COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 1, No. 3, pp. 305-332, September 2001

ON THE REDUNDANCY OF UNIVERSAL LOSSLESS CODING FOR GENERAL PIECEWISE STATIONARY SOURCES*

GIL I. SHAMIR[†] AND DANIEL J. COSTELLO, JR.[‡]

Abstract. The redundancy capacity theorem is used to obtain a lower bound on the achievable redundancy for universal coding of parametric sources with abruptly changing statistics. Unlike the previously known bound for a problem that assumes a fixed number of changes in the statistics, the new bound is general and can be used even if the number of changes increases with the length of the input string. In particular, it is shown that for any uniquely decipherable code, for almost every combination of the statistical parameters governing each segment, and for almost every vector of transition times, the minimum achievable redundancy is composed of $0.5 \log m$ extra code bits for each unknown segmental parameter in each segment and $\log m$ extra code bits for each unknown transition, where m is the average length of a segment of the input string. The same result is true also in the minimax and maximin senses. The new bound confirms the asymptotic optimality of recently proposed low complexity strongly sequential encoders (i.e., encoders that do not utilize knowledge of a prescribed value of the data string length) that were shown to achieve the same performance.

Index Terms: Piecewise stationary source, universal coding, redundancy, capacity, redundancy capacity theorem, minimax and maximin redundancies.

1. Introduction.

The universal lossless coding problem of *Piecewise Stationary Sources* (PSS's), namely, sources with abruptly changing statistics, has a significant practical importance. This results from the fact that data sequences, that are obtained from a large family of practical applications, can be modeled as being emitted from a source in this class. Particularly, this is the case if the sequence to be encoded is a concatenation of strings each drawn from a different stationary source. Such sequences may occur in almost any application area, and may be such as computer data files composed of different data types, images with different regions, audio, video, text, or even the output of transforms performed on stationary data strings, as the Burrows-Wheeler transform [1].

In spite of the importance of the class of PSS's and of the fact that universal coding schemes designed for either stationary sources or for non-stationary sources with slowly varying statistics are not optimal for PSS's, only recently has the universal coding problem of PSS's been given much attention, see [12]-[13], [16]-[18], [20]-[21]. While in [12]-[13] the emphasis was on a lower bound for universal coding of PSS's,

*Received on February 15, 2001, accepted for publication on August 17, 2001. This work was supported by NSF Grant NCR95-22939 and NASA Grant NAG5-8355. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Sorrento, Italy, June 25-30, 2000.

[†] Department of Electrical Engineering, University of Utah, Salt Lake City, UT 84112, U.S.A. The research presented in this paper was performed while Gil I. Shamir was with the Department of Electrical Engineering at the University of Notre Dame, Notre Dame, IN 46556, U.S.A. Email: gshamir@ee.utah.edu

[‡] Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, U.S.A. Email: Daniel.J.Costello.2@nd.edu

the work in [16]-[18] and [20]-[21] was directed to developing coding schemes that achieve good performance with low complexity. The bound in [12]-[13] was derived for a particular case of PSS's, where the number of changes in the statistics is fixed even if the sequence length n grows. A bound for the general case, for which the number of changes can increase with n, was not obtained. However, some of the schemes in [16]-[18] and [20]-[21] were proposed and analyzed for the general class of sources, and not only for the particular case of a fixed number of changes.

In this paper, we derive a lower bound on the redundancy of the general class, where the number of changes is not necessarily fixed, and *stationary segments* may be asymptotically shorter than the order of the complete sequence. The new bound shows that some of the schemes proposed in [16]-[18] are indeed asymptotically optimal, not only when the number of changes is fixed, but also in the general case. Despite the resemblance between the new bound and the bound for a fixed number of changes, the derivation of the new bound is not a straightforward extension of the bound in [12]-[13]. This is because some of the techniques employed to derive the bound in [12]-[13] fail with the change of the asymptotic behavior in the general case. The difficulties in using the techniques of the old bound will become clear throughout this paper.

