Toward a Source Coding Theory for Sets

Lav R. Varshney Vivek K Goyal*

Abstract

The problem of communicating (unordered) sets, rather than (ordered) sequences is formulated. Elementary results in all major branches of source coding theory, including lossless coding, high-rate and low-rate quantization, and rate distortion theory are presented. In certain scenarios, rate savings of $\log n!$ bits for sets of size n are obtained. Asymptotically in the set size, the entropy rate is zero and for sources with an ordered parent alphabet, the $(0, 0)$ point is the rate distortion function.

1 Introduction

In launching information theory, Shannon [1] wrote:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

What is a "message" in this formulation? Shannon dismisses the *meaning* of the message as "irrelevant to the engineering problem," and proceeds to develop a theory that applies to (ordered) sequences of samples. This paper explores the limits of communication when messages are (unordered) sets of samples.

Though we are accustomed to the processing of vectors and time series, order is not always relevant in a communication problem. The list of ingredients in a recipe, for example, can be rearranged with no ill effect. In a more technological realm, a communication link in a packet network might be free to reorder packets within the scope of the packet headers. An application is the compression of databases. It is known that reordering techniques can significantly improve compression efficiency [2], but achievable performance limits are not known. Other applications are mentioned in the final section.

Communicating a set of size n should, of course, require fewer bits than communicating a vector of length n since there is no need to distinguish between vectors that are permuted versions of each other. How many bits can be saved? Since there are $n!$ possible permutations, it would seem that $\log_2 n!$ bits would enter the picture; indeed for continuous-valued sources and high rates we establish that this is precisely the savings. But starting with an nR-bit source code for a vector of length n and saving $\log_2 n!$ bits does not make sense if $nR < \log_2 n!$, so there must be more to the story.

Notation and basic concepts. Let X be a (possibly multivariate) random variable that takes values in $\mathcal{X}$. We refer to X and $\mathcal{X}$ as the parent random variable

*The authors are with the Signal Transformation and Information Representation Group in the Research Laboratory of Electronics at MIT (`http://www.rle.mit.edu/stir/`). Work supported in part by the NSF Graduate Research Fellowship and Grant CCR-0325774.

and parent alphabet, respectively. We are interested in communicating multisets $\{x_i\}_{i=1}^n = \{x_1, x_2, \ldots, x_n\}$, where each x_i is an independent realization of X. In some instances, the i.i.d. assumption is relaxed. Rates are measured in bits per parent letter, and distortion measures are also normalized by set size.

Since a multiset of samples $\{x_i\}_{i=1}^n$ can be seen as a letter drawn from an alphabet of multisets, our theory does not lie outside of Shannon's theory. However, we obtain scaling with respect to n that is quite different than the problem of communicating the vector $(x_i)_{i=1}^n = (x_1, x_2, \ldots, x_n)$.

We have not thus far assumed that there is a total order defined on $\mathcal{X}$. When $\mathcal{X}$ is discrete (Section 2), an order is not important. Our results for continuous $\mathcal{X}$ (Section 3) requires a total order. Then distortion measures for multiset communication are naturally transformed into distortion measures on the vector of order statistics. Order statistics are detailed in Section 3.

Relationship with permutation codes. Permutation source codes are codes that only preserve a partial ordering relationship among elements of a vector; see [3] and references therein. As we show in (1), the ordering information and the value information of a vector are independent. Hence the residual uncertainty associated with the distortion-minimizing permutation code is exactly the uncertainty that is coded in multiset communication. The complementarity of permutation codes and multiset communication may be used for multiple descriptions or successive refinement.

