Elasticity Conditions for Storage Versus Error Exchange in Question–Answering Systems

ALAIN CROLOTTE, MEMBER, IEEE, AND JUDEA PEARL, SENIOR MEMBER, IEEE

Abstract—It has been conjectured that error-allowance could improve dramatically the performance of data processing systems. This hypothesis is tested in the framework of question–answering (QA) systems with storage requirements as a complexity measure. Shannon's rate distortion function $R(D)$ represents the minimum amount of memory a system must employ in order to achieve an average distortion less than D (the distortion can be, for example, the average proportion of erroneous answers produced by the system). The ability of a system to convert an amount D of distortion into memory savings is measured by the ratio $R(D)/R(0)$. A system will be called elastic if this ratio goes to zero as the size of the dataset ensemble goes to infinity. Asymptotic bounds to $R(D)$ are derived giving rise to elasticity conditions invoking the structure of the distortion matrix associated with the system. The bounds established represent a marked improvement over former results by narrowing the gap between the necessary and sufficient conditions for elasticity. Moreover, conditions are established under which the amount of computation required for testing elasticity can be substantially reduced.

I. Introduction

THE PRESENT WORK concerns itself with the following problem. Under what general conditions can the storage requirements of a question–answering (QA) system be significantly reduced by tolerating a small amount of error?

In QA systems a representative summary of an input data is stored in the computer memory and is consulted for answering queries about the data. This stored summary can be viewed as a channel-code connecting the input messages (data) and the output messages (answers to queries) reproduced at the receiving end. However, unlike ordinary transmission problems, the quality of the code is not judged by its ability to reproduce faithfully the input data on a symbol-by-symbol basis, but rather by the quality of the answers it helps generate. A distortion of the input data is judged to be significant only if it causes many queries to be answered incorrectly. In this respect a QA system can be regarded as a communication channel serving many users, with each query representing a user interested in a different aspect of the data.

The rationale and practical motivations for studying the memory versus error trade-offs in QA systems are described in our previous paper [1], which also establishes several general principles underlying the trade-off curves of such an exchange. As in our previous work, we also seek to provide simple tests for determining under what conditions a small error tolerance would result in large savings of memory space, an exchange condition which we termed *elastic*. Unlike our previous work, however, the criteria examined in this paper are based not only on general system parameters (e.g., the input statistics, the total number of queries, and the choice of distortion measure), but also on the logical relationships between the various queries the system ought to answer. Consequently, the criteria established in this paper are tighter and more effective; i.e., many QA systems whose elastic character remained undecided by the previous criteria could now be given a decisive elasticity test.

In Section II we introduce an information-theoretic model for QA systems which views the latter as communication channels and permits the use of Shannon's rate-distortion theory. The problem of elasticity is then stated in terms of a rate-distortion function $R(D)$, namely, under what conditions $\lim_{M\to\infty} R(D)/R(0)=0$ where M is the size of the dataset ensemble.

Section III contains the derivation of bounds which govern the memory–error exchange as measured by the ratio $R(D)/R(0)$. Both upper and lower bounds are shown to have identical mathematical formats, but incorporate different system parameters.

In Section IV the asymptotic behavior of the bounds is examined, and general criteria for elasticity are established. We show that, under quite general conditions, a necessary condition for elasticity could be established without examining the entire distortion matrix but only that portion of the matrix which corresponds to answer-strings reflecting some realizable dataset. As a result, the computation required for testing elasticity could be substantially curtailed.

Section V demonstrates the applicability of the criteria established in IV to the testing of elasticity in three QA systems. We show that systems which admit all binary valued questions on $\{0,1\}^m$, and those admitting singly conjunctive questions on $\{0,1\}^m$, are inelastic. On the other hand, a system which admits size comparison questions on the integers $\{1,\cdots,m\}$ is highly elastic.

Manuscript received April 18, 1978; revised March 12, 1979. This work was supported in part by the National Science Foundation under Grant MCS-18734.

A. Crolotte was with the University of California, Los Angeles, CA. He is now with Perceptronics, Inc., 6271 Variel Ave., Woodland Hills, CA 91367.

J. Pearl is with the University of California, 4731 Boelter Hall, Los Angeles, CA 90024.

0018-9448/79/1100-0653$00.75 ©1979 IEEE

II. Problem Definition

A. Background and Nomenclature

Following Minsky *et al.* [2] and Pearl [4], a QA system can be viewed as a device reflecting the pattern of behavior described in Fig. 1. It is characterized by two ensembles: a dataset ensemble M and a query ensemble Q,

$$M=\{\mu_1,\cdots,\mu_M\}$$
$$Q=\{q_1,\cdots,q_Q\}.$$

During a "filing" phase, a storage procedure B_{file} examines a dataset $\mu\in M$, summarizes it, and then transfers the summary into the memory S. Later, during a "finding" phase, a retrieval procedure B_{find} uses the information in the memory to answer queries from Q.

In order to define an overall performance for the system described above, we first define a degree of inconvenience for the user caused by answering $a(\mu,q)$ to query q about dataset μ, i.e., a real-valued function:

$$\delta: V\times M\times Q\rightarrow\mathbb{R}^+$$

where V is the answer vocabulary, and $\mathbb{R}^+$ stands for the nonnegative real numbers.

We assume that for every pair (μ,q) there exists a correct answer to query q about dataset μ, i.e., an $a^T\in V$ such that

$$\delta(a^T,\mu,q)=0.$$

The set

$$A^T=\{\bar{a}^T|\bar{a}^T=\bar{a}^T(\mu)=(a^T(\mu,q_1),\cdots,a^T(\mu,q_Q)) \quad \text{for some } \mu\in M\}$$

will be referred to as the set of the admissible answers, and the system for which only admissible answer strings are allowed will be referred to as the restricted or admissible system. Note that, under certain circumstances, it might be advantageous to employ a nonadmissible answer string. For example, consider the case where M consists of three equiprobable elements $\mu_1=(100)$, $\mu_2=(010)$, $\mu_3=(001)$, and the questions seek the identification of each data bit, i.e., q_j: "Is the jth bit a 'one'?" The string of correct answers about any dataset would reproduce that dataset so that $\bar{a}^T(\mu_i)$ is identical to μ_i. Using the Hamming distance as distortion criterion we have $d(\mu_i,\mu_j)=2(1-\delta_{ij})$, so that with no additional information about μ, a random selection of any admissible answer string would produce an average distortion (over the elements of M) of $4/3$. However, the choice of the nonadmissible answer string (000) would produce a lower average distortion, since $d[\mu_i,(000)]=1$, $i=1,2,3$. Similarly, if the stored summary of a particular dataset contains only the fact that it has more ones than zeroes, then the average number of errors would be minimized by generating the answer string $11\cdots1$, even though such a string may not represent the correct answers to any dataset in the ensemble M.

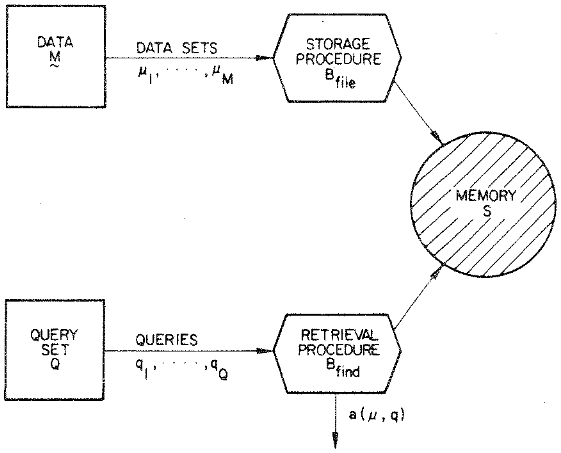


Fig. 1. QA system model.

