Information Theoretic Interpretation of Error Criteria

CHEN Ba-Dong[1] HU Jin-Chun[2] ZHU Yu[1] SUN Zeng-Qi[2]

Abstract Error criteria (or error cost functions) play significant roles in statistical estimation problems. In this paper, we study error criteria from the viewpoint of information theory. The relationships between error criteria and error′s entropy criterion are investigated. It is shown that an error criterion is equivalent to the error′s entropy criterion plus a Kullback-Leibler information divergence (KL-divergence). Based on this result, two important properties of the error criteria are proved. Particularly, the optimum error criterion can be interpreted via the meanings of entropy and KL-divergence. Furthermore, a novel approach is proposed for the choice of p-power error criteria, in which a KL-divergence based cost is minimized. The proposed method is verified by Monte Carlo simulation experiments.

Key words Estimation, error criteria, entropy, Kullback-Leibler information divergence (KL-divergence), adaptive filtering

Let $k \in \mathbf{N}$, (Ω, B, P) be a probability space, X be an integrable random variable defined on (Ω, B, P), $\boldsymbol{Y}$ be a random vector defined on (Ω, B, P) taking values in $\mathbf{R}^k$. $\mathbf{G}$ denotes the collection of all Borel measurable functions with respect to the σ-field $\sigma(\boldsymbol{Y})$ generated by $\boldsymbol{Y}$. A commonly encountered problem involves estimating X via $g(\boldsymbol{Y})$, $g \in \mathbf{G}$, so as to minimize a certain error criterion (or error cost)[1−4]:

$$\mathrm{E}\left[\phi\left(X - g\left(\mathbf{Y}\right)\right)\right] = \mathrm{E}_e\left[\phi\left(e\right)\right] = \int_{\mathbf{R}} \phi\left(e\right) p_e\left(e\right) \mathrm{d}e \tag{1}$$

where E denotes the expectation operator, $e = X - g\left(\boldsymbol{Y}\right)$ is the estimation error, ϕ denotes the Borel measurable cost function, and $p_e\left(e\right)$ denotes the probability density function (PDF) of e. We refer to g as an estimator of X based on $\boldsymbol{Y}$. Under error criterion $\mathrm{E}_e\left[\phi\left(e\right)\right]$, the optimum estimator, denoted by g^*, is

$$g^* = \arg\min_{g \in \mathbf{G}} \mathrm{E}_e\left[\phi\left(e\right)\right] = \arg\min_{g \in \mathbf{G}} \mathrm{E}\left[\phi\left(X - g\left(\mathbf{Y}\right)\right)\right] \tag{2}$$

The optimization problem can also reduce to a parameter search procedure in which a suitable structure of the estimator is assumed. Let $g_{\boldsymbol{W}}$ ($\boldsymbol{W} \in \boldsymbol{\Theta} \subset \mathbf{R}^d$) be a parameterization of g, in which $\boldsymbol{W}$ is the d-dimensional parameter vector and $\boldsymbol{\Theta}$ is the parameter space. Then, the optimum parameter $\boldsymbol{W}^*$ is

$$\begin{aligned} \boldsymbol{W}^* = &\arg\min_{\boldsymbol{W} \in \boldsymbol{\Theta}} \mathrm{E}_e\left[\phi\left(e\right)\right] = \\ &\arg\min_{\boldsymbol{W} \in \boldsymbol{\Theta}} \mathrm{E}_e\left[\phi\left(X - g_{\boldsymbol{W}}\left(\boldsymbol{Y}\right)\right)\right] \end{aligned} \tag{3}$$

In most practical applications, the optimum parameter can be solved by the stochastic gradient (SG) algorithm[3, 5−7]:

$$\begin{aligned} \boldsymbol{W}\left(k+1\right) = &\boldsymbol{W}\left(k\right) - \eta \left.\frac{\partial \phi\left(e\right)}{\partial \boldsymbol{W}}\right|_{\boldsymbol{W}=\boldsymbol{W}(k)} = \\ &\boldsymbol{W}\left(k\right) - \eta \left.\frac{\partial \phi\left(e\right)}{\partial e} \times \frac{\partial e}{\partial \boldsymbol{W}}\right|_{\boldsymbol{W}=\boldsymbol{W}(k)} \end{aligned} \tag{4}$$

where $\boldsymbol{W}(k)$ denotes the parameter vector at iteration k, and $\eta > 0$ is the step-size (or adaptation gain).

Received April 7, 2008; in revised form December 20, 2008
Supported by National Basic Research Program (973 Program) of China (2007CB724205), National Natural Science Foundation of China (60604010), and China Postdoctoral Science Foundation Funded Project (20080440384)
1. Institute of Manufacturing Engineering, Department of Precision Instruments and Mechanology, Tsinghua University, Beijing 100084, P. R. China 2. State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P. R. China
DOI: 10.3724/SP.J.1004.2009.01302

In the above estimation problem, the error criterion $\mathrm{E}_e\left[\phi\left(e\right)\right]$ (or the cost function $\phi\left(e\right)$) plays a central role[2−3, 5]. Among various error criteria, the mean-square error (MSE) ($\phi\left(e\right) = e^2$) is the most popular one due to its mathematical tractability[3, 7−8]. With the basic finite impulse response (FIR) filter structure, MSE yields a simple optimization problem, whose analytical solution is provided by the Wiener-Hopf equation[3, 7]. However, MSE is not always the optimum error criterion, especially for the non-linear or non-Gaussian situations[1−2, 8−9]. Thus many non-MSE criteria have been studied. In an early work, Sherman[1] showed that in the case of Gaussian processes, a large family of non-MSE criteria yields the same predictor as the linear minimum mean-square predictor of Wiener. Later, Sherman′s results and several extensions were revisited by Brown[10], Zakai[11], and Hall[2], et al. In order to take into account higher-order statistics, Walach proposed the mean fourth error (MFE) criterion, and derived the least mean fourth (LMF) algorithm[12], while Pei investigated the least mean p-power (LMP) criterion[6]. Further, the fractional lower order moments (FLOM) of the error have also been used in adaptive filtering in the presence of impulse alpha-stable noises[13].

Besides the error criteria, the entropy, which is a central notion in information theory[14], has also been used as a cost function in estimation problems. As the entropy measures the average uncertainty contained in a random variable, its minimization forces the error to gather. Weidemann and Stear studied the parameter estimation problem for non-linear and non-Gaussian discrete-time systems by using the error entropy as a criterion functional, and it was shown that the reduced error entropy is upper bounded by the amount of information obtained by observation[15−16]. Minamide[17] extended Weidemann′s results to a continuous-time estimation model. Recently, the minimum error entropy (MEE) criterion has been used by Erdogmus[18−21], Principe[22], Chen[23] and Han[24] et al. in the areas of supervised learning, and the stochastic information gradient (SIG) algorithms have been developed. Under MEE criterion, the optimum estimator is given by[15, 25−26]

$$\begin{aligned} g^{\#} = &\arg\min_{g \in \mathbf{G}} H\left(e\right) = \\ &\arg\min_{g \in \mathbf{G}} \left\{ -\int_{\mathbf{R}} p_e\left(e\right) \log p_e\left(e\right) \mathrm{d}e \right\} \end{aligned} \tag{5}$$

The error entropy is a functional of the error′s distribution and is related to various statistical behaviors of

it. Numerical examples indicated that compared with the MSE criterion, the MEE criterion could be able to achieve a better performance in the adaptive system training[19−21].