2 Sets with a Discrete Parent

Sets of any size. Lossless coding of a multiset is equivalent to lossless coding of a sequence that has the same elements ordered in a deterministic way. We can decompose the entropy of an arbitrary sequence source into two independent parts: the entropy of the multiset (values) and the entropy of the ordering. Define $H((X_i)_{i=1}^n)$ as vector entropy, $H(\{X_i\}_{i=1}^n)$ as multiset entropy, J_n as a random variable representing ordering, and $H(J_n)$ its entropy. Suppressing subscripts,

$$\begin{aligned} H((X)) &\overset{(a)}{=} H((X)) + H(\{X\}) - H((X)|J) \\ &= H(\{X\}) + I((X);J) \\ &\overset{(b)}{=} H(\{X\}) + H(J) - H(J|(X)) \\ &= H(\{X\}) + H(J), \end{aligned} \tag{1}$$

where (a) follows from noting that $H(\{X\}) = H((X)|J)$ and (b) from $H(J|(X)) = 0$. Since there are only $n!$ possible orderings, an upper bound on $H(J_n)$ leads to the lower bound

$$H(\{X_i\}_{i=1}^n) \geq H((X_i)_{i=1}^n) - \log n!. \tag{2}$$

The lower bound is not tight due to the positive chance of ties among members of a multiset drawn from a discrete parent. If the chance of ties is small (if $|\mathcal{X}|$ is sufficiently large and n is sufficiently small), the lower bound is a good approximation.

One interpretation of sequence entropy reduction by order entropy to yield multiset entropy is of a multiset as an equivalence class of sequences. Since an error event only occurs when a sequence is represented by a sequence outside of its equivalence class, uncertainty within the class is allowable without incurring error.

Rather than bounding $H(\{X\})$, we can compute it exactly. Notice that the number of occurrences of each $x \in \mathcal{X}$ fully specifies a multiset, i.e. the type (empirical frequency) is sufficient to describe a multiset. If the multiset is drawn i.i.d., the distribution of types is given by a multinomial distribution. Suppose $x_i \in \mathcal{X}$ has probability p_i and let K_i be the number of occurrences of x_i in n independent trials. Then

$$\Pr[K_i = k_i] = \binom{n}{k_1, k_2, \ldots, k_{|\mathcal{X}|}} \prod_{i=1}^{|\mathcal{X}|} p_i^{k_i}, \qquad \text{for } i = 1, \ldots, |\mathcal{X}|,$$

for any type $(k_1, k_2, \ldots, k_{|\mathcal{X}|})$ of non-negative integers with sum n. Thus,

$$H(\{X\}_{i=1}^n) = H(K_1, K_2, \ldots, K_{|\mathcal{X}|}; n), \tag{3}$$

where dependence on n is made explicit. Denote the alphabet of distinct types as $\mathcal{K}(\mathcal{X}, n)$; its size may be computed and bounded through simple combinatorics [4]:

$$|\mathcal{K}(\mathcal{X}, n)| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \leq (n+1)^{|\mathcal{X}|}. \tag{4}$$

By the source coding theorem [1, Theorem 9], we need a rate of at least $H(\{X\}) = H((K); n)$ to code the multiset with arbitrarily small probability of error.

Large-set asymptotics. The previous entropy calculation was for fixed and finite n; now we turn to coding properties asymptotic in n. Define the entropy rate of a multiset as

$$H(\mathfrak{X}) = \lim_{n\to\infty} \tfrac{1}{n} H(\{X_i\}_{i=1}^n). \tag{5}$$

We can show that the entropy rate is in fact zero.

Theorem 1. *The entropy rate for any multiset drawn from a finite-symbol parent is zero.*

Proof. The entropy rate is given by $H(\mathfrak{X}) = \lim_{n\to\infty} \frac{1}{n} H(K_1, K_2, \ldots, K_{|\mathcal{X}|}; n)$, because of the equivalence relation (3). Using the logarithm of the alphabet size upper bound and the bound on the number of types (4),

$$H(\mathfrak{X}) \;\leq\; \lim_{n\to\infty} \tfrac{1}{n} \log |\mathcal{K}| \;\leq\; \lim_{n\to\infty} \tfrac{1}{n} \log \left((n+1)^{|\mathcal{X}|} \right).$$

Since $|\mathcal{X}|$ is a finite constant, evaluating the limit yields $H(\mathfrak{X}) \leq 0$. Furthermore, by the information inequality $H(\mathfrak{X}) \geq 0$, so $H(\mathfrak{X}) = 0$. □

Note that the theorem holds for any multiset, not just for multisets drawn i.i.d. In fact, if the multiset is drawn i.i.d., the bounding technique yields an upper bound that is quite loose. To achieve the bound with equality, each of the types would have to be equiprobable; however by the strong AEP [5], collectively, all non-strongly typical types will occur with probability as close to zero as desired. The number of types in the strongly typical set is polynomial in n, so we cannot get a rate of convergence to the zero entropy rate faster than $O(\log n/n)$.

