Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 589260, 13 pages
doi:10.1155/2009/589260

Research Article

A Unified View of Adaptive Variable-Metric Projection Algorithms

Masahiro Yukawa[1] and Isao Yamada[2]

[1] *Mathematical Neuroscience Laboratory, BSI, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan*
[2] *Department of Communications and Integrated Systems, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8552, Japan*

Correspondence should be addressed to Masahiro Yukawa, myukawa@riken.jp

Received 24 June 2009; Accepted 29 October 2009

Recommended by Vitor Nascimento

We present a unified analytic tool named *variable-metric adaptive projected subgradient method (V-APSM)* that encompasses the important family of adaptive variable-metric projection algorithms. The family includes the transform-domain adaptive filter, the Newton-method-based adaptive filters such as quasi-Newton, the proportionate adaptive filter, and the Krylov-proportionate adaptive filter. We provide a rigorous analysis of V-APSM regarding several invaluable properties including *monotone approximation*, which indicates stable tracking capability, and convergence to an asymptotically optimal point. Small metric-fluctuations are the key assumption for the analysis. Numerical examples show (i) the robustness of V-APSM against violation of the assumption and (ii) the remarkable advantages over its constant-metric counterpart for colored and nonstationary inputs under noisy situations.

1. Introduction

The adaptive projected subgradient method (APSM) [1–3] serves as a unified guiding principle of many existing projection algorithms including the normalized least mean square (NLMS) algorithm [4, 5], the affine projection algorithm (APA) [6, 7], the projected NLMS algorithm [8], the constrained NLMS algorithm [9], and the adaptive parallel subgradient projection algorithm [10, 11]. Also, APSM has been proven a promising tool for a wide range of engineering applications: interference suppression in the code-division multiple access (CDMA) and multi-input multioutput (MIMO) wireless communication systems [12, 13], multichannel acoustic echo cancellation [14], online kernel-based classification [15], nonlinear adaptive beamforming [16], peak-to-average power ratio reduction in the orthogonal frequency division multiplexing (OFDM) systems [17], and online learning in diffusion networks [18]. However, APSM does not cover the important family of algorithms that are based on iterative projections with its metric controlled adaptively for better performance. Such a family of *variable-metric projection algorithms* includes the transform-domain adaptive filter (TDAF) [19–21], the LMS-Newton adaptive filter (LNAF) [22–24] (or quasi-Newton adaptive filter (QNAF) [25, 26]), the proportionate adaptive filter (PAF) [27–33], and Krylov-proportionate adaptive filter (KPAF) [34–36]; it has been shown, respectively, in [34, 37] that TDAF and PAF perform iterative projections onto hyperplanes (the same as used by NLMS) *with variable metric*. The variable-metric projection algorithms enjoy significantly faster convergence compared to their constant-metric counterparts with reasonable computational complexity. At the same time, however, the variability of metric causes major difficulty in analyzing this family of algorithms. It is of great interests and importance to reveal the convergence mechanism.

The goal of this paper is to build a unified analytic tool that encompasses the family of adaptive variable-metric projection algorithms. The key to achieve this goal is the assumption of *small metric-fluctuations*. We extend APSM into *the variable-metric adaptive projected subgradient method (V-APSM)* that allows the metric to change in time.

V-APSM includes TDAF, LNAF/QNAF, PAF, and KPAF as its particular examples. We present a rigorous analysis of V-APSM regarding several properties. First, we show that V-APSM enjoys *monotone approximation*, which indicates stable tracking capability. Second, we prove that the vector sequence generated by V-APSM converges to a point in a certain desirable set. Third, we prove that both the vector sequence and its limit point minimize a sequence of cost functions to be designed by the user asymptotically; each cost function determines each iteration procedure of the algorithm. The analysis gives us an interesting view that TDAF, LNAF/QNAF, PAF, or KPAF asymptotically minimizes the metric distance to the data-dependent hyperplane which makes the instantaneous output-error be zero. The impacts of metric-fluctuations on the performance of adaptive filter are investigated by simulations.

The remainder of the paper is organized as follows. Preliminary to the major contributions, we present a brief review of APSM starting with a connection to the widely used NLMS algorithm in Section 2. We present V-APSM and its examples in Section 3, the analysis in Section 4, the numerical examples in Section 5, and the conclusion in Section 6.

2. Adaptive Projected Subgradient Method: Asymptotic Minimization of a Sequence of Cost Functions

Throughout the paper, $\mathbb{R}$ and $\mathbb{N}$ denote the sets of all real numbers and nonnegative integers, respectively, and vectors (matrices) are represented by bold-faced lower-case (upper-case) letters. Let $\langle \cdot, \cdot \rangle$ be an inner product defined on the N-dimensional Euclidean space $\mathbb{R}^N$ and $\| \cdot \|$ its induced norm. *The projected gradient method* [38, 39] is a simple extension of the popular gradient method (also known as the steepest descent method) to convexly constrained optimization problems. Precisely, it solves the minimization problem of *a differentiable convex function* $\varphi : \mathbb{R}^N \to \mathbb{R}$ over a given closed convex set $C \subset \mathbb{R}^N$, based on *the metric projection*:

$$P_C : \mathbb{R}^N \longrightarrow C, \quad \mathbf{x} \longmapsto P_C(\mathbf{x}) \in \arg\min_{\mathbf{a} \in C} \|\mathbf{a} - \mathbf{x}\|. \tag{1}$$

To deal with a (*possibly nondifferentiable*) *continuous convex function*, a generalized method named *the projected subgradient method* has been developed in [40]. For convenience, a brief review of the projected gradient and projected subgradient methods is given in Appendix A.

In 2003, Yamada has started to investigate the generalized problem in which φ is replaced by *a sequence of continuous convex functions* $(\varphi_k)_{k \in \mathbb{N}}$ [1]. We begin by explaining how this formulation is linked to the adaptive filtering.

2.1. NLMS from a Viewpoint of Asymptotic Minimization. Let $\langle \cdot, \cdot \rangle_2$ and $\| \cdot \|_2$ be the standard inner product and the Euclidean norm, respectively. We consider the following linear system [41, 42]:

$$d_k := \mathbf{u}_k^T \mathbf{h}^* + n_k, \quad k \in \mathbb{N}. \tag{2}$$

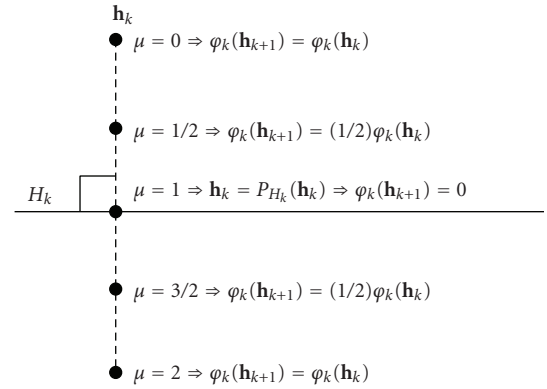

Figure 1: Reduction of the metric distance function $\varphi_k(\mathbf{x}) := d(\mathbf{x}, H_k)$ by the relaxed projection.

Here, $\mathbf{u}_k := [u_k, u_{k-1}, \ldots, u_{k-N+1}]^T \in \mathbb{R}^N$ is the input vector at time k with $(u_k)_{k \in \mathbb{N}}$ being the observable input process, $\mathbf{h}^* \in \mathbb{R}^N$ the unknown system, $(n_k)_{k \in \mathbb{N}}$ the noise process, and $(d_k)_{k \in \mathbb{N}}$ the observable output process. In the parameter estimation problem, for instance, the goal is to estimate $\mathbf{h}^*$. Given an initial $\mathbf{h}_0 \in \mathbb{R}^N$, the NLMS algorithm [4, 5] generates the vector sequence $(\mathbf{h}_k)_{k \in \mathbb{N}}$ recursively as follows:

$$\mathbf{h}_{k+1} := \mathbf{h}_k - \mu \frac{e_k(\mathbf{h}_k)}{\|\mathbf{u}_k\|_2^2} \mathbf{u}_k \tag{3}$$

$$= \mathbf{h}_k + \mu \left(P_{H_k}(\mathbf{h}_k) - \mathbf{h}_k \right), \quad k \in \mathbb{N}, \tag{4}$$

where $\mu \in [0, 2]$ is the step size (In the presence of noise, $\mu > 1$ would never be used in practice due to its unacceptable misadjustment without increasing the speed of convergence.) and

$$e_k(\mathbf{h}) := \langle \mathbf{u}_k, \mathbf{h} \rangle_2 - d_k, \quad \mathbf{h} \in \mathbb{R}^N, \ k \in \mathbb{N}, \tag{5}$$

$$H_k := \left\{ \mathbf{h} \in \mathbb{R}^N : e_k(\mathbf{h}) = 0 \right\}, \quad k \in \mathbb{N}. \tag{6}$$

The right side of (4) is called *the relaxed projection* due to the presence of μ, and it is illustrated in Figure 1. We see that for any $\mu \in (0, 2)$ the update of NLMS decreases the value of the metric distance function:

$$\varphi_k(\mathbf{x}) := d(\mathbf{x}, H_k) := \min_{\mathbf{a} \in H_k} \|\mathbf{x} - \mathbf{a}\|_2, \quad \mathbf{x} \in \mathbb{R}^N, \ k \in \mathbb{N}. \tag{7}$$

Figure 2 illustrates several steps of NLMS for $\mu = 1$. In noiseless case, it is readily verified that $\varphi_k(\mathbf{h}^*) = d(\mathbf{h}^*, H_k) = 0$, for all $k \in \mathbb{N}$, implying that (i) $\mathbf{h}^* \in \bigcap_{k \in \mathbb{N}} H_k$ and (ii) $\|\mathbf{h}_{k+1} - \mathbf{h}^*\|_2 \leq \|\mathbf{h}_k - \mathbf{h}^*\|_2$, for all $k \in \mathbb{N}$, due to the Pythagorean theorem. The figure suggests that $(\mathbf{h}_k)_{k \in \mathbb{N}}$ would converge to $\mathbf{h}^*$; namely, it would minimize $(\varphi_k)_{k \in \mathbb{N}}$ asymptotically. In noisy case, the properties (i) and (ii) shown above are *not* guaranteed, and NLMS can only compute an approximate solution. APA [6, 7] can be viewed in a similar way [10]. The APSM presented below is an extension of NLMS and APA.

