Generalized Correntropy for Robust Adaptive Filtering

Badong Chen, Senior Member, IEEE, Lei Xing, Haiquan Zhao, Member, IEEE
Nanning Zheng, Fellow, IEEE, José C. Príncipe, Fellow IEEE

Abstract—As a robust nonlinear similarity measure in kernel space, correntropy has received increasing attention in domains of machine learning and signal processing. In particular, the maximum correntropy criterion (MCC) has recently been successfully applied in robust regression and filtering. The default kernel function in correntropy is the Gaussian kernel, which is, of course, not always the best choice. In this work, we propose a generalized correntropy that adopts the generalized Gaussian density (GGD) function as the kernel, and present some important properties. We further propose the generalized maximum correntropy criterion (GMCC), and apply it to adaptive filtering. An adaptive algorithm, called the GMCC algorithm, is derived, and the stability problem and steady-state performance are studied. We show that the proposed algorithm is very stable and can achieve zero probability of divergence (POD). Simulation results confirm the theoretical expectations and demonstrate the desirable performance of the new algorithm.

Key Words: Correntropy; generalized correntropy; adaptive filtering; GMCC algorithm.

I. INTRODUCTION

Selecting a proper cost function (usually a statistical measure of the error signals) is a key issue in adaptive filtering theory and applications [1][2][3]. The mean square error (MSE) is widely used as a cost function since it has attractive features, such as smoothness, convexity, mathematical tractability, low computational burden and optimality under Gaussian assumption. The well-known least mean square (LMS) algorithm and its variants, such as normalized LMS (NLMS) and variable step-size LMS (VSSLMS), were developed under this criterion [1][2]. The MSE is desirable if the signals are Gaussian distributed. In non-Gaussian situations, however, its performance may degrade considerably and in these cases, a non-quadratic cost will be, in general, better than MSE [3].

Generally speaking, there are two types of non-Gaussian distributions: light-tailed (e.g. uniform, binary, etc.) and heavy-tailed (e.g. Laplace, α-stable, etc.) distributions. When the desired signals are disturbed by light-tailed non-Gaussian noises, a higher-order statistical (HOS) measure of the error is usually more desirable. A typical example is the least mean fourth (LMF) family algorithms, which use the mean even power of the error as the cost function [4]. Compared with the LMS algorithm, the LMF may achieve a faster convergence speed and a lower steady-state mean square deviation (MSD) especially in light-tailed noises. One drawback of the LMF algorithm however is that the stability is not guaranteed, which depends on the input and noise powers, and the initial values of the weights. A more general class of algorithms are the least mean p-power (LMP) family algorithms, which adopt the p-order absolute moment of the error as the adaptation cost [5].

When the desired signals are disturbed by heavy-tailed impulsive noises (which may cause large outliers), a lower-order statistical (LOS) measure of the error is usually more robust (i.e. less sensitive to impulsive interferences). For example, the sign algorithm (SA), which employs the mean absolute value of the error as the cost function, is rather robust to the presence of large noises [6][7][8]. The convergence speed and steady-state performance of the SA algorithm is however not so good in general. In the literature many other robust cost functions have been proposed to develop robust adaptive filtering algorithms. Typical examples include mixed-norm [9][10], M-estimate cost [11][12], and error entropy [13][14][15][16][17][18]. Particularly in recent years, the maximum correntropy criterion (MCC) has been successfully used in robust adaptive filtering, wherein the filter weights are adapted such that the correntropy between the desired signal and filter output is maximized [19][20][21][22][23][24]. The correntropy is a nonlinear and local similarity measure directly related to the probability of how similar two random variables are in a neighborhood of the joint space controlled by the kernel bandwidth, which also has its root in Renyi’s entropy (hence the name “correntropy”) [13][19]. Since correntropy is insensitive to outliers especially with a small kernel bandwidth, it is naturally a robust adaptation cost in presence of heavy-tailed impulsive noises.

The kernel function in correntropy is usually a Gaussian kernel, which is desirable due to its smoothness and strict positive-definiteness. With a Gaussian kernel, the correntropy induces a nonlinear metric called the correntropy induced metric (CIM) which behaves like an $L_2$ norm when data are relatively small compared with the kernel bandwidth, an $L_1$ norm as data get larger, and an $L_0$ norm when data are far away from the origin [19][25]. However, Gaussian kernel is, of course, not always the best choice. In the present work, we propose to use the generalized Gaussian density (GGD) [26][27] function as a kernel function in correntropy, and the new correntropy is called the generalized correntropy.

Some important properties of the generalized correntropy are presented. In particular, we show that the order-α generalized correntropy induced metric (GCIM) or generalized correntropy...
ic loss (GC-loss) function behaves like different norms (from $L_\alpha$ to $L_0$) of the data in different regions.

Similar to the usual correntropy with Gaussian kernel, the generalized correntropy can also be used as an optimization tool in estimation-related problems. In this work, we focus mainly on applying the generalized maximum correntropy criterion (GMCC) to adaptive filtering. We show that the optimal solution of GMCC filtering is in form similar to the well-known Wiener solution, except that the autocorrelation matrix and cross-correlation vector are weighted by an error nonlinearity. If the signals involved are zero-mean Gaussian, the optimal solution will equal to the Wiener solution. Under the GMCC criterion, a stochastic gradient based adaptive filtering algorithm, called the GMCC algorithm, is developed. The stability problem and the steady-state performance of the GMCC algorithm are analyzed. In particular, we present a simple example to show that the GMCC may have a zero probability of divergence (POD). A theoretical value of the steady-state excess mean square error (EMSE) of the GMCC algorithm is also derived. Simulation results confirm the theoretical expectations and the desirable performance of the GMCC.

The rest of the paper is organized as follows. In section II, we define the generalized correntropy, and present some important properties. In section III, we apply the generalized maximum correntropy criterion (GMCC) to adaptive filtering and develop the GMCC algorithm. In section IV, we analyze the stability and steady-state performance of the GMCC algorithm. In section V, we present illustrative examples to verify the theoretical results and demonstrate the good performance of the new algorithm. Finally in section VI, we give the conclusion.

