Storage Space Versus Validity of Answers in Probabilistic Question-Answering Systems

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

Abstract—The trade-offs between the required storage space and the validity of answers produced by probabilistic question-answering (PQA) systems are studies. If the correct answer to a given query is "yes" and the system, instead, issues the estimate that the query has a probability π of being true, then the economic loss to the user can be measured by a distortion function that decreases monotonically with π. A system will be called elastic if a drastic memory saving may be achieved by tolerating a small level of average distortion. The main result reported is that for a large class of distortion measures (i.e., measures exceeding $(1-\pi)^\alpha$ with $\alpha>0$), if a binary QA system (true-false answers) is inelastic then the corresponding probabilistic QA system must also be inelastic. This result implies, for example, that a PQA system that is designed to answer all binary questions on an arbitrary dataset is inelastic. Similarly a PQA system admitting singly conjunctive questions such as "Are both x and y in the dataset?" is also inelastic.

I. Introduction

A BINARY question-answering (QA) system, after consulting a summary of the input data stored in its memory, produces true-false answers to a given set of binary queries. Probabilistic question-answering (PQA) systems permit the answers to be cast in probabilistic phrasing (e.g., "Yes, but I am only 80 percent sure"). The use of PQA systems is justified when the size of the data exceeds the system's storage capacity, in which case storage space is economized at the expense of the answers' quality.

Whereas the user of a binary QA system with insufficient storage must be resigned to the fact that a certain fraction of the answers will be utterly erroneous, the user of a PQA system obtains estimates of each answer's reliability. An answer such as, "Yes, Mr. Smith is still employed by the X Corporation" is far more damaging than, "Yes, Mr. Smith is probably still with X, but I am only 85 percent sure," when in fact he is no longer with X. In the first case the user may make critical decisions to his detriment. In the second case the user is alerted to proceed with caution and is able to judge, given the probability estimates, what level of precautionary effort is warranted in any specific situation.

If the correct answer to a given query is "yes" $(i=1)$ and the system, instead, issues the estimate that the query has a probability π of being true $(0\leqslant\pi\leqslant 1)$, then the economic loss to the user is assumed to be expressed by a distortion function $\delta(1,\pi)=\delta(0,1-\pi)$. Similarly, we assume that the overall degradation in the validity of the system's answers is measured by a distortion criterion given by the mean value of δ, averaged over all queries and all possible datasets.

We study the trade-offs between the required storage space and the quality of answers produced by PQA systems with the primary objective of devising simple tests for determining whether tolerating a low level of the answers' indefiniteness may result in drastic savings of memory space.

We treat a PQA system as a communication channel whereby the content of the system's memory is viewed as a channel code connecting the input messages (datasets) and the output messages (answers to queries) reproduced at the receiving end. This permits us to lower bound the memory requirement by Shannon's rate distortion function $R(D)$ if an average distortion less than D has to be insured.

The ability of a PQA system to convert an amount D of validity degradation into a large memory saving is measured by the ratio $R(D)/R(0)$. If this ratio goes to zero as M, the size of the input ensemble, tends to infinity, then the system will be called *elastic*, since a drastic saving in memory might be feasible by tolerating a small level of indefiniteness. On the other hand, if $R(D)/R(0)$ does not tend to zero, the system will be called *inelastic*. No coding scheme exists which drastically reduces memory requirements with only small degradation in answers validity for the latter system.

Criteria for testing the elasticity of binary QA systems were established in [1] and [2] using the proportion of erroneous answers as a distortion measure. A previous analysis of storage-fidelity exchange in probabilistic QA systems demonstrated a very weak (inelastic) exchange for systems admitting only logically independent queries under both logarithmic and quadratic distortion functions [3]. This paper extends the analysis to PQA systems involving arbitrary (e.g., logically dependent) query sets by studying the asymptotic behavior of a lower bound $R_L(D)$ to $R(D)$.

Manuscript received July 24, 1979; revised January 17, 1980. This work was supported in part by the National Science Foundation under Grant MCS 78-18924. A preliminary version of this paper was presented at the 1979 IEEE International Conference on Information Sciences and Systems, Patras, Greece, July 9–13, 1979.

J. Pearl is with the School of Engineering and Applied Science, University of California, 4731 Boelter Hall, Los Angeles, CA 90024.

A. Crolotte was with the School of Engineering and Applied Science, University of California, Los Angeles, CA. He is now with Perceptronics, Inc., 6271 Variel Avenue, Woodland Hills, CA 91367.