Both error criteria and error entropy criterion measure the concentration of the error, and quantify how similar the two random variables are. There are close relationships between the two criteria. The goal of this paper is to study the error criteria from the viewpoint of information theory, and investigate the connections between the error criterion (MSE is the special case) and the MEE criterion. The paper is organized as follows. In the next section, we show that the error criterion is equivalent to the error′s entropy plus a Kullback-Leibler information divergence (KL-divergence), that is,

$$\mathrm{E}_e\left[\phi\left(e\right)\right] \propto \left\{H\left(e\right) + D_{KL}\left(p_e\left(e\right) \| q_\phi\left(e\right)\right)\right\} \tag{6}$$

where $q_\phi(e)$ is the worst case density related to the error cost $\phi(e)$. Then, in Section 2, based on (6), we give the information theoretic proofs of two important properties of the error criteria. Further, in Section 3, we interpret the optimum error criterion by the meanings of entropy and KL-divergence, and in Section 4, we propose an information theoretic approach to the choice of error criterion. For the family of p-power error criteria, the optimum p value is determined by

$$p_{\mathrm{opt}} = \arg\min_{p\in\mathbf{R}_+}\left\{\min_{\tau\in\mathbf{R}_+} D_{KL}\left(p_n\left(x\right) \| \exp\left(-\gamma_0 - \gamma_1\left|x\right|^p\right)\right)\right\} \tag{7}$$

where $p_n(\cdot)$ is the PDF of the interfering noise. Simulation experiments are presented to verify the proposed method. Finally, in Section 5, we give the concluding remarks.

1 Relationship between error criteria and error entropy criterion

Before proceeding, we give the definition of equivalence between two error criteria.

Definition 1. Two error criteria $\mathrm{E}_e[\phi_1(e)]$ and $\mathrm{E}_e[\phi_2(e)]$ are said to be equivalent, if and only if $\mathrm{E}_e[\phi_1(e)] = \gamma_0 + \gamma_1 \mathrm{E}_e[\phi_2(e)]$ (or equivalently, $\phi_1(e) = \gamma_0 + \gamma_1\phi_2(e)$), where $-\infty < \gamma_0 < \infty$ and $0 < \gamma_1 < \infty$.

Remark 1. If $\mathrm{E}_e[\phi_1(e)]$ and $\mathrm{E}_e[\phi_2(e)]$ are equivalent, we denote $\mathrm{E}_e[\phi_1(e)] \propto \mathrm{E}_e[\phi_2(e)]$ or $\phi_1(e) \propto \phi_2(e)$. Clearly, if $\mathrm{E}_e[\phi_1(e)] \propto \mathrm{E}_e[\phi_2(e)]$, we have $g_1^* = g_2^*$, where $g_1^* = \arg\min_{g\in\mathbf{G}} \mathrm{E}_e[\phi_1(e)]$ and $g_2^* = \arg\min_{g\in\mathbf{G}} \mathrm{E}_e[\phi_2(e)]$. Thus, $\mathrm{E}_e[\phi_1(e)]$ and $\mathrm{E}_e[\phi_2(e)]$ yield the same optimum solutions (or reflect the same fidelity demand). Similarly, for the parameterization situation, we have $\boldsymbol{W}_1^* = \boldsymbol{W}_2^*$, where $\boldsymbol{W}_1^* = \arg\min_{\boldsymbol{W}\in\Theta} \mathbf{E}_e[\phi_1(e)]$ and $\boldsymbol{W}_2^* = \arg\min_{\boldsymbol{W}\in\Theta} \mathrm{E}_e[\phi_2(e)]$. In this case, by choosing suitable step-sizes, $\mathrm{E}_e[\phi_1(e)]$ and $\mathrm{E}_e[\phi_2(e)]$ yield the same stochastic gradient algorithm. In fact, since $\phi_1(e) = \gamma_0 + \gamma_1\phi_2(e)$, if we choose the step-size $\eta_2 = \lambda_1\eta_1$, we have

$$\begin{aligned}\boldsymbol{W}\left(k+1\right) =& \boldsymbol{W}\left(k\right) - \eta_1 \left.\frac{\partial\phi_1\left(e\right)}{\partial\boldsymbol{W}}\right|_{\boldsymbol{W}=\boldsymbol{W}(k)} = \\ & \boldsymbol{W}\left(k\right) - \eta_2 \left.\frac{\partial\phi_2\left(e\right)}{\partial\boldsymbol{W}}\right|_{\boldsymbol{W}=\boldsymbol{W}(k)}\end{aligned} \tag{8}$$

Here, both gradient algorithms are identical and have the same performance characteristics (stability, adaptation speed, excess MSE, etc.).

The following theorem relates the error criterion with a certain probability density function.

Theorem 1. Given any error criterion $\mathrm{E}_e[\phi(e)]$ (or the cost function $\phi(e)$), there exists a probability density function $q_\phi(e)$, such that $q_\phi(e) = \exp[-\gamma_0 - \gamma_1\phi(e)]$, where γ_0 and γ_1 are determined by

$$\begin{cases} \exp\left(\gamma_0\right) = \displaystyle\int_{\mathbf{R}} \exp\left[-\gamma_1\phi\left(e\right)\right]\mathrm{d}e \\ \mathrm{E}_e\left[\phi\left(e\right)\right]\exp\left(\gamma_0\right) = \displaystyle\int_{\mathbf{R}} \phi\left(e\right)\exp\left[-\gamma_1\phi\left(e\right)\right]\mathrm{d}e \end{cases} \tag{9}$$

Proof. In fact, $q_\phi(e)$ is just the worst case density function according to Jaynes′ maximum entropy principle (MEP)[27−29]. Mathematically, the problem is to pick up a probability density $q_\phi(e)$, which maximizes Shannon′s entropy $H(q_\phi) = -\int_{\mathbf{R}} q_\phi(e)\log q_\phi(e)\,\mathrm{d}e$ subject to the constraints

$$\begin{cases} \int_{\mathbf{R}} q_\phi\left(e\right)\mathrm{d}e = 1 \\ \int_{\mathbf{R}} q_\phi\left(e\right)\phi\left(e\right)\mathrm{d}e = \int_{\mathbf{R}} p_e\left(e\right)\phi\left(e\right)\mathrm{d}e = \mathrm{E}_e\left[\phi\left(e\right)\right] \end{cases} \tag{10}$$

We create an unconstrained expression for the entropy using Lagrange multipliers:

$$\begin{aligned} L = & -\int_{\mathbf{R}} q_\phi\left(e\right)\log q_\phi\left(e\right)\mathrm{d}e - \left(\gamma_0 - 1\right)\left\{\int_{\mathbf{R}} q_\phi\left(e\right)\mathrm{d}e - 1\right\} - \\ & \gamma_1\left\{\int_{\mathbf{R}} q_\phi\left(e\right)\phi\left(e\right)\mathrm{d}e - \mathrm{E}_e\left[\phi\left(e\right)\right]\right\} \end{aligned} \tag{11}$$

Using calculus of variations[30], we maximize L with respect to $q_\phi(e)$:

$$\begin{aligned} & \frac{\partial}{\partial q_\phi}\left[-q_\phi\log q_\phi - \left(\gamma_0 - 1\right)q_\phi - \gamma_1 q_\phi\phi\right] = 0 \Rightarrow \\ & \qquad -\log q_\phi - 1 - \left(\gamma_0 - 1\right) - \gamma_1\phi = 0 \end{aligned}$$

Thus, the worst case density is

$$q_\phi\left(e\right) = \exp\left[-\gamma_0 - \gamma_1\phi\left(e\right)\right] \tag{12}$$

where γ_0 and γ_1 are determined by

$$\begin{cases} \displaystyle\int_{\mathbf{R}} \exp\left[-\gamma_0 - \gamma_1\phi\left(e\right)\right]\mathrm{d}e = 1 \\ \displaystyle\int_{\mathbf{R}} \exp\left[-\gamma_0 - \gamma_1\phi\left(e\right)\right]\phi\left(e\right)\mathrm{d}e = \mathrm{E}_e\left[\phi\left(e\right)\right] \end{cases} \Leftrightarrow$$

$$\begin{cases} \exp\left(\gamma_0\right) = \displaystyle\int_{\mathbf{R}} \exp\left[-\gamma_1\phi\left(e\right)\right]\mathrm{d}e \\ \mathrm{E}_e\left[\phi\left(e\right)\right]\exp\left(\gamma_0\right) = \displaystyle\int_{\mathbf{R}} \phi\left(e\right)\exp\left[-\gamma_1\phi\left(e\right)\right]\mathrm{d}e \end{cases}$$

