Universal Noiseless Compression for Noisy Data

Gil I. Shamir, Tjalling J. Tjalkens, and Frans M. J. Willems

***Abstract*— We study universal compression for discrete data sequences that were corrupted by noise. We show that while, as expected, there exist many cases in which the entropy of these sequences increases from that of the original data, somewhat surprisingly and counter-intuitively, universal coding redundancy of such sequences cannot increase compared to the original data. We derive conditions that guarantee that this redundancy does not decrease asymptotically (in first order) from the original sequence redundancy in the stationary memoryless case. We then provide bounds on the redundancy for coding finite length (large) noisy blocks generated by stationary memoryless sources and corrupted by some specific memoryless channels. Finally, we propose a sequential probability estimation method that can be used to compress binary data corrupted by some noisy channel. While there is much benefit in using this method in compressing short blocks of noise corrupted data, the new method is more general and allows sequential compression of binary sequences for which the probability of a bit is known to be limited within any given interval (not necessarily between 0 and 1). Additionally, this method has many different applications, including, prediction, sequential channel estimation, and others.**

I. Introduction

Lossless (noiseless) universal compression of data generated by some unknown source in a known class has been studied extensively (see, e.g., [2], [8], [12]). The underlying assumption in all the literature on noiseless compression is that the "clean" data produced by the source is to be compressed. However, suppose that before we compress a data sequence which was generated by an unknown source, it has been corrupted by a known noisy channel. Now, we attempt to use a lossless code to compress this noisy sequence. Consider, for example, a novice typist typing some text document. The typist makes typical errors that can be modeled by some form of a typewriter noisy channel, and the data compressed is not the original text, but its noisy version typed by the typist. The parameters of the source generating the data are no longer the ones actually governing the compressed data. How would the noisy data compress compared to the clean data? In this paper, we consider this question. We study universal coding redundancy for such noisy data, the entropy of the noisy data, as well as a method for *sequentially* coding binary memoryless sequences after they have been corrupted by noise. The sequential method proposed turns out to be much more general than for the specific application discussed, and can be used for various different problems.

[1]G. I. Shamir is with ECE Department, University of Utah, Salt Lake City, UT 84112, U.S.A., e-mail: gshamir@ece.utah.edu. T. J. Tjalkens and F. M. J. Willems are with the Eindhoven University of Technology, Electrical Engineering Department, 5600 MB Eindhoven, The Netherlands, e-mails: T.J.Tjalkens@tue.nl, F.M.J.Willems@tue.nl. The work of the first author was partially supported by NSF Grant CCF-0347969.

A different motivation for the present work can be found in recent work on denoising of discrete sequences generated by a discrete (unknown) source and then corrupted by a discrete output (known) channel [17]. Followup work in [10] studied conditions in which the "say-what-you-see" scheme was optimal for denoising a received noisy sequence. In this case, one will use the received data to represent the actual data. This noisy data can now be compressed without prior knowledge of the statistics of the source that generated the original noiseless sequence. Other related work considers prediction of the clean data which is based on noisy observations (see, e.g. [16]).

Basic intuition implies that a noisy sequence should be harder to compress. While this is true in many cases, as far as the source entropy is considered, there are somewhat surprisingly cases in which the entropy of the data sequence can actually decrease from the original data sequence, allowing the noisy sequence to compress better than the original sequence. Also somewhat counterintuitive is the fact that in universal coding, the redundancy for coding the noisy sequence cannot increase from the redundancy of coding the noiseless sequence. In fact, it can only decrease. The reason for that behavior is that a *known* channel can only reduce the richness of a source class, rather than increase it. This can be reflected in two effects: 1) reduction in the cardinality of the parameter vector governing the source statistics, and 2) reduction of the range of some parameters of the source.

We first derive a simple theorem that shows that the redundancy of universal coding of the noisy data cannot be greater than that of the original clean data. While we limit the discussion to stationary memoryless (independently and identically distributed - i.i.d.) sources, this theorem is more general. We then show that if the channel is relatively good (in the i.i.d. case), the decrease in redundancy is reflected only in second order performance, and asymptotically redundancy equal to that of the clean data is obtained. We next derive specific bounds for the redundancy of coding i.i.d. (long) finite length sequences corrupted by some specific channels, with focus on symmetric channels. Next, we study the entropy of memoryless sequences corrupted by some channels. We show that for symmetric channels the entropy must increase, but demonstrate that there exist channels for which the noisy source entropy can decrease.

The last part of the paper considers a sequential probability assignment method for binary i.i.d. sequences. Unlike standard probability assignment methods, such as the *Krichevsky-Trofimov* (KT) [8], here we do not limit the parameter to the standard interval $[0, 1]$, but allow the source parameter to be within any interval $[0 \leq \alpha, \beta \leq 1]$. This method is specifically useful for coding short binary sequences corrupted

by a noisy channel, since, as we see, for such sequences the parameter is limited to different intervals than the standard $[0,1]$ one. The gain of this method in compression is limited to short sequences, because the decrease in redundancy is insignificant asymptotically, as long as $\beta-\alpha$ is not extremely small. However, this method can have much more significant benefits in prediction problems, sequential channel estimation, and more.

The outline of the paper is as follows. Section II describes the notation. In Section III, an upper bound is derived. Section IV studies the asymptotics of the redundancy, specifically focusing on conditions on the noisy channel under which the redundancy of coding noisy sequences does not asymptotically decrease from that of clean sequences in the i.i.d. source case. Next, in Section V, bounds on the redundancy are derived for specific channels, demonstrating the decrease from the clean data redundancy. Section VI considers the entropy of the noisy data. Finally, in Section VII, a sequential probability assignment method for binary memoryless sequences with limited parameters is proposed.

II. Notation and System Description

Let $x^n \triangleq (x_1, x_2, \ldots, x_n)$ be a sequence of n symbols over the alphabet $\mathcal{X}$ with cardinality $k = |\mathcal{X}|$. We assume that $k = o(n)$, i.e., it is either fixed or can grow with n but slower than n. Without loss of generality, where needed, we will assume that $\mathcal{X} = \{1, 2, \ldots, k\}$. The sequence x^n is generated by an i.i.d. source, whose parameter vector is given by $\tilde{\boldsymbol{\theta}} \triangleq (\theta_1, \theta_2, \ldots, \theta_{k-1})^T$, where T denotes the transpose operator. The value $k-1$ gives the total number of parameters governing the source (the cardinality of the parameter vector). Specifically, θ_i is the probability of X taking the ith letter in the alphabet. For convenience, although the probability of the last alphabet letter can be given by subtracting the sum of all the parameters in $\tilde{\boldsymbol{\theta}}$ from 1, we also define θ_k as the kth component of the parameter vector. We use $\boldsymbol{\theta} \triangleq (\theta_1, \theta_2, \ldots, \theta_k)^T$ to define the complete probability vector of an i.i.d. source over the alphabet $\mathcal{X}$. The class of all possible sources $\boldsymbol{\theta}$ will be denoted by Λ, and is known *a-priori*.

In general, small letters will denote deterministic values, and capital letters will be used to denote random variables. Boldface letters will denote vectors, whose components will be denoted by their indices in the vector. For convenience, parameter vectors of all types will be assumed to be column vectors. Deterministic matrices will be denoted by bold capital letters. Estimators will be denoted by the *hat* sign. For example, the *Maximum Likelihood* (ML) estimator of $\boldsymbol{\theta}$ out of X^n will be denoted by $\hat{\boldsymbol{\theta}}$.

The probability of some sequence x^n generated by $\boldsymbol{\theta}$ is given by $P_\theta(x^n) \triangleq \Pr(x^n \mid \boldsymbol{\Theta} = \boldsymbol{\theta})$. The average nth-order (per-symbol) redundancy obtained by a code that assigns length function $L(\cdot)$ for $\boldsymbol{\theta}$ is

$$R_n(L, \boldsymbol{\theta}) \triangleq \frac{1}{n}\left\{E_\theta L[X^n] - H_\theta[X^n]\right\}, \tag{1}$$

where E_θ denotes expectation with respect to (w.r.t.) $\boldsymbol{\theta}$, and $H_\theta[X^n]$ is the nth order block entropy of the source (which equals $nH_\theta[X]$ for i.i.d. sources). With entropy coding techniques, assigning a universal probability $Q(x^n)$ is identical to designing a universal code for coding x^n where, up to usually negligible integer length constraints, the negative logarithm to the base of 2 of the assigned probability is considered as the code length.

