Preprints.org

Article

Not peer-reviewed version

MMSE Estimation and Mutual Information Gain

Jerry Gibson *

Posted Date: 5 August 2024

doi: 10.20944/preprints202408.0096.v1

Keywords: Mutual information gain; entropy power; minimum mean squared error estimation

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

MMSE Estimation and Mutual Information Gain

Jerry Gibson

University of California, Santa Barbara, Department of Electrical and Computer Engineering, Santa Barbara, CA 93106-9560; gibson@ece.ucsb.edu

Abstract: Information theoretic quantities such as entropy, entropy rate, information gain, and relative entropy are often used to understand the performance of intelligent agents in learning applications. Mean squared error has not played a role in these analyses, primarily because it is not felt to be a viable performance indicator in these scenarios. We build on a new quantity, the log ratio of entropy powers, to establish that minimum mean squared error (MMSE) estimation, prediction, and smoothing are directly connected to Mutual Information Gain or Loss in an agent learning system modeled by a Markov chain for many probability distributions of interest. Expressions for mutual information gain or loss are developed for MMSE estimation, prediction, and smoothing and an example for fixed lag smoothing is presented.

Keywords: mutual information gain; entropy power; minimum mean squared error estimation

1. Introduction

Minimum mean squared error estimation, prediction, and smoothing [1], whether as point estimation, batch least squares, recursive least squares [2], Kalman filtering[3], numerically stable square root filters, recursive least squares lattice structures [4], or stochastic gradient algorithms[3], are a staple in signal processing applications. However, even though stochastic gradient algorithms are the workhorses in the inner workings of machine learning, it is felt that mean squared error does not capture the performance of a learning agent [5]. We begin to address this assertion here and show the close relationship between mean squared estimation error and information theoretic quantities such as differential entropy and mutual information. We consider the problem of estimating a random scalar signal $x[k]$ (the extension to vectors will be obvious to the reader) given the perhaps noisy measurements $z[j]$, where $j = k$ is filtering, $j > k$ is smoothing, and $j < k$ is prediction, based on a minimum mean squared error cost function.

Information theoretic quantities such as entropy, entropy rate, information gain, and relative entropy are often used to understand the performance of intelligent agents in learning applications [6,7]. A relatively newer quantity called Mutual Information Gain or Loss has recently been introduced and shown to provide new insights into the process of agent learning [8]. We build on expressions for Mutual Information Gain that involve ratios of mean squared errors, and establish that minimum mean squared error (MMSE) estimation, prediction, and smoothing are directly connected to Mutual Information Gain or Loss for sequences modeled by many probability distributions of interest. The key quantity in establishing these relationships is the log ratio of entropy powers.

We begin in Sec. 2 by establishing the fundamental information quantities of interest and setting the notation. In Sec. 3, we review information theoretic quantities that have been defined and used in some agent learning analyses in the literature. Some prior work with similar results but based on the minimax entropy of the estimation error is discussed in Sec. 4. The following section, Sec. 5, introduces the key tool in our development, the log ratio of entropy powers, and derives its expression in terms of mutual information gain. In Sec. 6 the log ratio of entropy powers is used to characterize the performance of MMSE smoothing, prediction, and filtering in terms of ratios of entropy powers and Mutual Information Gain. For many probability distributions of interest, we are able to substitute MMSE into the entropy power expressions as shown in Sec. 7. A simple fixed lag smoothing example is presented in Sec. 8 that illustrates the power of the approach. Section 9 presents some properties and families of distributions that commonly occur in applications and that have desirable characterizations and implications. Lists of distributions that satisfy the log ratio of entropy power property and these

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

properties and fall in the classes of interest are given. Final discussions of the results are presented in Sec. 10.

2. Differential Entropy, Mutual Information and Entropy Rate: Definitions and Notation

Given a continuous random variable X with probability density function $p(x)$, the differential entropy is defined as

$$h(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx \tag{1}$$

where we assume X has the variance $var(X) = \sigma^2$. The differential entropy of a Gaussian sequence with mean zero and variance σ^2 is given by [9],

$$h(X) = \frac{1}{2} \log 2\pi e\sigma^2 \tag{2}$$

An important quantity for investigating structure and randomness is the differential entropy rate [9]

$$h(\mathcal{X}) = \lim_{N\to\infty} \frac{1}{N} h(X_1, \ldots, X_N) \tag{3}$$

which is the long term average differential entropy in bits/symbol for the sequence being studied. The differential entropy rate is a simple indicator of randomness that has been used in agent learning papers [6,7].

An alternative definition of differential entropy rate is [9]

$$h(\mathcal{X}) = \lim_{N\to\infty} h(X_N | X_{N-1}, \ldots, X_1) \tag{4}$$

which for the Gaussian process yields

$$h(\mathcal{X}) = \frac{1}{2} \log 2\pi e\sigma_\infty^2 \tag{5}$$

where σ_∞^2 is the minimum mean squared error of the best estimate given the infinite past, expressible as

$$\sigma_\infty^2 = \frac{1}{(2\pi e)} e^{2h(\mathcal{X})} \leq \sigma^2 \tag{6}$$

with σ^2 and $h(\mathcal{X})$ the variance and differential entropy rate of the original sequence, respectively [9]. In addition to defining entropy power, this equation shows that the entropy power is the minimum variance that can be associated with the not-necessarily-Gaussian differential entropy $h(X)$.