□

Example 1. Consider the LMP error criterion[6]

$$J = \mathrm{E}_e\left(\left|e\right|^p\right), \quad p > 0 \tag{13}$$

where the cost function $\phi(e) = |e|^p$. When $p = 2$, this criterion reduces to the well-known MSE criterion. By Theorem 1, we have

$$q_\phi\left(e\right) = \exp\left[-\gamma_0 - \gamma_1\left|e\right|^p\right] \tag{14}$$

where γ_0 and γ_1 are determined by

$$\begin{cases} \exp\left\{\gamma_0\right\} = \displaystyle\int_{-\infty}^{+\infty} \left(\exp\left\{-\gamma_1\left|e\right|^p\right\}\right)\mathrm{d}e \\ \mathrm{E}_e\left(\left|e\right|^p\right)\exp\left\{\gamma_0\right\} = \displaystyle\int_{-\infty}^{+\infty} \left|e\right|^p\left(\exp\left\{-\gamma_1\left|e\right|^p\right\}\right)\mathrm{d}e \end{cases} \tag{15}$$

It is easy to derive γ_0 and γ_1 as follows:

$$\begin{cases} \gamma_0 = \log\left\{\dfrac{2}{p}\Gamma\left(\dfrac{1}{p}\right)\right\} - \dfrac{1}{p}\log\left\{\dfrac{\Gamma\left(\dfrac{p+1}{p}\right)}{\mathrm{E}_e\left(|e|^p\right)\times\Gamma\left(\dfrac{1}{p}\right)}\right\} \\ \gamma_1 = \dfrac{\Gamma\left(\dfrac{p+1}{p}\right)}{\mathrm{E}_e\left(|e|^p\right)\times\Gamma\left(\dfrac{1}{p}\right)} \end{cases} \tag{16}$$

where $\Gamma(\cdot)$ represents the gamma function defined as

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x \tag{17}$$

Given probability density function $p_e(e)$, we can calculate $\mathrm{E}_e\left(|e|^p\right)$, and get the exact values of γ_0 and γ_1. For different p values, the cost functions $\phi(e) = |e|^p$ are shown in Fig. 1, and the corresponding worst case densities $q_\phi(e)$ are depicted in Fig. 2 (assuming $p_e(e) \sim \mathrm{N}(0,1)$).

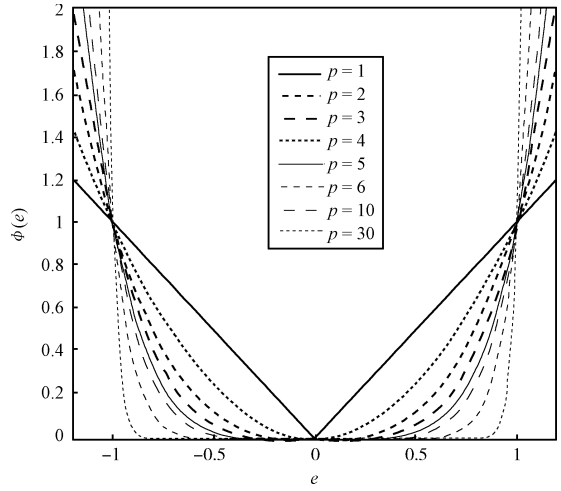

Fig. 1 Cost functions of p-power error criteria with different p values

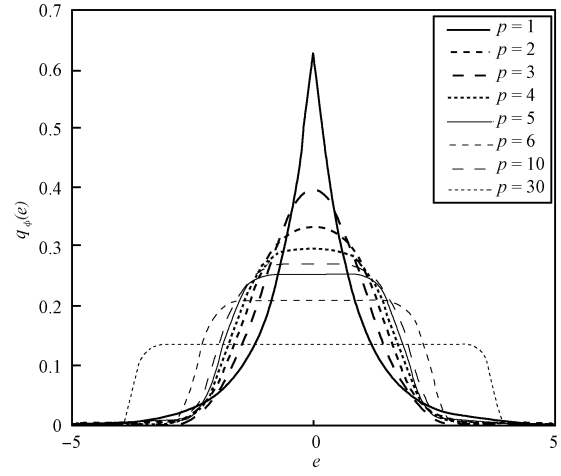

Fig. 2 Worst case densities $q_\phi(e)$ for different p values

Remark 2. It is worth noting that the worst case density $q_\phi(e) = \exp\left[-\gamma_0 - \gamma_1|e|^p\right]$ is actually the generalized Gaussian density (GGD)[31−32], in which p is called the shape parameter. The GGD model includes Laplace ($p = 1$) and Gaussian ($p = 2$) distributions as special cases, and can be used to approximate a large number of distributions in the areas of image coding, speech recognition, blind source separation (BSS), and so on.

Theorem 2. Any cost function $\phi(e)$, which satisfies $\lim_{|e|\to+\infty}\phi(e) = +\infty$, is equivalent to minus logarithm of a certain density function $q_\phi(e)$, i.e., $\phi(e) \propto -\log\left[q_\phi(e)\right]$.

Proof. By Theorem 1, for any cost function $\phi(e)$, there exists a PDF $q_\phi(e)$, such that

$$q_\phi(e) = \exp\left[-\gamma_0 - \gamma_1\phi(e)\right]$$

The minus logarithm of $q_\phi(e)$ is

$$-\log\left[q_\phi(e)\right] = \gamma_0 + \gamma_1\phi(e) \tag{18}$$

As $q_\phi(e) \geq 0$, $\int_{\mathbf{R}} q_\phi(e)\,\mathrm{d}e = 1$, we have $\lim_{|e|\to+\infty} q_\phi(e) = 0$, and it follows that

$$\begin{aligned} &\lim_{|e|\to+\infty} q_\phi(e) = 0 \Rightarrow \\ &\lim_{|e|\to+\infty}\left[\log q_\phi(e)\right] = -\infty \Rightarrow \\ &\lim_{|e|\to+\infty}\left[-\log q_\phi(e)\right] = +\infty \Rightarrow \\ &\lim_{|e|\to+\infty}\left[\gamma_0 + \gamma_1\phi(e)\right] = +\infty \Rightarrow \\ &\gamma_1 \lim_{|e|\to+\infty}\left[\phi(e)\right] = +\infty \overset{(a)}{\Rightarrow} \gamma_1 > 0 \end{aligned} \tag{19}$$

where (a) follows from the condition $\lim_{|e|\to+\infty}\phi(e) = +\infty$. By Definition 1, we have $\phi(e) \propto -\log\left[q_\phi(e)\right]$. □

Remark 3. Note that the condition $\lim_{|e|\to+\infty}\phi(e) = +\infty$ is not restrictive, because for most error criteria, such as the p-power error criterion ($\phi(e) = |e|^p$, $p > 0$) and the risk sensitivity criterion[33] ($\phi(e) = \exp(\lambda|e|^p)$, $\lambda > 0$, $p > 0$), the cost functions $\phi(e)$ increase rapidly when $|e|$ goes to infinity.

Now we arrive at the main theorem of this section.