The average (per-symbol) *minimax* redundancy of some class Λ is defined as the one obtained by the best code for the worst source $\boldsymbol{\theta}$ in the class Λ,

$$R_n^+(\Lambda) \triangleq \min_L \sup_{\boldsymbol{\theta}\in\Lambda} R_n(L, \boldsymbol{\theta}). \tag{2}$$

The *maximin* redundancy of Λ is the one obtained for the worst *prior* over the best code for a given prior in the class,

$$R_n^-(\Lambda) \triangleq \sup_{w_{n,\theta}} \min_L \int_\Lambda w_{n,\theta}(d\boldsymbol{\theta})\, R_n(L, \boldsymbol{\theta}), \tag{3}$$

where $w_{n,\theta}(\cdot)$ is an nth order prior on Λ. It was demonstrated in [2] that

$$R_n^+(\Lambda) \geq R_n^-(\Lambda). \tag{4}$$

It was later shown [5], [13], that under some mild conditions, the two redundancy measures are, in fact, equal. Davisson also showed in [2] that the maximin redundancy is bounded by

$$\sup_{w_{n,\theta}} \frac{1}{n} I(\boldsymbol{\Theta}; X^n) \leq R_n^-(\Lambda) < \sup_{w_{n,\theta}} \frac{1}{n} I(\boldsymbol{\Theta}; X^n) + \frac{1}{n} \tag{5}$$

where $I(\boldsymbol{\Theta}; X^n)$ is the mutual information between the parameter vector $\boldsymbol{\Theta}$ and the observed data sequence X^n induced by the joint distribution $w_{n,\theta}(\boldsymbol{\Theta}) \cdot P_\theta(X^n)$. The supremum on both sides of the equation is the nth order capacity of the channel between the parameter space Λ and the observed sequence space $\mathcal{X}^n$. For $n \to \infty$, this is the capacity of this channel. The statement in (5) can thus be referred to as the *weak version of the redundancy-capacity theorem*.

Another notion of redundancy considers the *redundancy for most sources* [12]. In [9], this notion of redundancy was tied to the random coding capacity of the channel between $\boldsymbol{\Theta}$ and X^n. Specifically, the following was shown: Let $\mu_{n,\theta}(\cdot)$ denote the *uniform prior* in Λ. Let $\boldsymbol{\omega}$ be a set of points in Λ, and cover Λ with sets $\boldsymbol{\omega}$, such that each $\boldsymbol{\theta} \in \Lambda$ belongs to a distinct set $\boldsymbol{\omega}_\theta$. Randomly select one set $\boldsymbol{\Omega}$ of points in Λ, where the distribution over the sets is dictated by $\mu_{n,\theta}(\cdot)$. Now, select a point $\boldsymbol{\Theta} \in \boldsymbol{\Omega}$ from $\boldsymbol{\Omega}$ under a uniform prior, and let $\boldsymbol{\Theta}$ generate X^n. Let $\hat{\boldsymbol{\Theta}}_\Omega \in \boldsymbol{\Omega}$ be an estimator of $\boldsymbol{\Theta}$ from X^n. Define P_e as

$$\begin{aligned} P_e &\triangleq \Pr\left[\hat{\boldsymbol{\Theta}}_\Omega \neq \boldsymbol{\Theta}\right] \\ &= \sum_{\theta\in\Lambda} \mu_{n,\theta}(\boldsymbol{\Omega} = \boldsymbol{\omega}_\theta)\, \mu_{n,\theta}(\Theta = \theta | \boldsymbol{\Omega} = \boldsymbol{\omega}_\theta)\, P_\theta\left(\hat{\boldsymbol{\Theta}}_\Omega \neq \boldsymbol{\theta}\right). \end{aligned} \tag{6}$$

Now, let M be the maximal cardinality of all possible sets $\boldsymbol{\omega}$ for which $P_e \to 0$ as $n \to \infty$, then

$$R_n(L, \boldsymbol{\theta}) \geq (1-\varepsilon)\frac{\log M}{n} \tag{7}$$

for every code $L(\cdot)$ and almost every $\boldsymbol{\theta} \in \Lambda$ except for a set $\mathcal{B}$ for which $\mu_{n,\theta}(\mathcal{B}) \to 0$, i.e., only for a diminishing set $\mathcal{B}$ can any code achieve smaller redundancy than in (7). This result is referred to as the *random coding strong version of the redundancy-capacity theorem*.

Unlike in standard universal compression, in the setting in this paper, the sequence x^n is corrupted by a discrete input discrete output memoryless channel given by the conditional probability mass function $\Pr(Y|X)$, and is observed and compressed as the sequence y^n over the alphabet $\mathcal{Y}$ that can differ from $\mathcal{X}$. The cardinality of the output alphabet is $m = |\mathcal{Y}| = o(n)$, that may also differ from the cardinality of the input alphabet k. Again, without loss of generality, where needed, we will assume that $\mathcal{Y} = \{1, 2, \ldots, m\}$. The channel $\Pr(Y|X)$ is given by a transition matrix

$$\Pr(Y|X) \triangleq \mathbf{P} \triangleq [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k] \tag{8}$$
$$\triangleq \begin{bmatrix} \Pr(Y=1|X=1) & \ldots & \Pr(Y=1|X=k) \\ \Pr(Y=2|X=1) & \ldots & \Pr(Y=2|X=k) \\ \vdots & \ddots & \vdots \\ \Pr(Y=m|X=1) & \ldots & \Pr(Y=m|X=k) \end{bmatrix}.$$

The columns of $\mathbf{P}$ are denoted by $\mathbf{p}_i$. Let $\rho \triangleq \text{Rank}(\mathbf{P})$ be the rank of $\mathbf{P}$. To keep consistent with the notation convention, the components of $\mathbf{P}$ will be compacted into a vector $\boldsymbol{\alpha}$. Since we consider only the case in which the channel is *known*, the parameter $\boldsymbol{\alpha}$ will always be deterministic.

III. A Maximin Upper Bound

Since the channel is deterministic and known, one can now consider a new parameter vector $\boldsymbol{\psi} \triangleq (\psi_1, \psi_2, \ldots, \psi_m)^T$, which is a *deterministic* function of $\boldsymbol{\theta}$ given $\boldsymbol{\alpha}$, given by

$$\boldsymbol{\psi} = \mathbf{P} \cdot \boldsymbol{\theta}. \tag{9}$$

Since one component of $\boldsymbol{\psi}$ is constrained by the other $m-1$, we can define, for every j, $1 \le j \le m$, $\tilde{\boldsymbol{\psi}}^{(j)} \triangleq (\psi_1, \psi_2, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_m)^T$ as the set of $m-1$ components of $\boldsymbol{\psi}$ excluding the jth one. The parameter $\boldsymbol{\psi}$ governs the sequence Y^n, and can be seen to be a point in a *new* parameter space Ξ.

Consider, for example, the binary case for both $\mathcal{X}$ and $\mathcal{Y}$, where a *Binary Symmetric Channel* (BSC) with crossover $\alpha \le 0.5$ separates between the two sequences, i.e.,

$$\mathbf{P}_{\text{BSC}} = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}. \tag{10}$$

Then, $\boldsymbol{\theta} = (\theta, 1-\theta)^T$, where θ is a probability of X taking one of the bit values, and $\boldsymbol{\psi} = (\psi, 1-\psi)^T$, where ψ is the probability of Y taking one of the bit values. Then, the relation $\psi = (1-\alpha)\theta + \alpha(1-\theta) = \theta + \alpha - 2\alpha\theta$ holds. It is easy to see that if $\theta \in [0,1]$, then we must have $\psi \in [\alpha, 1-\alpha]$. Hence, for $\alpha > 0$, the parameter spaces Λ and Ξ differ. In fact, the parameter space of the noisy sequence shrinks from that of the noiseless sequence, hinting to reduction in coding redundancy. Note that if α is not known *a-priori*, then we are back at $\Xi = \Lambda$, because α can take any value. A more general example is a *Binary Channel* (BC) with crossovers α and β,

$$\mathbf{P}_{\text{BC}} = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix}. \tag{11}$$

Here, the parameter governing Y^n satisfies $\psi \triangleq \psi_1 \in [\min\{1-\alpha, \beta\}, \max\{1-\alpha, \beta\}]$.