In the layout of the universal coding problem, we assume that any data sequence, emitted from a PSS, is divided into independent stationary segments between which abrupt changes (or transitions) in the statistics occur. The goal of the coding problem is to find an optimal encoder that can represent such a sequence with a codeword of minimal length, where neither the parameter vectors θ_i that govern the statistics in each stationary segment nor the time instants t_i of the changes between different segments and the number c of these changes are known in advance. In a more compact traditional universal coding setting, we want to design a uniquely decipherable code, with length function $L(\cdot)$, that is required to be as small as possible for any value of the parameter vector ψ in the PSS class Λ, where the code is not allowed to depend on the unknown vector ψ that defines the source in the class. The parameter $\psi \in \Lambda$ is defined for PSS's as the extended vector that contains the number of transitions c, the set of *segmental parameters* θ_i, and the set of transition times t_i, henceforth referred to as the *transition path*. The space Λ that represents the class contains all possible combinations of these parameters for a given sequence length n.

In the general universal coding problem, the performance of any given code for a particular class of sources is judged on the basis of the *redundancy* function $R_n(L, \psi)$, which is defined as the difference between the expected code length of $L(\cdot)$ with respect to (w.r.t.) a given source in the class P_ψ and the nth-order entropy of P_ψ normalized by the length n of the uncoded sequence.

Naturally, the lack of knowledge of the source parameters in universal coding results in some redundancy when coding data emitted by any or almost any unknown source from a known class. To measure the universality of such a class, some notion of this redundancy is used to represent the best possible performance for some worst case, i.e., the redundancy expected from the best code for the worst case. This notion of redundancy thus serves as a lower bound on the worst case redundancy of any code for this class of sources. Two such notions are the *maximin* universality and the *minimax* universality, first defined by Davisson [5]. In the maximin Bayesian approach the parameter ψ is considered random, and the *maximin redundancy* is obtained by the worst distribution that maximizes the minimum *expected* redundancy, i.e., the worst distribution for the best code. The minimax approach considers the parameter to be deterministic, and defines the *minimax redundancy* as the redundancy of the best code for the worst choice of ψ. In [5], Davisson showed that the maximin redundancy is equal to the normalized capacity of the "channel" whose input is the parameter ψ and whose output is the data sequence $x^n \overset{\triangle}{=} (x_1, x_2, \ldots, x_n)$ which is to be coded. Later, Gallager [8] showed that the minimax and maximin redundancies are essentially equivalent. Then, it was shown by Davisson [4] that under certain regularity conditions, the minimax redundancy, and hence also the maximin redundancy and the normalized capacity of the corresponding channel, for parametric stationary sources, for which ψ is a k-dimensional vector, is given by $0.5k\,(\log n)\,/n$, where the base of the logarithmic function is 2.

A third stronger notion of redundancy for "most" sources in a class was later established by Rissanen [14]. Rissanen showed that for the parametric case, the lower bound of $0.5k\,(\log n)\,/n$ is not only the bound in the minimax sense, but also for every $\psi \in \Lambda$, except for a set of values of ψ whose volume (Lebesgue measure) vanishes as n grows. Rissanen's result was generalized by Merhav and Feder [13], who showed that the normalized capacity of the channel between the space Λ and the space of the uncoded data sequence is not only a lower bound on the redundancy in the maximin and minimax sense, but it is also a lower bound on the redundancy for a set of sources whose probability under the capacity achieving *prior* (i.e., distribution of ψ in Λ) goes to 1 as $n \to \infty$. In addition, it was shown that for any arbitrarily chosen prior, except for a set of sources with vanishing probability, the redundancy is lower bounded by the *random coding* capacity, which equals to the normalized logarithm of the maximum number M of randomly chosen points $\psi^1, \ldots, \psi^M$, which form, with high probability, a set of distinguishable sources $P_{\psi^1}, \ldots, P_{\psi^M}$. Sources in a set of sources are distinguishable if the probability of estimating that a sequence x^n generated by ψ^i was generated by ψ^j, where $j \neq i$, goes to zero. This result can be intuitively explained by the fact that $\log M$ bits are required by the encoder to convey to the decoder which of the M distinguishable sources is most likely to have generated the sequence x^n. Particularly, this result can be applied to a uniform prior to extend Rissanen's result. These last results, which are a strong version of the *redundancy-capacity theorem*, are very important because they link the universal coding problem to the well-established theory of channel capacity, when almost all sources in a class are considered.

