Information Sciences 181 (2011) 1384–1402

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Δ-Entropy: Definition, properties and applications in system identification with quantized data

Badong Chen [a,*], Yu Zhu [b], Jinchun Hu [b], José C. Príncipe [a]

[a] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA
[b] Department of Precision Instruments and Mechanology, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 23 January 2010
Received in revised form 21 November 2010
Accepted 29 November 2010
Available online 7 December 2010

Keywords:
Δ-Entropy
Minimum error entropy criterion
System identification
Estimation of distribution algorithm

ABSTRACT

Recently, the minimum error entropy criterion, an information theoretic alternative to the traditional mean square error criterion, has been successfully used in the contexts of machine learning and signal processing. For system identification, however, the MEE criterion will be no longer suitable if the training data are discrete-valued, since minimizing error's discrete entropy cannot constrain error's dispersion. In this paper, to make the MEE criterion suitable for the discrete-valued data cases, we give a new entropy definition for the discrete random variables, i.e. the Δ-entropy, based on Riemann sums for finite size partitions. A probability weighted formula is established to calculate the average partition.This new entropy retains some important properties of the differential entropy and reduces to discrete entropy under certain conditions. Unlike discrete entropy, the Δ-entropy is sensitive to the dynamic range of the data, and can be used as a superior optimality criterion in system identification problems. Also, we present a plug-in estimate of Δ-entropy, analyze its asymptotic behavior and explore the links to the kernel based and m-spacing based estimates for differential entropy. Finally, the Δ-entropy criterion is applied in system identification with coarsely quantized input–output data to search for the optimum parameter set. Monte Carlo simulations demonstrate the performance improvement that may be achieved with the Δ-entropy criterion.

Published by Elsevier Inc.

1. Introduction

For the univariate discrete random variable X with M discrete values $\boldsymbol{S} = (s_1, s_2, \ldots, s_M)$, and corresponding probability distribution $\boldsymbol{P} = (p_1, p_2, \ldots, p_M)$, the discrete entropy, denoted by $H(X)$ or $H(\boldsymbol{P})$, is defined by [4]

$$H(X) = -\sum_{i=1}^{M} p_i \log p_i \tag{1}$$

which is a non-negative concave function of $\boldsymbol{P}$. This definition can be extended to the case where the discrete variable takes a countable infinite set of values.

The discrete entropy measures the average uncertainty (information) contained in the probability distribution,[1] and can be used to measure many other concepts such as equality, disorder, diversity, similarity, unbiasedness, randomness, and so

* Corresponding author.
E-mail address: chenbd04@mails.tsinghua.edu.cn (B. Chen).

[1] This entropy is called probabilistic entropy. In order to measure non-probabilistic uncertainty contained in a fuzzy set, a variety of fuzzy entropies are also defined (see e.g. [45–48]).

0020-0255/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.ins.2010.11.037

on [15]. However, as the discrete entropy depends only on the distribution $\boldsymbol{P}$, and takes no account of the discrete values, it is independent of the dynamic range of the random variable. Therefore, we conclude that discrete entropy is unable to differentiate between two random variables that have different dynamic ranges and the same distribution. In fact, the discrete random variables with the same entropy may have arbitrarily small or large variance, a typical measure for dispersion of the random variable, where dispersion here is defined loosely as the concentration spread around the mean value.

If X is a continuous random variable with probability density function (PDF) $f(x)$, the differential entropy, denoted by $h(X)$ or $h(f)$, is defined as [4]

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx \tag{2}$$

Different from the discrete entropy, the differential entropy can be negative and even minus infinite. So strictly speaking, the differential entropy cannot represent a measure of uncertainty since uncertainty should in general be positive, although it can be used as such if we take into consideration the dimension of the distribution [49]. However, differential entropy can be used to measure the dispersion of a continuous random variable. For example, if X is of Gaussian distribution with variance $Var(X)$, the differential entropy will be

$$h(X) = \frac{1}{2}\log(2\pi eVar(X)) \tag{3}$$

It is clear that smaller differential entropy implies smaller variance (dispersion).

As the differential entropy measures both the probabilistic uncertainty and dispersion, it can be used as an optimality criterion in the estimation and identification problems, which leads to the so called *minimum error entropy* (MEE) criterion [2,6–8,13,20,23,27,37]. The MEE criterion is concerned with the use of differential entropy as a cost function for system identification and parameter estimation. Traditionally, the *minimum mean square error* (MMSE) criterion has been the workhorse of estimation and identification because of simplicity and solid statistical foundation. However, recent studies suggest that, as an optimality criterion, MEE is superior to MMSE, since minimizing the error entropy constrains all moments of the error's PDF, whereas MMSE constrains only the first and second moments of the PDF [2,6–8,23,27]. The previous studies have demonstrated that the MEE criterion offers potentially significant performance improvement in system identification, particularly in nonlinear and non-Gaussian settings.

In many practical system identification scenarios, the unknown system's inputs and outputs may be discrete-valued for a variety of reasons:

(1) For many systems, especially in the field of digital communication, the input signals take values only in finite alphabetical sets [10,17,18,25].
(2) Coarsely quantized plant's inputs and outputs are commonly used, when the data are obtained from an A/D converter or from a communication channel [1,21,26,33,38]. Typical contexts involving quantized data include digital control systems (DCS), networked control systems (NCS), wireless sensor networks (WSN), etc.
(3) Binary-valued sensors occur frequently in practical systems [35,36,40]. Some typical examples of binary-valued sensors can be found in [36].
(4) Discrete-valued time-series are common in practice. In recent years, the count or integer-valued data time-series have gained increasing attention for point processes [41–44].

Sometimes, due to computational consideration, even if the observed input and output signals are continuous-valued, one may classify the data into groups and obtain the discrete-valued data [24, Chapter 5]. In all these situations one normally applies differential entropy to implement the MEE criterion, in spite of the fact that the random variable is indeed discrete. When the discretization is coarse (i.e. few levels) the use of differential entropy may carry a penalty in performance that is normally not quantified. Alternatively, the MEE implemented with discrete entropy will become ill-suited since the minimization fails to constrain the error's dispersion which should be pursued because the error dynamic range decreases over iterations.

The present paper augments the MEE criterion choices by providing a new entropy definition for discrete random variables, called Δ-entropy, which comprises two terms: one is the discrete entropy, and the other is the logarithm of the average interval between two successive discrete values. This new entropy retains important properties of the differential entropy and reduces to the traditional discrete entropy for a special case. More importantly, the proposed entropy definition can still be used to measure the dispersion of a discrete random variable, and hence can be used as an MEE optimality criterion in system identification with discrete-valued data.

The paper is organized as follows. In Section 2 we give the definition of Δ-entropy. In Section 3 we investigate the properties of Δ-entropy. In Section 4 we analyze the plug-in estimate of Δ-entropy. In Section 4.2, we apply the Δ-entropy criterion in system identification with quantized I/O data. The estimation of distribution algorithm (EDA) is used as the parameter search algorithm, and Monte Carlo simulations are performed to demonstrate the satisfactory performance. Finally, in Section 5, we draw some conclusions and discuss future work.

2. Definition of Δ-entropy

In this section, we propose an alternative entropy definition for discrete random variables. Before proceeding, let's review the relationship between the differential and the discrete entropy (see also [4] for details). Consider a continuous random variable X with PDF $f(x)$. We can produce a quantized random variable X^{Δ} (see Fig. 1), given by

$$X^{\Delta} = s_i, \quad \text{if } i\Delta \leqslant X < (i+1)\Delta \tag{4}$$

where s_i is one of countable values, which satisfies

$$i\Delta \leqslant s_i < (i+1)\Delta, \quad \text{and} \quad f(s_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx \tag{5}$$

The probability that $X^{\Delta} = s_i$ is

$$p_i = \Pr(X^{\Delta} = s_i) = f(s_i)\Delta \tag{6}$$

And the discrete entropy $H(X^{\Delta})$ can be calculated as

$$H(X^{\Delta}) = -\sum_{i=-\infty}^{\infty} p_i \log p_i = -\sum_{i=-\infty}^{\infty} \Delta f(s_i) \log f(s_i) - \log \Delta \tag{7}$$

If the density function $f(x)$ is Riemann integrable, the following limit holds

$$\lim_{\Delta \to 0} \left(H(X^{\Delta}) + \log \Delta\right) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx = h(X) \tag{8}$$

This suggests that, if the quantization interval Δ is small enough, we have

$$h(X) \approx H(X^{\Delta}) + \log \Delta \tag{9}$$

Thus the differential entropy of a continuous random variable X is approximately equal to the discrete entropy of the quantized variable X^{Δ} plus the logarithm of the quantization interval Δ. This important relationship explains why differential entropy is sensitive to dispersion. That is, compared with the discrete entropy, the differential entropy "contains" the term $\log \Delta$, which measures the average interval between two successive quantized values since

$$\Delta = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{i=-N}^{N} |s_{i+1} - s_i| \tag{10}$$

The above analysis inspired us to seek a new entropy definition for discrete random variables that will measure uncertainty as well as dispersion and is defined as follows:

Definition 1. For a discrete random variable X with values $\boldsymbol{S} = (s_1, s_2, \ldots, s_M)$, and the corresponding distribution $\boldsymbol{P} = (p_1, p_2, \ldots, p_M)$, the Δ-entropy, denoted by $H_{\Delta}(X)$ or $H_{\Delta}(\boldsymbol{S}, \boldsymbol{P})$, is defined as

$$H_{\Delta}(X) = -\sum_{i=1}^{M} p_i \log p_i + \log \Delta(X) \tag{11}$$

where $\Delta(X)$ (or $\Delta(\boldsymbol{S}, \boldsymbol{P})$) stands for the average interval (distance) between two successive values.

Remark 1. The Δ-entropy contains two terms, where the first term is identical to the classical discrete entropy and the second term equals the logarithm of the average interval between two successive values. Obviously, this new entropy can be used as an optimality criterion in the estimation and identification problems, because minimizing error's Δ-entropy will decrease the average interval and automatically force the error samples to concentrate without renormalization of the error variable through training.

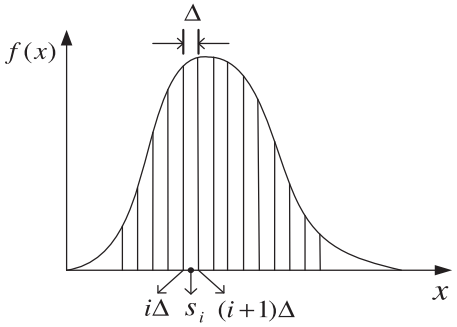

Fig. 1. Quantization of a continuous random variable.

