On Universal Coding of Unordered Data

Lav R. Varshney and Vivek K Goyal
Signal Transformation and Information Representation Group
Research Laboratory of Electronics
Massachusetts Institute of Technology

***Abstract*—There are several applications in information transfer and storage where the order of source letters is irrelevant at the destination. For these source-destination pairs, *multiset communication* rather than the more difficult task of *sequence communication* may be performed. In this work, we study universal multiset communication. For classes of countable-alphabet sources that meet Kieffer's condition for sequence communication, we present a scheme that universally achieves a rate of $n + o(n)$ bits per multiset letter for multiset communication. We also define redundancy measures that are normalized by the logarithm of the multiset size rather than per multiset letter and show that these redundancy measures cannot be driven to zero for the class of finite-alphabet memoryless multisets. This further implies that finite-alphabet memoryless multisets cannot be encoded universally with vanishing fractional redundancy.**

I. Introduction

When storing a collection of music in a portable media player, music files can be stored in any arbitrary order without causing change to the collection. The playlist (order permutation) is distinct from the songs themselves (value multiset), which are all that are usually needed. Similar to multimedia databases, the order invariance property also holds for databases of financial transactions, telephone records [1], and scientific data. The permutation invariance also applies to human memory, where order does not affect recognition or recall performance [2].

If a sample consists of independent observations from the same distribution, then associated minimum variance unbiased estimators are symmetric in the observations [3]. Therefore when coding for estimation, the multiset of observations (rather than the sequence of observations) is all that need be represented, cf. [4]. Moving beyond the point-to-point case, in distributed inference, particle-based [5], [6] and kernel-based [7], [8] representations of densities must often be communicated. As

$$p(x) = \sum_i \phi\left(x - x_i\right) = \sum_i \phi\left(x - x_{\pi(i)}\right)$$

for any permutation $\pi(\cdot)$, the multiset of representation coefficients $\{x_i\}$ may be communicated rather than the sequence of these values (x_i).

In previous work we looked at several source coding problems associated with the communication of multisets when the underlying source distribution is known [9], [10]. Most prominently in estimation and inference but also in other applications, however, the underlying distribution is not known. The underlying distribution to be estimated interacts with measurement noise to produce a composite source with unknown side information. Hence, we are also interested in what can be achieved universally and what cannot. In this short paper, first we review lossless coding results. Then we present some universal achievability results and also some results that show that there are no universal schemes that achieve arbitrarily small redundancy in a stronger sense that we define.

In addition to the standard uses of parentheses such as for arguments of functions, we also use parentheses and braces to distinguish between sequences and multisets. For the ordered sequence that is often denoted X_1^n, we also write $(X_i)_{i=1}^n$. When these n symbols are taken as an unordered multiset, we write $\{X_i\}_{i=1}^n$. In both cases, the limits 1 and n are sometimes omitted. We also use standard asymptotic notation such as $o(\cdot)$ and $O(\cdot)$ [11]. All logarithms are binary.

II. Lossless Multiset Coding: Known Distribution

In a source sequence, the order information (permutation) and value information (multiset) can be separated as nonlinear transform coefficients. When the sequence is i.i.d., the order and value are independent [10], [12], seen either through informational manipulations or as a comsequence of Basu's theorem on sufficient statistics [13]. Since order information is irrelevant to multiset communication, we need not discuss it further; our problem reduces to compressing value information. For discrete alphabets, the only information that we are interested in preserving is the type class of the source sequence, also known variously as the histogram, the empirical distribution, and the composition class.

A. Finite Alphabets

Since a multiset of samples $\{x_i\}$ can be cast as a superletter drawn from an alphabet of multisets, the lossless block-to-variable source coding theorem [14] applies and gives the asymptotically tight entropy lower bound on the rate required for representation. If the samples in the multiset are drawn i.i.d., then the distribution of types is given by a multinomial distribution. For a multiset with Bernoulli θ elements, we can use the de Moivre approximation of the binomial, $\mathcal{N}(n\theta, n\theta(1-\theta))$ evaluated at the integers [15, pp. 243–259], and the associated high rate approximation of Gaussian entropy with interval size 1 to get

$$H\left(\{X_i\}_{i=1}^n\right) = \tfrac{1}{2}\log(2\pi e n\theta(1-\theta)) + o(1). \qquad (1)$$

Supported in part by a National Science Foundation Graduate Research Fellowship and the Texas Instruments Leadership Universities Program.