Example 1. *Consider a set of size n drawn i.i.d. from a Bernoulli parent with parameter p. The entropy of the set is equal to the entropy of a binomial random variable with parameters n and p. An asymptotic expression for this entropy is given in [6] as*

$$H(\{X_i\}_{i=1}^n) \sim \tfrac{1}{2}\log_2\left(2\pi e n p(1-p)\right) + \sum_{k\geq 1} a_k n^{-k},$$

for some constants a_k. Thus the entropy rate is

$$H(\mathfrak{X}) = \lim_{n\to\infty} \tfrac{\log_2(2\pi e n p(1-p))}{2n} + \sum_{k\geq 1} a_k n^{-k-1} = 0.$$

Evidently, the rate of convergence is $O(\log n/n)$ as in the universal case of Theorem 1.

We know that if we can make the multiset large enough, we will require zero average rate to achieve arbitrarily small probability of error, but what if we cannot take large multisets and we have rate constraints tighter than entropy? Returning to the fixed and finite n regime, we find the rate distortion function for multisets with error frequency distortion. Through the equivalence of multisets and types, this is simply an i.i.d. discrete source with error frequency distortion, so the reverse waterfilling solution of Erokhin [7] applies. The rate distortion function is given parametrically as

$$\begin{aligned}
D_\theta &= 1 - S_\theta + \theta(N_\theta - 1)\\
R_\theta &= -\sum_{k:p(k)>\theta} p(k)\log p(k) + (1-D_\theta)\log(1-D_\theta) + (N_\theta - 1)\theta\log\theta,
\end{aligned}$$

where N_θ is the number of types whose probability is greater than θ and S_θ is the sum of the probabilities of these N_θ types. The parameter θ goes from 0 to $p(k^\ddagger)$ as D goes from 0 to $D_{max} = 1 - p(k^\dagger)$; the most probable type is denoted $k^\dagger$ and the second most probable type is denoted $k^\ddagger$.

3 Sets with a Continuous Parent

Distributions and entropies. Although the method of types does not extend well to continuous alphabets, the intuition that number of occurrences fully specify multisets continues to hold. When multisets are drawn i.i.d. from a continuous parent, however, the probability of ties is zero and multisets are sets with no multiplicity. Moreover, uncountably many distinct sets may occur. Rather than working directly from types, we use the sequence in which the elements are in ascending order, the natural representative of a permutation-invariant equivalence class. This canonical sequence representation is equivalent to type representation and naturally leads to the framework of order statistics. Assuming that the parent alphabet consists of the real line, the basic distribution theory of order statistics can be used [8].

When the sequence of random variables $X_1, \ldots, X_n$ is arranged in ascending order as $X_{(1)} \leq \ldots \leq X_{(n)}$, $X_{(r)}$ is called the rth order statistic. Suppose that $X_1, \ldots, X_n$

are i.i.d. from the parent with density $f(x)$. Then the marginal density of $X_{(r)}$ is well-known in closed form [8] and the joint density of all n order statistics is

$$f_{(1)\cdots(n)}(x_1, \dots, x_n) = \begin{cases} n! f(x_1) \cdots f(x_n), & \text{for } x_1 \leq \cdots \leq x_n; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The order statistics have the Markov property with transition probability

$$f_{X_{(r+1)}|X_{(r)}=x}(y) = (n-r) \left[\frac{1-F(y)}{1-F(x)} \right]^{n-r-1} \frac{f(y)}{1-F(x)}, \quad \text{for } y > x.$$

Based on these basic distributional properties of order statistics, we can derive the differential entropies of order statistics. The individual marginal differential entropies cannot be simplified from their integral forms unless the parent is specified. The average marginal differential entropy, however, can be expressed in terms of the parent differential entropy and a constant that depends only on n [9]:

$$\bar{h}(X_{(1)}, \dots, X_{(n)}) = \frac{1}{n} \sum_{i=1}^{n} h(X_{(i)}) = h(X_1) - \log n - \frac{1}{n} \sum_{i=1}^{n} \log \binom{n-1}{i-1} + \frac{n-1}{2}. \quad (7)$$

