2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Robust Adaptive Filtering Based on Adaptive Projected Subgradient Method: Moreau Enhancement of Distance Function

Daiki Sawada and Masahiro Yukawa
Department of Electronics and Electrical Engineering, Keio University, Japan
E-mail: sawada@ykw.elec.keio.ac.jp, yukawa@elec.keio.ac.jp, Phone: +81 (0)45 566 1521

***Abstract*—This paper presents an outlier-robust adaptive filtering algorithm based on the adaptive projected subgradient method (APSM). The proposed algorithm is derived from a cost function involving the Moreau enhancement of the distance function to a hyperslab which permits a prespecified amount of estimation residuals. A data-dependent penalty is also exploited to preserve the overall convexity of the instantaneous cost with a minimal amount of estimation bias. Our instantaneous cost function involves a constant term to make the minimum of the cost be exactly zero so that the monotone approximation property of APSM holds with a fixed step size. Numerical examples show remarkable robustness of the proposed algorithm for both white and colored input signals.**

I. Introduction

Robustness against outlier has been a crucial issue in signal processing and machine learning [1] as well as in statistics. In the signal processing community, adaptive filtering has a long history [3], [4], and the classical algorithms such as the least mean square (LMS) algorithm [5] and the normalized LMS (NLMS) algorithm [6], [7] are known to be sensitive to outliers (impulsive noise). A number of robust algorithms have therefore been proposed, such as the sign algorithm [8], the least logarithmic absolute difference algorithm [9], least mean mixed-norm algorithm [10], robust mixed-norm algorithm [11], the hyperbolic cosine adaptive filter (HCAF) [12], and the logarithmic HCAF (LHCAF) [12], to name just a few.

Despite mathematical tractability, those approaches based on convex loss functions have limited robustness, and non-convex loss functions have also been studied, such as the sigmoid LMS algorithm [13], the generalized maximum correntropy criterion (GMCC) algorithm [14], and the generalized modified Blake-Zisserman algorithm [15]. Those nonconvex approaches, however, have an issue of local minima. Recently, the minimax concave (MC) loss function has been studied for robust signal recovery [16]–[18]. While the MC function was proposed originally for reducing the estimation bias in sparse modeling [19], [20], its use in robust signal recovery brings remarkable robustness owing to the so-called *redescending* property as well as overall convexity of the whole cost function. It is therefore of great interest to extend the same idea to robust adaptive filtering.

This work was partially supported by JSPS Grants-in-Aid (22H01492).

The adaptive projected subgradient method (APSM) [21] is a unified mathematical framework covering a variety of adaptive filtering algorithms including LMS, NLMS, the affine projection algorithms [22]–[26], the adaptive parallel subgradient projection algorithm [27], [28], their constrained versions [29]–[31], among many others. See also [32] for a comprehensive account of APSM. APSM exploits the subgradient projection to suppress instantaneous cost functions at every time instant, and it generates a sequence of adaptive filters approaching the set of minimizers of the instantaneous cost function monotonically at each iteration. Despite its nice mathematical properties as well as a number of empirical evidences [33]–[40] which support its practical advantages, APSM has not been utilized so far in the context of robust adaptive filtering.

In this paper, we present a robust adaptive filtering algorithm exploiting the MC loss under the APSM framework. Our MC-based loss function is derived as the Moreau enhancement of the distance function with respect to a bounded-instantaneous-error hyperslab which permits instantaneous errors below a prespecified threshold. As a result, the proposed loss function returns zero if the instantaneous error is below the bound, and it returns a constant value if the error exceeds another bound (threshold) to detect outliers. While the loss function is weakly convex, an introduction of the standard Tikhonov regularization may cause serious estimation biases. To preserve overall convexity of the whole cost function without causing such large biases, we focus on the fact that the instantaneous loss function is constant over the subspace orthogonal to the input vector. This property indicates that the loss can be convexified by adding the square of the filtered input multiplied by an appropriate scaling factor. This additional term can be regarded as a minimal possible data-dependent penalty to convexify the weakly convex loss function. Finally, we subtract a constant from the cost to make the minimum value of the whole cost function be zero so that the monotone approximation holds with a fixed step size. Numerical examples show that the proposed method outperforms the existing methods for both white and colored input signals.

II. Preliminaries

This section gives the notation and some mathematical tools used in this paper.

A. Notation and definition

The sets of nonnegative integers and real numbers are denoted by $\mathbb{N}$ and $\mathbb{R}$, respectively. Transposition of vector/matrix is denoted by $(\cdot)^\top$. Given a vector $\boldsymbol{x} := [x_1, x_2, \cdots, x_n]^\top \in \mathbb{R}^n$, we define the ℓ_p norm $\|\boldsymbol{x}\|_p := (\sum_{k=1}^n |x_k|^p)^{1/p}$ for $p \in [1, +\infty)$. The inner product between vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y} = \sum_{k=1}^n x_k y_k$. The Fenchel conjugate of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is defined by $f^* : \mathbb{R}^n \to (-\infty, +\infty] : \boldsymbol{x} \mapsto \sup_{\boldsymbol{y} \in \mathbb{R}^n} (\langle \boldsymbol{x}, \boldsymbol{y} \rangle - f(\boldsymbol{y}))$.