Theorem 3. Given any error criterion $\mathrm{E}_e\left[\phi(e)\right]$, which satisfies $\lim_{|e|\to+\infty}\phi(e) = +\infty$, we have $\mathrm{E}_e\left[\phi(e)\right] \propto \{H(e) + D_{KL}\left(p_e(e)\,\|q_\phi(e)\right)\}$, where $q_\phi(e)$ is determined by Theorem 1, $H(e)$ and $D_{KL}\left(p_e(e)\,\|q_\phi(e)\right)$ are the entropy and Kullback-Leibler information divergence, respectively, i.e.,

$$\begin{cases} H(e) = -\displaystyle\int_{\mathbf{R}} p_e(e)\log p_e(e)\,\mathrm{d}e \\ D_{KL}\left(p_e(e)\,\|q_\phi(e)\right) = \displaystyle\int_{\mathbf{R}} p_e(e)\log\left(\frac{p_e(e)}{q_\phi(e)}\right)\mathrm{d}e \end{cases} \tag{20}$$

Proof. By Theorem 2, we have

$$\begin{aligned} &\mathrm{E}_e\left[\phi(e)\right] \propto \mathrm{E}\left[-\log q_\phi(e)\right] = \\ &\int p_e(e)\left[-\log q_\phi(e)\right]\mathrm{d}e = \\ &\int p_e(e)\left[-\log q_\phi(e)\right]\mathrm{d}e + \int p_e(e)\log p_e(e)\,\mathrm{d}e - \\ &\int p_e(e)\log p_e(e)\,\mathrm{d}e = \\ &\int p_e(e)\log\left(\frac{p_e(e)}{q_\phi(e)}\right)\mathrm{d}e - \int p_e(e)\log p_e(e)\,\mathrm{d}e = \\ &D_{KL}\left(p_e(e)\,\|q_\phi(e)\right) + H(e) \end{aligned}$$

□

Remark 4. Theorem 3 provides an interesting result. The error criterion $\mathrm{E}_e\left[\phi(e)\right]$ is equivalent to the error′s entropy $H(e)$ plus KL-divergence $D_{KL}\left(p_e(e)\,\|q_\phi(e)\right)$. Since

$D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e\right)\right)$ is always nonnegative, we have

$$D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e\right)\right)+H\left(e\right)\geq H\left(e\right) \tag{21}$$

with equality if and only if $p_e\left(e\right)=q_\phi\left(e\right)$. Thus, minimization of error criterion $\mathrm{E}_e\left[\phi\left(e\right)\right]$ is equivalent to minimization of an upper bound of the error entropy $H\left(e\right)$.

2 Information theoretic proofs of the properties of error criteria

In this section, we use Theorem 3 to prove two important properties of the error criteria. Although both properties can be proved by some other methods, here we give the information theoretic proofs. In the rest of the paper, we suppose the error cost function $\phi\left(e\right)$ always satisfies $\lim_{|e|\to+\infty}\phi\left(e\right)=+\infty$.

Before proceeding, we give the following lemma.

Lemma 1[14]. The entropy and KL-divergence is shift invariant, that is, for any random variables X and Y, and any constant $\mathrm{c}\in\mathbf{R}$, we have $H\left(X+\mathrm{c}\right)=H\left(X\right)$, $D_{KL}\left(X\|Y\right)=D_{KL}\left(X+\mathrm{c}\|Y+\mathrm{c}\right)$.

Property 1. If g^* minimizes the error criterion $\mathrm{E}_e\left[\phi\left(e\right)\right]$ over $\mathbf{G}$, i.e., $g^*=\arg\min_{g\in\mathbf{G}}\mathrm{E}_e\left[\phi\left(e\right)\right]$, then $\tilde{g}^*=g^*+\mathrm{c}$ minimizes $\mathrm{E}_e\left[\phi\left(e+\mathrm{c}\right)\right]$, where $\mathrm{c}\in\mathbf{R}$ is any constant.

Proof. By Theorem 3, we have

$$\begin{cases}\mathrm{E}_e\left[\phi\left(e\right)\right]\propto\left\{D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e\right)\right)+H\left(e\right)\right\}\\ \mathrm{E}_e\left[\phi\left(e+\mathrm{c}\right)\right]\propto\left\{D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e+\mathrm{c}\right)\right)+H\left(e\right)\right\}\end{cases}$$

It follows that

$$\begin{cases}\arg\min\limits_{g\in\mathbf{G}}\mathrm{E}_e\left[\phi\left(e\right)\right]=\\ \arg\min\limits_{g\in\mathbf{G}}\left\{D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e\right)\right)+H\left(e\right)\right\}\\ \arg\min\limits_{g\in\mathbf{G}}\mathrm{E}_e\left[\phi\left(e+\mathrm{c}\right)\right]=\\ \arg\min\limits_{g\in\mathbf{G}}\left\{D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e+\mathrm{c}\right)\right)+H\left(e\right)\right\}\end{cases}$$

Hence,

$$\begin{aligned}&\tilde{g}^*=\arg\min_{\beta\in\mathbf{G}}\left\{\mathrm{E}_e\left[\phi\left(e+\mathrm{c}\right)\right]\big|_{g=\beta}\right\}=\\ &\arg\min_{\beta\in\mathbf{G}}\left\{\begin{array}{l}D_{KL}\left(p_e\left(e\,|g=\beta\right)\|q_\phi\left(e+\mathrm{c}\right)\right)+\\ H\left(e\,|g=\beta\right)\end{array}\right\}\overset{(a)}{=}\\ &\arg\min_{\beta\in\mathbf{G}}\left\{\begin{array}{l}D_{KL}\left(p_e\left(e\,|g=\beta-\mathrm{c}\right)\|q_\phi\left(e\right)\right)+\\ H\left(e\,|g=\beta-\mathrm{c}\right)\end{array}\right\}=\\ &\arg\min_{\beta\in\mathbf{G}}\left\{\mathrm{E}_e\left[\phi\left(e\right)\right]\big|_{g=\beta-\mathrm{c}}\right\}=\\ &\arg\min_{\beta\in\mathbf{G}}\left\{\mathrm{E}_e\left[\phi\left(e\right)\right]\big|_{g=\beta}\right\}+\mathrm{c}=g^*+\mathrm{c}\end{aligned}$$

where (a) follows from the shift-invariance of entropy and KL -divergence (as stated in Lemma 1). □

Property 2. Let $m\in\mathbf{G}$, which equals a.e. $[\mu_{\boldsymbol{Y}}]$ to $\mathrm{E}\left[X\,|\boldsymbol{Y}=\boldsymbol{y}\right]$, and assume the conditional density function $p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)$ is symmetric around zero. Then, $g=m$ minimizes the error criterion $\mathrm{E}_e\left[\phi\left(e\right)\right]$ over $\mathbf{G}$, provided that the cost function $\phi\left(e\right)$ is even and convex.

Proof. This property is a direct consequence of the results of [2]. Here, we give an alternative proof using information theory. By Theorem 3, we have

$$\mathrm{E}_e\left[\phi\left(e\right)\right]\propto\mathrm{E}\left[-\log q_\phi\left(e\right)\right]=\left\{D_{KL}\left(p_e\left(e\right)\|q_\phi\left(e\right)\right)+H\left(e\right)\right\}$$

where $q_\phi\left(e\right)=\exp\left[-\gamma_0-\gamma_1\phi\left(e\right)\right]$, $\gamma_1>0$. For any $\beta\in\mathbf{G}$, and any $\boldsymbol{y}$ fixed, denote $\varepsilon\left(\boldsymbol{y}\right)=\beta\left(\boldsymbol{y}\right)-m\left(\boldsymbol{y}\right)$. If $\phi\left(e\right)$ is convex, we have

$$\phi\left(e\right)\leq\frac{1}{2}\left[\phi\left(e-\varepsilon\left(\boldsymbol{y}\right)\right)+\phi\left(e+\varepsilon\left(\boldsymbol{y}\right)\right)\right] \tag{22}$$

And hence (Note that $q_\phi\left(e\right)=\exp\left[-\gamma_0-\gamma_1\phi\left(e\right)\right]$)

$$-\log\left[q_\phi\left(e\right)\right]\leq-\log\left[\sqrt{q_\phi\left(e-\varepsilon\left(\boldsymbol{y}\right)\right)q_\phi\left(e+\varepsilon\left(\boldsymbol{y}\right)\right)}\right] \tag{23}$$

Then we derive

$$\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=m}\leq\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=m+\varepsilon=\beta} \tag{24}$$