Since Y^n is now governed by the parameter $\boldsymbol{\psi}$, it can best be compressed in the average to $H_\psi(Y^n|\boldsymbol{\alpha})$ bits, where the knowledge of the channel parameters can aid in compression. Therefore, the redundancy of a code that assigns length function $L(\cdot)$ to sequences now governed by $\boldsymbol{\psi}$ is given by

$$R_n(L, \boldsymbol{\psi}|\boldsymbol{\alpha}) \triangleq \frac{1}{n}\left\{E_\psi L[Y^n|\boldsymbol{\alpha}] - H_\psi[Y^n|\boldsymbol{\alpha}]\right\}. \tag{12}$$

We can also define the minimax and maximin redundancies $R_n^+(\Xi|\boldsymbol{\alpha})$ and $R_n^-(\Xi|\boldsymbol{\alpha})$, respectively, w.r.t. the class Ξ conditioned on the channel parameter vector $\boldsymbol{\alpha}$, using equations similar equations to (2) and (3), respectively, i.e.,

$$R_n^+[\Xi|\boldsymbol{\alpha}] \triangleq \min_L \sup_{\boldsymbol{\psi}\in\Xi} R_n(L, \boldsymbol{\psi}|\boldsymbol{\alpha}), \tag{13}$$

$$R_n^-[\Xi|\boldsymbol{\alpha}] \triangleq \sup_{w_{n,\psi}} \min_L \int_\Xi w_{n,\psi}(d\boldsymbol{\psi})\, R_n(L, \boldsymbol{\psi}|\boldsymbol{\alpha}), \tag{14}$$

where $w_{n,\psi}$ is a prior on the space Ξ.

We are ready to state a theorem.

Theorem 1:

$$R_n^-[\Xi|\boldsymbol{\alpha}] < R_n^-(\Lambda) + \frac{1}{n}. \tag{15}$$

Corollary 1: Let $L^*(\cdot|\boldsymbol{\alpha})$ be the length function of the code that achieves the maximin redundancy for noisy sequences Y^n conditioned on the known channel $\boldsymbol{\alpha}$, then,

$$R_n(L^*, \boldsymbol{\psi}|\boldsymbol{\alpha}) < R_n^-(\Lambda) + \frac{1}{n}. \tag{16}$$

Proof of Theorem 1: The key to the proof of Theorem 1 is the combination of the weak version of the redundancy-capacity theorem and the data processing inequality. Given the channel $\boldsymbol{\alpha}$, the random vectors $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}$, X^n, and Y^n form a Markov chain $\boldsymbol{\Psi} \to \boldsymbol{\Theta} \to X^n \to Y^n$, since $\boldsymbol{\Psi}$ is a deterministic function of $\boldsymbol{\Theta}$ given $\boldsymbol{\alpha}$, X^n depends on $\boldsymbol{\Theta}$ and given $\boldsymbol{\Theta}$ is independent of $\boldsymbol{\Psi}$, and Y^n given X^n is independent of $\boldsymbol{\Theta}$. Furthermore, we also have a Markov chain $\boldsymbol{\Theta} \to \boldsymbol{\Psi} \to Y^n$ given $\boldsymbol{\alpha}$ since $\boldsymbol{\Psi}$ is a deterministic function of $\boldsymbol{\Theta}$, and Y^n is independent of $\boldsymbol{\Theta}$ given $\boldsymbol{\Psi}$.

Now, let $w'_{n,\psi}(\cdot)$ be the prior on Ξ that achieves the capacity of the channel between $\boldsymbol{\Psi}$ and Y^n given $\boldsymbol{\alpha}$. This prior induces a prior $w'_{n,\theta}(\cdot)$ on Λ. (Note that $w'_{n,\theta}(\cdot)$ may not be unique if $\boldsymbol{\Psi}$ is not an invertible function of $\boldsymbol{\Theta}$.) Also, denote the capacity achieving prior for the channel between $\boldsymbol{\Theta}$ and X^n by $w^*_{n,\theta}(\cdot)$. Using the upper bound of the weak version of the

redundancy-capacity theorem in (5), we have

$$\begin{aligned} nR_n^- [\Xi|\boldsymbol{\alpha}] &< I_{w'_{n,\psi}} (\boldsymbol{\Psi}; Y^n|\boldsymbol{\alpha}) + 1 \\ &\overset{(a)}{=} I_{w'_{n,\theta}} (\boldsymbol{\Theta}; Y^n|\boldsymbol{\alpha}) + 1 \\ &\overset{(b)}{=} I_{w'_{n,\theta}} (\boldsymbol{\Theta}; X^n|\boldsymbol{\alpha}) - I_{w'_{n,\theta}} (\boldsymbol{\Theta}; X^n|Y^n, \boldsymbol{\alpha}) + 1 \\ &\overset{(c)}{\leq} I_{w^*_{n,\theta}} (\boldsymbol{\Theta}; X^n) + 1 \\ &\overset{(d)}{\leq} nR_n^- (\Lambda) + 1. \end{aligned} \tag{17}$$

Equality (a) follows from both $\boldsymbol{\Psi} \to \boldsymbol{\Theta} \to Y^n$ and $\boldsymbol{\Theta} \to \boldsymbol{\Psi} \to Y^n$ given $\boldsymbol{\alpha}$, (b) follows from the data processing over $\boldsymbol{\Theta} \to X^n \to Y^n$. Inequality (c) follows from the non-negativity of mutual information, from bounding the mutual information by that induced by the capacity achieving prior, and from the independence of the mutual information of the channel $\boldsymbol{\alpha}$. Finally, (d) follows from the left inequality in (5). Normalizing both sides of (17) by n concludes the proof of Theorem 1.

Corollary 1 follows since

$$\begin{aligned} R_n^- [\Xi|\boldsymbol{\alpha}] &= \sup_{w_{n,\psi}} \min_{L} R_n (L, w_{n,\psi}|\boldsymbol{\alpha}) \\ &= \sup_{w_{n,\psi}} R_n (L^*, w_{n,\psi}|\boldsymbol{\alpha}) \\ &\geq R_n (L^*, \boldsymbol{\psi}|\boldsymbol{\alpha}) \end{aligned} \tag{18}$$

where $R_n (L, w_{n,\psi}|\boldsymbol{\alpha})$ is the mixture over $R_n (L, \boldsymbol{\psi}|\boldsymbol{\alpha})$ induced by prior $w_{n,\psi}$. ∎

While we focus on the i.i.d. case in this paper, Theorem 1 and its proof are more general, and apply to other parametric families of sources, with Λ and Ξ properly defined. For example, if $\boldsymbol{\theta}$ describes parameters of a Markov source of some known order, the noisy version Y^n of X^n can be modeled by a *Hidden Markov Model* (HMM), where the parameters of the state sequence are given from $\boldsymbol{\theta}$ and the conditional probability of observing Y given X is given by the channel model $\boldsymbol{\alpha}$, and is *not* a component in the parameter vector $\boldsymbol{\psi}$, since the channel is assumed to be known.

Theorem 1 shows that the redundancy of compressing the noisy data cannot increase compared to that obtained when compressing the noiseless data. The mutual information $I_{w'_{n,\theta}} (\boldsymbol{\Theta}; X^n|Y^n, \boldsymbol{\alpha})$ obtained in step (b) of (17) constitutes the major decrease in redundancy when coding the noisy data. This indicates that the worse the channel is the greater the decrease in redundancy is. If the conditional entropy $H\left(X^N|Y^N, \boldsymbol{\alpha}\right)$ is small, we can infer X^n from observing Y^n, indicating a good channel. In such a case, $I_{w'_{n,\theta}} (\boldsymbol{\Theta}; X^n|Y^n, \boldsymbol{\alpha})$ is also small, and the decrease in redundancy is small as well. However, if the channel is bad, $H\left(X^N|Y^N, \boldsymbol{\alpha}\right)$ increases and with it $I_{w'_{n,\theta}} (\boldsymbol{\Theta}; X^n|Y^n, \boldsymbol{\alpha})$, thus decreasing the redundancy.

While perhaps counterintuitive at first, as shown in its proof, Theorem 1 follows directly from the fact that the noisy channel performs data processing on the coded sequence. Hence, it can only reduce the richness of the resulting parameter space. This can either reduce the total number of parameters, i.e., the cardinality of the parameter vector, or the ranges of parameters. If $m < k$, the number of parameters decreases. However, again, counter-intuitively, even if $m \geq k$, the redundancy can only decrease. The reason is again, the knowledge of the channel parameters. Consider, for example, a *Binary Erasure Channel* (BEC), for which

$$\mathbf{P}_{\text{BEC}} = \begin{bmatrix} 1-\alpha & 0 \\ \alpha & \alpha \\ 0 & 1-\alpha \end{bmatrix}. \tag{19}$$

The output alphabet is larger than the input alphabet. However, since the channel is known, the probability of the erasure symbol is known *a-priori* to be α independently of $\boldsymbol{\theta}$. The only remaining parameter is the probability of one of the other symbols, which is now limited to $[0, 1-\alpha]$ instead of $[0, 1]$ at the input of the channel. So, in all, while the additional symbol in the output does not increase the cardinality of $\boldsymbol{\psi}$, the reduction in the range of the other parameter decreases the overall redundancy.