Calling $P(\mu,q)$ the probability that dataset μ would be presented followed by query q, the overall performance of the system will be defined as

$$D=\sum_{\mu\in M}\sum_{q\in Q}P(\mu,q)\delta[a(\mu,q),\mu,q],$$

the average mean distortion relative to $P(\mu,q)$.

If the datasets and queries are independent, we will have

$$P(\mu,q)=p_\mu\pi_q, \qquad \mu\in M,\ q\in Q,$$

and D can be rewritten as

$$D=\sum_{\mu\in M}p_\mu d[\mu,\bar{a}(\mu)],$$

where

$$d[\mu,\bar{a}(\mu)]=\sum_{q\in Q}\pi_q\delta[a(\mu,q),\mu,q].$$

The developments in this paper assume a distortion measure δ based only on certain relationships between the generated answer and the correct one, but independent on the particular query or the dataset:

$$\delta[a(\mu,q),\mu,q]=\delta[a(\mu,q),a^T(\mu,q)].$$

It is assumed, as in [1], that δ is normalized, i.e., $\delta\leqslant1$, and that $\delta(x,y)$ is a distance on V.

For the purpose of calculating the minimum size of S it is convenient to regard a QA system as a communication channel which receives at its input the dataset μ, and reproduces at its output the answer string $\bar{a}$. Thus the source alphabet is M and the reproducing alphabet is V^Q.

To each dataset i $(i=1,\cdots,M)$ and each question q, a true answer $a^T(q)$ and an actual answer $a(q)$ are associated. We can then index each answer string capable of being generated by the system by an integer j varying from one to $N=V^Q$, denoted by $\bar{a}^j=(a^j(1),\cdots,a^j(Q))$.

With this convention the normalized distortion

$$\rho_{ij} = \sum_q \pi_q \delta\left[a^j(q,\mu_i), a^T(q,\mu_i) \right]$$

defines a distance between dataset i and answer string j. In this paper we will assume that all queries in Q are equally likely, i.e., $\pi_q = 1/Q$. Letting the set of conditional probability assignments which lead to an average distortion less than D be

$$\mathcal{P}_D = \left\{ \mathcal{P}(j|i) : \sum_{i,j} p_i \mathcal{P}(j|i) \rho_{ij} \leqslant D \right\}, \tag{1}$$

Shannon's rate-distortion function is defined by

$$R(D) = \min_{\mathcal{P}(j|i) \in \mathcal{P}_D} I(M,A) \tag{2}$$

where $I(M,A)$ is the mutual information between the source and the user associated with $\mathcal{P}(j|i)$, i.e.,

$$I(M,A) = \sum_{i,j} p_i \mathcal{P}(j|i) \log \frac{\mathcal{P}(j|i)}{\mathcal{P}_j} \tag{3}$$

where

$$\mathcal{P}_j = \sum_i p_i \mathcal{P}(j|i) \tag{4}$$

is the probability mass function of the output.

The definition of $R(D)$ takes its operational significance from the negative part of Shannon's source-coding theorem, stating that no code exists for which both the average distortion is less than D and the rate is less than $R(D)$. This implies, in particular, that any QA system must be provided with an average memory size of at least $R(D)$ nats per dataset in order to achieve a mean distortion at most D.

The positive form of Shannon's theorem, stating that codes exist which achieve a mean distortion D with an amount of memory arbitrarily close to $R(D)$, would be applicable only if simultaneous coding of very large numbers of datasets was allowed. In the model examined above, each dataset has to be coded individually; therefore $R(D)$ provides only a lower bound to the memory size, unless a filing procedure, achieving $R(D)$, can be exhibited. However, if the QA system serves many users connected to a central unit, each using a dataset μ in M, then $R(D)$ is a proper measure of the average storage space (per user) required to serve them with fidelity D.

As shown by Shannon, $R(D)$ is a continuous convex function for $D_{\min} \leqslant D \leqslant D_{\max}$ with

$$D_{\min} = \sum_i p_i \min_j \rho_{ij} \tag{5}$$

$$D_{\max} = \min_j \sum_i p_i \rho_{ij}. \tag{6}$$

For the QA systems considered, there always exists j such that $\rho_{ij} = 0$ for a given i, and therefore,

$$D_{\min} = 0. \tag{7}$$

Convexity implies strict monotonicity for $R(D)$, which is strictly decreasing from $R(0)$ to 0 (obtained for $D = D_{\max}$). Furthermore, $R'(D)$ is continuous, strictly increasing from $-\infty$ over the range $[0, D_{\max}]$.

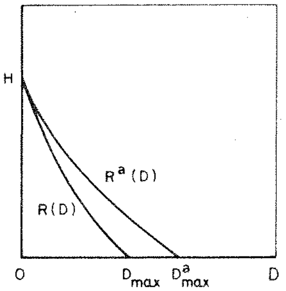


Fig. 2. Typical rate for QA system.

A QA system is said to be *identifiable* if every dataset can be identified knowing its true answer string, i.e., iff $\bar{a}^T(\mu_1) = \bar{a}^T(\mu_2)$ implies $\mu_1 = \mu_2$. $R(0)$ of identifiable systems coincides with H, the source entropy. Throughout this paper we will assume that we are treating identifiable systems. However, all our results will remain valid for nonidentifiable systems provided that $R(0)$ and H have the same asymptotic behavior, i.e., if $\lim_{M \to \infty} R(0)/H > 0$.

A typical situation for a QA system is depicted in Fig. 2, the superscript "a" referring to the admissible system (output vocabulary restricted to the admissible answer strings).

B. Problem Statement

To answer all queries without error requires $H = -\sum_{\mu \in M} p_\mu \log p_\mu$ nats of memory so that $R(D)/H$ is a measure of the impact of error-allowance on the memory requirements of QA systems. If this ratio is very small, error-allowance could be made beneficial.

Definition: A QA system such that $R(D)/H$ tends to zero for every $D > 0$ when $M \to \infty$ will be called *elastic*, while a system such that $R(D_0)/H$ is bounded away from zero for some $D_0 > 0$ will be called *inelastic*.

Although one can certainly conceive of cases neither elastic nor inelastic, i.e., cases where $R(D)/H$ oscillates, only systems exhibiting regularity properties will be treated here. In all practical cases, both $R(D)$ and H are monotonic functions of M, and the corresponding systems are either elastic or inelastic.

For an inelastic system, the fact that $R(D_0)/H$ does not tend to zero dispels any hope of achieving a drastic reduction in memory by allowing errors; for such systems we therefore will not try to find filing schemes achieving the lower bound $R(D)$.

For elastic systems, there is no theoretical impediment to the existence of filing schemes achieving a large reduction in memory by error allowance. However, as pointed out before, one is not guaranteed achieving $R(D)$ by filing schemes where each dataset is coded individually. A separate effort would be required to demonstrate the existence of such a scheme or at least a scheme which achieves a memory saving of the same order as $R(D)/H$.

There is a case of elastic systems which deserves special attention. If $D_{\max}\to 0$ when $M\to\infty$, we can eventually achieve relative error equal to zero without any memory. Such a case will be referred to as *trivially elastic*.

