EFFICIENT ADAPTIVE FILTERING FOR SMOOTH LINEAR FIR MODELS

Kristiaan Pelckmans[*], *Toon van Waterschoot*[+], *Johan A.K. Suykens*[+]

[*] Uppsala University, Department of IT, Syscon
Box 337, SE-751 05 Uppsala, Sweden, phone: +46 18 - 471 3393,
[+] K.U. Leuven, ESAT-SCD (SISTA),
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium, phone +32-16-32 19 27,
e-mail:kp@it.uu.se, toon.vanwaterschoot@esat.kuleuven.be, johan.suykens@esat.kuleuven.be

ABSTRACT

This[1] paper proposes the Smooth Gradient Descent (SGD) algorithm for recursively identifying a linear Finite Impulse Response (FIR) model with a large set of parameters ('long impulse response'). The main thesis is that a successful design of such adaptive filter must hinge on (i) the choice of a proper loss function, and (ii) the choice of a proper norm for the impulse response vector. Theoretical backup for this statement is found in slightly improving and interpreting the regret bound of the Gradient Descent (GD) algorithm presented in [3]. In practice, if the impulse response vector is known to be *smooth* in some apriori defined sense, the proposed algorithm will converge faster.

1. INTRODUCTION

The study of adaptive filtering or recursive identifying a linear FIR filters for dealing with (relatively) long impulse response vectors $\mathbf{h}_* \in \mathbb{R}^d$ requires the need for efficient adaptive filters [5, 2, 4], providing a benchmark setup to test new adaptive filters. We study a class of gradient descent algorithms which are based on a norm of the solution vector $\|\mathbf{h}_*\|$ and a proper loss function $L : \mathbb{R} \to \mathbb{R}$. If the norm were chosen as the standard Euclidean norm, the standard Gradient Descent (GD) algorithm were recovered. When in addition the squared loss function were employed - i.e. $L(e) = e^2$ for all $e \in \mathbb{R}$ - this algorithm reduces to the Least Mean Square (LMS) algorithm. If the absolute loss were used - i.e. $L(e) = |e|$ for all $e \in \mathbb{R}$ - the sign-error algorithm were recovered - see e.g. [4]. While previous research in deriving efficient adaptive filters has focused mainly on the design of (adaptive) step-sizes (with prototypical case the normalized LMS (NLMS) or the Affine Projections (AP) methods, see e.g. [4] and references), we focus here instead on different design decisions. Specifically we use a generic result described in [3] to motivate the proper choice of a norm where a corrsponding gradient descent algorithm is based on. It is found that a general algorithm - the Smooth Gradient Descent (SGD) - leads to a guaranteed performance when the norm is chosen properly. This theoretic property is then found to be relevant in practice, as it allows to incorporate general forms of prior knowledge of the vector $\mathbf{h}$ one aims at. The stepsize in this algorithm is fixed, and is considered to be given by an oracle.

The motivation for this study comes from investigations of Acoustic Echo Cancelation (AEC), see e.g. [5, 2, 1] and [9, 8]. Here one is after good estimates of the room echo system. In order to account for long echoes and complex dynamics, one uses typically a linear filter with $O(1000)$ timelags. Since (i) it is general found that such systems are not adequately described by a ratio of lower order polynomials (as IIR filters), (ii) the available hardware depreciates more involved calculations, and (iii) the straightforward parameterization of FIR filters is found to result in more reliable adaptive filters [2], one often resorts to FIR models with long impulse response vectors - denoted in this context as a Room Impulse Response (RIR) vector $\mathbf{h} \in \mathbb{R}^d$ for a FIR filter of order $d \in \mathbb{N}$ where $d = O(N)$. Here we treat the far-end signal $(u_t)_t$ (for example - the speech of a remote user) as the input to the local (echo) system. This input is transduced to an acoustic waveform by a loudspeaker, and this signal - perturbed and filtered by the room echo system - is picked up by a near-end microphone. This near-end signal is mixed with speech of a local user - represented as $(e_t)_t$, yielding the 'output' signal $(y_t)_t$. The main question of adaptive filtering applied to AEC is then how the RIR can be estimated from observation of $(u_t)_t$ and $(y_t)_t$. Using this estimate, an algorithm can be implemented canceling out the echoed signals (=uninformative part) from the signal which is transmitted to other (i.e. non-local) users. Since the inception in AT&T labs some fifty years ago, much progress has been made towards the use of different adaptive filters algorithms, see e.g. [2, 1] for a survey and recent work.

This paper revises some of the main difficulties, and argues how new insights in adaptive filtering and online estimation may lead to a constructive solution to the following inherent problems. (i) The RIRs are long with respect to the timespan the signal can be assumed to be stationary. (ii) The near-end 'noise' process disturbing the measurements of the echo cannot really be assumed to be of a Gaussian nature, nor to be uncorrelated to the echo. (iii) The system according to the loudspeaker is not linear in reality, and disturbances capturing the nonlinear effects troubles the typical statistical assumptions one makes on the noise. For these reasons, we aim at a framework not relying on a (restrictive) stochastic framework, and ask how one can incorporate prior knowledge in order to make efficient adaptive filters.

This paper is organized as follows. Section 2 describes the SGD algorithm and spells out the regret bound. Section 3 itemizes three different design principles for dealing with long impulse responses, and Section 4 gives a discussion of the result.

