entropy

MDPI

Article

Attention to the Variation of Probabilistic Events: Information Processing with Message Importance Measure

Rui She, Shanyun Liu and Pingyi Fan *

Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; sher15@mails.tsinghua.edu.cn (R.S.); liushany16@mails.tsinghua.edu.cn (S.L.)

* Correspondence: fpy@tsinghua.edu.cn; Tel.: +86-010-6279-6973

Received: 10 March 2019; Accepted: 23 April 2019; Published: 26 April 2019

Abstract: Different probabilities of events attract different attention in many scenarios such as anomaly detection and security systems. To characterize the events' importance from a probabilistic perspective, the message importance measure (MIM) is proposed as a kind of semantics analysis tool. Similar to Shannon entropy, the MIM has its special function in information representation, in which the parameter of MIM plays a vital role. Actually, the parameter dominates the properties of MIM, based on which the MIM has three work regions where this measure can be used flexibly for different goals. When the parameter is positive but not large enough, the MIM not only provides a new viewpoint for information processing but also has some similarities with Shannon entropy in the information compression and transmission. In this regard, this paper first constructs a system model with message importance measure and proposes the message importance loss to enrich the information processing strategies. Moreover, the message importance loss capacity is proposed to measure the information importance harvest in a transmission. Furthermore, the message importance distortion function is discussed to give an upper bound of information compression based on the MIM. Additionally, the bitrate transmission constrained by the message importance loss is investigated to broaden the scope for Shannon information theory.

Keywords: message importance measure; information theory; probabilistic events processing; message transmission and compression

1. Introduction

In recent years, massive data has attracted much attention in various realistic scenarios. Actually, there exist many challenges for data processing such as distributed data acquisition, huge-scale data storage and transmission, as well as correlation or causality representation [1–5]. Facing these obstacles, it is a promising way to make good use of information theory and statistics to deal with mass information. For example, a method based on Max Entropy in Metric Space (MEMS) is utilized for local features extraction and mechanical system analysis [6]; as an information measure different from Shannon entropy, Voronoi entropy is discussed to characterize the random 2D patterns [7]; Category theory, which can characterize the Kolmogorov–Sinai and Shannon entropy as the unique functors, is used in autonomous and networked dynamical systems [8].

To some degree, probabilistic events attract different interests according to their probability. For example, considering that small probability events hidden in massive data contain more semantic importance [9–13], people usually pay more attention to the rare events (rather than the common events) and design the corresponding strategies of their information representation and processing

in many applications including outliers detection in the Internet of Things (IoT), smart cities and autonomous driving [14–22]. Therefore, the probabilistic events processing has special values in the information technology based on semantics analysis of message importance.

In order to characterize the importance of probabilistic events, a new information measure named MIM is presented to generalize Shannon information theory [23–25]. Here, we shall investigate the information processing including compression (or storage) and transmission based on MIM to bring some new viewpoints in the information theory. Now, we first give a short review on MIM.

1.1. Review of Message Importance Measure

Essentially, the message importance measure (MIM) is proposed to focus on the probabilistic events importance [23]. In particular, the core idea of this information measure is that the weights of importance are allocated to different events according to the corresponding events' probability. In this regard, as an information measure, MIM may provide an applicable criterion to characterize the message importance from the viewpoint of inherent property of events without the human subjective factors. For convenience of calculation, an exponential expression of MIM is defined as follows.

Definition 1. *For a discrete distribution* $P(X) = \{p(x_1), p(x_2), ...,p(x_n)\}$*, the exponential expression of message importance measure (MIM) is given by*

$$L(\omega, X) = \sum_{x_i} p(x_i) e^{\omega\{1-p(x_i)\}}, \tag{1}$$

where the adjustable parameter ω *is nonnegative and* $p(x_i)e^{\omega\{1-p(x_i)\}}$ *is viewed as the self-scoring value of event* i *to measure its message importance.*

Actually, from the perspective of generalized Fadeev's postulates, the MIM is viewed as a rational information measure similar to Shannon entropy and Renyi entropy which are respectively defined by

$$H(X) = -\sum_{x_i} p(x_i) \log p(x_i), \tag{2a}$$

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{x_i} \{p(x_i)\}^\alpha, \quad (0 < \alpha < \infty, \alpha \neq 1), \tag{2b}$$

where the condition of variable X is the same as that described in Definition 1. In particular, a postulate for the MIM weaker than that for Shannon entropy and Renyi entropy is given by

$$F(PQ) \leq F(P) + F(Q), \tag{3}$$

while $F(PQ) = F(P) + F(Q)$ is satisfied for Shannon entropy and Renyi entropy [26], where P and Q are two independent random distributions and $F(\cdot)$ denotes a kind of information measure.

Moreover, the crucial operator of MIM to handle probability elements is exponential function while the corresponding operators of Shannon and Renyi entropy are logarithmic function and polynomial function respectively. In this case, MIM can be viewed as a map for the assignments of events' importance weights or the achievement for the self-scoring values of events different from conventional information measures.

As far as the application of MIM is concerned, it may be a better method by using this information measure to detect unbalanced events in signal processing. Ref. [27] has investigated the minor probability event detection by combining MIM and Bayes detection. Moreover, it is worth noting that the physical meaning of the components of MIM corresponds to the normalized optimal data recommendation distribution, which makes a trade-off between the users' preference and system revenue [28]. In this respect, MIM plays a fundamental role in the recommendation system (a popular applications of big data) from the theoretic viewpoint. Therefore, MIM does not come

from the imagination directly, whereas it is a meaningful information measure originated from the practical scenario.

1.2. The Importance Coefficient ϖ in MIM

In general, the parameter ϖ viewed as the importance coefficient has a great impact on the MIM. Actually, different parameter ϖ can lead to different properties and performances for this information measure. In particular, to measure a distribution $P(X) = \{p(x_1), p(x_2), ..., p(x_n)\}$, there are three kinds of work regions of MIM which can be classified by the parameters, whose details are discussed as follows.

(i) If the parameter satisfies $0 \leq \varpi \leq 2/\max\{p(x_i)\}$, the convexity of MIM is similar to Shannon entropy and Renyi entropy. Actually, these three information measures all have maximum value properties and allocate weights for probability elements of the distribution $P(X)$. It is notable that the MIM in this work region focuses on the typical sets rather than atypical sets, which implies that the uniform distribution reaches the maximum value. In brief, the MIM in this work region can be regarded as the same class of message measure as Shannon entropy and Renyi entropy to deal with the problems of information theory.

(ii) If we have $\varpi > 2/\max\{p(x_i)\}$, the small probability elements will be the dominant factor for MIM to measure a distribution. That is, the small probability events can be highlighted more in this work region of MIM than those in the first one. Moreover, in this work region, MIM can pay more attention to atypical sets, which can be viewed as a magnifier for rare events. In fact, this property corresponds to some common scenarios where anomalies catch more eyes such as anomalous detection and alarm. In this case, some problems (including communication and probabilistic events processing) can be rehandled from the perspective of rare events importance. Particularly, the compression encoding and maximum entropy rate transmission are proposed based on the non-parametric MIM (namely NMIM) [24]; in addition, the distribution goodness-of-fit approach is also presented by use of the differential MIM (namely DMIM) [29].

(iii) If the MIM has the parameter $\varpi < 0$, the large probability elements will be the main part contributing to the value of this information measure. In other words, the normal events attract more attention in this work region of MIM than rare events. In practice, this can be used in many applications where regular events are popular such as filter systems and data cleaning.

As a matter of fact, by selecting the parameter ϖ properly, we can exploit the MIM to solve several problems in different scenarios. The importance coefficient facilitates more flexibility of MIM in applications beyond Shannon entropy and Renyi entropy.

To focus on a concrete object, in this paper, we mainly investigate the first work region of MIM (namely $0 \leq \varpi \leq 2/\max\{p(x_i)\}$) and intend to dig out some novelties related to this metric for information processing.

1.3. Similarities and Differences between Shannon Entropy and MIM

In fact, when the parameter ϖ satisfies $0 \leq \varpi \leq 2/\max\{p(x_i)\}$, MIM is similar to Shannon entropy in regard to the expression and properties. The exponential operator of MIM is a substitute for the logarithm operator of Shannon entropy. As a kind of tool based on probability distributions, the MIM with parameter $0 \leq \varpi \leq 2/\max\{p(x_i)\}$ has the same concavity and monotonicity as Shannon entropy, which can characterize the information otherness for different variables.

By resorting to the exponential operator of MIM, the weights for small probability elements are amplified more in some degree than those for large probability ones, which is considered as message importance allocation based on the self-scoring values. In this regard, the MIM may add fresh factors to the information processing, which takes into account the effects of probabilistic events' importance from an objective viewpoint.

In the conventional Shannon information theory, data transmission and compression both can be viewed as the information transfer process from the variable X to Y. The capacity of information transmission is achieved by maximizing the mutual information between the X and Y. Actually, there exists distortion for probabilistic events during an information transfer process, which denotes the difference between the source and its corresponding reconstruction. Due to this fact, it is possible to compress data based on the allowable information loss in a certain extent [30–32]. In Shannon information theory, rate-distortion theory is investigated for lossy data compression, whose essence is mutual information minimization under the constraint of a certain distortion. However, in some cases involved with distortion, small probability events containing more message importance require higher reliability than those with large probability. In this sense, another aspect of information distortion may be essential, in which message importance is considered as a reasonable metric. Particularly, information transfer process is characterized by the MIM (rather than the entropy) with controlling the distortion, which can be viewed as a new kind of information compression, compared to the conventional scheme compressing redundancy to save resources. In fact, some information measures with respect to message importance have been investigated to extend the range of Shannon information theory [33–37]. In this regard, it is worthwhile exploring the information processing in the sense of MIM. Furthermore, it is also promising to investigate the Shannon mutual information constrained by the MIM in an information transfer process which may become a novel system invariant.

In addition, similar to Shannon conditional entropy, a conditional message importance measure for two distributions is proposed to process conditional probability.

Definition 2. *For the two discrete probability* $P(X) = \{p(x_1), p(x_2), ..., p(x_n)\}$ *and* $P(Y) = \{p(y_1), p(y_2), ..., p(y_n)\}$, *the conditional message importance measure (CMIM) is given by*

$$L(\omega, X|Y) = \sum_{y_j} p(y_j) \sum_{x_i} p(x_i|y_j) e^{\omega\{1-p(x_i|y_j)\}}, \tag{4}$$

where $p(x_i|y_j)$ *denotes the conditional probability between* y_j *and* x_i. *The component* $p(x_i|y_j)e^{\omega\{1-p(x_i|y_j)\}}$ *is similar to self-scoring value. Therefore, the CMIM can be considered as a system invariant which indicates the average total self-scoring value for an information transfer process.*

Actually, the MIM is a metric with different mathematical and physical meaning from Shannon entropy and Renyi entropy, which provides its own perspective to process probabilistic events. However, due to the similarity between the MIM and Shannon entropy, they may have analogous performance in some aspects. To this end, the information processing based on the MIM is discussed in this paper.

1.4. Motivation and Contributions

The purpose of this paper is to characterize the probabilistic events processing including compression and transmission by means of MIM. Particularly, in terms of the information processing system model shown in Figure 1, the message source φ (regarded as a random variable whose support set corresponds to the set of events' types) can be measured by the amount of information $H(\cdot)$ and the message importance $L(\cdot)$ according to the probability distribution. Then, the information transfer process whose details are presented in Section 2 can be characterized based on these two metrics. Different from the mathematically probabilistic characterization of traditional telecommunication system, this paper mainly discusses the information processing from the perspectives of message importance. In this regard, the information importance harvest in a transmission is characterized by the proposed message importance loss capacity. Moreover, the upper bound of information compression based on the MIM is described by the message importance distortion function. In addition, we also investigate the trade-off between bitrate transmission and message importance loss to bring some inspiration to the conventional information theory.