2.2. A Brief Review of Adaptive Projected Subgradient Method. We have seen above that asymptotic minimization of

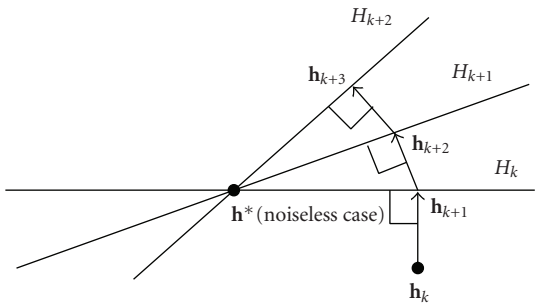

Figure 2: NLMS minimizes the sequence of the metric distance functions $\varphi_k(\mathbf{x}) := d(\mathbf{x}, H_k)$ asymptotically under certain conditions.

a sequence of functions is a natural formulation in the adaptive filtering. The task we consider now is asymptotic minimization of a sequence of (general) continuous convex functions $(\varphi_k)_{k\in\mathbb{N}}$, $\varphi_k : \mathbb{R}^N \to [0,\infty)$, over a possible constraint set $(\emptyset \neq) C \subset \mathbb{R}^N$, which is assumed to be closed and convex. In [2], it has been proven that APSM achieves this task under certain mild conditions by generating a sequence $(\mathbf{h}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ (for an initial vector $\mathbf{h}_0 \in \mathbb{R}^N$) recursively by

$$\mathbf{h}_{k+1} := P_C\Big[\mathbf{h}_k + \lambda_k\Big(T_{\mathrm{sp}(\varphi_k)}(\mathbf{h}_k) - \mathbf{h}_k\Big)\Big], \quad k \in \mathbb{N}, \tag{8}$$

where $\lambda_k \in [0,2]$, $k \in \mathbb{N}$, and $T_{\mathrm{sp}(\varphi_k)}$ denotes the subgradient projection relative to φ_k (see Appendix A). APSM reproduces NLMS by letting $C := \mathbb{R}^N$ and $\varphi_k(\mathbf{x}) := d(\mathbf{x}, H_k)$, $\mathbf{x} \in \mathbb{R}^N$, $k \in \mathbb{N}$, with the standard inner product. A useful generalization has been presented in [3]; this makes it possible to take into account multiple convex constraints in the parameter space [3] and also such constraints in multiple domains [43, 44].

3. Variable-Metric Extension of APSM

We extend APSM such that it encompasses the family of adaptive variable-metric projection algorithms, which have remarkable advantages in performance over their constant-metric counterparts. We start with a simplified version of the variable-metric APSM (V-APSM) and show that it includes TDAF, LNAF/QNAF, PAF, and KPAF as its particular examples. We then present the V-APSM that can deal with a convex constraint (the reader who has no need to consider any constraint may skip Section 3.3).

3.1. Variable-Metric Adaptive Projected Subgradient Method without Constraint. We present the simplified V-APSM which does not take into account any constraint (The full version will be presented in Section 3.3). Let $(\mathbb{R}^{N\times N} \ni)\mathbf{G}_k \succ 0$, $k \in \mathbb{N}$; we express by $\mathbf{A} \succ 0$ that a matrix $\mathbf{A}$ is symmetric and positive definite. Define the inner product and its induced norm, respectively, as $\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{G}_k} := \mathbf{x}^T\mathbf{G}_k\mathbf{y}$, for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$, and $\|\mathbf{x}\|_{\mathbf{G}_k} := \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_{\mathbf{G}_k}}$, for all $\mathbf{x} \in \mathbb{R}^N$. For convenience, we regard $\mathbf{G}_k$ as a metric. Recalling the definition, the subgradient projection depends on the inner product (and the norm), thus depending on the metric $\mathbf{G}_k$ (see (A.3) and (A.4) in Appendix A). We therefore specify the metric $\mathbf{G}_k$ employed in the subgradient projection by $T^{(\mathbf{G}_k)}_{\mathrm{sp}(\varphi_k)}$. The simplified variable-metric APSM is given as follows.

Scheme 1 (*Variable-metric APSM without constraint*). Let $\varphi_k : \mathbb{R}^N \to [0,\infty)$, $k \in \mathbb{N}$, be continuous convex functions. Given an initial vector $\mathbf{h}_0 \in \mathbb{R}^N$, generate $(\mathbf{h}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ by

$$\mathbf{h}_{k+1} := \mathbf{h}_k + \lambda_k\Big(T^{(\mathbf{G}_k)}_{\mathrm{sp}(\varphi_k)}(\mathbf{h}_k) - \mathbf{h}_k\Big), \quad k \in \mathbb{N}, \tag{9}$$

where $\lambda_k \in [0,2]$, for all $k \in \mathbb{N}$.

Recalling the linear system model presented in Section 2.1, a simple example of Scheme 1 is given as follows.

Example 1 (*Adaptive variable-metric projection algorithms*). An application of Scheme 1 to

$$\varphi_k(\mathbf{x}) := d_{\mathbf{G}_k}(\mathbf{x}, H_k) := \min_{\mathbf{a}\in H_k}\|\mathbf{x} - \mathbf{a}\|_{\mathbf{G}_k}, \quad \mathbf{x} \in \mathbb{R}^N, k \in \mathbb{N} \tag{10}$$

yields

$$\begin{aligned}\mathbf{h}_{k+1} &:= \mathbf{h}_k + \lambda_k\Big(P^{(\mathbf{G}_k)}_{H_k}(\mathbf{h}_k) - \mathbf{h}_k\Big)\\ &= \mathbf{h}_k - \lambda_k\frac{e_k(\mathbf{h}_k)}{\mathbf{u}_k^T\mathbf{G}_k^{-1}\mathbf{u}_k}\mathbf{G}_k^{-1}\mathbf{u}_k, \quad k \in \mathbb{N}.\end{aligned} \tag{11}$$

Equation (11) is obtained by noting that the normal vector of H_k with respect to the $\mathbf{G}_k$-metric is ${\mathbf{G}_k}^{-1}\mathbf{u}_k$ because $H_k = \{\mathbf{h} \in \mathbb{R}^N : \langle {\mathbf{G}_k}^{-1}\mathbf{u}_k, \mathbf{h}\rangle_{\mathbf{G}_k} = d_k\}$. More sophisticated algorithms than Example 1 can be derived by following the way in [2, 37]. To keep this work as simple as possible for better accessibility, such sophisticated algorithms will be investigated elsewhere.

3.2. Examples of the Metric Design. The TDAF, LNAF/QNAF, PAF, and KPAF algorithms have the common form of (11) with individual design of $\mathbf{G}_k$; interesting relations among TDAF, PAF, and KPAF are given in [34] based on the so-called error surface analysis. The $\mathbf{G}_k$-design in each of the algorithms is given as follows.

(1) Let $\mathbf{V} \in \mathbb{R}^{N\times N}$ be a prespecified transformation matrix such as the discrete cosine transform (DCT) and discrete Fourier transform (DFT). Given $s_0^{(i)} > 0$, $i = 1, 2, \ldots, N$, define $s_{k+1}^{(i)} := \gamma s_k^{(i)} + (\tilde{u}_k^{(i)})^2$, where $\gamma \in (0,1)$ and $[\tilde{u}_k^{(1)}, \tilde{u}_k^{(2)}, \ldots, \tilde{u}_k^{(N)}]^T := \mathbf{V}\mathbf{u}_k$ is the transform-domain input vector. Then, $\mathbf{G}_k$ for TDAF [19, 20] is given as follows:

$$\mathbf{G}_k := \mathbf{V}^T\,\mathrm{diag}\Big(s_k^{(1)}, s_k^{(2)}, \ldots, s_k^{(N)}\Big)\mathbf{V}. \tag{12}$$

Here, $\mathrm{diag}(\mathbf{a})$ denotes the diagonal matrix whose diagonal entries are given by the components of a vector $\mathbf{a} \in \mathbb{R}^N$. This metric is useful for colored input signals.

(2) $\mathbf{G}_k$s for LNAF in [23] and QNAF in [26] are given by $\mathbf{G}_k := \hat{\mathbf{R}}_{k,\mathrm{LN}}$ and $\mathbf{G}_k := \hat{\mathbf{R}}_{k,\mathrm{QN}}$, respectively, where for some initial matrices $\hat{\mathbf{R}}_{0,\mathrm{LN}}$ and $\hat{\mathbf{R}}_{0,\mathrm{QN}}$ their inverses are updated as follows:

$$\hat{\mathbf{R}}_{k+1,\mathrm{LN}}^{-1} := \frac{1}{1-\alpha}\left(\hat{\mathbf{R}}_{k,\mathrm{LN}}^{-1} - \frac{\hat{\mathbf{R}}_{k,\mathrm{LN}}^{-1}\mathbf{u}_k\mathbf{u}_k^T\hat{\mathbf{R}}_{k,\mathrm{LN}}^{-1}}{(1-\alpha)/\alpha + \mathbf{u}_k^T\hat{\mathbf{R}}_{k,\mathrm{LN}}^{-1}\mathbf{u}_k}\right), \quad \alpha \in (0,1),$$
$$\hat{\mathbf{R}}_{k+1,\mathrm{QN}}^{-1} := \hat{\mathbf{R}}_{k,\mathrm{QN}}^{-1} + \left(\frac{1}{2\mathbf{u}_k^T\hat{\mathbf{R}}_{k,\mathrm{QN}}^{-1}\mathbf{u}_k} - 1\right)\frac{\hat{\mathbf{R}}_{k,\mathrm{QN}}^{-1}\mathbf{u}_k\mathbf{u}_k^T\hat{\mathbf{R}}_{k,\mathrm{QN}}^{-1}}{\mathbf{u}_k^T\hat{\mathbf{R}}_{k,\mathrm{QN}}^{-1}\mathbf{u}_k}. \tag{13}$$

The matrices $\hat{\mathbf{R}}_{k,\mathrm{LN}}$ and $\hat{\mathbf{R}}_{k,\mathrm{QN}}$ well approximate the autocorrelation matrix of the input vector $\mathbf{u}_k$, which coincides with the Hessian of the mean squared error (MSE) cost function. Therefore, LNAF/QNAF is a stochastic approximation of the Newton method, yielding faster convergence than the LMS-type algorithms based on the steepest descent method.

(3) Let $\mathbf{h}_k =: [h_k^{(1)}, h_k^{(2)}, \ldots, h_k^{(N)}]^T$, $k \in \mathbb{N}$. Given small constants $\sigma > 0$ and $\delta > 0$, define $L_k^{\max} := \max\{\delta, |h_k^{(1)}|, |h_k^{(2)}|, \ldots, |h_k^{(N)}|\} > 0$, $\gamma_k^{(n)} := \max\{\sigma L_k^{\max}, |h_k^{(n)}|\} > 0$, $n = 1, 2, \ldots, N$, and $\alpha_k^{(n)} := \gamma_k^{(n)} / \sum_{i=1}^{N} \gamma_k^{(i)}$, $n = 1, 2, \ldots, N$. Then, $\mathbf{G}_k$ for the PNLMS algorithm [27, 28] is as follows:

$$\mathbf{G}_k := \mathrm{diag}^{-1}\left(\alpha_k^{(1)}, \alpha_k^{(2)}, \ldots, \alpha_k^{(N)}\right). \tag{14}$$

This metric is useful for sparse unknown systems $\mathbf{h}^*$. The improved proportionate NLMS (IPNLMS) algorithm [31] employs $\gamma_{\mathrm{ip},k}^{(n)} := 2[(1-\omega)\|\mathbf{h}_k\|_1/N + \omega|h_k^{(n)}|]$, $\omega \in [0,1)$, for $n = 1, 2, \ldots, N$ in place of $\gamma_k^{(n)}$; $\|\cdot\|_1$ denotes the ℓ_1 norm. IPNLMS is reduced to the standard NLMS algorithm when $\omega := 0$. Another modification has been proposed in, for example, [32].