In his landmark 1948 paper [10], Shannon defined the entropy power (also called entropy rate power) to be the power in a Gaussian white noise limited to the same band as the original ensemble and having the same entropy. He then used the entropy power in bounding the capacity of certain channels and for specifying a lower bound on the rate distortion function of a source.

Shannon gave the quantity σ_∞^2 the notation Q, which operationally is the power in a Gaussian process with the same differential entropy as the original random variable X [10]. Note that the original random variable or process does not need to be Gaussian. Whatever the form of $h(\mathcal{X})$ for the original process, the entropy power can be defined as in Eq. (6). In the following, we use $h(X)$ for both differential entropy and differential entropy rate unless a clear distinction is needed to reduce confusion.

The differential entropy is defined for continuous amplitude random variables and processes, and it is the appropriate quantity to study signals such as speech, audio, and biological signals. However, unlike discrete entropy, differential entropy can be negative or infinite, and is changed by scaling and similar transformations. Note that this is why mutual information is often the better choice for investigating learning applications.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

In particular, for continuous random variables X and Y with probability density functions $p(x,y)$, $p(x)$ and $p(y)$, respectively, the mutual information between X and Y is

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{7}$$

Mutual information is always greater than or equal to zero and is not impacted by scaling or similar transformations. Mutual information is the principal information theoretic indicator employed in this work.

3. Agent Learning and Mutual Information Gain

In agent learning, based on some observations of the environment, we develop an understanding of the structure of the environment, formulate models of this structure, and study any remaining apparent randomness or unpredictability [6,7]. Studies of agent learning have made use of the information theoretic ideas in Sec. 2, and have created variations on those information theoretic ideas to capture particular characteristics that are distinct to agent learning problems. These expressions and related results are discussed in detail in Gibson [8].

The agent learning literature explores the broad ideas of unpredictability and apparent randomness [6,7]. Toward this end, it is common to investigate the *total Shannon entropy* of length-N sequences $X^N = X_N, X_{N-1}, \ldots, X_1$ given by

$$h(X^N) = -\int P_X^{(N)}(X) \log P_X^{(N)}(X) dX^N \tag{8}$$

as a function of N, to characterize learning. The name total Shannon entropy is appropriate since it is not the usual per component entropy of interest in lossless source coding [9], for example.

In association with the idea of learning or discerning structure in an environment, the *entropy gain*, as defined in the literature, is the difference between the entropies of length N and length $N-1$ sequences as [7]

$$\Delta H(N) = h(X^N) - h(X^{N-1}) \tag{9}$$

Equation (9) was derived and studied much earlier by Shannon [10], not as an entropy gain, but as a conditional entropy.

In particular, Shannon [10] defined the conditional entropy of the next symbol when the $N-1$ preceding symbols are known as

$$h(X_N|X^{N-1}) = h(X_N, X^{N-1}) - h(X^{N-1}) = h(X^N) - h(X^{N-1}) \tag{10}$$

which is exactly Eq. (9); so the entropy gain from the agent learning literature is simply the conditional entropy expression developed by Shannon in 1948.

A recently introduced quantity, *Mutual Information Gain*, allows a more detailed parsing of what is happening in the learning process than observing changes in entropy [8]. Even though a relative entropy between two probability densities has been called the information gain in the agent learning literature [6,7], it is evident from Eqs. (9) and (10) that it is just a conditional entropy [8]. Thus, the nomenclature which defined information gain in terms of this conditional entropy is misleading.

In terms of information gain, the quantity of interest is the mutual information between the overall sequence and the growing history of the past given by

$$\begin{aligned} I(X_N; X^{N-1}) &= h(X_N) - h(X_N|X^{N-1}) \\ &= h(X_N) - [h(X^N) - h(X^{N-1})] \\ &= h(X_N) - \Delta H(N) \end{aligned} \tag{11}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

where $\Delta H(N)$ is defined in Eq. (9). The mutual information in Eq. (11) is a much more direct measure of information gained than entropy gain as a function of N and includes the entropy gain from agent learning as a natural component. We can obtain more insight by expanding Eq. (11) using the chain rule for mutual information [9] as

$$\begin{aligned} I(X_N; X^{N-1}) &= h(X_N) - h(X_N|X^{N-1}) = \sum_{k=1}^{N-1} I(X_N; X_k|X_{k-1}, \ldots, X_0) \\ &= I(X_N; X_{N-1}|X_{N-2}, \ldots, X_1, X_0) + \ldots + I(X_N; X_2|X_1, X_0) + I(X_N; X_1|X_0) \end{aligned} \tag{12}$$

Since $I(X_N; X_{k-1}|X_{k-2}, \ldots, X_0) \geq 0$, we see that $I(X_N; X^{N-1})$ is nondecreasing in N. ; however, what do these individual terms in Eq. (12) mean? The sequence X_N should be considered the input sequence to be analyzed with the block length N large but finite. The first term in the sum, $I(X_N; X_1|X_0)$ indicates the mutual information between the predicted value of X_1, given X_0, and the input sequence X_N. The next term $I(X_N; X_2|X_1, X_0)$ is the mutual information between the input sequence X_N and the predicted value of X_2, given the prior values X_1, X_0. Therefore, we can characterize the change in mutual information with increasing knowledge of the past history of the sequence as a sum of conditional mutual informations $I(X_N; X_{k-1}|X_{k-2}, \ldots, X_0)$ [8].

We denote $I(X_N; X^{N-1})$ as the *total mutual information gain*, and $I(X_N; X_{k-1}|X_{k-2}, \ldots, X_0)$ as the *incremental mutual information gain*. We utilize these terms in the following developments.