Now we discuss how to calculate the average interval $\Delta(X)$ to preserve the known properties of entropy. In the rest of the paper, we assume, without loss of generality, that the discrete values satisfy $s_1 < s_2 < \cdots < s_M$. Naturally, we immediately think of the arithmetic and geometric means, that is

$$\begin{cases} \Delta(X) = \frac{1}{M-1}\sum_{i=1}^{M-1}|s_{i+1}-s_i| & \text{for arithmetic mean} \\ \Delta(X) = \left(\prod_{i=1}^{M-1}|s_{i+1}-s_i|\right)^{1/(M-1)} & \text{for geometric mean} \end{cases} \tag{12}$$

Both arithmetic and geometric means take no account of the distribution $\boldsymbol{P}$. A more reasonable approach is to calculate the average interval $\Delta(X)$ by a *probability-weighted method*. For example, we can use the following formula:

$$\Delta(X) = \sum_{i=1}^{M-1}|s_{i+1}-s_i|\frac{p_i+p_{i+1}}{2} \tag{13}$$

However, if $(p_1 + p_M) > 0$, the sum of weights will be less than one, because

$$\sum_{i=1}^{M-1}\frac{p_i+p_{i+1}}{2} = 1 - \frac{p_1+p_M}{2} < 1 \tag{14}$$

To address this issue, we propose the formula:

$$\Delta(X) = \sum_{i=1}^{M-1}|s_{i+1}-s_i|\frac{p_i+p_{i+1}}{2} + \frac{|s_M-s_1|}{M-1}\frac{p_1+p_M}{2} \tag{15}$$

The second term of (15) equals the arithmetic mean multiplied by $(p_1 + p_M)/2$, which normalizes the weight sum to one. Substituting (15) into (11), we obtain

$$H_\Delta(X) = -\sum_{i=1}^{M} p_i \log p_i + \log\left(\sum_{i=1}^{M-1}|s_{i+1}-s_i|\frac{p_i+p_{i+1}}{2} + \frac{|s_M-s_1|}{M-1}\frac{p_1+p_M}{2}\right) \tag{16}$$

The above Δ-entropy can be immediately extended to the infinite value-set case, that is

$$H_\Delta(X) = -\sum_{i=-\infty}^{\infty} p_i \log p_i + \log\left(\sum_{i=-\infty}^{\infty}|s_{i+1}-s_i|\frac{p_i+p_{i+1}}{2} + \lim_{N\to\infty}\frac{|s_N-s_{-N}|}{2N}\frac{p_{-N}+p_N}{2}\right) \tag{17}$$

In the rest of the paper, we use (16) and (17) as the Δ-entropy expression, which has strong links with the differential entropy, and reduces to the traditional discrete entropy under an obvious condition. In the next section, we explore some important properties of the Δ-entropy.

3. Some important properties

As discussed in the previous section, the definition of Δ-entropy is enlightened by the connection between the differential entropy and its quantized discrete entropy, so Δ-entropy should maintain a close connection to the differential entropy. We first show that the Δ-entropy and the differential entropy have the following relationship in the limit:

Theorem 1. *For any continuous random variable X with Riemann integrable PDF f(x), we have* $\lim_{\Delta\to 0} H_\Delta(X^\Delta) = h(X)$, *where the quantized discrete variable* X^Δ *is given by (4).*

Proof. See Appendix A. □

Remark 1. By Theorem 1, the differential entropy of X is the limit of the Δ-entropy of X^Δ as $\Delta \to 0$. To some extent, we can regard the Δ-entropy as a "quantized version" of the differential entropy.

Theorem 2

$$\log\left(\max_{j=1,2,\ldots,M-1}|s_{j+1}-s_j|\right) \geqslant H_\Delta(X) - H(X) \geqslant \log\left(\min_{j=1,2,\ldots,M-1}|s_{j+1}-s_j|\right).$$

Proof. Omitted due to simplicity. □

Remark 2. An appealing feature of Theorem 2 is that, if the minimum interval between two successive discrete values is larger than one, we have $H_\Delta(X) > H(X)$, whereas if the maximum interval between two successive discrete values is smaller than one, we have $H_\Delta(X) < H(X)$.

Theorem 3. *If X is a discrete random variable with equally spaced values, i.e. $\forall i,\ 1 \leqslant i \leqslant M-1,\ |s_{i+1}-s_i| \equiv \Delta$, and $\Delta = 1$, then $H_\Delta(X) = H(X)$.*

Proof. For equally spaced intervals, the difference between the Δ-entropy and the discrete entropy equals $\log\Delta$. Hence, the statement follows directly. □

Now we can understand why we chose the name Δ-entropy for the new measure: when the scale goes to zero we end up with differential entropy and when the scale defaults to the natural numbers the measure is indistinguishable from discrete entropy. Classification is a typical example of the error variable distributed on equally spaced values $\{0, 1, 2, 3, \ldots\}$. Therefore, in classification, the error's discrete entropy is equivalent to the Δ-entropy. This fact also gives an interpretation for why the discrete entropy can be successfully used in the test and classification problems [14,31].

We can use Theorem 2 and the bound of the discrete entropy to obtain a bound on the Δ-entropy.

Corollary 1

$$\log\left(\min_{j=1,2,\ldots,M-1}|s_{j+1}-s_j|\right) \leqslant H_\Delta(X) \leqslant \frac{1}{2}\log\left(2\pi e\left(\sum_{i=1}^{M}p_i i^2-\left(\sum_{i=1}^{M}ip_i\right)+\frac{1}{12}\right)\left(\max_{j=1,2,\ldots,M-1}|s_{j+1}-s_j|\right)^2\right).$$

Proof. In information theory, it has been proved that (see [4, p. 489])

$$0 \leqslant H(X) \leqslant \frac{1}{2}\log\left(2\pi e\left(\sum_{i=1}^{M}p_i i^2-\left(\sum_{i=1}^{M}ip_i\right)+\frac{1}{12}\right)\right) \tag{18}$$

Combining the above result and Theorem 2, the corollary is proved. □

The lower bound of the Δ-entropy can also be expressed in term of the variance $Var(X)$, as given in the following theorem.

Theorem 4. *If $p_{min} = min\{p_i\} > 0$, then $H_\Delta(X) \geqslant \log\left(\frac{2Mp_{\min}}{M-1}\right)+\frac{1}{2}\log(Var(X))$.*

Proof. See Appendix B. □

The above lower bound confirms the fact that minimizing the Δ-entropy will constrain the variance. This is a key difference between the Δ-entropy and the classical discrete entropy.

Theorem 5. *For any discrete random variable X, $\forall c \in \mathbb{R}$, $H_\Delta(X+c) = H_\Delta(X)$.*

Proof. Since $H(X+c) = H(X)$, and $\Delta(X+c) = \Delta(X)$, we have $H_\Delta(X+c) = H_\Delta(X)$. □

Theorem 6

$$\forall\alpha \in \mathbb{R},\quad \alpha \neq 0,\quad H_\Delta(\alpha X) = H_\Delta(X) + \log|\alpha|.$$

Proof. Since $H(\alpha X) = H(X)$, and $\Delta(\alpha X) = |\alpha|\Delta(X)$, we have $H_\Delta(\alpha X) = H_\Delta(X) + \log|\alpha|$. □

Remark 3. Theorems 5 and 6 indicate that the Δ-entropy has the same shifting and scaling properties of differential entropy.

Theorem 7. *The Δ-entropy is a concave function of $\boldsymbol{P} = (p_1, p_2, \ldots, p_M)$.*

Proof. See Appendix C. □

Remark 4. The concavity of the Δ-entropy is a desirable property for the entropy optimization problem. Specifically, this property ensures that when a stationary value of the Δ-entropy subject to linear constraints is found, it gives the global maximum value [15].

Now we solve the maximum Δ-entropy distribution. Consider the constrained optimization problem:

$$\begin{cases}\max\limits_{\boldsymbol{P}} H_\Delta(X)\\ \text{s.t.}\begin{cases}\sum\limits_{i=1}^{M}p_i = 1\\ \sum\limits_{i=1}^{M}p_i g_k(s_i) = a_k,\quad k = 1,2,\ldots,K\end{cases}\end{cases} \tag{19}$$

in which a_k is the expected value of function $g_k(X)$. The Lagrangian is given by

$$L = H_{\Delta}(X) - (\lambda_0 - 1)\left(\sum_{i=1}^{M} p_i - 1\right) - \sum_{k=1}^{K} \lambda_k \left(\sum_{i=1}^{M} p_i g_k(s_i) - a_k\right) \tag{20}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_K$ are the $(K+1)$ Lagrange multipliers corresponding to the $(K+1)$ constraints. Here $\lambda_0 - 1$ is used as the first Lagrange multiplier instead of λ_0 as a matter of convenience. Let $\partial L/\partial p_i = 0$, we have

$$\Delta(X)\left(-\lambda_0 - \sum_{k=1}^{K} \lambda_k g_k(s_i) - \log p_i\right) + c_i = 0, \quad i = 1, 2, \ldots, M \tag{21}$$

where

$$c_i = \begin{cases} \frac{|s_M - s_1|}{2(M-1)} + \frac{|s_2 - s_1|}{2}, & i = 1 \\ \frac{|s_{i+1} - s_{i-1}|}{2}, & i = 2, \ldots, M-1 \\ \frac{|s_M - s_1|}{2(M-1)} + \frac{|s_M - s_{M-1}|}{2}, & i = M \end{cases} \tag{22}$$

Solving Eq. (21), we have the theorem:

Theorem 8. *The distribution* $\boldsymbol{P}$ *that maximizes the* Δ*-entropy subject to the constraints of* (19) *is given by*

$$p_i = \exp\left(-\lambda_0 - \sum_{k=1}^{K} \lambda_k g_k(s_i) + \frac{c_i}{\Delta(X)}\right), \quad i = 1, 2, \ldots, M \tag{23}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_K$ *are determined by substituting for* p_i *from* (23) *into the constraints of* (19).

Remark 5. For the case in which the discrete values are equally spaced, we have $c_1 = c_2 = \cdots = c_M = \Delta$, and (23) becomes

$$p_i = \exp\left(\left(1 - \lambda_0 - \sum_{k=1}^{K} \lambda_k g_k(s_i)\right)\right) \tag{24}$$

In this case, the maximum Δ-entropy distribution is identical to the maximum discrete entropy distribution [15].

4. Estimation of Δ-entropy

In practical applications, the discrete values $\{s_i\}$ and probabilities $\{p_i\}$ are usually unknown, so we have to estimate them from sample data $\{x_1, x_2, \ldots, x_n\}$, which is straight forward for the new measure. An immediate approach is to group the sample data into different values $\{\hat{s}_i\}$ and calculate the corresponding relative frequencies $\{\hat{p}_i\}$, given by

$$\hat{p}_i = n_i/n, \quad i = 1, 2, \ldots, M \tag{25}$$

in which n_i denotes the number of these outcomes belonging to the value $\hat{s}_i$, with $\sum_{i=1}^{M} n_i = n$.

Based on the estimated values $\{\hat{s}_i\}$ and probabilities $\{\hat{p}_i\}$, a "plug-in" estimate of Δ-entropy can be obtained as follows:

$$H_{\Delta}\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{P}}\right) = -\sum_{i=1}^{M} \hat{p}_i \log \hat{p}_i + \log\left(\sum_{i=1}^{M-1} |\hat{s}_{i+1} - \hat{s}_i| \frac{\hat{p}_i + \hat{p}_{i+1}}{2} + \frac{|\hat{s}_M - \hat{s}_1|}{M-1} \frac{\hat{p}_1 + \hat{p}_M}{2}\right) \tag{26}$$

where $\widehat{\boldsymbol{S}} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_M)$, and $\widehat{\boldsymbol{P}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_M)$.