(4) Let $\hat{\mathbf{R}}$ and $\hat{\mathbf{p}}$ be the estimates of $\mathbf{R} := E\{\mathbf{u}_k\mathbf{u}_k^T\}$ and $\mathbf{p} := E\{\mathbf{u}_k d_k\}$. Also let $\mathbf{Q} \in \mathbb{R}^{N\times N}$ be a matrix obtained by orthonormalizing (from left to right) the Krylov matrix $[\hat{\mathbf{p}}, \hat{\mathbf{R}}\hat{\mathbf{p}}, \ldots, \hat{\mathbf{R}}^{N-1}\hat{\mathbf{p}}]$. Define $[\widetilde{h}_k^{(1)}, \widetilde{h}_k^{(2)}, \ldots, \widetilde{h}_k^{(N)}]^T := \mathbf{Q}^T\mathbf{h}_k$, $k \in \mathbb{N}$. Given a proportionality factor $\overline{\omega} \in [0,1)$ and a small constant $\varepsilon > 0$, define

$$\beta_k^{(n)} := \frac{1-\overline{\omega}}{N} + \overline{\omega}\frac{\left|\widetilde{h}_k^{(n)}\right|}{\sum_{i=1}^{N}\left|\widetilde{h}_k^{(i)}\right| + \varepsilon} > 0, \quad n = 1, 2, \ldots, N, \quad k \in \mathbb{N}. \tag{15}$$

Then, $\mathbf{G}_k$ for KPNLMS [34] is given as follows:

$$\mathbf{G}_k := \mathbf{Q}\mathrm{diag}^{-1}\left(\beta_k^{(1)}, \beta_k^{(2)}, \ldots, \beta_k^{(N)}\right)\mathbf{Q}^T. \tag{16}$$

This metric is useful even for dispersive unknown systems $\mathbf{h}^*$, as $\mathbf{Q}^T$ sparsifies it. If the input signal is highly colored and the eigenvalues of its autocorrelation matrix are *not* clustered, then this metric is used in combination with the metric of TDAF (see [34]). We mention that this is not exactly the one proposed in [34]. The transformation $\mathbf{Q}^T$ makes the optimal filter into a special sparse system of which only a few first components would have large magnitude and the rest is nearly zero. This information (which is much more than only that the system is sparse) is exploited to reduce the computational complexity.

Finally, we present below the full version of V-APSM, which is an extension of Scheme 1 for dealing with a convex constraint.

3.3. The Variable-Metric Adaptive Projected Subgradient Method—A Treatment of Convex Constraint. We generalize Scheme 1 slightly so as to deal with a constraint set $\mathfrak{K} \subset \mathbb{R}^N$, which is assumed to be closed and convex. Given a mapping $T : \mathbb{R}^N \to \mathbb{R}^N$, $\mathrm{Fix}(T) := \{\mathbf{x} \in \mathbb{R}^N : T(\mathbf{x}) = \mathbf{x}\}$ is called *the fixed point set* of T. The operator $P_{\mathfrak{K}}^{(\mathbf{G}_k)}$, $k \in \mathbb{N}$, which denotes the metric projection onto $\mathfrak{K}$ with respect to the $\mathbf{G}_k$-metric, is 1-*attracting nonexpansive* (with respect to the $\mathbf{G}_k$-metric) with $\mathrm{Fix}(P_{\mathfrak{K}}^{(\mathbf{G}_k)}) = \mathfrak{K}$, for all $k \in \mathbb{N}$ (see Appendix B). It holds moreover that $P_{\mathfrak{K}}^{(\mathbf{G}_k)}(\mathbf{x}) \in \mathfrak{K}$ for any $\mathbf{x} \in \mathbb{R}^N$. For generality, we let $T_k : \mathbb{R}^N \to \mathbb{R}^N$, $k \in \mathbb{N}$, be an η-attracting nonexpansive mapping ($\eta > 0$) with respect to the $\mathbf{G}_k$-metric satisfying

$$T_k(\mathbf{x}) \in \mathfrak{K} = \mathrm{Fix}(T_k), \quad \forall k \in \mathbb{N}, \ \forall \mathbf{x} \in \mathbb{R}^N. \tag{17}$$

The full version of V-APSM is then given as follows.

Scheme 2 (*The Variable-metric APSM*). Let $\varphi_k : \mathbb{R}^N \to [0,\infty)$, $k \in \mathbb{N}$, be continuous convex functions. Given an initial vector $\mathbf{h}_0 \in \mathbb{R}^N$, generate $(\mathbf{h}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ by

$$\mathbf{h}_{k+1} := T_k\left[\mathbf{h}_k + \lambda_k\left(T_{\mathrm{sp}(\varphi_k)}^{(\mathbf{G}_k)}(\mathbf{h}_k) - \mathbf{h}_k\right)\right], \quad k \in \mathbb{N}, \tag{18}$$

where $\lambda_k \in [0,2]$, for all $k \in \mathbb{N}$.

Scheme 2 is reduced to Scheme 1 by letting $T_k := I$ ($\mathfrak{K} = \mathbb{R}^N$), for all $k \in \mathbb{N}$, where I denotes the identity mapping. The form given in (18) was originally presented in [37] without any consideration of the convergence issue. Moreover, a partial convergence analysis for $T_k := I$ was presented in [45] with no proof. In the following section, we present a more advanced analysis for Scheme 2 with a rigorous proof.

4. A Deterministic Analysis

We present a deterministic analysis of Scheme 2. In the analysis, small metric-fluctuations is the key assumption to be employed. The reader not intending to consider any constraint may simply let $\mathfrak{K} := \mathbb{R}^N$.

4.1. Monotone Approximation in the Variable-Metric Sense. We start with the following assumption.

Assumption 1. (a) (Assumption in [2]). There exists $K_0 \in \mathbb{N}$ s.t.

$$\varphi_k^* := \min_{\mathbf{x}\in\mathfrak{K}} \varphi_k(\mathbf{x}) = 0, \quad \forall k \geq K_0,$$
$$\Omega := \bigcap_{k\geq K_0} \Omega_k \neq \emptyset, \tag{19}$$

where

$$\Omega_k := \left\{\mathbf{x} \in \mathfrak{K} : \varphi_k(\mathbf{x}) = \varphi_k^*\right\}, \quad k \in \mathbb{N}. \tag{20}$$

(b) There exist $\varepsilon_1, \varepsilon_2 > 0$ s.t. $\lambda_k \in [\varepsilon_1, 2-\varepsilon_2] \subset (0,2)$, $k \geq K_0$.

The following fact is readily verified.

Fact 1. Under Assumption 1(a), the following statements are equivalent (for $k \geq K_0$):

(a) $\mathbf{h}_k \in \Omega_k$,

(b) $\mathbf{h}_{k+1} = \mathbf{h}_k$,

(c) $\varphi_k(\mathbf{h}_k) = 0$,

(d) $\mathbf{0} \in \partial_{\mathbf{G}_k}\varphi_k(\mathbf{h}_k)$.

V-APSM enjoys a sort of monotone approximation in the $\mathbf{G}_k$-metric sense as follows.

Proposition 1. *Let $(\mathbf{h}_k)_{k\in\mathbb{N}}$ be the vectors generated by Scheme 2. Under Assumption 1, for any $\mathbf{z}_k^* \in \Omega_k$,*

$$\left\|\mathbf{h}_k - \mathbf{z}_k^*\right\|_{\mathbf{G}_k}^2 - \left\|\mathbf{h}_{k+1} - \mathbf{z}_k^*\right\|_{\mathbf{G}_k}^2 \geq \varepsilon_1\varepsilon_2 \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}_k}^2} \tag{21}$$
$$(\forall k \geq K_0 \text{ s.t. } \mathbf{h}_k \notin \Omega_k),$$

$$\left\|\mathbf{h}_k - \mathbf{z}_k^*\right\|_{\mathbf{G}_k}^2 - \left\|\mathbf{h}_{k+1} - \mathbf{z}_k^*\right\|_{\mathbf{G}_k}^2 \geq \frac{\eta\varepsilon_2}{\varepsilon_2 + (2-\varepsilon_2)\eta}\|\mathbf{h}_k - \mathbf{h}_{k+1}\|_{\mathbf{G}_k}^2, \quad \forall k \geq K_0. \tag{22}$$

Proof. See Appendix C. □

Proposition 1 will be used to prove the theorem in the following.

4.2. Analysis under Small Metric-Fluctuations. To prove the deterministic convergence, we need the property of *monotone approximation* in a certain "constant-metric" sense [2]. Unfortunately, this property is not ensured automatically for the adaptive variable-metric projection algorithm unlike the constant-metric one. Indeed, as described in Proposition 1, the *monotone approximation* is only ensured *in the* $\mathbf{G}_k$*-metric* sense at each iteration; this is because the strongly attracting nonexpansivity of T_k and the subgradient projection $T_{\mathrm{sp}(\varphi_k)}^{(\mathbf{G}_k)}$ are both dependent on $\mathbf{G}_k$. Therefore, considerably different metrics may result in totally different directions of update, suggesting that under large metric-fluctuations it would be impossible to ensure the *monotone approximation* in the "constant-metric" sense. Small metric-fluctuations is thus the key assumption to be made for the analysis.

Given any matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, its spectral norm is defined by $\|\mathbf{A}\|_2 := \sup_{\mathbf{x}\in\mathbb{R}^N}\|\mathbf{A}\mathbf{x}\|_2/\|\mathbf{x}\|_2$ [46]. Given $\mathbf{A} \succ 0$, let $\sigma_{\mathbf{A}}^{\min} > 0$ and $\sigma_{\mathbf{A}}^{\max} > 0$ denote its minimum and maximum eigenvalues, respectively; in this case $\|\mathbf{A}\|_2 = \sigma_{\mathbf{A}}^{\max}$. We introduce the following assumptions.

Assumption 2. (a) Boundedness of the eigenvalues of $\mathbf{G}_k$. There exist $\delta_{\min}, \delta_{\max} \in (0,\infty)$ s.t. $\delta_{\min} < \sigma_{\mathbf{G}_k}^{\min} \leq \sigma_{\mathbf{G}_k}^{\max} < \delta_{\max}$, for all $k \in \mathbb{N}$.

(b) Small metric-fluctuations. There exist $(\mathbb{R}^{N\times N} \ni)\mathbf{G} \succ 0$, $K_1 \geq K_0$, $\tau > 0$, and a closed convex set $\Gamma \subseteq \Omega$ s.t. $\mathbf{E}_k := \mathbf{G}_k - \mathbf{G}$ satisfies

$$\frac{\|\mathbf{h}_{k+1} + \mathbf{h}_k - 2\mathbf{z}^*\|_2\|\mathbf{E}_k\|_2}{\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_2} < \frac{\varepsilon_1\varepsilon_2\sigma_{\mathbf{G}}^{\min}\delta_{\min}^2}{(2-\varepsilon_2)^2\sigma_{\mathbf{G}}^{\max}\delta_{\max}} - \tau$$
$$(\forall k \geq K_1 \text{ s.t. } \mathbf{h}_k \notin \Omega_k), \quad \forall \mathbf{z}^* \in \Gamma. \tag{23}$$

We now reach the convergence theorem.