Another example is a poor speller (or typewriter) channel. Assume that this speller would randomly pick 'c' or 'k' uniformly, wherever one of these letters should appear. Assuming X^n is i.i.d., the probability assigned to 'c' or 'k' by the parameter $\boldsymbol{\psi}$ will satisfy $\psi_c = \psi_k = (\theta_c + \theta_k)/2$. Hence, instead of two parameters, we only have one parameter now, since knowing $\theta_c + \theta_k$ is sufficient for knowing both ψ_c and ψ_k. The situation does not change even if 'c' and 'k' are not uniformly distributed, but still randomly picked with a known ratio. Unlike the BEC example, here the decrease in redundancy is caused by a loss of a parameter rather than shrinkage of its range.

The situation described in the above example implies that the matrix $\mathbf{P}$ is not a rank k matrix. The noise causes loss of parameters due to the linear dependence of rows of $\mathbf{P}$. This is also the reason that there is no increase in parameter vector cardinality if $m > k$, because we then must still have $\text{Rank}(\mathbf{P}) \leq k$ regardless of m.

IV. Asymptotic Redundancy

In Section III, the redundancy was upper bounded by that of the noiseless parameter space. This section focuses on conditions in which the decrease in redundancy for noisy sequences from the noiseless case is asymptotically negligible, and focuses on lower bounds on the redundancy in such cases. The section is partitioned into two parts. First, we discuss asymptotic bounds on the redundancy, showing conditions in which the noisy redundancy decreases negligibly. In the second part, we give examples of some channels and demonstrate how the theorems can be used for these channels.

A. Redundancy Bounds

In [14], it was shown that for an arbitrarily small $\varepsilon > 0$,

$$R_n^- (\Lambda) \geq \frac{k-1}{2n} \log \frac{n^{1-\varepsilon}}{k-1}. \tag{20}$$

Note that the logarithm in [14] takes argument $n^{1-\varepsilon}/k$. However, the proof method, in fact, results in $k-1$ in the denominator, where the -1 term is absorbed in the low order terms. The difference is insignificant in the asymptotic regime for $n \to \infty$, as well as when k grows with n, but is more significant, otherwise, for short sequences and small k. It was also shown that a similar bound

$$R_n\left(L, \boldsymbol{\theta}\right) \geq \frac{k-1}{2n} \log \frac{n^{1-\varepsilon}}{k-1} \tag{21}$$

holds for every code $L(\cdot)$ and almost every $\boldsymbol{\theta} \in \Lambda$ except for a set $\mathcal{B}$ of sources for which $\mu_{n,\theta}\left(\mathcal{B}\right) = o(1)$, where $\mu_{n,\theta}(\cdot)$ is the uniform prior over Λ. Upper bounds on the redundancy of the same form, where k is allowed to grow with n, were obtained in several works (see, e.g., [11], [14]), and showed that there exists a code with length function $L^*(\cdot)$ for which

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \left(1+\varepsilon\right) \frac{k-1}{2n} \log \frac{n}{k-1} \tag{22}$$

for every $\boldsymbol{\theta} \in \Lambda$. If the KT probability estimates $Q_{KT}\left(x^n\right)$, described in Section VII, are used, redundancy of

$$R_n\left(Q_{KT}, \boldsymbol{\theta}\right) \leq \frac{k-1}{2n} \log \frac{n}{k} + \frac{7k \log e}{12n} + \frac{k^2 \log e}{4n^2} - O\left(\frac{1}{n}\right) \tag{23}$$

is achieved for every $\boldsymbol{\theta} \in \Lambda$ (see, e.g., [14]). Specifically, for $k=2$, it was shown in [18] that

$$R_n\left(Q_{KT,k=2}, \boldsymbol{\theta}\right) \leq \frac{1}{2n} \log n + \frac{1}{n}. \tag{24}$$

The upper bounds of (22)-(23) hold for every $\boldsymbol{\theta} \in \Lambda$. Hence, they also hold for the maximin redundancy $R_n^-\left(\Lambda\right)$. By Theorem 1, they also hold for $R_n^-\left(\Xi|\boldsymbol{\alpha}\right)$, and thus by Corollary 1, there exists a code (the one achieving the minimum in (14)) for which the bounds also hold for every $\boldsymbol{\psi} \in \Xi$. Specifically, if $m \leq k$, one can use the KT estimates on Y^n to obtain the redundancy of (23) w.r.t. m. Thus, if $m < k$, there is an obvious decrease in redundancy.

While Theorem 1 obtains an upper bound on the redundancy for coding the noisy data, it does not completely characterize it. One would expect that a necessary condition for the redundancy not to decrease asymptotically is for $\text{Rank}(\mathbf{P}) = k$ if k is fixed, and $\text{Rank}(\mathbf{P}) \to k$ for large k. However, it turns out that this is *not* a sufficient condition even for very large n. While this condition guarantees that the cardinality of Ξ remains of the same order as that of Λ, a second condition is necessary to guarantee that the range of each parameter in the new noisy parameter space does not significantly decrease. We continue by showing when the redundancy asymptotically decreases negligibly.

Define

$$\mathbf{A} \stackrel{\triangle}{=} \mathbf{P}^T\mathbf{P}. \tag{25}$$

A symmetric real (or Hermitian) k-by-k matrix $\mathbf{M}$ is said to be *Positive Semi Definite* (PSD) (see, e.g., [6]) if for all nonzero (real) column vectors $\mathbf{v} \in \mathbb{R}^k$

$$\mathbf{v}^T\mathbf{M}\mathbf{v} \geq 0. \tag{26}$$

The matrix $\mathbf{M}$ is *Positive Definite* (PD) if a strict inequality holds.

Theorem 2: Fix $\varepsilon' > \varepsilon > 0$, and let $n \to \infty$. Then, if there exists a positive $\lambda > 1/n^{\varepsilon/2}$ such that the matrix $\mathbf{A} - \lambda\mathbf{I}$ is PSD, where $\mathbf{I}$ is the identity matrix, then

$$R_n^-\left(\Xi|\boldsymbol{\alpha}\right) \geq \frac{k-1}{2n} \log \frac{n^{1-\varepsilon'}}{k-1}, \tag{27}$$

and

$$R_n\left(L, \boldsymbol{\psi}|\boldsymbol{\alpha}\right) \geq \frac{k-1}{2n} \log \frac{n^{1-\varepsilon'}}{k-1} \tag{28}$$

for every code $L(\cdot)$ and almost every $\boldsymbol{\psi} \in \Xi$ except for a set $\mathcal{B}$ of sources for which $\mu_{n,\psi}\left(\mathcal{B}\right) = o(1)$, where $\mu_{n,\psi}(\cdot)$ is the uniform prior over Ξ.

Theorem 3: Fix $\varepsilon' > \varepsilon > 0$, let $n \to \infty$, and let λ denote the smallest eigenvalue of $\mathbf{A}$. Then, if $\lambda > 1/n^{\varepsilon/2}$ both (27) and (28) hold.

Theorems 2 and 3 are very related. Specifically, the condition on the positivity of the minimal eigenvalue of $\mathbf{A}$ implies that $\mathbf{A}$ is PD and that $\text{Rank}(\mathbf{P}) = k$ (see, e.g., [6]). This implies that if $\rho = \text{Rank}(\mathbf{P}) < k$, the condition does not hold. This is expected because if $\rho < k$ the cardinality of the parameter space Ξ must be smaller. Again, however, this is a necessary condition and not sufficient, because if the cardinality of the parameter space does not decrease but the range of each component of the parameter does, the redundancy may still decrease. The condition of $\lambda > 1/n^{\varepsilon/2}$ covers both cases. Note that it can be loosened to having a larger fraction of ε in the exponent, as long as it is smaller than 1.