B. Problem statement

Let $(\boldsymbol{u}_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n$ be the sequence of input vectors $\boldsymbol{u}_k := [u_k, u_{k-1}, \cdots, u_{k-n+1}]^\top$, where $k \in \mathbb{N}$ is the time index and $n \in \mathbb{N}$ is the tap length. Let $(n_k)_{k \in \mathbb{N}}$ and $(o_k)_{k \in \mathbb{N}}$ be the sequences of Gaussian noise and outlier, respectively. We consider the following linear system model:

$$d_k := \boldsymbol{u}_k^\top \boldsymbol{h}^* + n_k + o_k, \qquad k \in \mathbb{N}, \tag{1}$$

where $\boldsymbol{h}^* \in \mathbb{R}^n$ is the estimand (the system to be estimated). The estimation residual (the error function) is defined by

$$e_k : \mathbb{R}^n \to \mathbb{R} : \boldsymbol{h} \mapsto \boldsymbol{u}_k^\top \boldsymbol{h} - d_k, k \in \mathbb{N}. \tag{2}$$

The task of adaptive filter is to estimate $\boldsymbol{h}^*$ by $\boldsymbol{h}_k := [h_1^{(k)}, h_2^{(k)}, \cdots, h_n^{(k)}]^\top \in \mathbb{R}^n$ in a recursive fashion. Hereafter, we assume that $\boldsymbol{u}_k \neq \boldsymbol{0}$, because $\boldsymbol{u}_k = \boldsymbol{0}$ means that d_k only contains noise (and outliers) and such a data may not be used for adaptation.

III. Proposed Algorithm

To derive the proposed outlier-robust adaptive filtering algorithm, we first show that the Moreau enhancement of the distance function coincides with its composition with the MC penalty. Based on this, we present the proposed formulation to which APSM is applied to derive the proposed algorithm.

A. Moreau enhancement of distance function

We define the bounded instantaneous error hyperslab

$$S_k(\rho) \;\; := \;\; \{\boldsymbol{h} \in \mathbb{R}^n \mid e_k^2(\boldsymbol{h}) \leq \rho\}, \tag{3}$$

where $\rho > 0$ is the error bound. The distance function from a point $\boldsymbol{h} \in \mathbb{R}^n$ to $S_k(\rho)$ is given by

$$d_{S_k(\rho)}(\boldsymbol{h}) \;\; := \;\; \begin{cases} \dfrac{|e_k(\boldsymbol{h})| - \sqrt{\rho}}{\|\boldsymbol{u}_k\|_2}, & \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ 0, & \text{if } |e_k(\boldsymbol{h})| \leq \sqrt{\rho}. \end{cases} \tag{4}$$

Given a convex function $f : \mathbb{R}^m \to \mathbb{R}$ defined on the Euclidean space $\mathbb{R}^m$ of arbitrary dimension m, its Moreau envelope of index $\gamma > 0$ is defined by

$$^\gamma f(\boldsymbol{x}) := \min_{\boldsymbol{y} \in \mathbb{R}^m} \left(f(\boldsymbol{y}) + \frac{1}{2\gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right). \tag{5}$$

For instance, the Moreau envelope $^\gamma|\cdot|$ of the absolute-value function is the Huber loss function used widely in robust estimation. Subtracting the Moreau envelope $^\gamma|\cdot|$ from $|\cdot|$ yields the MC penalty [19], [20]:

$$\varphi_\gamma^{\mathrm{MC}}(x) = |x| - \, ^\gamma|\cdot|(x), \quad x \in \mathbb{R}, \tag{6}$$

which is γ^{-1}-weakly convex; i.e., $\varphi_\gamma^{\mathrm{MC}} + (\gamma^{-1}/2)\|\cdot\|^2$ is convex. The function $\varphi_\gamma^{\mathrm{MC}}(x)$ is referred to as the *Moreau enhancement* of $|\cdot|$ [20], [42]. In fact, the Moreau enhancement of the distance function $d_{S_k(\rho)}$ can be written by using the MC penalty $\varphi_\gamma^{\mathrm{MC}}$.

