FLEXIBLE ADAPTIVE FILTERING BY MINIMIZATION OF ERROR ENTROPY BOUND
AND ITS APPLICATION TO SYSTEM IDENTIFICATION

Xi-Lin Li and Tülay Adalı

University of Maryland Baltimore County, Baltimore, MD 21250

ABSTRACT

It has been shown that using minimum error entropy as the cost function leads to important performance gains in adaptive filtering, especially when the Gaussianity assumptions on the error distribution do not hold. In this paper, we show that by using the entropy bound rather than the entropy, we can derive an efficient algorithm for supervised training. We demonstrate its effectiveness by a system identification problem using a generalized Gaussian noise model.

Index Terms— Adaptive filtering, supervised training, identification, minimum error entropy

1. INTRODUCTION

Mean-square error cost, leading to very practical adaptive algorithms, most important of which is the least-mean square algorithm, has been the standard cost function for many adaptive filtering solutions. However, it is optimal only when the error is Gaussian and information theoretic learning provides a more general framework for non-Gaussian and nonlinear signal processing [2]. Within this framework, the Renyi’s entropy of error signal $e$,

$$H_{\alpha}(e) = \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} p_E(e)^{\alpha} de$$

of order $\alpha > 0$ is used as the cost in [3, 4], where $p_E(e)$ is the probability density function (pdf) of the error term, $e$. The equivalence between minimization of Renyi entropy of order $\alpha$ for the error and the minimization of a Csiszar distance measure between the densities of desired and system outputs is established in [3, 4]. In particular, the Shannon entropy of $e$ can be obtained as

$$H(e) = \lim_{\alpha \to 1} H_{\alpha}(e) = - \int p_E(e) \log p_E(e) de.$$  

Since $H(e) = -E[\log p_E(e)]$, it is clear that minimization of the Shannon entropy is equivalent to the maximization of the likelihood $\prod_t p_E[e(t)]$. Unfortunately, neither the entropies nor the maximum likelihood function can be easily estimated, since $p_E(e)$ is unknown in general. In [2–4], nonparametric density and entropy estimators are proposed for the task. However, many nonparametric estimators are computationally demanding and require large sample sizes, which may limit their application for online learning or for adaptation in a time-varying environment, where it may be impossible to accumulate enough samples of the same distribution.

In this paper, we propose to use the Shannon entropy bound of error as the cost function for designing adaptive algorithms. We introduce a flexible and effective entropy estimator, and derive a new learning algorithm using this estimator, and show its effectiveness by a system identification problem. In the rest of this paper, we simply refer to Shannon entropy as entropy.

2. ESTIMATION OF ENTROPY AND ITS BOUND

2.1. Estimation of entropy

Given the observed samples of a random variable $e$, the problem is the estimation of its entropy $H(e) = -E[\log p_E(e)]$. Without the prior knowledge of the error distribution, nonparametric estimators are commonly used to estimate $H(e)$. One of the most widely used nonparametric estimator is the Parzen window estimator, where first the error pdf is estimated by using a Parzen window [8], and then the entropy estimate is obtained by using this pdf estimate. Estimates of entropy based on sample-spacings and nearest neighbor distances directly yield the entropy estimates, and do not need to estimate the pdf [7]. However, such nonparametric estimators use order statistics, which are not differentiable. Thus they cannot be easily used in adaptive filtering, where the gradient information is required for adjusting the parameters of the adaptive filter. Although the nonparametric estimators are asymptotically consistent, they are computationally demanding and typically require large sample sizes.

In parametric estimators, the error is assumed to be drawn from a pdf with a certain known parametric form. In this way, the problem of entropy estimation becomes a parameter estimation problem for the pdf, since the entropy is a function of the density parameters. Such an estimator works well if the assumed density model is a good match for the true distribution. Otherwise, it may fail or the performance degrades.
Increasing the complexity of the model on the other hand increases the difficulty of the estimation problem. The estimator we introduce next approaches the problem from a slightly different angle and poses the problem as the estimation of the entropy bound, which can be solved numerically and provides a very good tradeoff between effectiveness and computational complexity, which can be defined through a number of competing measuring functions.

2.2. Estimation of entropy bound

Instead of estimating the entropy, we estimate its upper bound and use it as an approximation of the entropy. This is an attractive solution for several reasons. First, underestimation of entropy is naturally avoided, which implies that the matching of the associated maximum entropy density is less committed with respect to unseen data [9]. Secondly, estimation of entropy bound is a well defined mathematical problem, and can be numerically solved [6]. In this subsection, we briefly review the entropy estimator introduced in [6].

It is well known that under the mean and variance constraints $E[e] = 0$ and $E[e^2] = \sigma^2$, the maximum entropy distribution is the Gaussian distribution $q(e) = \exp(-0.5e^2/\sigma^2)/\sqrt{2\pi\sigma^2}$, and the maximum entropy is given by

$$H_{\text{[Gaussian bound]}}(e) = 0.5 + 0.5 \log(2\pi\sigma^2).$$

Thus $H_{\text{[Gaussian bound]}}(e)$ is an estimate of the upper bound of the true entropy of $e$. It can be used as a cost function for adaptive filtering. In fact, this cost function is equivalent to the MSE cost since logarithm is a monotonic function.

