AN EFFICIENT ROBUST ADAPTIVE FILTERING SCHEME BASED ON PARALLEL SUBGRADIENT PROJECTION TECHNIQUES

Isao Yamada, Konstantinos Slavakis and Kenyu Yamada

Tokyo Institute of Technology,
Dept. of Communications and Integrated Systems,
Ookayama, Meguro-ku, Tokyo 152-8552, JAPAN
E-mails: {isao,slavakis}@ss.titech.ac.jp

ABSTRACT

This paper presents a novel robust adaptive filtering scheme based on the interactive use of statistical noise information and an extension of the ideas developed originally for efficient algorithmic solutions to the convex feasibility problems. The statistical noise information is quantitatively formulated as stochastic property closed convex sets by the simple design formulae developed in this paper. The proposed adaptive algorithm is computationally efficient and robust to noise because it requires only an iterative parallel projection onto a series of closed half spaces highly expected to contain the unknown system to be identified. The numerical examples show that the proposed adaptive filtering scheme achieves low estimation error and realizes dramatically fast and stable convergence even for highly colored excited input signals in severely noisy situations.

1. INTRODUCTION

In this paper, a novel direction toward efficient robust adaptive filtering is addressed by recasting the similarity between one of the major adaptive algorithms, the *affine projection algorithm* (APA) [1, 2, 3, 4, 5], and a standard algorithmic solution (POCS) [6] to the convex feasibility problem for finding a point in the intersection of a family of closed convex sets in a real Hilbert space [7, 8, 9, 10].

With the intent of improving the convergence speed of the classical *normalized least mean squares* algorithm (NLMS) [11, 12] mainly for highly colored excited input signals, the APA was proposed as an iterative relaxed projection onto a series of linear varieties $V_k \subset \mathbb{R}^N$ $(k \in \mathbb{Z})$ where the unknown system parameter $\boldsymbol{h}^* \in \mathbb{R}^N$ surely exists in noiseless situations (See (3) for the definition of the APA). Each linear variety V_k is generated as the intersection of certain number, say $r \in \mathbb{N}^*$, of hyperplanes determined by the *instantaneous* input-output relations of the unknown system to be estimated (The NLMS, known as a variation of the standard LMS, is the simplest APA, for $r = 1$, based on an iterative relaxed projection onto a series of hyperplanes themselves. The rate of convergence of the NLMS is potentially faster than that of the standard LMS algorithm for both uncorrelated and correlated input signals [11, 12]. For recent advance on the convergence analysis of the variations of the LMS algorithm, see for example [13] and the references therein). Although it has been reported that the increase of the number r of the hyperplanes to generate V_k improves the convergence speed of APA, in particular for highly colored excited input signals in relatively high SNR, the increase of r causes a serious growth of the computational complexity. To save the computational cost of APA, a great deal of effort has been devoted for example by using a sliding windowed FRLS [4] or selective-partial-updating techniques [3, 5].

Obviously the underlying idea of the APA is to efficiently find a vector, as an estimate of the unknown system parameter $\boldsymbol{h}^*$ to be identified, in the intersection of hyperplanes expected to contain $\boldsymbol{h}^*$. From this point of view, the APA can be seen as a variation of the classical *Kaczmarz's iterative method* for system of linear equations [8] or its well-known generalization POCS [6]. Although there is a big difference between APA and POCS because POCS repeatedly uses all the linear varieties (more generally all the *closed convex sets*) but the APA does not, one of the most important properties, the *monotonicity of the estimation errors* (in noiseless case) of APA, is essentially derived from the same mathematical reason as that for the *Fejér-monotonicity* of the POCS [8].

However, in practice, the observed signal is corrupted by the additive noise and thus each hyperplane has low reliability to contain the system parameter $\boldsymbol{h}^*$—this causes serious instability or notable decline of the convergence in the learning process of the APA in relatively low SNR.

The apparent similarity of APA and POCS suggests that a straightforward strategy to overcome the sensitiveness of APA to noise would be the introduction of the idea of POCS and the replacement of the linear varieties in APA by alternative closed convex sets with higher reliability to contain the system parameter $\boldsymbol{h}^*$. Indeed, as will be shown in the next section, based on the statistical information on the additive noise, a set-theoretic formulation provides such a closed convex set by $C_i := \{\boldsymbol{x} \in \mathbb{R}^N : g_i(\boldsymbol{x}) \leq 0\}$ where $g_i : \mathbb{R}^N \to \mathbb{R}$ is a differentiable convex function for all $i \in I \subset \mathbb{Z}$. However since such a closed convex set C_i is not simple enough to have closed form expression of the exact projection onto itself, the necessary computation for such a projection is much more costly than that of the orthogonal projection onto the linear variety in the original APA. To circumvent the shortcoming of POCS caused by the use of exact projections, the subgradient projection methods have been developed as alternative algorithmic solutions to the convex feasibility problems [7, 8, 14]. Since the subgradient projection methods do not require the exact projection onto each closed convex set C_i but the projection onto closed halfspaces defined simply by the gradient or subgradient of the convex function g_i, significant reduction of the computational cost is achieved.