For the large sample case, the estimated value set $\widehat{\boldsymbol{S}}$ will match the true value set $\boldsymbol{S}$ with probability one, i.e. $\Pr\left(\widehat{\boldsymbol{S}} = \boldsymbol{S}\right) = 1$, as $n \to \infty$. In fact, assume $\{x_1, x_2, \ldots, x_n\}$ is an i.i.d. sample from the distribution $\boldsymbol{P}$, and $p_i > 0$, $i = 1, \ldots, M$, we have

$$\Pr\left(\widehat{\boldsymbol{S}} \neq \boldsymbol{S}\right) \leqslant \sum_{i=1}^{M} \Pr(s_i \notin \{x_1, x_2, \ldots, x_n\}) = \sum_{i=1}^{M} \left(\prod_{j=1}^{n} \Pr(X_j \neq s_i)\right) = \sum_{i=1}^{M} (1 - p_i)^n \to 0 \text{ as } n \to \infty \tag{27}$$

In the following, we investigate the asymptotic behavior of the Δ-entropy in random sampling. For tractability, we assume the value set $\boldsymbol{S}$ is known (or has been exactly estimated). Following the similar derivation of the asymptotic distribution for the ϕ-entropy (see [24, Chapter 2]), we denote the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{M-1})^T = (p_1, p_2, \ldots, p_{M-1})^T$, and rewrite (26) as

$$\begin{aligned} H_{\Delta}\left(\widehat{\boldsymbol{\theta}}\right) = &-\sum_{i=1}^{M-1} \hat{\theta}_i \log \hat{\theta}_i - \left(1 - \sum_{j=1}^{M-1} \hat{\theta}_j\right) \log\left(1 - \sum_{j=1}^{M-1} \hat{\theta}_j\right) \\ &+ \log\left(\sum_{i=1}^{M-2} |s_{i+1} - s_i| \frac{\hat{\theta}_i + \hat{\theta}_{i+1}}{2} + |s_M - s_{M-1}| \frac{\hat{\theta}_{M-1} + \left(1 - \sum_{j=1}^{M-1} \hat{\theta}_j\right)}{2} + \frac{|s_M - s_1|}{M-1} \frac{\hat{\theta}_1 + \left(1 - \sum_{j=1}^{M-1} \hat{\theta}_j\right)}{2}\right) \end{aligned} \tag{28}$$

The first order Taylor expansion of $H_\Delta\left(\widehat{\boldsymbol{\theta}}\right)$ around $\boldsymbol{\theta}$ gives

$$H_\Delta\left(\widehat{\boldsymbol{\theta}}\right) = H_\Delta(\boldsymbol{\theta}) + \sum_{i=1}^{M-1} \frac{\partial H_\Delta(\boldsymbol{\theta})}{\partial \theta_i}\left(\hat{\theta}_i - \theta_i\right) + o\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|\right) \tag{29}$$

where $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\| = \sqrt{\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)}$, and $\partial H_\Delta(\theta)/\partial\theta_i$ is calculated as

$$\frac{\partial H_\Delta(\theta)}{\partial \theta_i} = \begin{cases} -\log\theta_i + \log\left(1 - \sum_{j=1}^{M-1}\theta_j\right) + \dfrac{s_{i+1} - s_{i-1} - (s_M - s_{M-1})}{2\Delta} - \dfrac{|s_M - s_1|}{2(M-1)\Delta}, & i \neq 1,\ M-1 \\ -\log\theta_1 + \log\left(1 - \sum_{j=1}^{M-1}\theta_j\right) + \dfrac{s_2 - s_1 - (s_M - s_{M-1})}{2\Delta}, & i = 1 \\ -\log\theta_{M-1} + \log\left(1 - \sum_{j=1}^{M-1}\theta_j\right) + \dfrac{s_{M-1} - s_{M-2}}{2\Delta} - \dfrac{|s_M - s_1|}{2(M-1)\Delta}, & i = M-1 \end{cases} \tag{30}$$

in which

$$\Delta = \left(\sum_{i=1}^{M-2}|s_{i+1} - s_i|\frac{\hat{\theta}_i + \hat{\theta}_{i+1}}{2} + |s_M - s_{M-1}|\frac{\hat{\theta}_{M-1} + \left(1 - \sum_{j=1}^{M-1}\hat{\theta}_j\right)}{2} + \frac{|s_M - s_1|}{M-1}\frac{\hat{\theta}_1 + \left(1 - \sum_{j=1}^{M-1}\hat{\theta}_j\right)}{2}\right) \tag{31}$$

According to [24, Chapter 2], we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \underset{n\to\infty}{\overset{L}{\to}} N\left(0, I_F(\theta)^{-1}\right) \tag{32}$$

where the inverse of the Fisher information matrix of $\boldsymbol{\theta}$ is given by $I_F(\boldsymbol{\theta})^{-1} = diag(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T$. It follows that $\sqrt{n}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|$ is bounded in probability, and

$$\sqrt{n}\left(o\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|\right)\right) \underset{n\to\infty}{\overset{P}{\to}} 0 \tag{33}$$

Therefore, the random variable $\sqrt{n}\left(H_\Delta\left(\widehat{\boldsymbol{\theta}}\right) - H_\Delta(\theta)\right)$ and $\sqrt{n}\sum_{i=1}^{M-1}\frac{\partial H_\Delta(\theta)}{\partial\theta_i}\left(\hat{\theta}_i - \theta_i\right)$ have the same asymptotic distribution. Then the following theorem holds.

Theorem 9. *The estimate $H_\Delta\left(\boldsymbol{S}, \widehat{\boldsymbol{P}}\right)$, obtained by replacing the $\{p_i\}$ by their relative frequencies $\{\hat{p}_i\}$, in a random sample of size n, satisfies*

$$\sqrt{n}\left(H_\Delta\left(\boldsymbol{S}, \widehat{\boldsymbol{P}}\right) - H_\Delta(\boldsymbol{S}, \boldsymbol{P})\right) \underset{n\to\infty}{\overset{L}{\to}} N\left(0, \boldsymbol{U}^{\boldsymbol{T}} I_F(\boldsymbol{\theta})^{-1}\boldsymbol{U}\right) \tag{34}$$

provided $\boldsymbol{U}^{\boldsymbol{T}} I_F(\boldsymbol{\theta})^{-1}\boldsymbol{U} > 0$, where $\boldsymbol{\theta} = (p_1, p_2, \ldots, p_{M-1})^T$, $I_F(\boldsymbol{\theta})^{-1} = diag(\theta) - \boldsymbol{\theta}\boldsymbol{\theta}^T$, and

$$\boldsymbol{U} = \left(\partial H_\Delta(\boldsymbol{\theta})/\partial\theta_1, \partial H_\Delta(\boldsymbol{\theta})/\partial\theta_2, \ldots, \partial H_\Delta(\boldsymbol{\theta})/\partial\theta_{M-1}\right)^T \tag{35}$$

where $\partial H_{\boldsymbol{\Delta}}(\boldsymbol{\theta})/\partial\theta_i$ is calculated as (30).

Theorem 9 suggests that the Δ-entropy can be effectively estimated from random samples by replacing the probabilities $\{p_i\}$ by their relative frequencies $\{\hat{p}_i\}$.

It can also be shown that the plug-in estimate of the Δ-entropy has close relationships with certain estimates of the differential entropy.

4.1. Relation to differential entropy estimate based on kernel density estimation

Assume $\{x_1, x_2, \ldots, x_n\}$ are samples from a discrete random variable X, we rewrite the plug-in estimate (26) as

$$H_\Delta\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{P}}\right) = -\sum_{i=1}^{M}\hat{p}_i\log\hat{p}_i + \log\hat{\Delta} \tag{36}$$

where $\hat{\Delta} = \sum_{i=1}^{M-1}|\hat{s}_{i+1} - \hat{s}_i|\frac{\hat{p}_i + \hat{p}_{i+1}}{2} + \frac{|\hat{s}_M - \hat{s}_1|}{M-1}\frac{\hat{p}_1 + \hat{p}_M}{2}$.

Denote $\hat{\Delta}_{\min} = \min_{i=1,\ldots,M-1}|\hat{s}_{i+1} - \hat{s}_i|$, and let $\tau = \hat{\Delta}/\Delta_{\min}$, we construct another set of samples:

$$\left\{x'_1, x'_2, \ldots, x'_n\right\} = \{\tau x_1, \tau x_2, \ldots, \tau x_n\} \tag{37}$$

which can be regarded as the samples from discrete random variable τX. It follows easily that

$$\forall x'_i \neq x'_j, \quad \left|x'_i - x'_j\right| \geqslant \hat{\Delta} \tag{38}$$

We now consider $\{x'_1, x'_2, \ldots, x'_n\}$ as samples from a "continuous" random variable X'. The PDF of X' can be estimated by the kernel approach [5]

$$\hat{p}(x') = \frac{1}{n}\sum_{i=1}^{n} K(x' - x'_i) \tag{39}$$

The kernel function $K: \mathbb{R} \to [0,\infty)$ satisfies $K \geqslant 0$ and $\int_{-\infty}^{\infty} K(x)dx = 1$. Here we use the following uniform kernel:

$$K_{\hat{\Delta}}(x) = \begin{cases} 1/\hat{\Delta}, & x \in \left[-\hat{\Delta}/2, \hat{\Delta}/2\right] \\ 0 & \text{otherwise} \end{cases} \tag{40}$$

Then the kernel density estimation (KDE) of (39) becomes

$$\hat{p}(x') = \frac{1}{n}\sum_{j=1}^{n} K_{\hat{\Delta}}(x' - x'_j) = \frac{1}{n}\sum_{i=1}^{M} n_i K_{\hat{\Delta}}(x' - s'_i) \overset{(a)}{=} \begin{cases} \frac{p_i}{\hat{\Delta}}, & x' \in \left[s'_i - \hat{\Delta}/2, s'_i + \hat{\Delta}/2\right] \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

where (a) follows from $\hat{p}_i = n_i/n$, and $\forall x'_i \neq x'_j, \left|x'_i - x'_j\right| \geqslant \hat{\Delta}$. The differential entropy of X' can be estimated by the plug-in approach:

$$\begin{aligned} \hat{h}(X') &= -\int_{-\infty}^{\infty} \hat{p}(x')\log\hat{p}(x')dx' = -\sum_{i=1}^{M}\int_{s'_i-\hat{\Delta}/2}^{s'_i+\hat{\Delta}/2} \hat{p}(x')\log\hat{p}(x')dx' = -\sum_{i=1}^{M}\int_{s'_i-\hat{\Delta}/2}^{s'_i+\hat{\Delta}/2} \frac{p_i}{\hat{\Delta}}\log\frac{p_i}{\hat{\Delta}}dx' \\ &= -\sum_{i=1}^{M} p_i \log p_i + \log\hat{\Delta} = H_{\Delta}\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{P}}\right) \end{aligned} \tag{42}$$

As a result, the plug-in estimate of the Δ-entropy equals a uniform kernel based estimate for the differential entropy from the scaled samples (37).