The detailed derivation of (24)is given by (25) at the bottom of the page, in which (a) follows from the symmetry of $p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)$ and evenness of q_ϕ, and (b) follows from the shift-invariance of KL-divergence and entropy. Therefore, we have

$$\begin{aligned}&\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{g=m}=\mathrm{E}_{\boldsymbol{y}}\left\{\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=m}\right\}=\\ &\int_{\mathbf{R}^k}p_{\boldsymbol{y}}\left(\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=m}\right)\mathrm{d}y\leq\\ &\int_{\mathbf{R}^k}p_{\boldsymbol{y}}\left(\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=\beta}\right)\mathrm{d}y=\\ &\mathrm{E}_{\boldsymbol{y}}\left\{\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=\beta}\right\}=\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{g=\beta}\end{aligned}$$

It follows that

$$m=\arg\min_{g\in\mathbf{G}}\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]=\arg\min_{g\in\mathbf{G}}\mathrm{E}_e\left[\phi\left(e\right)\right]$$

□

Remark 5. It is well-known that the optimum solution for the minimum mean-square error (MMSE) estimation is the conditional mean $m\left(\boldsymbol{Y}\right)=\mathrm{E}\left[X\,|\boldsymbol{Y}\right]$. By placing a mild restriction on the conditional density function of the error, Property 2 suggests that the conditional mean $m\left(\boldsymbol{Y}\right)=\mathrm{E}\left[X\,|\boldsymbol{Y}\right]$, which minimizes $\mathrm{E}_e\left[e^2\right]$, also minimizes $\mathrm{E}_e\left[\phi\left(e\right)\right]$, where ϕ is even and convex.

3 Information theoretic interpretation of the optimum error criterion

Consider the system identification scheme of Fig. 3, in which the transfer functions of the plant and the adaptive

$$\begin{aligned}&\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=m}=\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left[-\log q_\phi\left(e\right)\right]\mathrm{d}e\leq\\ &\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left\{-\log\left[\sqrt{q_\phi\left(e-\varepsilon(\boldsymbol{y})\right)q_\phi\left(e+\varepsilon(\boldsymbol{y})\right)}\right]\right\}\mathrm{d}e=\\ &\tfrac{1}{2}\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left[-\log q_\phi\left(e-\varepsilon(\boldsymbol{y})\right)\right]\mathrm{d}e+\tfrac{1}{2}\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left[-\log q_\phi\left(e+\varepsilon(\boldsymbol{y})\right)\right]\mathrm{d}e\overset{(a)}{=}\\ &\tfrac{1}{2}\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left[-\log q_\phi\left(e-\varepsilon(\boldsymbol{y})\right)\right]\mathrm{d}e+\tfrac{1}{2}\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left[-\log q_\phi\left(e-\varepsilon(\boldsymbol{y})\right)\right]\mathrm{d}e=\\ &\int_{\mathbf{R}}p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\left[-\log q_\phi\left(e-\varepsilon(\boldsymbol{y})\right)\right]\mathrm{d}e=\\ &D_{KL}\left(p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\|q_\phi\left(e-\varepsilon(\boldsymbol{y})\right)\right)+H\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\overset{(b)}{=}\\ &D_{KL}\left(p_e\left(e+\varepsilon(\boldsymbol{y})\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)\|q_\phi\left(e\right)\right)+H\left(e-\varepsilon(\boldsymbol{y})\,|\boldsymbol{Y}=\boldsymbol{y},g=m\right)=\\ &D_{KL}\left(p_e\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m+\varepsilon\right)\|q_\phi\left(e\right)\right)+H\left(e\,|\boldsymbol{Y}=\boldsymbol{y},g=m+\varepsilon\right)=\mathrm{E}_e\left[-\log q_\phi\left(e\right)\right]\big|_{\boldsymbol{Y}=\boldsymbol{y},g=m+\varepsilon=\beta}\end{aligned} \tag{25}$$

filter are both represented in the FIR form by $\boldsymbol{W}(z) = \sum_{i=1}^{m} w_i z^{-i+1}$, where we define the m-dimensional parameter (or weight) vector $\boldsymbol{W} = [w_1, w_2, \cdots, w_m]^{\mathrm{T}}$ such that $\boldsymbol{W} = \boldsymbol{W}^*$ for the unknown plant and $\boldsymbol{W} = \boldsymbol{W}(k)$ for the adaptive filter at each iteration k. Our goal is to estimate the parameter vector $\boldsymbol{W}^*$ of the unknown plant given the noisy observation $d(k)$, which may be considered as the desired response of the adaptive filter when driven by the stochastic input $x(k)$ and perturbed by additive noise $n(k)$. In this case, the error signal $e(k)$ is formed as

$$\begin{aligned} & e(k) = d(k) - y(k) = \\ & \boldsymbol{W}^{*\mathrm{T}} \boldsymbol{X}(k) + n(k) - \boldsymbol{W}^{\mathrm{T}}(k) \boldsymbol{X}(k) = \\ & \boldsymbol{V}^{\mathrm{T}}(k) X(k) + n(k) \end{aligned}$$

where $y(k)$ is the output of the adaptive filter, $\boldsymbol{X}(k) = [x(k), x(k-1), \cdots, x(k-m+1)]^{\mathrm{T}}$ is the input data vector, and $\boldsymbol{V}(k) = \boldsymbol{W}^* - \boldsymbol{W}(k)$ is the weight error vector.

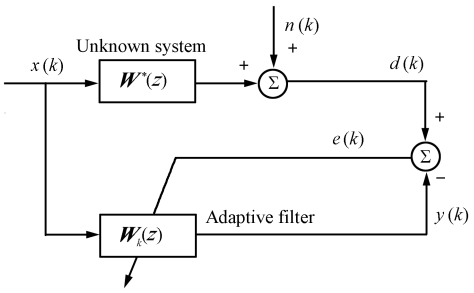

Fig. 3 Scheme of system identification with adaptive FIR filter

Given an error criterion $\mathrm{E}_e[\phi(e)]$, the above parameter estimation can be solved by the following stochastic gradient algorithm:

$$\begin{aligned} \boldsymbol{W}(k+1) = & \boldsymbol{W}(k) - \eta \left. \frac{\partial}{\partial \boldsymbol{W}} \phi(e(k)) \right|_{\boldsymbol{W}=\boldsymbol{W}(k)} = \\ & \boldsymbol{W}(k) - \eta \left. \frac{\partial \phi(e(k))}{\partial e(k)} \frac{\partial e(k)}{\partial \boldsymbol{W}} \right|_{\boldsymbol{W}=\boldsymbol{W}(k)} = \\ & \boldsymbol{W}(k) + \eta \left. \frac{\partial \phi(e(k))}{\partial e(k)} \boldsymbol{X}(k) \right|_{\boldsymbol{W}=\boldsymbol{W}(k)} \end{aligned} \tag{26}$$

The adaptation algorithm brings the problem of how to choose a suitable error criterion to maximize the performance, which may be measured in terms of both adaptation speed and minimum excess MSE. Among the literatures in this direction, the work of Douglas[5] deserves special attention. According to [5], under certain mild assumptions, the optimum error criterion for the slow adaptation (small η) satisfies

$$\frac{\partial \phi_{\mathrm{opt}}(x)}{\partial x} = -\frac{\gamma p_n'(x)}{2 p_n(x)} \tag{27}$$

where $p_n(\cdot)$ is the PDF of the interfering noise $n(k)$. By indefinite integral approach, $\phi_{\mathrm{opt}}(\cdot)$ is given by

$$\phi_{\mathrm{opt}}(x) = \gamma_0 + \gamma_1 \{-\log p_n(x)\} \tag{28}$$

where $\gamma_1 = \gamma/2$. As $\lim_{|x| \to +\infty} \phi_{\mathrm{opt}}(x) = +\infty$, we have $\gamma_1 > 0$, and by Definition 1, we get

$$\phi_{\mathrm{opt}}(e(k)) \propto \{-\log p_n(e(k))\} \tag{29}$$

It follows that

$$\begin{aligned} \mathrm{E}_e[\phi_{\mathrm{opt}}(e(k))] \propto & \mathrm{E}_e\{-\log p_n(e(k))\} = \\ & D_{KL}(p_e(e(k)) \| p_n(e(k))) + H(e(k)) \end{aligned} \tag{30}$$

Thus, the optimum error criterion $\mathrm{E}_e[\phi_{opt}(e(k))]$ is equivalent to the error′s entropy plus the KL-divergence between the PDFs of the error $e(k)$ and interfering noise $n(k)$.