4. Minimum Error Entropy

Minimum error entropy approaches to estimation, prediction, and smoothing are studied by Kalata and Priemer [11], and minimax error entropy stochastic approximation is investigated by Kalata and Priemer [12]. They consider the estimation error, $\tilde{X}[k] = X[k] - \widehat{X}[k|k]$, and study random variables with probability density functions that have the differential entropy $h(X) = \frac{1}{2} log[A\sigma_X^2]$. The authors point out that random variables with densities of this form are Gaussian, Laplacian, Uniform, triangular, exponential, Rayleigh, and Poisson [11,12].

They show, among other results, that minimizing the estimation error entropy is equivalent to minimizing the mutual information between the estimation error and the observations, that is,

$$min\ h(\tilde{X}) \Leftrightarrow\ min\ I(\tilde{X}; Z). \tag{13}$$

For differential entropies of the form $h(X) = \frac{1}{2} log[A\sigma_X^2]$, they also show that the MMSE estimate is the minimax error entropy estimate, that is,

$$min\ max\ h(\tilde{X}) \Leftrightarrow\ min\ \sum_{i=1}^{N} \sigma_{X_i}^2 \tag{14}$$

This allows the development of standard MMSE estimators for filtering, smoothing, and prediction based on the minimax error entropy approach.

The authors also develop an expression for the change in smoothing error with a new observation [13]. They note that there is a change in the error entropy with a new observation, $z(j+1)$, that is given by

$$\Delta h(\tilde{X}(k)|z(j+1)) = h(\tilde{X}(k)|Z(j)) - h(\tilde{X}(k)|Z(j+1)) \tag{15}$$

Using the definition of mutual information in terms of entropies, Eq. (7), and the given form of the differential entropy, it is shown that the minimum error entropy optimum smoothing error variance decreases as

$$\sigma^2_{\tilde{X}(k)|Z(j+1)} = \sigma^2_{\tilde{X}(k)|Z(j)} \exp\left[-2I(X(k); z(j+1)|Z(j))\right] \tag{16}$$

with the new observation $z(j+1)$ for the stated distributions.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

We will see in the following that we can obtain similar results but in terms of mutual information for the same probability distributions with MMSE estimation methods without considering the minimax error entropy approach.

5. Log Ratio of Entropy Powers

We can use the definition of the entropy power in Equation (6) to express the logarithm of the ratio of two entropy powers in terms of their respective differential entropies as [14]

$$\log \frac{Q_X}{Q_Y} = 2[h(X) - h(Y)]. \tag{17}$$

The conditional version of Equation (6) is

$$Q_{X|Y_N} = \frac{1}{(2\pi e)} \exp 2h(X|Y_N) \leq Var(X|Y_N), \tag{18}$$

and from which we can express Equation (17) in terms of the entropy powers at the outputs of successive stages in a signal processing Markov chain Y_{N-1}, Y_N that satisfy the Data Processing Inequality as

$$\frac{1}{2} \log \frac{Q_{X|Y_N}}{Q_{X|Y_{N-1}}} = h(X|Y_N) - h(X|Y_{N-1}). \tag{19}$$

It is important to notice that many signal processing systems satisfy the Markov chain property and thus the Data Processing Inequality so Eq. (19) is potentially very useful and insightful.

We can expand our insights if we add and subtract $h(X)$ to the right-hand side of Equation (19), so we then obtain an expression in terms of the difference in mutual information between the two successive stages as

$$\frac{1}{2} \log \frac{Q_{X|Y_N}}{Q_{X|Y_{N-1}}} = I(X;Y_{N-1}) - I(X;Y_N). \tag{20}$$

From the the entropy power in Equation (18), we know that both expressions in Equations (19) and (20) are greater than or equal to zero. So, from this result we see that we can now associate a change in mutual information as data passes through a Markov chain with the log ratio of entropy powers.

These results are from [14] and extend the data processing inequality by providing a new characterization of the mutual information gain or loss between stages in terms of the entropy powers of the two stages. Since differential entropies are difficult to calculate, it is useful to have expressions for the entropy power at two stages and then use Equations (19) and (20) to find the difference in differential entropy and mutual information between these stages.

To get some idea of how useful Eq. (20) can be, we turn to a few special cases. In many signal processing operations, a Gaussian assumption is accurate and can provide deep insights. Thus, considering two i.i.d. Gaussian distributions with zero mean and variances σ_X^2 and σ_Y^2, we have directly that $Q_X = \sigma_X^2$ and $Q_Y = \sigma_Y^2$, so

$$\frac{1}{2} \log \frac{Q_X}{Q_Y} = \frac{1}{2} \log \frac{\sigma_X^2}{\sigma_Y^2} = [h(X) - h(Y)], \tag{21}$$

which satisfies Equation (17) exactly.