In this paper we propose a novel adaptive filtering scheme by extending the ideas of an extrapolated parallel version of the subgradient projection methods [14] for the iterative parallel incorporation of the noise information into the instantaneous input-output relations. The monotonicity of the estimation error by the proposed adaptive scheme is highly expected because of elementary facts on the subgradient projection method. Moreover a simple design of closed convex sets of significant effectiveness to the proposed scheme is presented in this paper which also demonstrates an essential reason why the APA is sensitive to noise.

The numerical examples show that the proposed algorithm achieves low estimation error and realizes dramatically fast and stable convergence even for highly colored excited speech like input signals in severely noisy situations, which is hard task even for the RLS [12] because RLS suffers from certain model mismatch problem causing serious degradation of the learning performance [12].

2. PRELIMINARIES

Let $\mathbb{Z}, \mathbb{N}$ and $\mathbb{R}$ denote the sets of all integers, nonnegative integers and real numbers respectively. Define, also, $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$. Given $N \in \mathbb{N}^*$, we consider the Euclidean space $\mathcal{H} := \mathbb{R}^N$ that is a real Hilbert space equipped with the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^t \boldsymbol{y}$, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$, and its induced norm $\|\boldsymbol{x}\| := \left(\boldsymbol{x}^t \boldsymbol{x}\right)^{1/2}$, $\forall \boldsymbol{x} \in \mathcal{H}$, where the superscript t stands for transposition (We will also use the same notation $\|\cdot\|$ for the standard norm defined on the different finite dimensional space $\mathbb{R}^r, r \in \mathbb{N}^*$). A set $C \subset \mathcal{H}$ is *convex* provided that $\forall \boldsymbol{x}, \boldsymbol{y} \in C$, $\forall \nu \in [0,1]$, $\nu \boldsymbol{x} + (1-\nu)\boldsymbol{y} \in C$ (Note that a linear variety is a closed convex set). For any closed convex set $C \subset \mathcal{H}$, the *projection operator* $P_C : \mathcal{H} \to C$ is defined by $\|\boldsymbol{x} - P_C(\boldsymbol{x})\| = \min_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x} \in \mathcal{H}$. The mapping $T_C := I + \lambda(P_C - I)$, where $I : \mathcal{H} \to \mathcal{H}$ is the *identity operator*, for $\lambda \in (0,2)$ is called the *relaxed projection*. A function $g : \mathcal{H} \to \mathbb{R}$ is said to be *convex* if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$ and $\forall \nu \in [0,1]$, $g(\nu \boldsymbol{x} + (1-\nu)\boldsymbol{y}) \leq \nu g(\boldsymbol{x}) + (1-\nu) g(\boldsymbol{y})$. The *subdifferential* of a convex function g at $\boldsymbol{y}$, denoted $\partial g(\boldsymbol{y})$, is the set of all the *subgradients* of g at $\boldsymbol{y}$: $\partial g(\boldsymbol{y}) := \left\{\boldsymbol{s} \in \mathcal{H} : (\boldsymbol{x} - \boldsymbol{y})^t \boldsymbol{s} + g(\boldsymbol{y}) \leq g(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{H}\right\}$. The convex function $g : \mathcal{H} \to \mathbb{R}$ has a unique subgradient at $\boldsymbol{y} \in \mathcal{H}$ if g is (Gâteaux) differentiable at $\boldsymbol{y}$ [8]. This unique subgradient is nothing but the *gradient* $\nabla g(\boldsymbol{y})$, i.e., $\partial g(\boldsymbol{y}) = \{\nabla g(\boldsymbol{y})\}$.

In this paper we elaborate on the following *adaptive filtering* estimation scheme. Let $(u_k)_{k \in \mathbb{Z}} \subset \mathbb{R}$. Define the *input sequence* $(\boldsymbol{u}_k)_{k \in \mathbb{Z}} \subset \mathcal{H}$ as $\boldsymbol{u}_k := [u_k, u_{k-1}, \ldots, u_{k-N+1}]^t \in \mathcal{H}, \forall k \in \mathbb{Z}$. For $r \in \mathbb{N}^*$, let $\boldsymbol{U}_k := [\boldsymbol{u}_{k-r+1}, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{N \times r}$. Its *column* and *row spaces* are denoted by $\mathcal{R}(\boldsymbol{U}_k)$ and $\mathcal{R}\left(\boldsymbol{U}_k^t\right)$ respectively. Let, also, $(n_k)_{k \in \mathbb{Z}} \subset \mathbb{R}$ denote the *noise process*. If $\boldsymbol{h}^* \in \mathcal{H}$ stands for the system to be estimated, or *estimandum*, and $\boldsymbol{n}_k := [n_k, \ldots, n_{k-r+1}]^t \in \mathbb{R}^r$, $\forall k \in \mathbb{Z}$, we introduce the following linear model for the *data process* $(\boldsymbol{d}_k)_{k \in \mathbb{Z}} \subset \mathbb{R}^r$:

$$\boldsymbol{d}_k := \boldsymbol{U}_k^t \boldsymbol{h}^* + \boldsymbol{n}_k, \quad \text{a.s.}, \quad \forall k \in \mathbb{Z}, \tag{1}$$

where a.s. stands for *almost surely*. Let $\boldsymbol{h} \in \mathcal{H}$ denote an *estimate* of $\boldsymbol{h}^*$. We define then the *estimation residual* functions $\boldsymbol{e}_k : \mathcal{H} \to \mathbb{R}^r$, $k \in \mathbb{Z}$, by

$$\boldsymbol{e}_k(\boldsymbol{h}) := \boldsymbol{U}_k^t \boldsymbol{h} - \boldsymbol{d}_k, \quad \text{a.s.}, \quad \forall k \in \mathbb{Z}. \tag{2}$$

The APA scheme generates a sequence $(\boldsymbol{h}_k)_{k \in \mathbb{Z}} \subset \mathbb{R}^N$, as the estimates of $\boldsymbol{h}^*$ in (1), by

$$\boldsymbol{h}_{k+1} := \boldsymbol{h}_k + \lambda_k \left(P_{V_k}(\boldsymbol{h}_k) - \boldsymbol{h}_k\right), \quad \forall k \in \mathbb{Z}, \tag{3}$$

where $\lambda_k \in (0,2)$ and V_k is the linear variety defined by $V_k := \arg\min_{\boldsymbol{h} \in \mathcal{H}} \left\|\boldsymbol{U}_k^t \boldsymbol{h} - \boldsymbol{d}_k\right\|$ (NOTE: The equivalence between (3) and the original formulation in [2] is obvious). For highly colored excited signal $(\boldsymbol{u}_k)_{k \in \mathbb{Z}}$ in relatively high SNR, it has been reported that the use of $r > 1$ significantly improves the convergence speed of $(\boldsymbol{h}_k)$ generated by the NLMS algorithm defined by (3) with $r = 1$. Another characterization of V_k can be given by noticing that $\delta_k^{1/2} := \min_{\boldsymbol{h} \in \mathcal{H}} \left\|\boldsymbol{U}_k^t \boldsymbol{h} - \boldsymbol{d}_k\right\| = \left\|P_{\mathcal{R}\left(\boldsymbol{U}_k^t\right)}(\boldsymbol{d}_k) - \boldsymbol{d}_k\right\|$. Clearly, if $\boldsymbol{d}_k \in \mathcal{R}\left(\boldsymbol{U}_k^t\right) \subseteq \mathbb{R}^r$, then $\delta_k = 0$. By the fact $\boldsymbol{U}_k^t \in \mathbb{R}^{r \times N}$, we deduce that if $N \gg r$, a condition which complies with the demands of nowadays applications, then $\mathcal{R}\left(\boldsymbol{U}_k^t\right)$ will most likely extend over the space $\mathbb{R}^r$ and thus include $\boldsymbol{d}_k$ or at least be large enough to force δ_k take values close to zero.

For the noiseless case, a simple inspection leads to

$$\|\boldsymbol{h}_{k+1} - \boldsymbol{h}^*\| = \left\|(\boldsymbol{h}_k - \boldsymbol{h}^*) - \lambda_k P_{\mathcal{R}(\boldsymbol{U}_k)}(\boldsymbol{h}_k - \boldsymbol{h}^*)\right\|, \tag{4}$$

which is minimized at $\lambda_k = 1$. Moreover, for any $\lambda_k \in [0,2]$, the *monotonicity* property follows:

$$\|\boldsymbol{h}^* - \boldsymbol{h}_{k+1}\| \leq \|\boldsymbol{h}^* - \boldsymbol{h}_k\|, \quad \forall k \in I \subset \mathbb{Z}, \tag{5}$$

being one of the most desired properties for adaptive filtering. The property (5) is guaranteed due to the membership $\boldsymbol{h}^* \in \bigcap_{k \in I} V_k \neq \emptyset$, or more generally due to the following lemma:

Lemma 1 *Let $C_{k+1} \subset \mathcal{H}$ be a closed convex set satisfying $\boldsymbol{h}^* \in C_{k+1} \neq \emptyset$. For any $\boldsymbol{h}_k \in \mathcal{H}$ and any $\lambda_k \in [0,2]$, define $\boldsymbol{h}_{k+1} := \boldsymbol{h}_k + \lambda_k \left(P_{C_{k+1}}(\boldsymbol{h}_k) - \boldsymbol{h}_k\right)$. Then, $\|\boldsymbol{h}^* - \boldsymbol{h}_{k+1}\| \leq \|\boldsymbol{h}^* - \boldsymbol{h}_k\|$. In particular, $\|\boldsymbol{h}^* - \boldsymbol{h}_{k+1}\| < \|\boldsymbol{h}^* - \boldsymbol{h}_k\|$ for $\lambda_k \in (0,2)$, if $\boldsymbol{h}_k \notin C_{k+1}$.*