2. SMOOTH GRADIENT DESCENT

Let $N \in \mathbb{N}$ be a known constant. Let $(y_1, y_2, \ldots, y_N) \in \mathbb{R}^N$ and $(u_1, u_2, \ldots, u_N) \in \mathbb{R}^N$ be two given timeseries of length N. In some cases it is useful to consider the *exact* FIR system of the form $y_t = \sum_{\tau=1}^{d} h_{0,\tau} u_{t-\tau}$, where $d \in \mathbb{N}$ denotes the order of the system, $\mathbf{h}_0 = (h_{0,1}, \ldots, h_{0,d}) \in \mathbb{R}^d$ are the unknown *true* parameters of the model. In the context of this paper, we study the problem of estimating a good *approximate* model defined as

$$y_t = \sum_{\tau=1}^{d} h_\tau u_{t-\tau} + e_t, \tag{1}$$

where $\mathbf{h} = (h_1, \ldots, h_d) \in \mathbb{R}^d$ denote *appropriate* unknown parameters, and where $\mathbf{e} = (e_1, e_2, \ldots, e_{N,\mu})^T \in \mathbb{R}^N$ is its corresponding (unobserved) timeseries of residuals. The understanding is that $\mathbf{e}$ is small (in some norm), or that the different elements are not causally predictable (innovation sequence). In many acoustic applications, the sequence $\mathbf{e}$ denotes the actual speech-signal, and hence the *informative* component. Often, it can be assumed to be an innovation

[1] Acknowledgments: Johan Suykens is supported by K.U.Leuven, the Flemish government, FWO and the Belgian federal science policy office (FWO G.0302.07, CoE EF/05/006, GOA AMBioRICS, IUAP DYSCO)

 ISSN 2076-1465

sequence, but it is more realistically modeled as a stochastic process with nontrivial impulse response on its own.

Introduce for all $t = d+1\ldots,T$ the vector $\mathbf{u}_t = (q^{-1}u_t,\ldots,q^{-d}u_t)^T \in \mathbb{R}^d$ where q^{-1} denotes the back-shift operator. This paper considers cases where d is typically very large, say $O(N)$. We are interested in 'good' predictors, i.e. in an algorithm yielding a sequence of hypotheses denoted as $\mathbf{h}_0,\mathbf{h}_1,\ldots,\mathbf{h}_{T-1}$ such that

$$\frac{1}{N-d}\sum_{t=d+1}^{N} L\left(y_t-\mathbf{h}_t^T\mathbf{u}_t\right), \tag{2}$$

is small, here $L:\mathbb{R}\to\mathbb{R}^+$ is a positive, convex *loss function* with $L(0)=0$, with existing first derivative $L':\mathbb{R}\to\mathbb{R}$ defined as $L'(e)=\dfrac{\partial L(e)}{\partial e}$ for all $e\in\mathbb{R}$. Let $(\mathbf{h}_1,\ldots,\mathbf{h}_N)\in\mathbb{R}^d$ be a sequence of hypotheses generated by an algorithm. Then we will be interested in the *regret*, defined as the cumulative loss the algorithm has compared to the *best* fixed alternative, or

$$R_N=\sum_{d+1}^{N} L\left(y_t-\mathbf{h}_t^T\mathbf{u}_t\right)-\inf_{\mathbf{h}\in\mathbb{R}^d}\sum_{t=d+1}^{N} L\left(y_t-\mathbf{h}^T\mathbf{u}_t\right). \tag{3}$$

We will consider two adaptive filters giving an estimate $\mathbf{h}_t$ of $\mathbf{h}$ when observing the timeseries up to instant $t\le N$: the standard Gradient Descent (GD) algorithm is given as

$$\hat{\mathbf{h}}_{t,\mu}=\hat{\mathbf{h}}_{t-1,\mu}+\mu L'(y_t-\hat{\mathbf{h}}_{t-1}^T\mathbf{u}_t)\mathbf{u}_t,\ \forall t=d+1,\ldots,N, \tag{4}$$

where μ is an appropriate constant, and $\hat{\mathbf{h}}_{d,\mu}=0_d$. The estimates of this algorithm are denoted as $\hat{\mathbf{h}}_{d,\mu},\ldots,\hat{\mathbf{h}}_{N,\mu}$ (using *hat*). In the standard LMS version, one takes the squared loss, or $L(e)=e^2$ for all $e\in\mathbb{R}$. Let $\mathbf{A}\in\mathbb{R}^{d\times d}$ be an appropriate symmetric, positive definite matrix. The Smooth GD with weighting matrix $\mathbf{A}$ (abbreviated as SGD) is defined as

$$\tilde{\mathbf{h}}_{t,\nu}=\tilde{\mathbf{h}}_{t-1,\nu}+\nu L'(y_t-\tilde{\mathbf{h}}_{t-1,\nu}^T\mathbf{A}\mathbf{u}_t)\mathbf{u}_t,\ \forall t=d+1,\ldots,N, \tag{5}$$

where ν is an appropriate constants, and $\tilde{\mathbf{h}}_{d,\nu}=0_d$. The estimates of this algorithm are denoted as $\tilde{\mathbf{h}}_{d,\nu},\ldots,\tilde{\mathbf{h}}_{N,\nu}$ (using *tilde*). Note that we predict the outcome of the filter at time t based on the hypothesis $\tilde{\mathbf{h}}_{t-1,\nu}$ as $\tilde{\mathbf{h}}_{t-1,\nu}^T\mathbf{A}\mathbf{u}_t$. Note that one may rewrite the algorithm SGD in a slightly different form, or $\forall t=d+1,\ldots,N$ one has

$$\tilde{\mathbf{p}}_{t,\nu}=\tilde{\mathbf{p}}_{t-1,\nu}+\nu L'(y_t-\tilde{\mathbf{p}}_{t-1,\nu}^T\mathbf{u}_t)\mathbf{A}^T\mathbf{u}_t, \tag{6}$$