Looking at Fig. 2 a legitimate question to ask is "Is it possible that $D_{\max}\to 0$ while $D^a_{\max}$ is bounded away from zero?" In other words, is it possible that the entire system is trivially elastic while the restricted system is not? The answer is no, as can be seen from the triangular inequality

$$\rho_{ij}+\rho_{i'j}\geqslant\rho_{i'i}, \qquad \text{for all } i,i' \text{ admissible.} \tag{8}$$

Multiplying by p_i and summing on i' yields

$$\rho_{ij}+\sum_{i'}p_{i'}\rho_{i'j}\geqslant\sum_{i'}p_{i'}\rho_{i'i}\geqslant D^a_{\max}. \tag{9}$$

Let j_0 be such that

$$D_{\max}=\min_{j}\sum_{i'}p_{i'}\rho_{i'j}=\sum_{i'}p_{i'}\rho_{i'j_0}. \tag{10}$$

From (9) and (10)

$$\rho_{ij_0}+D_{\max}\geqslant D^a_{\max}, \qquad \text{for all admissible } i. \tag{11}$$

In the column j_0 there exists an element $\rho_{i_0j_0}$ which is lower than the weighted average $D_{\max}$, and consequently

$$2D_{\max}\geqslant\rho_{i_0j_0}+D_{\max}\geqslant D^a_{\max}. \tag{12}$$

Therefore, if $D_{\max}$ tends to zero, so does $D^a_{\max}$.

In this paper our objective will be to investigate what qualities of a QA system render it elastic or inelastic and to apply the results to simple yet typical QA systems.

III. Bounds for $R(D)$

It can easily be shown [3] that if u_s is a function of s such that

$$u_s\geqslant\max_{1<j<N}\sum_i e^{s\rho_{ij}} \tag{13}$$

where $N=V^Q$, then

$$R_L(D)=\max_{s<0}\left(H+sD-\log u_s\right) \tag{14}$$

is a lower bound to $R(D)$. If u_s is differentiable and log-convex, then (14) reduces to the following set of parametric equations:

$$D=\frac{d}{ds}\log u_s, \tag{15a}$$

$$R_L(D)=H+sD-\log u_s, \tag{15b}$$

which define $R_L(D)$ as a lower bound to $R(D)$.

Suppose now that we define u_s and $j^*(s)$ so that equality holds in (13), i.e., write

$$u_s\triangleq\max_j\sum_i e^{s\rho_{ij}}\triangleq\sum_i e^{s\rho_{ij^*(s)}}, \qquad j^*(s)\in\{1,\cdots,N\}. \tag{16}$$

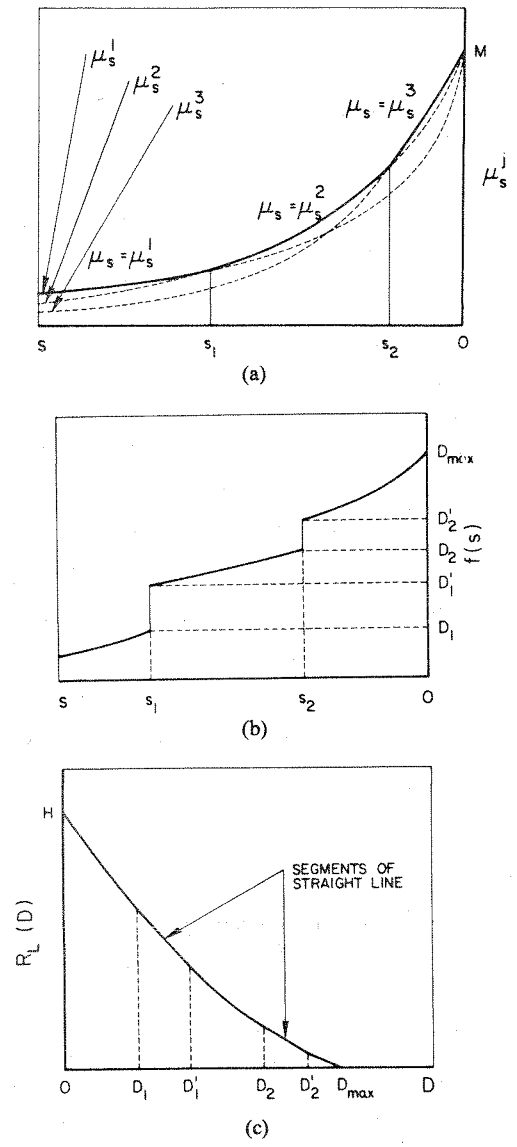


Fig. 3. Typical functions (a) μ_s, (b) $f(s)$, and (c) $R_L(D)$ participating in the solution of (18).

In this case u_s is log-convex but not differentiable in certain points and (15a) has to be modified. More specifically, $f(s)$ defined by

$$\frac{1}{u_s}\sum_i\rho_{ij^*}e^{s\rho_{ij^*}}\triangleq f(s) \tag{17}$$

can replace $(d/ds)\log u_s$ in (15a) (See [5, exercise 2.13, p. 62]), giving rise to the lower bound

$$D=\frac{1}{u_s}\sum_i\rho_{ij^*}e^{s\rho_{ij^*}}, \tag{18a}$$

$$R_L(D)=H+sD-\log u_s. \tag{18b}$$

Whenever a change in the value of j^* occurs, $f(s)$ "jumps" as illustrated in Fig. 3(a) and (b). However, it is still possible to define a surjection via (18a) using the following convention. Let s_k be a point of discontinuity of $f(s)$ and let

$$D_k=f(s_k^-) \tag{19a}$$

$$D_k'=f(s_k^+). \tag{19b}$$

Define next

$$s(D)=\begin{cases} f^{-1}(D), & \text{for } D\notin[D_k,D_k'],\\ s_k, & \text{for } D\in[D_k,D_k'],\end{cases} \tag{20}$$

for all k such that s_k is a discontinuity point of $f(s)$. Then

$$D=f(s)\Leftrightarrow s=s(D), \tag{21}$$

and

$$R_L(D)=H+Ds(D)-\log u_{s(D)}. \tag{22}$$

$R_L(D)$ is convex decreasing from H to 0 and has a continuous derivative equal to $s(D)$ over its entire range of definition $[0,D_{\max}]$ (see Fig. 3(c)).

To find an upper bound to $R(D)$, the following formula due to Haskell [7] will be used:

$$R(D)=\max_{s<0}\left[sD+\min_{\overline{\mathcal{P}}}\left\{-\sum_i p_i\log\left(\sum_j \mathcal{P}_j e^{s\rho_{ij}}\right)\right\}\right] \tag{23}$$

where $\overline{\mathcal{P}}$ varies over the set of all probability N-vectors.