Proposition 1. Let C be a nonempty closed convex subset of $\mathbb{R}^n$. Then, the Moreau enhancement of the distance function d_C of index $\gamma > 0$ is given by

$$(d_C)_\gamma(\boldsymbol{h}) := d_C(\boldsymbol{h}) - \, ^\gamma d_C(\boldsymbol{h}) = \varphi_\gamma^{\mathrm{MC}}(d_C(\boldsymbol{h})). \tag{7}$$

B. Proposed formulation

Based on Proposition 1, the Moreau enhancement of $d_{S_k(\rho)}$ is given by

$$(d_{S_k(\rho)})_\gamma(\boldsymbol{h}) = \varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h})). \tag{8}$$

Since $\varphi_\gamma^{\mathrm{MC}} \circ d_{S_k(\rho)}$ is γ^{-1}-weakly convex, an addition of $(\gamma^{-1}/2)\|\cdot\|^2$ as a regularizer makes the whole cost convex. This straightforward approach, however, causes a non-negligible bias, resulting in degraded performance. We therefore introduce the following alternative:

$$f_k(\boldsymbol{h}) := \varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h})) + \frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2, \; \boldsymbol{h} \in \mathbb{R}^n, \tag{9}$$

where $\tilde{\boldsymbol{u}}_k := \boldsymbol{u}_k / \|\boldsymbol{u}_k\|_2$. The loss function $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$ is highly robust against large outliers owing to the saturation property of the MC function $\varphi_\gamma^{\mathrm{MC}}$. However, since $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$ is concave over the subspace $\mathrm{span}\{\boldsymbol{u}_k\}$, it is difficult to guarantee the monotone approximation property of the global minimizer at each time instant, which is important to ensure stable performance of adaptive algorithms [21], [33]–[39]. Therefore, the second term $\frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2$ is introduced as a "minimal" possible regularizer, in some sense, to convexify $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$. The point here is that $\frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2$ gives no impact on those components over the orthogonal compliment $\mathrm{span}^\perp\{\boldsymbol{u}_k\}$, thus yielding lower bias compared to the Tikhonov regularization $\|\boldsymbol{h}\|_2^2$. We mention here that $\boldsymbol{h}$ is implicitly assumed to be a random vector obeying a generalized Gaussian distribution with the precision matrix $\tilde{\boldsymbol{R}} := E(\tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{u}}_k^\top)$, where $E(\cdot)$ stands for expectation. Our simulation will show that our method works well even though such an implicit assumption is violated (see Section IV).

We now focus on the minimum value

$$f_k^{\min} := \min_{\boldsymbol{h} \in \mathbb{R}^n} f_k(\boldsymbol{h}) \tag{10}$$

$$= \begin{cases} \min\left\{ \dfrac{\gamma}{2}, \dfrac{(|d_k| - \sqrt{\rho})^2}{2\gamma\|\boldsymbol{u}_k\|_2^2} \right\}, & \text{if } |d_k| > \sqrt{\rho}, \\ 0, & \text{if } |d_k| \leq \sqrt{\rho}, \end{cases} \tag{11}$$

of the function f_k. Subtracting $f_k^{\min}$ from f_k, our instantaneous cost function $\Theta_k : \mathbb{R}^n \to [0, +\infty)$ is defined by

$$\Theta_k(\boldsymbol{h}) := \underbrace{\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))}_{\text{robustifying the data fidelity term}} + \underbrace{\frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2}_{\text{convexifying the whole cost}} \underbrace{- f_k^{\min}}_{\text{making the minimum be zero}} \quad (12)$$

so that

$$\Theta_k^{\min} := \min_{\boldsymbol{h}\in\mathbb{R}^n} \Theta_k(\boldsymbol{h}) = 0. \quad (13)$$

The property in (13) allows the use of a constant step size in APSM (see Proposition 2 below). We remark that the first and second terms of (12) brings remarkable outlier-robustness while preserving the convexity of the whole cost.

C. Adaptive Projected Subgradient Method for proposed formulation

The nonnegative continuous convex function $\Theta_k(\boldsymbol{h})$ is minimized by APSM [21]:

$$\boldsymbol{h}_{k+1} = \boldsymbol{h}_k + \lambda_k \Big(T_{\mathrm{sp}(\Theta_k)}(\boldsymbol{h}_k) - \boldsymbol{h}_k\Big), \quad (14)$$

where $\lambda_k \in (0, 2)$ is the step size, and

$$T_{\mathrm{sp}(\Theta_k)}(\boldsymbol{h}) := \begin{cases} \boldsymbol{h} - \dfrac{\Theta_k(\boldsymbol{h})}{\|\Theta_k'(\boldsymbol{h})\|_2^2}\Theta_k'(\boldsymbol{h}), & \text{if } \Theta_k(\boldsymbol{h}) > 0, \\ \boldsymbol{h}, & \text{otherwise,} \end{cases} \quad (15)$$

with a subgradient

$$\Theta_k'(\boldsymbol{h}) := \begin{cases} \mathrm{sign}(e_k(\boldsymbol{h}))\left(1 - \dfrac{1}{\gamma}\mathrm{clip}_\gamma(d_{S_k(\rho)}(\boldsymbol{h}))\right)\tilde{\boldsymbol{u}}_k \\ \qquad + \dfrac{\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top}{\gamma}\boldsymbol{h}, \quad \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ \dfrac{\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top}{\gamma}\boldsymbol{h}, \quad \text{if } |e_k(\boldsymbol{h})| \le \sqrt{\rho}, \end{cases} \quad (16)$$

of Θ_k at $\boldsymbol{h}$. Here,

$$\mathrm{sign} : \mathbb{R} \to \{-1, 1\} : x \mapsto \begin{cases} 1, & \text{if } x \ge 0, \\ -1, & \text{if } x < 0, \end{cases} \quad (17)$$

$$\mathrm{clip}_\gamma : \mathbb{R}_+ \to [0, \gamma] : x \mapsto \begin{cases} \gamma, & \text{if } x \in (\gamma, +\infty), \\ x, & \text{if } x \in [0, \gamma]. \end{cases} \quad (18)$$