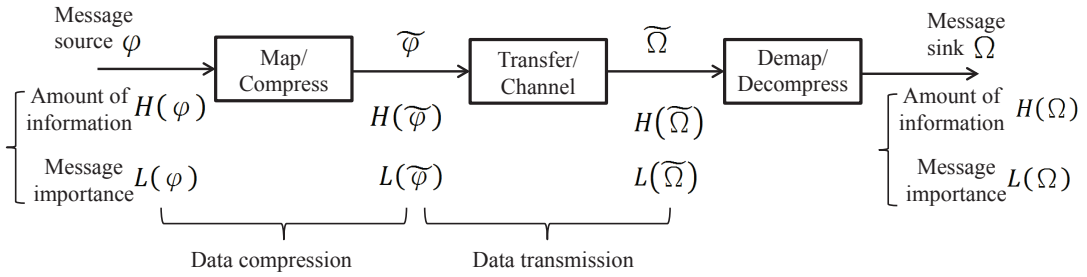

Figure 1. Information processing system model.

1.5. Organization

The rest of this paper is discussed as follows. In Section 2, a system model involved with message importance is constructed to help analyze the data compression and transmission in big data. In Section 3, we propose a kind of message transfer capacity to investigate the message importance loss in the transmission. In Section 4, message importance distortion function is introduced and its properties are also presented to give some details. In Section 5, we discuss the bitrate transmission constrained by message importance to widen the horizon for the Shannon theory. In Section 6, some numerical results are presented to validate propositions and the analysis in theory. Finally, we conclude this paper in Section 7. Additionally, the fundamental notations in this paper are summarized in Table 1.

Table 1. Notations.

Notation	Description
$P(X) = \{p(x_1), p(x_2), ..., p(x_n)\}$	The discrete probability distribution with respect to the variable X
φ	The message source in the information processing system model
$\widetilde{\varphi}$	The mapped or compressed message with respect to the φ
$\widetilde{\Omega}$	The received message transferred from the $\widetilde{\varphi}$
Ω	The recovered message with respect to the φ by the decoding process
ω	The importance coefficient
$L(\cdot)$	The message importance measure (MIM) described as Definition 1
$H(\cdot)$	The Shannon entropy, $H(X) = -\sum_{x_i} p(x_i) \log p(x_i)$ or $H(p) = -p \log p - (1-p) \log(1-p), (0 \leq p \leq 1)$
$H_\alpha(\cdot)$	The Renyi entropy with the parameter α $H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{x_i} \{p(x_i)\}^\alpha$
$L(\cdot\|\cdot)$	The CMIM described as Definition 2
$H(\cdot\|\cdot)$	The conditional Shannon entropy, $H(X\|Y) = \sum_{x_i} \sum_{y_j} p(x_i, y_j) \log \frac{1}{p(x_i\|y_j)}$
$\Phi_\omega(\cdot\|\|\cdot)$	The message importance loss described as Definition 3
C	the message importance loss capacity (MILC) described as Definition 4
$p(y\|x)$	An information transfer matrix from the variable X to Y
$\{X, p(y\|x), Y\}$	An information transfer process from the variable X to Y
β_s,β_e, β_k	The parameters in the binary symmetric matrix, binary eraser matrix and k-ary symmetric matrix respectively
$d(x, y)$	The distortion function, $d(x, y) \geq 0$
D	The allowable distortion ($D_{\min} \leq D \leq D_{\max}$)
$\bar{D}$	The average distortion, $\bar{D} = \sum_{x_i} \sum_{y_j} p(x_i) p(y_j\|x_i) d(x_i, y_j)$
B_D	The the allowable information transfer matrix set $B_D = \{q(y\|x) : \bar{D} \leq D\}$
$R_\omega(D)$	The message importance distortion function described as Definition 5
$I(X\|\|Y)$	Mutual information, $I(X\|\|Y) = \sum_{x_i} \sum_{y_j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)}$

2. System Model with Message Importance

Considering an information processing system model shown in Figure 1, the information transfer process is discussed as follows. At first, a message source φ follows a distribution $P_\varphi = \{p(\varphi_1), p(\varphi_2), ..., p(\varphi_n)\}$ whose support set is $\{\varphi_1, \varphi_2, ..., \varphi_n\}$ corresponding to the events types. Then, the message φ is encoded or compressed into the variable $\widetilde{\varphi}$ following the distribution $P_{\widetilde{\varphi}} = \{p(\widetilde{\varphi}_1), p(\widetilde{\varphi}_2), ..., p(\widetilde{\varphi}_n)\}$ whose alphabet is $\{\varphi_1, \varphi_2, ..., \varphi_n\}$. After the information transfer process denoted by matrix $p(\widetilde{\Omega}_j|\widetilde{\varphi}_i)$, the received message $\widetilde{\Omega}$ originating from $\widetilde{\varphi}$ is observed as a random variable, where the distribution of $\widetilde{\Omega}$ is $P_{\widetilde{\Omega}} = \{p(\widetilde{\Omega}_1), p(\widetilde{\Omega}_2), ..., p(\widetilde{\Omega}_n)\}$ whose alphabet is $\{\widetilde{\Omega}_1, \widetilde{\Omega}_2, ..., \widetilde{\Omega}_n\}$. Finally, the receiver recovers the original message φ by decoding $\Omega = g(\widetilde{\Omega})$ where $g(\cdot)$ denotes the decoding function and Ω is the recovered message with the alphabet $\{\Omega_1, \Omega_2, ..., \Omega_n\}$.

From the viewpoint of generalized information theory, a two-layer framework is considered to understand this model, where the first layer is based on the amount of information characterized by Shannon entropy denoted by $H(\cdot)$, while the second layer reposes on message importance measure of events denoted by $L(\cdot)$. Due to the fact that the former is discussed pretty entirely, we mainly investigate the latter in the paper.

Considering the source-channel separation theorem [38], the above information processing model consists of two problems, namely data compression and data transmission. On one hand, the *data compression* of the system can be achieved by using classical source coding strategies to reduce more redundancy, in which the information loss is described by $H(\varphi) - H(\varphi|\widetilde{\varphi})$ under the information transfer matrix $p(\widetilde{\varphi}|\varphi)$. Similarly, from the perspective of message importance, the data can be further compressed by discarding worthless messages, where the message importance loss can be characterized by $L(\varphi) - L(\varphi|\widetilde{\varphi})$. On the other hand, the *data transmission* is discussed to obtain the upper bound of the mutual information $H(\widetilde{\varphi}) - H(\widetilde{\varphi}|\widetilde{\Omega})$, namely the information capacity. In a similar way, $L(\widetilde{\varphi}) - L(\widetilde{\varphi}|\widetilde{\Omega})$ means the income of message importance in the transmission.

In essence, it is apparent that the data compression and transmission are both considered as an information transfer processes $\{X, p(y|x), Y\}$, and they can be characterized by the difference between $\{X\}$ and $\{X|Y\}$. In order to facilitate the analysis of the above model, the message importance loss is introduced as follows.

Definition 3. *For two discrete probability* $P(X) = \{p(x_1), p(x_2), ..., p(x_n)\}$ *and* $P(Y) = \{p(y_1), p(y_2), ..., p(y_n)\}$*, the message importance loss based on MIM and CMIM is given by*

$$\Phi_\varpi(X||Y) = L(\varpi, X) - L(\varpi, X|Y), \tag{5}$$

where $L(\varpi, X)$ *and* $L(\varpi, X|Y)$ *are given by the Definitions 1 and 2.*

In fact, according to the intrinsic relationship between $L(\varpi, X)$ and $L(\varpi, X|Y)$, it is readily seen that

$$\Phi_\varpi(X||Y) \geq 0, \tag{6}$$

where $0 < \varpi \leq 2 \leq 2/\max\{p(x_i|y_j)\}$.

Proof. Considering a function $f(x) = xe^{\varpi(1-x)}$ $(0 \leq x \leq 1$ and $0 < \varpi)$, it is easy to have $\frac{\partial^2 f(x)}{\partial x} = -\varpi e^{\varpi(1-x)}(2 - \varpi x)$, which implies if $\varpi \leq 2 \leq 2/x$, the function $f(x)$ is concave.

In the light of Jensen's inequality, if $0 < \varpi \leq 2 \leq 2/\max\{p(x_i|y_j)\}$ is satisfied, it is not difficult to see

$$\begin{aligned} L(\varpi, X) &= \sum_{x_i} p(x_i) e^{\varpi(1-p(x_i))} \\ &= \sum_{x_i} \{\sum_{y_j} p(y_j) p(x_i|y_j)\} e^{\varpi(1-\{\sum_{y_j} p(y_j) p(x_i|y_j)\})} \\ &\geq \sum_{y_j} p(y_j) \sum_{x_i} \{p(x_i|y_j) e^{\varpi(1-p(x_i|y_j))}\} = L(\varpi, X|Y). \end{aligned} \tag{7}$$

□

3. Message Importance Loss in Transmission

In this section, we will introduce the CMIM to characterize the information transfer processing. To do so, we define a kind of message transfer capacity measured by the CMIM as follows.

Definition 4. *Assume that there exists an information transfer process as*

$$\{X, p(y|x), Y\}, \tag{8}$$

where the $p(y|x)$ denotes a probability distribution matrix describing the information transfer from the variable X to Y. We define the message importance loss capacity (MILC) as

$$\begin{aligned} C &= \max_{p(x)} \{\Phi_\varpi(X||Y)\} \\ &= \max_{p(x)} \{L(\varpi, X) - L(\varpi, X|Y)\}, \end{aligned} \tag{9}$$

where $L(\varpi, X) = \sum_{x_i} p(x_i) e^{\varpi\{1-p(x_i)\}}$, $p(y_j) = \sum_{x_i} p(x_i) p(y_j|x_i)$, $p(x_i|y_j) = \frac{p(x_i)p(y_j|x_i)}{p(y_j)}$, $L(\varpi, X|Y)$ is defined by Equation (4), and $\varpi < 2 \leq 2/\max\{p(x_i)\}$.

In order to have an insight into the applications of MILC, some specific information transfer scenarios are discussed as follows.

3.1. Binary Symmetric Matrix

Consider the binary symmetric information transfer matrix, where the original variables are complemented with the transfer probability which can be seen in the following proposition.

Proposition 1. *Assume that there exists an information transfer process $\{X, p(y|x), Y\}$, where the information transfer matrix is*

$$p(y|x) = \begin{bmatrix} 1-\beta_s & \beta_s \\ \beta_s & 1-\beta_s \end{bmatrix}, \tag{10}$$

which indicates that X and Y both follow binary distributions. In that case, we have

$$C(\varpi, \beta_s) = e^{\frac{\varpi}{2}} - L(\varpi, \beta_s), \tag{11}$$

where $L(\varpi, \beta_s) = \beta_s e^{\varpi(1-\beta_s)} + (1-\beta_s) e^{\varpi\beta_s}$ ($0 \leq \beta_s \leq 1$) and $\varpi < 2 \leq 2/\max\{p(x_i)\}$.