The subtractive constant is positive and increasing in n. The individual conditional differential entropies also cannot be simplified much without parent specification. Again, as in the marginal case, the total conditional differential entropy can be expressed in terms of the parent differential entropy and a constant that depends only on n. Due to Markovianity, the sum of the individual conditional differential entropies is in fact the joint differential entropy.

$$h(X_{(1)}, \dots, X_{(n)}) = h(X_{(1)}) + \sum_{i=1}^{n-1} h(X_{(i+1)}|X_{(i)}) = nh(X_1) - \log n!.$$

Notice that an analogous statement (2) was a lower bound in the discrete parent case; equality holds in the continuous case since there are no ties. If we generalize the expression $\log n!$ to $H(J)$, this equality holds for all sets, not just those drawn i.i.d.

High rate quantization. Having computed marginal and conditional differential entropies of order statistics, asymptotic high rate quantization results follow directly. We introduce four quantization schemes in turn, focusing on high rate quantizer approximations under squared error fidelity. In particular, we sequentially introduce a *shape advantage*, a *memory advantage*, and a *space-filling advantage* as in [10].[1] As a baseline, take the naïve scheme of directly scalar quantizing the randomly ordered sequence. The average rate and distortion per source symbol of the naïve scheme are $R_1 = h(X) - \log \Delta$, and $D_1 = \Delta^2/12$, where Δ is the quantization step size. Now scalar quantize the deterministically ordered sequence (the order statistics). This

[1]Note that vector quantizer advantages are discussed in terms of distortion for fixed rate in [10], but we discuss some of these advantages in terms of rate for fixed distortion.

Table 1: Comparison between the Scheme 1 and several other quantization schemes. The symbols (s),(m), and (f) denote shape, memory, and space-filling advantages.

	Rate Reduction ($-$)	Distortion Reduction ($\times$)
Scheme 1	0	1
Scheme 2 (s)	$\log n + \frac{1}{n}\sum_{i=1}^{n} \log \binom{n-1}{i-1} - \frac{n-1}{2}$	1
Scheme 3 (s,m)	$\log n!/n$	1
Scheme 4 (s,m,f)	$\log n!/n$	$1/G(n)$

changes the shape of the marginal distributions resulting in shape advantage. The average rate per source symbol for this scheme is

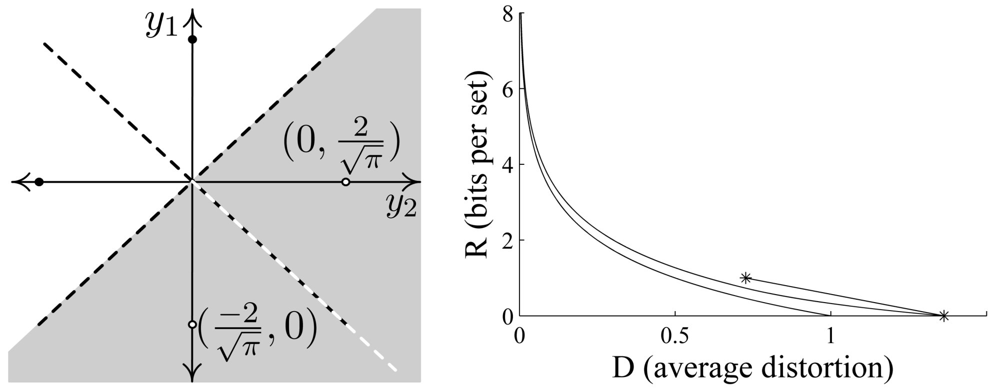
$$R_2 = \bar{h}(X_{(1)}, \ldots, X_{(n)}) - \log \Delta = R_1 - \log n - \frac{1}{n}\sum_{i=1}^{n} \log \binom{n-1}{i-1} + \frac{n-1}{2}. \quad (8)$$