Theorem 1. *Let $(\mathbf{h}_k)_{k\in\mathbb{N}}$ be generated by Scheme 2. Under Assumptions 1 and 2, the following holds.*

(a) *Monotone approximation in the constant-metric sense. For any $\mathbf{z}^* \in \Gamma$,*

$$\|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}}^2 - \|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}}^2 \geq \frac{(2-\varepsilon_2)^2\sigma_{\mathbf{G}}^{\max}}{\delta_{\min}^2}\tau\frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2} \quad (\forall k \geq K_1 \text{ s.t. } \mathbf{h}_k \notin \Omega_k) \tag{24}$$

$$\|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}}^2 - \|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}}^2 \geq \frac{\tau}{\sigma_{\mathbf{G}}^{\max}}\|\mathbf{h}_k - \mathbf{h}_{k+1}\|_{\mathbf{G}}^2, \quad \forall k \geq K_1. \tag{25}$$

(b) *Asymptotic minimization. Assume that $(\varphi_k'(\mathbf{h}_k))_{k\in\mathbb{N}}$ is bounded. Then,*

$$\lim_{k\to\infty}\varphi_k(\mathbf{h}_k) = 0. \tag{26}$$

(c) *Convergence to an asymptotically optimal point. Assume that Γ has a relative interior with respect to a hyperplane $\Pi \subset \mathbb{R}^N$; that is, there exists $\widetilde{\mathbf{h}} \in \Pi \cap \Gamma$ s.t. $\{\mathbf{x} \in \Pi : \|\mathbf{x} - \widetilde{\mathbf{h}}\| < \varepsilon_{r.i.}\} \subset \Gamma$ for some $\varepsilon_{r.i.} > 0$. (The norm $\|\cdot\|$ can be arbitrary due to the norm equivalency for finite-dimensional vector spaces.) Then, $(\mathbf{h}_k)_{k\in\mathbb{N}}$ converges to a point $\widehat{\mathbf{h}} \in \mathfrak{K}$. In addition, under the assumption in Theorem 1(b),*

$$\lim_{k\to\infty}\varphi_k\left(\widehat{\mathbf{h}}\right) = 0 \tag{27}$$

provided that there exists bounded $(\varphi_k'(\widehat{\mathbf{h}}))_{k\in\mathbb{N}}$ where $\varphi_k'(\widehat{\mathbf{h}}) \in \partial_{\mathbf{G}_k}\varphi_k(\widehat{\mathbf{h}})$, for all $k \in \mathbb{N}$.

(d) *Characterization of the limit point. Assume the existence of some interior point* $\widetilde{\mathbf{h}}$ *of* Ω. *In this case, under the assumptions in (c), if for all* $\varepsilon > 0$, *for all* $r > 0$, $\exists \delta > 0$ *s.t.*

$$\inf_{\substack{d(\mathbf{h}_k, \mathrm{lev}_{\le 0}\varphi_k) \ge \varepsilon, \\ \|\widetilde{\mathbf{h}} - \mathbf{h}_k\| \le r, \\ k \ge K_1}} \varphi_k(\mathbf{h}_k) \ge \delta, \tag{28}$$

then $\widehat{\mathbf{h}} \in \overline{\liminf_{k\to\infty} \Omega_k}$, *where* $\liminf_{k\to\infty} \Omega_k := \bigcup_{k=0}^{\infty} \bigcap_{n\ge k} \Omega_n$ *and the overline denotes the closure (see Appendix A for the definition of* $\mathrm{lev}_{\le 0}\varphi_k$*). Note that the metric for* $\|\cdot\|$ *and* $d(\cdot,\cdot)$ *is arbitrary.*

Proof. See Appendix D. □

We conclude this section by giving some remarks on the assumptions and the theorem.

Remark 1 (On Assumption 1). (a) Assumption 1(a) is required even for the simple NLMS algorithm [2].

(b) Assumption 1(b) is natural because the step size is usually controlled so as not to become too large nor small for obtaining reasonable performance.

Remark 2 (On Assumption 2). (a) In the existing algorithms mentioned in Example 1, the eigenvalues of $\mathbf{G}_k$ are controllable directly and usually bounded. Therefore, Assumption 2(a) is natural.

(b) Assumption 2(b) implies that the metric-fluctuations $\|\mathbf{E}_k\|_2$ should be sufficiently small to satisfy (23). We mention that the constant metric (i.e., $\mathbf{G}_k := \mathbf{G} \succ 0$, for all $k \in \mathbb{N}$, thus $\|\mathbf{E}_k\|_2 = 0$) surely satisfies (23): note that $\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_2 \neq 0$ by Fact 1. In the algorithms presented in Example 1, the fluctuations of $\mathbf{G}_k$ tend to become small as the filter adaptation proceeds. If in particular a constant step size $\lambda_k := \lambda \in (0, 2)$, for all $k \in \mathbb{N}$, is used, we have $\varepsilon_1 = \lambda$ and $\varepsilon_2 = 2 - \lambda$ and thus (23) becomes

$$\frac{\|\mathbf{h}_{k+1} + \mathbf{h}_k - 2\mathbf{z}^*\|_2 \|\mathbf{E}_k\|_2}{\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_2} < \left(\frac{2}{\lambda} - 1\right) \frac{\sigma_{\mathbf{G}}^{\min} \delta_{\min}^2}{\sigma_{\mathbf{G}}^{\max} \delta_{\max}} - \tau. \tag{29}$$

This implies that the lower the value of λ is, the larger amount of metric-fluctuations would be acceptable in the adaptation. In Section 5, it will be shown that the use of small λ makes the algorithm relatively insensitive to large metric-fluctuations. Finally, we mention that multiplication of $\mathbf{G}_k$ by any scalar $\xi > 0$ does not affect the assumption, because (i) $\sigma_{\mathbf{G}}^{\min}$, $\sigma_{\mathbf{G}}^{\max}$, $\delta_{\min}$, $\delta_{\max}$, and $\|\mathbf{E}_k\|_2$ in (23) are equally scaled, and (ii) the update equation (23) is unchanged (as $\varphi_k'(\mathbf{x})$ is scaled by $1/\xi$ by the definition of subgradient).

Remark 3 (On Theorem 1). (a) Theorem 1(a) ensures the monotone approximation in the "constant" $\mathbf{G}$-metric sense; that is, $\|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}} \le \|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}}$ for any $\mathbf{z}^* \in \Gamma$. This remarkable property is important for stability of the algorithm.

(b) Theorem 1(b) tells us that the variable-metric adaptive filtering algorithm in (11) asymptotically minimizes the sequence of the metric distance functions $\varphi_k(\mathbf{x}) = d_{\mathbf{G}_k}(\mathbf{x}, H_k)$, $k \in \mathbb{N}$. This intuitively means that the output error $e_k(\mathbf{h}_k)$ diminishes, since H_k is the zero output-error hyperplane. Note however that this does *not* imply the convergence of the sequence $(\mathbf{h}_k)_{k\in\mathbb{N}}$ (see Remark 3(c)). The condition of boundedness is automatically satisfied for the metric distance functions [2].

(c) Theorem 1(c) ensures the convergence of the sequence $(\mathbf{h}_k)_{k\in\mathbb{N}}$ to a point $\widehat{\mathbf{h}} \in \mathfrak{K}$. An example that the NLMS algorithm does not converge without the assumption in Theorem 1(c) is given in [2]. Theorem 1(c) also tells us that the limit point $\widehat{\mathbf{h}}$ minimizes the function sequence φ_k asymptotically; that is, the limit point is asymptotically optimal. In the special case where $n_k = 0$ (for all $k \in \mathbb{N}$) and the autocorrelation matrix of $\mathbf{u}_k$ is nonsingular, $\mathbf{h}^*$ is the unique point that makes $\varphi_k(\mathbf{h}^*) = 0$ for all $k \in \mathbb{N}$. The condition of boundedness is automatically satisfied for the metric distance functions [2].

(d) From Theorem 1(c), we can expect that the limit point $\widehat{\mathbf{h}}$ should be characterized by means of the intersection of Ω_ks, because Ω_k is the set of minimizers of φ_k on $\mathfrak{K}$. This intuition is verified by Theorem 1(d), which provides an explicit characterization of $\widehat{\mathbf{h}}$. The condition in (28) is automatically satisfied for the metric distance functions [2].

5. Numerical Examples

We first show that V-APSM outperforms its constant-metric (or Euclidean-metric) counterpart with the design of $\mathbf{G}_k$ presented in Section 3.2. We then examine the impacts of metric-fluctuations on the performance of adaptive filter by taking PAF as an analogy; recall here that metric-fluctuations were the key in the analysis. We finally consider the case of nonstationary inputs and present numerical studies on the properties of the monotone approximation and the convergence to an asymptotically optimal point (see Theorem 1).

5.1. Variable Metric versus Constant Euclidean Metric. First, we compare TDAF [19, 20] and PAF (specifically, IPNLMS) [31] with their constant-metric counterpart, that is, NLMS. We consider a sparse unknown system $\mathbf{h}^* \in \mathbb{R}^N$ depicted in Figure 3(a) with $N = 256$. The input is the colored signal called USASI and the noise is white Gaussian with the signal-to-noise ratio (SNR) 30 dB, where SNR $:= 10\log_{10}(E\{z_k^2\}/E\{n_k^2\})$ with $z_k := \langle \mathbf{u}_k, \mathbf{h}^* \rangle$ (The USASI signal is a wide sense stationary process and is modeled on the autoregressive moving average (ARMA) process characterized by $H(z) := (1 - z^{-2})/(1 - 1.70223z^{-1} + 0.71902z^{-2})$, $z \in \mathbb{C}$, where $\mathbb{C}$ denotes the set of all complex numbers. In the experiments, the average eigenvalue-spread of the input autocorrelation-matrix was 1.20×10^6.). We set $\lambda_k = 0.2$, for all $k \in \mathbb{N}$, for all algorithms. For TDAF, we set $\gamma = 1 - 10^{-3}$ and employ the DCT matrix for $\mathbf{V}$. For PAF (IPNLMS), we set $\omega = 0.5$. We use the performance measure of MSE $10\log_{10}(E\{e_k^2\}/E\{z_k^2\})$. The expectation operator is approximated by an arithmetic average over 300 independent trials. The results are depicted in Figure 3(b).

Next, we compare QNAF [26] and KPAF [34] with NLMS. We consider the noisy situation of SNR 10 dB and

nonsparse unknown systems $\mathbf{h}^*$ drawn from a normal distribution $\mathcal{N}(0,1)$ randomly at each trial. The other conditions are the same as the first experiment. We set $\lambda_k = 0.02$, for all $k \in \mathbb{N}$, for KPAF and NLMS, and use the same parameters for KPAF as in [34]. Although the use of $\lambda_k = 1.0$ for QNAF is implicitly suggested in [26], we instead use $\lambda_k = 0.04$ with $\hat{\mathbf{R}}_{0,QN}^{-1} = \mathbf{I}$ to attain the same steady-state error as the other algorithms ($\mathbf{I}$ denotes the identity matrix). The results are depicted in Figure 4.

Figures 3 and 4 clearly show remarkable advantages of the V-APSM-based algorithms (TDAF, PAF, QNAF, and KPAF) over the constant-metric NLMS. In both experiments, NLMS suffers from slow convergence because of the high correlation of the input signals. The metric designs of TDAF and QNAF accelerate the convergence by reducing the correlation. On the other hand, the metric design of PAF accomplishes it by exploiting the sparse structure of $\mathbf{h}^*$, and that of KPAF does it by sparsifying the nonsparse $\mathbf{h}^*$.