The $o(1)$ term may be obtained by analytic information theory methods [16]. Similarly, a Gaussian approximation to multinomial distributions with larger alphabet size $|\mathcal{X}|$ yields

$$H\left(\{X_i\}_{i=1}^n\right) = \tfrac{|\mathcal{X}|-1}{2}\log(Kn) + o(1), \qquad (2)$$

for constant K that depends on the source parameters. Again, the $o(1)$ term may be determined by analytic information theoretic means. For arbitrary probabilistic generation of multiset elements, we can find an alphabet-size upper bound to the entropy by enumerating the type classes:

$$H\left(\{X_i\}_{i=1}^n\right) \le \log\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} \le (n+1)^{|\mathcal{X}|}.$$

Defining the entropy rate of the multiset as,

$$H(\mathfrak{X}) = \lim_{n\to\infty} \tfrac{1}{n} H\left(\{X_i\}_{i=1}^n\right),$$

we readily see that $H(\mathfrak{X}) = 0$ for any source distribution. In fact this result is the foundation of Fitingof's universal sequence coding scheme [17]. Although memoryless sources and sources with equiprobable types both require zero rate, there is clearly a difference between these requirements; we will examine the distinction in more detail in later sections.

B. Countable Alphabets

The previous discussion has dealt with the entropy of finite-alphabet multisets, but what about countable alphabets? The enumeration of types does not yield a sublinear upper bound on the entropy. To see what happens, we invoke a sequence decomposition not dissimilar to the decomposition into order and value. A sequence may be decomposed into a *dictionary*, Δ, and a *pattern*, (Ψ_i), where the dictionary specifies which letters from the countable alphabet have appeared and the pattern specifies the order in which these letters have appeared [18]. For a sequence (x_i), dictionary entry δ_k (from $\mathcal{X}$) is the kth distinct letter in the sequence and pattern entry ψ_i (from $\mathbb{Z}^+$) is the dictionary index of the ith letter in the sequence. Note that the type of a pattern, denoted as $\{\Psi_i\}$, and the pattern of a multiset are the same. It can be seen that a multiset is determined by Δ and $\{\Psi_i\}$; the order of the pattern, $J(\Psi)$, is not needed.

Based on [11], we show that the entropy rate of a multiset generated by a discrete finite-entropy stationary process and the entropy rate of its pattern are equal. First note that the dictionary of a sequence and the dictionary of its associated multiset are the same; we use Δ to signify either one. Since $\{X_i\}$ determines $\{\Psi_i\}$ and since given $\{\Psi_i\}$, there is a one-to-one correspondence between $\{X_i\}$ and Δ,

$$\begin{aligned} H(\{X_i\}) &= H(\{\Psi_i\}) + H(\{X_i\} \mid \{\Psi_i\}) \\ &= H(\{\Psi_i\}) + H(\Delta \mid \{\Psi_i\}). \end{aligned}$$

If we can show that $H(\Delta|\{\Psi_i\})$ is $o(n)$, then it will follow that the entropy rate of the multiset, $H(\mathfrak{X})$, is equal to the entropy rate of the pattern of the multiset

$$H(\Psi) = \lim_{n\to\infty} \tfrac{1}{n} H(\{\Psi_i\}_{i=1}^n).$$

Noting the fact that the dictionary Δ is independent of the order of the pattern, $J(\Psi)$, we establish that

$$\begin{aligned} \lim_{n\to\infty} \tfrac{1}{n} H(\Delta|(\Psi_i)_{i=1}^n) &= \lim_{n\to\infty} \tfrac{1}{n} H(\Delta|J((\Psi_i)_{i=1}^n), \{\Psi_i\}_{i=1}^n) \\ &= \lim_{n\to\infty} \tfrac{1}{n} H(\Delta \mid \{\Psi_i\}_{i=1}^n), \end{aligned}$$