4.2. Relation to differential entropy estimate based on sample-spacing

It is also interesting to note that the plug-in estimate of the Δ-entropy has a close connection with the sample-spacing estimate [23,34] of differential entropy. Suppose the sample data are different from each other, and have been rearranged in an increasing order: $\{x_1 < x_2 < \cdots < x_n\}$, the m-spacing estimate is given by [23]

$$\hat{h}_m(X) = \frac{1}{n}\sum_{i=1}^{n-m} \log\left(\frac{n}{m}(x_{i+m} - x_i)\right) \tag{43}$$

where $m \in \mathbb{N}$, and $m < n$. Let $m = 1$, we obtain the *1*-spacing estimate:

$$\hat{h}_1(X) = \frac{1}{n}\sum_{i=1}^{n-1} \log\left(n(x_{i+1} - x_i)\right) \tag{44}$$

In general, regarding $\{x_1, x_2, \ldots, x_n\}$ as samples from a discrete distribution, we estimate the value set and probabilities as

$$\begin{cases} \widehat{\boldsymbol{S}} = (x_1, x_2, \ldots, x_n) \\ \widehat{\boldsymbol{P}} = (1/n, 1/n, \ldots, 1/n) \end{cases} \tag{45}$$

Then the plug-in estimate of Δ-entropy can be calculated as

$$H_{\Delta}\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{P}}\right) = -\sum_{i=1}^{n}\frac{1}{n}\log\frac{1}{n} + \log\left(\sum_{i=1}^{n-1}(x_{i+1} - x_i)\frac{1}{n} + \frac{(x_n - x_1)}{n-1}\frac{1}{n}\right) \tag{46}$$

It follows that

$$\begin{aligned} H_{\Delta}\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{P}}\right) &= -\sum_{i=1}^{n}\frac{1}{n}\log\frac{1}{n} + \log\left(\sum_{i=1}^{n-1}(x_{i+1} - x_i)\frac{1}{n} + \frac{(x_n - x_1)}{n-1}\frac{1}{n}\right) \overset{(a)}{\geqslant} \frac{1}{n}\sum_{i=1}^{n}\log n \\ &\quad + \frac{1}{n}\left(\sum_{i=1}^{n-1}\log(x_{i+1} - x_i) + \log\frac{(x_n - x_1)}{n-1}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n-1}\log\left(n(x_{i+1} - x_i)\right) + \frac{1}{n}\log\frac{n(x_n - x_1)}{n-1} = \hat{h}_1(X) + \frac{1}{n}\log\frac{n(x_n - x_1)}{n-1} \end{aligned} \tag{47}$$

where (a) follows from the concavity of the logarithm function. If $\{x_i\}$ is bounded, we have

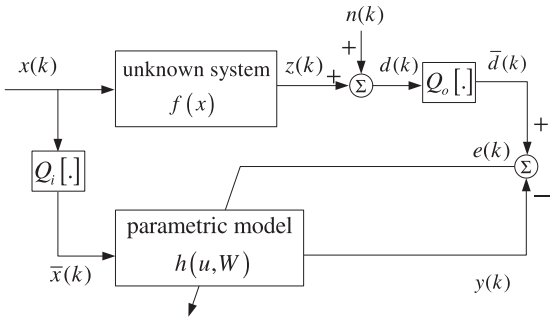

Fig. 2. System identification with quantized I/O data.

$$\lim_{n\to\infty} H_{\Delta}\left(\widehat{\boldsymbol{S}},\widehat{\boldsymbol{P}}\right) \geqslant \lim_{n\to\infty}\left(\hat{h}_1(X)+\frac{1}{n}\log\frac{n(x_n-x_1)}{n-1}\right)=\lim_{n\to\infty}\hat{h}_1(X) \tag{48}$$

In this case, the plug-in estimate of Δ-entropy provides an asymptotic upper bound to the *1*-spacing entropy estimate.

5. Application to system identification with quantized data

As Δ-entropy measures both the probabilistic uncertainty and the dispersion of a discrete random variable, it can be used as an optimality criterion in system identification where the error signal is distributed on an unknown (and varying with iterations) countable value set.

Consider the system identification scheme with quantized I/O data, as shown in Fig. 2, in which $f(\cdot)$ and $h(\cdot,W)$ denote the unknown system and the parametric model respectively, $W=[w_1,\ldots,w_d]^T$ is the d-dimensional parameter vector (weight vector) of the model. $x(k)$ and $z(k)$ are the actual input and output signals (usually continuous-valued) of the unknown system at k time; $n(k)$ is the additive noise and $d(k)$ the noisy output. $\bar{x}(k)$ and $\bar{d}(k)$ represent the quantized I/O observations, which are obtained via quantizers Q_i and Q_o. With uniform quantization box-sizes q_i and q_o, $\bar{x}(k)$ and $\bar{d}(k)$ are given by

$$\begin{cases}\bar{x}(k)=\lceil x(k)/q_i+1/2\rceil\times q_i\\ \bar{d}(k)=\lceil d(k)/q_o+1/2\rceil\times q_o\end{cases} \tag{49}$$

where $\lceil x\rceil$ gives the largest integer that is less than or equal to x.

The error signal $e(k)$ is calculated as

$$e(k)=\bar{d}(k)-y(k) \tag{50}$$

where $y(k)$ is the output of the model driven by quantized data $\bar{x}(k)$. Now the problem is to identify the unknown system using the quantized I/O data $\{(\bar{x}(k),\bar{d}(k)),\ k=1,2,\ldots,L\}$.

In the above identification setting, the error $e(k)$ will be discrete-valued, and hence the Δ-entropy can be used as the training cost. The optimum parameters of the model would be

$$W_{opt}=\arg\min_{W\in\boldsymbol{W}} H_{\Delta}(e) \tag{51}$$

where $\boldsymbol{W}\subset\mathbb{R}^d$ denotes the parameter space. That is, we propose to determine the model parameters by minimizing the Δ-entropy of the discrete-valued error residuals.[2]

In a practical application, Δ-entropy cannot be in general analytically computed, since the error's values and corresponding probabilities are unknown. We need to estimate the Δ-entropy by the plug-in method as discussed in the previous section. Therefore, the optimization of (51) becomes

$$W_{opt}=\arg\min_{W\in\boldsymbol{W}} H_{\Delta}\left(\widehat{\boldsymbol{S}}^{(e)},\widehat{\boldsymbol{P}}^{(e)}\right) \tag{52}$$

where $\widehat{\boldsymbol{S}}^{(e)}=\left(\hat{s}_1^{(e)},\hat{s}_2^{(e)},\ldots,\hat{s}_M^{(e)}\right)$ and $\widehat{\boldsymbol{P}}^{(e)}=\left(\hat{p}_1^{(e)},\hat{p}_2^{(e)},\ldots,\hat{p}_M^{(e)}\right)$ denote the estimated value-set of the error and the corresponding relative frequencies.

The classical gradient based methods cannot be used to solve the optimization problem of (52), since the objective function $H_{\Delta}\left(\widehat{\boldsymbol{S}}^{(e)},\widehat{\boldsymbol{P}}^{(e)}\right)$ is usually not differentiable. So we have to resort to other methods such as evolutionary algorithms (EAs), although they are usually more computationally complex. Here we adopt the estimation of distribution algorithms (EDAs)

[2] For the case in which the underlying distribution of the error residual is continuous, we can still use the Δ-entropy as the optimization criterion if we classify the errors into groups and obtain the quantized error data.

[16], which is a new class of EAs. The EDAs use the probability model built from the objective function to generate the promising search points instead of crossover and mutation as done in traditional genetic algorithms (GAs). Compared with other EAs, the EDAs may achieve better evolutionary performances [12,19,30]. Some theoretical results related to the convergence and time complexity of the EDAs can be found in [3,11,29,39].

Based on Delta-entropy and EDAs, we propose the parameter search algorithm as summarized in Table 1.

Usually, we use a Gaussian model with diagonal covariance matrix (GM/DCM) [16] to estimate the density function $f_g(W)$ of the gth generation. With GM/DCM model, we have

$$f_g(W) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j^{(g)}} \exp\left(-\left(w_j - \mu_j^{(g)}\right)^2 \Big/ \left(2\left(\sigma_j^{(g)}\right)^2\right)\right) \tag{53}$$

where the means $\mu_j^{(g)}$ and the deviations $\sigma_j^{(g)}$ can be estimated as

$$\begin{cases} \mu_j^{(g)} = \frac{1}{N}\sum_{l=1}^{N} W_{B(l)}^{(g-1)}(j) \\ \sigma_j^{(g)} = \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(W_{B(l)}^{(g-1)}(j) - \mu_j^{(g)}\right)^2} \end{cases} \tag{54}$$

We now perform a series of Monte-Carlo simulations of system identification based on quantized I/O data to demonstrate the performance of EDAs with Δ-entropy criterion. In all of the experiments below, we set the quantization box-sizes $q_i = q_o = q$. For the EDAs, we set $R = 100$ and $N = 30$.

5.1. Comparison with other optimality criteria

We will contrast the performances of Δ-entropy criterion and several other optimality criteria by using EDAs as parameter search algorithm. Consider the linear system identification case, in which we assume the unknown system and the parametric model are both two-tap FIR filters, that is

$$\begin{cases} z(k) = w_1^* x(k) + w_2^* x(k-1) \\ y(k) = w_1 \bar{x}(k) + w_2 \bar{x}(k-1) \end{cases} \tag{55}$$

Here we set the true weight vector of the unknown system $W^* = [1.0, 0.5]^T$, and the initial weight vector of FIR model $W(0) = [0, 0]^T$. In addition, we assume the input signal $\{x(k)\}$ and the additive noise $\{n(k)\}$ are both white Gaussian processes with variances 1.0 and 0.04, respectively. The length of training data is $L = 500$.

First, we compare the performances of three entropy criteria: Δ-entropy, differential entropy[3] and discrete entropy, using the same adaptation method. For different entropy criteria and different quantization box-sizes, the average evolution curves of weight error norm ($\|W^* - W\|$) over 100 Monte Carlo runs are shown in Fig. 3. Evidently, the Δ-entropy criterion achieves the best performance with the fastest convergence speed and the smallest error (weight error norm) at the final stage of learning. Further, we have two observations: (1) the discrete entropy criterion fails to converge as seen from the evolution of the learning curves, which agrees with the fact that the discrete entropy cannot constrain the error's dispersion; (2) the performance (especially final bias) of the differential entropy approaches that of the Δ-entropy when the quantization box-size q becomes smaller. This result agrees with the limiting relationship between the Δ-entropy and differential entropy.

In order to compare the error dispersions, we plot in Fig. 4 the probability mass functions of the error samples for different entropy criteria ($q = 1.0$). Here the error samples are obtained using 500 test data after training the model. From Fig. 4, it is clear that the Δ-entropy produces the most concentrated error samples, while the discrete entropy yields errors with the largest dispersion. The average intervals (calculated as (15)) of error samples for the Δ-entropy, differential entropy and discrete entropy are 0.0068, 0.0097 and 0.0622, respectively.

The performance of Δ-entropy criterion is also compared with that of the MSE criterion, with the same experimental setup. As one can see in Fig. 5, the Δ-entropy achieves again significantly smaller bias at the final stage of evolution, although its convergence speed is slightly slower than that of the MSE criterion.

5.2. Comparison with LMS and SIG algorithms

In the following we will contrast the performances of the Δ-entropy based EDA, LMS, and SIG (stochastic information gradient) algorithm [9] in a 4-tap FIR identification problem. The LMS and SIG are stochastic gradient based algorithms under MSE and MEE criterion respectively, so they both display a performance penalty called the misadjustment that is proportional to the stepsize utilized, while EDA is immune to this phenomenon. So the difference in algorithm performance cannot be only allotted to the difference in cost functions. The true weight vector of the unknown system is set as $W^* = [-0.2, 0.9, 0.7, -0.5]^T$. The input signal $\{x(k)\}$ is white and uniformly distributed on the interval $[-2, 2]$ with data length $L = 2000$. The

[3] Strictly speaking, the differential entropy criterion is invalid in this example, because the observed error is a discrete random variable. However, in the simulation, we can still use the m-spacing estimator of differential entropy.

Table 1
EDAs based parameter search algorithm with Δ-entropy criterion.