A general technique to find upper bounds to $R(D)$ is to use (23) and upper bound the second term by fixing $\overline{\mathcal{P}}$ as an arbitrary value independent of s:

$$\min_{\overline{\mathcal{P}}}-\sum_i p_i\log\left(\sum_j \mathcal{P}_j e^{s\rho_{ij}}\right)\leqslant-\sum_i p_i\log\left(\sum_j \mathcal{P}_j e^{s\rho_{ij}}\right) \tag{24}$$

for $s\leqslant 0$ and $\overline{\mathcal{P}}$ fixed. In particular, taking

$$\mathcal{P}_j=\begin{cases} 1/M, & j \text{ admissible},\\ 0, & j \text{ nonadmissible},\end{cases} \tag{25}$$

yields (noting that $\rho_{ij}=\rho_{ji}$ for i and j admissible)

$$R(D)\leqslant\max_{s<0}\left[sD+\log M-\sum_{j=1}^{M}p_j\log\sum_{j=1}^{M}e^{s\rho_{ij}}\right]. \tag{26}$$

Let

$$u_s^j=\sum_{i=1}^{M}e^{s\rho_{ij}} \tag{27}$$

and u_s be the geometric average of the u_s^j's, i.e.,

$$\log u_s=\sum_j p_j\log u_s^j \tag{28}$$

where the p_j are the *a priori* probabiltiies of the datasets. Equation (26) can be rewritten

$$R(D)\leqslant R_U(D)=\max_{s<0}\left[\log M+sD-\log u_s\right]. \tag{29}$$

Since each u_s^j is log-convex, so is u_s. Therefore, $R_U(D)$ is given by the following set of parametric equations:

$$D=\frac{d}{ds}\log u_s \tag{30a}$$

$$R_U(D)=\log M+sD-\log u_s \tag{30b}$$

If, instead of (28), we were to define u_s by

$$u_s\leqslant\min_{1<j<N}\sum_{i=1}^{M}e^{s\rho_{ij}}, \tag{31}$$

(29) would still hold for the new function u_s, but not (30). Equation (30) holds only if u_s is log-convex.

Thus the upper and lower bounds to $R(D)$ have similar analytical expressions, which become formally identical when the datasets are equally likely (though they involve a different function u_s).

IV. General Elasticity Conditions

A. Admissible Systems with Balanced Distortion Matrices and Equally Likely Datasets

If the datasets are equally likely and the admissible distortion matrix is balanced, $R^a(D)$ can be computed via (15) with u_s given by

$$u_s=\sum_{i=1}^{M}e^{s\rho_{ij}}=\sum_{i=1}^{M}e^{s\rho_i}\quad(\text{independent of } j). \tag{32}$$

We now analyze the asymptotic behavior of the solution of (15a) by assuming M and Q to be functions of an intrinsic parameter m, and studying the behavior of the system when $m\to\infty$. The function u_s as defined by (32) is clearly log-convex, and equation (15a) defines a unique solution for D fixed. Let

$$f_m^a(s)=\frac{d}{ds}\log u_s=\frac{\sum\rho_i e^{s\rho_i}}{\sum e^{s\rho_i}}. \tag{33}$$

Then

$$D=f_m^a(s)\Leftrightarrow s=s_m^a(D)\qquad\text{for } D\in\left[0,D_{\max}^a\right]. \tag{34}$$

We shall see further that conditions for elasticity can be derived from the asymptotic behavior of the (always negative) function $s_m^a(D)$, and consequently we develop a few lemmas concerning this behavior.

Lemma 1: As $m\to\infty$ we have[1]

$$O(1)\lesssim|s_m^a(D)|\lesssim 0(\log M)$$

for all $0<D<D_{\max}$.

Proof:

$$\int_0^{D_{\max}^a}|s_m^a(D)|\,dD=\int_{-\infty}^{0}f_m^a(s)\,ds$$
$$=\int_{-\infty}^{0}d(\log u_s)=\log M.$$

Since $|s_m^a(D)|$ is a decreasing function of D we have

$$\int_0^{D_{\max}^a}|s_m^a(x)|\,dx\geqslant D|s_m^a(D)|,\qquad D\text{ for all }\left[0,D_{\max}^a\right].$$

Consequently,

$$|s_m^a(D)|\leqslant\frac{1}{D}\log M, \tag{35}$$

which establishes the first part of the lemma. Note that since $s<0$ and $\rho_i<1$, we now have

$$\left[\sum e^{s\rho_i}\right]D_{\max}^a e^s\leqslant MD_{\max}^a e^s\leqslant\left[\sum\rho_i\right]e^s\leqslant\sum\rho_i e^{s\rho_i},$$

[1] Throughout this paper we use the following $O(\cdot)$ notation. We say that $u_m=O(v_m)$ when $m\to\infty$ iff for m large enough $A\leqslant u_m/v_m\leqslant B$ for some A and B such that $A\cdot B>0$.

and consequently $D^a_{\max} e^s \leqslant f^a_m(s)$. Therefore,

$$D^a_{\max} e^{s^a_m(D)} \leqslant f_m[s^a_m(D)] = D$$

which implies $s^a_m(D) \leqslant \log(D/D_{\max})$, i.e.,

$$|s^a_m(D)| \geqslant \log \frac{D_{\max}}{D}. \tag{36}$$

We shall see that the bounds expressed in Lemma 1 are tight, i.e., there exist systems which actually achieve them.

Lemma 2: $|s^a_m(D)| = o(\log M) \Rightarrow R^a(D)/\log M = o(1)$.[2]

Proof: Assume that $s^a_m(D) = o(\log M)$. Using the inequality

$$\sum e^{s\rho_{ij}} \geqslant Me^s,$$

and taking the logarithm of both sides and inserting in (15b) yields

$$R^a(D) \leqslant (D-1)s^a_m(D) = (1-D)|s^a_m(D)|,$$

i.e.,

$$\frac{R^a(D)}{\log M} \lesssim (1-D)\frac{|s^a_m(D)|}{\log M} = o(1). \tag{37}$$

Furthermore, it can be seen from (37) that the faster the ratio $|s^a_m(D)|/\log M$ tends to zero, the more elastic the system.

Lemma 3: $|s^a_m(D)| = O(\log M) \Rightarrow R^a(D/2)/\log M$ bounded away from zero.

Proof: Using the convexity of $R(D)$, one gets

$$R^a(D/2) \geqslant R^a(D) + \left[\frac{D}{2} - D\right]\frac{dR^a(D)}{dD}$$

$$\geqslant -\frac{D}{2}\frac{dR^a(D)}{dD} = -\frac{D}{2}s^a_m(D). \tag{38}$$

Thus for $D < D_{\max}/2$

$$\frac{R^a(D/2)}{\log M} \geqslant \frac{D}{2}\frac{|s^a_m(D)|}{\log M} = O(1) \tag{39}$$

which proves the lemma.

Combining Lemmas 2 and 3 yields:

Theorem 1: In the balanced case with equally likely datasets, a necessary and sufficient condition for the inelasticity of the admissible system is that $s^a_m(D)$, the unique solution to the first fundamental equation $D = f^a_m(s)$, satisfies $s^a_m(D) = O(\log M)$ for some $D > 0$.

B. Admissible Systems with Unbalanced Distortion Matrices and Equally Likely Datasets

If the admissible distortion matrix is not balanced, it is no longer possible to compute explicitly $R^a(D)$; a lower bound to $R^a(D)$ is, however, available, and conditions similar to those found previously can be established. Define

$$u^a_s = \max_{j \in A^T} \sum_i e^{s\rho_{ij}} = \sum_i e^{s\rho_{ij^*(s)}}. \tag{40}$$

Then $R^a_L(D)$ defined by

$$D = \frac{1}{u^a_s} \sum_i \rho_{ij^*} e^{s\rho_{ij^*}} \triangleq f^a_m(s), \tag{41a}$$

$$R^a_L(D) = \log M + sD - \log u^a_s \tag{41b}$$

is a lower bound to $R^a(D)$. Conditions sufficient for the inelasticity of the admissible system can therefore be drawn from (41). Let $s^a_m(D)$ be the solution to (41a) defined as in (20).

