding both in convergence speed and steady-state mismatch, while the extra computational complexity due to the additional soft-thresholding is negligibly low.

Index Terms— sparse adaptive filtering, proximal forward-backward splitting, soft-thresholding, variable metric, parallel projection

1. INTRODUCTION

Underlying sparseness in the system to be estimated has been recently exploited for performance improvements of adaptive filtering algorithms [1–6]. These studies are motivated by the fact that the system is often sparse in many applications including network/acoustic echo cancellation (E/AEC) and channel estimation/equalization. In particular, the proportionate normalized least mean squares (PNLMS) [1, 7] and its further recent advancements (See, e.g., [4, 6]) are known to exhibit excellent performance with reasonably low computational complexity. On the other hand, the soft-thresholding techniques, proposed originally for de-noising [8], have been widely applied to sparse recovery in the context of inverse imaging, compressed sensing, etc. There is no reason not to use it for exploiting the sparsity in adaptive filters. The goal of this paper is to develop such an adaptive algorithm using soft-thresholding for empowering the various conventional algorithms to achieve notable performance improvements.

In [9, 10], the authors proposed the adaptive parallel variable-metric projection algorithm (APVP), which includes PNLMS (and PAPA [2, 4]) as a particular case. Indeed, PNLMS projects — in a variable-metric sense — the instantaneous estimate onto the same hyperplane as used by the standard NLMS algorithm [11] at each iteration; the metric depends on the instantaneous estimate thus it is time-varying. APVP aims at minimizing a sequence of data-dependent cost functions, and its remarkable property is the monotone approximation [12, 13] (in the sense of a reasonably designed metric [9, 10]) to the set of all points minimizing the corresponding convex cost function at each iteration. It has been shown that APVP accelerates the convergence of PNLMS with the aid of parallel projection, hence we use the idea of APVP in this paper as well.

We focus meanwhile on the interesting fact that the soft-thresholding operator is an example of the proximity operator [14, 15] specially for the (weighted) $\ell_1$ norm; the proximity operator is a nontrivial generalization of orthogonal projection. In fact, the ideal sparse measure is the $\ell_0$ norm that counts the number of nonzero components. Unfortunately, however, the minimization of the $\ell_0$ norm is inherently combinatorial, thus becomes NP-hard [16]. This complexity issue has been circumvented by introducing an $\ell_1$ penalty term in the cost function [17] because it is the convex relaxation of the $\ell_0$ norm. This is a widely accepted strategy for sparse recovery in compressed sensing.

The first contribution of this paper is to present a unified adaptive filtering scheme composed of the parallel variable-metric projection and the proximity operator, enjoying the monotone approximation property. The proposed scheme is derived by applying the proximal forward-backward splitting (PFBS) techniques [15, 18] (posed originally for a fixed cost function) to the time-varying cost function defined as the weighted sum of squared distances to data-dependent convex sets plus a continuous convex penalty term, which is possibly nonsmooth. The second contribution is to propose, as a simple sparse adaptive filtering algorithm, an efficient algorithm composed of the standard/proportionate affine projection algorithm (APA [19, 20] / PAPA [2, 4]) and a time-varying soft-thresholding; any adaptive filtering algorithm, e.g., adaptive parallel subgradient projection (adaptive-PSG) [13], can be used in place of APA. The extra computational complexity due to the additional soft-thresholding is fairly low. The numerical examples for the E/AEC problem show that each of the proposed algorithms noticeably outperforms its counterpart without soft-thresholding both in convergence speed and steady-state mismatch.

2. PRELIMINARIES

2.1. Notation

Throughout this paper, we use the following notation. Let $\mathbb{R}$ and $\mathbb{N}$ denote the sets of all real numbers and nonnegative integers, respectively, and the superscript $(\cdot)^T$ transposition. For a positive definite matrix $Q \in \mathbb{R}^{N \times N}$ ($N \in \mathbb{N} := \mathbb{N} \cup \{0\}$) (denoted as $Q \succ 0$), define an inner product and its induced (quadratic) norm by $\langle x, y \rangle_Q := x^T Q y$ ($\forall x, y \in \mathbb{R}^N$) and $\|x\|_Q := \sqrt{\langle x, x \rangle_Q}$ ($\forall x \in \mathbb{R}^N$). The quadratic-metric distance between an arbitrary point $x \in \mathbb{R}^N$ and a closed convex set $C \subset \mathbb{R}^N$ is defined as $d_Q(x, C) := \min_{y \in C} \|x - y\|_Q$ and the projection of $x \in \mathbb{R}^N$ onto $C$ is defined as $P_C^Q(x) := \arg\min_{y \in C} \|x - y\|_Q$.