Proof of Proposition 1. Assume that the distribution of variable X is a binary distribution $(p, 1-p)$. According to Equation (10) and Bayes' theorem (namely, $p(x|y) = \frac{p(x)p(y|x)}{p(y)}$), it is not difficult to see that

$$p(x|y) = \begin{bmatrix} \frac{p(1-\beta_s)}{p(1-\beta_s)+(1-p)\beta_s} & \frac{(1-p)\beta_s}{p(1-\beta_s)+(1-p)\beta_s} \\ \frac{p\beta_s}{p\beta_s+(1-p)(1-\beta_s)} & \frac{(1-p)(1-\beta_s)}{p\beta_s+(1-p)(1-\beta_s)} \end{bmatrix}. \tag{12}$$

Furthermore, in accordance with Equations (4) and (9), we have

$$\begin{aligned} &C(\varpi, \beta_s) \\ &= \max_p \{C(p, \varpi, \beta_s)\} \\ &= \max_p \Big\{ L(\varpi, p) - \big\{ p(1-\beta_s) e^{\frac{\varpi(1-p)\beta_s}{p(1-\beta_s)+(1-p)\beta_s}} + (1-p)\beta_s e^{\frac{\varpi p(1-\beta_s)}{p(1-\beta_s)+(1-p)\beta_s}} + p\beta_s e^{\frac{\varpi(1-p)(1-\beta_s)}{p\beta_s+(1-p)(1-\beta_s)}} \\ &\quad + (1-p)(1-\beta_s) e^{\frac{\varpi p\beta_s}{p\beta_s+(1-p)(1-\beta_s)}} \big\} \Big\}, \end{aligned} \tag{13}$$

where $L(\varpi, p) = pe^{\varpi(1-p)} + (1-p)e^{\varpi p}$ $(0 < p < 1)$. Then, it is readily seen that

$$\begin{aligned} \frac{\partial C(p, \varpi, \beta_s)}{\partial p} =& (1-\varpi p)e^{\varpi(1-p)} + [(1-p)\varpi - 1]e^{\varpi p} \\ &- \Big\{ (1-\beta_s) \Big\{ 1 - \frac{\varpi p(1-\beta_s)\beta_s}{[p(1-\beta_s)+(1-p)\beta_s]^2} \Big\} e^{\frac{\varpi(1-p)\beta}{p(1-\beta)+(1-p)\beta}} \\ &+ (1-\beta_s) \Big\{ \frac{\varpi(1-p)\beta_s(1-\beta_s)}{[p\beta_s+(1-p)(1-\beta_s)]^2} - 1 \Big\} e^{\frac{\varpi p\beta_s}{p\beta_s+(1-p)(1-\beta_s)}} \\ &+ \beta_s \Big\{ \frac{\varpi(1-p)\beta_s(1-\beta_s)}{[p(1-\beta_s)+(1-p)\beta_s]^2} - 1 \Big\} e^{\frac{\varpi p(1-\beta_s)}{p(1-\beta_s)+(1-p)\beta_s}} \\ &+ \beta_s \Big\{ 1 - \frac{\varpi p(1-\beta_s)\beta_s}{[p\beta_s+(1-p)(1-\beta_s)]^2} \Big\} e^{\frac{\varpi(1-p)(1-\beta_s)}{p\beta_s+(1-p)(1-\beta_s)}} \Big\}. \end{aligned} \tag{14}$$

In the light of the positivity for $\frac{\partial C(p,\beta_s)}{\partial p}$ in $\{p|p \in (0, 1/2)\}$ and the negativity in $\{p|p \in (1/2, 1)\}$ (if $\beta_s \neq 1/2$), it is apparent that $p = 1/2$ is the only solution for $\frac{\partial C(p,\beta_s)}{\partial p} = 0$. That is, if $\beta_s \neq 1/2$, the extreme value is indeed the maximum value of $C(p, \varpi, \beta_s)$ when $p = 1/2$. Similarly, if $\beta_s = 1/2$, the solution $p = 1/2$ also results in the same conclusion. □

Remark 1. *According to Proposition 1, on one hand, when $\beta_s = 1/2$, that is, the information transfer process is just random, we will gain the lower bound of the MILC namely $C(\beta_s) = 0$. On the other hand, when $\beta_s = 0$, namely there is a certain information transfer process, we will have the maximum MILC. As for the distribution selection for the variable X, the uniform distribution is preferred to gain the capacity.*

3.2. Binary Erasure Matrix

The binary erasure information transfer matrix is similar to the binary symmetric one; however, in the former, a part of information is lost rather than corrupted. The MILC of this kind of information transfer matrix is discussed as follows.

Proposition 2. *Consider an information transfer process $\{X, p(y|x), Y\}$, in which the information transfer matrix is described as*

$$p(y|x) = \begin{bmatrix} 1-\beta_e & 0 & \beta_e \\ 0 & 1-\beta_e & \beta_e \end{bmatrix}, \tag{15}$$

which indicates that X follows the binary distribution and Y follows the 3*-ary distribution. Then, we have*

$$C(\varpi, \beta_e) = (1-\beta_e)\{e^{\frac{\varpi}{2}} - 1\}, \tag{16}$$

where $0 \leq \beta_e \leq 1$ *and* $0 < \varpi < 2 \leq 2/\max\{p(x_i)\}$.

Proof of Proposition 2. Assume the distribution of variable X is $(p, 1-p)$. Furthermore, according to the binary erasure matrix and Bayes theorem, we have that the transmission matrix conditioned by the variable Y as follows:

$$p(x|y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ p & 1-p \end{bmatrix}. \tag{17}$$

Then, it is not difficult to have

$$L(\varpi, X|Y) = \beta_e p e^{\varpi(1-p)} + \beta_e(1-p)e^{\varpi p} + 1 - \beta_e. \tag{18}$$

Furthermore, it is readily seen that

$$\begin{aligned} C(p, \varpi, \beta_e) &= \max_p \Big\{ L(\varpi, p) - \big\{ \beta_e p e^{\varpi(1-p)} + \beta_e(1-p)e^{\varpi p} + 1 - \beta_e \big\} \Big\} \\ &= (1-\beta_e)\big\{ \max_p \{L(\varpi, p)\} - 1 \big\}, \end{aligned} \tag{19}$$

where $L(\varpi, p) = pe^{\varpi(1-p)} + (1-p)e^{\varpi p}$. Moreover, we have the solution $p = 1/2$ leads to $\frac{\partial L(\varpi, p)}{\partial p} = 0$ and the corresponding second derivative is

$$\frac{\partial^2 L(\varpi, p)}{\partial p^2} = e^{\varpi(1-p)}(\varpi p - 2)\varpi + e^{\varpi p}[(1-p)\varpi - 2]\varpi < 0, \tag{20}$$

which results from the condition $0 < \varpi < 2 \leq 2/\max\{p(x_i)\}$.

Therefore, it is readily seen that, in the case $p = 1/2$, the capacity $C(p, \varpi, \beta_e)$ reaches the maximum value. □

Remark 2. *Proposition 2 indicates that, in the case* $\beta_e = 1$*, the lower bound of the capacity is obtained, that is* $C(\beta_e) = 0$*. However, if a certain information transfer process is satisfied (namely* $\beta_e = 0$*), we will have the maximum MILC. Similar to Proposition 1, the uniform distribution is selected to reach the capacity in practice.*

3.3. Strongly Symmetric Backward Matrix

As for a strongly symmetric backward matrix, it is viewed as a special example of information transmission. The discussion for the message transfer capacity in this case is similar to that in the symmetric matrix, whose details are given as follows.

Proposition 3. *For an information transmission from the source X to the sink Y, assume that there exists a strongly symmetric backward matrix as follows:*

$$p(x|y) = \begin{bmatrix} 1-\beta_k & \frac{\beta_k}{K-1} & \cdots & \frac{\beta_k}{K-1} \\ \frac{\beta_k}{K-1} & 1-\beta_k & \cdots & \frac{\beta_k}{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_k}{K-1} & \cdots & \frac{\beta_k}{K-1} & 1-\beta_k \end{bmatrix}, \tag{21}$$

which indicates that X and Y both obey K-ary distribution. We have

$$C(\omega, \beta_k) = e^{\frac{\omega(K-1)}{K}} - \{(1-\beta_k)e^{\omega\beta_k} + \beta_k e^{\omega(1-\frac{\beta_k}{K-1})}\}, \tag{22}$$

where $0 \leq \beta_k \leq 1$, $K \geq 2$ *and* $0 < \omega < 2 \leq 2/\max\{p(x_i)\}$.

Proof of Proposition 3. For given K-ary variables X and Y whose distribution are $\{p(x_1), p(x_2), ..., p(x_K)\}$ and $\{p(y_1), p(y_2), ..., p(y_K)\}$ respectively, we can use the strongly symmetric backward matrix to obtain the relationship between the two variables as follows:

$$p(x_i) = (1-\beta_k)p(y_i) + \frac{\beta_k}{K-1}[1-p(y_i)] \quad (i = 1, 2, ..., K), \tag{23}$$

which implies $p(x_i)$ is a one-to-one onto function for $p(y_i)$.

In accordance with Definition 2, it is easy to see that

$$\begin{aligned} L(\omega, X|Y) &= \sum_{x_i}\sum_{y_j} p(y_j)p(x_i|y_j)e^{\omega(1-p(x_i|y_j))} \\ &= \sum_{y_j} p(y_j)\{(1-\beta_k)e^{\omega\beta_k} + \beta_k e^{\omega(1-\frac{\beta_k}{K-1})}\} \\ &= (1-\beta_k)e^{\omega\beta_k} + \beta_k e^{\omega(1-\frac{\beta_k}{K-1})}. \end{aligned} \tag{24}$$

Moreover, by virtue of the definition of MILC in Equation (9), it is readily seen that

$$C(\omega, \beta_k) = \max_{p(x)}\{L(\omega, X)\} - [\,(1-\beta_k)e^{\omega\beta_k} + \beta_k e^{\omega(1-\frac{\beta_k}{K-1})}], \tag{25}$$

where $L(\omega, X) = \sum_{x_i} p(x_i)e^{\omega\{1-p(x_i)\}}$.

Then, by using Lagrange multiplier method, we have

$$G(p(x_i), \lambda_0) = \sum_{x_i} p(x_i)e^{\omega(1-p(x_i))} + \lambda_0\big[\sum_{x_i} p(x_i) - 1\big]. \tag{26}$$

By setting $\frac{\partial G(p(x_i),\lambda_0)}{\partial p(x_i)} = 0$ and $\frac{\partial G(p(x_i),\lambda_0)}{\partial \lambda_0} = 0$, it can be readily verified that the extreme value of $\sum_{y_j} p(y_j)e^{\omega(1-p(y_j))}$ is achieved by the uniform distribution as a solution, that is $p(x_1) = p(x_2) = ... = p(x_K) = 1/K$. In the case that $0 < \omega < 2 \leq 2/\max\{p(x_i)\}$, we have $\frac{\partial^2 G(p(x_i),\lambda_0)}{\partial p^2(x_i)} < 0$ with respect to $p(x_i) \in [0, 1]$, which implies that the extreme value of $\sum_{x_i} p(x_i)e^{\omega(1-p(x_i))}$ is the maximum value.

In addition, according to the Equation (23), the uniform distribution of variable X is resulted from the uniform distribution for variable Y.

Therefore, by substituting the uniform distribution for $p(x)$ into Equation (25), we will obtain the capacity $C(\omega, \beta_k)$. □

Furthermore, in light of Equation (22), we have

$$\frac{\partial C(\omega, \beta_k)}{\partial \beta_k} = \{1-\omega(1-\beta_k)\}e^{\omega\beta_k} + \Big\{\frac{\omega\beta_k}{K-1} - 1\Big\}e^{\omega(1-\frac{\beta_k}{K-1})}. \tag{27}$$

By setting $\frac{\partial C(\omega,\beta_k)}{\partial \beta_k} = 0$, it is apparent that $C(\omega, \beta_k)$ reaches the extreme value in the case that $\beta_k = \frac{K-1}{K}$. Additionally, when the parameter ω satisfies $0 < \omega < 2 \leq 2/\max\{p(x_i)\}$, we also have the second derivative of the $C(\omega, \beta_k)$ as follows:

$$\frac{\partial^2 C(\omega, \beta_k)}{\partial \beta_k^2} = \omega[2 - (1 - \beta_k)\omega]e^{\omega\beta_k} + \frac{\omega}{K-1}\Big\{2 - \frac{\omega\beta_k}{K-1}\Big\}e^{\omega(1-\frac{\beta_k}{K-1})} > 0, \tag{28}$$

which indicates that the convex $C(\omega, \beta_k)$ reaches the minimum value 0 in the case $\beta_k = \frac{K-1}{K}$.

Remark 3. *According to Proposition 3, when $\beta_k = \frac{K-1}{K}$, namely, the channel is just random, we gain the lower bound of the capacity namely $C(\omega, \beta_k) = 0$. On the contrary, when $\beta_k = 0$ (that is, there is a certain channel), we will have the maximum capacity.*

4. Distortion of Message Importance Transfer

In this section, we will focus on the information transfer distortion, a common problem of information processing. In a real information system, there exists inevitable information distortion caused by noises or other disturbances, though the devices and hardware of telecommunication systems are updating and developing. Fortunately, there are still some bonuses from allowable distortion in some scenarios. For example, in conventional information theory, rate distortion is exploited to obtain source compression such as predictive encoding and hybrid encoding, which can save a lot of hardware resources and communication traffic [39].