where the first step follows since the order and type of pattern determine the pattern, and the second step is due to independence. The result [11, Theorem 9] shows that for all discrete finite-entropy stationary processes, the asymptotic per-letter values of $H((\Psi_i)))$ and $H((X_i))$ are equal. Hence,

$$\lim_{n\to\infty} \tfrac{1}{n} H(\Delta \mid (\Psi_i)_{i=1}^n) = 0,$$

and so

$$\lim_{n\to\infty} \tfrac{1}{n} H(\Delta \mid \{\Psi_i\}_{i=1}^n) = 0.$$

This implies that $H(\Delta|\{\Psi_i\})$ is $o(n)$ and yields the following theorem.

Theorem 1: The entropy rate of a multiset generated by a discrete finite-entropy stationary process and the entropy rate of its pattern coincide:

$$H(\mathfrak{X}) = H(\Psi).$$

Computing the entropy rate of the multiset or equivalently of the pattern of the multiset can be difficult. See [19] and references therein for a discussion on computing the entropy of patterns; the entropy computation for patterns of multisets is closely related.

III. Universal Achievability

In this section, we propose a source coding scheme that achieves some degree of compression for all members of a source class at the same time. We will not compare to the entropy bound, holding off detailed discussion of redundancy until Section IV.

Consider classes of countable-alphabet i.i.d. sources that meet Kieffer's condition for universal encodability for the sequence representation problem [20], [21]. For these source classes, the redundancy for encoding $(X_i)_{i=1}^n$ is $o(n)$. We formulate a universal scheme for the multiset representation problem and demonstrate an achievability result, making use of the dictionary-pattern decomposition.

As we saw in Section II, a multiset can be represented as the concatenation of the pattern of the multiset and the dictionary. Consider the rate requirements of these two parts separately. First, let us bound the rate that is required to represent the pattern of the multiset or the type of the pattern. We can make use of the fact that there are 2^{n-1} types of patterns. This enumeration follows because the types are sequences of positive integers that sum to n. These can appear in any order, thus we are counting ordered partitions. It is well known that there are 2^{n-1} ordered partitions, which can be seen as determining arrangements of $n-1$ possible separations of n places. Thus the rate requirement for an enumerative universal scheme representing the type of pattern is $n-1$ bits.

Now to determine the rate requirement of the dictionary given the type of pattern. If the underlying distribution were known, we saw in Theorem 1 that $H(\Delta|\{\Psi_i\}_{i=1}^n)$ is $o(n)$. It was shown in [18], that there is an $O(\sqrt{n})$ upper bound on the pattern redundancy, independent of $|\mathcal{X}|$. Since this is sublinear, the asymptotic per-letter redundancy in coding a class of sequences $(X_i)_{i=1}^n$ from a countable alphabet coincides with the asymptotic per-letter redundancy in coding the dictionary given the pattern. Since we are considering a class that meets Kieffer's condition, we find the redundancy in coding the dictionary given the pattern is $o(n)$. This also carries over to coding the dictionary given the pattern of the multiset, due to the independence between dictionary and order of pattern that we had promulgated in Section II. Since both $H(\Delta|\{\Psi_i\}_{i=1}^n)$ and the redundancy in coding the dictionary given the pattern of the multiset are $o(n)$, the total rate requirement for coding the dictionary given the pattern of the multiset is $o(n)$.

Adding together the rate requirements for the two parts yields the following achievability theorem.

Theorem 2: Consider any i.i.d. source class that is universally encodable, then the multiset $\{X_i\}_{i=1}^n$ can be encoded with $n + o(n)$ bits.

Proof: A representation consists of the concatenation of the type of pattern and the dictionary given the type of pattern. The first part requires $n-1$ bits. The second part requires $o(n)$ bits. The total rate is then $n + o(n)$. ■

The rate requirement is universally reduced from $[0, \infty)$ bits per multiset letter for the sequence problem to 1 for the multiset problem.