Proximal Forward-Backward Splitting. Let $\varphi : \mathbb{R}^N \to \mathbb{R}$ be a smooth convex function and let $\psi : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ be a lower semicontinuous, convex, but possibly nonsmooth function. Consider the following problem:

\[
\min_{h \in \mathbb{R}^N} \varphi(h) + \psi(h). \tag{1}
\]
Here we assume that $Ω := \text{argmin}_{h \in \mathbb{R}^N} \{\varphi(h) + \psi(h)\} = \emptyset$, and the gradient $\nabla \varphi$, defined in the inner product space $(\mathbb{R}^N, \langle \cdot, \cdot \rangle_Q)$, is Lipschitz continuous with its Lipschitz constant $L > 0$, i.e.,

$$||\nabla \varphi(x) - \nabla \varphi(y)||_Q \leq L||x - y||_Q, \forall x, y \in \mathbb{R}^N.$$  

Then, for an arbitrarily chosen $h_0 \in \mathbb{R}^N$ and $\mu \in (0, 2)$, the sequence $(h_k)_{k \in \mathbb{N}}$ generated by

$$h_{k+1} := \text{prox}_{\frac{\mu}{L} \nabla \varphi} \left( I_N - \frac{\mu}{L} \nabla \varphi \right) (h_k), \quad (2)$$

converges to a solution of the problem (1), where $I_N \in \mathbb{R}^{N \times N}$ stands for the identity matrix and for any $\gamma > 0$

$$\text{prox}_{\gamma \psi} (x) := \text{argmin}_{y \in \mathbb{R}^N} \left( \psi(y) + \frac{1}{2\gamma} ||x - y||_Q^2 \right), \forall x \in \mathbb{R}^N$$

is called the proximity operator of $\psi$ of index $\gamma$. This property is verified simply by applying the nonexpansivity of

$$T := \text{prox}_{\frac{\mu}{L} \nabla \varphi} \left( I_N - \frac{\mu}{L} \nabla \varphi \right), \quad (3)$$

i.e., $||Tx - Ty||_Q \leq ||x - y||_Q, \forall x, y \in \mathbb{R}^N$ and the fact:

Fix $(T) := \{x \in \mathbb{R}^N | T(x) = x\} = Ω$, to Mann’s fixed point theorem (See for example [21]). Moreover, the averaged nonexpansivity of $T$ [12] ensures the (strictly) monotone approximation property of (2), i.e.,

$$||h_{k+1} - h_{\psi, \gamma}||_Q < ||h_k - h_{\psi, \gamma}||_Q,$$  

for any $h_{\psi, \gamma} \in Ω$, if $h_k \not\in Ω$. The structure of the operator $T$ in (3) is so-called the PFBS: proximal-forward-backward splitting for $\varphi$ and $\psi$ [15, 18].

### 2.2. Adaptive Filtering Problem

Let $k \in \mathbb{N}$ denote the time index and $N \in \mathbb{N}^*$ the tap length. With a sequence of input signals $(u_k)_{k \in \mathbb{N}}$, let $(u_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^N$ be a sequence of input vectors defined as $u_k := [u_{k-1}, \ldots, u_{k-N+1}]^T$. For $r \in \mathbb{N}^*$, define $U_k := [u_k, u_{k-1}, \ldots, u_{k-r+1}] \in \mathbb{R}^{N \times r}$ and the noise vector as $n_k := [n_{k}, n_{k-1}, \ldots, n_{k-r+1}]^T \in \mathbb{R}^r$ for all $k \in \mathbb{N}$, where $(n_k)_{k \in \mathbb{N}}$ is a sequence of an additive noise process. We assume the following linear model for the data process $(d_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^r$:

$$d_k := U_k h^* + n_k,$$

where $h^* \in \mathbb{R}^N$ stands for the system to be estimated (e.g., echo impulse response). In addition, as depicted in Fig. 1, we define the estimation residual functions $e_k : \mathbb{R}^N \to \mathbb{R}^r$, $k \in \mathbb{N}$, by

$$e_k(h) := U_k^T h - d_k.$$

The goal of the adaptive filtering problem is to estimate $h^*$ by $h_k := [h_{1,k}^*, h_{2,k}^*, \ldots, h_{N,k}^*]^T$, which is designed with $[u_k, d_k, e_k(h_{i,k-1})]_{i=1}^{N}$ for arbitrarily chosen $h_0 \in \mathbb{R}^N$.