5.2. Impacts of Metric-Fluctuations on the MSE Performance. We examine the impacts of metric-fluctuations on the MSE performance under the same simulation conditions as the first experiment in Section 5.1. We take IPNLMS because of its convenience in studying the metric-fluctuations as seen below. The metric employed in IPNLMS can be obtained by replacing $\mathbf{h}^*$ in

$$\mathbf{G}_{\text{ideal}} := 2\left(\frac{1}{N}\mathbf{I} + \frac{\operatorname{diag}(|\mathbf{h}^*|)}{\|\mathbf{h}^*\|_1}\right)^{-1} \tag{30}$$

by its instantaneous estimate $\mathbf{h}_k$, where $|\cdot|$ denotes the elementwise absolute-value operator. We can thus interpret that IPNLMS employs an approximation of $\mathbf{G}_{\text{ideal}}$. For ease of evaluating the metric-fluctuations $\|\mathbf{E}_k\|_2$, we employ a test algorithm which employs the metric $\mathbf{G}_{\text{ideal}}$ with cyclic fluctuations as follows:

$$\mathbf{G}_k^{-1} := \mathbf{G}_{\text{ideal}}^{-1} + \frac{\varrho}{N}\operatorname{diag}(\hat{\mathbf{e}}_{\iota(k)}), \quad k \in \mathbb{N}. \tag{31}$$

Here, $\iota(k) := (k \bmod N) + 1 \in \{1, 2, \ldots, N\}$, $k \in \mathbb{N}$, $\varrho \geq 0$ determines the amount of metric-fluctuations, and $\hat{\mathbf{e}}_j \in \mathbb{R}^N$ is a unit vector with only one nonzero component at the jth position. Letting $\mathbf{G} := \mathbf{G}_{\text{ideal}}$, we have

$$\|\mathbf{E}_k\|_2 = \frac{\varrho\left(g_{\text{ideal}}^{\iota(k)}\right)^2}{N + \varrho g_{\text{ideal}}^{\iota(k)}} \in \left[0, g_{\text{ideal}}^{\iota(k)}\right), \quad \forall k \in \mathbb{N}, \tag{32}$$

where g_{ideal}^n, $n \in \{1, 2, \ldots, N\}$, denotes the nth diagonal element of $\mathbf{G}_{\text{ideal}}$. It is seen that (i) for a given $\iota(k)$, $\|\mathbf{E}_k\|_2$ is monotonically increasing in terms of $\varrho \geq 0$, and (ii) for a given ϱ, $\|\mathbf{E}_k\|_2$ is maximized by $g_{\text{ideal}}^{\iota(k)} = \min_{j=1}^N g_{\text{ideal}}^j$.

First, we set $\lambda_k = 0.2$, for all $k \in \mathbb{N}$, and examine the performance of the algorithm for $\varrho = 0, 10, 40$. Figure 5(a) depicts the learning curves. Since the test algorithm has the knowledge about $\mathbf{G}_{\text{ideal}}$ (subject to the fluctuations depending on the ϱ value) from the beginning of adaptation, it achieves faster convergence than PAF (and of course than NLMS). There is a fractional difference between $\varrho = 0$ and $\varrho = 10$, indicating robustness of the algorithm against a moderate amount of metric-fluctuations. The use of $\varrho = 40$, on the other hand, causes the increase of steady-state error and the instability at the end. Meanwhile, the good steady-state performance of IPNLMS suggests that the amount of its metric-fluctuations is sufficiently small.

Next, we set $\lambda_k = 0.1, 0.2, 0.4$, for all $k \in \mathbb{N}$, and examine the MSE performance in the steady-state for each value of $\varrho \in [0, 50]$. For each trial, the MSE values are averaged over 5000 iterations after convergence. The results are depicted in Figure 5(b). We observe the tendency that the use of smaller λ_k makes the algorithm less sensitive to metric-fluctuations. This should not be confused with the well-known relations between the step size and steady-state performance in the standard algorithms such as NLMS. Focusing on $\varrho = 25$ in Figure 5(b), the steady-state MSE of $\lambda_k = 0.2$ is slightly higher than that of $\lambda_k = 0.1$, while the steady-state MSE of $\lambda_k = 0.4$ is unacceptably high compared to that of $\lambda_k = 0.2$. This does not usually happen in the standard algorithms. *The analysis presented in the previous section offers a rigorous theoretical explanation for the phenomena observed in Figure 5.* Namely, the larger the metric-fluctuations or the step size, the more easily Assumption 2(b) is violated, resulting in worse performance. Also, the analysis clearly explains that the use of smaller λ_k allows a larger amount of metric-fluctuations $\|\mathbf{E}_k\|_2$ [see (29)].

5.3. Performance for Nonstationary Input. In the previous subsection, we changed the amount of metric-fluctuations in a cyclic fashion and studied its impacts on the performance. We finalize our numerical studies by considering more practical situations in which Assumption 2(b) is easily violated. Specifically, we examine the performance of TDAF and NLMS for nonstationary inputs of female speech sampled at 8 kHz (see Figure 6(a)). Indeed, TDAF controls its metric to reduce the correlation of inputs, whose statistical properties change dynamically due to the nonstationarity. The metric therefore would tend to fluctuate dynamically by reflecting the change of statistics. For better controllability of the metric-fluctuations, we slightly modify the update of $s_k^{(i)}$ in (12) into $\hat{s}_{k+1}^{(i)} := \hat{\gamma}\hat{s}_k^{(i)} + (1-\hat{\gamma})(\tilde{u}_k^{(i)})^2$ for $\hat{\gamma} \in (0,1)$, $i = 1, 2, \ldots, N$. The amount of metric-fluctuations can be reduced by increasing $\hat{\gamma}$ up to one. Considering the acoustic echo cancellation problem (e.g., [33]), we assume SNR 20 dB and use the impulse response $\mathbf{h}^* \in \mathbb{R}^N$ ($N = 1024$) described in Figure 6(b), which was recorded in a small room.

For all algorithms, we set $\lambda_k = 0.02$. For TDAF, we set (A) $\hat{\gamma} = 1 - 10^{-4}$, (B) $\hat{\gamma} = 1 - 10^{-4.5}$, and (C) $\hat{\gamma} = 1 - 10^{-5}$, and were employ the DCT matrix for $\mathbf{V}$. In noiseless situations, V-APSM enjoys the monotone approximation of $\mathbf{h}^*$ and the convergence to the asymptotically optimal point $\mathbf{h}^*$ under Assumptions 1 and 2 (see Remark 3). To illustrate how these properties are affected by the violation of the assumptions due mainly to the noise and the input nonstationarity, Figure 6(c) plots the system mismatch $10\log_{10}(\|\mathbf{h}_k - \mathbf{h}^*\|_2^2/\|\mathbf{h}^*\|_2^2)$ for one trial. We mention that, although Theorem 1(a) indicates

the monotone approximation in the **G**-metric sense, **G** is unavailable and thus we employ the standard Euclidean metric (note that the convergence does *not* depend on the choice of metric). For (B) $\hat{\gamma} = 1 - 10^{-4.5}$ and (C) $\hat{\gamma} = 1 - 10^{-5}$, it is seen that $\mathbf{h}_k$ is approaching $\mathbf{h}^*$ monotonically. This implies that the monotone approximation and the convergence to $\mathbf{h}^*$ are *not seriously* affected from a practical point of view. For (A) $\hat{\gamma} = 1 - 10^{-4}$, on the other hand, $\mathbf{h}_k$ is approaching $\mathbf{h}^*$ *but not monotonically*. This is because the use of $\hat{\gamma} = 1 - 10^{-4}$ makes Assumption 2(b) violated easily due to the relatively large metric-fluctuations. Nevertheless, the observed nonmonotone approximation of (A) $\hat{\gamma} = 1 - 10^{-4}$ would be acceptable in practice; on its positive side, it yields the great benefit of faster convergence because it reflects the statistics of latest data more than the others.

6. Conclusion

This paper has presented a unified analytic tool named variable-metric adaptive projected subgradient method (V-APSM). The small metric-fluctuations has been the key for the analysis. It has been proven that V-APSM enjoys the invaluable properties of monotone approximation and convergence to an asymptotically optimal point. Numerical examples have demonstrated the remarkable advantages of V-APSM and its robustness against a moderate amount of metric-fluctuations. Also the examples have shown that the use of small step size robustifies the algorithm against a large amount of metric-fluctuations. This phenomenon should be distinguished from the well-known relations between the step size and steady-state performance, and our analysis has offered a rigorous theoretical explanation for the phenomenon. The results give us a useful insight that, in case an adaptive variable-metric projection algorithm suffers from poor steady-state performance, one could either reduce the step size or control the variable-metric such that its fluctuations become smaller. We believe—and it is our future task to prove—that V-APSM serves as a guiding principle to derive effective adaptive filtering algorithms for a wide range of applications.

Appendices

A. Projected Gradient and Projected Subgradient Methods

Let us start with the definitions of a convex set and a convex function. A set $C \subset \mathbb{R}^N$ is said to be *convex* if $\nu\mathbf{x} + (1-\nu)\mathbf{y} \in C$, for all $(\mathbf{x},\mathbf{y}) \in C \times C$, for all $\nu \in (0,1)$. A function $\varphi : \mathbb{R}^N \to \mathbb{R}$ is said to be *convex* if $\varphi(\nu\mathbf{x}+(1-\nu)\mathbf{y}) \le \nu\varphi(\mathbf{x}) + (1-\nu)\varphi(\mathbf{y})$, for all $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$, for all $\nu \in (0,1)$.

A.1. Projected Gradient Method. The projected gradient method [38, 39] is an algorithmic solution to the following convexly constrained optimization:

$$\min_{\mathbf{h}\in C} \varphi(\mathbf{h}), \tag{A.1}$$

where $C \subset \mathbb{R}^N$ is a closed convex set and $\varphi : \mathbb{R}^N \to \mathbb{R}$ a differentiable convex function with its derivative $\varphi' : \mathbb{R}^N \to \mathbb{R}^N$ being κ-*Lipschitzian*: that is, there exists $\kappa > 0$ s.t. $\|\varphi'(\mathbf{x}) - \varphi'(\mathbf{y})\| \le \kappa\|\mathbf{x}-\mathbf{y}\|$, for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^N$. For an initial vector $\mathbf{h}_0 \in \mathbb{R}^N$ and the step size $\lambda \in (0, 2/\kappa)$, the projected gradient method generates a sequence $(\mathbf{h}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ by

$$\mathbf{h}_{k+1} := P_C[\mathbf{h}_k - \lambda\varphi'(\mathbf{h}_k)], \quad k \in \mathbb{N}. \tag{A.2}$$

It is known that the sequence $(\mathbf{h}_k)_{k\in\mathbb{N}}$ converges to an arbitrary solution to the problem (A.1). *If, however, φ is nondifferentiable, how should we do?* An answer to this question has been given by Polyak in 1969 [40], which is described below.

A.2. Projected Subgradient Method. For a continuous (but not necessarily differentiable) convex function $\varphi : \mathbb{R}^N \to \mathbb{R}$, it has been proven that the so-called *projected subgradient method* solves the problem (A.1) iteratively under certain conditions. The interested reader is referred to, for example, [3] for its detailed results. We only explain the method itself, as it is helpful to understand APSM.