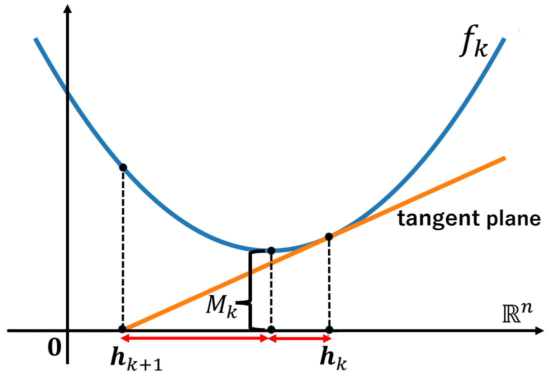
Proposition 2. (Monotone approximation [27, Theorem 2]) Assume that

$$\boldsymbol{h}_k \notin \Omega_k := \left\{\boldsymbol{h} \in \mathbb{R}^n \mid \Theta_k(\boldsymbol{h}) = \Theta_k^{\min} = 0\right\} \neq \emptyset. \quad (19)$$

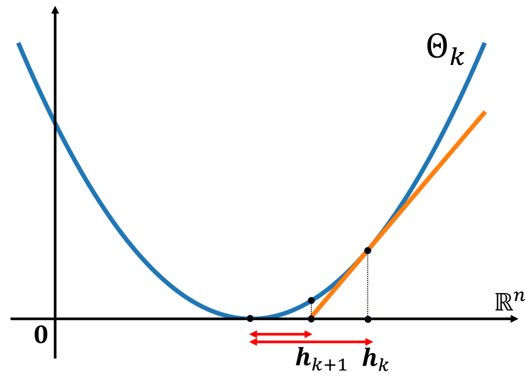
Then, for an arbitrary step size $\lambda_k \in (0, 2)$, it holds that

$$\|\boldsymbol{h}_{k+1} - \boldsymbol{h}_k^{(*)}\|_2 < \|\boldsymbol{h}_k - \boldsymbol{h}_k^{(*)}\|_2, \quad \forall \boldsymbol{h}_k^{(*)} \in \Omega_k. \quad (20)$$

Proposition 2 states that the sequence $(\boldsymbol{h}_k)_{k\in\mathbb{N}}$ monotonically approaches the solution set Ω_k at every time instant that is expected to contain the estimand. The role of $f_k^{\min}$ for the monotone approximation is illustrated in Fig. 1. Roughly speaking, the subgradient projection is the orthogonal projection onto the intersection of the tangent plane and $\mathbb{R}^n$. Viewing Fig. 1(a), one can see that $T_{\mathrm{sp}(f_k)}(\boldsymbol{h})$ is far from the minimizer of $f_k(\boldsymbol{h})$ than $\boldsymbol{h}$, meaning that the operation of $T_{\mathrm{sp}(f_k)}$ may put the current estimate away from the set Ω_k. This happens



(a) f_k



(b) Θ_k

Fig. 1: Illustrations of the subgradient projections $T_{\mathrm{sp}(f_k)}$ and $T_{\mathrm{sp}(\Theta_k)}$.

because the graph of f_k is distant (by $f_k^{\min}$) from the 'ground' $\mathbb{R}^n$, and, in this case, the step size must be smaller than a bound depending on $f_k^{\min}$ to ensure the monotone approximation property [21]. In contrast, the operator $T_{\mathrm{sp}(\Theta_k)}$ in Fig. 1(b) pushes $\boldsymbol{h}$ towards Ω_k so that the monotone approximation holds.

Proposition 3. For every $\gamma > 0$, $\Theta_k(\boldsymbol{h})$ is convex.