We can also consider the MMSE error variances in a Markov chain when X and Y_{N-1}, Y_N are Gaussian with the error variances at successive stages denoted as $Var(X|Y_{N-1}) = \sigma^2_{X|Y_{N-1}}$ and $Var(X|Y_N) = \sigma^2_{X|Y_N}$, then

$$\frac{1}{2} \log \frac{Q_{X|Y_N}}{Q_{X|Y_{N-1}}} = \frac{1}{2} \log \frac{\sigma^2_{X|Y_N}}{\sigma^2_{X|Y_{N-1}}} = h(X|Y_N) - h(X|Y_{N-1}) = I(X;Y_{N-1}) - I(X;Y_N). \tag{22}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

Perhaps surprisingly, this result holds for two i.i.d. Laplacian distributions with variances $\lambda^2_{X|Y_N}$ and $\lambda^2_{X|Y_{N-1}}$ [15], since their corresponding entropy powers $Q_{X|Y_N} = 2e\lambda^2_{X|Y_N}/\pi$ and $Q_{X|Y_{N-1}} = 2e\lambda^2_{X|Y_{N-1}}/\pi$, respectively, so we form

$$\frac{1}{2}\log\frac{Q_{X|Y_N}}{Q_{X|Y_{N-1}}} = \frac{1}{2}\log\frac{\lambda^2_{X|Y_N}}{\lambda^2_{X|Y_{N-1}}} = h(X|Y_N) - h(X|Y_{N-1}) = I(X;Y_{N-1}) - I(X;Y_N). \tag{23}$$

Since $h(X) = \ln(2e\lambda_X)$, the Laplacian distribution also satisfies Equation (17) through Eq. (20) exactly [14].

Using mean squared errors or variances in Equations (17) through (20) is accurate for many other distributions as well. It is straightforward to show that Equation (17) holds with equality when the differential entropy takes the form

$$h(X) = \frac{1}{2}log[A\sigma^2_X] \tag{24}$$

so the entropy powers can be replaced by the mean squared error for the Gaussian, Laplacian, logistic, Cauchy, uniform, symmetric triangular, exponential, and Rayleigh distributions. Equation (24) is of the same form of the distributions considered in [11,12] when considering the minimax error entropy estimate. Note here that we can work directly with MMSE estimates.

Therefore, the satisfaction of Equations (17) through (20) with equality when substituting the variance for entropy power occurs for several distibutions of significant interest for applications, and it is the log ratio of entropy powers that enables the use of the mean squared error to calculate the loss or gain in mutual information at each stage.

6. Minimum Mean Squared Error (MMSE) Estimation

Using the results from Sec. 5, a tight connection between mean squared estimation error, denoted as MSEE, and mutual information gain or loss in common applications is established here and in the following subsections. Then in Sec. 7 these results are specialized to the use of the error variances.

In minimum mean squared estimation (MMSE), the estimation error to be minimized is

$$\epsilon^2 = E(X[k] - \widehat{X}[k|j])^2 \tag{25}$$

at time instant k given observations up to and including time instant j where we may have $j = k$, $j > k$, or $j < k$, depending on whether the problem is classical estimation, smoothing, or prediction, respectively.

Using the estimation counterpart to Fano's Inequality, we can write [9]

$$E(X[k] - \widehat{X}[k|j])^2 \geq \frac{1}{2\pi e}\exp 2[h(X[k]|\widehat{X}[k|j]) \equiv Q_{X[k]|\widehat{X}[k|j]} \tag{26}$$

where we have used the classical notation for entropy power defined by Shannon [10]. Taking the logarithm of the right side of Eq. (26) past the inequality, we obtain

$$h(X[k]|\widehat{X}[k|j]) = \frac{1}{2}\log(2\pi e Q_{X[k]|\widehat{X}[k|j]}) \tag{27}$$

Subtracting $h(X[k]|\widehat{X}[k|l])$, $l \neq j$, from the left side of Eq. (27) and the corresponding entropy power expression from the right side, we get

$$h(X[k]|\widehat{X}[k|j]) - h(X[k]|\widehat{X}[k|l]) = \frac{1}{2}\log(\frac{Q_{X[k]|\widehat{X}[k|j]}}{Q_{X[k]|\widehat{X}[k|l]}}) \tag{28}$$

Note that the $2\pi e$ divides out in the ratio of entropy powers.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

Adding and subtracting $h(X[k])$ from both sides of Eq. (28), we can write

$$I(X[k];\widehat{X}[k|j]) - I(X[k];\widehat{X}[k|l]) = \frac{1}{2}\log(\frac{Q_{X[k]|\widehat{X}[k|j]}}{Q_{X[k]|\widehat{X}[k|l]}}) \tag{29}$$

Therefore, the difference between the mutual information of $X[k]$ and $\widehat{X}[k|j]$ and the mutual information of $X[k]$ and $\widehat{X}[k|l]$ can be expressed as one half the log ratio of conditional entropy powers. This allows us to characterize Mutual Information Gain or Loss in terms of the minimum mean squared error in filtering, smoothing, and prediction, as we demonstrate in Sec. 7.

6.1. MMSE Smoothing

We want to estimate a random scalar signal $x[k]$ given the perhaps noisy measurements $z[j]$ for $j > k$, where k is fixed and j is increasing, based on a minimum mean squared error cost function. So, the smoothing error to be minimized is Eq. (25), and again using the estimation counterpart to Fano's Inequality [9] we get Eq. (26), both with $j > k$ for smoothing. As j increases, the optimal smoothing estimate will not increase the MMSE so

$$Q_{X[k]|\widehat{X}[k|j]} \geq Q_{X[k]|\widehat{X}[k|j+1]} \tag{30}$$

Moving $Q_{X[k]|\widehat{X}[k|j+1]}$ over to the left side of Eq. (30) and substituting the definition of entropy power for each produces

$$\frac{Q_{X[k]|\widehat{X}[k|j]}}{Q_{X[k]|\widehat{X}[k|j+1]}} = \exp 2[h(X[k]|\widehat{X}[k|j])] - h(X[k]|\widehat{X}[k|j+1])] \geq 1 \tag{31}$$