Similar to the rate distortion theory for Shannon entropy [38], a kind of information distortion function based on MIM and CMIM is defined to characterize the effect of distortion on the message importance loss. In particular, there are some details of discussion as follows.

Definition 5. *Assume that there exists an information transfer process $\{X, p(y|x), Y\}$ from the variable X to Y, where the $p(y|x)$ denotes a transfer matrix (distributions of X and Y are denoted by $p(x)$ and $p(y)$ respectively). For a given distortion function $d(x, y)$ ($d(x, y) \geq 0$) and an allowable distortion D, the message importance distortion function is defined as*

$$\begin{aligned} R_\omega(D) &= \min_{p(y|x)\in B_D} \Phi_\omega(X||Y) \\ &= \min_{p(y|x)\in B_D} \{L(\omega, X) - L(\omega, X|Y)\}, \end{aligned} \tag{29}$$

in which $L(\omega, X) = \sum_{x_i} p(x_i)e^{\omega\{1-p(x_i)\}}$, $L(\omega, X|Y)$ is defined by Equation (4), $0 < \omega \leq \frac{2\min_j\{p(y_j)\}}{\max_i\{p(x_i)\}}$ and $B_D = \{q(y|x) : \bar{D} \leq D\}$ denotes the allowable information transfer matrix set where

$$\bar{D} = \sum_{x_i}\sum_{y_j} p(x_i)p(y_j|x_i)d(x_i, y_j), \tag{30}$$

which is the average distortion.

In this model, the information source X is given and our goal is to select an adaptive $p(y|x)$ to achieve the minimum allowable message importance loss under the distortion constraint. This provides a new theoretical guidance for information source compression from the perspective of message importance.

In contrast to the rate distortion of Shannon information theory, this new information distortion function just depends on the message importance loss rather than entropy loss to choose an appropriate information compression matrix. In practice, there are some similarities and differences between the rate distortion theory and the message importance distortion in terms of the source compression. On one hand, both two information distortion encodings can be regarded as special information

transfer processes just with different optimization objectives. On the other hand, the new distortion theory tries to keep the rare events as high as possible, while the conventional rate distortion focuses on the amount of information itself. To some degree, by reducing more redundant common information, the new source compression strategy based on rare events (viewed as message importance) may save more computing and storage resources in big data.

4.1. Properties of Message Importance Distortion Function

In this subsection, we shall discuss some fundamental properties of rate distortion function based on message importance in details.

4.1.1. Domain of Distortion

Here, we investigate the domain of allowable distortion, namely $[D_{\min}, D_{\max}]$, and the corresponding message importance distortion function values as follows.

(i) The lower bound $D_{\min}$: Due to the fact $0 \leq d(x_i, y_j)$, it is easy to obtain the non-negative average distortion, namely $0 \leq \bar{D}$. Considering $\bar{D} \leq D$, we readily have the minimum allowable distortion, that is

$$D_{\min} = 0, \tag{31}$$

which implies the distortionless case, namely Y is the same as X.

In addition, when the lower bound $D_{\min}$ (namely the distortionless case) is satisfied, it is readily seen that

$$L(\omega, X|Y) = L(\omega, X|X) = \sum_{x_i} p(x_i)p(x_i|x_i)e^{\omega\{1-p(x_i|x_i)\}} = 1, \tag{32}$$

and according to the Equation (29) the message importance distortion function is

$$R_\omega(D_{\min}) = L(\omega, X) - L(\omega, X|X) = L(\omega, X) - 1, \tag{33}$$

where $L(\omega, X) = \sum_{x_i} p(x_i)e^{\omega\{1-p(x_i)\}}$ and $0 < \omega \leq \frac{2\min_j\{p(y_j)\}}{\max_i\{p(x_i)\}}$.

(ii) The upper bound $D_{\max}$: When the allowable distortion satisfies $D \geq D_{\max}$, it is apparent that the variables X and Y are independent, that is, $p(y|x) = p(y)$. Furthermore, it is not difficult to see that

$$\begin{aligned} D_{\max} &= \min_{p(y)} \Big\{ \sum_{x_i}\sum_{y_j} p(x_i)p(y_j)d(x_i, y_j) \Big\} \\ &= \sum_{y_j} p(y_j) \min_{p(y)} \Big\{ \sum_{x_i} p(x_i)d(x_i, y_j) \Big\} \\ &\geq \min_{y_j} \Big\{ \sum_{x_i} p(x_i)d(x_i, y_j) \Big\}, \end{aligned} \tag{34}$$

which indicates that when the distribution of variable Y follows $p(y_j) = 1$ and $p(y_l) = 0$ $(l \neq j)$, we have the upper bound

$$D_{\max} = \min_{y_j} \Big\{ \sum_{x_i} p(x_i)d(x_i, y_j) \Big\}. \tag{35}$$

Additionally, on account of the independent X and Y, namely $p(x|y) = p(x)$, it is readily seen that

$$R_\omega(D_{\max}) = L(\omega, X) - \sum_{y_j} p(y_j)L(\omega, X) = 0. \tag{36}$$

4.1.2. The Convexity Property

For two allowable distortions D_a and D_b, whose optimal allowable information transfer matrixes are $p_a(y|x)$ and $p_b(y|x)$ respectively, we have

$$R_\omega(\delta D_a + (1-\delta)D_b) \leq \delta R_\omega(D_a) + (1-\delta)R_\omega(D_b), \tag{37}$$

where $0 \leq \delta \leq 1$ and $0 < \varpi \leq \frac{2\min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

Proof. Refer to the Appendix A. □

4.1.3. The Monotonically Decreasing Property

For two given allowable distortions D_a and D_b, if $0 \leq D_a < D_b < D_{\max}$ is satisfied, we have $R_\varpi(D_a) \geq R_\varpi(D_b)$, where $0 < \varpi \leq \frac{2\min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

Proof. Considering that $0 \leq D_a < D_b < D_{\max}$, we have $D_b = \gamma D_a + (1-\gamma) D_{\max}$ where $\gamma = \frac{D_{\max} - D_b}{D_{\max} - D_a}$. On account of the Equation (36) and the convexity property mentioned in Equation (37), it is not difficult to see that

$$R_\varpi(D_b) \leq \gamma R_\varpi(D_a) + (1-\gamma) R_\varpi(D_{\max}) = \gamma R_\varpi(D_a) < R_\varpi(D_a), \tag{38}$$

where $0 < \gamma < 1$. □

4.1.4. The Equivalent Expression

For an information transfer process $\{X, p(y|x), Y\}$, if we have a given distortion function $d(x, y)$, an allowable distortion D and a average distortion $\bar{D}$ defined in Equation (30), the message importance distortion function defined in Equation (29) can be rewritten as

$$R_\varpi(D) = \min_{\bar{D}=D} \{L(\varpi, X) - L(\varpi, X|Y)\}, \tag{39}$$

where $L(\varpi, X)$ and $L(\varpi, X|Y)$ are defined by the Equations (1) and (4), as well as $0 < \varpi \leq \frac{2\min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

Proof. For a given allowable distortion D, if there exists an allowable distortion D^* ($D_{\min} \leq D^* < D < D_{\max}$) and the corresponding optimal information transfer matrix $p^*(y|x)$ leads to $R_\varpi(D)$, we will have $R_\varpi(D) = R_\varpi(D^*)$, which contradicts the monotonically decreasing property. □

4.2. Analysis for Message Importance Distortion Function

In this subsection, we shall investigate the computation of message importance distortion function, which has a great impact on the probabilistic events analysis in practice. Actually, the definition of message importance distortion function in Equation (29) can be regarded as a special function, which is the minimization of the message importance loss with the symbol error less than or equal to the allowable distortion D. In particular, Definition 5 can also be expressed as the following optimization:

$$\begin{aligned} \mathcal{P}_1 : \min_{p(y_j|x_i)} \quad & \{L(\varpi, X) - L(\varpi, X|Y)\} \\ \text{s.t.} \quad & \sum_{x_i} \sum_{y_j} p(x_i) p(y_j|x_i) d(x_i, y_j) \leq D, \\ & \sum_{y_j} p(y_j|x_i) = 1, \\ & p(y_j|x_i) \geq 0, \end{aligned} \tag{40}$$

where $L(\varpi, X)$ and $L(\varpi, X|Y)$ are MIM and CMIM defined in Equations (1) and (4), as well as $0 < \varpi \leq \frac{2\min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

To take a computable optimization problem as an example, we consider Hamming distortion as the distortion function $d(x, y)$, namely

$$d(x,y) = \begin{bmatrix} 0 & 1 & ... & 1 \\ 1 & 0 & ... & 1 \\ ... & ... & & ... \\ 1 & 1 & ... & 0 \end{bmatrix}, \tag{41}$$

which means $d(x_i, y_i) = 0$ and $d(x_i, y_j) = 1$ $(i \neq j)$. In order to reveal some intrinsic meanings of $R_\omega(D)$, we investigate an information transfer of Bernoulli source as follows.

Proposition 4. *For a Bernoulli(p) source denoted by a variable X and an information transfer process* $\{X, p(y|x), Y\}$ *with Hamming distortion, the message importance distortion function is given by*

$$R_\omega(D) = \{pe^{\omega(1-p)} + (1-p)e^{\omega p}\} - \{De^{\omega(1-D)} + (1-D)e^{\omega D}\}, \tag{42}$$

and the corresponding information transfer matrix is

$$p(y|x) = \begin{bmatrix} \frac{(1-D)(p-D)}{p(1-2D)} & \frac{(1-p-D)D}{p(1-2D)} \\ \frac{D(p-D)}{(1-p)(1-2D)} & \frac{(1-p-D)(1-D)}{(1-p)(1-2D)} \end{bmatrix}, \tag{43}$$

where $0 < \omega \leq \frac{2\min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$ *and* $0 \leq D \leq \min\{p, 1-p\}$.

Proof of Proposition 4. Refer to the Appendix B. □

5. Bitrate Transmission Constrained by Message Importance

We investigate the information capacity in the case of a limited message importance loss in this section. The objective is to achieve the maximum transmission bitrate under the constraint of a certain message importance loss ϵ. The maximum transmission bitrate is one of system invariants in a transmission process, which provides a upper bound of amount of information obtained by the receiver.

In an information transmission process, the information capacity is the mutual information between the encoded signal and the received signal with the dimension bit/symbol. In a real transmission, there always exists an allowable distortion between the sending sequence X and the received sequence Y, while the maximum allowable message importance loss is required to avoid too much distortion of important events. From this perspective, message importance loss is considered to be another constraint for the information transmission capacity beyond the information distortion. Therefore, this might play a crucial role in the design of transmission in information processing systems.

In particular, we characterize the maximizing mutual information constrained by a controlled message importance loss as follows:

$$\begin{aligned} \mathcal{P}_2: \max_{p(x)} \ & I(X||Y) \\ \text{s.t.} \ & L(\omega, X) - L(\omega, X|Y) \leq \epsilon, \\ & \sum_{y_j} p(x_i) = 1, \\ & p(x_i) \geq 0, \end{aligned} \tag{44}$$

where $I(X||Y) = \sum_{x_i, y_j} p(x_i)p(y_j|x_i) \log \frac{p(x_i)p(y_j|x_i)}{p(y_j)}$, $p(y_j) = \sum_{x_i} p(x_i)p(y_j|x_i)$, $L(\omega, X)$ and $L(\omega, X|Y)$ are MIM and CMIM defined in Equations (1) and (4), as well as $0 < \omega \leq \frac{2\min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

Actually, the bitrate transmission with a message importance loss constraint has a special solution for a certain scenario. In order to give a specific example, we investigate the optimization problem in the Bernoulli(p) source with a symmetric or erasure transfer matrix as follows.