### 3. PROPOSED ADAPTIVE ALGORITHM

#### 3.1. Adaptive Filtering Algorithm by PFBS

Let $(S_k)_{k \in \mathbb{N}}$ be a sequence of closed convex sets, each of which has its elements consistent with the data available at time instant $k$. Assume that $Q_k \succ 0$ is used to define the inner product $(\cdot, \cdot)_{Q_k}$ at time $k$. Let $I_k := \{l_1, l_2, \ldots, l_k\} \subset \mathbb{N}, l_k \in \mathbb{N}^*$, be the indexes of the closed convex sets to be processed at time $k$, and define a time-varying smooth convex function (See Example 2 in [12])

$$\varphi_k(h) := \frac{1}{2} \sum_{i \in I_k} w_i^{(k)} d_{Q_k}(h, S_i),$$

where the weights $w_i^{(k)} \in [0, 1]$, $i \in I_k$ are chosen to satisfy $\sum_{i \in I_k} w_i^{(k)} = 1$. Now the proposed time-varying cost function $\Theta_k : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ is defined by

$$\Theta_k(h) := \varphi_k(h) + \psi_k(h),$$

where $\psi_k : \mathbb{R}^N \to [0, \infty)$, which we call sparseness penalty, is a lower semicontinuous and (possibly nonsmooth) convex function.

The function $\psi_k$ is introduced as a penalty to exploit the sparseness in the learning.

Using the PFBS in (3) for $\varphi_k$ and $\psi_k$, leads to the following adaptive filtering algorithm.

**Proposed Adaptive Algorithm.** For an arbitrarily chosen $h_0 \in \mathbb{R}^N$, generate a sequence $(h_k)_{k \in \mathbb{N}}$ by

$$h_{k+1} := \text{prox}_{\frac{\mu}{L} \nabla \varphi} \left( I_N + \mu \sum_{i \in I_k} w_i^{(k)} p_i^{(k)\ast} \right) (h_k),$$

where $\mu \in (0, 2)$ is the step-size.

**Remark.** The Lipschitz constant of $\nabla \varphi_k$ is one for all $k$. The algorithm (6) for any $\mu \in (0, 2)$ satisfies time-varying version of the (strictly) monotone approximation property (See (4)):

$$||h_{k+1} - h_{\varphi_k, \psi_k}||_Q < ||h_k - h_{\varphi_k, \psi_k}||_Q,$$

for any $h_{\varphi_k, \psi_k} \in Ω_k := \text{argmin}_{h \in \mathbb{R}^N} \Theta_k$, if $h_k \not\in Ω_k \neq \emptyset$.

The proposed algorithm (6) covers many existing algorithms as its special cases. For example, if we set

$$I_k := \{k\}, S_k := \text{argmin}_{h \in \mathbb{R}^N} \|U_k^T h - d_k\|_r, \psi_k := 0,$$

the APA (in particular, NLMS if $r = 1$) is reproduced for $Q_k := I_N, k \in \mathbb{N}$. The PAPA (in particular, PNLM if $r = 1$) is reproduced for $Q_k := \delta_{\varphi_k} := \text{diag}(\{(\delta_{\varphi_k})^{(k)}_1, \ldots, (\delta_{\varphi_k})^{(k)}_N\}) \succ 0$, $k \in \mathbb{N}$, where, for small positive constants $\rho$ and $\delta_{\varphi_k}^{(k)} := \frac{\gamma_k^{(k)}}{\sum_{j=1}^{N} \gamma_j^{(k)}}$ is defined via $\gamma_k^{(k)} := \max\{\rho L_{\max}, |h_k^{(k)}|\}$ and $L_{\max} := \max\{|h_k^{(k)}|, \ldots, |h_k^{(k)}|\}$. Other possible choices for the matrix $Q_k$ are found in recent studies (See, e.g., in [4, 6]).