IV. Unattainability of Zero Redundancy

For finite alphabets, we showed that the entropy rate is zero for any source and that if we simply enumerate the type classes, this requires zero rate asymptotically per multiset letter. Hence such a universal scheme requires zero rate for any source. If we take a finer look at the asymptotic rate, normalizing by the logarithm of the multiset size rather than by the multiset size, we see that there is a distinction between zero and zero. In this section, we define new information and redundancy measures. We will find that zero-redundancy universal coding of multisets is *not* possible with respect to the class of memoryless multisets, using the more stringent redundancy requirement.

A. Log-Blocklength Normalized Information Measures

We formulate several definitions and extend the source coding theorems to these definitions. We define the log-blocklength normalized entropy rate to be

$$\mathfrak{H}(\mathfrak{Z}) = \lim_{n\to\infty} \frac{H(Z_1^n)}{\log(n)}$$

when the limit exists. The conditional entropy of a random vector Z_1^n given another random variable Θ is denoted $H(Z_1^n|\Theta)$. The log-blocklength normalized conditional entropy rate is defined to be

$$\mathfrak{H}(\mathfrak{Z} \mid \Theta) = \lim_{n\to\infty} \frac{H(Z_1^n \mid \Theta)}{\log(n)}$$

when the limit exists. The log-blocklength normalized information rate is defined to be

$$\mathfrak{I}(\mathfrak{Z}; \Theta) = \lim_{n\to\infty} \frac{I(Z_1^n; \Theta)}{\log(n)}$$

when the limit exists.

For a sequence of source codes ϕ_n, the average codeword lengths are

$$C_{\phi,n} = \sum_{\mathcal{Z}^n} p_{Z_1^n}(z_1, \ldots, z_n)\ell(z_1, \ldots, z_n),$$

where $\ell(\cdot)$ is the length of the codeword assigned to the source realization z_1^n and $\mathcal{Z}$ is the source alphabet. Shannon's fixed-to-variable source coding theorem [14] establishes that there exists a sequence of source codes that satisfy the following inequalities for all n:

$$H(Z_1^n) \leq C_{\phi,n} \leq H(Z_1^n) + 1.$$

Dividing through by $\log(n)$ yields

$$\frac{H(Z_1^n)}{\log(n)} \leq \frac{C_{\phi,n}}{\log(n)} \leq \frac{H(Z_1^n) + 1}{\log(n)}.$$

Taking the limit of large blocklength, we see that there is a sequence of source codes that achieve $\mathfrak{H}(\mathfrak{Z})$.

B. Log-Blocklength Normalized Redundancy Measures

We define the redundancy of a source code, $r_{\phi,n}$, as the excess average codeword length that is required over the minimum $H(Z_1^n)$:

$$r_{\phi,n} = C_{\phi,n} - H(Z_1^n).$$

Finally, we define the log-blocklength normalized redundancy of a sequence of source codes to be

$$\begin{aligned} \lim_{n\to\infty} \frac{r_{\phi,n}}{\log(n)} &= \lim_{n\to\infty} \frac{C_{\phi,n} - H(Z_1^n)}{\log(n)} \\ &= \lim_{n\to\infty} \frac{C_{\phi,n}}{\log(n)} - \mathfrak{H}(\mathfrak{Z}) \\ &\triangleq \mathfrak{C}_\phi - \mathfrak{H}(\mathfrak{Z}). \end{aligned}$$

By the manipulations of the source coding theorem that we had made previously, we know that there is a sequence of codes with $\mathfrak{C} = 0$. The code used to develop the upper bound in the source coding theorem, however, requires that $p_{Z_1^n}(z_1, \ldots, z_n)$ is known. Now we define performance measures for source coding for a class of source distributions, rather than just a single source distribution.