II. GENERALIZED CORRENTROPY

A. Definition

Given two random variables $X$ and $Y$, the correntropy is defined by [19][28]

$$V(X,Y) = E[\kappa(X,Y)] = \int \kappa(x,y) dF_{XY}(x,y)$$

(1)

where $E$ denotes the expectation operator, $\kappa(\cdot, \cdot)$ is a shift-invariant Mercer kernel, and $F_{XY}(x,y)$ denotes the joint distribution function of $(X,Y)$. Without mentioned otherwise, the kernel function of correntropy is the Gaussian kernel:

$$\kappa(x,y) = G_\sigma(e) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{e^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\lambda e^2\right)$$

(2)

where $e = x - y$, $\sigma > 0$ is the kernel bandwidth, and $\lambda = 1/2\sigma^2$ is the kernel parameter. The correntropy $V(X,Y)$ can also be expressed as

$$V(X,Y) = E[\varphi(X)^T \varphi(Y)] = \text{trace}(R_{\varphi(X)\varphi(Y)})$$

(3)

where $R_{\varphi(X)\varphi(Y)} = E[\varphi(X)\varphi(Y)^T]$, and $\varphi(\cdot)$ denotes a nonlinear mapping induced by $\kappa$, which transforms its argument into a high-dimensional (infinite for Gaussian kernels) Hilbert space $\mathcal{F}_\kappa$, satisfying $\varphi(X)^T \varphi(Y) = \kappa(X,Y)$ [19]. The correntropy is therefore essentially a second-order statistic of the mapped feature space data.

There is a well-known generalization of Gaussian density function, called the generalized Gaussian density (GGD) function, which with zero-mean is given by [26][27]

$$G_{\alpha,\beta}(e) = \frac{\alpha}{2\beta \Gamma(1/\alpha)} \exp\left(-\frac{|e|^{\alpha}}{\beta}\right) \gamma_{\alpha,\beta} \exp\left(-\lambda |e|^{\alpha}\right)$$

(4)

where $\Gamma(\cdot)$ is the gamma function, $\alpha > 0$ is the shape parameter, $\beta > 0$ is the scale (bandwidth) parameter, $\lambda = 1/\beta^\alpha$ is the kernel parameter, and $\gamma_{\alpha,\beta} = \alpha/(2\beta \Gamma(1/\alpha))$ is the normalization constant. This parametric family of symmetric distributions include the Gaussian ($\alpha = 2$) and Laplace ($\alpha = 1$) distributions as the special cases. As $\alpha \to \infty$, the GGD density converges point-wise to a uniform density on $(-\beta, \beta)$.

In this work, we use the GGD density function as the kernel function of correntropy, and define

$$V_{\alpha,\beta}(X,Y) = E[G_{\alpha,\beta}(X - Y)]$$

(5)

To make a distinction between (5) and the correntropy with Gaussian kernel, we call it the generalized correntropy. Clearly, the correntropy with Gaussian kernel corresponds to the generalized correntropy with $\alpha = 2$.

Remark 1: It is worth noting that in the generalized correntropy, the kernel function does not necessarily satisfy the Mercer’s condition. Actually, the kernel function $\kappa(x, y) = G_{\alpha,\beta}(x - y)$ is positive definite if and only if $0 < \alpha \leq 2$ (see [29], page 434).

In practice, the joint distribution of $X$ and $Y$ is usually unknown, and only a finite number of samples $\{(x_i, y_i)\}_{i=1}^N$ are available. In this case, the sample mean estimator of the generalized correntropy is

$$\hat{V}_{\alpha,\beta}(X,Y) = \frac{1}{N} \sum_{i=1}^N G_{\alpha,\beta}(x_i - y_i)$$

(6)

B. Properties

Below we present several basic properties of the generalized correntropy. Some of them are simple extensions of the properties in [19], and hence will not be proved here.

Property 1: $V_{\alpha,\beta}(X,Y)$ is symmetric, i.e. $V_{\alpha,\beta}(X,Y) = V_{\alpha,\beta}(Y,X)$.

Property 2: $V_{\alpha,\beta}(X,Y)$ is positive and bounded: $0 < V_{\alpha,\beta}(X,Y) \leq G_{\alpha,\beta}(0) = \gamma_{\alpha,\beta}$, and it reaches its maximum if and only if $X = Y$.

Property 3: The generalized correntropy involves higher-order absolute moments of the error $X - Y$: $V_{\alpha,\beta}(X,Y) = \gamma_{\alpha,\beta} \sum_{n=0}^\infty \frac{(-\lambda)^n}{n!} E \left[|X - Y|^{\alpha n}\right]$

Remark 2: When the kernel parameter $\lambda$ is small enough, we have $V_{\alpha,\beta}(X,Y) \approx \gamma_{\alpha,\beta} (1 - \lambda E[|X - Y|^{\alpha}])$. In this case, the generalized correntropy is, approximately, an affine linear function of the $\alpha$-order absolute moment of the error.
Property 4: Assume that the samples \( \{(x_i, y_i)\}_{i=1}^N \) are drawn from the joint PDF \( p_{XY}(x, y) \). Let \( \hat{p}_e(\cdot) \) be the Parzen estimate of the error PDF from the samples \( \{e_i = x_i - y_i\}_{i=1}^N \), with the GGD density function \( G_{\alpha,\beta} \) as the Parzen window kernel. Then \( \hat{V}_{\alpha,\beta}(X, Y) \) is the value of \( \hat{p}_e(\cdot) \) evaluated at the point zero, that is

\[
\hat{V}_{\alpha,\beta}(X, Y) = \hat{p}_e(0)
\]  

(7)

where \( \hat{p}_e(\varepsilon) = \frac{1}{N} \sum_{i=1}^N G_{\alpha,\beta}(\varepsilon - e_i) \).

Property 5: For the case \( 0 < \alpha \leq 2 \), the generalized correntropy is a second-order statistic of the mapped feature space data.

Proof: When \( 0 < \alpha \leq 2 \), the kernel function \( \kappa(x, y) = G_{\alpha,\beta}(x - y) \) is a Mercer kernel, and hence we have

\[
V_{\alpha,\beta}(X, Y) = E\left[ p_{\alpha,\beta}(X)^T p_{\alpha,\beta}(Y) \right]
\]

where \( p_{\alpha,\beta}(\cdot) \) is a nonlinear mapping induced by \( G_{\alpha,\beta} \).

In data analysis such as regression and classification, a measure called the correntropic loss (C-loss) can be used instead of using the correntropy [30][31]. A generalized C-loss (GC-loss) function between \( X \) and \( Y \) can be defined as

\[
J_{GC-loss}(X, Y) = G_{\alpha,\beta}(0) - V_{\alpha,\beta}(X, Y)
\]  

(8)

the GC-loss satisfies \( J_{GC-loss}(X, Y) \geq 0 \), and when \( 0 < \alpha \leq 2 \), it can be expressed as

\[
J_{GC-loss}(X, Y) = \frac{1}{2} E\left[ \| p_{\alpha,\beta}(X) - p_{\alpha,\beta}(Y) \|^2 \right]
\]  

(9)

which is a mean-square loss in the feature space \( \mathcal{F}_\alpha \) induced by the Mercer kernel \( \kappa(x, y) = G_{\alpha,\beta}(x - y) \). Clearly, minimizing the GC-loss will be equivalent to maximizing the generalized correntropy.