The universal coding problem of PSS's was first addressed by Merhav [12]. Merhav extended Rissanen's concept [14] of the relation between estimation and universal coding to the class of PSS's. By proposing this extension, Merhav showed that the average per-letter universal coding redundancy over all sequences of n letters, drawn from almost any PSS with a *fixed* $q = c + 1$ number of segments, where the statistics in each segment are governed by a k-dimensional parameter vector, is lower bounded by

$$R_n\left(L, \psi\right) \geq (1 - \varepsilon)\left(\frac{1}{2}kq + q - 1\right)\frac{\log n}{n}, \tag{1.1}$$

where $\varepsilon > 0$ is any positive number that can be arbitrarily small. This bound was presented as a sum of two terms: The first term, henceforth referred to as *parameter redundancy*, corresponds to universality w.r.t. the unknown source parameters within each stationary segment. Since each segment is assumed to be of the same order as n, this term consists of $(0.5 \log n)/n$ bits per symbol for each component of the parameter vector, as was established by Rissanen. The second term, henceforth referred to as *transition redundancy*, corresponds to universality w.r.t. the unknown transition times from one stationary segment to another. This term consists of $(\log n)/n$ bits per symbol for each such transition. The lower bound on the transition redundancy for the simple case of a PSS with a single transition between two known distributions was also derived as an example of the use of the redundancy-capacity theorem in [13], using the random coding capacity.

While Merhav demonstrated in [12] high complexity schemes that can achieve the lower bound, the recent work following [12] on universal coding of PSS's was directed to finding low complexity *sequential* and *strongly sequential* schemes that achieve small redundancy for the PSS class. That is, coding schemes that perform a small number of operations for each data symbol they code, while the coding procedures are independent of future data, and in the latter case also of the *horizon* (or sequence length) n. Consequently, various universal coding schemes for the memoryless subclass, that can be extended to the more general classes of Markov and finite-state sources, were obtained and their redundancies analyzed and compared to the lower bound of (1.1) in [16]-[18], and [20]-[21].

As proposed in [12], the lower bound of (1.1) can be achieved by having the encoder test the code length that is required to code the transitions and each of the segments for each of the possible partitionings of the data sequence into segments. Then, the partitioning that can be represented by the shortest code is used. Its transition times are encoded, and then some optimal universal code for stationary data sequences is used to code each segment separately. However, for sequential coding the method of mixtures was used due to its appealing sequential implementation simplicity and the fact that no universal optimality is lost (see [13]). Using the mixture method, the coding problem is replaced by the universal probability assignment problem, and a universal probability is assigned to any sequence by averaging over all possible values of the parameters and all or a subset of all possible transition paths. Then, arithmetic coding (see [15]) is used to code the universally assigned probability, and therefore, up to integer length constraints, the redundancy can be evaluated w.r.t. the *ideal code length*, i.e., the negative logarithm of the assigned probability.