What is subgradient, and does it always exist? The subgradient is a generalization of gradient, and it always exists for any continuous (*possibly nondifferentiable*) convex function (To be precise, the subgradient is a generalization of *Gâteaux differential*.). In a differentiable case, the gradient $\varphi'(\mathbf{y})$ at an arbitrary point $\mathbf{y} \in \mathbb{R}^N$ is characterized as the unique vector satisfying $\langle \mathbf{x}-\mathbf{y}, \varphi'(\mathbf{y})\rangle + \varphi(\mathbf{y}) \le \varphi(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^N$. In a nondifferentiable case, however, such a vector is nonunique in general, and the set of such vectors

$$\partial\varphi(\mathbf{y}) := \left\{\mathbf{a} \in \mathbb{R}^N : \langle \mathbf{x}-\mathbf{y}, \mathbf{a}\rangle + \varphi(\mathbf{y}) \le \varphi(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^N\right\} \neq \emptyset \tag{A.3}$$

is called *subdifferential* of φ at $\mathbf{y} \in \mathbb{R}^N$. Elements of the subdifferential $\partial\varphi(\mathbf{y})$ are called *subgradients* of φ at $\mathbf{y}$.

The projected subgradient method is based on *subgradient projection*, which is defined formally as follows (see Figure 7 for its geometric interpretation). Suppose that $\mathrm{lev}_{\le 0}\varphi := \{\mathbf{x} \in \mathbb{R}^N : \varphi(\mathbf{x}) \le 0\} \neq \emptyset$. Then, the mapping $T_{\mathrm{sp}(\varphi)} : \mathbb{R}^N \to \mathbb{R}^N$ defined as

$$T_{\mathrm{sp}(\varphi)} : \mathbf{x} \longmapsto \begin{cases} \mathbf{x} - \dfrac{\varphi(\mathbf{x})}{\|\varphi'(\mathbf{x})\|^2}\varphi'(\mathbf{x}) & \text{if } \varphi(\mathbf{x}) > 0, \\ \mathbf{x} & \text{otherwise} \end{cases} \tag{A.4}$$

is called *subgradient projection* relative to φ, where $\varphi'(\mathbf{x}) \in \partial\varphi(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^N$. For an initial vector $\mathbf{h}_0 \in \mathbb{R}^N$, the projected subgradient method generates a sequence $(\mathbf{h}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ by

$$\mathbf{h}_{k+1} := P_C\Big[\mathbf{h}_k + \lambda_k\Big(T_{\mathrm{sp}(\varphi)}(\mathbf{h}_k) - \mathbf{h}_k\Big)\Big], \quad k \in \mathbb{N}, \tag{A.5}$$

where $\lambda_k \in [0,2]$, $k \in \mathbb{N}$. Comparing (A.2) with (A.4) and (A.5), one can see similarity between the two methods. However, it should be emphasized that $\varphi'(\mathbf{h}_k)$ is (not the gradient but) a subgradient.

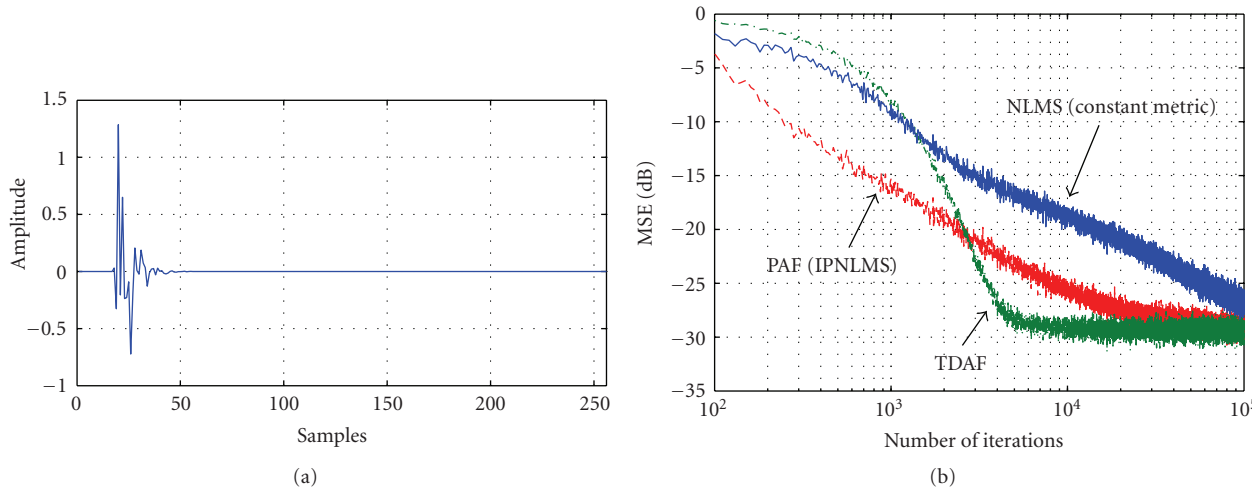

FIGURE 3: (a) Sparse impulse response and (b) MSE performance of NLMS, TDAF, and IPNLMS for $\lambda_k = 0.2$. SNR = 30 dB, $N = 256$, and colored inputs (USASI).

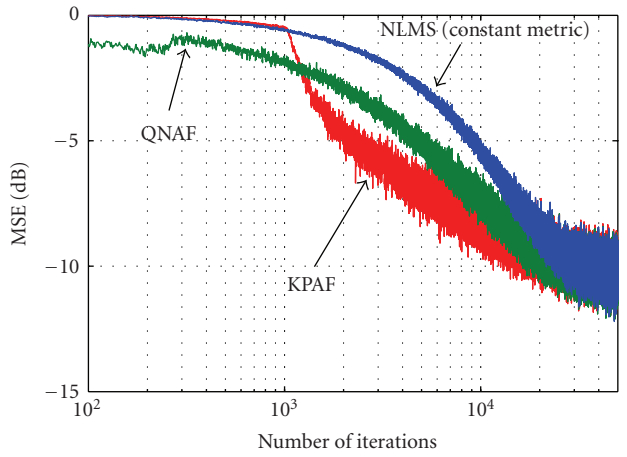

FIGURE 4: MSE performance of NLMS ($\lambda_k = 0.02$), QNAF ($\lambda_k = 0.04$), and KPAF ($\lambda_k = 0.02$) for nonsparse impulse responses and colored inputs (USASI). SNR = 10 dB, $N = 256$.

B. Definitions of Nonexpansive Mappings

(a) A mapping T is said to be *nonexpansive* if $\|T(\mathbf{x}) - T(\mathbf{y})\| \le \|\mathbf{x}-\mathbf{y}\|$, for all $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$; intuitively, T does *not* expand the distance between any two points $\mathbf{x}$ and $\mathbf{y}$.

(b) A mapping T is said to be *attracting nonexpansive* if T is nonexpansive with $\mathrm{Fix}(T) \neq \varnothing$ and $\|T(\mathbf{x}) - \mathbf{f}\|^2 < \|\mathbf{x} - \mathbf{f}\|^2$, for all $(\mathbf{x},\mathbf{f}) \in [\mathbb{R}^N \setminus \mathrm{Fix}(T)] \times \mathrm{Fix}(T)$; intuitively, T attracts any exterior point $\mathbf{x}$ to $\mathrm{Fix}(T)$.

(c) A mapping T is said to be *strongly attracting nonexpansive* or *η- attracting nonexpansive* if T is nonexpansive with $\mathrm{Fix}(T) \neq \varnothing$ and there exists $\eta > 0$ s.t. $\eta\|\mathbf{x} - T(\mathbf{x})\|^2 \le \|\mathbf{x} - \mathbf{f}\|^2 - \|T(\mathbf{x}) - \mathbf{f}\|^2$, for all $(\mathbf{x},\mathbf{f}) \in \mathbb{R}^N \times \mathrm{Fix}(T)$. This condition is stronger than that of *attracting nonexpansivity*, because, for all $(\mathbf{x},\mathbf{f}) \in [\mathbb{R}^N \setminus \mathrm{Fix}(T)] \times \mathrm{Fix}(T)$, the difference $\|\mathbf{x} - \mathbf{f}\|^2 - \|T(\mathbf{x}) - \mathbf{f}\|^2$ is bounded by $\eta\|\mathbf{x} - T(\mathbf{x})\|^2 > 0$.

A mapping $T : \mathbb{R}^N \to \mathbb{R}^N$ with $\mathrm{Fix}(T) \neq \varnothing$ is called *quasi-nonexpansive* if $\|T(\mathbf{x}) - T(\mathbf{f})\| \le \|\mathbf{x} - \mathbf{f}\|$ for all $(\mathbf{x},\mathbf{f}) \in \mathbb{R}^N \times \mathrm{Fix}(T)$.

C. Proof of Proposition 1

Due to the nonexpansivity of T_k with respect to the $\mathbf{G}_k$-metric, (21) is verified by following the proof of [2, Theorem 2]. Noticing the property of the subgradient projection $\mathrm{Fix}(T_{\mathrm{sp}(\varphi_k)}^{(\mathbf{G}_k)}) = \mathrm{lev}_{\le 0}\varphi_k$, we can verify that the mapping $\hat{T}_k := T_k[I + \lambda_k(T_{\mathrm{sp}(\varphi_k)}^{(\mathbf{G}_k)} - I)]$ is $(2-\lambda_k)\eta/(2-\lambda_k(1-\eta))$-attracting quasi-nonexpansive with respect to $\mathbf{G}_k$ with $\mathrm{Fix}(\hat{T}_k) = \mathfrak{K} \cap \mathrm{lev}_{\le 0}\varphi_k = \Omega_k$ (cf. [3]). Because $((2-\lambda_k)\eta)/(2-\lambda_k(1-\eta)) = [1/\eta + (\lambda_k/(2-\lambda_k))]^{-1} = [1/\eta + (2/\lambda_k - 1)^{-1}]^{-1} \ge (\eta\varepsilon_2)/(\varepsilon_2 + (2-\varepsilon_2)\eta)$, (22) is verified.