The distortion is the same as the naïve scheme, $D_2 = D_1$. As a third scheme, scalar quantize the order statistics sequentially, using the previous order statistic as a form of side information (assuming perfect side information). Since the order statistics form a Markov chain, this single-letter sequential transmission exploits all available memory advantage. The rate for this scheme is

$$R_3 = \tfrac{1}{n}h(X_{(1)}, \ldots, X_{(n)}) - \log \Delta = R_1 - \tfrac{1}{n}\log n!. \quad (9)$$

Again, $D_3 = D_1$. Finally, the fourth scheme would vector quantize the entire sequence of order statistics collectively, exploiting space-filling gain. The rate is the same as the third scheme, $R_4 = R_3$, however the distortion is less. This distortion reduction is related to the best packing of polytopes and is not known in closed form for most values of n; see Table I of [10] and more recent work. We denote the distortion as $D_4 = D_1/G(n)$, where $G(n)$ is a function greater than unity. The performance improvements of the presented schemes are summarized in Table 1. Notice that all values in Table 1 depend only on the set length n and not on the parent distribution.

Low rate quantization. Having characterized the high rate regime, we mention some properties of optimal MSE quantization of order statistics. Optimal MSE (Lloyd-Max) quantization for order statistics was studied in [11] for the separate scalar quantization of order statistics, like Scheme 2. Here we comment on full vector quantization, like Scheme 4. Unlike scalar quantization, where the operations of sorting and quantizing can always be interchanged without loss of optimality [11], this is true only under certain conditions for vector quantization. As given in (6), the joint density of the n order statistics is a scaled version of the n-fold product of the parent density supported over a cone. The cone of support is one of $n!$ cones that form a disjoint partition of $\mathbb{R}^n$ such that each equivalence class consists of one point from each cone; the geometry is related to permutation polyhedra. Now if the representation points for an MSE-optimal (k bit, n dimension) order statistic quantizer are the intersection of the representation points for an MSE-optimal ($k + \log n!, n$) quantizer for the unordered variates and the cone of support, then we can interchange sorting



(a) (b)

Figure 1: Quantization and rate distortion for bivariate standard normal order statistics. (a) Optimal one-bit quantizer (white) achieves $\left(R = 1, D = \frac{2\pi-4}{\pi}\right)$. Optimal two-bit quantizer (black) for unordered variates achieves $\left(R = 2, D = \frac{2\pi-4}{\pi}\right)$. Since representation points for order statistic quantizer are the intersection of the cone (shaded) and the representation points for the unordered quantizer, the distortion performance is the same. (b) Shannon upper and lower bounds for the order statistic rate distortion function. The point achievable by quantizer in (a) is also shown connected to the zero rate point, which is known to be tight. Note that this is not normalized per source letter.

and quantization without loss of optimality. This condition can be interpreted as a requirement on permutation polyhedral symmetry of the quantizer of the unordered variates. In fact, the distortion performance of the MSE-optimal (k, n) order statistic quantizer is equal to the distortion performance of the best $(k + \log n!, n)$ unordered quantizer constrained to have the required permutation symmetry. An example where the symmetry condition is met is shown in Figure 1(a).

Rate distortion for large sets. In our discussion of multisets from discrete parents, we established that the entropy rate is zero. In a similar vein we will make the average distortion negligible using zero rate, with asymptotically large set size. In particular, we will look at the squared error distortion measure

$$D = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[(X_{(i)} - \hat{X}_{(i)})^2\right].$$

Since there is no rate, the best choice is $\hat{X}_{(i)} = \mathrm{E}[X_{(i)}]$, thus the distortion reduces to

$$D_n(R = 0) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{var}[X_{(i)}],$$

where dependence on n is explicitly noted. We will show that $\lim_{n\to\infty} D_n(0) = 0$ for a very large class of parent distributions, but first an example.

Example 2. *Consider a parent distribution, $p_X \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$. The average variance may be computed in closed form as $D_n(0) = 2/(n + 1)$, so $\lim_{n\to\infty} D_n(0) = 0$.*

The general theorem on zero rate, zero distortion will be based on the parent quantile function, the generalized inverse of the distribution function

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\},$$

and the empirical quantile function, defined in terms of order statistics,

$$Q_n(u) = X_{(\lfloor un \rfloor + 1)} = F_n^{-1}(u),$$

where $F_n(\cdot)$ is the empirical distribution function. The quantile function $Q(\cdot)$ is continuous if and only if the distribution function has no flat portions in the interior, i.e. the density is strictly positive over its region of support except perhaps on isolated points. The main step of the proof will be a Glivenko-Cantelli like theorem for empirical quantile functions [12].