A high complexity double mixture approach that achieves the bound of (1.1) was demonstrated by Merhav in [12] for the memoryless case. Then, Willems showed in [20] how its complexity can be reduced, although without achieving the bound. Shamir and Merhav demonstrated in [18] how Willems' approach can achieve the bound in a strongly sequential manner with linear per-letter computational complexity (quadratic over the whole sequence). Suboptimal schemes with lower complexity were proposed by Willems and Krom in [21], and by Shamir and Merhav [18]. An additional estimation mechanism based on the observed data was added to the double mixture techniques in [16]-[18] in order to reduce the complexity by mixing over a subset of the possible transition paths that contains a good estimate of the actual true path. This led to a scheme that achieves the lower bound of (1.1) with a fixed number of operations per-letter in [17], and then to the fixed per-letter complexity scheme proposed by Shamir and Costello in [16] that achieves redundancy of

$$R_n(L,\psi) \le (1+\varepsilon)\left(\frac{1}{2}kq + q - 1\right)\frac{\log m}{n} \tag{1.2}$$

for any memoryless PSS with $c = q - 1$ transitions, where $m \triangleq n/q$ is the average segment length, $k = r - 1$ for an alphabet of size r, and $\varepsilon > 0$ can be made arbitrarily small. The same performance was also obtained in [18] with a linear per-letter complexity scheme.

As the lower bound in (1.1), the upper bound in (1.2) can be decomposed into two terms, the parameter and the transition redundancies. Again, the parameter redundancy reflects universality w.r.t. the unknown source parameters, and the transition redundancy reflects universality w.r.t. the unknown transition times. However, unlike the bound in (1.1), the parameter redundancy now consists only of $0.5 \log m$ extra code bits which are required to code each unknown segmental parameter in each of the segments, and the transition redundancy consists of $\log m$ extra code bits for each unknown transition time. While the details of all the schemes described above can be found in the respective references, the question that rises is whether or not the bound in (1.2) can be replaced by a lower bound for most sources in a PSS class $\Lambda_c \subset \Lambda$ of sources with c transitions, where ε is replaced by $-\varepsilon$. In other words, we would like to know if the schemes that achieve the performance of (1.2) are optimal for almost all sources in the class Λ.

It turns out that the answer to this question is yes as long as the order of q is smaller than the order of n, i.e., we allow only subclasses $\Lambda_c \subset \Lambda$ with $c = o(n)$, i.e., as $n \to \infty$, $(c/n) \to 0$, although c may go to infinity. Note that if q is of the order of n, the source is no longer considered a PSS but an arbitrarily varying source. Despite the fact that at first glance the general bound appears to be just an extension of the bound in (1.1), this is not the case. Of course, if c is fixed, the lower bound derived from (1.2) with negative ε reduces to the lower bound of (1.1). However, in the general case, the difficulty of showing that the proposed bound is indeed a lower bound on the redundancy of *almost all* sources in the class lies in the fact that, unlike the case of a fixed c, the subset of sources, for which there exist segments shorter than the order of m or statistical transitions that are too small, is *not negligible* w.r.t. the set of all sources in the class. Therefore, under a uniform prior, the probability of sources with segments or transitions contributing redundancy smaller than the average redundancy in (1.2) per segment or per transition is not vanishing. For such sources, the lower bound on the redundancy is expected to be smaller than the expression in (1.2) with a negative ε. However, it turns out that the probability that the quantity of such segments and transitions is larger than δq for some arbitrarily small $\delta > 0$ vanishes under the uniform prior. Therefore, a lower bound of

$$R_n(L,\psi) \geq (1-\varepsilon)\left(\frac{1}{2}kq + q - 1\right)\frac{\log m}{n} \tag{1.3}$$

can be obtained for most sources in a class Λ_c, where ε can be arbitrarily small, but must be larger than the ε in (1.1) if identical behavior is required w.r.t. the mean segment length. The difference in ε is the cost of generalizing the bound to include the non-vanishing group of sources that contain a small number of segments shorter than the order of m or a small number of very small transitions.

For the same reasons as these discussed above, the techniques used in [12] to derive the bound in (1.1) fail when the number of transitions is allowed to grow with n. The new bound must be derived by treating the subclass of Λ_c that contains sources for which *most* segments are sufficiently long and *most* transitions are sufficiently large, whereas the bound for the case where c is fixed can be derived by treating the subclass of Λ_c that contains all sources for which *all* segments are sufficiently long and *all* transitions sufficiently large.