with $\tilde{\mathbf{p}}_{t,\nu}=\mathbf{A}^T\tilde{\mathbf{h}}_{t,\nu}$ for $t=d,\ldots,N$. One now predicts the outcome of the filter at time t based on the hypothesis $\tilde{\mathbf{p}}_{t-1,\nu}$ as $\tilde{\mathbf{p}}_{t-1,\nu}^T\mathbf{u}_t$. This formulation is equivalent to (5), even if the matrix $\mathbf{A}$ is not invertible: this is seen as the prediction rule $\mathbf{A}\tilde{\mathbf{h}}_{t-1,\nu}$ or $\tilde{\mathbf{p}}_{t,\nu}$ lies always in the span of $\mathbf{A}$ by construction, implying invertibility at the point of interest. An intuitive way to comprehend this formulation is to minimize the following instantaneous cost function

$$\tilde{\mathbf{p}}_{t,\nu}=\operatorname*{argmin}_{\mathbf{p}}\frac{1}{2}\left\|\tilde{\mathbf{p}}_{t-1,\nu}-\mathbf{p}\right\|_{\mathbf{B}}+\nu L(y_t-\mathbf{p}^T\mathbf{u}_t), \tag{7}$$

for given symmetric, positive definite matrix $\mathbf{B}\in\mathbb{R}^{d\times d}$. The minimizer is then

$$\mathbf{B}\tilde{\mathbf{p}}_{t,\nu}=\mathbf{B}\tilde{\mathbf{p}}_{t-1,\nu}+\nu L'(y_t-\tilde{\mathbf{p}}_{t-1,\nu}^T\mathbf{u}_t)\mathbf{u}_t, \tag{8}$$

which is equal to the update step (6) in the SGD algorithm in case $\mathbf{A}^T\mathbf{B}=I_d$, or to (5) when $\mathbf{B}\tilde{\mathbf{p}}_{t-1,\nu}=\tilde{\mathbf{h}}_{t-1,\nu}$ and $\tilde{\mathbf{p}}_{t-1,\nu}=\mathbf{A}\tilde{\mathbf{h}}_{t-1,\nu}$. Observe that this interpretation led [8] to the similar algorithm LMR-NLMS, presenting additional empirical evidence for such scheme.

2.1 Regret and Satisfaction

The performance of an algorithm is defined in terms of its cumulative loss, defined as $L_N(\mathbf{h}),\hat{L}_{N,\mu}$ and $\tilde{L}_{N,\nu}$:

$$\begin{cases}\tilde{L}_{N,\nu}=\sum_{t=d+1}^{N}L\left(y_t-\tilde{\mathbf{h}}_{t-1,\nu}^T\mathbf{A}\mathbf{u}_t\right)\\ \hat{L}_{N,\mu}=\sum_{t=d+1}^{N}L\left(y_t-\hat{\mathbf{h}}_{t-1,\mu}^T\mathbf{u}_t\right)\\ L_N(\mathbf{h})=\sum_{t=d+1}^{N}L\left(y_t-\mathbf{h}^T\mathbf{u}_t\right).\end{cases} \tag{9}$$

Now let the regret of either algorithm with respect to an hypothesis be given as

$$\begin{cases}\tilde{R}_{N,\nu}(\mathbf{h})=\tilde{L}_{N,\nu}-L_N(\mathbf{h})\\ \hat{R}_{N,\mu}(\mathbf{h})=\hat{L}_{N,\mu}-L_N(\mathbf{h}).\end{cases} \tag{10}$$

The regret with respect to the hypothesis $\tilde{\mathbf{h}}_{N,\nu}$ or $\hat{\mathbf{h}}_{N,\mu}$ - defined as $\tilde{R}_{N,\nu}(\tilde{\mathbf{h}}_{N,\nu})$ and $\hat{R}_{N,\mu}(\hat{\mathbf{h}}_{N,\mu})$ respectively - reflects what loss one suffers since one did not know the 'learned' solution before seeing the data. One could interpret terms $\tilde{R}_{N,\nu}(\tilde{\mathbf{h}}_{N,\nu})$ and $\hat{R}_{N,\mu}(\hat{\mathbf{h}}_{N,\mu})$ as 'the pain of learning'. The *actual* regret is defined as the regret of either algorithm with respect to the best (unknown) hypothesis ('expert'), or

$$\begin{cases}\tilde{R}_{N,\nu}=\tilde{L}_{N,\nu}-\inf_{\mathbf{h}\in\mathbb{R}^d}L_N(\mathbf{h})\\ \hat{R}_{N,\mu}=\hat{L}_{N,\mu}-\inf_{\mathbf{h}\in\mathbb{R}^d}L_N(\mathbf{h}).\end{cases} \tag{11}$$

This paper will investigate how to design a GD algorithm which minimizes a bound on $\tilde{R}_{N,\nu}$, potentially leading to much better properties than hold for $\hat{R}_{N,\nu}$. The main result is established following the proof in [3], using the following two basic results. In order to deal with general convex loss functions L, we need the following result.