D. Proof of Theorem 1

Proof of (a). In the case of $\mathbf{h}_k \in \Omega_k$, Fact 1 suggests $\mathbf{h}_{k+1} = \mathbf{h}_k$; thus (25) holds with equality. In the following, we assume $\mathbf{h}_k \notin \Omega_k (\Leftrightarrow \mathbf{h}_{k+1} \neq \mathbf{h}_k)$. For any $\mathbf{x} \in \mathbb{R}^N$, we have

$$\mathbf{x}^T\mathbf{G}_k\mathbf{x} = \left(\frac{\mathbf{y}^T\mathbf{H}_k\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right)\mathbf{x}^T\mathbf{G}\mathbf{x}, \tag{D.1}$$

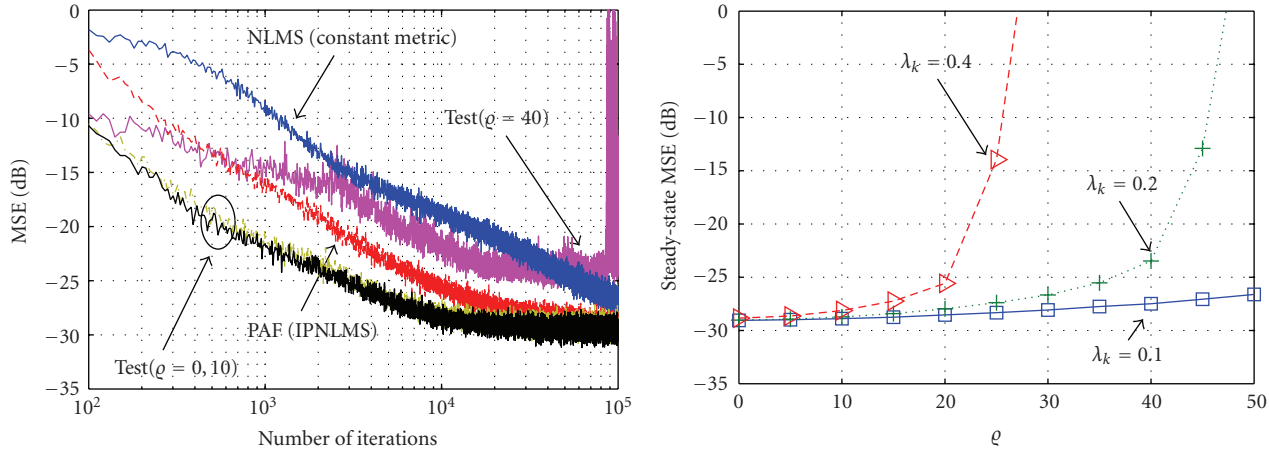

FIGURE 5: (a) MSE learning curves for $\lambda_k = 0.2$ and (b) steady-state MSE values for $\lambda_k = 0.1, 0.2, 0.4$. SNR = 30 dB, $N = 256$, and colored inputs (USASI).

where $\mathbf{y} := \mathbf{G}^{1/2}\mathbf{x}$ and $\mathbf{H}_k := \mathbf{G}^{-1/2}\mathbf{G}_k\mathbf{G}^{-1/2} \succ 0$. By Assumption 2(a), we obtain

$$\sigma_{\mathbf{H}_k}^{\max} = \|\mathbf{H}_k\|_2 \le \left\|\mathbf{G}^{-1/2}\right\|_2 \|\mathbf{G}_k\|_2 \left\|\mathbf{G}^{-1/2}\right\|_2 = \frac{\sigma_{\mathbf{G}_k}^{\max}}{\sigma_{\mathbf{G}}^{\min}} < \frac{\delta_{\max}}{\sigma_{\mathbf{G}}^{\min}}$$

$$\left(\sigma_{\mathbf{H}_k}^{\min}\right)^{-1} = \left\|\mathbf{H}_k^{-1}\right\|_2 \le \left\|\mathbf{G}^{1/2}\right\|_2 \left\|\mathbf{G}_k^{-1}\right\|_2 \left\|\mathbf{G}^{1/2}\right\|_2 = \frac{\sigma_{\mathbf{G}}^{\max}}{\sigma_{\mathbf{G}_k}^{\min}} < \frac{\sigma_{\mathbf{G}}^{\max}}{\delta_{\min}}. \tag{D.2}$$

By (D.1) and (D.2), it follows that

$$\frac{\delta_{\min}}{\sigma_{\mathbf{G}}^{\max}}\|\mathbf{x}\|_{\mathbf{G}}^2 < \|\mathbf{x}\|_{\mathbf{G}_k}^2 < \frac{\delta_{\max}}{\sigma_{\mathbf{G}}^{\min}}\|\mathbf{x}\|_{\mathbf{G}}^2, \quad \forall k \ge K_1,\ \forall \mathbf{x} \in \mathbb{R}^N. \tag{D.3}$$

Noting $\mathbf{E}_k^T = \mathbf{E}_k$, for all $k \ge K_1$ (because $\mathbf{G}_k^T = \mathbf{G}_k$ and $\mathbf{G}^T = \mathbf{G}$), we have, for all $\mathbf{z}^* \in \Gamma \subseteq \Omega \subset \Omega_k$ and (for all $k \ge K_1$ s.t. $\mathbf{h}_k \notin \Omega_k$),

$$\begin{aligned}
&\|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}}^2 - \|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}}^2 \\
&= \|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}_k}^2 - \|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}_k}^2 \\
&\quad - (\mathbf{h}_k - \mathbf{z}^*)^T\mathbf{E}_k(\mathbf{h}_k - \mathbf{z}^*) + (\mathbf{h}_{k+1} - \mathbf{z}^*)^T\mathbf{E}_k(\mathbf{h}_{k+1} - \mathbf{z}^*) \\
&\ge \varepsilon_1\varepsilon_2 \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}_k}^2} + (\mathbf{h}_{k+1} + \mathbf{h}_k - 2\mathbf{z}^*)^T\mathbf{E}_k(\mathbf{h}_{k+1} - \mathbf{h}_k) \\
&\ge \frac{\varepsilon_1\varepsilon_2\sigma_{\mathbf{G}}^{\min}}{\delta_{\max}} \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2} - \|\mathbf{h}_{k+1} + \mathbf{h}_k - 2\mathbf{z}^*\|_2 \|\mathbf{E}_k\|_2 \\
&\quad \times \|\mathbf{h}_{k+1} - \mathbf{h}_k\|_2.
\end{aligned} \tag{D.4}$$

The first inequality is verified by Proposition 1 and the second one is verified by (D.3), the Cauchy-Schwarz inequality, and the basic property of induced norms. Here, $\delta_{\min} < \sigma_{\mathbf{G}_k}^{\min} \le (\mathbf{x}^T\mathbf{G}_k\mathbf{x})/(\mathbf{x}^T\mathbf{x})$ implies

$$\begin{aligned}
\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_2^2 &< (\delta_{\min})^{-1}\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_{\mathbf{G}_k}^2 \\
&\le (\delta_{\min})^{-1}\lambda_k^2 \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}_k}^2} \\
&< \frac{(2-\varepsilon_2)^2\sigma_{\mathbf{G}}^{\max}}{\delta_{\min}^2} \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2},
\end{aligned} \tag{D.5}$$

where the second inequality is verified by substituting $\mathbf{h}_{k+1} = T_k[\mathbf{h}_k - \lambda_k(\varphi_k(\mathbf{h}_k)/\|\varphi_k'(\mathbf{h}_k)\|_{\mathbf{G}_k}^2)\varphi_k'(\mathbf{h}_k)]$ and $\mathbf{h}_k = T_k(\mathbf{h}_k)$ ($\Leftarrow \mathbf{h}_k \in \mathfrak{K} = \mathrm{Fix}(T_k)$; see (17)) and noticing the nonexpansivity of T_k with respect to the $\mathbf{G}_k$-metric. By (D.4), (D.5), and Assumption 2(b), it follows that, for all $\mathbf{z}^* \in \Gamma$, for all $k \ge K_1$ s.t. $\mathbf{h}_k \notin \Omega_k$,

$$\begin{aligned}
&\|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}}^2 - \|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}}^2 \\
&\ge \left(\frac{\varepsilon_1\varepsilon_2\sigma_{\mathbf{G}}^{\min}}{\delta_{\max}} - \frac{\|\mathbf{h}_{k+1} + \mathbf{h}_k - 2\mathbf{z}^*\|_2\|\mathbf{E}_k\|_2}{\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_2} \frac{(2-\varepsilon_2)^2\sigma_{\mathbf{G}}^{\max}}{\delta_{\min}^2}\right) \\
&\quad \times \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2} > \frac{(2-\varepsilon_2)^2\sigma_{\mathbf{G}}^{\max}}{\delta_{\min}^2}\tau \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2}
\end{aligned} \tag{D.6}$$

which verifies (24). Moreover, from (D.3) and (D.5), it is verified that

$$\begin{aligned}
\frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2} &> \frac{\delta_{\min}}{(2-\varepsilon_2)^2\sigma_{\mathbf{G}}^{\max}}\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_{\mathbf{G}_k}^2 \\
&> \frac{1}{(2-\varepsilon_2)^2}\left(\frac{\delta_{\min}}{\sigma_{\mathbf{G}}^{\max}}\right)^2\|\mathbf{h}_{k+1} - \mathbf{h}_k\|_{\mathbf{G}}^2.
\end{aligned} \tag{D.7}$$

By (D.6) and (D.7), we can verify (25). □

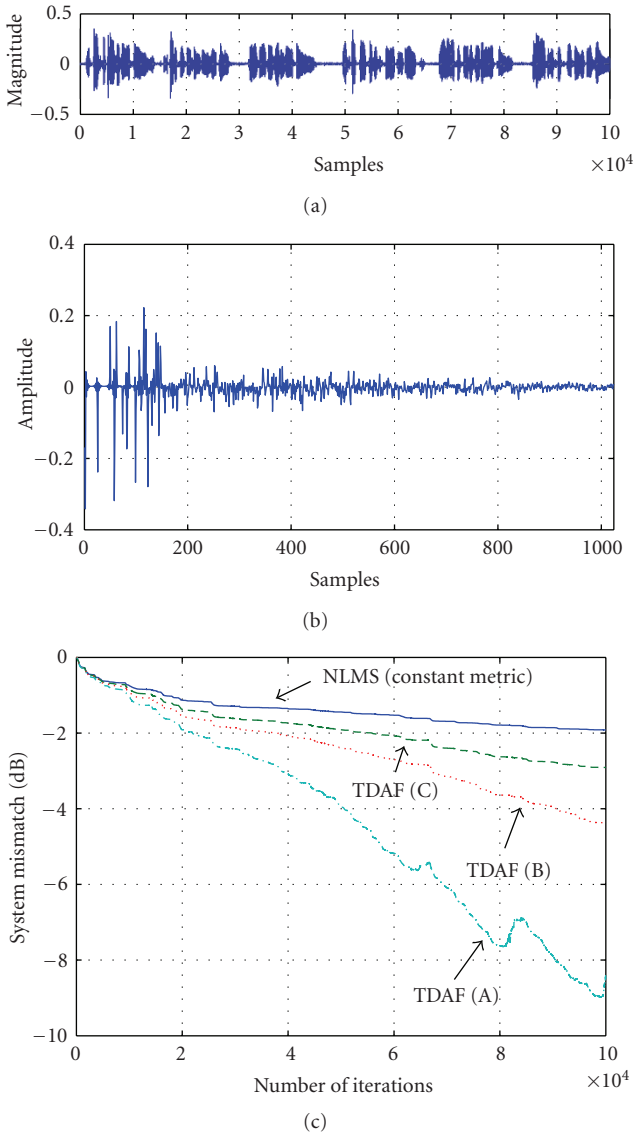

Figure 6: (a) Speech input signal, (b) recorded room impulse response, and (c) system mismatch performance of NLMS and TDAF for $\lambda_k = 0.02$, SNR = 20 dB, and $N = 1024$. For TDAF, (A) $\hat{\gamma} = 1 - 10^{-4}$, (B) $\hat{\gamma} = 1 - 10^{-4.5}$, and (C) $\hat{\gamma} = 1 - 10^{-5}$.