As we will see in the examples at the end of this section, it will sometimes be simpler to consider a different matrix from $\mathbf{A}$. Let

$$\mathbf{Q}' \stackrel{\triangle}{=} \left[\mathbf{p}_1 - \mathbf{p}_k, \mathbf{p}_2 - \mathbf{p}_k, \ldots, \mathbf{p}_{k-1} - \mathbf{p}_k\right] \tag{29}$$

be a matrix whose $k-1$ columns are the differences between the first $k-1$ column vectors of $\mathbf{P}$ and the last column of $\mathbf{P}$. Define

$$\mathbf{B}' \stackrel{\triangle}{=} \mathbf{Q}'^T\mathbf{Q}'. \tag{30}$$

Now, let $\mathbf{Q}_j; 1 \leq j \leq m$, be the matrix $\mathbf{Q}'$ with the jth row removed. Define

$$\mathbf{B}_j \stackrel{\triangle}{=} \mathbf{Q}_j^T\mathbf{Q}_j. \tag{31}$$

Theorem 4: a) Fix $\varepsilon' > \varepsilon > 0$, and let $n \to \infty$. Then, if there exists a positive $\lambda > 1/n^{\varepsilon/2}$ such that the matrix $\mathbf{B}' - \lambda\mathbf{I}$ is PSD then both (27) and (28) hold. Alternatively, if the smallest eigenvalue λ of $\mathbf{B}'$ satisfies $\lambda > 1/n^{\varepsilon/2}$, then both (27) and (28) hold.

b) If there exists j, $1 \leq j \leq m$, such that either the first condition or the second condition of part a are satisfied for $\mathbf{B}_j$, then both (27) and (28) hold.

Theorem 4 allows us to remove the dependence on the constraint component of the parameter vector in Λ (part a). It also allows us to remove dependence on a constraint component of the parameter vector in Ξ (part b). The importance of Theorem 4, however, is in the implementation of the condition.

As the examples at the end of the section show, it will result in much simpler matrices, specifically the $\mathbf{B}_j$ matrices, that will provide the limiting value of λ immediately in many cases. We continue by outlining the proofs of the three theorems. The complete proofs are found in [15].

Proof of Theorem 2: The proof relies on the proof of the bound (21) on the redundancy for most sources in [14]. The idea for proving (27) is to take one set $\mathbf{\Omega}_\theta$ of points $\boldsymbol{\theta} \in \Lambda$ that are placed at centers of $k-1$ dimensional spheres with radius $1/\sqrt{n}^{1-\varepsilon}$ packed in Λ. The number of spheres is bounded as in (21). Each $\boldsymbol{\theta} \in \mathbf{\Omega}_\theta$ can be transformed to $\boldsymbol{\psi} \in \Xi$ with (9) forming a set $\mathbf{\Omega}_\psi \subset \Xi$. Then, it is shown that the condition of Theorem 2 guarantees that the error probability P_e based on estimation of $\mathbf{\Psi}$ by observing Y^n diminishes. Using the weak version of the redundancy-capacity theorem, the proof of (27) is complete. Using the fact that two spheres of the same volume in Λ transform to two objects of the same volume in Ξ, this proof also carries over to most sources, implying (28). ■

Proof of Theorem 3: We show that the condition on the eigenvalues of $\mathbf{A}$ leads to the same bound on the square distance between two points in Ξ as the same distance obtained with the PSD condition in Theorem 2. This results in the same diminishing error probability on estimation of $\mathbf{\Psi}$ by observing Y^n. The proof is concluded similarly to that of Theorem 2. ■

Proof of Theorem 4: Since a probability parameter in Λ is defined by its first $k-1$ components, we can consider the Euclidean distance in the $k-1$ dimensional space, including only the first $k-1$ components of the parameter. To prove part a of Theorem 4, one can use the distance of the m components of a point $\boldsymbol{\psi} \in \Xi$ from another point $\boldsymbol{\psi}' \in \Xi$ induced by the respective points $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in Λ. The m components of the distance can be obtained from the first $k-1$ components of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. If the distance is above some threshold, there will be diminishing probability of estimating $\boldsymbol{\psi}'$ by observing Y^n, if $\boldsymbol{\psi}$ is the point corresponding to $\boldsymbol{\theta}$ which generated X^n. This leads to part a of the theorem in the same manner as Theorems 2-3 are obtained, but this time w.r.t. the first $k-1$ free components of $\boldsymbol{\theta}$. Part b is proved in a similar way by only considering $m-1$ free parameters in the space Ξ. ■

B. Examples

1) The Binary (Symmetric) Channel: The transition matrix for a general binary channel is given in (11). The matrix for the special BSC case is given in (10). By definition, it is shown that

$$\mathbf{B}_{\text{BC}} = \left[(1-\alpha-\beta)^2\right] \tag{32}$$

where the index j is omitted because an identical matrix is obtained for both $j=1$ and $j=2$. For the special BSC case, we have

$$\mathbf{B}_{\text{BSC}} = \left[(1-2\alpha)^2\right]. \tag{33}$$

For $\lambda \leq (1-\alpha-\beta)^2$ for the general BC case and $\lambda \leq (1-2\alpha)^2$ for the specific BSC case, $\mathbf{B}-\lambda\mathbf{I}$ is PSD (and thus also $\mathbf{A}-\lambda\mathbf{I}$ is also PSD). Specifically, it can be shown that the smallest eigenvalue of $\mathbf{A}_{\text{BSC}}$ equals $(1-2\alpha)^2$. Hence, in order for the condition of Theorems 4 (and of Theorem 3) to hold, we must have $v^2 \stackrel{\triangle}{=} (1-\alpha-\beta)^2 > 1/n^{\varepsilon/2}$ for the general BC and $(1-2\alpha)^2 > 1/n^{\varepsilon/2}$ for the BSC, where v is the range of the single parameter ψ in Ξ.

If

$$0.75v^2n^{\varepsilon} - \ln(n+1) > 5, \tag{34}$$

then $(n+1)e^{-0.75(1-\alpha-\beta)^2 n^{\varepsilon}} < e^{-5} < 0.007$. With tighter bounding than used in the proofs of Theorems 2-4, one can show [15] that in the binary case, the expression on the left hand side of (34) bounds the exponent of the error probability. The decrease in redundancy can then be considered negligible w.r.t. the redundancy if (34) holds. For $n = e^{70}$ and $\varepsilon = 0.1$, this is true for $\alpha \leq 0.349$ in the BSC case and for $\alpha + \beta \leq 0.698$ in the BC case. Thus for the above values of n and ε, the decrease in redundancy is less than 0.7% of the redundancy as long as BSC crossover is not greater than 0.349. If the crossover is greater, so is the decrease in redundancy. In a similar manner, the redundancy will decrease by less than 0.7% for $n = 10^{20}$ and the same ε for a BSC with $\alpha \leq 0.087$.

2) The Z Channel: The Z Channel is a special case of the BC, with transition matrix

$$\mathbf{P}_{\text{Z}} = \begin{bmatrix} 1 & \alpha \\ 0 & 1-\alpha \end{bmatrix}. \tag{35}$$

By definition, we obtain,

$$\mathbf{B}_{\text{Z}} = \left[(1-\alpha)^2\right]. \tag{36}$$

Thus for the conditions of Theorems 2-4 to hold, we must have $(1-\alpha)^2 > 1/n^{\varepsilon/2}$. For more accurate results we can use (34) with $v = 1-\alpha$ to guarantee decrease of less than 0.7% in redundancy.

3) The Binary Symmetric Binary Erasure Channel: The *Binary Symmetric Binary Erasure Channel* (BSC-BEC) is given by the transition matrix

$$\mathbf{P}_{\text{BSC-BEC}} = \begin{bmatrix} 1-\alpha-\beta & \beta \\ \alpha & \alpha \\ \beta & 1-\alpha-\beta \end{bmatrix}. \tag{37}$$

Here, $m > k$. The BSC in (10) and BEC in (19) are special cases of (37).

There are two different matrices $\mathbf{B}_j$ that can be obtained for this channel,

$$\begin{aligned} \mathbf{B}_{1,\text{BSC-BEC}} &= \left[(1-\alpha-2\beta)^2\right] \\ \mathbf{B}_{2,\text{BSC-BEC}} &= \left[2\,(1-\alpha-2\beta)^2\right]. \end{aligned} \tag{38}$$

Matrix $\mathbf{B}_{1,\text{BSC-BEC}}$ is obtained from $\mathbf{Q}'$ by removing the first or third row, and matrix $\mathbf{B}_{2,\text{BSC-BEC}}$ is obtained from $\mathbf{Q}'$ by removing the second row. By Theorem 4, either one can be used. Hence, for this channel if $2\,(1-\alpha-2\beta)^2 > 1/n^{\varepsilon/2}$, negligible asymptotic redundancy decrease is obtained. Note that the factor 2 can also be obtained for the BC if one uses the matrix $\mathbf{B}'$ and part a of Theorem 4. Again, we can use

(34) with $v = (1 - \alpha - 2\beta)$ to guarantee decrease of less than 0.7% in redundancy. (Note that (34) must be used w.r.t. one component of the parameter, and thus the 2 factor does not apply.)