Taking logarithms, we see that

$$\frac{1}{2}\log\frac{Q_{X[k]|\widehat{X}[k|j]}}{Q_{X[k]|\widehat{X}[k|j+1]}} = [h(X[k]|\widehat{X}[k|j]) - h(X[k]|\widehat{X}[k|j+1])] \geq 0 \tag{32}$$

Adding and subtracting $h(X[k])$ to the right hand side of Eq. (32), yields

$$\frac{1}{2}\log\frac{Q_{X[k]|\widehat{X}[k|j]}}{Q_{X[k]|\widehat{X}[k|j+1]}} = I(X[k];\widehat{X}[k|j+1]) - I(X[k];\widehat{X}[k|j]) \geq 0 \tag{33}$$

Equation (33) shows that the mutual information is nondecreasing for increasing $j > k$. Thus we have an expression for the *Mutual Information Gain* due to smoothing as a function of lookahead j in terms of entropy powers.

We can also use Eq. (33) to obtain the rate of decrease of the entropy power in terms of the mutual information as

$$Q_{X[k]|\widehat{X}[k|j+1]} = Q_{X[k]|\widehat{X}[k|j]} \exp\left[-2(I(X[k];\widehat{X}[k|j+1]) - I(X[k];\widehat{X}[k|j]))\right] \tag{34}$$

Here we see that the rate of decrease in the entropy power is exponentially related to the Mutual Information Gain due to smoothing.

We note that this result is obtained only using entropy power expressions rather than minimal error entropy. In fact, it can be shown that Eq, (34) and Eq. (16) are the same by employing the chain rule to prove that $I(X(k);z(j+1)|Z(j)) = I(X[k];\widehat{X}[k|j+1]) - I(X[k];\widehat{X}[k|j]$.

6.2. MMSE Prediction

We want to predict a random scalar signal $x[k]$ given the perhaps noisy measurements $z[j]$ for $j < k$, where k is fixed and j is decreasing from $k-1$, based on a minimum mean squared error

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

cost function. So, the prediction error to be minimized is Eq. (25) and again using the estimation counterpart to Fano's Inequality [9] we get Eq. (26), both with $j < k$ for prediction. As j decreases, the optimal prediction will increase the minimum mean squared prediction error since the prediction is further ahead, so

$$Q_{X[k]|\widehat{X}[k|j]} \leq Q_{X[k]|\widehat{X}[k|j-1]} \tag{35}$$

Moving $Q_{X[k]|\widehat{X}[k|j]}$ over to the right side of Eq. (35) and substituting the definition of entropy power for each produces

$$\frac{Q_{X[k]|\widehat{X}[k|j-1]}}{Q_{X[k]|\widehat{X}[k|j]}} = \exp 2[h(X[k]|\widehat{X}[k|j-1])] - h(X[k]|\widehat{X}[k|j])] \geq 1 \tag{36}$$

Taking logarithms, we see that

$$\frac{1}{2}\log \frac{Q_{X[k]|\widehat{X}[k|j-1]}}{Q_{X[k]|\widehat{X}[k|j]}} = [h(X[k]|\widehat{X}[k|j-1]) - h(X[k]|\widehat{X}[k|j])] \geq 0 \tag{37}$$

Adding and subtracting $h(X[k])$ to the right hand side of Eq. (37), yields

$$\frac{1}{2}\log \frac{Q_{X[k]|\widehat{X}[k|j-1]}}{Q_{X[k]|\widehat{X}[k|j]}} = I(X[k];\widehat{X}[k|j]) - I(X[k];\widehat{X}[k|j-1]) \geq 0 \tag{38}$$

This result shows that there is a *Mutual Information Loss* with further lookahead in prediction and this loss is expressible in terms of a ratio of entropy powers. Equation (38) shows that the mutual information is decreasing for decreasing $j < k$, that is, for prediction further ahead, since $I(X[k];\widehat{X}[k|j]) \geq I(X[k];\widehat{X}[k|j-1])$. As a result, the observations are becoming less relevant to the variable to be predicted. We can also use Eq. (38) to obtain the rate of increase of the entropy power as the prediction is further ahead in terms of the mutual information as

$$Q_{X[k]|\widehat{X}[k|j-1]} = Q_{X[k]|\widehat{X}[k|j]} \exp\left[2(I(X[k];\widehat{X}[k|j]) - I(X[k];\widehat{X}[k|j-1]))\right] \tag{39}$$

Thus, the entropy power increase grows exponentially with the Mutual Information Loss corresponding to increasing lookahead in prediction.

6.3. MMSE Filtering

We want to estimate a random scalar signal $x[k]$ given the perhaps noisy measurements $z[k]$, based on a minimum mean squared error cost function. So, the estimation error to be minimized is Eq. (25). From the estimation counterpart to Fano's Inequality [9] we get Eq. (26), both with $j = k$ for filtering.