However, in practice, the inescapable presence of noise $(\boldsymbol{n}_k)$ randomly dislocates each linear variety from its original position, which in general makes the membership $\boldsymbol{h}^* \in \bigcap_{k \in I} V_k \neq \emptyset$ questionable. This is the main reason why the performance of the APA is strongly influenced by additive noise. A very small relaxation parameter $\lambda_k (\approx 0.05)$ has been empirically chosen even for SNR $= 30$dB [3]. To design therefore a robust adaptive filtering scheme, the information about the noise process has to be incorporated somehow into the formulation—a strategy that is not followed in the APA. We do so by using the set theoretic estimation frame [15].

Suppose that we are in the situation where $(n_k)_{k \in \mathbb{Z}}$ is the noise process of zero mean, independent identically distributed Gaussian random variables $\mathcal{N}\left(0, \sigma^2\right)$, and where the estimandum $\boldsymbol{h}^*$ is perfectly estimated, i.e., $\boldsymbol{h}_k = \boldsymbol{h}^*$. Then, it becomes obvious by (1) and (2) that $\boldsymbol{e}_k(\boldsymbol{h}^*) = -\boldsymbol{n}_k$, a.s., $\forall k \in \mathbb{Z}$. Thus, the stochastic processes $(\boldsymbol{e}_k(\boldsymbol{h}^*))_{k \in \mathbb{Z}}$ and $(-\boldsymbol{n}_k)_{k \in \mathbb{Z}}$ have the same probability theoretic properties. As a result, the random variable $\xi := \|\boldsymbol{e}_k(\boldsymbol{h}^*)\|^2$ follows the χ^2 statistic whose probability density function (pdf) is given by

$$f_r(\xi) = \begin{cases} \frac{1}{(\sigma\sqrt{2})^r \Gamma(r/2)} \xi^{(r-2)/2} e^{-\xi/2\sigma^2}, & \text{for } \xi > 0, \\ 0, & \text{for } \xi \leq 0. \end{cases} \tag{6}$$

The pdf $f_r(\xi)$ of (6) is strictly monotone decreasing over $\xi \geq 0$ for $r = 1, 2$, whereas for $r \geq 3$ it has its unique maximum at $\xi = (r-2)\sigma^2$ and $f_r(0) = \lim_{\xi \to \infty} f_r(\xi) = 0$. Moreover the mean and the variance of ξ are given respectively by $m_\xi = r\sigma^2$ and $\sigma_\xi^2 = 2r\sigma^4$ [16]. In this case, the probability theoretic property of the process $(\boldsymbol{n}_k)_{k \in \mathbb{Z}} \subset \mathbb{R}^r$ is quantitatively formulated by the following *stochastic property set*.

$$C_k(\rho) := \left\{\boldsymbol{h} \in \mathcal{H} : \left\|\boldsymbol{U}_k^t \boldsymbol{h} - \boldsymbol{d}_k\right\|^2 - \rho \leq 0\right\}, \tag{7}$$

where $\rho \geq 0$ determines the reliability on the membership $\boldsymbol{h}^* \in C_k(\rho)$ by $\int_0^\rho f_r(\xi) d\xi \in [0,1]$ (For non-Gaussian process $(\boldsymbol{n}_k)_{k \in \mathbb{Z}}$, the asymptotic Gaussian approximation of ξ based on Lindeberg's *central limit theorem* was used to evaluate the above integral in set theoretic estimation schemes [17]). By (7) and the convexity of the function $\|\cdot\|^2$, we remark that $C_k(\rho) \neq \emptyset$ iff $\delta_k = \min_{\boldsymbol{h} \in \mathcal{H}} \left\|\boldsymbol{U}_k^t \boldsymbol{h} - \boldsymbol{d}_k\right\|^2 \leq \rho$. Thus, by increasing ρ and by letting $N \gg r$ according to the discussion following (3), we can highly expect $\boldsymbol{h}^* \in \bigcap_{k \in J} C_k(\rho) \neq \emptyset$, where J is a sufficiently large subset of $\mathbb{Z}$, and thus the stable convergence of $(\boldsymbol{h}_k)_k$ due to Lemma 1. Obviously the APA with $\lambda_k = 1$ is based on the projection onto $C_k(\delta_k) = V_k$ in (3). The pdf $f_r(\xi)$ clearly demonstrates that the APA for $r \geq 3$ can hardly ensure the desired relation $\boldsymbol{h}^* \in \bigcap_{k \in J} V_k \neq \emptyset$, even for a small subset $J \subset \mathbb{Z}$, causing thus serious decline of the convergence of the APA. On the other hand, by the properties of the pdf $f_1(\xi)$ of (6), we can expect stable convergence for NLMS with $\lambda_k = 1$ regardless noisy situations because it projects onto $C_k(\delta_k)$ $(k \in \mathbb{Z})$ for $r = 1$ ($\delta_k = 0$ for $\boldsymbol{u}_k \neq \boldsymbol{0}$, $C_k(\delta_k) = \mathcal{H}$ for $\boldsymbol{u}_k = \boldsymbol{0}$). This agrees with the H^∞ optimality of the NLMS [18].