Proof: For the sake of accessibility, we present a primitive proof. It suffices to show convexity of $f_k(\boldsymbol{h})$. By virtue of the Moreau decomposition [41] and by the definition of $\varphi_\gamma^{\mathrm{MC}}$ in (6), $f_k(\boldsymbol{h})$ can be rewritten as

$$f_k(\boldsymbol{h}) = \underbrace{d_{S_k(\rho)}(\boldsymbol{h}) + {}^{\gamma^{-1}}(|\cdot|^*)(\gamma^{-1}d_{S_k(\rho)}(\boldsymbol{h}))}_{\text{convex}} \\ \underbrace{-\frac{1}{2\gamma}d^2_{S_k(\rho)}(\boldsymbol{h}) + \frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2}_{g_k(\boldsymbol{h})}. \quad (21)$$

Here, the conjugate function $|\cdot|^*$ of $|\cdot|$ and the distance function $d_{S_k(\rho)}(\boldsymbol{h})$ are convex [21]. It can also be verified that the composition ${}^{\gamma^{-1}}(|\cdot|^*)(\gamma^{-1}d_{S_k(\rho)}(\boldsymbol{h}))$ is also convex. It is therefore sufficient to show convexity of $g_k(\boldsymbol{h})$. Recall that

TABLE I: Computational complexity per iteration

proposed method	GMCC	LHCAF	NLMS
$3n+25$	$2n+4$	$2n+6$	$3n+2$

it is assumed that $\boldsymbol{u}_k \neq \mathbf{0}$. The gradient $\nabla g_k(\boldsymbol{h})$ of $g_k(\boldsymbol{h})$ is given by

$$\nabla g_k(\boldsymbol{h}) = \begin{cases} \dfrac{1}{\gamma}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h} - \mathrm{sign}(e_k(\boldsymbol{h}))\dfrac{|e_k(\boldsymbol{h})| - \sqrt{\rho}}{\gamma\|\boldsymbol{u}_k\|_2}\tilde{\boldsymbol{u}}_k, \\ \qquad\qquad \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ \dfrac{1}{\gamma}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h}, \quad \text{if } |e_k(\boldsymbol{h})| \leq \sqrt{\rho}, \end{cases} \tag{22}$$

and the Hessian $\nabla^2 g_k(\boldsymbol{h})$ of $g_k(\boldsymbol{h})$ is given by

$$\nabla^2 g_k(\boldsymbol{h}) = \begin{cases} \boldsymbol{O}, & \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ \dfrac{1}{\gamma}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top, & \text{if } |e_k(\boldsymbol{h})| \leq \sqrt{\rho}, \end{cases} \tag{23}$$

where $\boldsymbol{O} \in \mathbb{R}^{n\times n}$ is the zero matrix. From (23), $\nabla^2 g_k(\boldsymbol{h})$ is positive semi-definite as $\boldsymbol{h}^\top(\boldsymbol{u}_k\boldsymbol{u}_k^\top)\boldsymbol{h} = (\boldsymbol{u}_k^\top\boldsymbol{h})^2 \geq 0$ for all $\boldsymbol{h} \in \mathbb{R}^n$, and hence $g_k(\boldsymbol{h})$ is convex. ■

D. Complexity

The computational complexities (the number of multiplications) of the proposed and existing algorithms are given in Table I, where it can be seen that all methods have linear-order complexities. It will be shown in Section IV that the proposed method significantly improves outlier robustness with the moderate computational complexity.

E. Parameter design

Possible choices of ρ are given as follows [27]: $\rho_1 := (1+\sqrt{2})\sigma^2$, $\rho_2 := \sigma^2$, and $\rho_3 := 0$, where n_k is assumed to obey the zero-mean i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$. An increase of the error bound ρ enhances the "thickness" of the hyperslab $S_k(\rho)$, while the step size λ_k governs the degree of relaxation for the projection onto $S_k(\rho)$. In fact, an increase of ρ may give an impact, similar to a decrease of λ_k, on the performance. We mention here another aspect that an increase of ρ also decreases the frequency of update, thereby reducing the computational loads. Our recommended choice is setting ρ to ρ_1, ρ_2, or ρ_3, and tuning λ_k possibly depending on the noise variance.

The saturation factor γ could be selected based on the the variance of Gaussian noise and the outlier statistics. As depicted in Fig. 2, the Gaussian noise concentrates around the mean value, while outliers reside in those regions apart from the Gaussian noise. The saturation factor γ is desired to be tuned in such a way that the Gaussian noise falls, with high probability, into the central region between the two dotted lines in the figure, and, at the same time, the outliers arise in the saturated regions where the gradient vanishes.

IV. Numerical Examples

We compare the performance of the proposed method to those of the robust adaptive filtering algorithms: LHCAF [12] and GMCC [14]. For reference, we also test the NLMS algorithm. We consider two types of input signal $(u_k)_{k\in\mathbb{N}}$, white and colored. The white input signal is generated from the standard Gaussian distribution. As a colored input, the USASI signal (speech-like wide-sense stationary process) is used, modeled on the autoregressive moving average process characterized by the transfer function [27]

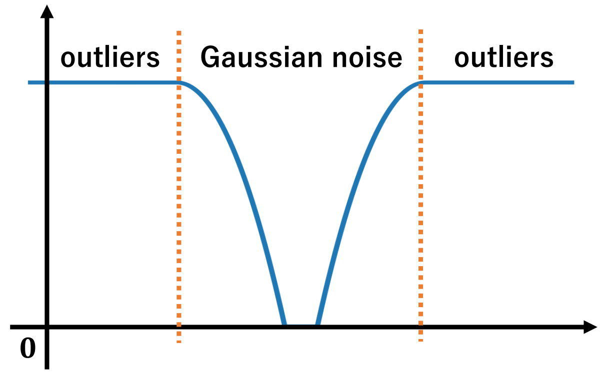


Fig. 2: Graph of $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$ and the saturation factor γ.