The problem becomes even more complicated when the number of segments q is much larger than the average segment length, i.e., $q \gg m$, where the notation $q \gg m$ will be used to indicate that for every constant ν, $q > m^{\nu}$. Using the random coding version of the redundancy-capacity theorem, the redundancy for most sources can be lower bounded by the normalized logarithm of the maximum number of distinguishable sources which are randomly selected in a manner that satisfies their prior. A group of sources is distinguishable if the probability of estimating a source with parameter $\hat{\psi}$ from the data sequence x^n, which was generated by a source with parameter $\psi \neq \hat{\psi}$, goes to zero for any source ψ in the set. Since the segments become shorter, the probability of error in estimating the segmental parameters of a particular segment or the transition time between two particular adjacent segments increases. This results in reducing the set of distinguishable sources to a set that contains only sources that differ in more than a single segmental parameter vector or a single transition time, thus increasing ε in (1.3) even more. However, it is shown that even in this case, the increased ε can still be made arbitrarily small.

We begin the next sections with definitions of the model and notation. For the benefit of the reader, the notation is defined to resemble the notation in [12], which treated the bound for a fixed q. For the sake of completeness, some material from [12] and [13] is repeated but in a manner that is best suitable for the presentation of the new results. The lower bound of (1.3) is then shown in two main steps. First, we show that this is the lower bound in the minimax sense, and hence also in the maximin sense and for a set of sources with probability that goes to 1 under the capacity achieving prior. Then, the bound is extended to apply to most sources in Λ_c, for any $c = o(n)$. In each step, the two cases $q \not\gg m$ and $q \gg m$ are treated separately. The redundancy-capacity theorem is used w.r.t. the class Λ_c to obtain the bound in (1.3) for each c. However, it is also shown that the additional redundancy required to represent the number of transitions c, i.e., to determine the subclass Λ_c from the class of PSS's Λ, can be made negligible w.r.t. the lower bound in (1.3), and in fact the lower bound is tight and achievable even if the number of transitions is unknown in advance.

The outline of this paper is as follows. In Section 2, we define the model and the notation. Then, in Section 3, we review the basic results on the maximin and minimax redundancies and on the redundancy-capacity theorem. The derivation of the lower bound is presented in Section 4, that begins with a presentation of a basic regularity condition, which must be assumed about the source family and is satisfied for common parametric families. Finally, the achievability of the lower bound is demonstrated in Section 5.

2. Notation and Definitions. Let $\{p_\theta\}$ be a parametric family of stationary probability mass functions (PMF's) of vectors whose components take on values in a finite alphabet Σ. The parameter θ is assumed to be a k-dimensional parameter vector taking on values in a compact set $\boldsymbol{\Theta} \subset \mathbb{R}^k$. The parameter can be a set of probabilities of a memoryless source over an alphabet of size $k+1$, a set of transition probabilities of a Markov or finite state source, or any other vector that defines a parametric source. For simplicity, let us assume that each component of the parameter is limited to the closed interval $[0, 1]$. (If this is not the case, the parameters can be normalized onto this interval).