Proposition 1 *Let $L:\mathbb{R}\to\mathbb{R}$ be a continuous, twice differentiable convex loss function. Then for any $y,\bar{y}\in\mathbb{R}$ one has*

$$L'(y-\bar{y})(\bar{y}-\hat{y})\le(L(y-\hat{y})-L(y-\bar{y}))\le L'(y-\hat{y})(\bar{y}-\hat{y}). \tag{12}$$

Proof: The first inequality follows from the mean value theorem, as there exists a c inbetween $\hat{y},\bar{y}$ that gives

$$L(y-\bar{y})=L(y-\hat{y})+L'(y-\hat{y})(\hat{y}-\bar{y})+\frac{L''(y-c)(\hat{y}-\bar{y})^2}{2}, \tag{13}$$

and using the fact that $L''(e)\ge 0$ since L is convex, or $L(y-\bar{y})-L(y-\hat{y})\ge L'(y-\hat{y})(\hat{y}-\bar{y})$, or $L(y-\hat{y})-L(y-\bar{y})\le L'(y-\hat{y})(\bar{y}-\hat{y})$. Similarly $\exists c'\in\mathbb{R}$ such that

$$L(y-\hat{y})=L(y-\bar{y})+L'(y-\bar{y})(\bar{y}-\hat{y})+\frac{L''(y-c')(\bar{y}-\hat{y})^2}{2}, \tag{14}$$

or $L(y-\hat{y})-L(y-\bar{y})\ge L'(y-\bar{y})(\bar{y}-\hat{y})$, proving the second inequality in (12). Furthermore, we have

Proposition 2 *Let $\mathbf{w},\mathbf{x},\bar{\mathbf{w}}\in\mathbb{R}^d$ be vectors for $d\in\mathbb{N}$, and let $\mathbf{A}\in\mathbb{R}^{d\times d}$ be a positive definite matrix. Then one has for any $z\in\mathbb{R}$ that*

$$z(\bar{\mathbf{w}}^T\mathbf{A}\mathbf{x}-\mathbf{w}^T\mathbf{A}\mathbf{x})=\frac{\|\bar{\mathbf{w}}-\mathbf{w}\|_{\mathbf{A}}^2}{2}-\frac{\|\bar{\mathbf{w}}-(\mathbf{w}+z\mathbf{x})\|_{\mathbf{A}}^2}{2}+\frac{z^2}{2}\|\mathbf{x}\|_A^2, \tag{15}$$

where $\|\mathbf{x}\|_{\mathbf{A}}^2=\mathbf{x}^T\mathbf{A}\mathbf{x}$. Specifically, this holds for the case $\mathbf{A}=\operatorname{diag}(1,\ldots,1)=I_d\in\mathbb{R}^{d\times d}$, reducing $\|\cdot\|_{\mathbf{A}}$ to the standard Euclidean norm $\|\cdot\|_2$.

This follows by working out the terms.

Proposition 3 *Let $a,b>0$ be constants, then*

$$\inf_{\xi>0} \frac{a}{\xi} + \xi b = 2\sqrt{ab}. \tag{16}$$

This is seen by choosing $\xi = \sqrt{\frac{a}{b}}$, obtained by equating the derivative to zero. As in [3], the previous 2 propositions can be used to bound the regret of both GD and SGD. We will spell this result out for the SGD case, since the GD case can be recovered by taking $\mathbf{A} = \text{diag}(1,\ldots,1) = I_d \in \mathbb{R}^{d\times d}$.

Lemma 1 (Regret Bounds) *Given a symmetric, strictly positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with inverse $\mathbf{A}^{-1} \in \mathbb{R}^{n\times n}$. Let $R_{\mathbf{A}}^2 \geq \sum_{t=d+1,\ldots,N} \|\mathbf{u}_t\|_{\mathbf{A}}^2$ and $C_1^2 \geq \sup_e L'(e)^2$. When performing SGD with $\nu = \frac{\|\mathbf{h}\|_{\mathbf{A}^{-1}}}{R_{\mathbf{A}} C_1 \sqrt{N-d}}$, one obtains predictions with regret for any $\mathbf{h} \in \mathbb{R}^d$ bounded by*

$$\inf_{\nu>0} \tilde{L}_{N,\nu} - L_N(\mathbf{h}) \leq \|\mathbf{h}\|_{\mathbf{A}^{-1}} C_1 R_{\mathbf{A}}. \tag{17}$$

or $\inf_{\nu>0} \tilde{R}_{N,\nu} \leq R_{\mathbf{A}} C_1 \|\mathbf{h}_*\|_{\mathbf{A}^{-1}}$ with $h_* = \text{argmin}_{\mathbf{h}} L_N(\mathbf{h})$.

Proof: As in [3], one has for any $\mathbf{h} \in \mathbb{R}^d$ that

$$\begin{aligned}
&2L\left(y_t - \tilde{\mathbf{h}}_{t-1,\nu}^T \mathbf{A}\mathbf{u}_t\right) - 2L\left(y_t - \mathbf{h}^T\mathbf{u}_t\right) \\
&\quad \leq 2L'\left(y_t - \tilde{\mathbf{h}}_{t-1,\nu}^T \mathbf{A}\mathbf{u}_t\right)\left(\mathbf{h}^T\mathbf{A}^{-T}\mathbf{A}\mathbf{u}_t - \tilde{\mathbf{h}}_{t-1,\nu}^T\mathbf{A}\mathbf{u}_t\right) \\
&\quad = \frac{1}{\nu}\left\|\mathbf{A}^{-1}\mathbf{h} - \tilde{\mathbf{h}}_{t-1,\nu}\right\|_{\mathbf{A}}^2 - \frac{1}{\nu}\left\|\mathbf{A}^{-1}\mathbf{h} - \tilde{\mathbf{h}}_{t,\nu}\right\|_{\mathbf{A}}^2 \\
&\qquad + \nu L'\left(y_t - \tilde{\mathbf{h}}_{t-1,\nu}^T\mathbf{A}\mathbf{u}_t\right)^2 \|\mathbf{u}_t\|_{\mathbf{A}}^2,
\end{aligned} \tag{18}$$

from the previous two propositions, and having $z = \nu L'\left(y_t - \tilde{h}_{t-1,\nu}^T \mathbf{A}\mathbf{u}_t\right)$. When taking the sum '$\sum_{t=d+1}^N$' over both sides of the inequality, one sees that most difference terms $\|\mathbf{A}^{-1}\mathbf{h} - \tilde{\mathbf{h}}_{t,\nu}\|_{\mathbf{A}}$ cancel out (the 'telescoping property') and one has for any $\mathbf{h} \in \mathbb{R}^d$ that

