A Variable Step-Size Adaptive Algorithm under Maximum Correntropy Criterion

Ren Wang, Badong Chen, Nanning Zheng, Jose C. Principe
School of Electronic and Information Engineering
Xi’an Jiaotong University, Xi’an, 710049, China
chenbd@mail.xjtu.edu.cn

Abstract—Correntropy, a novel localized similarity measure defined in kernel space, has been successfully used as a cost function in adaptive system training. The adaptive algorithms under the maximum correntropy criterion (MCC) have been shown to be robust to impulsive non-Gaussian noises. However, they may converge slowly especially at a region far from the optimal solution. In this paper, we propose a new MCC algorithm with a variable step-size (VSS) called the VSS-MCC algorithm, which may achieve a much faster convergence speed while maintaining similar steady-state performance. In the new algorithm, the step-size is updated based on an approximation for the curvature of performance surface. Simulation results demonstrate the superior performance of VSS-MCC compared with the original MCC algorithm.

Keywords—correntropy; maximum correntropy criterion (MCC) curvature; variable step size

I. INTRODUCTION

Adaptive filtering techniques have been successfully applied in many different fields of signal processing and machine learning, such as channel equalization, noise cancellation, system identification, and so on. In general, an adaptive filtering algorithm is designed to update the parameters of an adaptive filter (the model) such that the filter output gets as close to the desired response as possible. Fig.1 shows a typical configuration of an adaptive filtering system. The cost function for an adaptive filtering algorithm is usually based on a second-order statistics, such as the mean square error (MSE), which measures the similarity between two random variables in a quadratic form. The well-known least mean square (LMS) algorithm [1] is a stochastic gradient based adaptive algorithm under MSE criterion. The second-order statistics based costs are mathematically tractable and computationally simple, and perform well with Gaussian distributed data, their performance will, however, degrade seriously with non-Gaussian data, especially when data contain large outliers or are contaminated by impulsive noises [2].

To improve the performance in non-Gaussian situations, many non-quadratic cost functions were proposed to develop new adaptive filtering algorithms. A typical example is the least mean fourth (LMF) criterion [3-6], which uses the fourth-order moment of the error as the adaptation cost function. The LMF algorithm can achieve excellent performance in a light-tailed non-Gaussian noise. However, it is rather sensitive to an impulsive noise and may converge very slowly to an optimal solution, although the initial convergence speed can be much faster. The mixed norm error criterion, which uses a weighted sum of the squared error and the fourth-order error, can improve the convergence behavior [7], but the performance can still be poor in the presence of an impulsive noise. To develop an adaptive algorithm with robustness against impulsive noises, the lower-order statistics can be applied [8-10]. For example, instead of using the squared error, the sign algorithm (SA) uses the absolute value of the error as the cost function and has been proven to be a very robust algorithm in impulsive noises. However, one drawback of the SA is the slower convergence performance especially for highly correlated input signals.

Correntropy, which is a localized similarity measure between two random variables, can also be used as an alternative cost function in adaptive filtering [11-15]. Since correntropy contains all even order moments of the data, it is particularly useful for non-Gaussian signal processing. Recent studies suggest that the adaptive filtering algorithms under maximum correntropy criterion (MCC) can be very robust in impulsive noises environments. Many important properties of the MCC criterion have been studied. In [16], it is proved that the maximum correntropy estimation is in essence a smoothed maximum a posteriori (MAP) estimation. The steady-state performance for adaptive filtering under MCC criterion has been analyzed in [17].
The MCC cost is invex which guarantees the global optimality of the solution [18]. However, there is a crucial drawback that the performance surface under MCC can be rather flat at the region far from the optimal solution (see Fig. 2). Thus, with a fixed step-size the gradient based adaptive algorithm may converge very slowly at the beginning. To overcome this problem, we propose in this work a variable step-size for MCC adaptive algorithm. We estimate the curvature of the performance surface, and use a larger step-size at the region with small curvature, and a small step-size at the region with large curvature. The new method can improve the convergence speed while maintaining similar steady-state performance.

The rest of the paper is organized as follows. In section II, we give a brief introduction about the maximum correntropy criterion. In section III, we develop the VSS-MCC algorithm. In section IV, we present simulation results to demonstrate the performance of the new method. Finally, a conclusion is given in section V.