Let \( \{(x_i, y_i)\}_{i=1}^N \) be \( N \) samples drawn from \( p_{XY} \). An estimator of the GC-loss is

\[
\hat{J}_{GC-loss}(X, Y) = G_{\alpha,\beta}(0) - \hat{V}_{\alpha,\beta}(X, Y)
\]

\[
= \gamma_{\alpha,\beta} - \frac{1}{N} \sum_{i=1}^N G_{\alpha,\beta}(x_i - y_i)
\]

(10)


Property 6: Let \( X = [x_1, \cdots, x_N]^T, Y = [y_1, \cdots, y_N]^T \).

Then the function \( GCIM(X, Y) = \sqrt{J_{GC-loss}(X, Y)} \), called the generalized correntropy induced metric (GCIM), defines a metric in the \( N \)-dimensional sample vector space when \( 0 < \alpha \leq 2 \).

Proof: See Appendix A.

For the case in which \( X = [x_1, x_2]^T, Y = [0, 0]^T, \alpha = 4 \), and \( \lambda = 1 \), the surface of the GCIM\( (X, Y) \) with respect to \( x_1 \) and \( x_2 \) is shown in Fig. 1. As one can see, the GCIM behaves like different norms (from \( L_\alpha \) to \( L_0 \)) of the data in different regions. This observation, which is similar to that obtained with Gaussian kernel [19], is confirmed by Properties 7 and 8 below.

Property 7: As \( \lambda \to 0^+ \) (or \( x_i \to 0, i = 1, \cdots, N \)), the function \( L_{\alpha,\beta}(X) = \left( \frac{1}{N} \hat{J}_{GC-loss}(X, 0) \right)^{1/\alpha} \) will approach the \( L_\alpha \)-norm of \( X \), that is

\[
L_{\alpha,\beta}(X) \approx \|X\|_\alpha = \left( \sum_{i=1}^N |x_i|^\alpha \right)^{1/\alpha}, \text{as}\lambda \to 0
\]  

(11)

Proof: See Appendix B.

Property 8: Assume that \( |x_i| > \delta, \forall i : x_i \neq 0 \), where \( \delta \) is a small positive number. As \( \lambda \to \infty \) (or \( \beta \to 0^+ \)), minimizing the function \( L_{\alpha,\beta}(X) \) will be, approximately, equivalent to minimizing the \( L_0 \)-norm of \( X \), that is

\[
\min_{X \in \Omega} L_{\alpha,\beta}(X) \sim \min_{X \in \Omega} \|X\|_0, \text{as}\lambda \to \infty
\]  

(12)

where \( \Omega \) denotes a feasible set of \( X \).

Proof: Similar to the one presented in [32]. See Appendix C for the detailed derivation.

Below we present some optimization properties of the GCIM. Similar results for the C-loss can be found in [31].

Property 9: Let \( e = [e_1, \cdots, e_N]^T \). Then the following statements hold:

1) if \( 0 < \alpha \leq 1 \), then the GC-loss \( \hat{J}_{GC-loss} \) is concave at any \( e \) with \( e_i \neq 0 \) (\( i = 1, \cdots, N \));

2) if \( \alpha > 1 \), then the GC-loss \( \hat{J}_{GC-loss} \) is convex at any \( e \) with \( 0 < |e_i| \leq \left( \frac{1}{\alpha - 1} \right)^{1/\alpha} (i = 1, \cdots, N) \);

3) if \( \lambda \to 0^+ \), then for any \( e \) with \( e_i \neq 0 \) (\( i = 1, \cdots, N \)), the GC-loss \( \hat{J}_{GC-loss} \) is concave at \( e \) for \( 0 < \alpha \leq 1 \), and convex at \( e \) for \( \alpha > 1 \).

Proof: See Appendix D.

Property 10: For \( \alpha > 1 \), the GC-loss \( \hat{J}_{GC-loss} \) is a differentiable invex function of \( e = [e_1, \cdots, e_N]^T \) with \( e_i \leq M \) (\( i = 1, \cdots, N \)), where \( M \) is an arbitrary positive number.

Proof: See Appendix E.

Fig. 1. Surface of the GCIM in 3D space (\( \alpha = 4, \lambda = 1 \))
III. ADAPTIVE FILTERING UNDER GMCC CRITERION

A. Cost function

Similar to the correntropy, the generalized correntropy can also be used as an optimization cost in estimation-related problems. In the context of linear adaptive filtering, under the generalized maximum correntropy criterion (GMCC), the optimal weight vector of the filter can be solved by minimizing the following GC-loss function:

$$J_{GC-loss} = \gamma_{\alpha,\beta} - E\left[ G_{\alpha,\beta}(e(i)) \right]$$

where the error

$$e(i) = d(i) - y(i) = d(i) - W^T X(i)$$

with $d(i) \in \mathbb{R}$ being the desired value at time $i$, $y(i) = W^T X(i)$ the output of the filter, $W = [w_1, w_2, \ldots, w_m]^T \in \mathbb{R}^m$ the weight vector, and $X(i) \in \mathbb{R}$ the input vector, usually given by

$$X(i) = [x(i), x(i-1), \ldots, x(i-m+1)]^T$$

where $x(i)$ is the input signal.

B. Optimal solution

On the optimal weight vector of the GMCC adaptive filter, we summarize the main results as the following theorems.

Theorem 1: The optimal weight vector that minimizes $J_{GC-loss}$ can be expressed as

$$W_{opt} = [R^h_{XX}]^{-1} P^h_{dx}$$

where $R^h_{XX} = E\left[ h(e(i)) X(i) X(i)^T \right]$ is a weighted autocorrelation matrix of the input signal, in which the weighting is a function of $h(e(i)) = \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-2}$, and $P^h_{dx} = E\left[ h(e(i)) d(i) X(i) \right]$ is a weighted cross-correlation vector between the desired and the input vector.

Proof: Let $\frac{\partial}{\partial W} J_{GC-loss} = 0$, we have

$$E\left[ \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-2} \cdot \text{sign}(e(i)) X(i) \right] = 0$$

$$\Rightarrow E\left[ h(e(i)) (d(i) - W^T X(i)) X(i) \right] = 0$$

$$\Rightarrow W = \left[ R^h_{XX} \right]^{-1} P^h_{dx}$$

Remark 3: Note that the above solution is not a closed-form solution, since both matrices $R^h_{XX}$ and $P^h_{dx}$ depend on the weight vector $W$ through the error $e(i)$. The solution (16) is actually a fixed-point equation.

For the case $\alpha = 2$, we have $h(e(i)) = \exp(-\lambda e^2(i))$. In this case, as $\lambda \to 0+$, we have $h(e(i)) \to 1$, and $W_{opt} \to R^{-1}_{XX} P_{dx}$, with $R_{XX} = E\left[ X(i) X(i)^T \right]$, $P_{dx} = E\left[ d(i) X(i) \right]$ , which corresponds to the well-known Wiener solution.

Theorem 2: If and are both zero-mean Gaussian processes, then the optimal solution under GMCC criterion is equal to the Wiener solution.