TABLE II: Parameter setting for the case of white signals

Methods	noise 10 dB	noise 20 dB
proposed	λ_k=0.4, γ=3.5, ρ=0	λ_k=0.5, γ=3.5, ρ=0
GMCC	η=0.025, α=1, λ=0.001	η=0.029, α=1, λ=0.005
LHCAF	λ=0.12, μ= 0.03	λ=0.14, μ=0.035
NLMS	μ=0.03	μ=0.03

$$H_{\mathrm{USASI}}(z) := \frac{1-z^{-2}}{1-1.70233z^{-1}+0.71902z^{-2}}, \quad z \in \mathbb{C}, \tag{24}$$

where $\mathbb{C}$ denotes the set of complex numbers.

The estimand $\boldsymbol{h}^*$ of length $n = 256$ is generated from the i.i.d. standard Gaussian distribution. The noise n_k is white Gaussian with zero mean and signal-to-noise ratio (SNR) 10 dB and 20 dB, where SNR $:= 10\log_{10}(E[z_k^2]/E[n_k^2])$ with $z_k := \boldsymbol{u}_k^\top\boldsymbol{h}^*$. An impulsive noise is generated by $o_k := b_k v_k$, where b_k obeys the Bernoulli process with success and fail probabilities $P(b_k = 1) = 0.05$ and $P(b_k = 0) = 0.95$, respectively, and v_k is zero-mean Gaussian with variance $\sigma_v^2 = 1000E[z_k^2]$. We adopt the system mismatch defined as

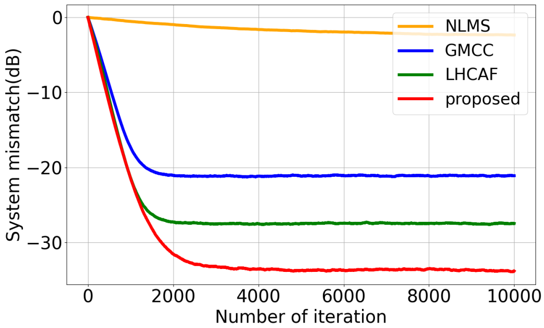
$$\xi(\boldsymbol{h}_k) := 20\log_{10}\frac{\|\boldsymbol{h}^* - \boldsymbol{h}_k\|_2}{\|\boldsymbol{h}^*\|_2}. \tag{25}$$

All result are averaged over 300 trials. The adaptive filters are initialized as $\boldsymbol{h}_0 := \mathbf{0}$. The parameters for each input are given in Tables I and II, respectively.

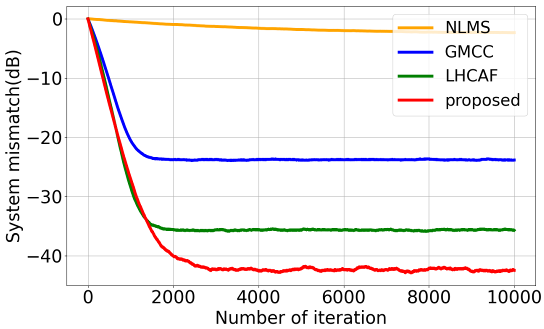
Figs. 3 and 4 depict the results for the white and colored input signals, respectively. It can be seen that the proposed method significantly outperforms the existing robust methods in both cases.

V. Conclusion

This paper presented the robust adaptive filtering algorithm based on APSM. The instantaneous cost function consists of three terms: (i) the weakly convex loss given by the Moreau enhancement of the distance function, (ii) the data-dependent

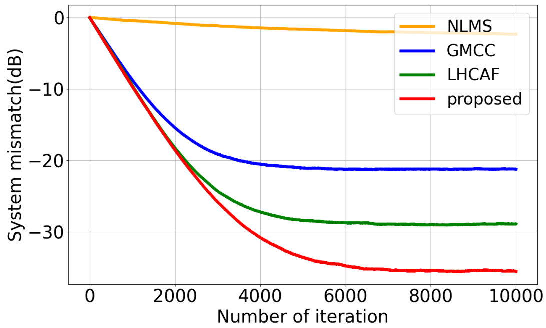


(a) noise 10 dB

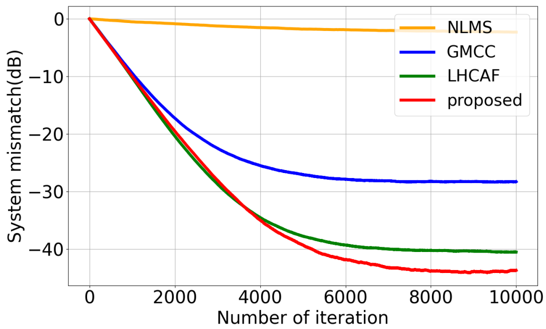


(b) noise 20 dB

Fig. 3: Learning curves for white signals.