5.1. Binary Symmetric Matrix

Proposition 5. *For a Bernoulli(p) source X whose distribution is* $\{p, 1-p\}$ $(0 \leq p \leq 1/2)$ *and an information transfer process* $\{X, p(y|x), Y\}$ *with transfer matrix*

$$p(y|x) = \begin{bmatrix} 1-\beta_s & \beta_s \\ \beta_s & 1-\beta_s \end{bmatrix}, \tag{45}$$

we have the solution for $\mathcal{P}_2$ *defined in Equation (44) as follows:*

$$\begin{aligned} &\max_{p(x)} I(X||Y) \\ &= \begin{cases} 1 - H(\beta_s), \ (\epsilon \geq C_{\beta_s}) \\ H(p_s(1-\beta_s) + (1-p_s)\beta_s) - H(\beta_s), \ (0 < \epsilon \leq C_{\beta_s}), \end{cases} \end{aligned} \tag{46}$$

where p_s *is the solution of* $L(\varpi, X) - L(\varpi, X|Y) = \epsilon$ ($L(\varpi, X)$ *and* $L(\varpi, X|Y)$ *mentioned in the optimization problem* $\mathcal{P}_2$*), whose approximate value is*

$$p_s \doteq \frac{1 - \sqrt{\Theta}}{2}, \tag{47}$$

in which the parameter Θ *is given by*

$$\Theta = 1 - \frac{4\epsilon}{4\varpi + \varpi^2} - \frac{4\sqrt{(1-2\beta_s)^2\epsilon^2 + 2(4\varpi + \varpi^2)\beta_s(1-\beta_s)\epsilon}}{(4\varpi + \varpi^2)|1-2\beta_s|}, \tag{48}$$

and $H(\cdot)$ *denotes the operator for Shannon entropy, that is* $H(p) = -[(1-p)\log(1-p) + p\log p]$ *,* $C_{\beta_s} = e^{\frac{\varpi}{2}} - \{\beta_s e^{\varpi(1-\beta_s)} + (1-\beta_s)e^{\varpi\beta_s}\}$ $(0 \leq \beta_s \leq 1)$ *and* $\varpi < 2 \leq 2/\max\{p(x_i)\}$.

Proof of Proposition 5. Considering the Bernoulli(p) source X following $\{p, 1-p\}$ and the binary symmetric matrix, it is not difficult to gain

$$\begin{aligned} I(X||Y) &= H(Y) - H(Y|X) \\ &= -\{p(y_0)\log p(y_0) + p(y_1)\log p(y_1)\} - H(\beta_s), \end{aligned} \tag{49}$$

where $p(y_0) = p(1-\beta_s) + (1-p)\beta_s$, $p(y_1) = p\beta_s + (1-p)(1-\beta_s)$ and $H(\beta_s) = -[(1-\beta_s)\log(1-\beta_s) + \beta_s\log\beta_s]$.

Moreover, define the Lagrange function as $G_s(p) = I(X||Y) + \lambda_s(L(\varpi, X) - L(\varpi, X|Y) - \epsilon)$ where $\epsilon > 0$, $0 \leq p \leq 1/2$ and $\lambda_s \geq 0$. It is not difficult to have the partial derivative of $G_s(p)$ as follows:

$$\frac{\partial G_s(p)}{\partial p} = \frac{\partial I(X||Y)}{\partial p} + \lambda_s \frac{\partial C(p, \varpi, \beta_s)}{\partial p}, \tag{50}$$

where $\frac{\partial C(p,\varpi,\beta_s)}{\partial p}$ is given by the Equation (14) and

$$\frac{\partial I(X||Y)}{\partial p} = (1-2\beta_s)\log\left\{\frac{(2\beta_s - 1)p + 1 - \beta_s}{(1-2\beta_s)p + \beta_s}\right\}. \tag{51}$$

By virtue of the monotonic increasing function $\log(x)$ for $x > 0$, it is easy to see the nonnegativity of $\frac{\partial I(X||Y)}{\partial p}$ is equal to $(1-2\beta_s)\{(2\beta_s - 1)p + 1 - \beta_s - [(1-2\beta_s)p + \beta_s]\} = (1-2p)(1-2\beta_s)^2 \geq 0$ in

the case $0 \leq p \leq 1/2$. Moreover, due to the nonnegative $\frac{\partial C(p,\omega,\beta_s)}{\partial p}$ in $p \in [0, 1/2]$ which is mentioned in the proof of Proposition 1, it is readily seen that $\frac{\partial G_s(p)}{\partial p} \geq 0$ is satisfied under the condition $0 \leq p \leq 1/2$.

Thus, the optimal solution p_s^* is the maximal available p ($p \in [0, 1/2]$) as follows:

$$p_s^* = \begin{cases} \frac{1}{2}, \text{ for } \quad \epsilon \geq C_{\beta_s}, \\ p_s, \text{ for } \quad 0 < \epsilon \leq C_{\beta_s}, \end{cases} \tag{52}$$

where p_s is the solution of $L(\omega, X) - L(\omega, X|Y) = \epsilon$, and C_{β_s} is the MILC mentioned in Equation (11).

By using Taylor series expansion, the equation $L(\omega, X) - L(\omega, X|Y) = \epsilon$ can be expressed approximately as follows:

$$(2\omega + \frac{\omega^2}{2})\left\{(1-p)p - \frac{p(1-p)\beta_s(1-\beta_s)}{[(2\beta_s-1)p+1-\beta_s][(1-2\beta_s)p+\beta_s]}\right\} = \epsilon, \tag{53}$$

whose solution is the approximate p_s as the Equation (47).

Therefore, by substituting the p_s^* into Equation (49), we have Equation (46). □

Remark 4. *Proposition 5 gives the maximum transmission bitrate under the constraint of message importance loss. Particularly, there are growth regions and smooth regions for the maximum transmission bitrate in the receiver with respect to message importance loss ϵ. When the message importance loss ϵ is constrained in a little range, the real bitrate is less than the Shannon information capacity, which is involved with the entropy of the symmetric matrix parameter β_s.*

5.2. Binary Erasure Matrix

Proposition 6. *Assume that there is a Bernoulli(p) source X following distribution $\{p, 1-p\}$ $(0 \leq p \leq 1/2)$ and an information transfer process $\{X, p(y|x), Y\}$ with the binary erasure matrix*

$$p(y|x) = \begin{bmatrix} 1-\beta_e & 0 & \beta_e \\ 0 & 1-\beta_e & \beta_e \end{bmatrix}, \tag{54}$$

where $0 \leq \beta_e \leq 1$. In this case, the solution for $\mathcal{P}_2$ described in Equation (44) is

$$\begin{aligned} &\max_{p(x)} I(X||Y) \\ &= \begin{cases} 1-\beta_e, \ (\epsilon \geq C_{\beta_e}) \\ (1-\beta_e)H(p_e), \ (0 < \epsilon \leq C_{\beta_s}), \end{cases} \end{aligned} \tag{55}$$

where p_e is the solution of $(1-\beta_e)\{pe^{\omega(1-p)} + (1-p)e^{\omega p} - 1\} = \epsilon$, whose approximate value is

$$p_e \doteq \frac{1 - \sqrt{1 - \frac{8\epsilon}{(1-\beta_e)(4\omega+\omega^2)}}}{2}, \tag{56}$$

and $H(x) = -[(1-x)\log(1-x) + x\log x]$, $C_{\beta_e} = (1-\beta_e)(e^{\frac{\omega}{2}} - 1)$ and $\omega < 2 \leq 2/\max\{p(x_i)\}$.

Proof of Proposition 6. In the binary erasure matrix, considering the Bernoulli(p) source X whose distribution is $\{p, 1-p\}$, it is readily seen that

$$\begin{aligned} I(X||Y) &= H(Y) - H(Y|X) \\ &= (1-\beta_e)H(p), \end{aligned} \tag{57}$$

where $H(\cdot)$ denotes the Shannon entropy operator, namely $H(p) = -[(1-p)\log(1-p) + p\log p]$.

Moreover, according to the Definitions 1 and 2, it is easy to see that

$$L(\varpi, X) - L(\varpi, X|Y) = (1 - \beta_e)\{L(\varpi, p) - 1\}, \tag{58}$$

where $L(\varpi, p) = pe^{\varpi(1-p)} + (1-p)e^{\varpi p}$.

Similar to the proof of the Proposition 5 and considering the monotonically increasing $H(p)$ and $L(\varpi, p)$ in $p \in [0, 1/2]$, it is not difficult to see that the optimal solution p_e^* is the maximal available p in the case $0 \leq p \leq \frac{1}{2}$, which is given by

$$p_e^* = \begin{cases} \frac{1}{2}, & \text{for} \quad \epsilon \geq C_{\beta_e}, \\ p_e, & \text{for} \quad 0 < \epsilon \leq C_{\beta_e}, \end{cases} \tag{59}$$

where p_e is the solution of $(1 - \beta_e)\{L(\varpi, p) - 1\} = \epsilon$, and the upper bound C_{β_e} is gained in Equation (16).

By resorting to Taylor series expansion, the approximate equation for $(1 - \beta_e)\{L(\varpi, p) - 1\} = \epsilon$ is given by

$$(1 - \beta_e)(2\varpi + \frac{\varpi^2}{2})(1 - p)p = \epsilon, \tag{60}$$

from which the approximate solution p_e in Equation (56) is obtained.

Therefore, Equation (55) is obtained by substituting the p_e^* into the Equation (57). □

Remark 5. *From Proposition 6, there are two regions for the maximum transmission bitrate with respect to message importance loss. The one depends on the message importance loss threshold ϵ. The other is just related to the erasure matrix parameter β_e.*

Note that single-letter models are discussed to show some theoretical results for information transfer under the constraint of massage importance loss, which may be used in some special potential applications such as maritime international signal or switch signal processing. As a matter of fact, in practice, it is preferred to operate multi-letters models which can be applied to more scenarios such as the multimedia communication, cooperative communications and multiple access, etc. As for these complicated cases which may be different from conventional Shannon information theory, we shall consider it in the near future.

6. Numerical Results

This section shall provide numerical results to validate the theoretical results in this paper.

6.1. The Message Importance Loss Capacity

First of all, we give some numerical simulation with respect to the MILC in different information transmission cases. In Figure 2, it is apparent to see that if the Bernoulli source follows the uniform distribution, namely $p = 0.5$, the message importance loss will reach the maximum in the cases of different matrix parameter β_s. That is, the numerical results of MILC are obtained as $\{0.4081, 0.0997, 0, 0.2265\}$ in the case of parameter $\beta_s = \{0.1, 0.3, 0.5, 0.8\}$ and $\varpi = 1$, which corresponds to Proposition 1. Moreover, we also know that if $\beta_s = 0.5$, namely the random transfer matrix is satisfied, the MILC reaches the lower bound that is $C = 0$. In contrast, if the parameter β_s satisfies $\beta_s = 0$, the upper bound of MILC will be gained such as $\{0.1618, 0.4191, 0.6487, 1.7183\}$ in the case $\varpi = \{0.3, 0.7, 1.0, 2.0\}$.

Figure 3 shows that, in the binary erasure matrix, the MILC is reached under the same condition as that in the binary symmetric matrix, namely $p = 0.5$. For example the numerical results of MILC with $\varpi = 1$ are $\{0.5838, 0.4541, 0.3244, 0.1297\}$ in the cases $\beta_e = \{0.1, 0.3, 0.5, 0.8\}$. However, if $\beta_e = 1$,

the lower bound of MILC ($C = 0$) is obtained in the erasure transfer matrix, different from the symmetric case.

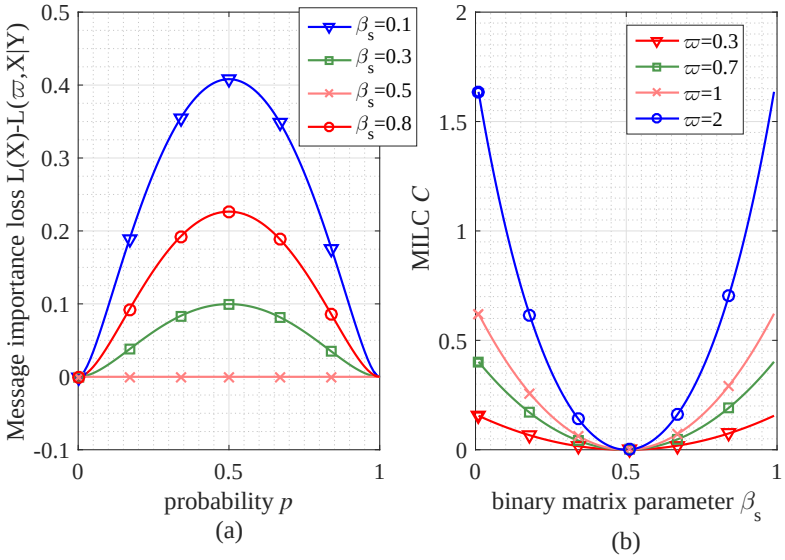

Figure 2. The performance of message importance loss and MILC (mentioned in Definition 4) in the Binary symmetric matrix. (**a**) the performance of message importance loss (with $\varpi = 1$) versus probability p in the cases of different symmetric matrix parameter ($\beta_s = 0.1, 0.3, 0.5, 0.8$); (**b**) the performance of MILC versus matrix parameter β_s in the cases of different parameter ϖ.