For the identification scheme depicted in Fig. 3, the ideal error signal (when $\boldsymbol{W} = \boldsymbol{W}^*$) equals the disturbance noise $n(k)$. So the desired PDF of the error is $p_e^*(e_k) = p_n(e_k)$. Hence, (30) can be rewritten as

$$\mathrm{E}_e[\phi_{\mathrm{opt}}(e(k))] \propto D_{KL}(p_e(e(k)) \| p_e^*(e(k))) + H(e(k)) \tag{31}$$

The meanings of the KL-divergence and entropy enables us to understand that the first part (the KL-divergence) of the right side of (31) provides a "force" to shape the error′s PDF into the desired one, and that the second part (the entropy) provides another "force" to decrease the dispersion (or uncertainty) of the error. Obviously, both "forces" make the parameter vector $\boldsymbol{W}(k)$ approach the optimum one ($\boldsymbol{W}^*$). Thus, the optimum error criterion owns abilities of both PDF shaping and dispersion decreasing. Further, near the convergence, we have $\boldsymbol{W}(k) \approx \boldsymbol{W}^*$, and hence,

$$e(k) = (\boldsymbol{W}^* - \boldsymbol{W}(k))^{\mathrm{T}} \boldsymbol{X}(k) + n(k) \approx n(k) \tag{32}$$

It follows immediately that $p_e(x) \approx p_n(x)$, and $D_{KL}(p_e(e(k)) \| p_n(e(k))) \approx 0$. Then

$$\begin{aligned} \mathrm{E}_e[\phi_{\mathrm{opt}}(e(k))] \propto & D_{KL}(p_e(e(k)) \| p_e^*(e(k))) + H(e(k)) \approx \\ & H(e(k)) \end{aligned} \tag{33}$$

Therefore, the optimum error criterion $\mathrm{E}_e[\phi_{\mathrm{opt}}(e(k))]$ is equivalent to the error entropy criterion near the convergence. This gives an interesting interpretation for why the error entropy criteria perform well in the areas of adaptive system training or the supervised learning.

4 Information theoretic approach for the choice of p-power error criteria

We now consider the choice of error criterion as a parameter search in which a suitable structure of the criterion is assumed. Among parameterized error criteria, the family of p-power error criteria ($\mathrm{E}_e[\phi(e)] = \mathrm{E}_e[|e|^p]$, see also Example 1), in which $p > 0$ are the parameter, is widely used due to its low computation requirement and good performance[6, 34]. The cost functions and adaptation algorithms under p-power error criteria for $p = 1, 2, 3, 4$ are listed in Table 1.

Table 1 The cost functions and adaptation algorithms for $p = 1, 2, 3, 4$

p	$\phi(e)$	Adaptation algorithms
1	$\lvert e \rvert$	Least absolute difference (LAD)[12]
2	e^2	Least mean square (LMS)[3]
3	$\lvert e^3 \rvert$	Least mean absolute third (LMAT)[36]
4	e^4	Least mean fourth (LMF)[12]

With p-power error criteria, one is often confronted with the problem of how to choose a suitable value of p to improve the performance of the adaptation algorithm. For cases in where $p = 2K$ ($K = 1, 2, \cdots$), the problem has been solved by Walach[12]. In his approach, the optimum choice of K can be by minimizing a cost function $\alpha(K)$, which depends on the moments of the interfering noise $n(k)$. Here, we give an information theoretic approach to the choice of p, in which the number p is not limited to the form of $p = 2K$. By Theorem 3, we have

$$\mathrm{E}_e[|e|^p] \propto D_{KL}(p_e(e) \| q_\phi(e)) + H(e) \tag{34}$$

where $q_\phi(e) = \exp[-\gamma_0 - \gamma_1 |e|^p]$. Comparing (34) with (30), we may easily conclude that if the probability density functions $q_\phi(x)$ and $p_n(x)$ are "close" enough, the p-power error criterion will be approximately equivalent to the optimum error criterion. Since the KL-divergence expresses a "similarity" or a "distance" of two probability measures, we may choose number p such that the KL-divergence between $q_\phi(x)$ and $p_n(x)$ is minimized, that is,

$$p_{\text{opt}} = \arg \min_{p \in \mathbf{R}_+} D_{KL}(p_n(x) \| q_\phi(x)) = \arg \min_{p \in \mathbf{R}_+} D_{KL}(p_n(x) \| \exp[-\gamma_0 - \gamma_1 |x|^p]) \tag{35}$$

As constants γ_0 and γ_1 depend on both p and $\mathrm{E}_e(|x|^p)$ (see Example 1), the optimum p value will depend on $\mathrm{E}_e(|x|^p)$, too. In order to optimize p over all range of $\mathrm{E}_e(|x|^p)$, we propose the following optimization:

$$p_{\text{opt}} = \arg \min_{p \in \mathbf{R}_+} \{\Psi(p)\} = \arg \min_{p \in \mathbf{R}_+} \left\{ \min_{\tau \in \mathbf{R}_+} D_{KL}(p_n(x) \| \exp[-\gamma_0 - \gamma_1 |x|^p]) \right\} \tag{36}$$

where $\Psi(p) = \min_{\tau \in \mathbf{R}_+} D_{KL}(p_n(x) \| \exp[-\gamma_0 - \gamma_1 |x|^p])$ and $\tau = \mathrm{E}_e(|x|^p)$. Clearly, the smaller value $\Psi(p)$, the better p.

Consider several specific probability density functions of $n(k)$, which are depicted in Fig. 4. Note that the term "MixN" denotes the mixed normal (Gaussian) distribution. For each noise distribution, the functions $\Psi(p)$ are plotted in Fig. 5. Clearly, for the Gaussian and Laplace cases, $\Psi(p)$ achieves its global minima at $p = 2$ and $p = 1$, respectively; for the Uniform case, the larger the p value, the smaller the function $\Psi(p)$; for the mixed normal distribution, $\Psi(p)$ is in its global minimum at approximately $p = 3$. Table 2 gives the exact values of $\Psi(p)$ $(p = 1, 2, 3, 4)$ for each noise distribution. Therefore, if p is limited in set $\{1, 2, 3, 4\}$, then $p = 1, 2, 3, 4$ will be the optimum choices for the Laplace, Gaussian, MixN, and Uniform noise, respectively.