A string drawn by the source from time instant i to time instant j $(x_i, x_{i+1}, \ldots, x_j$ $)$, $j \geq i$ will be denoted by x_i^j, and if $i = 1$ by x^j. Let $x^n \overset{\triangle}{=} (x_1, x_2, \ldots, x_j, \ldots, x_n)$ be a string emitted from a PMF whose parameter θ takes on a particular value θ_1 from $j = 1$ to $j = t_1$; then $\theta = \theta_2$ from $j = t_1 + 1$ until $j = t_2$, and so on. Finally, from $j = t_c + 1$ to $j = n$, θ is equal to θ_q, where $q \overset{\triangle}{=} c + 1$. The vectors $\{x_1, \ldots, x_{t_1}\}$, $\{x_{t_1+1}, \ldots, x_{t_2}\}, \cdots, \{x_{t_c+1}, \ldots, x_n\}$ will be referred to as the *stationary segments*, and correspondingly, $\theta_1, \theta_2, \cdots, \theta_q$ will be called the *segmental parameters*. It will be assumed that the different segments are statistically independent. The extended vector $(\theta_1, \theta_2, \ldots, \theta_q)$ will be denoted by $\boldsymbol{\theta}$. The c dimensional vector, representing the c transition time instants $(t_1, t_2, \ldots, t_c)$, will be denoted by $\mathbf{t}$, and referred to as the *transition path.* For convenience, we define $t_0 \triangleq 0$ and $t_q \triangleq n$. The extended vector $\psi \triangleq (\boldsymbol{\theta}, \mathbf{t}) \in \Lambda \triangleq \Lambda_n$ will uniquely define a PSS in the nth order class Λ. The subclass $\Lambda_c \subset \Lambda$, which is a subset of the space $\boldsymbol{\Theta}^q \times (1, 2, \ldots, n-1)^c$ that is restricted to c transition times satisfying $t_i < t_{i+1}$, will contain all PSS's for n-sequences with c transitions. The regime of the asymptotics will be such that the number of segments q must be of smaller order than n, but can grow with n. Consequently, the mean segment length $m \triangleq n/q$ goes to infinity as $n \to \infty$, but at a slower rate, unless q is fixed, in which case $m \to \infty$ at the same rate as n.

The probability of a measurable event F under P_ψ will be denoted by $P_\psi(F)$, and under the segmental PMF p_θ by $p_\theta(F)$. Correspondingly, the probability of x^n for $\psi \in \Lambda_c$ is defined as

$$P_\psi(x^n) = \prod_{i=1}^{q} p_{\theta_i}\left(x_{t_{i-1}+1}, \ldots, x_{t_i}\right), \tag{2.1}$$

where $p_{\theta_i}\left(x_{t_{i-1}+1}, \ldots, x_{t_i}\right)$ is the probability of the string in the ith segment, defined by the k parameters of θ_i. Similarly, $E_\psi\{\cdot\}$ and $E_\theta\{\cdot\}$ will denote the expectations under the respective two PMF's. The per-letter entropy of the ith segment is defined as

$$H_{\theta_i}\left(X_{t_{i-1}+1}^{t_i}\right) \triangleq -\frac{1}{t_i - t_{i-1}} E_{\theta_i} \log p_{\theta_i}\left(X_{t_{i-1}+1}^{t_i}\right), \tag{2.2}$$

and the average per-letter entropy of a PSS $\psi \in \Lambda_c$ is defined as

$$H_\psi(X^n) \triangleq \frac{1}{n} \sum_{i=1}^{q} (t_i - t_{i-1}) H_{\theta_i}\left(X_{t_{i-1}+1}^{t_i}\right), \tag{2.3}$$

where X_j is used to denote the random variable of the jth letter in the random sequence X^n. We will also use the notation $h(\alpha)$ for the binary entropy of α, $0 \le \alpha \le 1$, where

$$h(\alpha) \triangleq -\alpha \log \alpha - (1-\alpha) \log (1-\alpha). \tag{2.4}$$

A *length function* $L(x^n)$ of a uniquely decipherable lossless code is a map from Σ^n to the positive integers that satisfies Kraft's inequality

$$\sum_{x^n \in \Sigma^n} 2^{-L(x^n)} \le 1. \tag{2.5}$$

The nth order *redundancy* of an encoder that generates a length function $L(\cdot)$ for n-sequences governed by $\psi \in \Lambda$ is defined for any class Λ as

$$R_n(L, \psi) \triangleq \frac{1}{n} E_\psi L(X^n) - H_\psi(X^n). \tag{2.6}$$

Similarly, we also define the nth order *pointwise redundancy* as

$$R_n(L, \psi, x^n) \triangleq R(L, \psi, x^n) \triangleq \frac{1}{n}\left[L(x^n) + \log P_\psi(x^n)\right]. \tag{2.7}$$