$$\begin{aligned}
2\tilde{L}_{N,\nu} - 2L_N(\mathbf{h}) &\leq \frac{1}{\nu}\|\mathbf{h}\|_{\mathbf{A}^{-1}}^2 + \sum_{t=d+1}^N \nu L'\left(y_t - \tilde{\mathbf{h}}_{t-1,\nu}^T\mathbf{A}\mathbf{u}_t\right)^2 \|\mathbf{u}_t\|_{\mathbf{A}}^2 \\
&\leq \frac{1}{\nu}\|\mathbf{h}\|_{\mathbf{A}^{-1}}^2 + (N-d)C_1^2 R_{\mathbf{A}}^2 \nu.
\end{aligned} \tag{19}$$

Now the upperbound is minimized by taking $\nu = \frac{\|\mathbf{h}\|_{\mathbf{A}^{-1}}}{R_{\mathbf{A}} C_1 \sqrt{N-d}}$ (as in proposition 3), or

$$\inf_{\nu>0} \tilde{L}_{N,\nu} - \inf_{\mathbf{h}\in\mathbb{R}^d} L_N(\mathbf{h}) \leq \sqrt{(N-d)}\, C_1 R_{\mathbf{A}} \|\mathbf{h}\|_{\mathbf{A}^{-1}}, \tag{20}$$

yielding the upperbund of (17).

3. PRACTICAL ILLUSTRATION OF SGD

This section aims to illustrate that the above derivation gives us an often very practical and useful result. In case studies where data is not abundant, poorly exciting, or is subject to significant noise, it is often useful to exploit prior information about the desired model in order to control the variance of the estimates as good as possible.

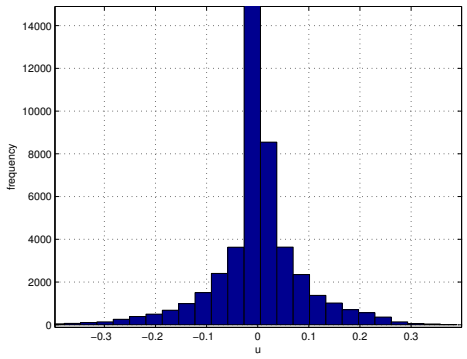

Figure 1: *Histogram of a real speech-signal. Note the fat tails, suggesting the use of a robust loss function for modeling in acoustic applications.*

3.1 On the Choice of the Loss Function L

While typically, one chooses the squared error loss function $L(e) = e^2$ for all $e \in \mathbb{R}$, it is often worthwhile to consider other appropriate loss functions. A strength of the above results is they hold for arbitrary (i) convex, with (ii) bounded first derivative (by C_1, see Lemma 1). The bound then suggest that the performance of an algorithm improves (or the regret decays) when C_1 is smaller.

Another motivation to make a choice for another loss function L than the squared loss is as follows. In the AEC examples we will be interested in, the residual terms $\mathbf{e}$ will reflect the near-end speech signal. Assuming that this signal is white and Gaussian effectively ignores the structure of typical acoustic signals. While coloring of the signal $\mathbf{e}$ has been treated in the context of double-talk robustness in [8], the non-Gaussian nature of $\mathbf{e}$ suggests the use of another loss function. Specifically, in speech signals, one typically has fat-tailed distributions (see Figure 1) modeling the occurrence of articulations or non-smooth signals, so typical for meaningful speech.

As such, a more robust loss function is motivated, and we will do so by considering the Huber loss function, proposed in the literature of robust statistics, dealing effectively with such fat-tailed data or significantly contaminated signals. The fact that this loss is not often used in practice may be dedicated to the need for computationally intensive methods in batch settings (since there is no closed-form solution resembling the normal equations, the problem of finding minimum huber loss can be done by solving a convex Quadratic Program, see e.g. [6]). The implementation of this loss in a Gradient Descent (GD) algorithm (4) is however straightforward. Huber's convex, continuous and differentiable loss function is defined for any parameter $\delta \geq 0$ as

$$L_\delta(e) = \begin{cases} \frac{1}{2\delta}e^2 & |e| < \delta \\ |e| - \frac{\delta}{2} & |e| \geq \delta. \end{cases} \tag{21}$$

Its derivative exists for all e, and becomes

$$L'_\delta(e) = \begin{cases} \frac{1}{\delta}e & |e| < \delta \\ \text{sign}(e) & |e| \geq \delta. \end{cases} \tag{22}$$

and $C_1 = 1$ in this case. This loss function basically has the advantage of having a bounded term C_1 in the bound of Lemma 1. Note that in this case, the second derivative does not exist everywhere, and this technical issue might be resolved using a twice differentiable proxy L_δ^+ instead.

3.2 On the Choice of the Complexity Term $\|\cdot\|_{\mathbf{A}^{-1}}$

3.2.1 *Transforming* $\mathbf{u}_t$

At first, let us consider the case that the inputs $\{\mathbf{u}_t\}_t \subset \mathbb{R}^d$ have a covariance matrix which is (nearly) rank-deficient, or $\mathbf{R} = \frac{1}{N-d}\sum_{t=d+1}^{N}\mathbf{u}_t\mathbf{u}_t^T$ has condition-number $\kappa(\mathbf{R}) = \frac{\lambda_{\max}(\mathbf{R})}{\lambda_{\min}(\mathbf{R})} \to \infty$. This is often the case in acoustic applications, due to the tonality of voiced speech and the high RIR orders often required [9]. Then it is advantageous to transform the inputs to a sequence which has covariance matrix with condition number nearly one. Suppose we have such a transformation given by $\mathbf{T} \in \mathbb{R}^{d\times d}$, hence one has for all $t = d+1,\ldots,N$ that

$$\bar{\mathbf{u}}_t = \mathbf{T}\mathbf{u}_t, \tag{23}$$

The SGD algorithm then becomes

$$\bar{\mathbf{h}}_{t,\mu} = \bar{\mathbf{h}}_{t-1,\mu} + \mu L'(y_t - \bar{\mathbf{h}}_{t-1}^T\mathbf{T}\mathbf{u}_t)\mathbf{T}\mathbf{u}_t, \tag{24}$$