4) The Symmetric Channel: Finally, we consider a symmetric channel, whose transition matrix is given by

$$\mathbf{P}_{\text{sym}} = \begin{bmatrix} 1-\alpha & \frac{\alpha}{k-1} & \dots & \frac{\alpha}{k-1} \\ \frac{\alpha}{k-1} & 1-\alpha & \dots & \frac{\alpha}{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\alpha}{k-1} & \frac{\alpha}{k-1} & \dots & 1-\alpha \end{bmatrix}. \tag{39}$$

The simplest matrix $\mathbf{B}$ that can be obtained for this channel is

$$\mathbf{B}_{\text{sym}} = \left(1 - \alpha - \frac{\alpha}{k-1}\right)^2 \mathbf{I}. \tag{40}$$

Using Theorem 4, it is immediate to see that if $v^2 = \left(1 - \alpha - \frac{\alpha}{k-1}\right)^2 > 1/n^{\varepsilon/2}$, the condition holds, and the asymptotic redundancy decreases negligibly. The examples presented illustrate the simplicity in using Theorem 4 instead of Theorems 2 or 3.

V. Redundancies with Specific Channels

In this section we consider symmetric channels, with transition matrix $\mathbf{P}$ as defined in (39). For such channels we upper and lower bound the maximin and most sources redundancies, and show that the lower bounds are achievable. The bounds are reduced from those of the noiseless case, demonstrating the decrease in redundancy obtained to the noisy case for finite length (although large) sequences. We then bound the redundancy for a general BC.

Theorem 5: Fix $\varepsilon' > \varepsilon > 0$, and let $n \to \infty$. Then, for the symmetric channel defined by the matrix in (39) the following hold:

a)

$$R_n^- \left(\Xi | \boldsymbol{\alpha}_{\text{sym}}\right) \geq \frac{k-1}{2n} \log \frac{\left(1 - \alpha - \frac{\alpha}{k-1}\right)^2 n^{1-\varepsilon'}}{k-1}. \tag{41}$$

b)

$$R_n \left(L, \boldsymbol{\psi} | \boldsymbol{\alpha}_{\text{sym}}\right) \geq \frac{k-1}{2n} \log \frac{\left(1 - \alpha - \frac{\alpha}{k-1}\right)^2 n^{1-\varepsilon'}}{k-1} \tag{42}$$

for every code $L(\cdot)$ and almost every $\boldsymbol{\psi} \in \Xi$ except for a set $\mathcal{B}$ of sources for which $\mu_{n,\psi}(\mathcal{B}) = o(1)$, where $\mu_{n,\psi}(\cdot)$ is the uniform prior over Ξ.

c) There exists a code with length function $L^*(\cdot)$, for which

$$R_n \left(L^*, \boldsymbol{\psi} | \boldsymbol{\alpha}_{\text{sym}}\right) \leq \frac{k-1}{2n} \log \frac{\left(1 - \alpha - \frac{\alpha}{k-1}\right)^2 n^{1+\varepsilon'}}{k-1} \tag{43}$$

for every $\boldsymbol{\psi} \in \Xi$.

Theorem 5 shows that the redundancy decreases by $-\log\left(1 - \alpha - \alpha/(k-1)\right)$ for each parameter. This is simply because of the decrease in the range that such a parameter can take. Namely, the matrix $\mathbf{P}_{\text{sym}}$ dictates that each component of the parameter vector is restricted to the interval $\psi_i \in [\alpha/(k-1), 1-\alpha]$. The decrease in the range leads to a decrease in the number of packed spheres in the parameter space proportional to the square of the decrease in range of each parameter. The same decrease is also reflected in nonuniform griding that leads to a similar upper bound. Note that the same square decrease is reflected in the matrices $\mathbf{A}$, $\mathbf{B}'$, or $\mathbf{B}_j$, and their eigenvalues. While Theorem 5 applies only to the symmetric channel defined in (39), it may be possible to extend the concept to more general channels, where the decreases in ranges of the parameters are reflected in the decrease in redundancy.

Proof of Theorem 5: To prove the lower bounds, we pack the parameter space Ξ with spheres of radius $1/\sqrt{n}^{1-\varepsilon}$. Sources at the centers of the spheres constitute $\boldsymbol{\Omega}_\psi$, where in order to use the strong random coding version of the redundancy-capacity theorem, the set $\boldsymbol{\Omega}_\psi$ is chosen by randomly shifting one set $\boldsymbol{\omega}_\psi$. All shifts of this set form a covering of Ξ. Unlike Theorems 2-4, here the spheres are packed in Ξ, and not in Λ, but with the same radius as those packed in Λ in the proofs of these theorems. Since the volume of Ξ reduces from that of Λ, a smaller number of spheres can be packed. This leads to the gain over the noiseless case. In order to compute the volume of Ξ, we need to consider displacement components from the minimum value each component can take, which is $\alpha/(k-1)$.

To prove the upper bound, a nonuniform grid, as used in [14] is used to quantize the components of an estimate of $\boldsymbol{\psi}$ from Y^n. The quantized estimate is the ML estimate as long as the empirical distribution of Y^n satisfies the constraints on $\boldsymbol{\psi}$ imposed by Ξ. Otherwise, it is a quantized vector in Ξ for which the representation of a sequence Y^n is the shortest possible among all quantized vectors possible. The quantization cost is shown to be negligible in the *average*, and the representation cost is bounded with a bound like that in (43) (with $\varepsilon < \varepsilon'$). The bound is achieved, again, by considering a displacement vector representing the displacement of each parameter from its minimum possible value. This concludes the proof of Theorem 5. The complete details of the proof of can be found in [15]. ■

In a similar manner to Theorem 5, we can show that in the binary case

$$R_n \left(L, \boldsymbol{\psi} | \boldsymbol{\alpha}_{\text{binary}}\right) \geq \frac{1}{2n} \log\left(v^2 n^{1-\varepsilon'}\right) \tag{44}$$

for every code $L(\cdot)$ and almost every $\boldsymbol{\psi} \in \Xi$ and also for the maximin redundancy, where v is the length of the interval of the single parameter determining $\boldsymbol{\psi}$. Similarly, there exists a code $L^*(\cdot)$, for which

$$R_n \left(L^*, \boldsymbol{\psi} | \boldsymbol{\alpha}_{\text{binary}}\right) \leq \frac{1}{2n} \log\left(v^2 n^{1+\varepsilon'}\right) \tag{45}$$

for every $\boldsymbol{\psi} \in \Xi$. The last two bounds apply to a general binary input channels, as the BC in (11), with the special cases of a BSC and the Z channel. It also applies to the BSC-BEC. In [3], the precise expression for the *individual sequence* minimax

redundancy for the binary case in which the parameter is limited to the interval $[0 \leq \alpha, \beta \leq 1]$ was computed. For an arbitrary small ε' and $n \to \infty$, the expressions in (44)-(45) are slightly smaller than the expression in [3] because they express the asymptotic behavior of the average case redundancy. The average redundancy is upper bounded by the individual sequence minimax redundancy.

For finite sequences the decrease in redundancy is rather significant. For example, if we have a BSC with $\alpha = 0.5 - 0.5n^{-\gamma}$, $\gamma < 0.5$, the redundancy becomes

$$R_n\left(L, \boldsymbol{\psi}|\boldsymbol{\alpha}\right) \approx \frac{(1-2\gamma)}{2n} \log\left(n^{1\pm\varepsilon''}\right) \tag{46}$$

where ε'' absorbs the lower order terms, and $\pm$ denotes addition for upper bound and subtraction for lower. This demonstrates a significant gain in redundancy for short sequences with such values of α, where the redundancy becomes a fraction of the noiseless redundancy.

For example, consider $n = 10^{20}$ bits coded over a BSC with crossover 0.45. For this "short" sequence, $0.5 \log n \approx 33.2$. The redundancy decreases here by $-\log 0.1 \approx 3.32$ bits. Hence, the noisy channel results in approximately 10% decrease in redundancy, even for sequences that are very long. For very short sequences, say, $n = 100$, this redundancy decrease is even not negligible w.r.t. the actual source entropy.