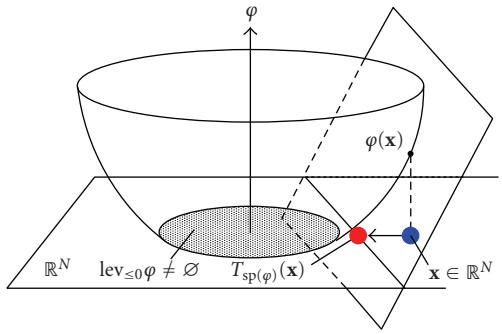

Figure 7: Subgradient projection $T_{\mathrm{sp}(\varphi)}(\mathbf{x}) \in \mathbb{R}^N$ is the projection of $\mathbf{x}$ onto the separating hyperplane (the thick line), which is the intersection of $\mathbb{R}^N$ and the tangent plane at $(\mathbf{x}, \varphi(\mathbf{x})) \in \mathbb{R}^N \times \mathbb{R}$.

Proof of (b). From Fact 1, for proving $\lim_{k\to\infty}\varphi_k(\mathbf{h}_k) = 0$, it is sufficient to check the case $\mathbf{h}_k \notin \Omega_k (\Rightarrow \varphi_k'(\mathbf{h}_k) \neq \mathbf{0})$. In this case, by Theorem 1(a),

$$\begin{aligned} &\|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}}^2 - \|\mathbf{h}_{k+1} - \mathbf{z}^*\|_{\mathbf{G}}^2 \\ &\geq \frac{(2-\varepsilon_2)^2 \sigma_{\mathbf{G}}^{\max}}{\delta_{\min}^2} \tau \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2} \geq 0. \end{aligned} \tag{D.8}$$

For any $\mathbf{z}^* \in \Gamma$, the nonnegative sequence $(\|\mathbf{h}_k - \mathbf{z}^*\|_{\mathbf{G}})_{k\geq K_1}$ is monotonically nonincreasing, thus convergent. This implies that

$$\lim_{\substack{k\to\infty \\ \varphi_k'(\mathbf{h}_k)\neq\mathbf{0}}} \frac{\varphi_k^2(\mathbf{h}_k)}{\left\|\varphi_k'(\mathbf{h}_k)\right\|_{\mathbf{G}}^2} = 0; \tag{D.9}$$

hence the boundedness of $(\varphi_k'(\mathbf{h}_k))_{k\in\mathbb{N}}$ ensures $\lim_{k\to\infty}\varphi_k(\mathbf{h}_k) = 0$. □

Proof of (c). By Theorem 1(a) and [2, Theorem 1], the sequence $(\mathbf{h}_k)_{k\geq K_1}$ converges to a point $\hat{\mathbf{h}} \in \mathbb{R}^N$. The closedness of $\mathfrak{K}(\ni \mathbf{h}_k$, for all $k \in \mathbb{N} \setminus \{0\})$ ensures $\hat{\mathbf{h}} \in \mathfrak{K}$.

By the definition of subgradients and Assumption 2(a), we obtain

$$\begin{aligned} 0 \leq \varphi_k\left(\hat{\mathbf{h}}\right) &\leq \varphi_k(\mathbf{h}_k) - \left\langle \mathbf{h}_k - \hat{\mathbf{h}}, \varphi_k'(\hat{\mathbf{h}}) \right\rangle_{\mathbf{G}_k} \\ &\leq \varphi_k(\mathbf{h}_k) + \left\|\mathbf{h}_k - \hat{\mathbf{h}}\right\|_2 \|\mathbf{G}_k\|_2 \left\|\varphi_k'(\hat{\mathbf{h}})\right\|_2 \\ &< \varphi_k(\mathbf{h}_k) + \delta_{\max}\left\|\mathbf{h}_k - \hat{\mathbf{h}}\right\|_2 \left\|\varphi_k'(\hat{\mathbf{h}})\right\|_2. \end{aligned} \tag{D.10}$$

Hence, noticing (i) Theorem 1(b) under the assumption, (ii) the convergence $\mathbf{h}_k \to \hat{\mathbf{h}}$, and (iii) the boundedness of $(\varphi_k'(\hat{\mathbf{h}}))_{k\in\mathbb{N}}$, it follows that $\lim_{k\to\infty}\varphi_k(\hat{\mathbf{h}}) = 0$. □

Proof of (d). The claim can be verified in the same way as in [2, Theorem 2(d)]. □

Acknowledgment

The authors would like to thank the anonymous reviewers for their invaluable suggestions which improved particularly the simulation part.

References

[1] I. Yamada, "Adaptive projected subgradient method: a unified view for projection based adaptive algorithms," *The Journal of IEICE*, vol. 86, no. 8, pp. 654–658, 2003 (Japanese).

[2] 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, no. 7-8, pp. 593–617, 2004.

[3] K. Slavakis, I. Yamada, and N. Ogura, "The adaptive projected subgradient method over the fixed point set of strongly attracting nonexpansive mappings," *Numerical Functional Analysis and Optimization*, vol. 27, no. 7-8, pp. 905–930, 2006.

[4] J. Nagumo and J. Noda, "A learning method for system identification," *IEEE Transactions on Automatic Control*, vol. 12, no. 3, pp. 282–287, 1967.

[5] A. E. Albert and L. S. Gardner Jr., *Stochastic Approximation and Nonlinear Regression*, MIT Press, Cambridge, Mass, USA, 1967.

[6] T. Hinamoto and S. Maekawa, "Extended theory of learning identification," *Transactions of IEE of Japan*, vol. 95, no. 10, pp. 227–234, 1975 (Japanese).

[7] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electronics & Communications in Japan A*, vol. 67, no. 5, pp. 19–27, 1984.

[8] S. C. Park and J. F. Doherty, "Generalized projection algorithm for blind interference suppression in DS/CDMA communications," *IEEE Transactions on Circuits and Systems II*, vol. 44, no. 6, pp. 453–460, 1997.

[9] J. A. Apolinário Jr., S. Werner, P. S. R. Diniz, and T. I. Laakso, "Constrained normalized adaptive filters for CDMA mobile communications," in *Proceedings of the European Signal Processing Conference (EUSIPCO '98)*, vol. 4, pp. 2053–2056, Island of Rhodes, Greece, September 1998.

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

[11] M. Yukawa and I. Yamada, "Pairwise optimal weight realization—acceleration technique for set-theoretic adaptive parallel subgradient projection algorithm," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4557–4571, 2006.

[12] M. Yukawa, R. L. G. Cavalcante, and I. Yamada, "Efficient blind MAI suppression in DS/CDMA systems by embedded constraint parallel projection techniques," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 8, pp. 2062–2071, 2005.

[13] R. L. G. Cavalcante and I. Yamada, "Multiaccess interference suppression in orthogonal space-time block coded MIMO systems by adaptive projected subgradient method," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1028–1042, 2008.

[14] M. Yukawa, N. Murakoshi, and I. Yamada, "Efficient fast stereo acoustic echo cancellation based on pairwise optimal weight realization technique," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 84797, 15 pages, 2006.

[15] K. Slavakis, S. Theodoridis, and I. Yamada, "Online kernel-based classification using adaptive projection algorithms," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, part 1, pp. 2781–2796, 2008.

[16] K. Slavakis, S. Theodoridis, and I. Yamada, "Adaptive constrained learning in reproducing kernel Hilbert spaces: the robust beamforming case," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4744–4764, 2009.

[17] R. L. G. Cavalcante and I. Yamada, "A flexible peak-to-average power ratio reduction scheme for OFDM systems by the adaptive projected subgradient method," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1456–1468, 2009.

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

[19] S. S. Narayan, A. M. Peterson, and M. J. Narasimha, "Transform domain LMS algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 3, pp. 609–615, 1983.

[20] D. F. Marshall, W. K. Jenkins, and J. J. Murphy, "The use of orthogonal transforms for improving performance of adaptive filters," *IEEE Transactions on Circuits and Systems*, vol. 36, no. 4, pp. 474–484, 1989.

[21] F. Beaufays, "Transform-domain adaptive filters: an analytical approach," *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 422–431, 1995.

[22] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1985.

[23] P. S. R. Diniz, M. L. R. de Campos, and A. Antoniou, "Analysis of LMS-Newton adaptive filtering algorithms with variable convergence factor," *IEEE Transactions on Signal Processing*, vol. 43, no. 3, pp. 617–627, 1995.

[24] B. Farhang-Boroujeny, *Adaptive Filters: Theory and Applications*, John Wiley & Sons, Chichester, UK, 1998.

[25] D. F. Marshall and W. K. Jenkins, "A fast quasi-Newton adaptive filtering algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 7, pp. 1652–1662, 1992.

[26] M. L. R. de Campos and A. Antoniou, "A new quasi-Newton adaptive filtering algorithm," *IEEE Transactions on Circuits and Systems II*, vol. 44, no. 11, pp. 924–934, 1997.

[27] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 508–517, 2000.

[28] S. L. Gay, "An efficient fast converging adaptive filter for network echo cancellation," in *Proceedings of the 32nd Asilomar Conference on Signals, Systems and Computers*, pp. 394–398, Pacific Grove, Calif, USA, November 1998.

[29] T. Gänsler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, 2000.

[30] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, Berlin, Germany, 2001.

[31] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, pp. 1881–1884, Orlando, Fla, USA, May 2002.

[32] H. Deng and M. Doroslovački, "Proportionate adaptive algorithms for network echo cancellation," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1794–1803, 2006.

[33] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing—Signals and Communication Technology*, Springer, Berlin, Germany, 2006.

[34] M. Yukawa, "Krylov-proportionate adaptive filtering techniques not limited to sparse systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 927–943, 2009.

[35] M. Yukawa and W. Utschick, "Proportionate adaptive algorithm for nonsparse systems based on Krylov subspace and constrained optimization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, pp. 3121–3124, Taipei, Taiwan, April 2009.

[36] M. Yukawa and W. Utschick, "A fast stochastic gradient algorithm: maximal use of sparsification benefits under computational constraints," to appear in *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E93-A, no. 2, 2010.

[37] M. Yukawa, K. Slavakis, and I. Yamada, "Adaptive parallel quadratic-metric projection algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1665–1680, 2007.

[38] A. A. Goldstein, "Convex programming in Hilbert space," *Bulletin of the American Mathematical Society*, vol. 70, pp. 709–710, 1964.

[39] E. S. Levitin and B. T. Polyak, "Constrained minimization methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 6, no. 5, pp. 1–50, 1966.

[40] B. T. Polyak, "Minimization of unsmooth functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29, 1969.
[41] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, USA, 4th edition, 2002.
[42] A. H. Sayed, *Fundamentals of Adaptive Filtering*, John Wiley & Sons, Hoboken, NJ, USA, 2003.
[43] M. Yukawa, K. Slavakis, and I. Yamada, "Signal processing in dual domain by adaptive projected subgradient method," in *Proceedings of the 16th International Conference on Digital Signal Processing (DSP '09)*, pp. 1–6, Santorini-Hellas, Greece, July 2009.
[44] M. Yukawa, K. Slavakis, and I. Yamada, "Multi-domain adaptive learning based on feasibility splitting and adaptive projected subgradient method," to appear in *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E93-A, no. 2, 2010.
[45] M. Yukawa and I. Yamada, "Adaptive parallel variable-metric projection algorithm—an application to acoustic ECHO cancellation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 3, pp. 1353–1356, Honolulu, Hawaii, USA, May 2007.
[46] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, NY, USA, 1985.