To refine the above Gaussian entropy bound, we add one more measurement constraint on $e$ as $E[G(e)] = \mu_g$, where $G(\cdot)$ is an arbitrary function except for the linear combination of $e$ and $e^2$, and $\bar{e} = e/\sigma$ is the normalized error signal. Now we have the following maximum entropy problem

$$\max_q \quad -\int q(e) \log q(e) de, \quad \text{s.t.} \quad \int q(e) de = 1, \quad \int e^2 q(e) de = \sigma^2, \quad \int G(\bar{e}) q(e) de = \mu_g$$

where $q(e)$ is the pdf to be optimized. As shown in [6], the optimal solution to the above problem is given by

$$q(e) = A \exp[-ae^2 - be - cG(\bar{e})]$$

$$H_{\text{[bounded]}}(e) = 0.5 + 0.5 \log(2\pi\sigma^2) - V \{E[G(e)]\}$$

if it exists, where constants $A$, $a$, $b$ and $c$ are to be determined by the four constraints given in (1), and function $V(\cdot)$ can be numerically determined and it is always nonnegative.

To match different types of densities, several measuring functions can be used, say $G_k(\cdot)$, $k = 1, \ldots, K$. Each measuring function will lead to an upper bound of $H(e)$, and the tightest one can be used as the final estimation of $H(e)$, i.e.,

$$\hat{H}(e) = \min_{1 \leq k \leq K} H_{k}^{\text{[bounded]}}(e)$$

where

$$H_k^{\text{[bounded]}}(e) = 0.5 + 0.5 \log(2\pi\sigma^2) - V_k \{E[G_k(e)]\}$$

is the maximum entropy bound obtained by using the $k$th measuring function. As shown in [6], entropies of a wide range of distributions, those that have sub- or super-Gaussian, unimodal, bimodal, symmetric or skewed pdfs, can be approximated by using a few simple measuring functions. Since functions $V_k(\cdot)$ can be solved and saved in advance, this entropy estimator is computationally simple and easy to use. In addition, it can be shown that the maximum entropy distribution and entropy bound always exist for bounded measuring functions.

3. FILTERING BY ERROR ENTROPY BOUND MINIMIZATION

3.1. Signal model and two existing solutions

Consider an adaptive filtering problem where $x(t)$ is the known input to an unknown system and an adaptive system with adjustable parameters, $d(t)$ is the desired response, and $e(t)$ is the error signal. Though in general the unknown system in consideration can be linear or nonlinear, time-invariant or slowly time-varying, in this paper, to simplify the discussions, we assume that the adaptive system is a time-invariant linear filter with a finite impulse response (FIR) given by $w(\ell), 0 \leq \ell \leq L - 1$. Thus the error signal is written as

$$e(t) = d(t) - \sum_{\ell=0}^{L-1} w(\ell)x(t-\ell) = d(t) - w^T x(t)$$

where $w = [w(0), \ldots, w(L-1)]^T$, and $x(t) = [x(t), \ldots, x(t-L+1)]^T$.

The well known minimum MSE (MMSE) solution for the filter taps is given by the Wiener-Hopf equations $w^{\text{[MMSE]}} = E[x x^T]^{-1} E[d x]$ obtained by minimizing the MSE cost $E[e^2]$, which can be adaptively computed using the well known least mean squares adaptive algorithm or its variants [1]. The solution proposed in [3,4] is obtained by minimizing the Renyi’s entropy of the error signal. Quadratic Renyi’s entropy estimated by using a Gaussian kernel is proposed in [3,4], and it can be estimated as

$$\hat{H}_2(e) = -\log \Lambda(e)$$

where

$$\Lambda(e) = \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} \Phi(e(t_1) - e(t_2), 2\varsigma^2) / T^2$$

is the information potential [2], $T$ is the number of samples, $\Phi(u, \varsigma^2) = \exp\left(-0.5u^2 / \varsigma^2\right) / \sqrt{2\pi\varsigma^2}$ is the Gaussian kernel, and $\varsigma$ is the kernel size (or bandwidth). From (6) it is clear that minimization of the quadratic Renyi’s entropy is equivalent to the maximization of $\Lambda(e)$, whose gradient with respect to $w$
can be shown to be
\[
\frac{\partial \Lambda(e)}{\partial w} = \frac{1}{T^2} \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} \phi \left[ e(t_1) - e(t_2), 2\varsigma^2 \right] \left[ x(t_2) - x(t_1) \right]
\]
where \( \phi(\cdot) \) is the first order derivative of \( \Phi(\cdot) \). We use a gradient ascent algorithm for optimization of \( w \). The stochastic information gradient (SIG) can be used for the online implementation of this algorithm [5].