Proof: See Appendix F.

C. Adaptive algorithm

Based on the cost function (13), a stochastic gradient based adaptive algorithm, called in this work the GMCC algorithm, can be simply derived as

$$W(i + 1) = W(i) + \mu \frac{\partial}{\partial W} \exp(-\lambda|e(i)|^\alpha)$$

$$= W(i) - \mu \lambda |e(i)|^\alpha \text{sign}(e(i)) \frac{\partial e(i)}{\partial W(i)}$$

$$= W(i) + \eta \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-2} \text{sign}(e(i)) X(i)$$

(18)

where $\eta = \mu \lambda$ is the step-size parameter.

We have the following observations:

1) When $\alpha = 2$, the GMCC algorithm becomes

$$W(i + 1) = W(i) + \eta \exp(-\lambda e^2(i)) e(i) X(i)$$

(19)

which is the original MCC algorithm [20].

2) The weight update equation (18) can be rewritten as

$$W(i + 1) = W(i) + \eta |e(i)|^{\alpha-1} \text{sign}(e(i)) X(i)$$

(20)

where $\eta = \eta \exp(-\lambda e^2(i))$. Therefore, the GMCC algorithm can be viewed as an LMP algorithm with $p = \alpha$ and a variable step-size $\eta(i)$. The LMP algorithm is derived under the LMP (least mean $P$-power) criterion, including the SA ($p = 1$), LMS ($p = 2$) and LMF ($p = 4$) as special cases (see [5] for the details).

3) When $\lambda \to 0+$, we have $\eta(i) \to \eta$. In this case, the GMCC algorithm reduces to the traditional LMP algorithm with $p = \alpha$:

$$W(i) = W(i - 1) + \eta |e(i)|^{\alpha-1} \text{sign}(e(i)) X(i)$$

(21)

In particular, when $\alpha = 2$, (21) becomes the well-known LMS algorithm:

$$W(i) = W(i - 1) + \eta e(i) X(i)$$

(22)

4) When $|e(i)| \to \infty$, we have $\eta(i) \to 0$. Thus, a large error will have little influence on the filter weights. This implies that the GMCC algorithm will be robust to large outliers (or impulsive noises), which often cause large errors.

Remark 4: One can derive various variants of the GMCC algorithm, such as the variable step-size GMCC and normalized GMCC, where the step-size is changed across iterations or divided by the squared norm of the input vector.

The computational complexity of the GMCC algorithm is almost the same as that of the LMP algorithm, and the only extra computational effort needed is to calculate the term $\exp(-\lambda|e(i)|^\alpha)$, which is obviously not expensive.

IV. STABILITY AND STEADY-STATE PERFORMANCE

In this section, we analyze the stability and steady-state mean-square performance of the proposed GMCC algorithm. The algorithm (18) can be written in a general form:

$$W(i) = W(i - 1) + \eta f(e(i)) X(i)$$

(23)

where $f(e(i))$ is a nonlinear function of $e(i)$, that is,

$$f(e(i)) = \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-1} \text{sign}(e(i))$$

(24)
Assume that the desired signal $d(i)$ can be expressed as

$$d(i) = W_0^T X(i) + v(i)$$  \hspace{1cm} (25)

where $W_0$ denotes an unknown weight vector that needs to be estimated, and $v(i)$ stands for the disturbance noise, with variance $\sigma_v^2$. Then, we have

$$e(i) = \tilde{W}(i - 1)^T X(i) + v(i) = e_a(i) + v(i)$$  \hspace{1cm} (26)

where $\tilde{W}(i - 1) = W_0 - W(i - 1)$ is the weight error vector at iteration $i - 1$, and $e_a(i) = \tilde{W}(i - 1)^T X(i)$ is referred to as the \textit{a priori} error. After some simple algebra, one can obtain [33]:

$$E \left[ \| \tilde{W}(i) \|^2 \right] = E \left[ \| \tilde{W}(i - 1) \|^2 \right] - 2\eta E \left[ e_a(i) f(e(i)) \right] + \eta^2 E \left[ \| X(i) \|^2 f^2(e(i)) \right]$$  \hspace{1cm} (27)

### A. Stability Analysis

From (27), if the step-size $\eta$ is chosen such that for all $i$

$$\eta \leq \frac{2E \left[ e_a(i) f(e(i)) \right]}{E \left[ \| X(i) \|^2 f^2(e(i)) \right]} = \frac{2E \left[ e_a(i) \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-1} \text{sign}(e(i)) \right]}{E \left[ \| X(i) \|^2 \exp(-2\lambda|e(i)|^\alpha) |e(i)|^{2(\alpha-1)} \right]}$$  \hspace{1cm} (28)

then the sequence of weight error power $E \left[ \| \tilde{W}(i) \|^2 \right]$ will be decreasing and converging.

The bound in (28) seems very complicated and is not easy to be verified in a practical situation. To better understand the stability of the GMCC, it’s more important to investigate the \textit{probability of divergence} (POD) [34]. Here the divergence means $\lim_{i \to \infty} \| \tilde{W}(i) \|^2 = \infty$ in a realization of an adaptive algorithm. For the LMF algorithm, the POD is nonzero when the input distribution has infinite support, no matter how small the step-size is chosen [34]. Below we present a simple example to show that the GMCC is very stable and may have zero POD, no matter what input distribution. Let’s consider a scalar filtering case as in [34], in which the desired signal is $d(i) = W_0 X(i)$, where $W_0$ and $X(i)$ are both scalars, and the noise $v(i)$ is assumed to be zero for simplicity. Then, we have

$$\tilde{W}(i) = \tilde{W}(i - 1) - \eta f(e(i)) X(i)$$

$$= \left[ 1 - \eta \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-2} X^2(i) \right] \tilde{W}(i - 1)$$

$$= \left[ 1 - \eta \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha} \tilde{W}(i - 1) \right]^{-2} \tilde{W}(i - 1)$$  \hspace{1cm} (29)

It is easy to show

$$0 \leq \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha} \leq \frac{1}{\lambda} \exp(-1)$$  \hspace{1cm} (30)

So it holds that if $\left| \tilde{W}(i - 1) \right|^2 \geq \frac{\eta}{\lambda} \exp(-1)$, then

$$0 \leq \eta \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha} \left| \tilde{W}(i - 1) \right|^2 \leq 2$$  \hspace{1cm} (31)

In this case, we have

$$\left| \tilde{W}(i) \right|^2 = \left| 1 - \eta \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha} \tilde{W}(i - 1) \right|^{-2} \tilde{W}(i - 1)$$

$$\times \left| \tilde{W}(i - 1) \right|^2 \leq \left| \tilde{W}(i - 1) \right|^2 \left( 1 - \eta \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha} \right)$$

$$\left| \tilde{W}(i) \right|^2 \leq \left| \tilde{W}(i - 1) \right|^2$$  \hspace{1cm} (32)

Therefore, the limit (if exists) of $\left| \tilde{W}(i) \right|$ always satisfies

$$\lim_{i \to \infty} \left| \tilde{W}(i) \right| < \infty$$

which implies that in this simple example the GMCC will never diverge (or its POD is zero). Note that in this case the LMF will have a non-zero POD [34].