Dividing $Q_{X[k]|\widehat{X}[k|k]}$ by $Q_{X[k-1]|\widehat{X}[k-1|k-1]}$ and substituting the definition of entropy power for each produces

$$\frac{Q_{X[k]|\widehat{X}[k|k]}}{Q_{X[k-1]|\widehat{X}[k-1|k-1]}} = \exp 2[h(X[k]|\widehat{X}[k|k])] - h(X[k-1]|\widehat{X}[k-1|k-1])] \tag{40}$$

Taking logarithms, we see that

$$\frac{1}{2}\log \frac{Q_{X[k]|\widehat{X}[k|k]}}{Q_{X[k-1]|\widehat{X}[k-1|k-1]}} = [h(X[k]|\widehat{X}[k|k]) - h(X[k-1]|\widehat{X}[k-1|k-1])] \tag{41}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

Adding and subtracting $h(X[k])$ and $h(X[k-1])$ to the right hand side of Eq. (41), yields

$$\frac{1}{2}\log\frac{Q_{X[k]|\widehat{X}[k|k]}}{Q_{X[k-1]|\widehat{X}[k-1|k-1]}} = I(X[k-1];\widehat{X}[k-1|k-1]) - I(X[k];\widehat{X}[k|k]) + [h(X[k]) - h(X[k-1])] \quad (42)$$

This equation involves the differential entropies of $X[k]$ and $X[k-1]$ unlike prior expressions for smoothing and prediction. This is because the reference points for the two entropy powers are different. However, for certain wide sense stationary processes, we will have simplications as shown in the next section on Entropy Power and MSE, where it is shown that for several important distributions, we can replace the entropy power with the variance.

7. Entropy Power and MSE

We know from Sec. 2 that the entropy power is the minimum variance that can be associated with a differential entropy $h(X)$. The key insight into relating mean squared error and mutual information comes from considering the (apparently not so special) cases of random variables whose differential entropy has the form in Eq. (24) and the log ratio of entropy powers. In these cases we do not have to explicitly calculate the entropy power since we can use the variance or mean squared error in the log ratio of entropy power expressions to find the mutual information gain or loss for these distributions.

Thus, all of the results in Sec. 6 in terms of log ratio of entropy powers can be expressed as ratios of variances or mean squared errors for continous random variables with differential entropies of the form in Eq. (24). In the following we use the more notationally bulky $var(X[k]|\widehat{X}[k|j])$ for the conditional variances rather than the simpler notation $\sigma^2_{X(k)|\widehat{X}(j)}$ since the σ^2 symbol could be confused as indicating a Gaussian assumption, which is not needed.

In particular, for the smoothing problem, we can rewrite Eq. (33) as

$$\frac{1}{2}\log\frac{var(X[k]|\widehat{X}[k|j])}{var(X[k]|\widehat{X}[k|j+1])} = I(X[k];\widehat{X}[k|j+1]) - I(X[k];\widehat{X}[k|j]) \geq 0 \quad (43)$$

and the decrease in MSE in terms of the change in mutual information as

$$var(X[k|\widehat{X}[k|j+1]) = var(X[k|\widehat{X}[k|j])\exp\left[-2(I(X[k];\widehat{X}[k|j+1]) - I(X[k];\widehat{X}[k|j]))\right] \quad (44)$$

Here we see that the rate of decrease in the MMSE is exponentially related to the Mutual Information Gain due to smoothing.

Rewriting the results for prediction in terms of variances, we have that Eq. (39) becomes

$$\frac{1}{2}\log\frac{var(X[k]|\widehat{X}[k|j-1])}{var(X[k]|\widehat{X}[k|j])} = I(X[k];\widehat{X}[k|j]) - I(X[k];\widehat{X}[k|j-1]) \geq 0 \quad (45)$$

and that the growth in MMSE with increasing lookahead is

$$var(X[k]|\widehat{X}[k|j-1]) = var(X[k]|\widehat{X}[k|j])\exp\left[2(I(X[k];\widehat{X}[k|j]) - I(X[k];\widehat{X}[k|j-1]))\right] \quad (46)$$

Thus, as lookahead in prediction is increased, the conditional error variance grows exponentially.

For the filtering problem, we have the two differential entropies, $h(X[k])$ and $h(X[k-1])$ in Eq.(42) in addition to the mutual information expressions. However, for wide sense stationary random processes with differential entropies of the form shown in Eq. (24), the two variances are equal so $var(X[k]) = var(X[k-1])$ so the difference in the two differential entropies is zero. This simplifies Eq.(42) to

$$\frac{1}{2}\log\frac{var(X[k]|\widehat{X}[k|k])}{var(X[k-1]|\widehat{X}[k-1|k-1])} = I(X[k-1];\widehat{X}[k-1|k-1]) - I(X[k];\widehat{X}[k|k]) \leq 0, \quad (47)$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

which if the error variance in monotonically nonincreasing, is less than or equal to zero as shown. Rewriting this last result in terms of increasing mutual information, we have

$$\frac{1}{2}\log\frac{var(X[k-1]|\widehat{X}[k-1|k-1])}{var(X[k]|\widehat{X}[k|k])} = I(X[k];\widehat{X}[k|k]) - I(X[k-1];\widehat{X}[k-1|k-1]) \geq 0 \tag{48}$$

Thus, we have related mean squared error from estimators to the change in gain or loss of mutual information.

It is important to recognize the power of the expressions in this section. They allow us to obtain the mutual information gain or loss by using the variances of MMSE estimators, the latter of which are easily calculated in comparison to direct calculation of differential entropy or mutual information. There is no need to utilize techniques to approximately compute differential entropies or mutual informations, which are fraught with difficulties. See Hudson [16] and Kraskov[17].

8. Fixed Lag Smoothing Example

To provide a concrete example of the preceding results in Sec. 7, we consider an example of finding the mutual information gain using the results of a fixed lag MMSE smoothing problem for a simple first order system model with noisy observations. Fixed lag smoothing is a popular approach since measurements $z(k)$ at time instants $k, k+1, \cdots, k+L$ are used to estimate the value of $x(k)$; that is, measurements L samples ahead of the present time k are used to estimate $x(k)$ [18].