![Fig. 1. Adaptive filtering scheme.](image-url)
3.2. Design of Sparseness Penalty \( \psi_k \)

Let \( Q_k = \text{diag}(q_1^{(k)}, q_2^{(k)}, \ldots, q_N^{(k)}) \succ 0 \). Assume that \( h^* \in \mathbb{R}^N \) is sparse, i.e., few coefficients are significantly different from zero (active coefficients) and many coefficients are zero or near-zero (inactive coefficients). In this paper, we present following simple designs of sparseness penalty \( \psi_k \) (of course, there are many other choices for \( \psi_k \)).

**Example 1.** Let \( \psi_k \) be the \( \ell_1 \) norm function:

\[
\psi_k (h) := \lambda \| h \|_1 = \lambda \sum_{i=1}^{N} |h_i|
\]  

(7)

for \( h = [h_1, h_2, \ldots, h_N]^T \in \mathbb{R}^N \), where \( \lambda > 0 \) is the so called regularization parameter. By the result of [15], we deduce the proximity operator for \( \psi_k \) as

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

(8)

where \( \text{sgn}(\cdot) \) is the signum function and \( \{e_i\}_{i=1}^{N} \) is the standard orthonormal basis of \( \mathbb{R}^N \). Intuitively, this operator cuts off components of fairly small absolute values, where the threshold \( \mu \lambda / q_i^{(k)} \) is influenced by the time-varying matrix \( Q_k \). This is a slight extension of the standard soft-thresholding operator [8].

In (8), coefficients of large absolute values are as well attracted to zero if \( Q_k = G_k \) (see Sec. 3.1). However, these coefficients should not be easily attracted to zero if they correspond to active coefficients. Next design of \( \psi_k \) is expected to alleviate this issue.

**Example 2.** Let \( \psi_k \) be the weighted \( \ell_1 \) norm function:

\[
\psi_k (h) := \lambda \sum_{i=1}^{N} \omega_i^{(k)} |h_i|,
\]

(9)

where \( \lambda > 0 \) is the regularization parameter and the weights \( \omega_i^{(k)} (i = 1, 2, \ldots, N) \) are defined by

\[
\omega_i^{(k)} := \begin{cases} 
\epsilon, & \text{if } |h_i^{(k)}| > \tau, \\
1, & \text{otherwise},
\end{cases}
\]

(10)

with the thresholding parameter \( \tau > 0 \) and a small positive constant \( \epsilon \). The thresholding parameter is chosen to distinguish between active coefficients and inactive coefficients. The parameter \( \tau \) is desired to be set with noise statistics. In a way similar to Example 1, the proximity operator for \( \psi_k \) is given by

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

By introducing the weights \( \omega_i^{(k)} (i = 1, 2, \ldots, N) \), the coefficients of large absolute values become less sensitive to the effect of the proximity operator for \( \psi_k \). Therefore, the operator (10) is expected to exploit the sparseness more efficiently than the operator (8).

Since the operator (8) and (10) are time-varying, we call them **time-varying soft-thresholding** operator. Note that the computational complexity of the time-varying soft-thresholding operator is negligible.

Next, we present a simple example of the algorithm (6) employing the time-varying soft-thresholding.

### Table 1. Parameter settings for the proposed algorithms, where \( G_k \)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>( \tau )</th>
<th>( \lambda )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed-(a)</td>
<td>( 1 )</td>
<td>( 5.0 \times 10^{-3} )</td>
<td>( \sigma_k^2 )</td>
</tr>
<tr>
<td>Proposed-(b)</td>
<td>( 2 )</td>
<td>( 2.0 \times 10^{-4} )</td>
<td>( \sigma_k^2 )</td>
</tr>
<tr>
<td>Proposed-(c)</td>
<td>( G_k^{-1} )</td>
<td>( 1.0 \times 10^{-2} )</td>
<td>( \sigma_k^2/N )</td>
</tr>
<tr>
<td>Proposed-(d)</td>
<td>( G_k^{-1} )</td>
<td>( 1.0 \times 10^{-1} )</td>
<td>( \sigma_k^2/N )</td>
</tr>
</tbody>
</table>

Fig. 2. An example of sparse echo impulse response \( h^* \) of length \( N = 512 \).