When $\tilde{W}(i)$ is a vector and there is a noise $v(i)$, a rigorous analysis of the POD of GMCC is very difficult and is left open in this work. However, we believe that the above scalar example explains clearly what is the mechanism of the zero POD of GMCC. Our simulation results also suggest that in most cases the POD of GMCC is zero, even when the noise signal contains large outliers.

### B. Steady-State Mean Square Performance

With a similar derivation presented in [35], one can analyze the mean square transient behaviors of the algorithm (18). This is a trivial but quite tedious task since we have to evaluate the expectations

$$E \left[ e_a(i) \exp(-\lambda|e(i)|^\alpha) |e(i)|^{\alpha-1} \text{sign}(e(i)) \right]$$

and

$$E \left[ \| X(i) \|^2 \exp(-2\lambda|e(i)|^\alpha) |e(i)|^{2(\alpha-1)} \right].$$

In the following, we only analyze the steady-state mean square performance by using a Taylor expansion method [23].

As the filter reaches the steady-state, we have

$$E \left[ \| \tilde{W}(i) \|^2 \right] = E \left[ \| \tilde{W}(i - 1) \|^2 \right].$$

By (27), it holds that

$$2E \left[ e_a(i) f(e(i)) \right] = \eta E \left[ \| X(i) \|^2 f^2(e(i)) \right]$$  \hspace{1cm} (33)

Assume that $\| X(i) \|^2$ is asymptotically uncorrelated with $f^2(e(i))$ (the rationality of this assumption has been discussed in [33]). Then (33) becomes

$$2E \left[ e_a(i) f(e(i)) \right] = \eta T_r (R_{XX}) E \left[ f^2(e(i)) \right]$$  \hspace{1cm} (34)

In the steady-state, the distributions of $e_a(i)$ and $e(i)$ are independent of $i$, thus one can omit the time index and simply write (34) as

$$2E \left[ e_a f(e) \right] = \eta T_r (R_{XX}) E \left[ f^2(e) \right]$$  \hspace{1cm} (35)

Let $S = \lim_{i \to \infty} E \left[ e_a^2(i) \right] = E \left[ e_a^2 \right]$ be the steady-state \textit{excess mean square error} (EMSE). An approximate analytical expression of $S$ can be derived. Before proceeding we give two common assumptions:

**A1**: The noise $v(i)$ is zero-mean, independent, identically distributed, and is independent of the input.
When the step-size $\eta$ to
in which
Substituting (37) and (38) into (35), we obtain
where $f'(v)$ and $f''(v)$ are the first and second derivatives, and $o(e_n^2)$ denotes the third and higher-order terms. If $\mathbb{E}[o(e_n^2)]$ is small enough, then based on the assumptions $A1$ and $A2$ we can derive
\[
\mathbb{E}[e_n f(e)] = \mathbb{E}[e_n f(v) + f'(v)e_n^2 + o(e_n^2)]
\approx \mathbb{E}[f'(v)] S
\]
(37)
\[
\mathbb{E}[f^2(v)] \approx \mathbb{E}[f^2(v)] + \mathbb{E}[f'(v)f''(v) + |f'(v)|^2] S
\]
(38)
Substituting (37) and (38) into (35), we obtain
\[
S = \frac{\eta \frac{\text{Tr}(R_{XX})}{2} \mathbb{E}[f^2(v)] - \eta \frac{\text{Tr}(R_{XX})}{2} \mathbb{E}[f'(v)f''(v) + |f'(v)|^2]}
\]
\[
= \frac{\mathbb{E}\left[\exp(-2\lambda|v|^\alpha)|v|^{2\alpha-2}\right] - \eta \frac{\text{Tr}(R_{XX})}{2} \mathbb{E}[\zeta(v)]}{\mathbb{E}[\exp(-\lambda|v|^\alpha)|v|^{\alpha-2}((\alpha-1) - \lambda|v|^\alpha)]}
\]
(39)
in which
\[
\zeta(v) = \left[f'(v)f''(v) + |f'(v)|^2\right]
\]
\[
= \exp\left(-2\lambda|v|^\alpha\right)|v|^{2\alpha-4} \times
\]
\[
\left[(\alpha-1)(2\alpha-3) - 5\lambda(\alpha-1)|v|^\alpha + 2\lambda^2\alpha^2|v|^{2\alpha}\right]
\]
(40)
When the step-size $\eta$ is small enough, (39) can be simplified to
\[
S \approx \frac{\eta \text{Tr}(R_{XX}) \mathbb{E}\left[\exp(-2\lambda|v|^\alpha)|v|^{2\alpha-2}\right] - \eta \frac{\text{Tr}(R_{XX})}{2} \mathbb{E}[\zeta(v)]}{\mathbb{E}[\exp(-\lambda|v|^\alpha)|v|^{\alpha-2}((\alpha-1) - \lambda|v|^\alpha)]}
\]
(41)

\textbf{Remark 5}: Given a noise distribution, one can evaluate the expectations in (39) and obtain a theoretical value of the steady-state EMSE. It is, however, worth noting that the steady-state EMSE of (39) is derived under the assumption that the steady-state \textit{a priori} error $e_n$ is small such that its third and higher-order terms are negligible. When the step-size or noise power is too large, the \textit{a priori} error will also be large. In this case, the derived EMSE value will not accurately enough characterize the performance.

\section{V. SIMULATION RESULTS}

Now we present simulation results to confirm the theoretical predictions and demonstrate the desirable performance of the proposed GMCC algorithm.

\section{A. Probability of divergence}

First, we investigate the stability problem of the GMCC algorithm. The steady-state performance is valid only when the algorithm does not diverge. In many cases, however, an adaptive algorithm may diverge especially at the initial convergence stage. Below we present some simulation results about the probability of divergence (POD) of the GMCC (with $\alpha = 4.0, \lambda = 2.0$), compared with that of the LMF algorithm, whose probability of divergence has been studied in [34]. The weight vector of the unknown system is assumed to be $W_0 = [0, 1, 0.2, 0.3, 0.4, 0.5, 0.4, 0.3, 0.2, 0.1]$, and the initial weight vector of the adaptive filter is a null vector. The input signal and the disturbance noise are both zero-mean Gaussian with variance 1.0. The PODs with different step-sizes are illustrated in Fig. 2. To evaluate the PODs, 1000 independent Monte Carlo simulations were performed and in each simulation, 1000 iterations were run. We labeled a learning curve as “diverging” if at the last iteration the weight error power $\|W_n - W(i)\|^2$ is larger than 100. As one can see clearly, compared with the LMF, the GMCC is very stable and does not diverge in this particular situation.