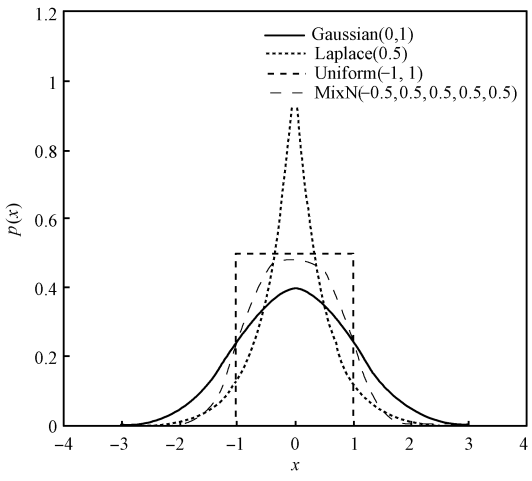

Fig. 4 Several probability density functions (Gaussian(0,1): $p(x) = \frac{1}{\sqrt{2\pi}} \exp(-0.5x^2)$; Laplace(0.5): $p(x) = \exp(-2|x|)$; Uniform(–1,1): $p(x) = 0.5$, $(-1 \le x \le 1)$; MixN: $p(x) = \frac{1}{\sqrt{2\pi}} \{\exp(-2(x-0.5)^2) + \exp(-2(x+0.5)^2)\}$)

Table 2 $\Psi(p)$ values for $p = 1, 2, 3, 4$ and each distribution

	$p=1$	$p=2$	$p=3$	$p=4$
$n(k) \sim$ Gaussian	0.0484	0	0.0163	0.0472
$n(k) \sim$ Laplace	0	0.0724	0.1837	0.2929
$n(k) \sim$ Uniform	0.3069	0.1765	0.1243	0.0962
$n(k) \sim$ MixN	0.0915	0.0097	0.0007	0.0113

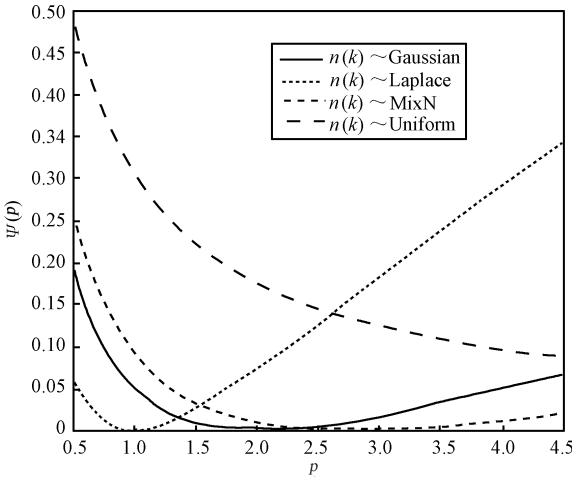

Fig. 5 Curves of $\Psi(p)$ for each noise distribution

We now perform Monte-Carlo simulations to verify the above conclusion. Let us consider the system identification scheme of Fig. 3, in which we assume $\boldsymbol{W}^* = [0.1, 0.3, 0.5, 0.3, 0.1]^{\mathrm{T}}$. In the experiments, the input signal $x(k)$ is unity-power white Gaussian noise, and the initial parameters of the adaptive filter are set to zero. For each noise distribution (Laplace, Gaussian, MixN, and Uniform) and each p value $(p = 1, 2, 3, 4)$, 100 Monte-Carlo simulations are run and the results are averaged. The average convergence curves for each noise distribution and each p value are shown in Fig. 6, in which the weight error vector norm is defined as

$$\|\boldsymbol{W}(k) - \boldsymbol{W}^*\| = \sqrt{(\boldsymbol{W}(k) - \boldsymbol{W}^*)^{\mathrm{T}} (\boldsymbol{W}(k) - \boldsymbol{W}^*)} \tag{37}$$

Note that for each noise distribution, the step-sizes for each p are chosen so that the initial convergence rates are visually identical. From Fig. 6, it is evident that for each noise, the smaller the $\Psi(p)$ value (see Table 2), the better performance (smaller misadjustment) the stochastic gradient algorithm. In order to compare the statistical results of the system training, we summarize in Table 3 the sample mean and standard deviation of the weight w_3 $(w_3^* = 0.5)$, from which we see that the bias and deviation of the optimum algorithms (with the optimum p values) are both smaller. Clearly, the simulation results agree with the previous analysis.

5 Concluding remarks

Relationships between the error criteria and information theoretic criteria (entropy, KL-divergence, etc.) were investigated. The error criterion was shown to be equivalent to the error′s entropy plus a KL-divergence. This basic result was used to prove two important properties of error criteria. The optimum error criterion, which owns the abilities of both PDF shaping and dispersion decreasing, can

be understood from the meanings of the entropy and KL-divergence. In order to choose a suitable p-power error criterion , a KL-divergence based optimization was proposed, and was verified by Monte-Carlo simulation experiments of FIR system training. This work bridges the traditional error criteria and information theoretic criteria together, and will help us in choosing or constructing a suitable error criterion for a specific application.

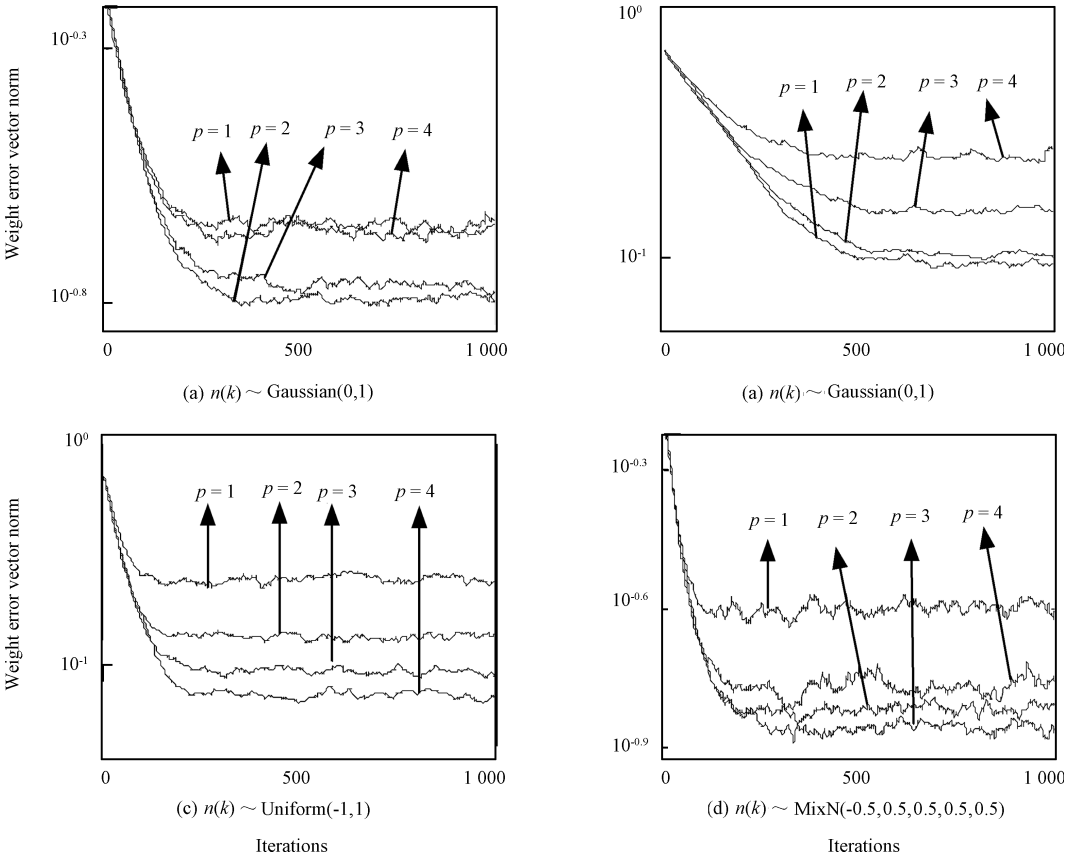

Fig. 6 Average convergence curves for each p value and each noise distribution with 100 Monte-Carlo simulations

Table 3 The mean ± deviation results of w_3 ($w_3^* = 0.5$) with 100 Monte-Carlo simulations

	$p=1$	$p=2$	$p=3$	$p=4$
$n(k) \sim$ Gaussian	0.4955 ± 0.0962	0.4983 ± 0.0664	0.5119 ± 0.0746	0.4964 ± 0.0991
$n(k) \sim$ Laplace	0.4971 ± 0.0415	0.4945 ± 0.0445	0.4841 ± 0.0751	0.4862 ± 0.0947
$n(k) \sim$ Uniform	0.5168 ± 0.1118	0.4958 ± 0.0656	0.5085 ± 0.0401	0.4978 ± 0.0303
$n(k) \sim$ MixN	0.5130 ± 0.1160	0.4924 ± 0.0784	0.5038 ± 0.0623	0.5091 ± 0.0730