The Euclidean norm of a generic vector v will be denoted by $||v||$. Consequently, the Euclidean distance between the segmental parameters θ and θ' will be denoted by $||\theta - \theta'||$. A *decision rule* will be defined as a partitioning of Σ^n into a set of decision regions $\Omega \subseteq \Sigma^n$, each corresponding to a different hypothesis. The complement of a generic set Ω will be denoted by overbar $\overline{\Omega}$. Bold-face letters will be used to denote (extended) vectors. We will use the hat sign to denote estimates of the PSS parameters. Correspondingly, $\hat{\psi}$, $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{t}}$, $\hat{\theta}_i$, and $\hat{t}_i$ will be used to denote estimates of ψ, $\boldsymbol{\theta}$, $\mathbf{t}$, θ_i, and t_i, respectively. Capital letters will be used to denote random vectors, and thus $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}$, $\mathbf{T}$, Θ_i, and T_i will denote the respective random vectors.

Using the concepts of the random coding version of the redundancy-capacity theorem, a random set of M PSS's will be chosen. To distinguish from the subscript i, that is used to denote the ith transition or the ith segment, the jth element in the set will be denoted by the superscript j. Correspondingly, the set $\boldsymbol{\Phi} \triangleq \left(\Psi^1, \Psi^2, \ldots, \Psi^M\right)$ will denote a set of M randomly chosen PSS's, that take on the values $\boldsymbol{\varphi} = \left(\psi^1, \psi^2, \ldots, \psi^M\right)$, respectively. The ith segmental parameter random vector Θ_i will be assumed to take on the values of a randomly chosen grid of points $\mathbf{G}_i \triangleq \left(G_i^1, G_i^2, \ldots, G_i^{M_\theta}\right)$, where the component G_i^j is a random vector in the k-dimensional space $\boldsymbol{\Theta}$, and M_θ is the number of grid points. Similarly, the random variable of the ith transition T_i will take on values from the random grid $\boldsymbol{\mathcal{T}}_i \triangleq \left(\mathcal{T}_i^1, \mathcal{T}_i^2, \ldots, \mathcal{T}_i^{M_t}\right)$, where M_t is the number of grid points. Deterministic grids of the ith segmental parameter and the ith transition will be denoted by $\mathbf{g}_i \triangleq \left(g_i^1, g_i^2, \ldots, g_i^{M_\theta}\right)$ and $\boldsymbol{\tau}_i \triangleq \left(\tau_i^1, \tau_i^2, \ldots, \tau_i^{M_t}\right)$, respectively.

3. Redundancy Capacity and Probability of Error. In this section, we review some of the results mainly from [5], [7], [8] and [13] that tie the universal coding redundancy with the capacity of the channel between Λ_c and X^n. We will use these results in the next section to obtain the lower bound. Although we present the results in the context of PSS's, the results presented in this section apply to any *general* class Λ, which is partitioned into disjoint subclasses Λ_c. For the sake of convenience, and essentially without any effect on the results, let us ignore the integer length constraint associated with the length function $L(\cdot)$, and allow any function from Σ^n to the nonnegative reals such that Kraft's inequality is satisfied.

The minimax redundancy of the subclass Λ_c is defined as

$$R_n^+(\Lambda_c) \triangleq \min_L \sup_{\psi \in \Lambda_c} R_n(L, \psi). \tag{3.1}$$