0018-9448/80/1100-0633$00.75

The main result in this paper is the following: for distortion measures of the type $\delta(1,\pi) \geq (1-\pi)^\alpha$, where $\alpha > 0$, if the inelasticity of a binary QA system (true–false answers) can be established on the basis of such a lower bound, then the corresponding probabilistic QA system must also be inelastic. The class of distortion measures to which this theorem applies encompasses all reproducing scoring rules [4] used in practice, such as truncated logarithmic, quadratic, and spherical scoring rules.

The theorem stated above implies, for example, that a PQA system that admits all binary questions on an arbitrary dataset is inelastic. Similarly a PQA system which admits singly conjunctive questions (e.g., "Are both x_i and x_j in the dataset?") is also inelastic. The results in [3] are special cases of this theorem since the inelasticity of a binary QA system with logically independent queries and probability of error distortion criterion can be established by $R_L(D)$, using the fact that the query set is nonredundant [1].

Section II introduces models and nomenclature used in the analysis of QA systems and defines the lower bound $R_L(D)$ to $R(D)$ which will be used in determining the elasticity of PQA systems. Section III summarizes previous results concerning the asymptotic behavior of $R_L(D)$ for binary QA systems. Section IV shows a connection between the results reported in Section III and the elasticity of PQA systems under distortion measures of the form $\delta(1,\pi)=(1-\pi)^\alpha$, where $\alpha>0$. Section V explores some implications of the above results.

II. Models and Nomenclature

Following Minsky and Papert [5] and Pearl [6], a QA system can be described by the model of Fig. 1. It is characterized by 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:

$$\Delta: V \times M \times Q \to \mathbb{R}^+,$$

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

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

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

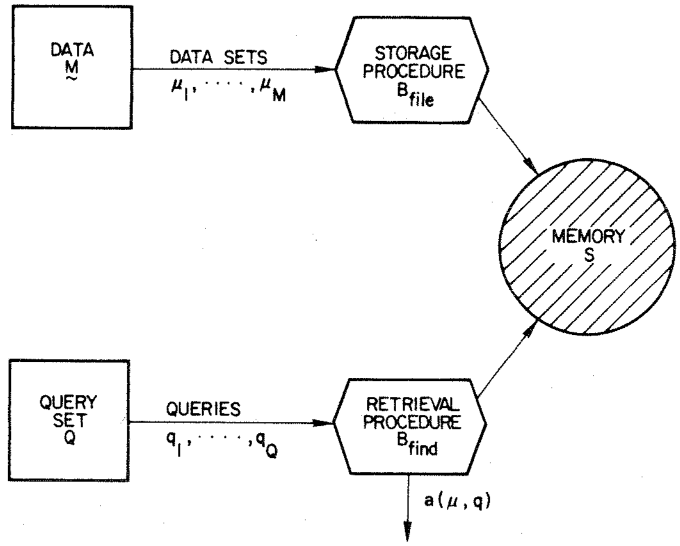

Fig. 1. QA system model.

contains strings of answers which are consistent with some realizable datasets, referred to as the set of the *admissible answers*. Similarly, a system for which only admissible answer strings are allowed will be referred to as a *restricted* or *admissible system*. For systems permitted to issue only binary answers, we have previously established relationships between the restricted and unrestricted models and have shown that under certain general circumstances the asymptotic behavior of the unrestricted system matches that of the restricted system [2]. We presently try to relate the behavior of a probabilistic system to the behavior of the corresponding binary system which is constrained to produce only admissible answers. The latter is much easier to analyze.

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

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

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

If the datasets and queries are independent, we will have

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

where p_μ and P_q are the marginal probabilities corresponding to the joint probability $P(\mu,q)$, and D can then be rewritten as

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

where

$$d[\mu, a(\mu)]=\sum_{q \in Q} P_q \Delta[a(\mu,q),\mu,q].$$

The developments in this paper assume distortion measures Δ based only on certain relationships between the generated answer and the correct one but independent of the particular query or the dataset:

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

If the true answer to question q about dataset μ is "yes" ($a^T(\mu, q)=1$), then $\delta(1, \pi)$ represents the damage caused to the user by assigning a credibility π to a proposition which is in fact true. By symmetry we should have

$$\delta(0, \pi)=\delta(1, 1-\pi). \quad (1)$$

Standard scoring rules have been developed [4] for the purpose of evaluating the quality of probability assessments. An important property of these scoring rules, called "reproducibility", is their tendency to keep the forecaster "honest". For present purposes, however, it is not necessary to limit ourselves to reproducing scoring rules since our scope is to evaluate from a general standpoint what reduction in memory requirements would be obtained by allowing probabilistic statements instead of true–false answers. Here we will consider two types of distortion measures:

- $\delta(1, \pi)$ concave,
- $\delta(1, \pi)$ convex of the form $(1-\pi)^\alpha$, $\alpha \geqslant 1$.