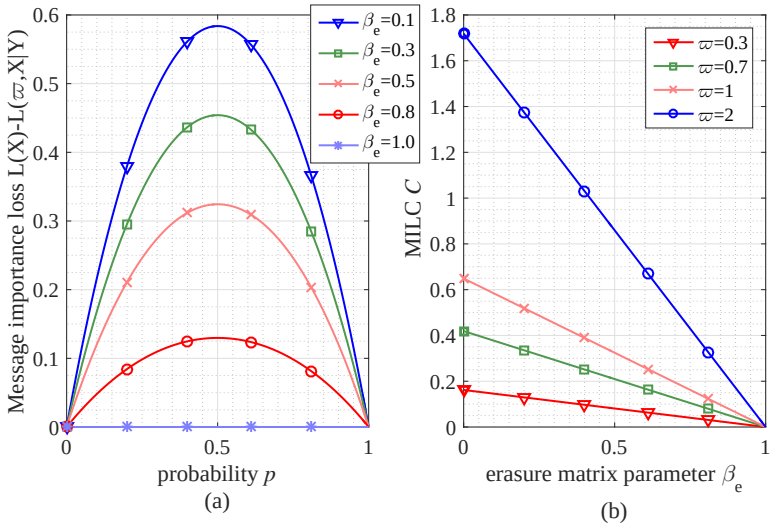

Figure 3. The performance of message importance loss and MILC in the Binary erasure matrix. (**a**) the performance of message importance loss (with $\varpi = 1$) versus probability p in the cases of different matrix parameter ($\beta_e = 0.1, 0.3, 0.5, 0.8$); (**b**) the performance of MILC versus erasure matrix parameter β_e in the cases of different parameter ϖ.

From Figure 4, it is not difficult to see that the certain transfer matrix (namely $\beta_k = 0$) leads to upper bound of MILC. For example, when the number of source symbols satisfies $K = \{4, 6, 8, 10\}$, the numerical results of MILC with $\varpi = 2$ are $\{3.4817, 4.2945, 4.7546, 5.0496\}$. In addition, the lower bound of MILC is reached in the case that $\beta_k = 1 - \frac{1}{K}$.

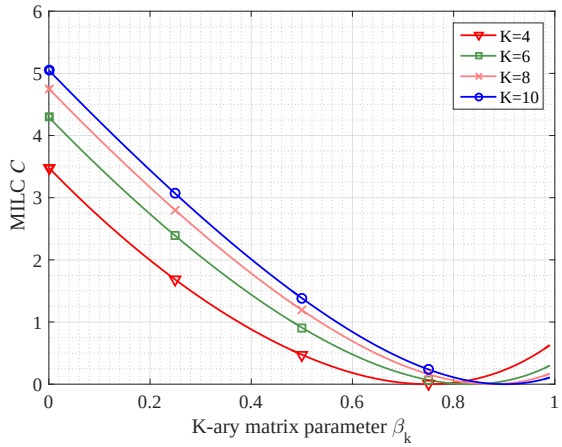

Figure 4. The performance of MILC in strongly symmetric matrix with $K = 4, 6, 8, 10$.

6.2. Message Importance Distortion

We focus on the distortion of message importance transfer and give some simulations in this subsection. From Figure 5, it is illustrated that the message importance distortion function $R_\omega(D)$ is monotonically non-increasing with respect to the distortion D, which can validate some properties mentioned in Section 4.1. Moreover, the maximum $R_\omega(D)$ is obtained in the case $D = 0$. Taking the Bernoulli(p) source as an example, the numerical results of $R_\omega(D)$ with $\omega = 0.2$ are $\{0.0379, 0.0674, 0.0884, 0.1010, 0.1052\}$ and the corresponding probability satisfies $p = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Note that the turning point of $R_\omega(D)$ is gained when the probability p equals to the distortion D, which conforms to Proposition 4.

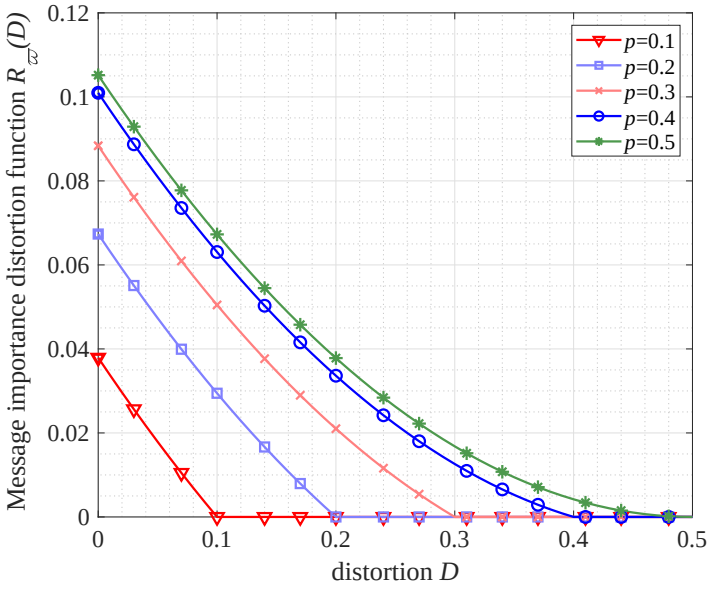

Figure 5. The performance of message importance distortion function $R_\omega(D)$ in the case of Bernoulli(p) source ($p = 0.1, 0.2, 0.3, 0.4$).

6.3. Bitrate Transmission with Message Importance Loss

Figure 6 shows the allowable maximum bitrate (characterized by mutual information) constrained by a message importance loss ϵ in a Bernoulli(p) source case. It is worth noting that there are two regions for the mutual information in the both transfer matrixes. In the first region, the mutual information is monotonically increasing with respect to the ϵ; however, in the second region, the mutual information is stable, namely the information transmission capacity is obtained. As for the numerical

results, the turning points are obtained at $\epsilon = \{0.0328, 0.0185, 0.0082, 0.0021\}$ and the maximum mutual information values are $\{0.5310, 0.2781, 0.1187, 0.0290\}$ in the binary symmetric matrix with the corresponding parameter $\beta_s = \{0.1, 0.2, 0.3, 0.4\}$, while the turning points of erasure matrix are at $\epsilon = \{0.0416, 0.0410, 0.0359, 0.0308\}$ in the case that $\beta_e = \{0.1, 0.2, 0.3, 0.4\}$ with the maximum mutual information values $\{0.9, 0.8, 0.7, 0.6\}$. Consequently, Propositions 5 and 6 are validated from the numerical results.

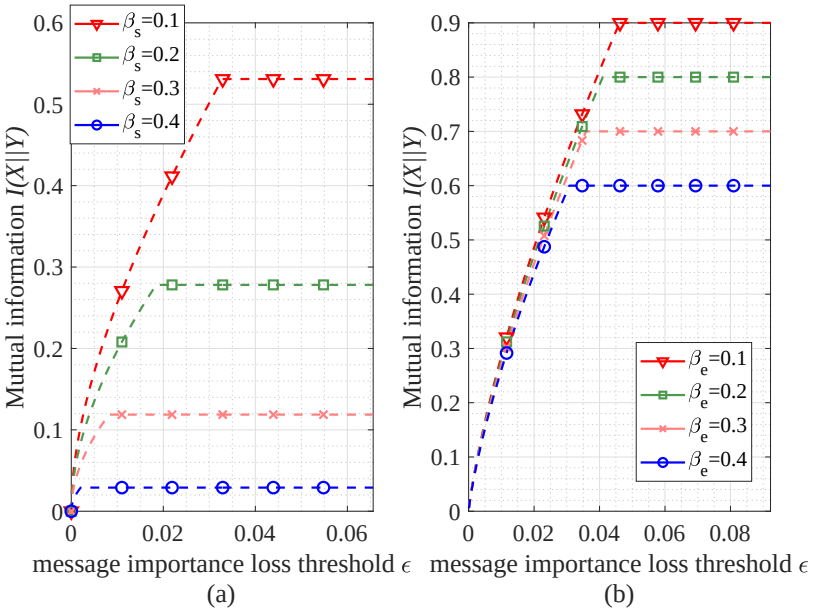

Figure 6. The performance of mutual information $I(X||Y)$ constrained by the message importance loss ϵ (the parameter $\varpi = 0.1$). (**a**) the performance of $I(X||Y)$ versus ϵ in the binary symmetric matrix; (**b**) the performance of $I(X||Y)$ versus ϵ in the erasure matrix.

6.4. Experimental Simulations

In this subsection, we take the binary stochastic process (in which the random variable follows Bernoulli distribution) as an example to validate theoretical results. In particular, the Bernoulli(p) source X (whose distribution is denoted by $P(X) = \{p, 1-p\}$ where $0 < p < 1$) with the symmetric or erasure matrix (described by Equations (10) and (15)) is considered to reveal some properties of message importance loss capacity (in Section 3), message importance distortion function (in Section 4) as well as bitrate transmission constrained by message importance (in Section 5).

From Figure 7, it is seen that the uniform information source X (that is $P(X) = \{1/2, 1/2\}$) leads to the maximum message importance loss (namely MILC) in both cases of symmetric matrix and erasure matrix, which implies Propositions 1 and 2. Moreover, with the increase of number of samples, the performance of massage importance loss tends to smooth. In addition, the MILC in symmetric transfer matrix is larger than that in the erasure one when the matrix parameters β_s and β_e are the same.

As for the distortion of message importance transfer, we investigate the message importance loss based on different transfer matrices, which is shown in Figure 8 where $p_{optimal}(y|x)$ is described as Equation (43), $p_{symmetric}(y|x) = \begin{bmatrix} 1-D & D \\ D & 1-D \end{bmatrix}$, $p_{random\ 1}(y|x) = \begin{bmatrix} 1-\frac{D}{10p} & \frac{D}{10p} \\ \frac{9D}{10(1-p)} & 1-\frac{9D}{10(1-p)} \end{bmatrix}$, $p_{random\ 2}(y|x) = \begin{bmatrix} 1-\frac{D}{5p} & \frac{D}{5p} \\ \frac{D}{5(1-p)} & 1-\frac{D}{5(1-p)} \end{bmatrix}$, $p_{random\ 3}(y|x) = \begin{bmatrix} 1-\frac{D}{10p} & \frac{D}{10p} \\ \frac{D}{10(1-p)} & 1-\frac{D}{10(1-p)} \end{bmatrix}$, $p_{certain}(y|x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, D is the allowable distortion and p is the probability element of Bernoulli(p) source. From

Figure 8, it is illustrated that, when the $p_{optimal}(y|x)$ is selected as the transfer matrix, the massage importance loss reaches the minimum, which corresponds to Proposition 4. In addition, if the transfer matrix is not certain (existing distortion), message importance loss is decreasing with the increase of allowable distortion.