The definitions that we make parallel those of [22]. Suppose that the source distribution is chosen from a class that is parameterized by $\Theta \in \mathcal{T}$. For each θ, there is a conditional distribution

$$p(z_1^n \mid \theta) = \Pr\left[Z_1^n = z_1^n \mid \Theta = \theta\right].$$

The parameter θ is fixed but unknown, when generating the source realization. Moreover, there may be a distribution on this parameter, $p_\Theta(\theta)$. Let Φ_n be the set of all uniquely decipherable codes of length n. Then, the average log-blocklength normalized redundancy of a code $\phi \in \Phi_n$ for the class of sources described by $p_\Theta(\theta)$ is

$$\mathcal{L}_{\phi,n}(p_\Theta) = \int_{\mathcal{T}} \frac{r_{\phi,n}}{\log(n)} p_\Theta(\theta)\, d\theta.$$

The minimum nth order average log-blocklength normalized redundancy is

$$\mathcal{L}_n^*(p_\Theta) = \inf_{\psi \in \Phi_n} \mathcal{L}_{\psi,n}(p_\Theta).$$

Finally, the minimum average log-blocklength normalized redundancy is

$$\mathcal{L}^*(p_\Theta) = \lim_{n\to\infty} \mathcal{L}_n^*(p_\Theta)$$

If $\mathcal{L}^*(p_\Theta) = 0$, then a sequence of codes that achieve the limit are called *weighted log-blocklength normalized universal*. Now let T be the set of all probability distributions defined on the alphabet $\mathcal{T}$. Then the nth order maximin log-blocklength normalized redundancy of T is

$$\mathcal{L}_n^- = \sup_{q_\Theta \in T} \mathcal{L}_n^*(q_\Theta).$$

If it exists, then the maximin log-blocklength normalized redundancy is

$$\mathcal{L}^- = \lim_{n\to\infty} \mathcal{L}_n^-.$$

If $\mathcal{L}^- = 0$, then a sequence of codes that achieve the limit are called *maximin log-blocklength normalized universal*. The nth order minimax log-blocklength normalized redundancy of $\mathcal{T}$ is

$$\mathcal{L}_n^+ = \inf_{\phi\in\Phi_n} \sup_{\theta \in \mathcal{T}} \frac{r_{\phi,n}(\theta)}{\log(n)}$$

and the minimax log-blocklength normalized redundancy of $\mathcal{T}$ is

$$\mathcal{L}^+ = \lim_{n\to\infty} \mathcal{L}_n^+.$$

If $\mathcal{L}^+ = 0$, then a sequence of codes that achieve the limit are called *minimax log-blocklength normalized universal*.

C. Redundancy-Capacity Theorems

The senses of universality that we define obey an ordering relation.

Theorem 3: The several log-normalized redundancy quantities satisfy

$$\mathcal{L}_n^+ \geq \mathcal{L}_n^- \geq \mathcal{L}_n^*(p_\Theta)$$

and

$$\mathcal{L}^+ \geq \mathcal{L}^- \geq \mathcal{L}^*(p_\Theta).$$

Proof: Minor modification of [22, Theorem 1]. ■

Armed with definitions and relations among several notions of log-blocklength normalized universality, we now study when it is possible to achieve universality. We give a theorem that gives a necessary and sufficient condition on the existence of weighted log-blocklength normalized universal codes.

Theorem 4: The minimum nth order average log-blocklength normalized redundancy is bounded as

$$\frac{I(Z_1^n;\Theta)}{\log(n)} \leq \mathcal{L}_n^*(p_\Theta) \leq \frac{I(Z_1^n;\Theta)}{\log(n)} + \frac{1}{\log(n)}.$$

A necessary and sufficient condition for the existence of weighted log-blocklength normalized universal codes is that

$$\mathcal{L}^*(p_\Theta) = \lim_{n\to\infty} \mathcal{L}_n^*(p_\Theta) = \lim_{n\to\infty} \frac{I(Z_1^n;\Theta)}{\log(n)} = \mathfrak{I}(Z;\Theta) = 0.$$

Proof: Minor modification of [22, Theorem 2]. ■

Theorem 4 can be extended to conditions for minimax and maximin log-blocklength normalized universality and also can be strengthened by suitable modification of theorems in [23], [24].