In the next section, we will show that the important idea of finding a point in the intersection of a family of closed convex sets can also serve as the key to realize an efficient adaptive filtering scheme. The robustness of the proposed adaptive filtering design can be achieved by the interactive use of ρ in $C_k(\rho)$ based on the variance of the additive noise. The problem regarding the suitable choice of $\rho \geq 0$ in (7) will be also addressed. It will be shown that by assigning simple values to ρ, which are directly related to r and to the variance of the noise process, low estimation error can be realized, and faster as well as stabler convergence than the one for APA, NLMS or RLS can be achieved even for severely noisy situations.

3. ADAPTIVE ALGORITHM AND DESIGN OF THE STOCHASTIC PROPERTY SETS

Given $q \in \mathbb{N}^*$, define $I_k := \left\{\iota_1^{(k)}, \iota_2^{(k)}, \ldots, \iota_q^{(k)}\right\} \subset \mathbb{N}, \forall k \in \mathbb{N}$, and $w_\iota > 0,\ \forall \iota \in I_k, \forall k \in \mathbb{N}$, to satisfy $\sum_{\iota \in I_k} w_\iota = 1, \forall k \in \mathbb{N}$. Let $I := \bigcup_{k\in\mathbb{N}} I_k$. We propose the following adaptive algorithm and a proposition regarding the monotonicity. The algorithm is based on the formulation given by Pierra [19] which was also used in [14]. Since the proposed algorithm does not require the repetitive use of the information on the same closed convex set $C_\iota(\rho)$, it can be seen as an extension of the algorithms [7, 8, 14, 19] for the convex feasibility problem. The proof of the proposition is omitted due to lack of space.

Algorithm 1 *(Adaptive parallel outer projection algorithm) Suppose that a sequence of closed convex sets $(C_\iota(\rho))_{\iota\in I} \subset \mathcal{H}$ is defined as in (7). Let $\boldsymbol{h}_0 \in \mathcal{H}$ and define a sequence $(\boldsymbol{h}_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ by*

$$\boldsymbol{h}_{k+1} = \boldsymbol{h}_k + \mu_k \left(\sum_{\iota \in I_k} w_\iota P_{S_\iota}(\boldsymbol{h}_k) - \boldsymbol{h}_k \right), \quad \forall k \in \mathbb{N}, \quad (8)$$

where $S_\iota \subset \mathcal{H}$ is a closed convex set satisfying

$$C_\iota(\rho) \subset S_\iota \quad \textit{and} \quad \boldsymbol{h}_k \notin C_\iota(\rho) \Rightarrow \boldsymbol{h}_k \notin S_\iota, \quad (9)$$

and the relaxation parameter $\mu_k \in [0, 2\mathcal{M}_k]$, *where*

$$\mathcal{M}_k := \begin{cases} \dfrac{\sum_{\iota\in I_k} w_\iota \left\| P_{S_\iota}(\boldsymbol{h}_k) - \boldsymbol{h}_k \right\|^2}{\left\| \sum_{\iota\in I_k} w_\iota P_{S_\iota}(\boldsymbol{h}_k) - \boldsymbol{h}_k \right\|^2}, & \boldsymbol{h}_k \notin \bigcap_{\iota\in I_k} S_\iota, \\ 1, & \textit{otherwise}. \end{cases} \quad (10)$$

(NOTES: $\mathcal{M}_k \geq 1$ by the convexity of $\|\cdot\|^2$. By normalizing an algorithm in [20], a special case of (8) for $S_\iota := C_\iota(\delta_\iota) = V_\iota$, $\forall \iota \in I_k, \forall k \in \mathbb{N}$, with $r = 1$ was derived in [21].)