\section{B. Steady-state performance}

Second, we show the theoretical and simulated steady-state performance of the GMCC. In the simulation, we set $\alpha = 4$, and $\lambda = 0.03$. The filter length is 20, the input signal is a zero-mean white Gaussian process with variance 1.0, and the disturbance noise is assumed to be zero-mean and uniform distributed over $[-\sqrt{3}, \sqrt{3}]$. Fig. 3 shows the steady-state EMSEs with different step-sizes and the noise variances, where the simulated EMSEs are computed as an average over 100 independent Monte Carlo simulations, and in each simulation, 50000 iterations were run to ensure the algorithm to reach the steady state, and the steady-state EMSE was obtained as an average over the last 1000 iterations. One can observe: i) the steady-state EMSEs are increasing with step-size and noise.
dent of (or impulsive disturbances). The noise processes substantially much larger variance to represent large outliers with smaller variance, and $B$ is an occurrence probability; whereas $A$ with zero-mean and unit variance; $d$) Uniform distribution over $\{−1, 1\}$ with probability mass $Pr\{x = −1\} = Pr\{x = 1\} = 0.5$; $c$) Laplace distribution with zero-mean and unit variance; $d$) Uniform distribution over $[−\sqrt{3}, \sqrt{3}]$. The convergence curves in terms of the weight error power $\|W_0 − W(i)\|^2$ averaged over 100 independent Monte Carlo runs are shown in Fig. 4. In the simulation, the step-sizes are chosen such that all the algorithms have almost the same initial convergence speed, and the parameter $\lambda$ in GMCC is experimentally chosen such that the algorithm achieves desirable performance. The final selected values of these parameters are reported in the figures. From simulation results we can observe: 1) the GMCC family algorithms are much more stable (robust) than the LMP family algorithms (In this example, when $p > 4$, the LMP will not converge); 2) the GMCC with $\alpha \neq 2$ may outperform significantly the original MCC ($\alpha = 2$) algorithm. In particular, the GMCC with $\alpha = 6$ achieves the best performance when $A(i)$ is of Binary or Uniform distribution. Additionally, in Fig. 5, we demonstrate the performance comparison among GMCC, least mean mixed-norm (LMMN) [9] and least mean M-estimate (LMM) [11] algorithms, where $A(i)$ is Binary distributed. The parameters in LMMN and LMM are selected by scanning for the best results. Evidently, the GMCC with a proper $\alpha$ value can outperform both LMMN and LMM significantly. To investigate how the parameter $\lambda$ affects the performance, we show in Fig. 6 the convergence curves of the GMCC with different $\lambda$, where $A(i)$ is Binary distributed and $\alpha = 6$. As one can see, the performance will become worse when $\lambda$ is too small or too large. In this example, the GMCC algorithm achieves the best performance when $\lambda$ is about 0.01. In order to further demonstrate the robustness of the GMCC against large outliers, we increase the variance of the outlier noise $B(i)$ from 15 to 100. In this case, the LMP family algorithms, except SA and LMS, will diverge, while the GMCC family algorithms can still work well. We present in Fig. 7 the simulation results when $A(i)$ is Uniform distributed.

C. Performance comparison with other algorithms

Third, we compare the performance of the GMCC and the LMP family algorithms with different $p$ values, namely SA ($p = 1$), LMS ($p = 2$), and the LMF ($p = 4$). The unknown system is assumed to be the same as that in the first simulation. In particular, we consider a noise model with form $v(i) = (1 − a(i))A(i) + a(i)B(i)$, where $a(i)$ is a binary independent and identically distributed process with $Pr\{a(i) = 1\} = c$, $Pr\{a(i) = 0\} = 1 − c$, and $0 ≤ c ≤ 1$ is an occurrence probability; whereas $A(i)$ is a noise process with smaller variance, and $B(i)$ is another noise process with substantially much larger variance to represent large outliers (or impulsive disturbances). The noise processes $A(i)$ and $B(i)$ are mutually independent and they are both independent of $a(i)$. In the simulation, $c$ is set at 0.06, and $B(i)$ is a white Gaussian process with zero-mean and variance 15. For the noise $A(i)$, we consider four distributions: $a$) Gaussian distribution with zero-mean and unit variance; $b$) Binary distribution over $\{-1, 1\}$ with probability mass $Pr\{x = −1\} = Pr\{x = 1\} = 0.5$; $c$) Laplace distribution with zero-mean and unit variance; $d$) Uniform distribution over $[−\sqrt{3}, \sqrt{3}]$. The convergence curves in terms of the weight error power $\|W_0 − W(i)\|^2$ averaged over 100 independent Monte Carlo runs are shown in Fig. 4. In the simulation, the step-sizes are chosen such that all the algorithms have almost the same initial convergence speed, and the parameter $\lambda$ in GMCC is experimentally chosen such that the algorithm achieves desirable performance. The final selected values of these parameters are reported in the figures. From simulation results we can observe: 1) the GMCC family algorithms are much more stable (robust) than the LMP family algorithms (In this example, when $p > 4$, the LMP will not converge); 2) the GMCC with $\alpha \neq 2$ may outperform significantly the original MCC ($\alpha = 2$) algorithm. In particular, the GMCC with $\alpha = 6$ achieves the best performance when $A(i)$ is of Binary or Uniform distribution. Additionally, in Fig. 5, we demonstrate the performance comparison among GMCC, least mean mixed-norm (LMMN) [9] and least mean M-estimate (LMM) [11] algorithms, where $A(i)$ is Binary distributed. The parameters in LMMN and LMM are selected by scanning for the best results. Evidently, the GMCC with a proper $\alpha$ value can outperform both LMMN and LMM significantly. To investigate how the parameter $\lambda$ affects the performance, we show in Fig. 6 the convergence curves of the GMCC with different $\lambda$, where $A(i)$ is Binary distributed and $\alpha = 6$. As one can see, the performance will become worse when $\lambda$ is too small or too large. In this example, the GMCC algorithm achieves the best performance when $\lambda$ is about 0.01. In order to further demonstrate the robustness of the GMCC against large outliers, we increase the variance of the outlier noise $B(i)$ from 15 to 100. In this case, the LMP family algorithms, except SA and LMS, will diverge, while the GMCC family algorithms can still work well. We present in Fig. 7 the simulation results when $A(i)$ is Uniform distributed.