1. BEGIN
2. Generate R individuals $A_0 = \left\{W_1^{(0)}, W_2^{(0)}, \ldots, W_R^{(0)}\right\}$ randomly from parameter space $\boldsymbol{W}$, $g \leftarrow 0$
3. WHILE the final stopping criterion is not met DO
4. $g \leftarrow g + 1$
5. For each parameter vector in A_{g-1}, estimate the error's Δ-entropy using a training data set $\{(\bar{x}(k), \bar{d}(k))\}$
6. Select $N (N \leqslant R)$ promising individuals $B_g = \left\{W_{B(1)}^{(g-1)}, W_{B(2)}^{(g-1)}, \ldots, W_{B(N)}^{(g-1)}\right\}$ from A_{g-1} according to the truncation selection method (using Δ-entropy as the fitness function)[a]
7. Estimate the probability density function $f_g(W)$ based on the statistical information extracted from the selected N individuals B_g
8. Sample R individuals $A_g = \left\{W_1^{(g)}, W_2^{(g)}, \ldots, W_R^{(g)}\right\}$ from $f_g(W)$
9. END WHILE
10. Calculate the estimated parameter: $W(g) = (1/N)\sum_{n=1}^{N} W_{B(n)}^{(g-1)}$
11. END

[a] The truncation selection is a widely-used selection method in EDAs. In the truncation selection, individuals are sorted according to their objective function (or fitness function) values and only the best individuals are selected.

additive noise $\{n(k)\}$ is a white Gaussian process with variance $\sigma_n^2 = 0.01$ or 0.25. In the simulations, the step-sizes (and the kernel size for SIG) of the two gradient algorithms are adjusted so as to produce the least weight error norm. Fig. 6 shows the histograms of error norm at the final stage of system training based on 100 Monte Carlo experiments. The inset plots in Fig. 6 give the summary of the mean and the spread of the histograms for each algorithm. Clearly, the EDA (with Δ-entropy criterion) outperforms both the LMS and SIG algorithms in terms of final bias, especially for the case of coarser quantization[4] ($q = 1.0$). It should be noted that although the EDA (with Δ-entropy criterion) can achieve significant improvements in the weight bias, it is much more computationally intensive than LMS and SIG algorithms. A runtime comparison shows that, for $q = 1.0$ case, the proposed method took on average 39 s to complete a single Monte Carlo run (with 15 generations), while the LMS and SIG only took 0.02 and 0.18 s, respectively (with 2000 iterations).

5.3. Comparison with TLS, EWC and IV methods

Considering the quantization noises, the system identification scheme of Fig. 2 is actually an errors-in-variables (EIV) identification problem where both input and output are corrupted by noises [32]. In this example, we will compare the proposed method with the total least squares (TLS), error whitening criterion (EWC) [28] and the instrumental variables (IV) method, which under certain conditions give an unbiased estimate in identification of the EIV model. Assume the unknown system is a 6-tap FIR system with $W^* = [-0.6, 0.4, 0.8, 0.6, 0.3, -0.5]^T$. The input signal is a first-order AR process given by

$$x(k) = 0.2x(k-1) + \varepsilon(k) \tag{56}$$

where $\varepsilon(k)$ is a white Gaussian noise with unit variance. The additive noise $\{n(k)\}$ is a white Gaussian process with variance $\sigma_n^2 = 0.04$. Further, the quantization box-size is $q = 1.0$ and the length of training data $L = 2000$. For the TLS, we use the recursive algorithm derived in [9] to obtain the solution. For the EWC, we adopt the stochastic gradient based algorithm, i.e. EWC–LMS [28]. For the IV method, we choose the delayed input vector as the instrument. The histograms of weight error norm for 100 Monte Carlo simulations are shown in Fig. 7, from which we see that the new approach outperforms the EWC and IV methods and achieves estimation accuracy comparable to that of TLS which, however, requires some *a priori* information on the noise statistics.

5.4. Nonlinear system identification case

Consider a nonlinear system identification case, in which we assume the unknown system and parametric model are both second-order polynomial basis function networks (second-order Volterra systems with 2-sample memory), that is

$$\begin{cases} z(k) = w_1^* + w_2^* x(k) + w_3^* x(k-1) + w_4^* x^2(k) + w_5^* x(k)x(k-1) + w_6^* x^2(k-1) \\ y(k) = w_1 + w_2 \bar{x}(k) + w_3 \bar{x}(k-1) + w_4 \bar{x}^2(k) + w_5 \bar{x}(k)\bar{x}(k-1) + w_6 \bar{x}^2(k-1) \end{cases} \tag{57}$$

Let the weight vector of unknown system be $W^* = [0.5, 0.8, -0.6, 0.2, -0.7, 0.1]^T$, and assume that the input signal, additive noise and training data length are the same as the previous example. As Δ-entropy is shift invariant (see Theorem 5), during simulation the bias weight w_1 is adjusted so as to yield zero-mean error.

First we set $q = 0.5$ and compare again the performance of the EDA (with Δ-entropy criterion), LMS and SIG algorithms. The histogram shown in Fig. 8 indicates that the new method still performs best in terms of bias (measured by the weight

[4] The coarse quantization frequently occurs in many practical systems. For example, in the wireless sensor networks (WSN), the cheap sensors might produce very roughly quantized measurements (less than 8 bit) due to energy consumption constraint or very limited data communication rate.

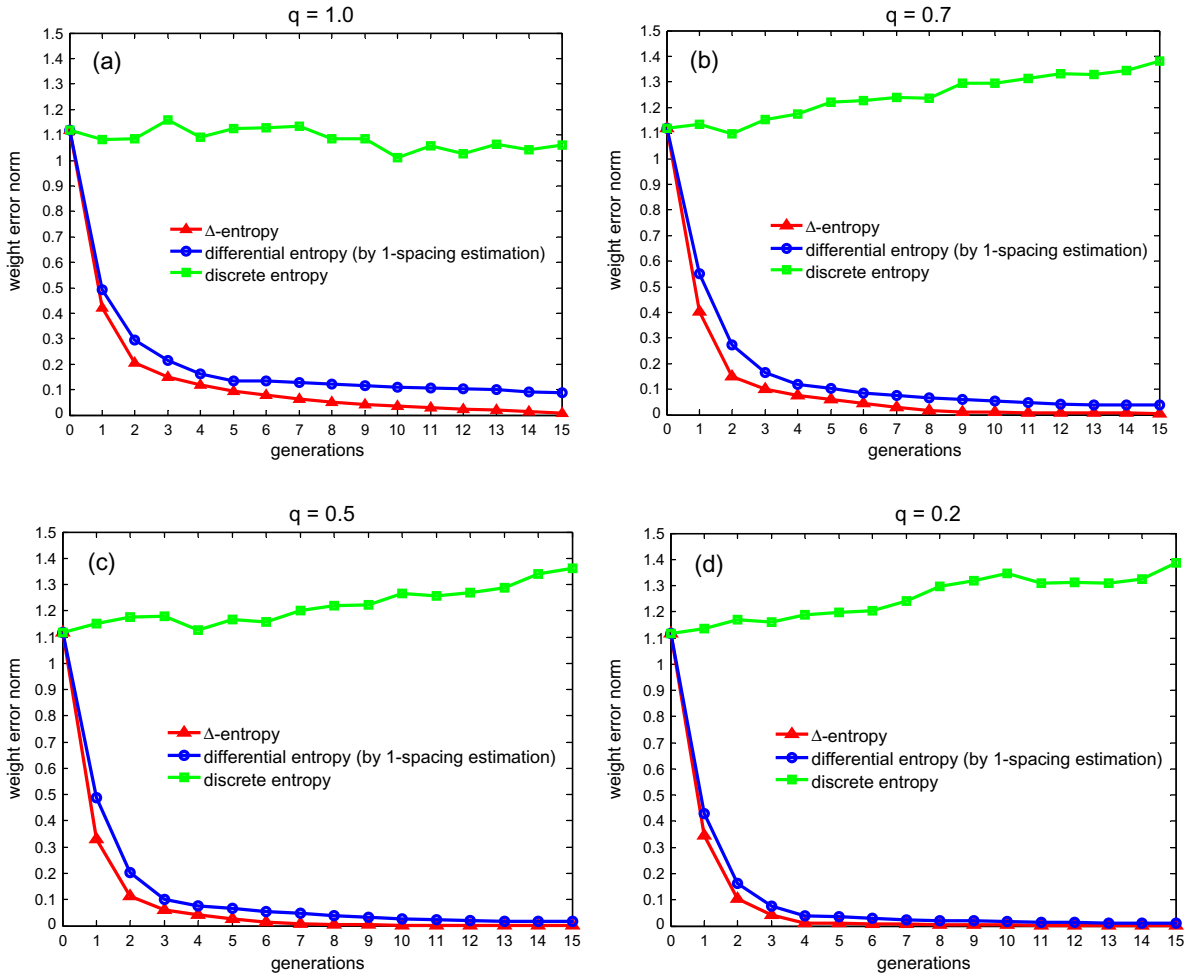

Fig. 3. Evolution curves of weight error norm for different entropy criteria and different quantization box-sizes.

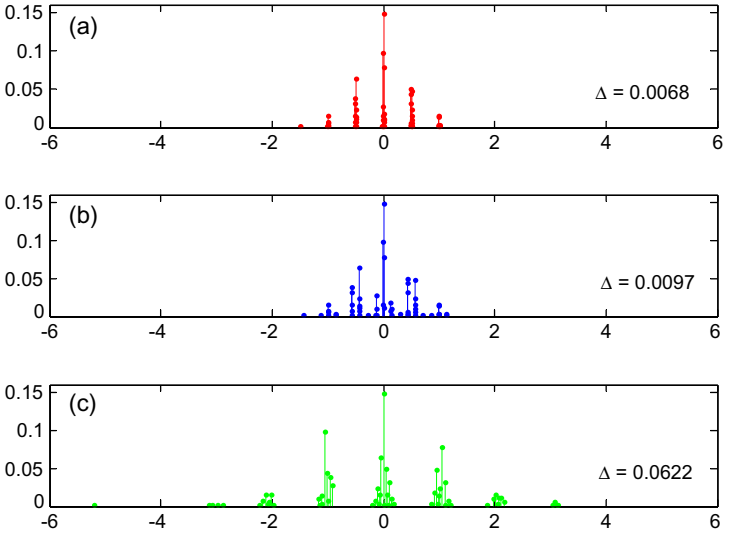

Fig. 4. Probability mass functions of errors for three entropy criteria: (a) Δ-entropy, (b) differential entropy, and (c) discrete entropy.

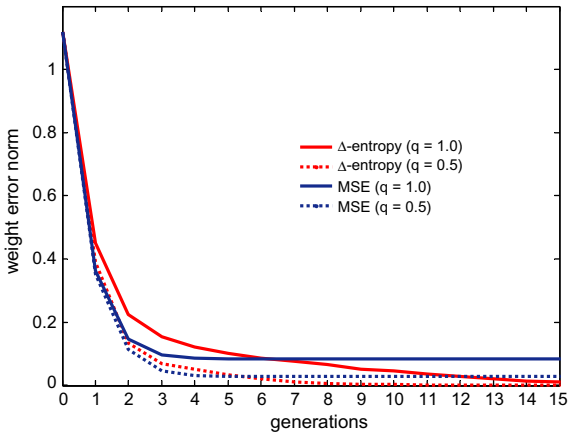

Fig. 5. Evolution curves of weight error norm for Δ-entropy and MSE criterion.