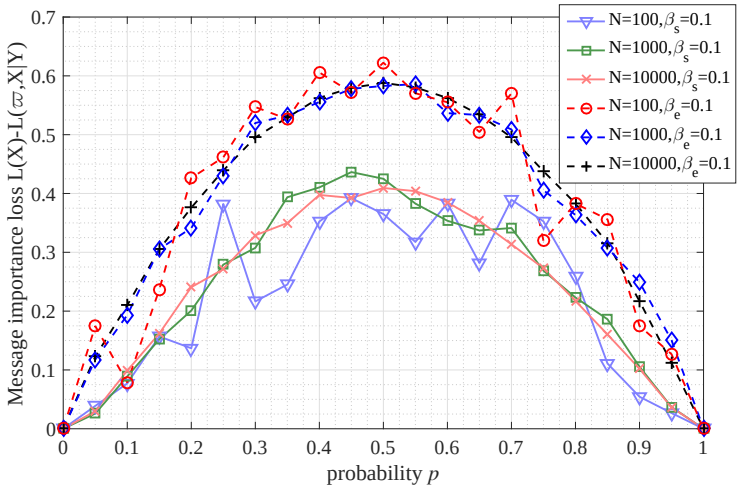

Figure 7. The message importance loss (with parameter $\varpi = 1$) versus the probability p of Bernoulli(p) source with number of samples N ($N = \{100, 1000, 10,000\}$). There are two different transfer matrices, namely the symmetric matrix with parameter $\beta_s = 0.1$ and the erasure matrix with parameter $\beta_e = 0.1$.

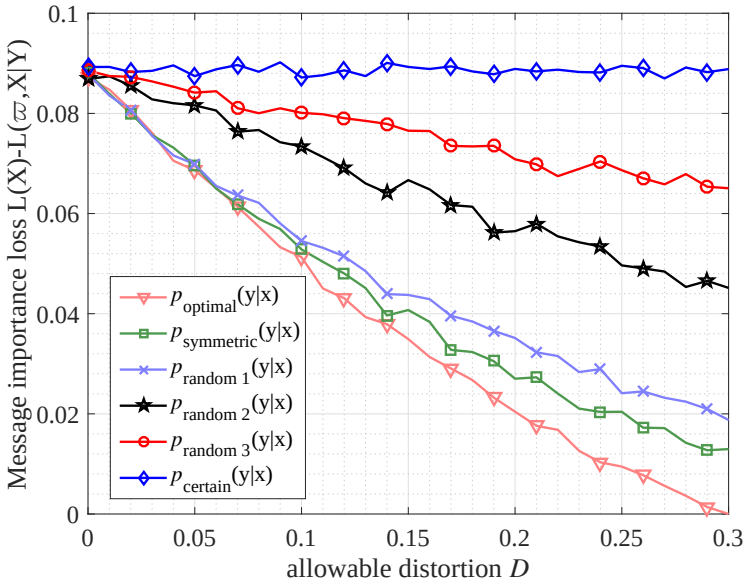

Figure 8. The message importance loss (with parameter $\varpi = 0.1$) versus allowable distortion D (the corresponding distortion function is Hamming distortion) in the case of different transfer matrices. The information source X follows Bernoulli(p) distribution (where $p = 0.3$, namely $P(X) = \{0.3, 0.7\}$) and the number of samples is $n = 10,000$.

Considering the transmission with a message importance loss constraint, Figure 9 shows that, when the p_s^* (given by Equation (52)) and p_e^* (given by Equation (59)) are selected as the probability elements for the Bernoulli(p) source in the symmetric matrix and erasure matrix respectively, the corresponding mutual information values are larger than those based on other probability (such as $p_{random\ 1} = (1 - \sqrt{1 - 8\epsilon})/2$ and $p_{random\ 2} = (1 - \sqrt{1 - 4\epsilon})/2$). In addition, it is not difficult to see that, when the parameter β_s is equal to β_e, the mutual information (constrained by a message importance loss) in symmetric transfer matrix is larger than that in the erasure one.

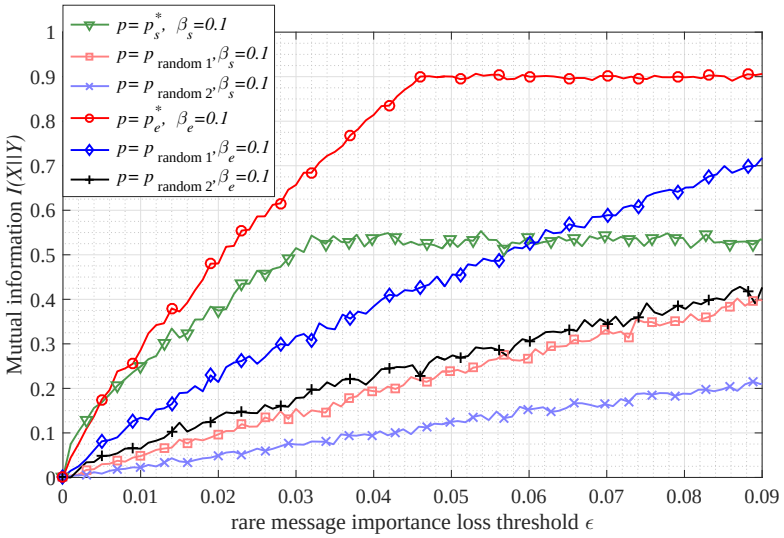

Figure 9. The mutual information $I(X||Y)$ versus the rare message importance loss threshold ϵ (the parameter $\varpi = 0.1$) in the case of Bernoulli(p) source X (that is $P(X) = \{p, 1-p\}$ with different probability p). The number of samples observed from the source X is $n = 10,000$, and transfer matrix is the symmetric matrix with parameter $\beta_s = 0.1$ or the erasure matrix with parameter $\beta_e = 0.1$.

7. Conclusions

In this paper, we investigated the information processing from the perspective of an information measure i.e., MIM. Actually, with the help of parameter ϖ, the MIM has more flexibility and can be used widely. Here, we just focused on the MIM with $0 \leq \varpi \leq 2/\max\{p(x_i)\}$ which not only has properties of self-scoring values for probabilistic events but also has similarities with Shannon entropy in information compression and transmission. In particular, based on a system model with message importance processing, a message importance loss was presented. This measure can characterize the information distinction before and after a message transfer process. Furthermore, we have proposed the message importance loss capacity which can provide an upper bound for the message importance harvest in the information transmission. Moreover, the message importance distortion function, which is to select an information transfer matrix to minimize the message importance loss, was discussed to characterize the performance of information lossy compression from the viewpoint of message importance of events. In addition, we exploited the message importance loss to constrain the bitrate transmission so that the combined factors of message importance and amount of information are considered to guide an information transmission. To give the validation for theoretical analyses, some numerical results and experimental simulations were also presented in details. As the next step research, we are looking forward to exploiting real data to design some applicable strategies for information processing based on the MIM, as well as investigating the performance of multivariate systems in the sense of MIM.

Author Contributions: R.S., S.L. and P.F. all contributed to this work on investigation and writing.

Funding: The authors appreciate the support of the National Natural Science Foundation of China (NSFC) No. 61771283.

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

Abbreviations

The following abbreviations are used in this manuscript:

MIM	Message Importance Measure
MEMS	Max Entropy in Metric Space
IoT	Internet of Things
NMIM	Non-Parametric MIM
DMIM	Differential MIM
CMIM	Conditional Message Importance Measure
MILC	Message Importance Loss Capacity

Appendix A. Proof of the Convexity Property of $R_\omega(D)$

As for an allowable distortion $D_0 = \delta D_a + (1-\delta)D_b$, we have the average distortion for the information transfer matrix $p_0(y|x) = \delta p_a(y|x) + (1-\delta)p_b(y|x)$ as follows:

$$\begin{aligned} \bar{D}_0 &= \delta \sum_{x_i}\sum_{y_j} p(x_i)p_a(y_j|x_i)d(x_i,y_j) + (1-\delta)\sum_{x_i}\sum_{y_j} p(x_i)p_b(y_j|x_i)d(x_i,y_j) \\ &\le \delta D_a + (1-\delta)D_b = D_0, \end{aligned} \tag{A1}$$

which indicates that the $p_0(y|x)$ is an allowable information transfer matrix for D_0.

Moreover, by using Jensen's inequality and Bayes' theorem, we have the CMIM with respect to $p_0(y|x)$ as follows:

$$\begin{aligned} &L_0(\omega, X|Y) \\ &= \sum_{x_i}\sum_{y_i} p(x_i)p_0(y_j|x_i)e^{\omega\{1-\frac{p(x_i)p_0(y_j|x_i)}{p_0(y_j)}\}} \\ &= \sum_{x_i}\sum_{y_i} p(x_i)[\delta p_a(y_j|x_i) + (1-\delta)p_b(y_j|x_i)]e^{\omega\{1-\frac{p(x_i)[\delta p_a(y_j|x_i)+(1-\delta)p_b(y_j|x_i)]}{p_0(y_j)}\}} \\ &\ge \sum_{x_i}\sum_{y_i} p(x_i)[\delta p_a(y_j|x_i)]e^{\omega\{1-\frac{p(x_i)[\delta p_a(y_j|x_i)]}{p_0(y_j)}\}} + \sum_{x_i}\sum_{y_i} p(x_i)[(1-\delta)p_b(y_j|x_i)]e^{\omega\{1-\frac{p(x_i)[(1-\delta)p_b(y_j|x_i)]}{p_0(y_j)}\}} \\ &\ge \delta\sum_{x_i}\sum_{y_i} p(x_i)p_a(y_j|x_i)e^{\omega\{1-\frac{p(x_i)p_a(y_j|x_i)}{p_a(y_j)}\}} + (1-\delta)\sum_{x_i}\sum_{y_i} p(x_i)p_b(y_j|x_i)e^{\omega\{1-\frac{p(x_i)p_b(y_j|x_i)}{p_b(y_j)}\}} \\ &= \delta L_a(\omega, X|Y) + (1-\delta)L_b(\omega, X|Y), \end{aligned} \tag{A2}$$

in which

$$\begin{aligned} p_0(y_j) &= \sum_{x_i} p(x_i)p_0(y_j|x_i) \\ &= \sum_{x_i} p(x_i)[\delta p_a(y_j|x_i) + (1-\delta)p_b(y_j|x_i)] \\ &= \delta p_a(y_j) + (1-\delta)p_b(y_j), \end{aligned} \tag{A3}$$

and the parameter ω is $0 < \omega \le \frac{2\min_j\{p(y_j)\}}{\max_i\{p(x_i)\}}$.

Furthermore, according to the Equations (29) and (A2), it is not difficult to have

$$\begin{aligned} R_\omega(D_0) &= \min_{p(y|x)\in B_{D_0}} \{L(\omega, X) - L(\omega, X|Y)\} \\ &\le \{L(\omega, X) - L_0(\omega, X|Y)\} \\ &\le \delta\{L(\omega, X) - L_a(\omega, X|Y)\} + (1-\delta)\{L(\omega, X) - L_b(\omega, X|Y)\} \\ &= \delta R_\omega(D_a) + R_\omega(D_b), \end{aligned} \tag{A4}$$

where $L(\omega, X)$ is the MIM for the given information source X, while $L_a(\omega, X|Y)$ and $L_b(\omega, X|Y)$ denote the CMIM with respect to $p_a(y|x)$ and $p_b(y|x)$, respectively.

Therefore, the convexity property is testified.

Appendix B. Proof of Proposition 4

Considering the fact that the Bernoulli source X is given and the equivalent expression is mentioned in Equation (39), the optimization problem $\mathcal{P}_1$ can be regarded as

$$\begin{aligned} \mathcal{P}_{1-A}: \max_{p(y_j|x_i)} \; & L(\omega, X|Y) \\ \text{s.t. } & p(x_0)p(y_1|x_0) + p(x_1)p(y_0|x_1) = D, \\ & p(y_0|x_0) + p(y_1|x_0) = 1, \\ & p(y_0|x_1) + p(y_1|x_1) = 1, \\ & p(y_j|x_i) \geq 0, \quad (i = 0, 1; j = 0, 1), \end{aligned} \tag{A5}$$

where $L(\omega, X|Y) = \sum_{x_i, y_j} p(x_i, y_j) e^{\omega(1-p(x_i|y_j))}$ and $0 < \omega \leq \frac{2 \min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

To simplify the above one, we have

$$\begin{aligned} \mathcal{P}_{1-B}: \max_{\alpha, \beta} \; & L_D(\omega, X|Y) \\ \text{s.t. } & p\alpha + (1-p)\beta = D, \\ & 0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, 0 \leq p \leq 1, \end{aligned} \tag{A6}$$

in which p and $(1-p)$ denote $p(x_0)$ and $p(x_1)$, α and β denote $p(y_1|x_0)$ and $p(y_0|x_1)$, and

$$\begin{aligned} & L_D(\omega, X|Y) \\ & = p(1-\alpha) e^{\frac{\omega(1-p)\beta}{p(1-\alpha)+(1-p)\beta}} + (1-p)\beta e^{\frac{\omega p(1-\alpha)}{(1-p)\beta+p(1-\alpha)}} + (1-p)(1-\beta) e^{\frac{\omega p\alpha}{p\alpha+(1-p)(1-\beta)}} + p\alpha e^{\frac{\omega(1-p)(1-\beta)}{p\alpha+(1-p)(1-\beta)}}, \end{aligned} \tag{A7}$$

where $0 < \omega \leq \frac{2 \min_j \{p(y_j)\}}{\max_i \{p(x_i)\}}$.