References

1 Sherman S. Non-mean-square error criteria. *IRE Transactions on Information Theory*, 1958, **4**(3): 125−126

2 Hall E B, Wise G L. On optimum estimation with respect to a large family of cost functions. *IEEE Transactions on Information Theory*, 1991, **37**(3): 691−693

3 Haykin S. *Adaptive Filtering Theory (Third Edition)*. New York: Prentice Hall, 1996

4 Kailath T, Sayed A H, Hassibi B. *Linear Estimation*. New Jersey: Prentice Hall, 2000

5 Douglas S C, Meng T H Y. Stochastic gradient adaptation under general error criteria. *IEEE Transactions on Signal Processing*, 1994, **42**(6): 1335−1351

6 Pei S C, Tseng C C. Least mean p-power error criterion for adaptive FIR filter. *IEEE Journal on Selected Areas in Communications*, 1994, **12**(9): 1540−1547

7 Widrow B, Walach E. *Adaptive Inverse Control*. New York: Prentice Hall, 1996

8 Wang Z, Bovik A C. Mean squared error: love it or leave it? *IEEE Signal Processing Magazine*, 2009, **26**(1): 98−117

9 Erdogmus D, Principe J C. From linear adaptive filtering to nonlinear information processing. *IEEE Signal Processing Magazine*, 2006, **23**(6): 14−33

10 Brown J L. Asymmetric non-mean-square error criteria. *IRE Transactions on Automatic Control*, 1962, **7**(1): 64−66

11 Zakai M. General error criteria. *IEEE Transactions on Information Theory*, 1964, **10**(1): 94−95

12 Walach E, Widrow B. The least mean fourth (LMF) adaptive algorithm and its family. *IEEE Transactions on Information Theory*, 1984, **30**(2): 275−283

13 Min S, Nikias C L. Signal processing with fractional lower order moments: stable processes and their applications. *Proceedings of the IEEE*, 1993, **81**(7): 986−1010

14 Cover T M, Thomas J A. *Element of Information Theory*. Chichester: Wiley and Son Inc., 1991

15 Weidemann H L, Stear E B. Entropy analysis of parameter estimation. *Information and Control*, 1969, **14**(6): 493−506

16 Weidemann H L, Stear E B. Entropy analysis of estimating systems. *IEEE Transactions on Information Theory*, 1970, **16**(3): 264−270

17 Minamide N. An extension of the entropy theorem for parameter estimation. *Information and Computation*, 1982, **53**(1-2): 81−90

18 Erdogmus D, Hild K E, Principe J C. Online entropy manipulation: stochastic information gradient. *IEEE Signal Processing Letters*, 2003, **10**(8): 242−245

19 Erdogmus D, Principe J C. Generalized information potential criterion for adaptive system training. *IEEE Transactions on Neural Networks*, 2002, **13**(5): 1035−1044

20 Erdogmus D, Principe J C. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Transactions on Signal Processing*, 2003, **51**(7): 1966−1978

21 Erdogmus D, Principe J C. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In: Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation. Berlin, German: Springer, 2000. 75−80

22 Principe J C, Xu D, Zhao Q, Fisher J W. Learning from examples with information theoretic criteria. *The Journal of VLSI Signal Processing*, 2000, **26**(1-2): 61−77

23 Chen B D, Hu J C, Pu L, Sun Z Q. Stochastic gradient algorithm under (h, ϕ)-entropy criterion. *Circuits, Systems, and Signal Processing*, 2007, **26**(6): 941−960

24 Han S, Rao S, Erdogmus D, Jeong K H, Principe J C. A minimum-error entropy criterion with self-adjusting step-size (MEE-SAS). *Signal Processing*, 2007, **87**(11): 2733−2745

25 Wolsztynski E, Thierry E, Pronzato L. Minimum entropy estimation in semi-parametric models. *Signal Processing*, 2005, **85**(5): 937−949

26 Silva L M, Felgueiras C S, Alexandre L A, Marques de Sa J. Error entropy in classification problems: a univariate data analysis. *Neural Computation*, 2006, **18**(9): 2036−2061

27 Kapur J N, Kesavan H K. *Entropy Optimization Principles with Applications*. New York: Academic Press, 1992

28 Karmeshu J. *Entropy Measures, Maximum Entropy Principle and Emerging Applications*. Berlin: Springer-Verlag, 2003

29 Ling T, Taniar D. Adaptive estimated maximum-entropy distribution model. *Information Science*, 2007, **177**(15): 3110−3128

30 Bryson A E, Ho Y C. *Applied and Optimal Control, Optimization, Estimation, and Control*. New York: Hemisphere Publishing, 1975. 47−48

31 Varanasi M K, Aazhang B. Parametric generalized Gaussian density estimation. *The Journal of the Acoustical Society of America*, 1989, **86**(4): 1404−1415

32 Kokkinakis K, Nandi A K. Exponent parameter estimation for generalized Gaussian probability density functions with application to speech modeling. *Signal Processing*, 2005, **85**(9): 1852−1858

33 Lo J T, Wanner T. Existence and uniqueness of risksensitivity estimates. *IEEE Transactions on Automatic Control*, 2002, **47**(11): 1945−1948

34 Xiao Y, Tadokoro Y, Shida K. Adaptive algorithm based on least mean p-power error criterion for Fourier analysis in additive noise. *IEEE Transactions on Signal Processing*, 1999, **47**(4): 1172−1181

35 Chambers J, Avlonitis A. A robust mixed-norm adaptive filter algorithm. *IEEE Signal Processing Letters*, 1997, **4**(2): 46−48

36 Cho S H, Kim S D. Adaptive filters based on the high order error statistics. In: Proceedings of IEEE Asia Pacific Conference on Circuits and Systems. Seoul, South Korea: IEEE, 1996. 109−112

CHEN Ba-Dong Received his bachelor and M.S. degrees in control theory and engineering from Chongqing University in 1997 and 2003, and the Ph. D. degree in computer science and technology from Tsinghua University in 2008, respectively. He is currently a postdoctor at the Institute of Manufacturing Engineering, Department of Precision Instruments and Mechanology, Tsinghua University. His research interest covers signal processing, adaptive control, and information theoretic aspects of control systems. Corresponding author of this paper.
E-mail: chenbd04@mails.tsinghua.edu.cn

HU Jun-Chun Received his Ph. D. degree in control science and engineering from Nanjing University of Science and Technology in 1998. He was a postdoctoral researcher in Nanjing University of Aeronautics and Astronautics in 1999 and Tsinghua University in 2002, respectively. He is currently an associate professor in the Department of Computer Science and Technology, Tsinghua University. His research interest covers flight control, aerial robot and intelligent control. E-mail: hujinchun@tsinghua.edu.cn

ZHU Yu Received his bachelor degree in radio electronics from Beijing Normal University in 1983, his M.S. degree in computer applications, and his Ph. D. degree in mechanical design and theory in 1993 and 2001 at China University of Mining & Technology, respectively. He is currently a professor at the Institute of Manufacturing Engineering of Department of Precision and Mechanology, Tsinghua University. His research interest covers parallel machanism and theory, two photon micro-fabrication, ultra-precision motion system and motion control.
E-mail: zhuyu@tsinghua.edu.cn

SUN Zeng-Qi Received his bachelor degree from the Department of Automatic Control, Tsinghua University in 1966 and the Ph. D. degree in control engineering from the Chalmas University of Technology, Sweden in 1981, respectively. He is currently a professor in the Department of Computer Science and Technology, Tsinghua University. He is the author or coauthor of more than 100 papers and eight books on control and robotics. His research interest covers robotics, intelligent control, fuzzy system, neural networks, and evolutionary computation.
E-mail: szq-dcs@mail.tsinghua.edu.cn