(a) noise 10 dB



(b) noise 20 dB

Fig. 4: Learning curves for USASI signals.

TABLE III: Parameter setting for the case of colored signals

Methods	noise 10 dB	noise 20 dB
proposed	λ_k=0.3, γ=5, ρ=0	λ_k=0.3, γ=5, ρ=0
GMCC	η=0.01, α=1, λ=0.005	η=0.01 α=1, λ=0.005
LHCAF	λ=0.12, μ=0.007	λ=0.12, μ=0.007
NLMS	μ=0.03	μ=0.03

penalty function, and (iii) the constant term. The loss function brings significant robustness against impulsive noises, while the penalty function makes the whole cost function be convex without causing serious estimation biases. The constant term makes the minimum of the whole cost be exactly zero, allowing the use of a fixed step size. The numerical examples showed that the proposed method outperformed the existing methods (GMCC, LHCAF) in the presence of impulsive noises for both white and colored input signals.

References

[1] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, Robust Statistics for Signal Processing. Cambridge: Cambridge Uni- versity Press, 2018.

[2] T. van Erven, S. Sachs, W. M. Koolen, and W. Kotlowski, “Robust Online Convex Optimization in the Presence of Outliers,” Proc. Machine Learning Research, vol. 134, pp. 1–21, 2021.

[3] S. Haykin, “Adaptive Filter Theory”, 4th ed., New Jersey: Prentice Hall, 2002.

[4] A. H. Sayed, “Fundamentals of Adaptive Filtering”, Wiley-Interscience, New York, 2003.

[5] B. Widrow and M. E. Hoff Jr., “Adaptive switching circuits,” in IRE WESCON Conv. Rec., vol. 4, pp. 96–104, 1960.

[6] J. Nagumo and A. Noda, “A learning method for system identification”, IEEE Trans. Autom. Control, vol. 12, no. 3, pp. 282–287, 1967.

[7] A. E. Albert and L. S. Gardner, Jr., “Stochastic Approximation and Nonlinear Regression”, Cambridge, MIT Press, 1967.

[8] N. Verhoeckx, H. Van Den Elzen, F. Snijders, and P. Van Gerwen, “Digital echo cancellation for baseband data transmission”, IEEE Trans. Acoust, vol. 27, no. 6, pp. 768–781, 1979.

[9] M. O. Sayin, N. D. Vanli, and S. S. Kozat, “A novel family of adaptive filtering algorithms based on the logarithmic cost”, IEEE Trans. Signal Process., vol. 62, no.17, 2014. 4411–4424.

[10] J. A. Chambers, O. Tanrikulu, and A.G. Constantinides, “Least mean mixed-norm adaptive filtering,” Electron Lett, vol. 30, no. 19, pp. 1574–1575, 1994.

[11] J. Chambers, and A. Avlonitis, “A robust mixed-norm adaptive filter algorithm,” IEEE Signal Process. Lett., vol. 4, no. 2, pp. 46–48, 1997.

[12] S. Wang, W. Wang, K. Xiong, H. H. C. Iu, and K.T. Chi, “Logarithmic hyperbolic cosine adaptive filter and its performance analysis”, IEEE Trans. Systems, Man., and Cybernetics: Systems, vol. 51, no. 4, pp. 2512–2524, Apr. 2021.

[13] F. Huang, J. Zhang, and S. Zhang, “A family of robust adaptive filtering algorithms based on sigmoid cost”, Signal Process., vol. 149, pp. 179–192, 2018.

[14] B. Chen, L. Xing, H. Zhao, N. Zheng, and J.C. Principe, “Generalized correntropy for robust adaptive filtering”, IEEE Trans. Signal Process., vol. 64, no.13, pp. 3376–3387, 2016.

[15] K. Kumar, M. L. N. S. Karthik, and N. V. George, ”Generalized Modified Blake–Zisserman Robust Sparse Adaptive Filters,” IEEE Trans. Systems, Man, and Cybernetics: Systems, vol. 53, no. 1, pp. 647-652, Jan. 2023.

[16] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, “Linearly-involved Moreau-enhanced-over-subspace model: Debiased sparse modeling and stable outlier-robust regression,” IEEE Trans. Signal Process., vol. 71, pp. 1232–1247, 2023.

[17] K. Suzuki and M. Yukawa, “Robust recovery of jointly-sparse signals

using minimax concave loss function," IEEE Trans. Signal Process., vol. 69, pp. 669–681, 2021.

[18] K. Suzuki and M. Yukawa, "Sparse stable outlier-robust signal recovery under Gaussian noise," IEEE Trans. Signal Process., vol.71, pp.372–387, 2023.