A first order autoregressive (AR) system model is given by

$$x(k+1) = \alpha x(k) + w(k+1), \tag{49}$$

where $\alpha \in [0,1]$ and $w(k+1)$ is a stationary, Gaussian, zero mean, white process with variance q. The observation model is expressed as

$$z(k+1) = x(k+1) + v(k+1), \tag{50}$$

where $v(k+1)$ is the zero mean Gaussian noise with variance r.

For this problem we compute the steady state errors in fixed-lag smoothing as a function of the smoothing lag. The steady-state expression for the fixed-lag smoothing error covariance as a function of the lag L is (details are available in [18,19] and are not included here)

$$P_L = P - \sum_{j=1}^{L}[(1-K)\alpha]^{2j}(\bar{P}-P), \tag{51}$$

where the components shown from the Kalman filter are

$$\bar{P} = \alpha^2 P + q \tag{52}$$

$$K = \bar{P}(\bar{P}+r)^{-1} \tag{53}$$

$$P = (1-K)\bar{P}. \tag{54}$$

with P the filtering or estimation error variance, $\bar{P}$ the apriori filter error variance, and K the Kalman filter gain .

Given α, q, and r, P can be computed in the steady state case as the positive root of the following quadratic equation

$$\alpha^2 P^2 + (q + r - \alpha^2 r)P - qr = 0. \tag{55}$$

Then $\bar{P}$, K, and P_L can be evaluated using Equations (52), (53), and (51) respectively.

The asymptotic expression for the smoothing error covariance as L gets large is given by

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

$$P_{L,\min} = P - \sum_{j=1}^{\infty} [(1-K)\alpha]^{2j} (\bar{P} - P),$$
$$= P - (\bar{P} - P) \cdot \frac{(1-K)^2\alpha^2}{1-(1-K)^2\alpha^2} \tag{56}$$

This result can be used to determine what value should be selected for the maximum delay to obtain near asymptotic performance.

Example: We now consider the specific case of the scalar models in Equations (49) and (50) with $\alpha = 0.98$, $q = 0.118$, and $r = 1.18$. The choice of this ratio of $q/r = 0.1$ corresponds an accurate model for the AR process but with a very noisy observation signal. Table 1 lists the smoothing error covariance as a function of smoothing lag L using Eq. (51) and the following. The result for $L = 15$ comes from Eq. (56).

Table 1. Mutual Information Gain due to smoothing with $\alpha = 0.98, q = 0.118, r = 1.18, P = 0.3045$

L	P_L	Incremental MI Gain	Total MI Gain
1	0.2485	0.1145	0.1145
2	0.2189	0.0633	0.1748
3	0.2033	0.0369	0.2117
4	0.1950	0.0209	0.2326
5	0.1906	0.01235	0.24795
15	$0.1857 \approx P_{L_{\min}}$	0.00255	0.2505

Therefore, we are able to obtain statements concerning the gain or loss of mutual information by calculating the much more directly available quantity, the minimum mean squared smoothing error.

We obtained the third column in the table labeled "Incremental MI Gain" from Eq. (43) where for simplicity of notation, we set $P = var(X[k]|\widehat{X}[k|k])$ and let $P_L = var(X[k]|\widehat{X}[k|k+L])$. The fourth column is obtained for a specific L, say $L_{totalgain}$, by adding all values of the Incremental MI Gain with $L \leq L_{totalgain}$. For example, with $L_{totalgain} = 5$, we sum up all values of Incremental MI Gain for $L \leq 5$ to get 0.24795.

The asymptotic reduction in MSE due to smoothing as L gets large is thus $([P - P_{L,\min}]/P) \cdot 100 = 39\%$, and the corresponding mutual information gain is 0.2505 bits.

9. Properties and Families (Classes) of Probability Densities

As we have seen, there are many common probability distributions that let us substitute mean squared error for entropy power in the log ratio of entropy power expression. While a general class of distributions that satisfy this property has not been established, many important and ubiquitous "named" continuous distributions do so. In particular, distributions that satisfy the log ratio of entropy power condition are Gaussian, Laplacian, Cauchy, Gamma, Logistic, exponential, Rayleigh, symmetric triangular, and uniform.

This group of distributions exhibits certain properties and fall into common families or classes of distributions that can prove useful in further studies . The following sections discuss these properties and familiies.

9.1. Properties

Given a continuous random variable X with cumulative probability distribution $P_X(x)$ and corresponding probability density function $p(x)$, then the distribution is said to be *unimodal* if for some $x = a$ such that $P_X(x)$ is convex for $x < a$ and concave for $x > a$ [20]. Example distributions that satisfy this condition and thus are unimodal are the Gaussian, Laplacian, Cauchy, Logistic, exponential, Rayleigh, symmetric triangular, and uniform distributions.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

Further, a unimodal distribution is called *strongly unimodal* if $-\log p(x)$ is convex. Distributions of the form $p(x) = k \exp |x|^{\alpha}$ for $\alpha \geq 1$ are strongly unimodal, and by inspection, we see that the Gaussian, Laplacian, and logistic distributions have this property [21].

9.2. Families or Classes

A number of families, or classes, of distributions have been defined to help categorize random variables. Families or classes that help clarify the scope of the log ratio of entropy power results are *location-scale* families and *exponential* families.

9.2.1. Location-Scale Family

Given a random variable Y with distribution $P_Y(y)$, then for a transformation $X = a + bY$, $b > 0$, a family that satisfies

$$P(X \leq x) = P_Y((x - a)/b) \tag{57}$$

is called a *location-scale* family [22]. Location-Scale Families include Gaussian, Laplacian, Cauchy, Logistic, exponential, symmetric triangular, and uniform distributions [22].