For the purpose of calculating the minimum storage requirement, it is convenient to regard a QA system as a communication channel which receives at its input the dataset μ and reproduces at this output the answer-string Π. Thus the source alphabet is $\boldsymbol{M}$ and the reproducing alphabet is $\boldsymbol{\Pi}=[0,1]^Q$. Note that the set of the extreme points of $\boldsymbol{\Pi}$ is merely $\boldsymbol{A}$, the set of all possible binary answer strings associated with the query ensemble $\boldsymbol{Q}$.

To remain consistent with the literature on rate distortion theory we index the datasets in the input ensemble by an integer i, varying from one to M, replacing the generic variable μ. To each dataset i $(i=1, \cdots, M)$ and each question q, a true answer $a^T(i, q)$ and an actual answer $\pi(i, q)$ are associated so that

$$\rho_{i\Pi}=\sum_q P_q \delta\left[a^T(i, q), \pi(i, q)\right]$$

defines a distortion between dataset i and the answer string

$$\Pi=\left[\pi(i, q_1), \pi(i, q_2), \cdots, \pi(i, q_Q)\right].$$

We assume that all queries in Q are equally likely, i.e., $P_q=1/Q$. Letting the set of conditional probability assignments which lead to an average distortion less than D be

$$\mathcal{P}_D=\left\{P(\Pi|i) \,\middle|\, \sum_{i, \Pi} p_i P(\Pi|i) \rho_{i\Pi} \leqslant D\right\}, \quad (2)$$

Shannon's rate distortion function is defined by

$$R(D)=\min_{P(\Pi|i) \in \mathcal{P}_D} I(M, \Pi), \quad (3)$$

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

$$I(\boldsymbol{M}, \boldsymbol{\Pi})=\sum_{i, \Pi} p_i P(\Pi|i) \log \frac{P(\Pi|i)}{P_\Pi} \quad (4)$$

and

$$P_\Pi=\sum_i p_i P(\Pi|i) \quad (5)$$

is the probability density 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 $\boldsymbol{M}$, 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

$$D_{\min}=\sum_i p_i \min_\Pi \rho_{i\Pi}, \quad (6)$$

$$D_{\max}=\min_\Pi \sum_i p_i \rho_{i\Pi}. \quad (7)$$

For the QA systems considered, there always exists an answer string Π such that $\rho_{i\Pi}=0$ for a given i, and therefore

$$D_{\min}=0. \quad (8)$$

Convexity implies strict monotonicity of $R(D)$, 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 $[0, D_{\max}]$.

A QA system is said to be *identifiable* if every dataset can be identified knowing its true answer string, i.e., if and only if $\boldsymbol{a}^T(\mu_1)=\boldsymbol{a}^T(\mu_2)$ implies $\mu_1=\mu_2$. For identifiable systems $R(0)$ coincides with H, the source entropy. Throughout this paper we assume we are treating identifiable systems. Our results will remain valid for nonidentifiable systems, however, 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).

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

Definition: A QA system such the $R(D)/H$ 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 (bafo) for some $D_0>0$ will be called *inelastic*.

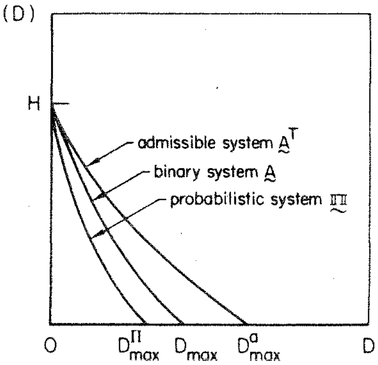

Fig. 2. Typical rate for PQA system.

Although one can certainly conceive of cases which are neither elastic nor inelastic, for instance 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 imprecision; for such systems, one should not attempt, therefore, to find filing schemes achieving more than a fractional saving in storage space.

For elastic systems, there is no theoretical impediment to the existence of filing schemes achieving a large reduction in memory by error allowance. As noted, however, one is not guaranteed to achieve $R(D)$ 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 $R(D)/H$.