Proposition 1 *For any $\boldsymbol{h}_0 \in \mathcal{H}$, let the sequence $(\boldsymbol{h}_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ generated by Algorithm 1. Then, for any $\boldsymbol{h}^* \in \bigcap_{\iota\in I_k} S_\iota$ (this of course holds if $\boldsymbol{h}^* \in \bigcap_{\iota\in I_k} C_\iota(\rho)$), $\|\boldsymbol{h}^* - \boldsymbol{h}_{k+1}\| \leq \|\boldsymbol{h}^* - \boldsymbol{h}_k\|$. If, in particular, $\mu_k \in (0, 2\mathcal{M}_k)$ and $\boldsymbol{h}_k \notin \bigcap_{\iota\in I_k} S_\iota$ ($\Leftarrow \boldsymbol{h}_k \notin \bigcap_{\iota\in I_k} C_\iota(\rho)$ by (9)), then for any $\boldsymbol{h}^* \in \bigcap_{\iota\in I_k} S_\iota$, $\|\boldsymbol{h}^* - \boldsymbol{h}_{k+1}\| < \|\boldsymbol{h}^* - \boldsymbol{h}_k\|$.*

Obviously, the critical point in Algorithm 1 is the systematic generation of the sequence of closed convex sets $S_\iota \subset \mathcal{H}$, $\iota \in I$, each of which must be simple enough to have closed form expression of P_{S_ι} as well as must satisfy (9). Such a systematic set generation is realized by applying the following elementary result on the *subgradient* to the stochastic property set in (7).

Lemma 2 *Suppose that a closed convex set $C \subseteq \mathcal{H}$ is defined by a convex function $g : \mathcal{H} \to \mathbb{R}$ as $C := \{\boldsymbol{x} \in \mathcal{H} : g(\boldsymbol{x}) \leq 0\}$. For $\boldsymbol{y} \in \mathcal{H}$, define the closed half space $H^-(\boldsymbol{y}) := \{\boldsymbol{x} \in \mathcal{H} : (\boldsymbol{x} - \boldsymbol{y})^t \boldsymbol{s} + g(\boldsymbol{y}) \leq 0\}$, where $\boldsymbol{s} \in \partial g(\boldsymbol{y})$. Then, $C \subset H^-(\boldsymbol{y})$ and $\boldsymbol{y} \notin C \Rightarrow \boldsymbol{y} \notin H^-(\boldsymbol{y})$.*

Now, define the convex functions $g_\iota : \mathcal{H} \to \mathbb{R}$, $\iota \in I$, by $g_\iota(\boldsymbol{h}) := \left\| \boldsymbol{U}_\iota^t \boldsymbol{h} - \boldsymbol{d}_\iota \right\|^2 - \rho$. The function g_ι is differentiable everywhere and $\nabla g(\boldsymbol{h}) = 2\boldsymbol{U}_\iota \left(\boldsymbol{U}_\iota^t \boldsymbol{h} - \boldsymbol{d}_\iota \right)$, $\forall \boldsymbol{h} \in \mathcal{H}$. In the context of Algorithm 1, let $S_\iota := H_\iota^-(\boldsymbol{h}_k) := \{\boldsymbol{h} \in \mathcal{H} : (\boldsymbol{h} - \boldsymbol{h}_k)^t \boldsymbol{s}_\iota + g_\iota(\boldsymbol{h}_k) \leq 0\}$, where $\boldsymbol{s}_\iota := \nabla g_\iota(\boldsymbol{h}_k)$, $\forall \iota \in I_k, \forall k \in \mathbb{N}$. The set $H_\iota^-(\boldsymbol{h}_k)$ is a closed half space and thus the associated projection operator has the simple closed form expression:

$$P_{H_\iota^-(\boldsymbol{h}_k)}(\boldsymbol{h}) = \begin{cases} \boldsymbol{h}, & \boldsymbol{h} \in H_\iota^-(\boldsymbol{h}_k), \\ \boldsymbol{h} + \frac{-g_\iota(\boldsymbol{h}_k) + (\boldsymbol{h}_k - \boldsymbol{h})^t \boldsymbol{s}_\iota}{\|\boldsymbol{s}_\iota\|^2} \boldsymbol{s}_\iota, & \boldsymbol{h} \notin H_\iota^-(\boldsymbol{h}_k). \end{cases} \quad (11)$$

Lemma 2 and (11) imply that if $\rho = \delta_\iota$, $r = 1$ and $\boldsymbol{h}_k \notin H_\iota^-(\boldsymbol{h}_k)$, then $P_{H_\iota^-(\boldsymbol{h}_k)}(\boldsymbol{h}_k) = 1/2 \left(P_{V_\iota}(\boldsymbol{h}_k) + \boldsymbol{h}_k \right)$, because $C_\iota(\delta_\iota) = V_\iota$. If, also, $q = 1$, (8) becomes: $\boldsymbol{h}_{k+1} = \boldsymbol{h}_k + \mu_k/2 \left(P_{V_k}(\boldsymbol{h}_k) - \boldsymbol{h}_k \right)$, which is NLMS if $\mu_k := 2\lambda_k$. This fact will be used in Section 4 below.

Clearly, by (8) and (11), the proposed design is free from the computational load of solving a system of linear equations to update the estimate $\boldsymbol{h}_{k+1}$ from $\boldsymbol{h}_k$, unlike the APA scheme for $r \geq 2$. A simple inspection of the summation in (8) implies that the proposed design is well suited for q concurrent processors. Due to the parallel implementation and the use of relaxation parameters that are bounded by (10) the speed of convergence can be also raised.