Theorem 2. *Let X_1 satisfy*

$$\mathrm{E}\,|\min(X_1, 0)|^{1/v_1} < \infty \quad \textit{and} \quad \mathrm{E}(\max(X_1, 0))^{1/v_2} < \infty \tag{10}$$

for some $v_1 > 0$ and $v_2 > 0$ and have continuous quantile function $Q(u)$. Then for the coding of size-n sets drawn with parent distribution of X_1 we have

$$\lim_{n \to \infty} D_n(R = 0) = 0. \tag{11}$$

Proof. For any nonnegative function w defined on $(0, 1)$, define a weighted Kolmogorov-Smirnov like statistic $\Delta_n(w) = \sup_{0<u<1} w(u)\,|Q_n(u) - Q(u)|$. For each $v_1 > 0$, $v_2 > 0$, and $u \in (0, 1)$, define the weight function $w_{v_1,v_2}(u) = u^{v_1}(1-u)^{v_2}$. Assume that Q is continuous, choose any $v_1 > 0$ and $v_2 > 0$, and define $\gamma = \limsup_{n\to\infty} \Delta_n(w_{v_1,v_2})$. Then by the result of Mason [12], $\gamma = 0$ with probability 1 when (10) holds. Our assumptions on the parent meet this condition, so $\gamma = 0$ with probability 1. This implies that

$$\limsup_{n\to\infty} |X_{(\lfloor un \rfloor+1)} - Q(u)| \leq 0 \text{ for all } u \in (0, 1) \text{ w.p.1},$$

and since the absolute value is nonnegative, the inequality holds with equality. So we have almost sure convergence of all order statistics to associated deterministic quantile function constants. The bounded moment condition on the parent, (10), implies a bounded moment condition on the order statistics. Almost sure convergence together with the bounded moment condition implies convergence in moment. Since there is convergence in moment to a Dirac delta function distribution, the variance of each order statistic is zero, and thus the average variance is zero. □

We have established that asymptotically in n, the point $(R = 0, D = 0)$ is achievable; combining with the information inequality lower bound, this is in fact the rate distortion function. Due to the generality of the Glivenko-Cantelli like theorem that we used, the result will stand for a very large class of distortion measures.

Rate distortion for a small set. Returning to the fixed and finite n regime, we give some bounds on the rate distortion function for the independent bivariate standard normal order statistics. For distortion

$$d(\vec{x}, \hat{\vec{x}}) = (x_{(1)} - \hat{x}_{(1)})^2 + (x_{(2)} - \hat{x}_{(2)})^2,$$

the Shannon lower bound is simply $R_{SLB}(D) = \log(1/D)$, the usual Gaussian rate distortion function reduced by $\log n!$ bits (one bit). Note that since the order statistic source cannot be written as the sum of two independent processes, one of which has the properties of a Gaussian with variance D,[2] the Shannon lower bound is loose everywhere [13], though it becomes asymptotically tight in the high rate limit.

The covariance matrix of the Gaussian order statistics can be computed as

$$\Lambda = \begin{bmatrix} 1 - 1/\pi & 1/\pi \\ 1/\pi & 1 - 1/\pi \end{bmatrix},$$

with eigenvalues 1 and $1 - 2/\pi$. Reverse waterfilling yields the Shannon upper bound

$$R_{SUB}(D) = \begin{cases} \frac{1}{2}\log\left(\frac{2-4/\pi}{D}\right) + \frac{1}{2}\log\left(\frac{2}{D}\right), & 0 \leq D \leq 2 - 4/\pi \\ \frac{1}{2}\log\left(\frac{1}{D-1+2/\pi}\right), & 2 - 4/\pi \leq D \leq 2 - 2/\pi \\ 0, & D \geq 2 - 2/\pi. \end{cases}$$

This bound is tight at the point achieved by zero rate. Since the Gaussian order statistics for $n = 2$ have small non-Gaussianity, the Shannon lower bound and the Shannon upper bound are close to each other, as shown in Figure 1(b). For moderately small distortion values, we can estimate the rate distortion function quite well.