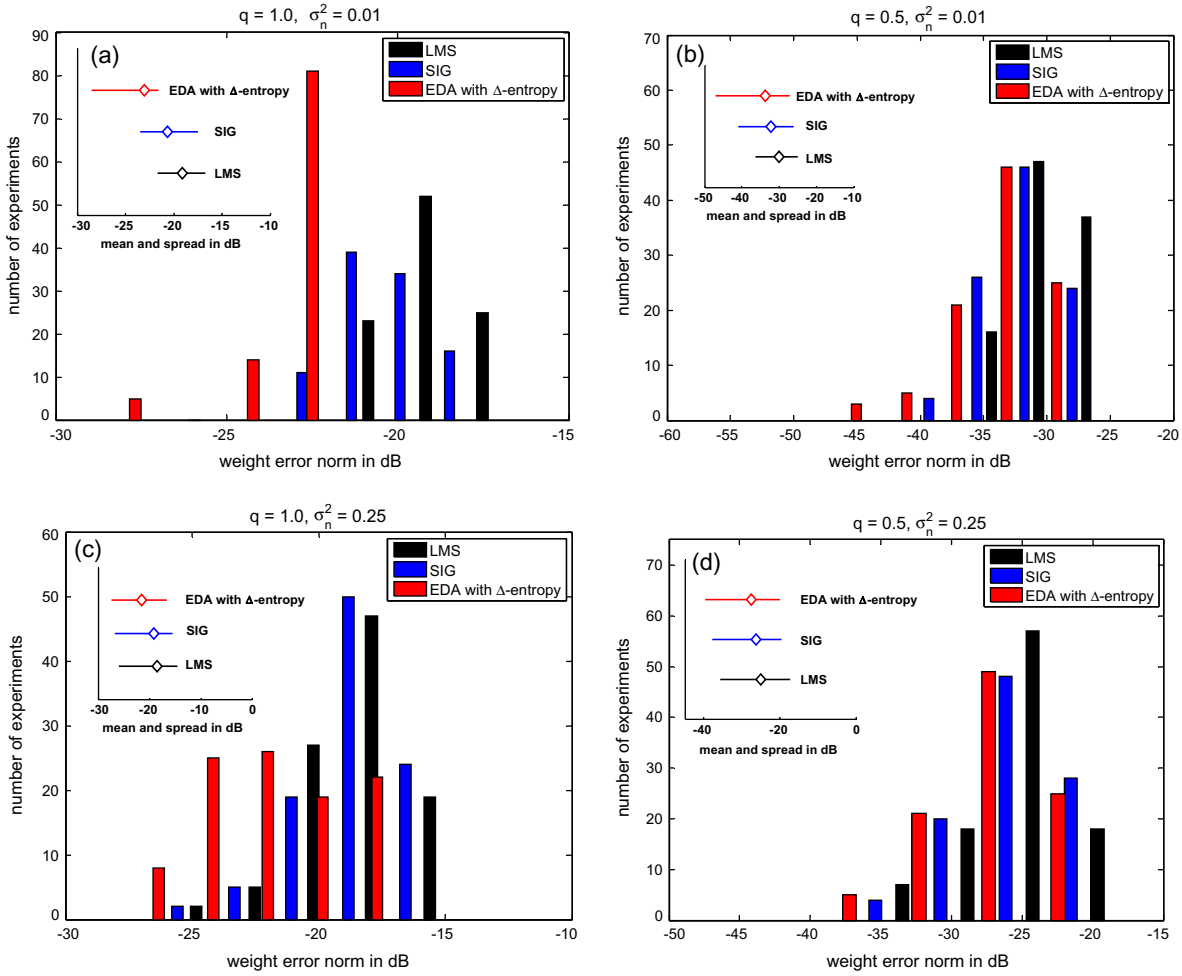

Fig. 6. Histogram plots of the weight error norm for LMS, SIG and the EDA with Δ-entropy criterion.

error norm). This result can also be confirmed by Fig. 9, in which the probability density functions of the intrinsic errors are plotted. Here the intrinsic error is defined as

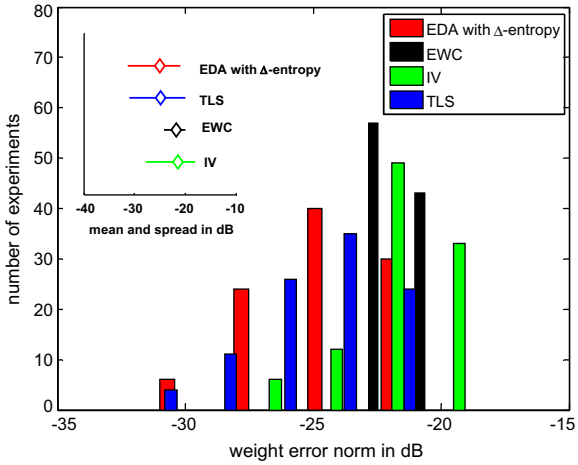

Fig. 7. Histogram plots of the weight error norm for TLS, EWC, IV method and the EDA with Δ-entropy criterion.

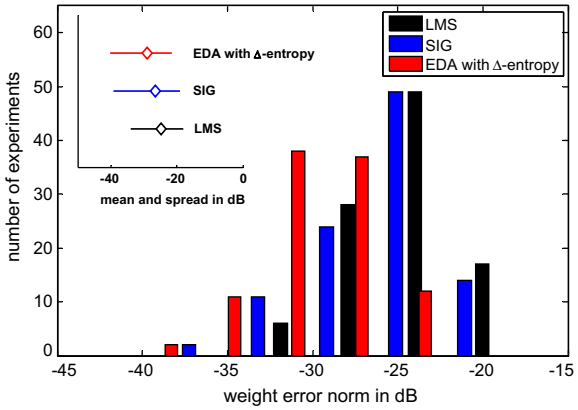

Fig. 8. Histogram plots of the weight error norm for LMS, SIG and the EDA with Δ-entropy criterion (nonlinear system case).

$$\widehat{e}(k) = z(k) - \widehat{y}(k) \tag{58}$$

where $\widehat{y}(k)$ denotes the output of the model (after training) driven by the unquantized input signal $\{x(k)\}$, that is

$$\widehat{y}(k) = w_1 + w_2 x(k) + w_3 x(k-1) + w_4 x^2(k) + w_5 x(k)x(k-1) + w_6 x^2(k-1) \tag{59}$$

As shown in Fig. 9, the EDA (Δ-entropy) produces the largest and most concentrated peak centered at the zero intrinsic error.

In order to further evaluate the performance of the new approach, we also compare it with a recently proposed method by Ozertem and Erdogmus [22] which under certain conditions provides an unbiased estimate of the true parameters of an order-2 Volterra model with noisy input–output measurements. Similar to [22], we use the angle between the estimated and the actual weight vector (coefficient vector) as the performance metric. Fig. 10 shows the angles averaged over 100 Monte Carlo simulations for the two approaches for different quantization box-sizes. One can see clearly that our approach achieves much smaller angles.

Remark 6. Our previous simulations show that the proposed Δ-entropy criterion performs better than the existing criteria like MSE and MEE. One intuitive reason for this is that, minimizing the Δ-entropy will decrease both the probabilistic uncertainty and the average interval, which enforces the residual errors to concentrate. A more rational explanation for this performance improvement is given in Appendix D, wherein we also report some supplementary simulation results.

6. Conclusion remarks

The minimum error entropy (MEE) criterion applied to discrete data minimizes statistical uncertainty in the error PDF but fails to constrain the error's dispersion, which penalizes performance. To address this problem, we develop a new definition

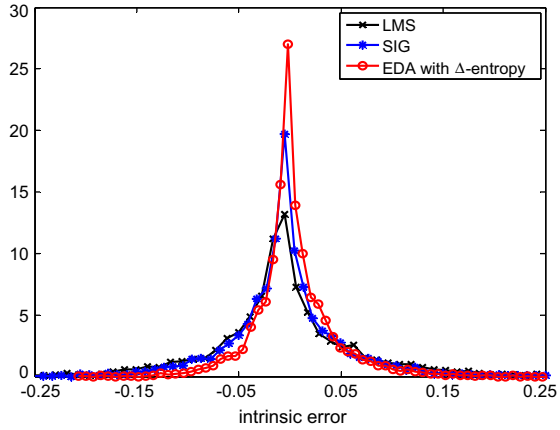

Fig. 9. Probability density functions of the intrinsic errors produced by three algorithms.

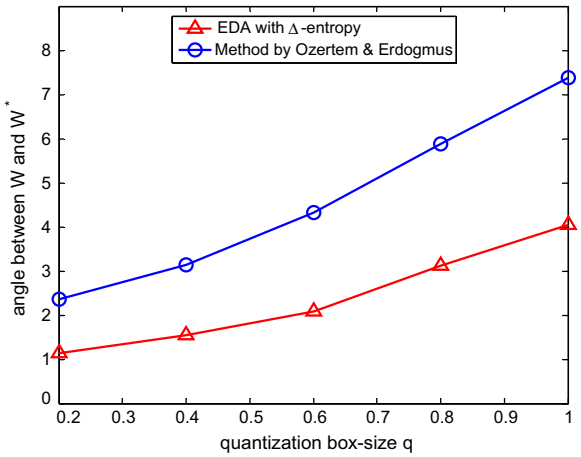

Fig. 10. Angles (degrees) between the estimated and the actual weight vector.

of entropy for discrete random variables, called Δ-entropy. Inspired by the connection between differential and discrete entropies, we define the Δ-entropy as the sum of the discrete entropy and the logarithm of average interval between two successive values, in which the first term measures the probabilistic uncertainty, and the second term measures the dispersion in the error variable.

The Δ-entropy has strong links both with differential entropy and discrete entropy, acting as a bridge between the differential and the discrete entropy, but when evaluated for a given analytic distribution its value will differ from them. However, when entropy is estimated directly from data, it is interesting that the plug-in estimate of the Δ-entropy ties in closely with the kernel based and m-spacing based estimates of the differential entropy. Specifically, the plug-in estimate of the Δ-entropy equals a uniform kernel based entropy estimate from the scaled samples, and establishes an asymptotic upper bound to the 1-spacing entropy estimate. Further work is necessary to establish the role of Δ-entropy as an estimator for differential entropy.

As the Δ-entropy measures both the probabilistic uncertainty and dispersion of the random variable, its role as a cost function in adaptation of linear or nonlinear systems with discretized inputs and desired responses is much clearer and has been validated in this study. At present, however, the gradient based algorithms have not been derived since the objective function is not differentiable. In this work, the estimation of distribution algorithm (EDA) is used as the parameter search algorithm, despite the increase in computational cost. Monte Carlo simulations are performed for the system identification with quantized input–output data. Simulation results confirm that the Δ-entropy criterion may achieve a significant improvement in estimation accuracy.

As a new entropy definition, the Δ-entropy will find applications in many other fields. Potential applications lie particularly in machine learning and signal processing with discrete-valued data. Typical examples are: (1) count data time-series modeling and prediction; (2) channel equalization in digital transmission and (3) blind separation of discrete-valued sources. Exploring these applications will be the goal of future research.

Acknowledgements

This work was partially supported by NSF Grant ECCS 0856441, NSF IIS 0964197 and ONR N00014-10-1-0375, and National Natural Science Foundation of China (No. 60904054), National Key Basic Research and Development Program (973) of China (No. 2009CB724205).