Fig. 2. An one-dimensional performance surface with optimal solution $W^* = 10$

II. MAXIMUM CORRENTROPY CRITERION

The correntropy between two random variables $X$ and $Y$ is defined by [11-13]:

$$V_c(X, Y) = E[k_c(x-y)] = \int \int k_c(x-y)dF_{xy}(x,y)$$

where $E$ denotes the expectation operator, $F_{xy}(x,y)$ denotes the joint distribution function of $X$ and $Y$, and $k_c(\cdot)$ is a translation-invariant Mercer kernel, with bandwidth $\sigma$. In this paper, we use the following Gaussian kernel

$$k_c(x-y) = \exp \left( -\frac{|x-y|^2}{2\sigma^2} \right)$$

which can be rewritten as

$$k_c(x-y) = 1 - \frac{1}{2} \| \varphi(x) - \varphi(y) \|^2$$

where $\varphi(\cdot)$ denotes a nonlinear mapping induced by the kernel $k_c(\cdot)$, and $\Phi$ denotes the corresponding feature space (kernel space). Thus, correntropy can be viewed as an L_2 distance in kernel space.

Correntropy can be used as a cost function in adaptive filtering. In practical applications, one usually has to use a finite number of samples to estimate the correntropy, which leads to the following correntropic cost function

$$J_{MCC} = E[k_c(d-y)] = \frac{1}{N} \sum_{i=r}^{\infty} k_c(d_i-y_i)$$

where $N$ is the number of samples (or sliding data length for online learning scenarios), $d_i$ denotes the desired signal, and $y_i$ is the filter output.

Assume that the filter weight vector at the $n^{th}$ iteration is $W_n$. An iterative gradient ascent method can be applied to search an optimal solution. Specifically, a new weight vector can be obtained by adding an incremental item to the current weight vector, expressed as

$$W_{n+1} = W_n + \eta \nabla J_{MCC}$$

Combining (4) and (5) yields

$$W_{n+1} = W_n + \eta \sum_{i=r}^{\infty} \exp \left( -\frac{e_i^2}{2\sigma^2} \right)e_iX_i$$

where $\eta$ is the step-size parameter that should be chosen to ensure the convergence.

For finite-impulse-response (FIR) adaptive filtering, we have $y_i = W_n^TX_i$, and hence $\frac{\partial J_{MCC}}{\partial W_n} = X_i$. In this case, a gradient based algorithm, called the MCC algorithm, can be derived as [14]

$$W_{n+1} = W_n + \eta_n e_n X_n$$

where $\eta_n = \eta \exp(-\frac{e_n^2}{2\sigma^2})$. So the MCC can be viewed as a variable step-size LMS algorithm. When an outlier causes a larger error, the step-size $\eta_n$ will reduce to zero. This implies that the MCC has strong outlier rejection capability. In addition, if $\sigma \to \infty$, we have $\eta_n \to \eta$. In this case, the MCC reduce to the LMS algorithm. Obviously, MCC has almost the same computational complexity as the LMS.

III. VSS-MCC ALGORITHM

As mentioned previously, at a region far from the optimal solution, the performance surface may get very flat, which leads to a slow convergence speed if adopting a fixed step-size. An intuitive idea is to use a larger step-size at that region. How
to select a proper step-size for a specific region is however a crucial problem that is still open and needed to be addressed. Inspired by the fact that the gradient descent (or ascent) method is based on the linear approximation \[ f(x + \Delta x) \approx f(x) + f'(x) \cdot \Delta x \] (9)

we define in this work a factor \( \beta_n \) at the \( n \)th iteration:

\[
\beta_n = \frac{J_{\text{MCC}}^{n-1} - J_{\text{MCC}}^n}{J_{\text{MCC}}^{n-1} - J_{\text{MCC}}^n \Delta W_{n-1}}
\]

(10)

where

\[
J_{\text{MCC}}^{n+1} = \frac{1}{N} \sum_{i=0}^{N-1} \exp\left(-\frac{(d_i - W_{n,1}^2 X_i)^2}{2\sigma^2}\right)
\]

(11)

\[
J_{\text{MCC}}^{n+1}' = \frac{1}{N\sigma^2} \sum_{i=0}^{N-1} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)e_i X_i
\]

(12)

\[
\Delta W_{n-1} = \frac{\mu_n}{N\sigma^2} \sum_{i=0}^{N-1} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)e_i X_i
\]

(13)

\[
W_n = W_{n-1} + \mu_n \Delta W_{n-1}
\]

(14)

\[
J_{\text{MCC}}^{n+1}' = \frac{1}{N} \sum_{i=0}^{N-1} \exp\left(-\frac{(d_i - W_{n,1}^2 X_i)^2}{2\sigma^2}\right)
\]

(15)

where \( e_i \) denotes \( d_i - W_{n,1}^2 X_i \) and \( J_{\text{MCC}}^{n+1}' \) denotes the derivative with respect to \( W_{n-1} \).