The results of the present paper are based on the asymptotic properties of the following lower bound [7] to $R(D)$. Let u_s be any function of s satisfying

$$u_s \geqslant \max_{\Pi \in \Pi} \sum_{i=1}^{M} e^{s\rho_i \Pi}. \tag{9}$$

Then, for all $s<0$, $R(D)$ satisfies

$$R(D) \geqslant H + sD - \log u_s \triangleq R_L(s, D). \tag{10}$$

Optimizing this family of lower bounds with respect to s yields

$$R(D) \geqslant R_L(D) = \max_{s<0} R_L(s, D). \tag{11}$$

If u_s is differentiable and log-convex, $R_L(D)$ is given by the following parametric equations:

$$D = \frac{d}{ds} \log u_s, \tag{12a}$$

$$R_L(D) = H + sD - \log u_s. \tag{12b}$$

These two equations define the lower bound to be used in elasticity tests and will be referred to as the set of *fundamental equations* associated with u_s.

III. Summary of Previous Results

In [2] we studied the relationship between the admissible system and the unrestricted binary system where both admissible and nonadmissible binary answers are permitted. In particular we established the following result [2 th. 4].

Theorem 1: In the case where the datasets are equally likely, if there exists a log-convex and differentiable function u_s such that

$$u_s \geqslant \max_{j \in A^T} \sum_i e^{s\rho_{ij}} \tag{13}$$

and $s_m(D)$, the solution to the first fundamental equation associated with u_s, satisfies[1]

$$s_m(D) = O(\log M), \tag{14}$$

then the system is inelastic.

The power of this result stems from the fact that it provides an inelasticity test which involves the admissible answers only. We wish to derive similar results here since it often happens that while the distortion matrix of the probabilistic system is intractable, the admissible matrix is easier to handle and the test for inelasticity becomes a simple computational task. For convenience we make the following definition.

Definition: A system such that its admissible distortion matrix fulfills the conditions of Theorem 1 will be called *tractably inelastic*.

The reason for coining the name "tractable" for such systems is that there appears to be no manageable method for testing inelasticity in cases where these conditions do not hold.

The analysis of the next section requires additional results, proved in [2], which will be restated here without proof.

Lemma 1: The asymptotic behavior of $s^a_m(D)$, the unique solution of (12a) associated with $u^a_s = \max_{j \in A^T} \sum_i e^{s\rho_{ij}}$, satisfies

$$O(1) \leqslant |s^a_m(D)| \leqslant O(\log M). \tag{15}$$

Theorem 2: If the datasets are equally likely, a necessary and sufficient condition for the inelasticity of the admissible system is that $s^a_m(D)$ satisfies $s^a_m(D) = O(\log M)$ for some $D>0$.

Lemma 2:

$$\frac{R^a_L(D)}{\log M} \not\to 0 \Rightarrow \frac{\log u^a_{s^a_m}(D)}{\log M} \not\to 1. \tag{16}$$

Lemma 3: If,[2] for all $0<D<D_0$, $|s^a_m(D)| \sim \phi(D) \log M$, then $\lim_{D \to 0} D\phi(D) = 0$.

[1] We say that $u_m = O(V_m)$ if and only if there exist A and B $(AB>0)$ such that for m large enough $A \leqslant u_m/v_m \leqslant B$, m is an intrinsic parameter tending to $+\infty$ with M and Q.

[2] We use the notation $u_m \sim v_m$ to denote $\lim_{m \to \infty} u_m/v_m = 1$.

IV. Relation Between Binary and Probabilistic QA Systems

In this section we consider the case where the function $\pi \to \delta(1,\pi)$ is convex of the form $\delta(1,\pi)=(1-\pi)^{\alpha}$ for $\alpha \geqslant 1$. Note that there are two subcases of great interest: $\alpha=1$ which yields the Hamming distance and $\alpha=2$ which is known as the quadratic scoring rule. The quadratic scoring rule is the only reproducing scoring rule of the family $\delta(1,\pi)=(1-\pi)^{\alpha}$ [4].

We first establish two lemmas which will be used in deriving the main theorem of this section.