VI. Noisy Source Entropy

Intuitively one would expect the source entropy of the noisy data to increase from that of the noiseless data. However, this is not always the case. If the channel is symmetric, the entropy will increase, but otherwise it may increase or decrease depending on the channel parameters. Consider, for example, the Z channel described in (35). If $\alpha = 1$, then $\boldsymbol{\psi} = (1, 0)^T$ regardless of $\boldsymbol{\theta}$ and then $H_\psi\left(Y^n\right) = 0$. This is the extreme case. Figure 1 shows the noiseless and noisy entropies for different values of α as function of the noiseless parameter θ. It shows that the behavior of the noisy entropy depends on the specific $\boldsymbol{\theta}$ and the parameter α. For alphabets $\mathcal{X} = \mathcal{Y} = \{1, 0\}$, if $\boldsymbol{\theta} = (\theta, 1-\theta)^T$, where θ is the probability of 1, we have $\boldsymbol{\psi} = (\psi, 1-\psi)^T$, where

$$\psi = \theta + \alpha(1-\theta) = (1-\alpha)\theta + \alpha. \tag{47}$$

In general, as we observe in Figure 1, for larger θ, the noisy entropy decreases from the noiseless one, whereas for smaller θ it decreases. Specifically, since $\psi \geq \theta$, then for $\theta > 0.5$, we observe a decrease in entropy since the binary entropy is monotonically decreasing in this region.

We define a general *symmetric channel* as a channel with a square transition matrix $\mathbf{P}$, where each row of $\mathbf{P}$ is a permutation of the first row, and each column of $\mathbf{P}$ is a permutation of the first column. We state the following theorem:

Theorem 6: Let $\mathbf{P}$ define a symmetric channel. Then,

$$H_\psi\left[Y^n|\boldsymbol{\alpha}\right] = H_\psi\left[Y^n\right] \geq H_\theta\left[X^n\right]. \tag{48}$$

Theorem 6 states that the entropy cannot decrease for a symmetric channel. The proof is based on the definition of a symmetric channel and Jensen's inequality and is included in the complete version of the paper [15]. (Note that once the parameter $\boldsymbol{\psi}$ is known, conditioning in $\boldsymbol{\alpha}$ is unnecessary.)

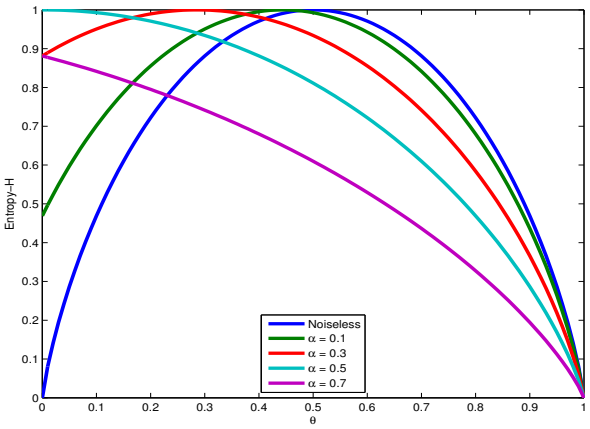

Fig. 1: Entropy of random variable Y at the output of a Z Channel with different crossover parameters as function of the noiseless parameter θ generating X^n.

VII. Sequential Probability Assignment

Consider binary input/output alphabets $\mathcal{X} = \mathcal{Y} = \{1, 0\}$. Let θ denote the probability of a noiseless 1, and ψ be the probability of a noisy 1, i.e., $\theta = \Pr(X = 1)$, and $\psi = \Pr(Y = 1)$. A BC with positive crossovers transforms θ into ψ that is restricted to an interval $[0 \leq \alpha, \beta \leq 1]$. In compressing Y^n, one can use the fact that ψ may be limited to a known interval shorter than $[0, 1]$ to benefit in redundancy performance (as shown in Section V). This section focuses on deriving a sequential probability assignment method that takes advantage of this shrinkage of the parameter space from Λ to Ξ. More generally, one can consider the material in this section as the derivation of a universal method for compressing binary sequences, for which the probability of 1 is not necessarily in $[0, 1]$, but can be in a smaller interval restricted by $[\alpha, \beta]$.

Using the *Dirichlet*-$1/2$ prior, which is used to derive the KT estimates [8], one can assign the following *universal mixture* probability to a sequence y^n:

$$Q\left(y^n\right) = \int_\alpha^\beta \frac{1}{C(\alpha,\beta)\sqrt{\psi(1-\psi)}} \psi^{n_y(1)}(1-\psi)^{n_y(0)} d\psi, \tag{49}$$

where $n_y(b)$ is the number of times bit b occurs in y^n, and the constant $C(\alpha, \beta)$ is given by

$$\begin{aligned} C &\triangleq C(\alpha,\beta) = \int_\alpha^\beta \frac{1}{\sqrt{x(1-x)}} dx \\ &= 2\left(\sin^{-1}\sqrt{\beta} - \sin^{-1}\sqrt{\alpha}\right). \end{aligned} \tag{50}$$

The constant C guarantees that the prior integrates to 1 over $[\alpha, \beta]$.

Theorem 7: The probability assigned to y^n in (49) can be computed sequentially by an initialization step

$$Q\left(y^0\right) = 1 \tag{51}$$

and an update step,

$$Q\left(y^{t+1}\right) = Q\left(y^{t}\right)\cdot\frac{n^t(y_{t+1})+0.5}{t+1} + (2y_{t+1}-1)\cdot\frac{\alpha^{n^t(1)+0.5}(1-\alpha)^{n^t(0)+0.5}}{C\cdot(t+1)} - (2y_{t+1}-1)\cdot\frac{\beta^{n^t(1)+0.5}(1-\beta)^{n^t(0)+0.5}}{C\cdot(t+1)} \quad (52)$$

where $n^t(b)$ is the occurrence count of bit b in y^t, and

$$2y_{t+1}-1 = \begin{cases} 1 & \text{if } y_{t+1}=1 \\ -1 & \text{if } y_{t+1}=0 \end{cases}. \quad (53)$$

Note that the KT estimates are a special case of the above sequential assignment with $[\alpha = 0, \beta = 1]$. Specifically, in that case, $C(\alpha,\beta)=\pi$, and (52) reduces to the binary form of the KT estimator,

$$Q\left(y^{t+1}\right) = Q\left(y^{t}\right)\cdot\frac{n^t(y_{t+1})+0.5}{t+1}. \quad (54)$$

Proof: Using (49), we can express the probability of y^t followed by a 1 bit, $Q\left(y^t 1\right)$, as

$$\begin{aligned} Q\left(y^t 1\right) &= \int_\alpha^\beta \frac{1}{C\sqrt{\psi(1-\psi)}}\psi^{n^t(1)+1}(1-\psi)^{n^t(0)}d\psi \\ &= \frac{1}{C}\int_\alpha^\beta \psi^{n^t(1)+0.5}(1-\psi)^{n^t(0)-0.5}d\psi \\ &\overset{(a)}{=} -\frac{1}{C}\frac{\psi^{n^t(1)+0.5}(1-\psi)^{n^t(0)+0.5}}{n^t(0)+0.5}\Bigg|_\alpha^\beta + \frac{n^t(1)+0.5}{C(n^t(0)+0.5)}\cdot \int_\alpha^\beta \psi^{n^t(1)-0.5}(1-\psi)^{n^t(0)+0.5}d\psi \\ &= \frac{1}{C(n^t(0)+0.5)}\cdot \left[\alpha^{n^t(1)+0.5}(1-\alpha)^{n^t(0)+0.5} - \beta^{n^t(1)+0.5}(1-\beta)^{n^t(0)+0.5}\right] + \frac{(n^t(1)+0.5)}{(n^t(0)+0.5)}\cdot Q\left(y^t 0\right). \end{aligned} \quad (55)$$

Equality (a) follows from integration by parts with $f(\psi)=\psi^{n^t(1)+0.5}$ and $g'(\psi)=(1-\psi)^{n^t(0)-0.5}$. Since $Q\left(Y^t\right)$ is a joint probability measure on all possible y^t for every t, we must have

$$Q\left(y^t\right) = Q\left(y^t 0\right) + Q\left(y^t 1\right). \quad (56)$$

Substituting (55) in (56), we obtain

$$Q\left(y^t\right) = \frac{1}{C(n^t(0)+0.5)}\cdot \left[\alpha^{n^t(1)+0.5}(1-\alpha)^{n^t(0)+0.5} - \beta^{n^t(1)+0.5}(1-\beta)^{n^t(0)+0.5}\right] + \frac{(t+1)}{(n^t(0)+0.5)}\cdot Q\left(y^t 0\right), \quad (57)$$