**APA/PAPA with Time-Varying Soft-Thresholding.** In algorithm (6), let \( T_k := \{ k \} \), \( S_k := \arg \min_{\xi \in \mathbb{R}^N} \| U_k^T h - d_k \| / \| \xi \|_2 \), \( \psi_k(h) \) be the same as Example 2. By using time-varying soft-thresholding (10) and introducing a regularization parameter \( \delta \) for numerical stability, we obtain the algorithm:

\[
h_k^{(1)} := \text{prox}_{\mu \psi_k}(I_N - \mu Q_k^{-1} U_k \Gamma^{-1} \varepsilon_k)(h_k),
\]

(11)

where \( \Gamma^{-1} = (U_k^T Q_k^{-1} U_k + \delta I)^{-1} \). Note: Setting \( Q_k := I_N \) or \( Q_k := G_k^{-1} \) in (11) yields APA/PAPA with the time-varying soft-thresholding.

### 4. NUMERICAL EXAMPLES

We compare the performance of the proposed sparse adaptive filtering algorithms (11) (see Table 1) with their counterparts without soft-thresholding. APA \( (r = 1, 2) \) and PAPA \( (r = 1, 2) \), respectively, in the EC problem. For all the simulations, we use the sparse echo impulse response \( h^* \) of length \( N = 512 \) with sampling rate 8 kHz shown in Fig. 2, which is initialized according to ITU-T G.168 [22].

The input signal \( u_k \) is the autoregressive signal generated by \( u_k = 0.8 u_{k-1} + v_k \) and then normalized to variance 1, where \( v_k \) is a white Gaussian noise with zero mean and variance \( 1 \). The noise \( n_k \) is a white Gaussian noise with zero mean and signal-to-noise ratio (SNR) = 30 dB, where \( \text{SNR} := 10 \log_{10}(E[z_k^2]/E[n_k^2]) \). in (11), we set \( \mu := 1, \epsilon := 1.0 \times 10^{-6}, \tau := 1.0 \times 10^{-4} \approx E[n_k^2] \) and the other parameters such as in Table 1. For APA and PAPA, we set the parameters in the same way as their proposed counterparts.

We adopt two measures: (i) system mismatch:

\[
\eta(h) := \log_{10} \frac{\| h^* - h_k \|^2 / N}{\| h^* \|^2 / N},
\]

and (ii) sparseness measure [4, 23]:

\[
\xi(h) := \frac{N}{N - \sqrt{N}} \left( 1 - \frac{\| h_k \|_1}{\sqrt{N} \| h_k \|_2} \right) \in [0, 1].
\]

The larger \( \xi(h) \) is, the sparser \( h_k \) is: e.g., if \( h \in \{ 0, 1 \}^N \) such that exactly one component \( h_i \) is nonzero, then \( \xi(h) = 1 \). The results are averaged over 100 simulation runs.

Figure 3 illustrates the simulation results for system mismatch. It is seen that each of the proposed algorithms noticeably outperforms its counterpart without soft-thresholding both in convergence.
As specified more quickly, the value of which depicts the sparseness measure. It is clearly shown that, which promotes sparseness. Such an effect is observed in Fig. 4, performance come because of the use of time-varying soft-thresholding, which promotes sparseness. Such an effect is observed in Fig. 4, which depicts the sparseness measure. It is clearly shown that, compared to its counterpart without soft-thresholding, the value of sparseness measure by the proposed algorithms approaches the true value more quickly.

5. CONCLUSIONS

This paper has proposed a unified adaptive filtering scheme composed of parallel variable-metric projection and the proximity operator. As specific use of our proposed algorithm, tailored to sparse system identification, we also have proposed an efficient sparse adaptive filtering algorithm composed of the standard/proportionate APA and time-varying soft-thresholding. The numerical examples for the EC problem have shown that each of the proposed algorithms significantly outperforms its counterpart without soft-thresholding both in convergence speed and steady-state mismatch, while the extra computational complexity is negligible. Our algorithm achieves notable performance improvements for sparse-system identification.

In this paper, although the experiments only for single cases (i.e., $|Z_k| = 1$ in Sec. 3.1) of the proposed algorithm (6) are reported, the notable performance improvements by proposed time-varying soft-thresholding for multiple cases (i.e., $|Z_k| > 1$) have been also examined. This result will be reported elsewhere.

6. REFERENCES