D. Class of Memoryless Multisets

Consider the class of memoryless, binary multisets. The parameter θ is the Bernoulli trial parameter. Now suppose that there is a distribution over the parameter space $q_\Theta(\theta) \in T$ that is uniform over $[0,1]$. This gives a mixed source where all type classes are equiprobable, as seen now. Let the realizations of $\{X_i\}_{i=1}^n$ be expressed as $z \in \{0,\ldots,n\}$, the number of ones.

$$p(\{X_i\}_{i=1}^n = z) = \int_0^1 q_\Theta(\theta) \Pr[\{X_i\}_{i=1}^n = z \mid \Theta = \theta]\, d\theta,$$

where $q_\Theta(\theta)$ simply equals one over the range of integration, and $\Pr[\{X_i\}_{i=1}^n = z | \Theta = \theta]$ is given by a binomial distribution:

$$\Pr[\{X_i\}_{i=1}^n = z \mid \Theta = \theta] = \binom{n}{z} \theta^z (1-\theta)^{n-z}.$$

So,

$$p(\{X_i\}_{i=1}^n = z) = \int_0^1 \binom{n}{z} \theta^z (1-\theta)^{n-z}\, d\theta = \binom{n}{z} B(z+1, n-z+1) = \frac{1}{1+n},$$

where the beta function $B(\cdot,\cdot)$ is used. As we can see, the result is that the type classes are equiprobable. Since the types are equiprobable, the entropy $H(\{X_i\}_{i=1}^n)$ is just the enumeration of the types:

$$H(\{X_i\}_{i=1}^n) = \log \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} = \log(n+1).$$

The entropy conditioned on $\Theta = \theta$ is given by (1), in which we can specify the $o(1)$ term as in [16] with given constants a_k, and so the conditional entropy is

$$\begin{aligned} H(\{X_i\}_{i=1}^n \mid \Theta) &= \int H(\{X_i\}_{i=1}^n \mid \theta) q_\Theta(\theta)\, d\theta \\ &= \int \left[\tfrac{1}{2} \log\left(2\pi e n \theta(1-\theta)\right) + \sum_{k\geq 1} a_k n^{-k} \right] q_\Theta(\theta)\, d\theta \\ &= \tfrac{1}{2}\log(n) + \sum_{k\geq 1} n^{-k} \int a_k q_\Theta(\theta)\, d\theta \\ &\quad + \int \tfrac{1}{2} \log\left(2\pi e \theta(1-\theta)\right) q_\Theta(\theta)\, d\theta. \end{aligned}$$

Now the mutual information is $I(\{X_i\}_{i=1}^n;\Theta) = H(\{X_i\}_{i=1}^n) - H(\{X_i\}_{i=1}^n \mid \Theta)$. The log-blocklength normalized information rate is given by

$$\begin{aligned}\mathfrak{I} &= \lim_{n\to\infty} \frac{I(\{X_i\}_{i=1}^n;\Theta)}{\log(n)} \\ &= \lim_{n\to\infty} \frac{\log(n+1) - \frac{1}{2}\log(n) - o(\log(n))}{\log(n)} \\ &= \lim_{n\to\infty} \frac{\log(n+1) - \frac{1}{2}\log(n)}{\log(n)} \\ &= \lim_{n\to\infty} -\frac{1}{2} + \frac{\log(n+1)}{\log(n)} = \frac{1}{2}.\end{aligned}$$

Since this is greater than 0, we have shown that the weakest form of universality is not possible, by Theorem 4. Thus by Theorem 3, stronger forms of universality are not possible either. Since the class of binary memoryless sets is a subset of more general source classes such as memoryless; Markov; and stationary, ergodic, universal source coding over these source classes is not possible either.