Actually, it is not easy to deal with the Equation (A6) directly; we intend to use an equivalent expression to describe this objective. By using Taylor series expansion of e^x, namely $e^x = 1 + x + \frac{x^2}{2} + o(x^2)$, we have

$$L_D(\omega, X|Y) \doteq 1 + (2\omega + \frac{\omega^2}{2}) \Big\{ \frac{p\alpha(1-p)(1-\beta)}{p\alpha + (1-p)(1-\beta)} + \frac{p(1-\alpha)(1-p)\beta}{p(1-\alpha) + (1-p)\beta} \Big\}. \tag{A8}$$

By substituting $\beta = \frac{D - p\alpha}{1-p}$ into the Equation (A8), it is easy to have

$$L_D(\omega, X|Y) \doteq 1 + p(2\omega + \frac{\omega^2}{2}) \Big\{ \frac{p\alpha^2 + (1-p-D)\alpha}{2p\alpha + (1-p-D)} + \frac{p\alpha^2 - (p+D)\alpha + D}{(p+D) - 2p\alpha} \Big\}, \tag{A9}$$

where $\max\{0, 1 + \frac{D-1}{p}\} \leq \alpha \leq \min\{1, \frac{D}{p}\}$ resulted from the constraints in Equation (A6).

Moreover, it is not difficult to have the partial derivative of $L_D(\omega, X|Y)$ in Equation (A9) with respect to α as follows:

$$\frac{\partial L_D(\omega, X|Y)}{\partial \alpha} \doteq 2p^2(2\omega + \frac{\omega^2}{2}) \Big\{ \frac{-p\alpha^2 - (1-p-D)\alpha}{[2p\alpha + (1-p-D)]^2} + \frac{p\alpha^2 - (p+D)\alpha + D}{[(p+D) - 2p\alpha]^2} \Big\}. \tag{A10}$$

By setting $\frac{\partial L_D(\omega, X|Y)}{\partial \alpha} = 0$, it is readily seen that the solutions of α in Equation (A10) are given by $\alpha_1 = \frac{(1-p-D)D}{p(1-2D)}$ and $\alpha_2 = \frac{1-D-p}{1-2p}$, respectively.

In addition, in the light of the domain of D mentioned in Equation (35), it is easy to have $D_{\max} = \min\{p, 1-p\}$ in the Bernoulli source case. That is, the allowable distortion satisfies $0 \le D \le \min\{p, 1-p\}$. Thus, the domain of α namely $\max\{0, 1+\frac{D-1}{p}\} \le \alpha \le \min\{1, \frac{D}{p}\}$, can be given by $0 \le \alpha \le \frac{D}{p}$.

Then, it is easy to have the appropriate solution of α as follows:

$$\alpha^* = \frac{(1-p-D)D}{p(1-2D)}, \tag{A11}$$

in which the second derivative $\frac{\partial^2 L_D(\omega, X|Y)}{\partial \alpha^2}$ is non-positive, namely maximum value is reached, and the corresponding information transfer matrix is

$$p(y|x) = \begin{bmatrix} \frac{(1-D)(p-D)}{p(1-2D)} & \frac{(1-p-D)D}{p(1-2D)} \\ \\ \frac{D(p-D)}{(1-p)(1-2D)} & \frac{(1-p-D)(1-D)}{(1-p)(1-2D)} \end{bmatrix}, \tag{A12}$$

where $0 \le D \le \min\{p, 1-p\}$.

Consequently, by substituting the matrix Equation (A12) into the Equation (40), it is not difficult to verify this proposition.

References

1. Ju, B.; Zhang, H.; Liu, Y.; Liu, F.; Lu, S.; Dai, Z. A feature extraction method using improved multi-scale entropy for rolling bearing fault diagnosis. *Entropy* **2018**, *20*, 212. [CrossRef]
2. Wei, H.; Chen, L.; Guo, L. KL divergence-based fuzzy cluster ensemble for image segmentation. *Entropy* **2018**, *20*, 273. [CrossRef]
3. Rehman, S.; Tu, S.; Rehman, O.; Huang, Y.; Magurawalage, C.M.S.; Chang, C.C. Optimization of CNN through novel training strategy for visual classification problems. *Entropy* **2018**, *20*, 290. [CrossRef]
4. Rui, S.; Liu, S.; Fan, P. Recognizing information feature variation: message importance transfer measure and its applications in big data. *Entropy* **2018**, *20*, 401. [CrossRef]
5. Hu, H.; Wen, Y.; Chua, T.S.; Li, X. Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access* **2017**, *5*, 7776–7797.
6. Villecco, F. On the evaluation of errors in the virtual design of mechanical systems. *Machines* **2018**, *6*, 36. [CrossRef]
7. Bormashenko, E.; Frenkel, M.; Legchenkova, I. Is the Voronoi Entropy a True Entropy? Comments on "Entropy, Shannon's Measure of Information and Boltzmann's H-Theorem". *Entropy* **2019**, *21*, 251. [CrossRef]
8. Delvenne, J. Category theory for autonomous and networked dynamical systems. *Entropy* **2019**, *21*, 301. [CrossRef]
9. Ramaswamy, S.; Rastogi, R.; Shim, K. Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Rec.* **2000**, *29*, 427–438. [CrossRef]
10. Harrou, F.; Kadri, F.; Chaabane, S.; Tahon, C.; Sun, Y. Improved principal component analysis for anomaly detection: Application to an emergency department. *Comput. Ind. Eng.* **2015**, *88*, 63–77. [CrossRef]
11. Xu, S.; Baldea, M.; Edgar, T.F.; Wojsznis, W.; Blevins, T.; Nixon, M. An improved methodology for outlier detection in dynamic datasets. *AIChE J.* **2015**, *61*, 419–433. [CrossRef]
12. Yu, H.; Khan, F.; Garaniya, V. Nonlinear Gaussian belief network based fault diagnosis for industrial processes. *J. Process Control* **2015**, *35*, 178–200. [CrossRef]
13. Prieto-Moreno, A.; Llanes-Santiago, O.; Garcia-Moreno, E. Principal components selection for dimensionality reduction using discriminant information applied to fault diagnosis. *J. Process Control* **2015**, *33*, 14–24. [CrossRef]
14. Christidis, K.; Devetsikiotis, M. Blockchains and Smart Contracts for the Internet of Things. *IEEE Access* **2016**, *4*, 2292–2303. [CrossRef]

15. Wu, J.; Zhao, W. Design and realization of winternet: From net of things to internet of things. *ACM Trans. Cyber Phys. Syst.* **2017**, *1*, 2. [CrossRef]
16. Lin, J.; Yu, W.; Zhang, N.; Yang, X.; Zhang, H.; Zhao, W. A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. *IEEE Internet Things J.* **2017**, *4*, 1125–1142. [CrossRef]
17. Sun, Y.; Song, H.; Jara, A.J.; Bie, R. Internet of things and big data analytics for smart and connected communities. *IEEE Access* **2016**, *4*, 766–773. [CrossRef]
18. Zanella, A.; Bui, N.; Zorzi, M. Internet of Things for smart cities. *IEEE Internet Things J.* **2014**, *1*, 22–32. [CrossRef]
19. Jain, R.; Shah, H. An anomaly detection in smart cities modeled as wireless sensor network. In Proceedings of the 2016 International Conference on Signal and Information Processing (IConSIP), Nanded, India, 6–8 October 2016; pp. 1–5.
20. Ramos, S.; Gehrig, S.; Pinggera, P.; Franke, U.; Rother, C. Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Redondo Beach, CA, USA, 11–14 June 2017; pp. 1025–1032.
21. Amaradi, P.; Sriramoju, N.; Dang, L.; Tewolde, G.S.; Kwon, J. Lane following and obstacle detection techniques in autonomous driving vehicles. In Proceedings of the 2016 IEEE International Conference on Electro Information Technology (EIT), Dekalb, IL, USA, 19–21 May 2016; pp. 674–679.
22. Gaikwad, V.; Lokhande, S. An improved lane departure method for advanced driver assistance system. In Proceedings of the 2012 International Conference on Computing, Communication and Applications (ICCCA), Dindigul, India, 22–24 February 2012; pp. 1–5.
23. Fan, P.; Dong, Y.; Lu, J.; Liu, S. Message importance measure and its application to minority subset detection in big data. In Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016; pp. 1–6.
24. Liu, S.; She, R.; Fan, P.; Letaief, K.B. Non-parametric Message Importance Measure: Storage Code Design and Transmission Planning for Big Data. *IEEE Trans. Commun.* **2018**, *66*, 5181–5196. [CrossRef]
25. She, R.; Liu, S.; Dong, Y.; Fan, P. Focusing on a probability element: Parameter selection of message importance measure in big data. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–6.
26. Renyi, A. On measures of entropy and information. In *Proceedings of the 4th Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
27. Wan, S.; Lu, J.; Fan, P.; Letaief, K. Minor probability events' detection in big data: An integrated approach with bayes detection and mim. *IEEE Commun. Lett.* **2019**, *23*, 418–421. [CrossRef]
28. Liu, S.; Dong, Y.; Fan, P.; She, R.; Wan, S. Matching users' preference under target revenue constraints in data recommendation systems. *Entropy* **2019**, *21*, 205. [CrossRef]
29. Liu, S.; She, R.; Fan, P. Differential message importance measure: A new approach to the required sampling number in big data structure characterization. *IEEE Access* **2018**, *6*, 42851–42867. [CrossRef]
30. Jalali, S.; Weissman, T. Block and sliding-block lossy compression via MCMC. *IEEE Trans. Commun.* **2012**, *60*, 2187–2198. [CrossRef]
31. Cui, T.; Chen, L.; Ho, T. Distributed distortion optimization for correlated sources with network coding. *IEEE Trans. Commun.* **2012**, *60*, 1336–1344. [CrossRef]
32. Koken, E.; Tuncel, E. Joint source–Channel coding for broadcasting correlated sources. *IEEE Trans. Commun.* **2017**, *65*, 3012–3022. [CrossRef]
33. Lee, W.; Xiang, D. Information-theoretic measures for anomaly detection. In Proceedings of the 2001 IEEE Symposium on Security and Privacy, Oakland, CA, USA, 13–16 May 2001; pp. 130–143.
34. Ando, S.; Suzuki, E. An information theoretic approach to detection of minority subsets in database. In Proceedings of the IEEE Sixth International Conference on Data Mining, Hong Kong, China, 13–15 December 2006; pp. 11–20.
35. Touchette, H. The large deviation approach to statistical mechanics. *Phys. Rep.* **2009**, *478*, 1–69. [CrossRef]
36. Curiel, R.P.; Bishop, S. A measure of the concentration of rare events. *Sci. Rep.* **2016**, *6*, 1–6.
37. Weinberger, N.; Merhav, N. A large deviations approach to secure lossy compression. *IEEE Trans. Inf. Theory* **2017**, *63*, 2533–2559. [CrossRef]

38. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley Series in Telecommunications and Signal Processing; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2006.
39. Sechelea, A.; Munteanu, A.; Cheng, S.; Deligiannis, N. On the rate-distortion function for binary source coding with side information. *IEEE Trans. Commun.* **2016**, *64*, 5203–5216. [CrossRef]