$\forall t = d+1,\ldots,N$. Now one may multiply both sides by $\mathbf{T}^T$ and replace $\mathbf{T}^T\bar{\mathbf{h}}_{t,\mu}$ by $\bar{\mathbf{p}}_{t,\mu}$, and this gives in turn

$$\bar{\mathbf{p}}_{t,\mu} = \bar{\mathbf{p}}_{t-1,\mu} + \mu L'(y_t - \bar{\mathbf{p}}_{t-1}^T\mathbf{u}_t)\mathbf{T}^T\mathbf{T}\mathbf{u}_t, \tag{25}$$

and so it follows that convergence properties are characterized by $\|\mathbf{h}_*\|_{\mathbf{A}^{-1}} = \mathbf{h}_*^T(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{h}_*$ or $\mathbf{A} = \mathbf{T}^T\mathbf{T}$.

The optimal transformation $\mathbf{T}_*$ in this respect is such that $\frac{1}{N-d}\sum_{t=d+1}^{N}\mathbf{T}_*\mathbf{u}_t\mathbf{u}_t^T\mathbf{T}_* = \mathbf{T}_*\mathbf{R}\mathbf{T}_*^T = I_d$. An optimal inverse transformation $\mathbf{T}_*^{-1}$ can hence be computed by a Cholesky decomposition as $\mathbf{R} = \mathbf{T}_*^{-T}\mathbf{T}_*^{-1}$. Computing the inverse of the uppertriangular matrix $\mathbf{T}_*^{-1}$ (assuming it exists) gives the optimal transformation $\mathbf{T}_* \in \mathbb{R}^{d\times d}$ (indeed we have that for this $\mathbf{T}_*$ the equality $\mathbf{T}_*\mathbf{R}\mathbf{T}_*^T = I_d$ holds). Moreover, we have that $\mathbf{T}_*^T\mathbf{T}_* = \mathbf{R}^{-1}$, that is, if the inverse to $\mathbf{R}$ exists.

Finally, it is important to note that in case we choose a time-dependent $\mathbf{A}_t$ as $\frac{1}{t}\mathbf{R}_t^{-1}$ in timesteps $t = k+1,\ldots,N$ - where $\mathbf{R}_t = \frac{1}{t-d}\sum_{k=d+1}^{t}\mathbf{u}_k\mathbf{u}_k^T$ for all $t = d+1,\ldots,N$- one recovers the classical Recursive Least Squares (RLS) algorithm.

3.2.2 *Inverse Covariance of* $\mathbf{h}_0$

Now, the choice of a matrix $\mathbf{A}$ may be guided by considerations on the typical vector $\mathbf{h}_* \in \mathbb{R}^d$ one aims at. This subsubsection adopts a probabilistic setup in the following non-standard sense. Consider that the vector $\mathbf{h}_*$ may be assumed to be a sample from a random process itself. Rather than having a fixed *true* impulse response, we have the case that impulse responses associated with all (possible) acoustic room and speaker-microphone setups follow a fixed but unknown distribution law. The 'target' $\mathbf{h}$ appropriate to the present task is seen as an (independent) sample of this probability distribution. This interpretation also led [8] to the construction of new algorithms. Now a convenient way to proceed is to assume that this probability law can be well approximated as a Gaussian process, i.e.

$$\mathbf{h} \sim \mathcal{N}(0_d, \mathbf{R}_\mathbf{h}), \tag{26}$$

with positive (semi-)definite covariance matrix $\mathbf{R}_\mathbf{h} = \mathbf{R}_\mathbf{h}^T$ defined as $\mathbf{R}_{\mathbf{h},ts} = E(\mathbf{h}_t\mathbf{h}_s)$. Note that we take 0_d as the mean impulse response, meaning effectively that when averaging out all possible impulse responses for different setups, one gets 0_d (it does not really say that we 'expect' the impulse response to be zero). Then the likelihood of $\mathbf{h}$ under the given model becomes

$$L(\mathbf{h}) = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\mathbf{R}_\mathbf{h})}\exp\left(-\frac{1}{2}\mathbf{h}^T\mathbf{R}_\mathbf{h}^{-1}\mathbf{h}\right), \tag{27}$$

and as such $\mathbf{h}^T\mathbf{R}_\mathbf{h}^{-1}\mathbf{h} = \|\mathbf{h}\|_{\mathbf{A}^{-1}}$ for $\mathbf{A} = \mathbf{R}_\mathbf{h} \in \mathbb{R}^{d\times d}$ is inversely proportional to how well $\mathbf{h}$ fits this model. The assumption of the parameter $\mathbf{h}$ being a random variable justifies the term 'Bayesian' for this model, with (26) representing a Gaussian prior.

The use of this model in practice goes as follows. Suppose we have identified RIRs $\{\mathbf{h}_k\}_{k=1}^n \subset \mathbb{R}^d$ for n different acoustic setups of interest, using FIR models of constant order d. From those one may estimate the covariance matrix $\mathbf{R}_h$ assuming the expected impulse response is zero. Now the above reasoning makes it advantage to apply the SGD algorithm in a new acoustic situation similar to the ones used for constructing $\mathbf{R}_{\mathbf{h},n}$. As such, we implicitly model the prior knowledge (up to second order) as $\mathbf{R}_{\mathbf{h},n} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{h}_k\mathbf{h}_k^T$ with n different RIRs collected during previous experiments.