The factor \( \beta_n \) is a measure of how well \( J_{n+1} + J_{n+1}' \Delta W_{n-1} \) approximates \( J_n \) in a sense that the closer \( \beta_n \) is to one, the better the approximation is. Actually, this factor reflects, to some extent, the magnitude of curvature at the point \( W_n \) in performance surface (see Fig. 3).

One can use the \( \beta_n \) to determine the step-size \( \mu_n \) at the \( n \)th iteration. In particular, we propose the following update rule for the step-size:

\[
\begin{cases} 
\mu_n = a\mu_n, & 1 - \delta \leq \beta_n \leq 1 + \delta (a > 1) \\
\mu_n = \mu_n, & \text{otherwise}
\end{cases}
\]

(16)

where \( \delta \) is a threshold with small positive value, \( a \) is a ratio larger than 1.0, and \( \mu_n \) is a predetermined step-size.

The proposed adaptive algorithm is called the variable step-size MCC (VSS-MCC) algorithm, whose pseudocode is summarized in Table 1.

Table 1. Pseudocode of VSS-MCC algorithm

**VSS-MCC**

**Initialization:**

predetermined step-size \( \mu_0 \), threshold \( \delta \), ratio \( a \), and initial weight vector \( W_0 \)

**Computation:**

While \( \{x_i, d_i\} \) available do

Compute \( e_n \)

Compute \( \beta_n \) using (10)

Compute \( \mu_n \) using (16)

Compute \( W_{n+1} = W_n + \mu_n \sum_{i=0}^{N-1} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)e_i X_i \)

End while

With the VSS-MCC algorithm, the overall configuration of the adaptive filtering is depicted in Fig. 4. There are two adaptation loops in the system, one for weights update, and the other for the step-size update. At each iteration, the step-size is automatically selected before updating the filter weights.

Fig. 4. Overall configuration for VSS-MCC algorithm

Compared with the original MCC algorithm, the VSS-MCC is still computationally simple. An extra computation is to compute the factor \( \beta_n \), whose computational complexity is only \( O(N) \), where \( N \) is the sliding data length.

**IV. SIMULATION RESULTS**

In this section, we present a simple two-tap FIR system identification example to demonstrate the performance of VSS-MCC, compared with the original MCC.
Assume that the desired signal is generated via
\[ d = W^T X + v, \]
where \( W^* = [10, 10]^T \) is the optimal weight vector with two dimensions, \( X = [x_{i-1}, x_i]^T \) is the tapped input vector in which \( \{x_i\} \) is a zero-mean white Gaussian process with variance 1.0, and \( v \) is an additive noise independent of the input, which is symmetric \( \alpha \)-stable (\( S\alpha S \)) distributed, with characteristic function
\[ \psi_{\nu, \gamma}(\omega) = \exp(-\gamma|\omega|^\alpha) \]
where scale parameter \( \gamma \) is 0.1 and exponential parameter \( \alpha \) is 1.1.

The adaptive filter is also a two-tap FIR system with weight vector \( W = [w_1, w_2]^T \) to be estimated. The kernel sizes for MCC and VSS-MCC are both set at 2.0, and the step-sizes (the predetermined step-size for VSS-MCC) are also both set at 2.0. The ratio parameter \( a \) and the threshold \( \delta \) in VSS-MCC are set at 10 and 0.05, respectively. In all the simulations below, we use a batch method (i.e. all the training data are used at each iteration) to update the filter weights and the performance measure is the weight error norm ratio (WENR) defined by
\[ \text{WENR} = \log\left(\frac{\|W^* - W\|}{\|W\|}\right) \]

Fig. 5 shows the convergence curves of MCC and VSS-MCC. It is clear that the VSS-MCC can achieve a much faster convergence speed while maintaining the same steady-state performance as that of MCC. Fig. 6 illustrates the convergence curves of VSS-MCC with different values of \( a \). As one can see, a larger \( a \) results in a faster convergence speed. However, if \( a \) is too large (e.g. \( a = 20 \)), the learning may fluctuate seriously or even diverge. In general, the choice of \( a \) is relatively easy as the new algorithm can work well with \( a \) in a wide range. Similar results can be observed from Fig. 7, which demonstrates the convergence curves of the VSS-MCC with different values of \( \delta \). As \( \delta \) becomes larger, the convergence speed will be faster.

V. CONCLUSION

The maximum correntropy criterion (MCC) has been successfully applied to adaptive system training in recent years, since it is robust with respect to impulsive noises. However, under MCC the convergence speed may be very slow at the initial stage. In this work, we propose a variable step-size method to improve the convergence speed of MCC. The new approach utilizes the curvature information of the performance surface. Compared with the original MCC algorithm, the variable step-size MCC (VSS-MCC) can achieve a much faster convergence speed while maintaining a desirable steady-state performance.

REFERENCES