D. Application to EEG denoising

Finally, we consider the denoising of an electroencephalograph (EEG) signal that contains interferences caused by the EOG signal. The task is to remove the interferences by subtracting the output of an adaptive filter from the original EEG signal. In our approach, the adaptive filter is trained by the GMCC algorithm, with the EOG signal being the input and the original EEG being the desired signal. The filter length is
Fig. 4. Convergence curves with different distributions of $A(i)$: (a) Gaussian; (b) Binary; (c) Laplace; (d) Uniform

set to ten. We use the power ratio ($10 \log \left( \frac{E[d_i^2]}{E(e_i^2)} \right)$) between the original and cleaned EEGs as a measure of performance. A larger power ratio means more interferences are removed from the original signal. For GMCC with different $\alpha$ values ($\alpha = 1, 2, 4$), the power ratios (after smoothing using a running window of 10000 samples) are illustrated in Fig. 8. In this figure, the GMCC with $\alpha = 1$ achieves the best performance. The results confirm again the fact that the GMCC with $\alpha \neq 2$ may outperform the original MCC (corresponding to the GMCC with $\alpha = 2$). The original EEG, EOG, filter output and cleaned EEG (all within a segment of 6000 samples) are shown in Fig. 9, in which the sample rate is 1000Hz and the unit is mV.

VI. CONCLUSION

Correntropy is a recently proposed similarity measure, and the maximum correntropy criterion (MCC) has been widely applied in domains of machine learning and signal processing. In previous studies, the kernel function in correntropy has been however limited to the Gaussian kernel. Gaussian kernel is desirable in many cases but obviously, it is not always the best choice. In this work, we proposed a generalized correntropy, using the generalized Gaussian density (GGD) function as the kernel. The new definition is very general and flexible, which includes the original correntropy with Gaussian kernel as a special case. Some important properties of the generalized correntropy were presented. The generalized maximum correntropy criterion (GMCC) was also proposed as an optimality criterion in estimation related problems. In particular, we applied the GMCC criterion to adaptive filtering. The optimal solution under GMCC and an adaptive algorithm, called the GMCC algorithm, were derived. Further, we investigated the stability and steady-state performance of the developed algorithm. A simple example was presented to
show that the GMCC algorithm is very stable and will have a zero probability of divergence (POD). A theoretical value of the steady-state excess mean square error (EMSE) was also derived. Theoretical predictions and excellent performance of the GMCC were confirmed by simulation results. Of course, similar to the original correntropy, the generalized correntropy can also be applied to other domains such as classification, principal components analysis (PCA), period estimation, spectral characterization and so on. This will be an interesting area of future study.

APPENDIX A
PROOF OF PROPERTY 6

Proof: When $0 < \alpha \leq 2$, we can construct a nonlinear mapping: $\Phi_{\alpha,\beta}(X) = \left[\varphi_{\alpha,\beta}(x_1), \cdots, \varphi_{\alpha,\beta}(x_N)\right]^T$, such that $\|\Phi_{\alpha,\beta}(X) - \Phi_{\alpha,\beta}(Y)\| = \sqrt{2N} \times GCIM(X,Y)$. In this case the GCIM function $GCIM(X,Y)$ defines a "Euclidean distance" in the Hilbert space $\mathcal{F}_k^N$, and hence is a metric in the sample vector space since it satisfies: i) $GCIM(X,Y) \geq 0$; ii) $GCIM(X,Y) = GCIM(Y,X)$; iii) $GCIM(X,Z) \leq GCIM(X,Y) + GCIM(Y,Z)$.

APPENDIX B
PROOF OF PROPERTY 7

Proof: As $\lambda \to 0+$ (or $x_i \to 0, i = 1, \cdots, N$), we have

$$L_{\alpha,\beta}(X) = \left(\frac{N}{\lambda^{\alpha}} \cdot \int_{\mathcal{F}_k} J_{GC-\alpha\beta}(X,0)\right)^{1/\alpha}$$

$$= \left[\frac{N}{\lambda^{\alpha}} \left(\sum_{i=1}^{N} G_{\alpha,\beta}(x_i)\right)\right]^{1/\alpha}$$

$$\approx \left[\frac{N}{\lambda^{\alpha}} \left(\sum_{i=1}^{N} \gamma_{\alpha,\beta} (1 - \lambda |x_i|^{\alpha})\right)\right]^{1/\alpha}$$

$$= \sum_{i=1}^{N} |x_i|^{\alpha} \right]^{1/\alpha}$$ (42)


**APPENDIX C**

**PROOF OF PROPERTY 8**

*Proof:* Let \( \mathbf{X}_0 \) be the solution obtained by minimizing \( \| \mathbf{X} \|_0 \) over \( \Omega \) and \( \mathbf{X}_i \) the solution achieved by minimizing \( L_{\alpha, \beta}(\mathbf{X}) \).

Then \( L_{\alpha, \beta}(\mathbf{X}_i) \leq L_{\alpha, \beta}(\mathbf{X}_0) \), and hence

\[
\sum_{i=1}^{N} g_{\alpha, \beta}((\mathbf{X}_i)_i) \geq \sum_{i=1}^{N} g_{\alpha, \beta}((\mathbf{X}_0)_i)
\]

(43)

where \((\mathbf{X}_i)_i\) denotes the \(i\)th component of \(\mathbf{X}_i\). It follows that

\[
(N - \| \mathbf{X}_i \|_0) + \sum_{i=1,(\mathbf{X}_i)_i \neq 0}^{N} \exp \left( -\lambda \| (\mathbf{X}_i)_i \|_0 \right)^\alpha
\]

\[
\geq (N - \| \mathbf{X}_0 \|_0) + \sum_{i=1,(\mathbf{X}_0)_i \neq 0}^{N} \exp \left( -\lambda \| (\mathbf{X}_0)_i \|_0 \right)^\alpha
\]

Thus we have

\[
\| \mathbf{X}_i \|_0 - \| \mathbf{X}_0 \|_0 \leq \sum_{i=1,(\mathbf{X}_i)_i \neq 0}^{N} \exp \left( -\lambda \| (\mathbf{X}_i)_i \|_0 \right)^\alpha
\]

(45)

Since \( |x_i| > \delta \), \( \forall i : x_i \neq 0 \), as \( \lambda \to \infty \) the right hand side of (45) will approach zero. Therefore, if \( \lambda \) is large enough, it holds that

\[
\| \mathbf{X}_i \|_0 \leq \| \mathbf{X}_0 \|_0 + \varepsilon
\]

(46)

where \( \varepsilon \) is a small positive number arbitrarily close to zero.