Appendix A

Proof of theorem 1. Combining (6) and (17), we have

$$\begin{aligned} H_{\Delta}(X^{\Delta}) &= -\sum_{i=-\infty}^{\infty} f(s_i)\Delta \log\left(f(s_i)\Delta\right) + \log\left(\sum_{i=-\infty}^{\infty} |s_{i+1}-s_i| \frac{f(s_i)\Delta + f(s_{i+1})\Delta}{2} + \lim_{N\to\infty} \frac{|s_N - s_{-N}|}{2N} \frac{f(s_{-N})\Delta + f(s_N)\Delta}{2}\right) \\ &= -\sum_{i=-\infty}^{\infty} \Delta f(s_i) \log f(s_i) + \log\left(\sum_{i=-\infty}^{\infty} |s_{i+1}-s_i| \frac{f(s_i)+f(s_{i+1})}{2}\right) \end{aligned} \tag{A.1}$$

As $f(x)$ is Riemann integrable, it follows that

$$\lim_{\Delta\to 0} H_{\Delta}(X^{\Delta}) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx + \log\left(\int_{-\infty}^{\infty} f(x)dx\right) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx = h(X) \tag{A.2}$$

which completes the proof. □

Appendix B

Proof of theorem 4. It is easy to derive

$$Var(X) = \sum_{i=1}^{M} (s_i - \bar{s})^2 p_i \leqslant \sum_{i=1}^{M} \left(s_i - \frac{s_M + s_1}{2}\right)^2 p_i \leqslant \sum_{i=1}^{M} \left(s_M - \frac{s_M + s_1}{2}\right)^2 p_i = \frac{1}{4}(s_M - s_1)^2 \tag{B.1}$$

It follows that $|s_M - s_1| \geqslant 2\sqrt{Var(X)}$, and hence

$$\begin{aligned} H_{\Delta}(X) &\geqslant \log\left(\sum_{i=1}^{M-1} |s_{i+1}-s_i| \frac{p_i + p_{i+1}}{2} + \frac{|s_M - s_1|}{M-1} \frac{p_1 + p_M}{2}\right) \geqslant \log\left(\sum_{i=1}^{M-1} |s_{i+1}-s_i| p_{\min} + \frac{|s_M - s_1|}{M-1} p_{\min}\right) \\ &= \log\left(\frac{M|s_M - s_1|}{M-1} p_{\min}\right) \geqslant \log\left(\frac{2M\sqrt{Var(X)}}{M-1} p_{\min}\right) = \log\left(\frac{2Mp_{\min}}{M-1}\right) + \frac{1}{2}\log\left(Var(X)\right) \quad \square \end{aligned} \tag{B.2}$$

Appendix C

Proof of theorem 7. $\forall \boldsymbol{P}_1 = \left(p_1^{(1)}, p_2^{(1)}, \ldots, p_M^{(1)}\right)$, $\boldsymbol{P}_2 = \left(p_1^{(2)}, p_2^{(2)}, \ldots, p_M^{(2)}\right)$, and $\forall 0 \leqslant \lambda \leqslant 1$, we have

$$\Delta(\boldsymbol{S}, \lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2) = \lambda\Delta(\boldsymbol{S}, \boldsymbol{P}_1) + (1-\lambda)\Delta(\boldsymbol{S}, \boldsymbol{P}_2) \tag{C.1}$$

By the concavity of the logarithm function, we get

$$\log\left(\Delta(\boldsymbol{S}, \lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2)\right) \geqslant \lambda\log\left(\Delta(\boldsymbol{S}, \boldsymbol{P}_1)\right) + (1-\lambda)\log\left(\Delta(\boldsymbol{S}, \boldsymbol{P}_2)\right) \tag{C.2}$$

Further, it is well-known the discrete entropy $H(\boldsymbol{P})$ is a concave function of the distribution $\boldsymbol{P}$, that is

$$H(\lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2) \geqslant \lambda H(\boldsymbol{P}_1) + (1-\lambda)H(\boldsymbol{P}_2), \quad \forall 0 \leqslant \lambda \leqslant 1 \tag{C.3}$$

Combining (B.2) and (B.3), we have

$$H_{\Delta}(\boldsymbol{S}, \lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2) \geqslant \lambda H_{\Delta}(\boldsymbol{S}, \boldsymbol{P}_1) + (1-\lambda)H_{\Delta}(\boldsymbol{S}, \boldsymbol{P}_2) \tag{C.4}$$

which implies Δ-entropy is a concave function of $\boldsymbol{P}$. □

Appendix D

D.1. An explanation for the satisfactory performance of the Δ-entropy criterion

Here we give an explanation on why Δ-entropy criterion performs well in system identification. To clarify the analysis, we consider the errors-in-variables (EIV) case in which the input signal $\{x(k)\}$, input noise $\{n_1(k)\}$, and the output noise

$\{n_2(k)\}$ are all discrete-valued data with finite value-sets. Further, we assume the unknown system and the parametric model are both m-tap FIR filters, that is

$$\begin{cases} G^*(z) = \sum_{i=1}^{m} w_i^* z^{-i+1} \\ G(z) = \sum_{i=1}^{m} w_i z^{-i+1} \end{cases} \tag{D.1}$$

where $G*(z)$ and $G(z)$ denote the transfer functions of the unknown system and the model, respectively. In this case, the error signal $e(k)$ is

$$e(k) = \sum_{i=1}^{m} (w_i^* - w_i) x(k-i+1) - \sum_{i=1}^{m} w_i n_1(k-i+1) + n_2(k) = \widehat{e}(k) + \widehat{e}(k) \tag{D.2}$$

where $\widehat{e}(k)$ and $\widehat{n}(k)$ stand for the intrinsic error and the "equivalent output noise", i.e.

$$\begin{cases} \widehat{e}(k) = \sum_{i=1}^{m} (w_i^* - w_i) x(k-i+1) \\ \widehat{n}(k) = -\sum_{i=1}^{m} w_i n_1(k-i+1) + n_2(k) \end{cases} \tag{D.3}$$

Let the value-sets and the probability distributions of $\widehat{e}(k)$ and $\widehat{n}(k)$ be

$$\begin{cases} \boldsymbol{S}^{(\widehat{e})} = \left(s_1^{(\widehat{e})}, s_2^{(\widehat{e})}, \ldots, s_{M_1}^{(\widehat{e})} \right) \\ \boldsymbol{S}^{(\widehat{n})} = \left(s_1^{(\widehat{n})}, s_2^{(\widehat{n})}, \ldots, s_{M_2}^{(\widehat{n})} \right) \end{cases} \tag{D.4}$$

and

$$\begin{cases} \boldsymbol{P}^{(\widehat{e})} = \left(p_1^{(\widehat{e})}, p_2^{(\widehat{e})}, \ldots, p_{M_1}^{(\widehat{e})} \right) \\ \boldsymbol{P}^{(\widehat{n})} = (p_1^{(\widehat{n})}, p_2^{(\widehat{n})}, \ldots, p_{M_2}^{(\widehat{n})}) \end{cases} \tag{D.5}$$

Then we have $\Delta(e(k)) \approx \Delta(\widehat{e}(k))$, provided the following assumptions hold:

Assumption 1

$\widehat{e}(k)$ is independent of $\widehat{n}(k)$.

Assumption 2. $\widehat{e}(k)$ is small enough such that $\left|\widehat{e}(k)\right| < \frac{1}{2}\left(\min_{j=1,2,\ldots,M_2-1} \left| s_{j+1}^{(\widehat{n})} - s_j^{(\widehat{n})} \right| \right)$.

Assumption 3. The probabilities of the extreme values of $\widehat{e}(k)$ are nearly zero, that is, $p_1^{(\widehat{e})} \approx 0$, $p_{M_1}^{(\widehat{e})} \approx 0$.

Assumption 1 will be valid if the input signal $\{x(k)\}$ is independent of the noises $\{n_1(k)\}$ and $\{n_2(k)\}$. Assumption 2 will hold if the model weight vector W is close enough to the true weight vector W^*. Assumption 3 is reasonable for most practical distributions.

Under Assumptions 1 and 2, the value-set and probability distribution of $e(k)$ will be

Table 2
Mean ± deviation results of w_1 and w_2 at the 10th EDA generation for different σ_{n_1} ($\sigma_{n_2} = 0.1$).

σ_{n_1}	Δ-entropy		MSE	
	w_1	w_2	w_1	w_2
0.1	1.0000 ± 0.0008	0.4999 ± 0.0007	0.9896 ± 0.0071	0.4952 ± 0.0067
0.2	0.9997 ± 0.0015	0.4995 ± 0.0015	0.9609 ± 0.0098	0.4800 ± 0.0097
0.3	0.9994 ± 0.0016	0.4993 ± 0.0017	0.9192 ± 0.0118	0.4617 ± 0.0140
0.4	0.9991 ± 0.0019	0.4991 ± 0.0019	0.8629 ± 0.0129	0.4316 ± 0.0163
0.5	0.9980 ± 0.0039	0.4972 ± 0.0077	0.8015 ± 0.0146	0.4016 ± 0.0200

Table 3
Mean ± deviation results of w_1 and w_2 at the 10th EDA generation for different σ_{n_2} ($\sigma_{n_1} = 0.1$).

σ_{n_2}	Δ-entropy		MSE	
	w_1	w_2	w_1	w_2
0.1	1.0000 ± 0.0008	0.4999 ± 0.0007	0.9896 ± 0.0071	0.4952 ± 0.0067
0.2	0.9999 ± 0.0009	0.4999 ± 0.0008	0.9887 ± 0.0096	0.4949 ± 0.0089
0.3	0.9999 ± 0.0012	0.4998 ± 0.0011	0.9908 ± 0.0139	0.4943 ± 0.0142
0.4	0.9999 ± 0.0020	0.4998 ± 0.0017	0.9880 ± 0.0192	0.4948 ± 0.0208
0.5	1.0001 ± 0.0039	0.4998 ± 0.0024	0.9926 ± 0.0231	0.4946 ± 0.0235

$$\begin{cases} \boldsymbol{S}^{(e)} = \left(s_1^{(e)}, s_2^{(e)}, \ldots, s_{M_1M_2}^{(e)}\right) \\ \boldsymbol{P}^{(e)} = \left(p_1^{(e)}, p_2^{(e)}, \ldots, p_{M_1M_2}^{(e)}\right) \end{cases} \tag{D.6}$$

where $s_l^{(e)} = s_j^{(n)} + s_i^{(\widehat{e})}$, $p_l^{(e)} = p_j^{(n)} p_i^{(\widehat{e})}$, in which $j = \lceil (l-1)/M_1 \rceil + 1$, $i = l - (j-1)M_1$.