[19] C. H. Zhang et al., "Nearly unbiased variable selection under minimax concave penalty," The Annals of statistics, vol. 38, no. 2, pp. 894–942, 2010.

[20] I. Selesnick, "Sparse regularization via convex analysis," IEEE Trans. Signal Process., vol. 65, no. 17, pp. 4481–4494, 2017.

[21] I. Yamada and N. Ogura, "Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions," Numerical Functional Analysis and Optimization, vol. 25, pp. 593–617, 2004.

[22] T. Hinamoto and S. Maekawa, "Extended theory of learning identification," (in Japanese) Trans. Inst. Elect. Eng. Jpn., vol. 95, no. 10, pp. 227–234, 1975.

[23] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," Electron. Commun. Jpn., vol. 67-A, no. 5, pp. 19–27, 1984.

[24] D. Slock, "The block underdetermined covariance (BUC) fast transversal filter (FTF) algorithm for adaptive filtering," in Proc. Asilomar Conf. Signals, Syst., Comput., 1992, pp. 550–554.

[25] D. R. Morgan and S. G. Kratzer, "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms," IEEE Signal Process. Lett., vol. 3, no. 8, pp. 245–247, Aug. 1996.

[26] M. Rupp, "A family of adaptive filter algorithms with decorrelating properties," IEEE Trans. Signal Process., vol. 46, no. 3, pp. 771–775, Mar. 1998.

[27] I. Yamada, K. Slavakis, and K. Yamada, "An efficient robust adaptive filtering algorithm based on parallel subgradient projection techniques," IEEE Trans. Signal Process., vol. 50, no. 5, pp. 1091–1101, 2002.

[28] M. Yukawa and I. Yamada, "Pairwise optimal weight realization — Acceleration technique for set-theoretic adaptive parallel subgradient projection algorithm," IEEE Trans. Signal Process., vol. 54, no. 12, pp. 4557–4571, Dec. 2006.

[29] J. A. Apolinário, Jr., S. Werner, P. S. R. Diniz, and T. I. Laakso, "Constrained normalized adaptive filters for CDMA mobile communications," in Proc. EUSIPCO, 1998, vol. IV, pp. 2053–2056.

[30] M. L. R. de Campos and J. A. Apolinário, Jr., "The constrained affine projection algorithm — Development and convergence issues," in Proc. Balkan Conf. Signal Process., Commun., Circuits Syst., 2000.

[31] M. Yukawa, R. L. G. Cavalcante, and I. Yamada, "Efficient blind MAI suppression in DS/CDMA systems by embedded constraint parallel projection techniques," IEICE Trans. Fundam., vol. E88-A, no. 8, pp. 2062–2071, Aug. 2005.

[32] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections: A unifying framework for linear and nonlinear classification and regression tasks," IEEE Signal Process. Mag., vol. 28, no. 1, pp. 97–123, Jan. 2011.

[33] M. Yukawa and I. Yamada, "Efficient adaptive stereo echo canceling schemes based on simultaneous use of multiple state data," IEICE Trans. Fundam., vol. E87-A, no. 8, pp. 1949–1957, Aug. 2004.

[34] R. L. G. Cavalcante, M. Yukawa, and I. Yamada, "Set-theoretic DS/CDMA receivers for fading channels by adaptive projected subgradient method," in Proc. IEEE GLOBECOM, SP06-1, 2005.

[35] K. Slavakis, M. Yukawa, and I. Yamada, "Robust Capon beamforming by the adaptive projected subgradient method," in Proc. IEEE ICASSP, 2006, pp. 1005–1008.

[36] M. Yukawa, K. Slavakis, and I. Yamada, "Adaptive parallel quadratic-metric projection algorithms," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1665–1680, Jul. 2007.

[37] R. L. G. Cavalcante and I. Yamada, "Multiaccess interference suppression in orthogonal space-time block coded MIMO systems by the adaptive projected subgradient method," IEEE Trans. Signal Process., vol. 56, no. 3, pp. 1028-1042, March 2008.

[38] 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, July 2009.

[39] R. L. G. Cavalcante, A. Rogers, N. R. Jennings, and I. Yamada, "Distributed asymptotic minimization of sequences of convex functions by a broadcast adaptive subgradient method," IEEE J. Selected Topics in Signal Process., vol. 5, no. 4, pp. 739-753, Aug. 2011.

[40] N. Agrawal, R.L.G. Cavalcante, M. Yukawa, and S. Stańczak, "Distributed Convex Optimization "Over-the-Air" in Dynamic Environments", IEEE Trans. Signal and Information Processing over Networks, vol. 10, pp. 610-625, 2024.

[41] H. H. Bauschke and P. L. Combettes, "Convex Analysis and Monotone Operator Theory in Hilbert Spaces", 2nd Ed., Springer, 2017.

[42] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition," Inverse Problems, vol. 36, no. 3, pp. 1–36, Feb. 2020.