Lemma 4: $O(1) \lesssim |s^a_m(D)| \lesssim O(\log M)$.

Proof: The proof proceeds exactly as that of Lemma 1, unaffected by the finite number of discontinuities of $f^a_m(s)$.

Lemma 5: $|s^a_m(D)| = o(\log M) \Rightarrow R^a_L(D)/\log M = o(1)$.

Proof:

$$R^a_L(D) = \log M + Ds^a_m(D) - \log u^a_{s^a_m(D)}. \tag{42}$$

Using

$$u^a_s = \sum e^{s\rho_{ij^*}} \geqslant Me^s$$

yields

$$R^a_L(D) \leqslant (1-D)|s^a_m(D)| \tag{43}$$

as in the proof of Lemma 2.

Lemma 6: $|s^a_m(D)| = O(\log M) \Rightarrow R^a(D/2)/\log M$ is bounded away from zero.

Proof: Since $R^a_L(D)$ is convex and has a derivative equal to $s^a_m(D)$, the proof is the same as for Lemma 3.

Theorem 2: If the datasets are equally likely, a necessary and sufficient condition for the inelasticity of the admissible system is that $s^a_m(D)$ (the unique solution to the first fundamental equation $D = f^a_m(s)$ associated with $u^a_s = \max_{j \in A^T} \Sigma_i e^{s\rho_{ij}}$) satisfies $s^a_m(D) = O(\log M)$ for some $D > 0$.

C. Systems with Equally Likely Datasets: Relations Between Admissible and Nonadmissible Answers

To establish the inelasticity of the admissible system, it is enough to show that $R^a_L(D)/\log M$ is bounded away from zero. In other words, if it is not possible to apply Theorem 2 directly because of the difficulties involved in computing $s^a_m(D)$, one can always try to find a function u_s log-convex and differentiable satisfying

$$u_s \geqslant \max_{j \in A^T} \sum_i e^{s\rho_{ij}}, \tag{44}$$

solve (15a) and show that $s_m(D)$ associated with (15a) satisfies $s_m(D) = O(\log M)$. This, however, insures only the inelasticity of the admissible system but not that of the entire unrestricted system. It is tempting to conjecture that inelasticity of the admissible subsystem implies inelasticity for the entire system. In this section we prove a slightly milder version of this conjecture. Actually, we show that if $R^a_L(D)/\log M$ is bounded away from zero, then the entire system is inelastic. In other words, if the inelasticity of the admissible system is proven on the basis of a lower bound

[2] We say that $u_m = o(v_m)$ when $m \to \infty$ iff $u_m/v_m \to 0$.

of type (15) associated with u_s satisfying (44), then the entire system is inelastic. A few lemmas will be needed.

Lemma 7: $R_L^a(D)/\log M \nrightarrow 0 \Rightarrow \log u^a_{s_m^a(D)}/\log M \nrightarrow 1$

Proof: Assume

$$\frac{\log u^a_{s_m^a(D)}}{\log M} \to 1.$$

Then

$$\log M - \log u^a_{s_m^a(D)} = o(\log M)$$

and $R_L^a(D) = Ds_m^a(D) + o(\log M)$. If $|s_m^a(D)| = O(\log M)$ then $R_L^a(D)$ would be equivalent to $Ds_m^a(D)$ which is negative. Since $R_L^a(D)$ cannot be negative on $[0, D^a_{\max}]$, we must have $s_m^a(D) = o(\log M)$, and consequently $R_L^a(D)/\log M$ tends to zero.

Lemma 8: If for all $0 < D < D_0$ $|s_m^a(D)| \sim \alpha(D)\log M$, then $\lim_{D\to 0} D\alpha(D) = 0$.

Proof: Clearly

$$\begin{aligned}\int_D^{D_{\max}} |s_m^a(x)|\,dx &= Ds_m^a(D) + \int_{s_m^a(D)}^0 f_m^a(s)\,ds\\ &= Ds_m^a(D) + \int_{s_m^a(D)}^0 d(\log u_s^a)\\ &= \log M + Ds_m^a(D) - \log u^a_{s_m^a(D)} = R_L^a(D).\end{aligned}$$

This shows that the Riemann integral $\int_0^{D_{\max}} |s_m^a(x)|\,dx$ exists and is equal to $R_L^a(0) = \log M$. Consequently $\int_0^{D_0} \alpha(x)$ exists. Since $|s_m^a(D)|$ is a decreasing function of D, so is $\alpha(D)$ and

$$\int_0^D \alpha(x)\,dx \geqslant D\alpha(D), \qquad \text{for all } D \leqslant D_0.$$

Therefore

$$\lim_{D\to 0} D\alpha(D) \leqslant \lim_{D\to 0} \int_0^D \alpha(x)\,dx = 0$$

Lemma 9: $u_s^j \leqslant M^{1/2}(u_s^a)^{1/2}$, for all $j \in A$.

Proof: If i_0 is defined by

$$\rho_{i_0 j} = \min_i \rho_{ij},$$

we have

$$u_s^j \leqslant M e^{s\rho_{i_0 j}}. \tag{45}$$

Applying the triangle inequality yields

$$\rho_{ij} + \rho_{i_0 j} \geqslant \rho_{ii_0}, \qquad \text{for } i,j \text{ admissible},$$

which implies, since $s \leqslant 0$,

$$e^{s\rho_{ij}} \cdot e^{s\rho_{i_0 j}} \leqslant e^{s\rho_{ii_0}},$$

and thus

$$e^{s\rho_{i_0 j}} \sum_i e^{s\rho_{ij}} \leqslant \sum_i e^{s\rho_{ii_0}} \leqslant u_s^a$$

or

$$u_s^j \leqslant e^{-s\rho_{i_0 j}} \cdot u_s^a. \tag{46}$$

Equations (45) and (46) together imply

$$(u_s^j)^2 \leqslant M u_s^a, \tag{47}$$

which establishes Lemma 9.

Theorem 3: If $s_m^a(D_0) = O(\log M)$ for some fixed $D_0 > 0$, the entire system is inelastic.

Proof: Let

$$u_s = \max_{j\in A} \sum_i e^{s\rho_{ij}}. \tag{48}$$

Note that

$$R(D) \geqslant R_L(D,s) \triangleq \log M + sD - \log u_s, \qquad \text{for } s < 0.$$

In particular,

$$R(D) \geqslant R_L(D, s_m^a(D)). \tag{49}$$

Since $s_m^a(D) = O(\log M)$ the ratio $R_L^a(D)/\log M$ does not tend to zero, and consequently

$$\frac{\log u^a_{s_m^a(D)}}{\log M} \nrightarrow 1 \tag{50}$$

by virtue of Lemma 7. From Lemma 9 we have

$$\log u^j_{s_m^a(D)} \leqslant \tfrac{1}{2}\log M + \tfrac{1}{2}\log u^a_{s_m^a(D)}.$$

Using (50) we obtain

$$\frac{\log u^j_{s_m^a(D)}}{\log M} \nrightarrow 1, \qquad \text{for } j \in A,$$

and consequently

$$\frac{\log u_{s_m^a(D)}}{\log M} = \frac{\max_j \log u^j_{s_m^a(D)}}{\log M} \nrightarrow 1.$$