We can calculate the weighted redundancy for classes of memoryless sources with larger alphabet sizes. Using the $p_\Theta(\theta)$ that yields equiprobable multisets and (2), we are interested in

$$\lim_{n\to\infty} \frac{\log\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} - \frac{|\mathcal{X}|-1}{2}\log(Kn)}{\log(n)} = \frac{|\mathcal{X}|-1}{2}.$$

As we can see, this redundancy grows without bound as the alphabet size increases. Perhaps unsurprisingly, this redundancy expression is reminiscent of the unnormalized redundancy expression for i.i.d. sequences [25]:

$$\frac{|\mathcal{X}|-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma^{|\mathcal{X}|}\left(\frac{1}{2}\right)}{\Gamma\left(\frac{|\mathcal{X}|}{2}\right)} + o_{|\mathcal{X}|}(1).$$

We see that the richness of a class of sources as sequences is the same as the richness of that class as multisets.

Let us also comment that in the deterministic case of individual multisets, rather than the probabilistic classes of sources that we have been considering, the same non-achievability result applies. This follows from the arguments summarized in [26].

V. Conclusion

Motivated by applications in distributed inference and elsewhere, we have investigated universal lossless source coding of multisets. We have established that a rate of no more than $n+o(n)$ bits per multiset letter is needed to represent multisets from members of classes that meet Kieffer's condition. We also saw, however, that when normalized by the logarithm of the multiset size, the redundancy cannot be made zero. An intuitive and reasonably rigorous interpretation is that there are no sufficient statistics for the class parameter that are of negligible (in the log-blocklength normalized sense) rate.

Acknowledgment

We thank Alon Orlitsky, especially for discussions leading to the results in Section III. We also thank Sanjoy K. Mitter and Kush R. Varshney.

References

[1] A. L. Buchsbaum, G. S. Fowler, and R. Giancarlo, "Improving table compression with combinatorial optimization," *J. ACM*, vol. 50, no. 6, pp. 825–851, Nov. 2003.

[2] B. R. Judd and N. S. Sutherland, "The information content of nonsequential messages," *Inf. Control*, vol. 2, no. 4, pp. 315–332, Dec. 1959.

[3] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons, Inc., 1965.

[4] T. S. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.

[5] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[6] A. T. Ihler, J. W. Fisher, III, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE J. Select. Areas Commun.*, vol. 23, no. 4, pp. 809–819, Apr. 2005.

[7] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, Sept. 1962.

[8] A. T. Ihler, J. W. Fisher, III, and A. S. Willsky, "Using sample-based representations under communications constraints," M.I.T. Lab. Inform. Decision Syst., Tech. Rep. 2601, Dec. 2004.

[9] L. R. Varshney and V. K. Goyal, "Toward a source coding theory for sets," in *Proc. Data Compression Conf.*, Mar. 2006, pp. 13–22.

[10] L. R. Varshney and V. K. Goyal, "Ordered and disordered source coding," in *Proc. Inform. Theory & Applications Inaugural Workshop*, Feb. 2006.

[11] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "Limit results on pattern entropy," *IEEE Trans. Inform. Theory*, vol. 52, no. 7, pp. 2954–2964, July 2006.

[12] Z. Ji, K. Tanaka, and S. Kitamura, "Block permutation coding of images using cosine transform," *IEEE Trans. Commun.*, vol. 43, no. 11, pp. 2833–2846, Nov. 1995.

[13] D. Basu, "On statistics independent of a complete sufficient statistic," *Sankhyā*, ser. A, vol. 15, no. 4, pp. 377–380, 1955.

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

[15] A. de Moivre, *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 3rd ed. London: A. Millar, 1756.

[16] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 1072–1081, May 1999.

[17] B. M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Probl. Inf. Transm.*, vol. 2, no. 2, pp. 1–7, Apr.-June 1966.

[18] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1469–1481, July 2004.

[19] G. I. Shamir, "Patterns of i.i.d. sequences and their entropy," *IEEE Trans. Inform. Theory*, submitted.

[20] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 6, pp. 674–682, Nov. 1978.

[21] D.-K. He and E.-H. Yang, "The universality of grammar-based codes for sources with countably infinite alphabets," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3753–3765, Nov. 2005.

[22] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.

[23] R. G. Gallager, "Source coding with side information and universal coding," M.I.T. Lab. Inform. Decision Syst., Tech. Rep. 937, Sept. 1979.

[24] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 714–722, May 1995.

[25] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2686–2707, Nov. 2004.

[26] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.