3.2. Filtering by entropy bound minimization

Let us assume that at \( n \)th iteration, the \( [k(n)] \)th measuring function leads to the tightest entropy bound. Then according to (3), the entropy of error signal is estimated as
\[
\hat{H}(e) = 0.5 + 0.5 \log(2\pi\sigma^2) - V_{k(n)} \left\{ E \left[ G_{k(n)}(\bar{e}) \right] \right\} . \tag{8}
\]
The gradient of \( \hat{H}(e) \) with respect to \( w \) is given by
\[
\frac{\partial \hat{H}(e)}{\partial w} = -\frac{E[ex]}{\sigma^2} - \nu \left\{ E \left[ G(\bar{e}) \right] \right\} E \left[ g(\bar{e}) \frac{\partial \bar{e}}{\partial w} \right] \tag{9}
\]
where \( \nu(\cdot) \) and \( g(\cdot) \) are the first order derivatives of \( V(\cdot) \) and \( G(\cdot) \) respectively, the subindex \( k(n) \) is suppressed for simplicity, and
\[
\frac{\partial \bar{e}}{\partial w} = \frac{eE[ex]}{\sigma^3} - \frac{x}{\sigma}.
\]
We use the covariance matrix \( E[x^2] \) as an approximation of the Hessian of the cost \( \hat{H}(e) \). Then the learning rule for \( w \) is
\[
w^{[\text{new}]} = w^{[\text{old}]} - \mu \left\{ E[x^2] \right\}^{-1} \frac{\partial \hat{H}(e)}{\partial w} \tag{10}
\]
where \( \mu \) is a positive step size. It is straightforward to obtain an online implementation of this algorithm by adopting learning rule (10) using a small block of data at each iteration, as done in the SIG algorithm [5].

4. APPLICATION TO SYSTEM IDENTIFICATION AND SIMULATION RESULTS

Suppose that the desired response is generated as
\[
d(t) = \sum_{\ell=0}^{L-1} h(\ell)x(t - \ell) + v(t)
\]
where \( h = [h(0), \ldots, h(L-1)]^T \) is the unknown impulse response to be identified, \( v(t) \) models the unknown additive noise, and \( d(t) \) and \( x(t) \) are known. Hence, for this problem \( w^{[\text{opt}]} = h \) minimizes the entropy of the error signal for the filter given in (5).

In our simulations, \( h \) is modeled as a random vector of length \( L = 5 \). At each run, the taps of \( h \) are independently drawn from the Gaussian distribution of zero mean and unit variance. The input signal \( x(t) \) is a white Gaussian noise of zero mean and unit variance. The additive noise \( v(t) \) is modeled as an independent and identically distributed (i.i.d.) sequence with samples drawn from the generalized Gaussian distribution (GGD) with pdf \( p_v(v) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{|v|^\beta}{\sigma^\beta}\right) \), where \( \beta > 0 \) is the shape parameter of GGD. Note that \( v \) is sub-Gaussian when \( \beta > 2 \), Gaussian when \( \beta = 2 \), and super-Gaussian when \( 0 < \beta < 2 \). The additive noise is scaled such that the signal to noise ratio (SNR) defined as
\[
E \left[ |d - v|^2 \right] / E \left[ |v|^2 \right] \]
is \(-5\) dB.

Our solution is compared with the MMSE solution and the solution obtained by minimizing the quadratic Renyi’s entropy of error signal [3, 4]. Only the batch processing results are presented and compared. For our entropy estimator, we use two measuring functions, \( G_1(e) = e^4 \) and \( G_2(e) = |e|/(1 + |e|) \). Measuring function \( G_1(e) \) is suitable for matching sub-Gaussian densities, and \( G_2(e) \) is proper for matching super-Gaussian densities. For the algorithm proposed in [3, 4], the bandwidth is set to \( \varsigma = 1.06\sqrt{E[|v|^2]/T^{-1/5}} \), the optimal value as suggested in [8]. The normalized identification error \( ||h - w||^2 / ||h||^2 \) is used as the performance index.

In Fig. 1 we show typical learning curves for the proposed algorithm, which shows its fast convergence for all three types of error distributions. Fig. 2 summarizes the average identification error with varying training sample sizes and shape parameters \( \beta \). We can draw several conclusions from Fig. 2. First, the performance of MMSE solution is rather insensitive to the statistic properties of noise as expected. However, it is important to note that this solution can be quite far from the best solution when the noise is non-Gaussian. The proposed algorithm performs the best, and consistently for both sub and super-Gaussian error values. It is also worth noting that the...
Gaussian noise leads to the largest identification error compared to the noise with other distributions at the same SNR level. The performance of the algorithm proposed in [3, 4] is between the MMSE solution and the proposed one for most cases. However, its performance is worse than the MMSE solution when the noise is close to Gaussian. This is especially true when the sample size is small, since it adopts a nonparametric estimator.

5. DISCUSSIONS AND CONCLUSIONS

We propose to use the entropy bound, rather than the entropy, of the error signal as the cost function for supervised training. A flexible entropy bound estimator is introduced, and an effective learning algorithm is proposed. We demonstrate its application by a system identification problem using a generalized Gaussian noise model. The proposed algorithm provides superior and reliable performance even at small sample sizes.

The further work will include the adaptive training of more general nonlinear systems as well design of efficient online learning algorithms.

6. REFERENCES