3.2.3 *Smooth* $\mathbf{h}_*$

The previous reasoning can also be applied without the Bayesian setup. To see this, consider the case that the vector $\mathbf{h}$ obeys the recursion

$$\mathbf{h}_k = a\mathbf{h}_{k-1} + z_k, \ \ \forall k = 1,\ldots,d, \tag{28}$$

for parameter $a \in \mathbb{R}$ with $|a| < 1$, $\mathbf{h}_{(0)} = 0$ and for appropriate terms $\{z_k\}_{k=1}^d$. In this case the RIR vector $\mathbf{h} \in \mathbb{R}^d$ is modeled as a first order AR process itself. Small terms $\{z_k\}_k$ and $a \approx 1$ effectively mean that coefficients of $\mathbf{h}$ associated to nearby timelags are not too different. The matrix $\mathbf{H}$ takes the form

$$\mathbf{H}_a = \begin{bmatrix} 1 & -a & a^2 & & & (-a)^{d-1} \\ 0 & 1 & a & & & (-a)^{d-2} \\ & & & \ddots & & \\ & & & & 1 & a \\ & & & 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{29}$$

Then $\mathbf{h} = \mathbf{H}_a\mathbf{z}$ where $\mathbf{z} = (z_1,\ldots,z_k)^T \in \mathbb{R}^d$. If one beliefs that $\mathbf{h}_*$ can be described by this $\mathbf{H}_a$ and a vector $\mathbf{z}$ which has small norm, then the efficacy result (Lemma 1) claims that the SGD algorithm using the norm $\|\mathbf{h}\|_{\mathbf{A}^{-1}}$ with $\mathbf{A}$ defined as $\mathbf{A} = (\mathbf{H}_a^T\mathbf{H}_a)^{-1}$ gives better estimates than a standard application of the GD (GD) algorithm. For example, in case of the AR(1) model (28) with parameter a, the following matrix $\mathbf{A}$ would be advocated

$$\mathbf{A} = \begin{bmatrix} 1 & a & 0 & & & \\ a & 1 & a & & & \\ 0 & a & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & a \\ & & & & a & 1 \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{30}$$

If the model AR(1) model does fit the desired $\mathbf{h}_*$, then one has that the norm $\|\mathbf{h}_*\|_{\mathbf{A}^{-1}} \sim \|\mathbf{h}_*\|_2$, Hence the SGD wil give the same performance as the standard GD. Note that this can be generalized to other models capturing the prior knowledge of the $\mathbf{h}$ vector of interest.

Suppose we believe the desired vector $\mathbf{h}_* \in \mathbb{R}^d$ is not too small in a Sobolev norm. The second order Sobolev norm of a twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is defined as $\|f_*\|_s^2 = \left(\int |Df_*(x)|^2 dx\right)$, where D denotes the differential operator. The equivalent of this norm on a vector $\mathbf{h}_*$ is then $\mathbf{h}_*^T\mathbf{A}_s\mathbf{h}_*$ where

$$\mathbf{A}_s = \begin{bmatrix} -1 & 1 & 0 & & & \\ 1 & -2 & 1 & & & \\ 0 & 1 & -2 & & & \\ & & & \ddots & & \\ & & & & -2 & 1 \\ & & & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{31}$$

The study falls in the realm of (smoothing) Spline theory, see e.g. [10]. This reasoning can be generalized as follows. Suppose $\mathbf{h}_*$ can be adequately described as

$$\mathbf{h}_{*,K} = \boldsymbol{\phi}_k^T\boldsymbol{\theta} + z_k, \tag{32}$$

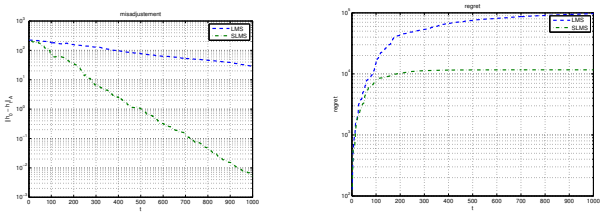

Figure 2: *An example of an experiment where $N = 1000$ and $d = 100$, a matrix $\mathbf{A}$ with $\|\mathbf{A}\|_2 = 1$ is chosen such that $10\mathbf{h}_*^T\mathbf{A}^{-1}\mathbf{h}_* \leq \mathbf{h}_*^T\mathbf{h}_*$. Panel (a) and (b) shows the evolution of the misadjustment and the regret of GD (blue dashed line), SGD (green dashed-dotted line).*

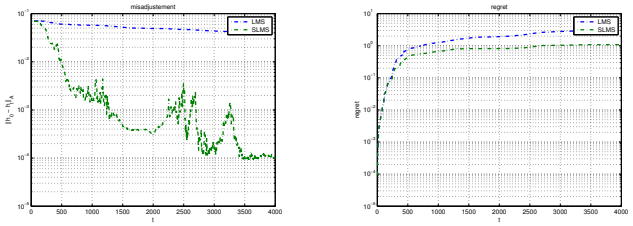

Figure 3: *An example of an AEC experiment with $d = 1001$ and $N = 4000$. Panel (a) and (b) shows the evolution of the misadjustment and the regret of GD (blue dashed line), SGD (green dashed-dotted line).*