We propose a systematic design of the stochastic property set $C_\iota(\rho)$ based on the following simple formulae for ρ that rely only on r and on the variance of the corrupting noise process $(n_k)_{k\in\mathbb{Z}}$.

Example 1 *(Design of Stochastic Property Sets)* $\rho_1 := m_\xi + \sigma_\xi = (r + \sqrt{2r})\sigma^2 \geq \rho_2 := m_\xi = r\sigma^2 \geq \rho_3 := \max\{(r-2)\sigma^2, 0\}$ *(Other possible choice may be $\rho(\alpha) := \rho_3 + \alpha\sigma_\xi$, $\alpha \geq 0$).*

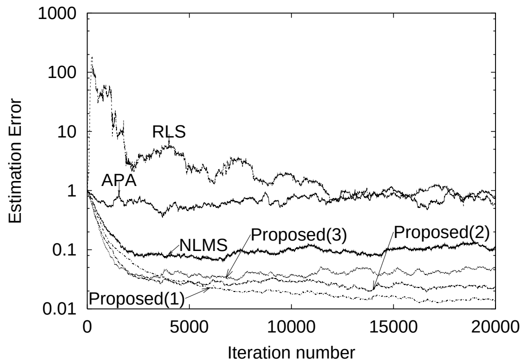
4. NUMERICAL EXAMPLES AND CONCLUDING REMARKS

To compare the proposed design with the APA, the NLMS and the RLS for estimating $\boldsymbol{h}^* \in \mathcal{H} := \mathbb{R}^{256}$, we use USASI signal, stationary with a speech-like spectrum, as the input $(u_k)_{k\in\mathbb{N}}$ (The USASI generation routine can be found in `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/txt/usasi.txt`).

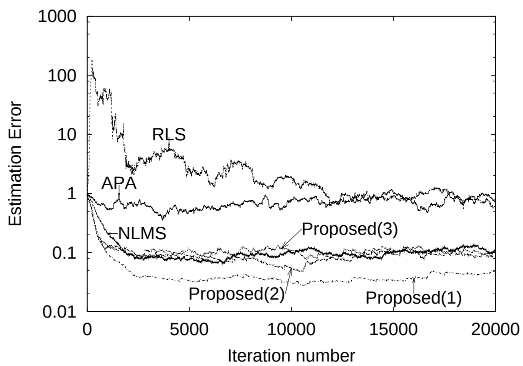
We set $r = 10$ for the APA frame (Of course, $r = 1$ for NLMS). Although the proposed design has the freedom of employing q parallel processors, for a meaningful comparison, we fix the number of processed data for each update by $rq = 10$. To ensure fair comparison, the relaxation parameters $\lambda_k = 1$, $\forall k \in \mathbb{N}$, are used for the NLMS (see (4)). For the APA, we adopt $\lambda_k = 0.05$, $\forall k \in \mathbb{N}$, which has been employed empirically in [3, 5] (See also the discussion following (7)). Regarding RLS, the *initial value* $\boldsymbol{P}_0 = 100\boldsymbol{I} \in \mathbb{R}^{256\times 256}$ ($\boldsymbol{I}$ denotes the *identity matrix*) is utilized according to the recommendation in [12, p. 570] and the *exponential weighting factor* 1, $\forall k \in \mathbb{N}$, which seems to give the best performance among our trials is implemented. According to the arguments following (11), $\mu_k = 2$, $\forall k \in \mathbb{N}$, for the proposed design. The degrees of freedom for implementing the proposed scheme for $q \in \mathbb{N}^*$ are represented by I_k which is given in its most general form in the beginning of Section 3. For the present numerical tests, where $rq = 10$ with $r = 1$ or $r = 10$, we focus on the special case of $I_k := \{k - ir\}_{i=0}^{q-1}$. Moreover, we let $w_\iota := 1/q$, $\forall \iota \in I_k$, $\forall k \in \mathbb{N}$. The simulation tests are performed under the noise situations of SNR $:= 10\log_{10}\left(E\left\{z_k^2\right\}/E\left\{n_k^2\right\}\right) = 10, 20$dB, where $z_k := \boldsymbol{u}_k^t \boldsymbol{h}^*$ and E denotes *expectation*. In Fig. 1, $\|\boldsymbol{h}^* - \boldsymbol{h}_k\|^2 / \|\boldsymbol{h}^*\|^2$, $\forall k \in \mathbb{N}$, is used as the estimation error. The notations Proposed(j), $j = 1, 2, 3$, correspond to ρ_j, $j = 1, 2, 3$, in Example 1.

As expected, the examples show the strong degradation of APA as the noise level increases. NLMS exhibits stable, but unfortunately slow convergence regardless SNR. The RLS suffers from estimating the statistics of the speech like signal. The proposed scheme seems to resolve successfully the tradeoff among speed, stability, accuracy and complexity.