Lemma 4: For all $a, b>0$ and $\alpha \geqslant 1$

$$(a^{1/\alpha}+b^{1/\alpha})^{\alpha} \leqslant \min(a,b)+(2^{\alpha}-1)\max(a,b).$$

Proof: Assume $a \leqslant b$. Then

$$\min(a,b)+(2^{\alpha}-1)\max(a,b) = a+(2^{\alpha}-1)b=b\left[\frac{a}{b}+2^{\alpha}-1\right].$$

Furthermore by writing

$$(a^{1/\alpha}+b^{1/\alpha})^{\alpha}=b\left[1+\left(\frac{a}{b}\right)^{1/\alpha}\right]^{\alpha},$$

the inequality stated in the lemma is equivalent to

$$(1+x^{1/\alpha})^{\alpha} \leqslant x+2^{\alpha}-1, \qquad 0<x\leqslant 1,$$

or

$$y=(1+x^{1/\alpha})^{\alpha}-x-(2^{\alpha}-1)\leqslant 0.$$

Note that $y'=u-1$ where

$$u=x^{1/\alpha-1}(1+x^{1/\alpha})^{\alpha-1}$$

and

$$\frac{u'}{u}=\frac{1-\alpha}{\alpha}\,\frac{1}{x(1+x^{1/\alpha})}<0.$$

Therefore u is decreasing, $u(x)\geqslant u(1)=2^{\alpha-1}$, and consequently $y'\geqslant 2^{\alpha-1}-1\geqslant 0$; $y(x)$ is therefore increasing and $y(x)\leqslant y(1)=0$.

Lemma 5:

$$u_s=\max_{\Pi\in\boldsymbol{\Pi}}\sum_{i=1}^{M} e^{s\rho_{i\Pi}} \leqslant M^{1/2}\left[u^{a}_{s/(2^{\alpha}-1)}\right]^{1/2}.$$

Proof: Let

$$\rho_{i_0\Pi} \triangleq \min_i \rho_{i\Pi} \leqslant \rho_{i\Pi}, \qquad \text{for all } i\in A^T,\ \Pi\in\boldsymbol{\Pi}. \tag{17}$$

Note that $\rho_{i\Pi}=Q^{-1}\Sigma_{q=1}^{Q}|i_q-\Pi_q|^{\alpha}$ is not a distance but $(Q\rho_{i\Pi})^{1/\alpha}$ is, since it satisfies the triangular inequality

$$(Q\rho_{i\Pi})^{1/\alpha}+(Q\rho_{i_0\Pi})^{1/\alpha}\geqslant(Q\rho_{ii_0})^{1/\alpha},$$

or

$$\rho_{ii_0}\leqslant\left(\rho_{i\Pi}^{1/\alpha}+\rho_{i_0\Pi}^{1/\alpha}\right)^{\alpha}.$$

Applying Lemma 4 with $a=\rho_{i_0\Pi}$ and $b=\rho_{i\Pi}$

$$\rho_{ii_0}\leqslant(2^{\alpha}-1)\rho_{i_\Pi}+\rho_{i_0\Pi}$$

or, since $s<0$,

$$e^{(2^{\alpha}-1)s\rho_{i\Pi}}\cdot e^{s\rho_{i_0\Pi}}\leqslant e^{s\rho_{ii_0}},$$

which implies

$$\sum_i e^{(2^{\alpha}-1)s\rho_{i\Pi}}\leqslant e^{-s\rho_{i_0\Pi}}\cdot u_s^{a}. \tag{18}$$

Exponentiating (17) and summing with respect to i yields

$$u_s=\max_{\Pi\in\boldsymbol{\Pi}}\sum_{i=1}^{M} e^{s\rho_{i\Pi}}\leqslant Me^{s\rho_{i_0\Pi}}.$$

Note that since $s<0$, $(2^{\alpha}-1)s<s$, $e^{(2^{\alpha}-1)s}<e^{s}$, and thus

$$u_{(2^{\alpha}-1)s}=\sum_{i=1}^{M} e^{(2^{\alpha}-1)s\rho_{i\Pi}}<\sum_{i=1}^{M} e^{s\rho_{i\Pi}}\leqslant Me^{s\rho_{i_0\Pi}}, \tag{19}$$

(18) and (19) together imply

$$u_{(2^{\alpha}-1)s}\leqslant M^{1/2}(u_s^{a})^{1/2}$$

or

$$u_s\leqslant M^{1/2}\left[u^{\alpha}_{s/(2^{\alpha}-1)}\right]^{1/2}.$$

This inequality gives rise to the following theorem.

Theorem 3: For distortion measures of the type $\delta(1,\pi)=(1-\pi)^{\alpha}$, $\alpha\geqslant 1$, and equally likely datasets, if $s_m^{a}(D_0)=O(\log M)$ for some fixed $D_0>0$, the probabilistic system is inelastic.