More precisely

$$\frac{\log u_{s_m^a(D)}}{\log M}$$

being bounded away from one implies the existence of m_1 and $1 > \eta > 0$ such that

$$\forall m \geqslant m_1 \Rightarrow \frac{\log u_{s_m^a(D)}}{\log M} \leqslant 1 - \eta, \qquad \text{for } D \in \left[0, D^a_{\max}\right].$$

Since $|s_m^a(D)|$ is a decreasing function of D, the assumption $s_m^a(D) = O(\log M)$ implies $s_m^a(D)$ cannot be of order less than $\log M$ for all $D \leqslant D_0$. Moreover, by Lemma 1, $s_m^a(D)$ cannot be of order larger than $\log M$, i.e.,

$$s_m^a(D) \sim -\alpha(D)\log M, \qquad \text{for } \alpha(D) > 0$$

with

$$\lim_{D\to 0} D\alpha(D) = 0$$

by virtue of Lemma 8. Therefore, there exists a $D_1 > 0$ such that

$$D \leqslant D_1 \Rightarrow D\alpha(D) \leqslant \eta/2.$$

Let $D = \min(D_0, D_1)$. Since

$$\left| \frac{s_m^a(D)}{\log M} \Big/ \alpha(D) \right| \to 1,$$

there is an m_1 such that, for $m \geqslant m_1$,

$$\left| \frac{s_m^a(D)}{\log M} \Big/ \alpha(D) \right| \leqslant \tfrac{3}{2},$$

i.e.,

$$\frac{s_m^a(D)}{\log M} \geqslant -\tfrac{3}{2}\alpha(D) \geqslant \eta/2D.$$

Consequently for $m \geqslant m_0 = \max(m_1, m_2)$ and $D = \min(D_0, D_1)$,

$$\begin{aligned}\frac{R(D)}{\log M} &\geqslant \frac{R_L[D, s_m^a(D)]}{\log M}\\ &= 1 + \frac{D s_m^a(D)}{\log M} - \frac{\log u_{s_m^a(D)}}{\log M}\\ &\geqslant 1 - \eta/2 - (1-\eta) = \eta/2,\end{aligned}$$

which establishes the proposition.

Theorem 4: In the case where the datasets are equally likely, if there exists a log-convex and differentiable function u_s such that

$$u_s \geqslant \max_{j \in A^T} \sum_i e^{s\rho_{ij}}$$

and if the corresponding $s_m(D)$ satisfies

$$s_m(D) = O(\log M),$$

the system is inelastic.

Proof: Equation (15) associated with u_s define a lower bound $R_L(D)$ to $R_L^a(D)$. Consequently $R_L^a(D)/\log M$ does not tend to zero, since $R_L(D)/\log M$ does not tend to zero, and thus $s_m^a(D) = O(\log M)$ by Theorem 2. By Theorem 3 the system is therefore inelastic.

A tighter connection between the admissible and nonadmissible systems can be established in case the former gives rise to a balanced distortion matrix. In this case, and with equally likely datasets, we have $R_L^a(D) = R^a(D)$. We can therefore state the next theorem.

Theorem 5: In the case of equally likely datasets, if the admissible answers form a balanced distortion matrix, then the elasticity properties of the entire system are completely determined by those of the admissible system.

The power of Theorem 5 will be demonstrated in Section V. It often happens that while the distortion matrix of the entire system is intractable, its admissible submatrix is balanced, and so the test for elasticity becomes a simple computational task. Even when the admissible submatrix is nonbalanced, it is much easier to find a function u satisfying (44) over the admissible submatrix than the entire matrix. One can then invoke Theorems 3 and 4 for establishing the system's inelasticity.

D. Sufficient Conditions for Elasticity

In Section IV-C we obtained a sufficient condition for inelasticity by lower bounding $R(D)$ using

$$u_s \geqslant \max_{J \in A^T} \sum_i e^{s\rho_{ij}}.$$

We can similarly find a sufficient condition for elasticity upper bounding $R(D)$ using a different u_s, which satisfies

$$\log u_s = \sum_{j=1}^{M} p_j \log u_s^j.$$

Since the resulting upper bound has exactly the same analytical form as $R^a(D)$ in A, the results found above, in particular Theorem 1, apply and lead to the next theorem.

Theorem 6: If $s_m(D)$ associated with u_s, the geometric average of the u_s^j, is such that

$$s_m(D) = o(\log M),$$

the system is elastic.

A simpler and weaker form of this theorem follows.

Theorem 7: If $s_m(D)$, associated with a log-convex and differentiable u_s such that

$$u_s \leqslant \min_{j \in A^T} u_s^j$$

satisfies

$$s_m(D) = o(\log M),$$

the system is elastic.

Proof: Let $R_U(D)$ be the upper bound to $R^a(D)$ associated with u_s via (30). Applying Lemma 2 shows that $R_U(D)/\log M \to 0$, and therefore $R^a(D)/\log M \to 0$. Since $R^a(D)$ is an upper bound to $R(D)$ the elasticity of the entire system is insured.

E. Generalization to Nonuniform Input Distributions

If the datasets are not equally likely, H is no longer equal to $\log M$, and the conditions previously found have to be slightly altered. The inequalities on $R^a(D)$ established in A are still valid, so that

$$D\frac{|s_m^a(2D)|}{H} \lesssim \frac{R^a(D)}{H} \lesssim (1-D)\frac{|s_m^a(D)|}{\log M}.$$

Consequently, Theorem 1 becomes the following.

Theorem 8: In terms of the conditions of Theorem 1, a necessary and sufficient condition for inelasticity of the admissible system is

$$|s_m^a(D)| = O(H).$$

We have not found a straightforward generalization of Theorem 4; however, we can still state the next theorem.

Theorem 9: If there exists a log-convex and differentiable function u_s such that

$$u_s \geqslant \max_{j \in A} \sum_i e^{s\rho_{ij}}$$

and if the corresponding $s_m(D)$ satisfies

$$s_m(D) = O(H),$$

the system is inelastic.

The function u_s might be hard to find, since its legitimacy must be tested for every answer string $j \in A$. The general upper bound, however, involves the admissible answers only, and we can state the following theorem.

Theorem 10: In the general case ($H<\log M$) if $s_m(D)$ associated with the geometric average of the u_s^j or with a lower bound to u_s for j admissible is such that $s_m(D)=o(H)$, the system is inelastic.

The proof follows from the inequality

$$\frac{R(D)}{H}\leqslant(1-D)\frac{|s_m(D)|}{H}.$$

V. Examples of Applications

In this section we shall analyze some simple yet typical QA systems using the techniques developed in Section IV. These conditions all involve functions of the type $\Sigma_i e^{s\rho_{ij}}$ for j admissible. A major deterrent to the analysis of even very simple QA systems is the great computational complexity involved in finding the admissible distortion matrix itself and *a fortiori* the functions $\Sigma_i e^{s\rho_{ij}}$. Checking the asymptotic behavior of $s_m(D)$ defined implicitly by the functions $\Sigma_i e^{s\rho_{ij}}$ can therefore be expected to be a difficult task. These points are illustrated in the examples selected throughout this section.