Thus we have

$$\begin{aligned}
\Delta(e(k)) &= \sum_{l=1}^{M_1M_2-1} \left| s_{l+1}^{(e)} - s_l^{(e)} \right| \frac{p_l^{(e)} + p_{l+1}^{(e)}}{2} + \frac{|s_{M_1M_2}^{(e)} - s_1^{(e)}|}{M_1M_2 - 1} \frac{p_1^{(e)} + p_{M_1M_2}^{(e)}}{2} \\
&= \sum_{j=1}^{M_2} \sum_{i=1}^{M_1-1} \left| (s_j^{(n)} + s_{i+1}^{(\widehat{e})}) - (s_j^{(n)} + s_i^{(\widehat{e})}) \right| \frac{p_j^{(n)} p_i^{(\widehat{e})} + p_j^{(n)} p_{i+1}^{(\widehat{e})}}{2} \\
&\quad + \sum_{j=1}^{M_2-1} \left| (s_j^{(n)} + s_{M_1}^{(\widehat{e})}) - (s_{j+1}^{(n)} + s_1^{(\widehat{e})}) \right| \frac{p_j^{(n)} p_{M_1}^{(\widehat{e})} + p_{j+1}^{(n)} p_1^{(\widehat{e})}}{2} \\
&\quad + \frac{|(s_{M_2}^{(n)} + s_{M_1}^{(\widehat{e})}) - (s_1^{(n)} + s_1^{(\widehat{e})})|}{M_1M_2 - 1} \frac{p_1^{(n)} p_1^{(\widehat{e})} + p_{M_2}^{(n)} p_{M_1}^{(\widehat{e})}}{2} \\
&\overset{(a)}{\approx} \sum_{j=1}^{M_2} \sum_{i=1}^{M_1-1} \left| (s_j^{(n)} + s_{i+1}^{(\widehat{e})}) - (s_j^{(n)} + s_i^{(\widehat{e})}) \right| \frac{p_j^{(n)} p_i^{(\widehat{e})} + p_j^{(n)} p_{i+1}^{(\widehat{e})}}{2} \\
&= \sum_{j=1}^{M_2} p_j^{(n)} \left\{ \sum_{i=1}^{M_1-1} \left| s_{i+1}^{(\widehat{e})} - s_i^{(\widehat{e})} \right| \frac{p_i^{(\widehat{e})} + p_{i+1}^{(\widehat{e})}}{2} \right\} \overset{(b)}{\approx} \sum_{j=1}^{M_2} p_j^{(n)} \Delta(\widehat{e}(k))
\end{aligned}$$

where (a) and (b) follows from Assumption 3 (i.e.$p_1^{(\widehat{e})} \approx 0$, $p_{M_1}^{(\widehat{e})} \approx 0$). The above result is promising, since it implies minimizing the Δ-entropy of the noise-corrupted error $e(k)$ will be approximately equivalent to minimizing the Δ-entropy of the intrinsic error, which is the ultimate objective function that needs to be minimized. Therefore, the Δ-entropy criterion may yield approximately an unbiased solution even if the input and output data are both corrupted by noises. To verify the "unbiasness" of the Δ-entropy criterion, we present herein a supplementary example. Consider again the identification of a two-tap FIR filter in which $G^*(z) = w_1^* + w_2^* z^{-1} = 1.0 + 0.5z^{-1}$. This time we assume the input signal $x(k)$, input noise $n_1(k)$, and output noise $n_2(k)$ are all zero-mean white Bernoulli processes with distributions below

$$\begin{cases} \Pr\{x(k) = \sigma_x\} = 0.5, \quad \Pr\{x(k) = -\sigma_x\} = 0.5 \\ \Pr\{n_1(k) = \sigma_{n_1}\} = 0.5, \quad \Pr\{n_1(k) = -\sigma_{n_1}\} = 0.5 \\ \Pr\{n_2(k) = \sigma_{n_2}\} = 0.5, \quad \Pr\{n_2(k) = -\sigma_{n_2}\} = 0.5 \end{cases} \tag{D.7}$$

where σ_x, σ_{n_1} and σ_{n_2} denote respectively, the standard deviations of $x(k)$, $n_1(k)$ and $n_2(k)$. In the simulation we set $\sigma_x = 1.0$, and the training data length $L = 500$.

Tables 2 and 3 list the "mean ± deviation" results (over 100 Monte Carlo runs) of the estimated parameters (w_1 and w_2) at the 10th EDA generation for different values of σ_{n_1} and σ_{n_2}. For comparison purpose we also include the results obtained using MSE criterion. From the tables, we observe that the Δ-entropy criterion produces nearly mean-unbiased estimates under various SNR conditions, whereas the MSE criterion yields mean-biased solution especially when the input noise power ($\sigma_{n_1}^2$) increasing.

References

[1] J.C. Aguero, G.C. Goodwin, J.I. Yuz, System identification using quantized data, in: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, USA, December 2007, pp. 4263–4268.
[2] B. Chen, J. Hu, L. Pu, Z. Sun, Stochastic gradient algorithm under (h,ϕ)-entropy criterion, Circuits Systems Signal Processing 26 (2007) 941–960.
[3] T. Chen, K. Tang, G. Chen, X. Yao, On the analysis of average time complexity of estimation of distribution algorithms, in: Proceedings of 2007 IEEE Congress on Evolutionary Computation (CEC2007), 2007, pp. 453–460.
[4] T.M. Cover, J.A. Thomas, Element of Information Theory, Wiley & Son, Inc., Chichester, 1991.
[5] L. Devroye, G. Lugosi, Combinatorial Methods in Density Estimation, Springer-Verlag, 2000.
[6] D. Erdogmus, E.H. Kenneth, J.C. Principe, Online entropy manipulation: stochastic information gradient, IEEE Signal Processing Letters 10 (2003) 242–245.
[7] D. Erdogmus, J.C. Principe, Generalized information potential criterion for adaptive system training, IEEE Transactions on Neural Networks 13 (2002) 1035–1044.
[8] D. Erdogmus, J.C. Principe, Convergence properties and data efficiency of the minimum error entropy criterion in Adaline training, IEEE Transactions on Signal Processing 51 (2003) 1966–1978.
[9] D.Z. Feng, X.D. Zhang, D.X. Chang, W.X. Zheng, A fast recursive total least squares algorithm for adaptive FIR filtering, IEEE Transactions on Signal Processing 52 (2004) 2729–2737.
[10] E. Gassiat, E. Gautherat, Identification of noisy linear systems with discrete random input, IEEE Transactions on Information Theory 44 (1998) 1941–1952.
[11] C. Gonzalez, A. Ramirez, J.A. Lozano, P. Larranaga, Average time complexity of estimation of distribution algorithms, in: The 18th International Work-Conference on Artificial Neural Networks (IWANN2005), Lecture Notes in Computer Science, vol. 3512, 2005, pp. 42–49.
[12] L. Hu, C. Zhou, Z. Sun, Estimating biped gait using spline-based probability distribution function with Q-learning, IEEE Transactions on Industrial Electronics 55 (2008) 1444–1452.
[13] M. Janzura, T. Koski, A. Otahal, Minimum entropy of error principle in estimation, Information Sciences 79 (1994) 123–144.
[14] M. Janzura, T. Koski, A. Otahal, Minimum entropy of error estimation for discrete random variables, IEEE Transactions on Information Theory 42 (4) (1996) 1193–1201.
[15] J.N. Kapur, H.K. Kesavan, Entropy Optimization Principles with Applications, Academic Press Inc., 1992.
[16] P. Larranaga, J.A. Lozano, Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation, Kluwer Academic Publishers, Boston, 2002.
[17] T.H. Li, Blind identification and deconvolution of linear systems driven by binary random sequences, IEEE Transactions on Information Theory 38 (1992) 26–38.
[18] T.H. Li, Finite-alphabet information and multivariate blind deconvolution and identification of linear systems, IEEE Transactions on Information Theory 49 (2003) 330–337.
[19] T. Ling, T. David, Adaptive estimated maximum-entropy distribution model, Information Science 177 (2007) 3110–3128.
[20] N. Minamide, An extension of the entropy theorem for parameter estimation, Information and Control 53 (1982) 81–90.
[21] A. Okao, M. Ikeda, R. Takahashi, System identification for nano control: a finite wordlength problem, in: Proceedings of Conference on Control Applications, Istanbul, Turkey, June 2003, pp. 49–53.
[22] U. Ozertem, D. Erdogmus, Second-order volterra system identification with noisy input–output measurements, IEEE Signal Processing Letters 16 (2009) 18–21.
[23] U. Ozertem, I. Uysal, D. Erdogmus, Continuously differentiable sample-spacing entropy estimation, IEEE Transactions on Neural Networks 19 (2008) 1978–1984.
[24] L. Pardo, Statistical Inference Based on Divergence Measures, Chapman & Hall/CRC, 2006.
[25] J.C. Patra, R.N. Pal, R. Baliarsingh, G. Panda, Nonlinear channel equalization for QAM signal constellation using artificial neural networks, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 29 (2) (1999) 262–271.
[26] C.L. Phillips, H.T. Nagle, Digital Control System Analysis and Design, fourth ed., Prentice Hall Press, 2007.
[27] J.C. Principe, D. Xu, Q. Zhao, et al, Learning from examples with information theoretic criteria, Journal of VLSI Signal Processing Systems 26 (2000) 61–77.
[28] Y.N. Rao, D. Erdogmus, J.C. Principe, Error whitening criterion for adaptive filtering: theory and algorithms, IEEE Transactions on Signal Processing 53 (2005) 1057–1069.
[29] R. Rastegar, M.R. Meybodi, A study on global convergence time complexity of estimation of distribution algorithms, in: Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC2005), Lecture Notes in Artificial Intelligence, vol. 3641, 2005, pp. 441–450.
[30] R. Santana, P. Larranaga, J.A. Lozano, Side chain placement using estimation of distribution algorithms, Artificial Intelligence in Medicine 39 (2007) 49–63.
[31] L.M. Silva, C.S. Felgueiras, L.A. Alexandre, J. Marques, Error entropy in classification problems: a univariate data analysis, Neural Computation 18 (2006) 2036–2061.
[32] T. Soerstrom, Errors-in-variables methods in system identification, Automatica 43 (2007) 939–958.
[33] H. Suzuki, T. Sugie, System identification based on quantized I/O data corrupted with noise and its performance improvement, in: Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, USA, December 2006, pp. 3684–3689.
[34] O. Vasicek, A test for normality based on sample entropy, Journal of the Royal Statistical Society Series A 38 (1976) 54–59.
[35] L.Y. Wang, G.G. Yin, Y. Zhao, J.F. Zhang, Identification input design for consistent parameter estimation of linear systems with binary-valued output observations, IEEE Transactions on Automatic Control 53 (2008) 867–880.
[36] L.Y. Wang, J.F. Zhang, G.G. Yin, System identification using binary sensors, IEEE Transactions on Automatic Control 48 (11) (2003) 1892–1907.
[37] H.L. Weidemann, E.B. Stear, Entropy analysis of estimating systems, IEEE Transactions on Information Theory 16 (1970) 264–270.
[38] T.C. Yang, Networked control system: a brief survey, Control Theory and Applications, IEE Proceedings 153 (4) (2006) 403–412.
[39] Q. Zhang, H. Muhlenbein, On the convergence of a class of estimation of distribution algorithms, IEEE Transactions on Evolutionary Computation 8 (2) (2004) 127–136.
[40] Y. Zhao, L.Y. Wang, G.G. Yin, J.F. Zhang, Identification of Wiener systems with binary-valued output observations, Automatica 43 (2007) 1752–1765.
[41] A.C. Harvey, C. Fernandez, Time series for count data or qualitative observations, Journal of Business and Economic Statistics 7 (1989) 407–417.
[42] M. Al-Osh, A. Alzaid, First order integer-valued autoregressive INAR(1) process, Journal of Time Series Analysis 8 (3) (1987) 261–275.
[43] K. Brannas, A. Hall, Estimation in integer-valued moving average models, Applied Stochastic Models in Business and Industry 17 (3) (2001) 277–291.
[44] C.H. Weiß, Thinning operations for modeling time series of counts—a survey, AStA Advances in Statistical Analysis 92 (3) (2008) 319–341.
[45] M. Sato-Ilic, Fuzzy regression models on entropy based blocking structures, International Journal of Innovative Computing, Information and Control 5 (6) (2009) 1475–1484.
[46] W. Zeng, F. Yu, X. Yu, H. Chen, S. Wu, Entropy of intuitionistic fuzzy set based on similarity measure, International Journal of Innovative Computing, Information and Control 5 (12A) (2009) 4737–4744.
[47] X. Gao, C. You, Maximum entropy membership functions for discrete fuzzy variables, Information Sciences 179 (14) (2009) 2353–2361.
[48] Q.S. Zhang, S.Y. Jiang, A note on information entropy measures for vague sets and its applications, Information Sciences 178 (21) (2008) 4184–4191.
[49] J. Balatoni, A. Renyi, On the notion of entropy (Hungarian), vol. 1, Publ. Math. Inst. Hungarian Acad. Sci., 1956, pp. 9–40 (English Translation: Selected Papers of Alfred Renyi, vol. 1, Akademiat Kiado, Budapest, 1976, pp. 558–584).