Proof: The proof parallels that of Theorem 3 in [2]. We first lower bound $R(D)/\log M$ using $R_L(D,s)$ of (10), then show that the condition of Theorem 3 restricts $R_L(D,s)$ to be bounded away from zero. Let

$$u_s=\max_{\Pi\in\boldsymbol{\Pi}}\sum_i e^{s\rho_{i\Pi}},$$

then for all $s<0$ we have

$$R(D)\geqslant R_L(D,s)\triangleq \log M+sD-\log u_s.$$

In particular

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

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

$$\frac{\log u^{a}_{s_m^{a}(D)}}{\log M}\not\to 1 \tag{20}$$

by virtue of Lemma 2. From Lemma 5 we have

$$\log u^{\Pi}_{s_m^{a}(D)}\leqslant\frac{1}{2}\log M+\frac{1}{2}\log u^{a}_{s_m^{a}(D)/(2^{\alpha}-1)},$$

where

$$u_s^{\Pi}=\sum_{i=1}^{M} e^{s\rho_{i\Pi}}.$$

Using an argument similar to the one which led to (20) it can be shown that

$$\frac{\log u^{a}_{s_m^{a}(D)/k}}{\log M}\not\to 1, \qquad \text{for } k\geqslant 1,$$

which implies

$$\frac{\log u^{\Pi}_{s_m^{a}(D)}}{\log M}\not\to 1, \qquad \text{for } \Pi\in\boldsymbol{\Pi}.$$

Consequently,

$$\frac{\log u_{s_m^a(D)}}{\log M} = \frac{\max_{\Pi} \log u^{\Pi}_{s_m^a(D)}}{\log M} \not\to 1.$$

More precisely, since $\log u_{s_m^a(D)}/\log M$ is bounded away from one, there exist m_1 and $1>\eta>0$ such that

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

Since $|s_m^a(D)|$ is a decreasing function of D, the assumption $s_m^a(D_0)=O(\log M)$ implies that $s_m^a(D)$ cannot be of an 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.,

$$s_m^a(D) \sim -\phi(D)\log M, \qquad \phi(D)>0,$$

with

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

by virtue of Lemma 3. Therefore, there is a number $D_1>0$ such that

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

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

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

there exists an m_1 such that for $m \geqslant m_1$

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

i.e.,

$$\frac{s_m^a(D)}{\log M} \geqslant -3\phi(D)/2 \geqslant -3\eta/4D.$$

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

$$\frac{R(D)}{\log M} \geqslant \frac{R_L[D, s_m^a(D)]}{\log M} = 1 + \frac{D s_m^a(D)}{\log M} - \frac{\log u_{s_m^a(D)}}{\log M}.$$

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

$$\frac{R(D)}{\log M} \geqslant \frac{R_L[D, s_m^a(D)]}{\log M} \simeq 1 - D\phi(D) - \frac{\log u_{s_m^a(D)}}{\log M} \geqslant 1 - 3\eta/4 - (1-\eta) = \eta/4$$

which establishes Theorem 3.

A somewhat more general (and useful) version of this result is the following.

Theorem 4: For distortion measures of the type $\delta(1,\pi) = (1-\pi)^\alpha$, $\alpha \geqslant 1$, and equally likely datasets, if the binary system is tractably inelastic, then the probabilistic system is inelastic.

Proof: Equations (12) 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.

Theorem 5: For distortion measures such that there exists $\alpha>0$ satisfying $\delta(1,\pi) \geqslant (1-\pi)^\alpha$, if the binary system is tractably inelastic, then the probabilistic system is inelastic.

Proof: Simply note that if $\delta_1(1,\pi) \geqslant \delta_2(1,\pi)$ and if $\delta_1(1,0)=\delta_2(1,0)=1$, $R_1(D) \geqslant R_2(D)$, where $R_1(D)$ is the rate distortion function for a QA system in which δ_1 is used and $R_2(D)$ is the rate distortion function for the same QA system in which δ_2 is used as distortion measure. Consequently, if there exists an $\alpha>0$ such that Theorem 4 holds for $\delta_2(1,\pi)=(1-\pi)^\alpha$, $R_2(D)/\log M$ is bounded away from zero. Theorem 4 would also hold for $\delta_1(1,\pi) \geqslant (1-\pi)^\alpha$ since $R_1(D)/\log M \geqslant R_2(D)/\log M$ is also bounded away from zero. In particular, taking $\delta_1(1,\pi)=(1-\pi)^\beta$, $0<\beta\leqslant 1$, extends the validity of Theorem 4 to the range $\alpha>0$.