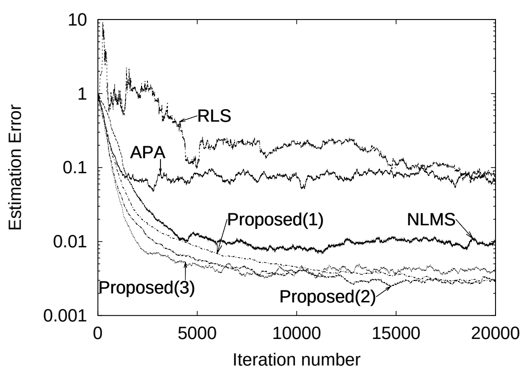
Acknowledgement: The authors would like to express their deep gratitude to Prof. K. Sakaniwa of Tokyo Institute of Technology for his encouraging discussions.



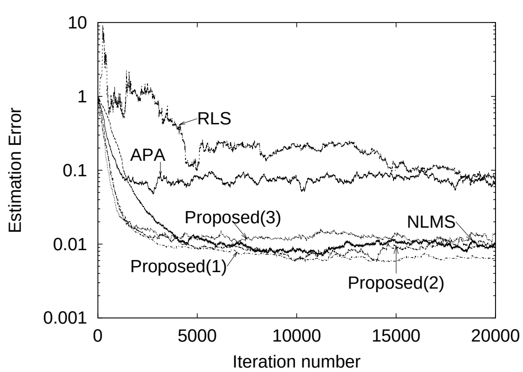
(a) SNR=10dB (r=1)



(b) SNR=10dB (r=10)



(c) SNR=20dB (r=1)



(d) SNR=20dB (r=10)

Fig. 1. Numerical Examples.

5. REFERENCES

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

[2] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *IEICE Trans.*, vol. 67-A, no. 5, pp. 126–132, 1984, in Japanese.

[3] T. Furukawa, H. Kubota, and S. Tsujii, "Orthogonal projection algorithm for block adaptive signal processing and some properties," *IEICE Trans.*, vol. 71-A, no. 12, pp. 2138–2146, 1988, in Japanese.

[4] S. L. Gay and S. Tavathia, "The fast affine projection algorithm," in *Proc. of IEEE ICASSP*, 1995, pp. 3023–3026.

[5] K. Doğancay and O. Tanrikulu, "Selective-partial-update NLMS and affine projection algorithms for acoustic echo cancellation," in *Proc. of IEEE ICASSP*, 2000, pp. 448–451.

[6] L. G. Gubin, B. T. Polyak, and E. V. Raik, "The method of projections for finding the common point of convex sets," *USSR Comput. Math. Phys.*, vol. 7, pp. 1–24, 1967.

[7] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review*, vol. 38, no. 3, pp. 367–426, 1996.

[8] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithm, and Optimization*, Oxford University Press, 1997.

[9] I. Yamada, N. Ogura, Y. Yamashita, and K. Sakaniwa, "Quadratic optimization of fixed points of nonexpansive mappings in Hilbert space," *Numerical Functional Analysis and Optimization*, vol. 19, no. 1 & 2, pp. 165–190, 1998.

[10] I. Yamada, "The Hybrid Steepest Descent Method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings," in *Inherently Parallel Algorithm for Feasibility and Optimization*, D. Butnariu, Y. Censor, and S. Reich, Eds. Elsevier, New York, 2001.

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

[12] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.

[13] A. H. Sayed and M. Rupp, "Error-energy bounds for adaptive gradient algorithms," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 1982–1989, 1996.

[14] P. L. Combettes, "Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections," *IEEE Trans. Image Processing*, vol. 6, no. 4, pp. 493–506, 1997.

[15] P. L. Combettes, "The foundations of set theoretic estimation," *Proc. IEEE*, vol. 81, no. 2, pp. 182–208, 1993.

[16] H. Cramér, *Mathematical methods of statistics*, Princeton University Press, 1946.

[17] P. L. Combettes and H. J. Trussell, "The use of noise properties in set theoretic estimation," *IEEE Trans. Signal Processing*, vol. 39, no. 7, pp. 1630–1641, 1991.

[18] B. Hassibi, A. H. Sayed, and T. Kailath, "H^{∞} optimality of the LMS algorithm," *IEEE Trans. Signal Processing*, vol. 44, no. 2, pp. 267–280, 1996.

[19] G. Pierra, "Decomposition through formalization in a product space," *Math. Programming*, vol. 28, pp. 96–115, 1984.

[20] E. R. Ferrara, "Fast implementation of LMS adaptive filters," *IEEE Trans. Acous., Speech, Signal Processing*, vol. 28, no. 4, pp. 474–475, 1980.

[21] K. Kurosawa and T. Furusawa, "A geometric interpretation of adaptive algorithms," *IEICE Trans.*, vol. 71-A, no. 2, pp. 343–347, 1988, in Japanese.