9.2.2. Exponential Family

Given a family of probability density functions $p(x, \theta)$ with $a < \theta < b$, a pdf of the form

$$p(x, \theta) = \exp[\eta(\theta)K(x) + S(x) + q(\theta)], c < x < d, \tag{58}$$

is said to be a member of the *exponential family* of distributions of continuous type [23]. Additionally, given the set of random variables $X_1, X_2, ..., X_n$, their joint pdf of the form

$$\exp[\eta(\theta] \sum_{i=1}^{n} K(x_i) + \sum_{i=1}^{n} S(x_i) + nq(\theta)], \tag{59}$$

$c < x_i < d$ with $a < \theta < b$, and zero elsewhere, is in the exponential family. A nice property of exponential distributions is that sufficient statistics exist for this family.

Examples of distributions in the exponential family are Gaussian, exponential, Gamma, and Poisson [22].

10. Discussion

Entropy and mutual information have been incorporated into many analyses of agent learning. However, mean squared error has mostly been viewed as suspect as a performance indicator in learning applications. It is shown here that the MMSE performance of smoothing, prediction, and filtering algorithms have direct interpretations in terms of the mutual information gained or lost in the estimation process for a fairly large set of probabilty densities that have differential entropies of the form in Eq. (24).

Not only are these results satisfying in terms of a performance indicator, but the expressions in Eqs. (43), (45), and (47) allow gains or losses in mutual information to be calculated from estimation error variances. This avoids the more cumbersome estimates of probability histograms to be used in mutual information expressions or direct approximations of mutual information from data.

These results open the door to explorations of mutual information gain or loss for additional classes of probability densities, perhaps by considering the properties and families briefly discussed in Sec. 9.

Funding: "This research received no external funding"

Data Availability Statement: "Not applicable".

Conflicts of Interest: "The authors declare no conflict of interest."

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2024 doi:10.20944/preprints202408.0096.v1

Abbreviations

The following abbreviations are used in this manuscript:

i.i.d.	independent and identically distributed
MMSE	minimum mean squared error
Q	entropy power
MMSPE(M)	minimum mean squared prediction error of order M
MSE	mean squared error
MI Gain	mutual information gain

References

1. Wiener, N. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*; MIT Press, 1949.
2. Ljung, L.; Soderstrom, T. *Theory and Practice of Recursive Identification*; MIT Press, 1983.
3. Haykin, S. *Adaptive Filter Theory*; Prentice-Hall, 2002.
4. Honig, M.L.; Messerschmitt, D.G. *Adaptive filters: structures, algorithms, and applications*; Kluwer Academic Publishers: Hingham, MA, 1984.
5. Tishby, N.; Zaslavsky, N. Deep Learning and the Information Bottleneck Principle. *CoRR* **2015**, *abs/1503.02406*.
6. Crutchfield, J.P.; Feldman, D.P. Synchronizing to the environment: Information-theoretic constraints on agent learning. *Advances in Complex Systems* **2001**, *4*, 251–264.
7. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **2003**, *13*, 25–54.
8. Gibson, J.D. Mutual Information Gain and Linear/Nonlinear Redundancy for Agent Learning, Sequence Analysis and Modeling. *Entropy* **2020**, *22*, 608–624. doi:10.3390/e22060608.
9. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience, 2006.
10. Shannon, C.E. A mathematical theory of communication. *Bell Sys. Tech. Journal* **1948**, *27*, 379–423.
11. Kalata, P.; Priemer, R. Linear prediction, filtering, and smoothing: An information-theoretic approach. *Information Sciences* **1979**, *17*, 1–14.
12. Kalata, P.; Priemer, R. On minimal error entropy stochastic approximation. *Int. Journal of Systems Sciences* **1974**, *5*, 895–906. doi:10.1080/00207727408920148.
13. Kalata, P.R.; Priemer, R. When should smoothing cease? *Proceedings of the IEEE* **1974**, *62*, 1289–1290.
14. Gibson, J.D. Log Ratio of Entropy Powers. Proc. UCSD Information Theory and Applications, 2018.
15. Shynk, J.J. *Probability, random variables, and random processes: theory and signal processing applications*; John Wiley & Sons, 2012.
16. Hudson, J.E. Signal Processing Using Mutual Information. *IEEE Signal Processing Magazine* **2006**, *23*, 50–54. doi:10.1109/SP-M.2006.248712.
17. Kraskov, A.; Stogbauer, A.; Grassberger, P. Estimating Mutual Information. *Physical Review: E* **2004**, *69*, 006138–1–006138–16. doi:10.1103/PhysRevE.69.006138.
18. Chirarattananon, S.; Anderson, B. The Fixed-Lag Smoother as a Stable, Finite-Dimensional Linear Filter. *Automatica* **1971**, *7*, 657–669.
19. Gibson, J.D.; Bhaskaranand, M. Performance improvement with decoder output smoothing in differential predictive coding. Proc. UCSD Information Theory and Applications, 2014.
20. Lukacs, E. *Characteristic Functions*; Griffin London, 1970.
21. Lehmann, E.L. *Testing Statistical Hypotheses*; John Wiley & Sons, Inc., 1986.
22. Lehmann, E.L. *Theory of Point Estimation*; John Wiley & Sons, Inc., 1983.
23. Hogg, R.V.; Craig, A.T. *Intorduction to Mathematical Statistics*; Macmillan, 1970.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.