Corollary 1: For all distortion measures which are concave, if the binary system is tractably inelastic, then the probabilistic system is inelastic.

Proof: If $\delta(1,\pi)$ is concave, then $\delta(1,\pi) \geqslant 1-\pi$, and the proof used for Theorem 5 applies directly.

The validity of Theorem 5 could be further extended if the answer vocabulary is limited to a finite set of estimates: $\pi \in \{\pi_1, \pi_2, \cdots, \pi_k\}$. In this case it is sufficient to require that the distortion associated with any uncertain answer ($\pi<1$) be a finite positive quantity

$$\delta(1,\pi)>0, \qquad \text{if } \pi<1. \tag{21}$$

If (21) is satisfied for a finite vocabulary we can always find an $\alpha>0$ such that $\delta(1,\pi) \geqslant (1-\pi)^\alpha$, leading to the following result.

Corollary 2: For all distortion measures satisfying $\delta(1,\pi)>0$ for $\pi<1$, if the binary system is tractably inelastic, then any corresponding probabilistic system which employs a finite set of probabilistic answers is also inelastic.

The reason that (21) was insufficient in the continuous case is due to pathological functions such as $\delta(1,\pi)=\exp(-\pi/1-\pi)$ which satisfy (21) but cannot be lower bounded by $(1-\pi)^\alpha$ for all π. Corollary 2 is useful for QA systems which issue linguistic characterizations of the answers' credibilities, e.g., "probably", "likely", "doubtfully", etc.

V. Applications

A. Classical Reproducing Scoring Rules

A scoring rule is called *reproducing* if it encourages a probability assessor to be honest. In other words, an assessor who minimizes his/her expected loss, $p\delta(1,\pi)+(1-p)\delta(0,\pi)$, can do so only by reporting his/her actual "degree of belief" p that an event will occur. There are three such rules commonly used:

quadratic $\delta(1,\pi)=(1-\pi)^2$;

spherical $\delta(1,\pi)=1-\dfrac{\pi}{\left(\pi^2+(1-\pi)^2\right)^{1/2}}$;

logarithmic $\delta(1,\pi)=-\log\pi$.

Since the logarithmic scoring rule is undefined for $\pi=0$ the truncated logarithmic scoring rule:

$$\delta(1,\pi)=\begin{cases}-\log\pi, & \pi>\epsilon,\\ -\log\epsilon, & \pi\leqslant\epsilon,\end{cases}$$

is often used in practice [4].

The quadratic scoring rule directly falls into the category of distortion measures studied in Section IV with $\alpha=2$. For the spherical scoring rule it can be shown that $\delta(1,\pi)\geqslant(1-\pi)^2/2$; consequently this scoring rule meets the conditions of Theorem 5, except for the factor 1/2. However, since two distortion measures which differ only by a constant multiplier yield identical rate distortion functions save for a scaling factor on the arguments, we conclude that Theorem 5 applies to the spherical scoring rule. For the truncated logarithmic scoring rule it is clear that

$$\delta(1,\pi)\geqslant 1-\pi,$$

and consequently Theorem 5 applies to the logarithmic scoring rule as well. To summarize, if a binary QA system is tractably inelastic, relaxing the requirement that answers be true or false by allowing probabilistic answers will not change the asymptotic properties of the memory–quality exchange for practically all distortion measures. In particular, the classical cases of quadratic, spherical, and truncated logarithmic scoring rules fall into that category. This generalizes the results of [3] to the case where questions are not logically independent.

We next demonstrate the usefulness of Theorem 5 in analyzing a probabilistic QA system with a highly interdependent query set.

B. Example: The Complete Binary System (CBS)

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

$$M=\{0,1\}^m, \qquad M=2^m,$$
$$Q=\{f\mid f\colon M\to\{0,1\}\}, \qquad Q=2^M.$$

For such a system, every two distinct datasets produce identical answers for exactly 50 percent of the questions. Thus, using the proportion of erroneous answers as distortion criterion, the admissible distortion matrix is balanced [7],

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

and

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

Consequently the CBS is tractably inelastic.

Consider now the probabilistic version of this system. All binary assertions about the data are still permissible as queries but probabilistic responses, indicating the likelihood of each assertion, may now be generated by the system. Since the queries are highly interrelated (all 2^{2^m} queries could be deduced from only m bits), the method used in [3] is no longer applicable. Using Theorem 5, however, it follows immediately that the probabilistic CBS is inelastic under distortion measures of the form $\delta(1,\pi)\geqslant(1-\pi)^\alpha$ ($\alpha\geqslant 1$). Without invoking this theorem it would be practically impossible to provide any kind of analysis as we would have to examine a huge output vocabulary consisting of all 2^{2^m}-dimensional vectors with entries between zero and one.