A. The Set of All Binary Valued Questions on $\{0,1\}^m$

A QA system whose query set consists of all binary valued questions on the data is called a complete binary system (CBS). The time-storage exchange in this system was analyzed by Elias and Flower [8]. If the data requires a code of m bits, then $\mathbf{M}=\{0,1\}^m$, $M=2^m$, $\mathbf{Q}=\{f|f\colon \mathbf{M}\to\{0,1\}$, and $Q=2^M$.

For such a system every two distinct datasets produce identical answers for exactly 50 percent of the questions. Thus, using the Hamming distance as distortion criterion, the admissible distortion matrix is balanced and

$$u_s^a=1+(M-1)e^{s/2},$$

$$s_m^a(D)=-2\log M+2\log\frac{D}{\frac{1}{2}-D}.$$

Consequently from Theorem 3 the system is inelastic.

Note that without invoking Theorem 2 it would be almost hopeless to provide a lower bound to $R(D)$ for the entire system. One would have to examine matrices of size $2^m\times 2^{2^m}$. In fact, in a previous paper [1] we established the CBS inelasticity using the following.

Theorem 11: Any QA system such that $H=\log M$ and the (normalized) distance between any two distinct admissible answer strings is bound away from zero, is inelastic.

We shall see that this is a direct consequence of Theorem 4.

Proof: Assume that

$$\rho_{ij}\geqslant r>0,\qquad \text{for } i,j \text{ admissible, } i\neq j.$$

Then for j admissible

$$\sum_{i=1}^{M} e^{s\rho_{ij}}\leqslant 1+(M-1)e^{sr}\triangleq u_s.$$

Equation (15a) associated with u_s is

$$D=r\frac{(M-1)e^{sr}}{1+(M-1)e^{sr}}$$

implying

$$s_m(D)=\frac{1}{r}\left[\log\frac{D}{r-D}-\log(M-1)\right].$$

Therefore

$$s_m(D)\sim-\frac{1}{r}\log M=O(\log M),$$

and consequently, using Theorem 4, the system is inelastic.

B. The Set of Singly Conjunctive Questions on $\{0,1\}^m$

Consider the following question–answering system. Let $C=\{x_1,\cdots,x_m\}$ be a collection of m objects and C_p a subset of C with p elements. During the filing phase, C_p is presented to a QA system and, later, queries such as "has x_i been observed in C_p?" are presented to the system. An answer is at fault ($\delta=1$) whenever an element of C_p is declared unobserved or an element of $C-C_p$ declared observed. Each C_p can be uniquely represented as a binary string of length m with a one in the ith position if $x_i\in C_p$, and a zero otherwise. For such a system we have $M=\binom{m}{p}$, $Q=m$ and, if p is kept constant, the system is trivially elastic since $D_{\max}\to 0$. When $p=\alpha m$, $0<\alpha<1$, we have $(\log M)/Q\sim H_b(\alpha)>0$, and consequently [1] the system is inelastic. Similarly, if p is unrestricted and any subset of C could be admitted as data, we have $M=2^m$ and $(\log M)/Q\to 1$, rendering the system inelastic.

Since a necessary condition for elasticity is $(\log M)/Q\to 0$ [1], one may wonder if increasing Q by allowing compound questions would cause the system to turn elastic (the distortion criterion still being the normalized Hamming distance, i.e., the probability of producing an erroneous answer). Section V-A demonstrates that the inclusion of all compound binary questions in the query set results in an inelastic system. In the following analysis, we limit the queries to be singly conjunctive, i.e., instead of "has x_i been observed?", each question is now "have both x_i and x_j been observed?". The condition $(\log M)/Q\to 0$ is still satisfied ($Q\sim m^2/2$) but, unlike the complete binary system, the distance between the two nearest admissible answer strings is no longer bounded away from zero. Two datasets which differ by only one element would produce at most $m-1$ conflicting answers and their normalized distance $2(m-1)/m^2$ would approach zero.

It turns out that, for unrestricted p, the admissible distortion matrix is extremely hard to express and the analysis seems intractable. A simpler version of this problem is formed by restricting the input to those subsets with exactly $m/2$ elements ($p=m/2$). This is justified by the fact that, for large m, a significant fraction of the subsets would have m /2 elements, since

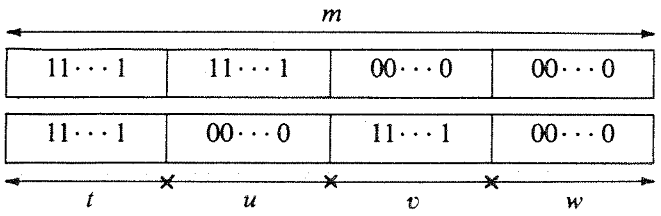
$$\log\binom{m}{m/2}\sim m\log 2=\log 2^m.$$

Note that for this input restricted system, as well as for the unrestricted one, we have

$$Q=\frac{m(m-1)}{2}\sim\frac{m^2}{2}\text{ and }\frac{\log M}{Q}\sim\frac{2\log 2}{m}\rightarrow 0,$$

which complies with the necessary conditions for elasticity. We will show nevertheless that the system remains inelastic. The merit of Theorem 3 will again become apparent, as it will permit us to establish inelasticity on the basis of the $M\times M$ (balanced) distortion matrix of the admissible answers, rather than the intractable $M\times 2^{m(m-1)/2}$ matrix representing all possible answers.

We first wish to compute the distortion between the answer strings generated by two datasets a distance d apart. Consider two binary strings of length m and weight p representing two arbitrary datasets. The different possible configurations are summarized on the following drawing.



Let d be the distance between the two strings. We have the constraints that $t+v=p$, $u+v=d$, $t+u=p$, $t+u+v+w=m$. These imply in particular $2t=d$, i.e., d must be even, let say $d=2i$. We then get $t=p-i$, $u=u=i$, $w=m-p-i$.

In terms of the present QA system a distance $2i$ between two datasets implies a distance

$$tu+tv+\frac{u(u-1)}{2}+\frac{v(v-1)}{2}=i(2p-1-i)$$

between the two corresponding answer strings, and for every given data set there are

$$\binom{p}{u}\binom{p}{v}=\binom{p}{i}\binom{m-p}{i}$$

data sets situated a distance $2i$ from it.

The admissible matrix is balanced and, for $p=m/2$, letting $z=e^{s/Q}$

$$u_s^a=\sum_{i=0}^{m/2}\left[\binom{m/2}{i}\right]^2 z^{i(m-1-i)}.$$

Note that

$$\binom{m/2}{i}\leq\binom{m/2}{m/4}$$

and

$$m-1-i\geq\frac{m}{2}-1\Rightarrow z^{i(m-1-i)}\leq z^{i(m/2-1)},$$

so that

$$u_s^a\leq\binom{m/2}{m/4}\sum_{i=0}^{m/2}\binom{m/2}{i}z^{(m/2-1)i}\triangleq u_s.$$

Therefore, the simpler function

$$u_s=\binom{m/2}{m/4}\left[1+z^{(m/2)-1}\right]^{m/2}$$

satisfies the condition of Theorem 4 and may be used to test the inelasticity of the entire system. Equation (15a) associated with u_s can be written

$$D=\frac{2}{m(m-1)}\cdot\frac{m}{2}\cdot\frac{\left(\frac{m}{2}-1\right)z^{(m/2)-1}}{1+z^{(m/2)-1}}.$$

In order to determine the asymptotic behavior of $s_m(D)$, we first guess a certain functional relationship and then check whether it leads to a fixed D in the equation above. Trying $s_m(D)\sim am$ $(a>0)$ and taking the limit as $m\rightarrow\infty$ yields the equality

$$D=\frac{1}{2}\frac{1}{e^{a/2}-1},$$

which establishes a bijection between a and D. Therefore, the assumption on the asymptotic behavior of $s_m(D)$ is validated and

$$s_m(D)\sim -am=-\frac{a}{\log 2}\log M,$$

i.e., $s_m(D)=O(\log M)$; this implies inelasticity by Theorem 4 (or Theorem 5, since the admissible system is balanced).