where (i) $\phi_K \in \mathbb{R}^{n_\phi}$ is a (possibly infinite dimensional) vector summarizing characteristics of $\mathbf{h}_{*,K}$, (ii) $\theta \in \mathbb{R}^{n_\phi}$ is a vector of unknowns and (iii) z_k is as before. For notational convenience, let $\Phi = (\phi_1, \ldots \phi_d)^T \in \mathbb{R}^{d \times n_\phi}$. Now let the norm of a vector $\mathbf{h}_*$ be defined as

$$\|\mathbf{h}_*\|_k = \inf_{\theta,\mathbf{z}} \frac{1}{2}\theta^T\theta + \frac{\gamma}{2}\mathbf{z}^T\mathbf{z} \text{ s.t. } \mathbf{h}_* = \Phi\theta + \mathbf{z}, \tag{33}$$

with γ a fixed parameter. Then the representer theorem (see e.g. [10] and citations, or [7]) implies that this norm can be written as

$$\|\mathbf{h}_*\|_K = \mathbf{h}_*^T \left(\mathbf{K} + \frac{1}{\gamma} I_d\right)^{-1} \mathbf{h}_*, \tag{34}$$

with the kernel matrix defined as $\mathbf{K} = \Phi\Phi^T \in \mathbb{R}^{d \times d}$. It is found that we never have to construct explicitly the vectors ϕ_k, but a definition of a positive definitive kernel function $K(i, j)$ between two timelags $1 \leq i, j \leq d$ is enough for our needs. The matrix $\mathbf{K} \in \mathbb{R}^{d \times d}$ is then defined as $\mathbf{K}_{ij} = K(i, j)$ for all $1 \leq i, j \leq d$. In general, this kernel expresses how well two different lags are related. A common choice could be the case $K(i, j) \propto \exp(-(i - j)^2)$. Again, we can hence apply the SGD algorithm (6) using matrix $\mathbf{A} = \left(\mathbf{K} + \frac{1}{\gamma} I_d\right)$. Not that in this way we nowhere need to invert the kernel matrix, which was the computational bottleneck in many existing kernel techniques (see e.g. [7]).

4. EXAMPLE

We finish the paper with two examples illustrating the use of the above algorithm. The first example is constructed as follows. Consider the task of identification of the linear parameters of a linear model of order $d = 100$, obeying $y_t = \mathbf{x}_t\mathbf{h}_0 + e_t$, where $\mathbf{x}_t \sim \mathcal{N}(0_d, I_d)$, $\mathbf{h}_0 \sim \mathcal{N}(0_d, \mathbf{R})$ (with $\mathbf{R}$ as in (31)), and $e_t \sim \mathcal{N}(0, 10^{-3})$. Suppose 1000 independent samples are given and fed to the GD algorithm and SGD algorithm using the squared error loss. Here SGD is implemented using the matrix $\mathbf{A} = \mathbf{R}$ such that $\|\mathbf{h}_0\|^2 \geq 10\|\mathbf{h}\|^2_{\mathbf{A}^{-1}}$. An appropriate stepsize here is given by $\mu = 0.1$ and $\nu = 0.01$, determined by using a cross-validation argument. Figure 2 gives how the misadjustment $\|\mathbf{h}_t - \mathbf{h}_0\|^2$ (panel a) and the regret (panel b) behaves when $t = 1, \ldots, 100$. The surprising bit is that even using O(100) samples, SGD will perform well while standard GD (LMS) is clearly suboptimal. The second example (Figure 3) is based on an actual AEC experiment. Here we use an adaptive filter of $d = 1001$, and a far-end and near-end signal of length $N = 4000$. Besides the measured signals, we have the 'true' RIR determined by an identification experiment. Both GD and SGD implement Huber's loss with $\delta = 0.001$. The SGD is implemented as the empirical covariance matrix given as $\mathbf{A} = \frac{1}{5}\sum_{k=1}^{5} \mathbf{h}_k\mathbf{h}_k^T$ using 5 independent experimentally defined impulse responses measured under similar conditions.

5. DISCUSSION

Since we are dealing with RIR vectors having (relatively) many (say $O(1000)$) coefficients, it is crucial to exploit every bit of prior knowledge ones has. In this paper, SGD is proposed to deal with this, and its performance is found to be captured by a norm $\|\mathbf{h}_*\|$ of the target RIR $\mathbf{h}_*$. The choice for a specific norm is up to the user, but if $\|\mathbf{h}_*\|$ happens to be small in that norm, good performances can be guaranteed for SGD. A second important outcome of the analysis is that one could do better than LMS when adopting a suitable loss function. For example, in AEC, disturbance terms are typically non-Gaussian, and the choice for the robust Huber loss function is suggested.

References

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

[2] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp. Acoustic echo control. an application of very-high-order adaptive filters. *IEEE Signal Process. Mag.*, 16(4):42–69, July 1999.

[3] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999.

[4] P. S. R. Diniz. *Adaptive filtering: Algorithms and practical implementation*. Springer, Boston, MA, 2008.

[5] A. P. Liavas and P. A. Regalia. Acoustic echo cancellation: Do IIR models offer better modeling capabilities than their FIR counterparts. *IEEE Trans. Signal Process.*, 46(9):2499–2504, September 1998.

[6] O.L. Mangasarian and D.R. Musicant. Robust linear and support vector regression. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22(9):950–955, 2000.

[7] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[8] T. van Waterschoot, G. Rombouts, and M. Moonen. Optimally regularized adaptive filtering algorithms for room acoustic signal enhancement. *Signal Processing*, 88(3):594–611, March 2008.

[9] T. van Waterschoot, G. Rombouts, P. Verhoeve, and M. Moonen. Double-talk-robust prediction error identification algorithms for acoustic echo cancellation. *IEEE Trans. Signal Process.*, 55(3):846–858, March 2007.

[10] G. Wahba. *Spline models for observational data*. SIAM, 1990.