**APPENDIX D**

**PROOF OF PROPERTY 9**

*Proof:* The Hessian (if exists) of \( \hat{J}_{GC-loss} \) with respect to \( \mathbf{e} \) is

\[
H_{\hat{J}_{GC-loss}}(\mathbf{e}) = -\frac{\alpha \lambda^{-1} \alpha \beta}{N} \cdot \left( T(e_1) \alpha \lambda \| e_1 \|_0 ^\alpha - (\alpha - 1) \right) \cdots 0 \cdots \left( T(e_N) \alpha \lambda \| e_N \|_0 ^\alpha - (\alpha - 1) \right)
\]

(47)

where \( T(x) = \exp \left( -\lambda \| x \|_0 \right) \| x \|_0 ^\alpha - 2. \) From (47) one can see:

i) if \( 0 < \alpha \leq 1 \), then \( H_{\hat{J}_{GC-loss}}(\mathbf{e}) \leq 0 \) for any \( \mathbf{e} \) with \( e_i \neq 0 \) (\( i = 1, \ldots, N \));

ii) if \( \alpha > 1 \), then \( H_{\hat{J}_{GC-loss}}(\mathbf{e}) \geq 0 \) for any \( \mathbf{e} \) with \( 0 < \| e_i \|_0 \leq \| (\alpha - 1)/\alpha \|^{1/\alpha} (\alpha = 1, \ldots, N) \);

iii) if \( \alpha \to 0^+ \), then for any \( \mathbf{e} \) with \( e_i = 0 \) (\( i = 1, \ldots, N \)), we have \( H_{\hat{J}_{GC-loss}}(\mathbf{e}) \leq 0 \) for \( 0 < \alpha \leq 1 \), and \( H_{\hat{J}_{GC-loss}}(\mathbf{e}) \geq 0 \) for \( \alpha > 1 \).

**APPENDIX E**

**PROOF OF PROPERTY 10**

*Proof:* A differentiable function \( f : S \to \mathbb{R} \) (\( S \subset \mathbb{R}^N \)) is said to be convex, if and only if [31]

\[
f(x_2) \geq f(x_1) + q(x_1, x_2)^T \nabla f(x_1)
\]

(48)

where \( \nabla f(x) \) denotes the gradient of \( f \) with respect to \( x \), and \( q(x_1, x_2) \) is some vector valued function. For \( \alpha > 1 \), the GC-loss \( \hat{J}_{GC-loss} \) is a differentiable function of \( \mathbf{e} \), and the gradient \( \nabla \hat{J}_{GC-loss}(\mathbf{e}) \) is

\[
\nabla \hat{J}_{GC-loss}(\mathbf{e}) = \frac{\lambda \alpha^{-1} \alpha \beta}{N} \left( \exp(-\lambda \| e_1 \|_0) \| e_1 \|_0 ^\alpha - \| e_1 \|_0 \right) \cdots \left( \exp(-\lambda \| e_N \|_0) \| e_N \|_0 ^\alpha - \| e_N \|_0 \right)^T
\]

(49)

where \( \text{sign}(\cdot) \) is the sign function. Since \( e_i \leq M \), we have \( \nabla \hat{J}_{GC-loss}(\mathbf{e}) = 0 \) if and only if \( \mathbf{e} = 0 \). On the other hand, we have \( \hat{J}_{GC-loss}(\mathbf{e}) \geq \hat{J}_{GC-loss}(0) = 0 \). So we can construct the following function

\[
q(e_1, e_2) = \left\{ \begin{array}{ll}
\nabla \hat{J}_{GC-loss}(\mathbf{e})^T \nabla \hat{J}_{GC-loss}(\mathbf{e}) & \text{if } e \neq 0 \\
0 & \text{if } e = 0
\end{array} \right.
\]

(50)

such that it holds

\[
\hat{J}_{GC-loss}(\mathbf{e}_2) \geq \hat{J}_{GC-loss}(\mathbf{e}_1) + q(e_1, e_2)^T \nabla \hat{J}_{GC-loss}(\mathbf{e}_1)
\]

(51)

**APPENDIX F**

**PROOF OF THEOREM 2**

*Proof:* First, we derive

\[
E \left[ g_{\alpha, \beta}(\mathbf{e}(i)) \right] = \int_{\mathbb{R}} g_{\alpha, \beta}(\mathbf{e}) p_{\mathbf{e}(i)}(\mathbf{e}) d\mathbf{e}
\]

\[
= \int_{\mathbb{R}} p_{\mathbf{e}(i)}(\mathbf{e}) \left( \int_{0}^{\gamma_{\beta}} \frac{1}{\xi} \left( \xi \leq g_{\alpha, \beta}(\mathbf{e}) \right) d\xi \right) d\mathbf{e}
\]

\[
= \int_{0}^{\gamma_{\beta}} \left( \int_{-\infty}^{\mathbf{e}(i)} p_{\mathbf{e}(i)}(\mathbf{e}) d\mathbf{e} \right) d\xi
\]

\[
= \int_{0}^{\gamma_{\beta}} \left( \int_{-\mathbf{e}(i)}^{\mathbf{e}(i)} p_{\mathbf{e}(i)}(\mathbf{e}) d\mathbf{e} \right) d\xi
\]

(52)

where \( \mathbb{I}(\cdot) \) denotes an indicator function, and \( \varepsilon \) is a certain positive number satisfying \( g_{\alpha, \beta}(\mathbf{e}) = g_{\alpha, \beta}(-\mathbf{e}) = \varepsilon \). Since \( x(i) \) and \( d(i) \) are both zero-mean Gaussian processes, then error \( e(i) \) is also a zero-mean Gaussian process, with PDF

\[
p_{\mathbf{e}(i)}(\mathbf{e}) = \frac{1}{\sqrt{2\pi \sigma^2_{\mathbf{e}(i)}}} \exp \left( -\frac{\varepsilon^2}{2\sigma^2_{\mathbf{e}(i)}} \right)
\]

(53)

where \( \sigma^2_{\mathbf{e}(i)} = E[e^2(i)] \) is the error variance. Hence

\[
\int_{-\varepsilon}^{\varepsilon} p_{\mathbf{e}(i)}(\mathbf{e}) d\mathbf{e} = \varepsilon f \left( \frac{\varepsilon}{\sqrt{2\sigma^2_{\mathbf{e}(i)}}} \right)
\]

(54)

with \( \varepsilon f(x) = \frac{1}{\sqrt{\pi}} \int_{-\varepsilon}^{\varepsilon} \exp(-t^2)dt \) being the error function, which is a monotonically increasing function of \( x \). Therefore, \( \int_{-\varepsilon}^{\varepsilon} p_{\mathbf{e}(i)}(\mathbf{e}) d\mathbf{e} \) is a monotonically decreasing function of \( \sigma^2_{\mathbf{e}(i)} \). It follows that

\[
\max E \left[ g_{\alpha, \beta}(\mathbf{e}(i)) \right] \leq \max_{-\varepsilon}^{\varepsilon} p_{\mathbf{e}(i)}(\mathbf{e}) d\mathbf{e} \leq \min \sigma^2_{\mathbf{e}(i)}
\]

(55)

That is, the maximization of the generalized correntropy \( E \left[ g_{\alpha, \beta}(\mathbf{e}(i)) \right] \) will be equivalent to the minimization of the mean square error \( \sigma^2_{\mathbf{e}(i)} = E[e^2(i)] \). In this case, the optimal solution under GMCC will be equal to the Wiener solution.
REFERENCES