C. The Set of Size-Comparison Questions on the Integers $\{1,\cdots,m\}$

Consider a system in which the data consist of one integer between one and m, and the questions consist of presenting an integer between one and m and asking if it is lower than the given integer. The admissible answers for $m=5$ are illustrated in the following table:

	q_1	q_2	q_3	q_4	q_5
$\mu_1=1$	0	0	0	0	0
$\mu_2=2$	1	0	0	0	0
$\mu_3=3$	1	1	0	0	0
$\mu_4=4$	1	1	1	0	0
$\mu_5=5$	1	1	1	1	0

The Hamming distance between the admissible answer strings associated with μ_i and μ_j is easily seen to be $|i-j|$. This is normalized to $|i-j|/m$.

The same distortion matrix is generated by the question: "What is the integer stored to j?" if $V=\{1,\cdots,m\}$ and an absolute-difference distortion criterion is used. However, we prefer to view the system in terms of binary valued question set, as it offers a more solid rationale for the use of the absolute-difference criterion, especially in cases where the size of the integers reflect only ordinal information.

In our case, $M=Q=m$, and the distortion matrix for the admissible answers, though unbalanced, is simple enough for computation. The jth column of the distortion matrix is

$$\begin{bmatrix} j-1 \\ \vdots \\ 1 \\ 0 \\ 1 \\ \vdots \\ m-j \end{bmatrix}$$

and therefore (taking $z=e^{s/m}$),

$$u_s^j=z^{j-1}+\cdots+z+1+z+\cdots+z^{m-j}.$$

In order to find a lower bound to u_s^j independent on j, we note that since

$$z^i \geqslant z^{m-j+i}, \qquad \text{for } i \text{ and } j \leqslant m,$$

$z+\cdots+z^{j-1}$ can be replaced by $z^{m-j+1}+\cdots+z^{m-1}$, and consequently

$$u_s^j > 1+z+\cdots+z^{m-1}=\frac{1-z^m}{1-z} \triangleq u_s, \qquad \text{for } j \text{ admissible.}$$

Using this u_s function leads to $s_m(D) \sim -a(D)$ where $a(D)$ satisfies

$$D=\frac{1}{a}-\frac{1}{e^a-1}.$$

By Theorem 10 the system is elastic for input statistics which give rise to $H \to \infty$. The rate of convergence of $R(D)/R(0)$ is the fastest possible, i.e., $O(1/H)$.

For the case of equally likely datasets, it is also easy to find a filing scheme achieving similar memory requirements. Since each dataset represents an integer between one and m we can use the following quantizing scheme. Partition the range from one to m into k intervals of length c and store the identity of the interval where the data integer occurs. During the answering phase, answers would correspond to an integer situated in the middle of the selected slot. That guarantees a maximum error of $c/2$ and an average error $c/4$ so that the asymptotic performance of this scheme is

$$R(D)=\log 1/4D.$$

VI. Summary

Several criteria were established for determining whether a given question–answering system is elastic, i.e., whether a small tolerance for errors could be exploited to yield a sharp reduction in storage requirements. The criteria established depend on the logical interaction between the admitted set of questions and are analytically tractable.

We examined the asymptotic behavior of lower and upper bounds to the rate-distortion function which are both defined parametrically and involve a function u_s of the sums of elements in the columns of the matrix $(e^{s\rho_{ij}})_{M\times N}$. To get a lower bound we take

$$u_s \geqslant \max_j \sum_i e^{s\rho_{ij}},$$

and to get an upper bound we take

$$\log u_s=\sum_{j=1}^{M} p_j \log \sum_i e^{s\rho_{ij}}.$$

Here $\{p_j\}$ is the probability density function of the datasets.

The first equation defining the bounds then reads

$$D=\frac{d}{ds}\log u_s,$$

and the second reads

$$R_B(D)=X+sD-\log u_s$$

with $X=H$ (the source entropy) for the lower bound, and $X=\log M$ for the upper bound. If the datasets are equally likely, then both upper and lower bounds have the same formal expression (although u_s is not the same).

It is then shown that the asymptotic behavior of these bounds depends solely upon the first parametric equation. More specifically, a necessary and sufficient condition for elasticity of each bound is $s_m(D)=o(\log M)$ where $s_m(D)$ is the solution to the first equation, and M the size of the dataset ensemble. Necessary conditions (i.e., elasticity of the lower bound) and sufficient conditions (i.e., elasticity of the upper bound) for the system's elasticity are then established. It is further shown that, when the datasets are equally likely, elasticity conditions can be determined solely on the basis of the distortion matrix between the admissible answers. Thus the computational work required for elasticity tests is reduced significantly.

These conditions are then applied to three simple QA systems:

1) a system which admits the set of all binary questions regarding an arbitrary binary data (complete binary system).
2) a system which admits singly conjunctive questions (e.g., "Are both x_i and x_j in the dataset?") and the data contains $m/2$ items.
3) a system answering size-comparison questions on the integers $\{1,\cdots,m\}$ (e.g., "Is integer j smaller than the one stored?").

It is shown that examples 1) and 2) are inelastic, while 3) is elastic.

The analysis of the second example demonstrates that redundancy $((\log M)/Q \to 0)$ and denseness $(\min_j \rho_{ij} \to 0$, for $i,j \in A^T)$, though necessary, are not sufficient to guarantee elasticity.

Acknowledgment

The authors greatly appreciate the insightful comments of one reviewer who pointed out an error in one of the earlier proofs. This work was performed while J. Pearl was visiting the Department of Applied Mathematics at the Weizmann Institute of Science, Rehovot, Israel.

References

[1] A. Crolotte and J. Pearl, "Bounds on memory versus error tradeoffs in question-answering systems," *UCLA-ENG-7787*, Dec. 1977, also in *IEEE Trans. Inform. Theory*, vol. IT-25, no. 2, pp. 193–202, Mar. 1979.

[2] M. Minsky and S. Papert, *Perceptrons*. Cambridge, MA: M.I.T. Press, 1969, Chap. 12.

[3] J. Pearl, "Error-bounds for inferential question-answering systems with limited memory," *UCLA-ENG-7481*, Oct. 1974.

[4] ——, "On the storage economy of inferential question-answering systems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, no. 6, pp. 595–602, Nov. 1975.

[5] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

[6] A. Crolotte, "Memory versus error tradeoffs in question-answering systems," Ph.D. dissertation, *UCLA-ENG-7753*, July 1977.

[7] B. Haskell, "The computation and bounding of rate distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 525-531, 1969.

[8] P. Elias and R. A. Flower, "The complexity of some simple retrieval problems," *J. Ass. Comput. Mach.*, vol. 22, no. 1, pp. 367–379, July 1975.

[9] J. Pearl, "An application of rate-distortion theory to pattern recognition and classification," *Pattern Recognition*, vol. 8, pp. 11–22, Jan. 1976.