The analysis of CBS can be generalized to a larger class of probabilistic QA systems.

Theorem 6: For any probabilistic QA system under distortion measures of the type $\delta(1,\pi)\geqslant(1-\pi)^\alpha$ $\alpha>0$, if $H=\log M$ and the (normalized) distance between any two distinct admissible answer strings is bafo, the system is inelastic.

Proof: Assume that

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

Then if j is admissible

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

Equation (16a) 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 from Theorem 5 the system is inelastic.

C. A System Admitting Singly Conjunctive Queries

In [2] we considered a binary QA which receives as data binary vectors $M=\{(x_1,x_2,\cdots,x_m)|x_i\in\{0,1\}\}$ and answers queries of the type, "Are both x_i and x_j ONE?" The rationale for studying this system was to test the elasticity conditions for a redundant query set where the (normalized) distance between any two admissible answer strings approached zero. Our analysis of this system used a fairly complex u_s function as a basis for demonstrating that the binary system (with the proportion of erroneous answers as a distortion criterion) is tractably inelastic. Now, by virtue of Theorem 3, we can conclude that the probabilistic version of the QA system is also inelastic for all distortion criteria of the type $\delta(1,\pi)\geqslant(1-\pi)^{\alpha}$, where $\alpha>0$.

VI. Summary

This paper investigated whether the storage requirements of question-answering systems can be greatly reduced by allowing some distortion in the form of indefinite or imprecise answers. Using an information theoretic model, it was shown that for true–false type questions, if significant reduction is impossible for a restricted set of true–false answers and a "fraction-of-wrong-answers" distortion criterion, then it is (for all practical purposes) also impossible for systems issuing indefinite answers cast in probabilistic phrasings. Thus, by giving partial credit for unsure answers rather than counting an answer as wrong if it is not both certain and correct, we are still unable to reduce greatly the storage requirement.

Although our results were derived for systems issuing probabilistic estimates, they remain largely applicable to nonnumerical answers as well. Systems producing linguistic phrases such as "almost certainly true" or "possibly true" would be subject to the same storage space constraint as those issuing numerical estimates, as long as the distortion associated with such messages can be expressed numerically. In our analysis we did not assume that the system's probabilistic answers comply with the rules of probability calculus. For example, if the truth of query q_1 subsumes the truth of q_2, then we did not insist that the respective responses π_1 and π_2 satisfy $\pi_1\geqslant\pi_2$. That leaves us the freedom of assigning a variety of interpretations to the responses issued by the system. The only requirement necessary for generalizing our results is that all answers conveying uncertainty ($\pi<1$) be assigned positive distortion measures $\delta(1,\pi)>0$.

The results established in this paper also have implications for complexity measures other than storage space. For example, one may wish to inquire whether by tolerating imprecise answers it is possible to shorten the average time required for generating the answers. In [9] it was shown that, under quite general conditions, Shannon's rate distortion function underbounds the extent to which many complexity measures such as number of gates, execution time, and state complexity can be reduced by allowing imprecision. A necessary condition for enabling the conversion of a small degree of imprecision into a drastic reduction of complexity is that the rate distortion function exhibit a similar drastic drop, i.e., that $R(D)/R(0)\to0$ as the size of the data increases. Hence the conditions for inelasticity established in this paper not only delimit the reduction of storage space due to imprecision but also impose an absolute bound on the amount of savings realizable in other computational resources. It appears that the use of probabilistic answers, although they benefit users by alerting them to variations in answers' credibility, is not sufficient to permit a significantly larger savings in computational resources than those realizable by true–false type answers.

References

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

[2] ——, "Elasticity conditions for storage versus error exchange in question-answering systems," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 6, pp. 653–664, Nov. 1979.

[3] J. Pearl, "On summarizing data using probabilistic assertions," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 4, pp. 459–465, July 1977.

[4] E. Shufford, A. Albert, and H. Massengill, "Admissible probability measurements procedures," *Psychometrica*, vol. 31, pp. 125–145, 1966.

[5] M. Minsky and S. Papert, *Perceptrons.* Cambridge, MA: MIT, 1969, ch. 12.

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

[7] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

[9] J. Pearl, "Theoretical bounds on the complexity of inexact computations," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 5, pp. 580–586, Sept. 1976.