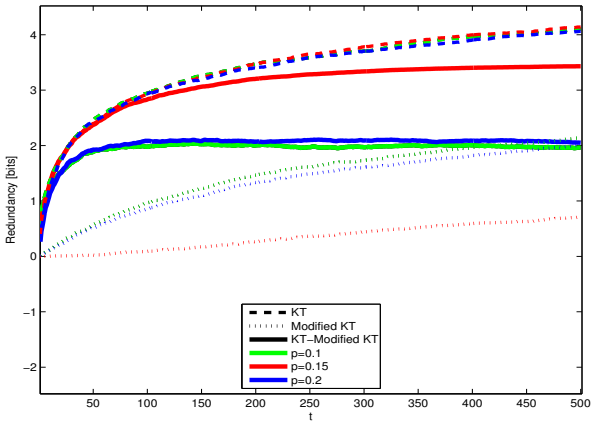

Fig. 2: Redundancy for the KT estimates and the sequential algorithm shown in Theorem 7 averaged over 1000 simulations for $n = 500$ bits where $\theta \in [0.1, 0.2]$. The gain of the proposed method over KT is shown in solid curves.

where the equality follows from $n^t(0)+n^t(1)=t$. Reorganizing (57), expressing $Q\left(y^t 0\right)$ in terms of $Q\left(y^t\right)$ yields (52) for $y_{t+1}=0$. Substituting $Q\left(y^t 0\right)$ in (55), again using $n^t(0)+n^t(1)=t$, yields (52) for $y_{t+1}=1$. The proof of Theorem 7 is concluded. ■

Figure 2 shows redundancy for the sequential probability assignment in (52) compared to that achieved by the standard KT estimates. Specifically, average curves over 1000 simulations are shown for up to 500 bits coded with $\theta \in [0.1, 0.2]$. The algorithm assumes $\alpha = 0.1$ and $\beta = 0.2$. The solid curves express the decrease in redundancy from the standard binary KT estimates with the proposed method. For short blocks, the gain in redundancy recovers the loss of the KT estimates, but for longer blocks it becomes constant. It is, however, not negligible w.r.t. the overall redundancy for the values of n demonstrated.

Figure 2 shows that the new probability estimator, based on a mixture over a limited interval, appears to reduce the redundancy at the center of the interval w.r.t. the KT estimator by about $-\log v$ bits, where $v=\beta-\alpha$. The same behavior is obtained for other intervals with the same value of v. This decrease in redundancy is the one anticipated by the bounds in (44)-(45). However, at the boundaries of the interval ($\psi=\alpha$ and $\psi=\beta$), the redundancy decrease is smaller. This may result from the extraction of part of the boundary neighborhood from the universal mixture. To improve performance at the boundary, a margin can be taken to include α and β inside the mixture interval. However, such a margin will increase v, trading off the gain from including α and β as inner points in the parameter interval. It is possible that a different prior from the Dirichlet-$1/2$ may be better here. However, there is no guarantee that if such a better prior is found it can be used to obtain a low-complexity sequential update procedure such as that described in Theorem 7.

While it reduces the redundancy in coding finite blocks of

data, larger benefits from the sequential estimator proposed in Theorem 7 may be gained in many other applications beyond that of compression of noisy data. It can be used in sequential prediction, where the parameter is limited in range (see, e.g., [7]). Specifically, it can be used in *Context Tree Weighting* (CTW) [18] based prediction, where the data contains memory, and within each memory context different ranges of statistical parameters govern the data. Specifically, recent work [19] used the context tree method for prediction. If the parameter is known to be limited as hypothesized, using the estimator in (52) may gain in terms of prediction error much more significantly than the gain one would expect in compression. Another related application is that of universal investment portfolios.

One potential compression application where larger gains are anticipated with the new method is bit decomposition of data, in which the parameters of a source over some alphabet whose cardinality is greater than 2 are decomposed into the bits that constitute them. Instead of having the probability of each letter as a parameter, the bits of the alphabet symbols are stored in a tree, and the overall probability of each bit constitutes a parameter. If bits are known to have probabilities within a given interval, the overall compression gain can accumulate over the various bit probabilities. Since many states have small occurrences in practice, the gains anticipated may be significant fractions of the actual redundancy.

Another application for which the new estimator is useful is sequential channel estimation with channel decoding. Consider a real time sequential decoder of a convolutional code. The channel is a BSC which occasionally changes abruptly (even at *a-priori* known instances). The crossover at each *segment* is unknown in advance. A Viterbi decoder must estimate the crossover in each segment. If one uses the KT estimates, the initial estimate of the crossover of the first bit in a segment is 0.5, implying that the received data is noise, and leading to potentially catastrophic performance. Usually with a BSC, one can assume that the crossover is less than 0.5. If the maximum possible crossover is β, using the estimator of Theorem 7, the crossover hypothesized for the first bit is

$$Q(Y_1 = 1) = 0.5 - \frac{\sqrt{\beta(1-\beta)}}{2\sin^{-1}\sqrt{\beta}}. \tag{58}$$

Specifically, if $\beta = 0.5$, the expression in (58) results in crossover estimate of 0.18 at the first time unit. For $\beta = 0.25$, it gives 0.087.

The application proposed above can even be extended to include inverting channels. One can consider two different estimators. For one $\psi \in [0, 0.5]$, and for the other $\psi \in [0.5, 1]$. Both can be updated, and the one that yields a greater probability in (52) can be used as the estimate of the channel. Partitioning like those described here (including finer ones for more intervals) can also be used for classification problems.

VIII. Summary and Conclusions

We studied universal lossless compression for noisy data. We showed that while the entropy can either increase or decrease, depending on the known noisy channel, universal compression redundancy can only decrease. We derived conditions on negligible asymptotic decrease in redundancy, and tightly bounded the redundancy for specific channels. For symmetric channels whose transition matrix is given in (39) each parameter cost reduces by the logarithm of $1 - \alpha - \alpha/(k-1)$, and for binary sources the redundancy reduces by $-\log v$, where v is the length of the noisy parameter interval. It was shown that for symmetric channels the entropy of the noisy data must increase. Finally, a sequential probability assignment algorithm was proposed for coding binary sequences for which the probability of 1 is limited within a known interval.

Acknowledgments

The authors thank W. Szpankowski for providing information about the results in [3], and A. Cohen for pointing out the work in [16].

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, second edition, John Wiley & Sons, 2006.

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

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

[4] P. Elias, "Universal codeword sets and representation of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 194-203, Mar. 1975.

[5] R. G. Gallager, "Source coding with side information and universal coding," unpublished manuscript, Sept. 1976.

[6] R. A. Horn, and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985 (reprinted 1999).

[7] S. S. Kozat, and A. C. Singer, "Universal piecewise linear prediction via context trees," submitted for publication.

[8] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199-207, Mar. 1981.

[9] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 714-722, May 1995.

[10] E. Ordentlich, and T. Weissman, "On the optimality of symbol-by-sybmol filtering and denoising," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 19-40, Jan. 2006.

[11] A. Orlitsky, and N. P. Santhanam, "Speaking of infinity," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2215-2230, Oct. 2004.

[12] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629-636, July 1984.

[13] B. Ya. Ryabko, "Coding of a source with unknown but ordered probabilities," *Problems of Information Transmission*, vol. 15, no. 2, pp. 134-138, Oct. 1979.

[14] G. I. Shamir, "On the MDL principle for i.i.d. sources with large alphabets," *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1939-1955, May 2006.

[15] G. I. Shamir, T. J. Tjalkens, and F. M. J. Willems, "Universal noiseless compression for noisy data," in preparations for *IEEE Trans. Inform. Theory*.

[16] T. Weissman, and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 47, no. 6, pp. 2151-2173, Sept. 2001.

[17] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, M. J. Weinberger, "Universal discrete denoising: known channel," *IEEE Trans. Info. Theory*, vol. 45, no. 1, pp. 5-28, Jan. 2005.

[18] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "The Context-Tree weighting method: basic properties," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 653-664, May 1995.

[19] J. Ziv, and N. Merhav, "On context tree prediction of individual sequences," in *Arxiv:cs.IT/0508127*.