4 Comments

An assumption in basic information theory is that the encoder and decoder share a knowledge of a code, which is essentially the same as sharing a probabilistic model for messages. Given this shared knowledge and the relative frequency interpretation of probability, with associated laws of large numbers and Glivenko-Cantelli theorems, our asymptotic results in set size are not unexpected. After all, if the empirical frequency converges to the true density (known to the decoder), then it makes sense that for large n, little rate is required to achieve low distortion: the decoder can produce a good estimate just by generating samples according to the known distribution. If the encoder does not know the model of the source, then universal source coding schemes that learn the source are necessary. In fact, the problem of learning a fixed but unknown parent distribution and the problem of coding sets are related. The asymptotic results on the redundancy of universal source codes going to zero [14] are quite similar to our results on the entropy rate itself going to zero.

One avenue for building upon this work was mentioned in the introduction: to develop a theory for database compression. Another is to consider cases in which encoder and decoder do not share perfect knowledge of the source distribution.

In the introduction, it was suggested that there are many source-destination pairs that have the order irrelevance property. If the destination is to perform a permutation-invariant function on the received sequence, then the order is irrelevant.

[2] Even though $X_{(1)} = \frac{1}{2}(X_1 + X_2) - \frac{1}{2}|X_1 - X_2|$ and $X_{(2)} = \frac{1}{2}(X_1 + X_2) + \frac{1}{2}|X_1 - X_2|$, and the first terms are Gaussian, the troublesome part is the independence.

If a database is used for simple recall tasks, the order of the retrieved records is irrelevant, so vectors of records can be stored as multisets of records. As another example, suppose that a continuous-time signal can be decomposed into the sum of n shifted versions of a kernel function, ϕ; the signal reconstruction $\hat{f}(t) = \sum_{i=1}^{n} \phi(t - t_i)$ from $[t_1, t_2, \ldots, t_n] \in \mathbb{R}^n$ is clearly permutation invariant. This is exploited in [15] for the transportation of kernel density estimates. Thus in this class of examples, the multiset communication problem offers savings not only due to the nature of the source, but also due to what Shannon [1] called "the nature of the final destination of the information."

Acknowledgment. The authors thank the reviewers for constructive comments and referring them to [6].

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July and Oct. 1948.

[2] A. L. Buchsbaum, G. S. Fowler, and R. Giancarlo, "Improving table compression with combinatorial optimization," *J. ACM*, vol. 50, pp. 825–851, Nov. 2003.

[3] V. K. Goyal, S. A. Savari, and W. Wang, "On optimal permutation codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2961–2971, Nov. 2001.

[4] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

[5] R. W. Yeung, *A First Course in Information Theory.* New York: Kluwer, 2002.

[6] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1072–1081, May 1999.

[7] V. Erokhin, "ε-entropy of a discrete random variable," *Theory Probab. Appl.*, vol. 3, pp. 97–100, 1958.

[8] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ: Wiley, 2003.

[9] K. M. Wong and S. Chen, "The entropy of ordered sequences and order statistics," *IEEE Trans. Inform. Theory*, vol. 36, pp. 276–284, Mar. 1990.

[10] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1020–1033, Sept. 1989.

[11] P. P. Gandhi, "Optimum quantization of order statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 2153–2159, Sept. 1997.

[12] D. M. Mason, "Some characterizations of almost sure bounds for weighted multidimensional empirical distributions and a Glivenko-Cantelli theorem for sample quantiles," *Z. Wahrscheinlichkeitstheor. verw. Geb.*, vol. 59, pp. 505–513, May 1982.

[13] A. M. Gerrish and P. M. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 265–271, Oct. 1964.

[14] L. D. Davisson, "Universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[15] A. T. Ihler, J. W. Fisher, III, and A. S. Willsky, "Using sample-based representations under communications constraints," MIT LIDS, 2601, Dec. 2004.