To define the maximin redundancy of Λ_c, let us assign a probability measure (prior) $w_c(\cdot)$ on Λ_c for each subclass Λ_c (where $w_c(\psi) = 0$ if $\psi \notin \Lambda_c$) and let us define the mixture source

$$P_{w_c}(x^n) \triangleq \int_{\Lambda_c} w_c(d\psi) P_\psi(x^n). \tag{3.2}$$

The average redundancy associated with a length function $L(\cdot)$ is defined as

$$R_n(L, w_c) \triangleq \int_{\Lambda_c} w_c(d\psi) R_n(L, \psi). \tag{3.3}$$

The minimum expected redundancy for a given prior w_c (which is attained by the ideal code length w.r.t. the mixture, $L(x^n) = -\log P_{w_c}(x^n)$) is defined as

$$R_n(w_c) \overset{\triangle}{=} \min_L R_n(L, w_c). \tag{3.4}$$

Finally, the maximin redundancy of the subclass Λ_c is the worst case minimum expected redundancy among all priors w_c, i.e.,

$$R_n^-(\Lambda_c) \overset{\triangle}{=} \sup_{w_c} R_n(w_c). \tag{3.5}$$

It was established by Davisson [5] that the maximin redundancy of Λ_c is identical to the normalized capacity of the channel defined by the conditional probability $P_\psi(x^n)$, i.e.,

$$R_n^-(\Lambda_c) = \frac{C_n(\Lambda_c)}{n} \overset{\triangle}{=} \sup_{w_c} \frac{1}{n} I_{w_c}(\Psi;\ X^n), \tag{3.6}$$

where $I_{w_c}(\Psi;\ X^n)$ is the mutual information induced by the joint measure $w_c(\psi) \cdot P_\psi(x^n)$, i.e.,

$$I_{w_c}(\Psi;\ X^n) \overset{\triangle}{=} \int_{\psi\in\Lambda_c} w_c(d\psi) \sum_{x^n\in\Sigma^n} P_\psi(x^n)\log\frac{P_\psi(x^n)}{P_{w_c}(x^n)}. \tag{3.7}$$

Gallager [8] then showed that if $P_\psi(x^n)$ is a measurable function of ψ for every x^n, as is assumed throughout this paper, then

$$R_n^+(\Lambda_c) = R_n^-(\Lambda_c) = \frac{1}{n}C_n(\Lambda_c). \tag{3.8}$$

Let w_c^* be the capacity achieving prior for Λ_c in (3.6). Merhav and Feder showed in [13] that if $C_n \to \infty$ as $n \to \infty$, then the probability under w_c^* that the redundancy of $\psi \in \Lambda_c$ is lower bounded by the capacity of (3.6) multiplied by a factor of $1-\varepsilon$ goes to 1, where ε is arbitrarily small. In other words,

$$R_n(L, \psi) \geq (1-\varepsilon)\frac{C_n(\Lambda_c)}{n} \tag{3.9}$$

for all $\psi \in \Lambda_c$, except for a subset $B_c \subset \Lambda_c$, where

$$w_c^*(B_c) \leq e \cdot 2^{-\varepsilon C_n}. \tag{3.10}$$

In particular, we can lower bound $R_n^-(\Lambda_c)$, $R_n^+(\Lambda_c)$, and the redundancy of any code for most sources under w_c^*, if we define a subclass $\Lambda_c' \subseteq \Lambda_c$, whose capacity is easier to lower bound. To do that, let us allow x^n to be generated only by a source from the set $\left(\psi^1, \psi^2, \ldots, \psi^M\right)$ of M sources in Λ_c', and define the decision rule

$$\Omega^i = \left\{x^n \in \Sigma^n\ :\ \text{decide } \psi^i\right\}. \tag{3.11}$$

Let the probability of error be defined as

$$P_e = \frac{1}{M}\sum_{i=1}^{M} P_{\psi^i}\left(\overline{\Omega}^i\right). \tag{3.12}$$

Then, if $P_e \to 0$, we use (3.9) and the fact that the capacity between $\Psi \in \Lambda_c'$ and X^n is always not smaller than the logarithm of the number M of points in Λ_c' that are distinguishable observing X^n to obtain

$$R_n(L,\psi) \geq (1-\varepsilon)\frac{C_n(\Lambda_c)}{n} \geq (1-\varepsilon)\frac{C_n(\Lambda_c')}{n} \geq (1-\varepsilon)\frac{\log M}{n} \tag{3.13}$$

for almost all $\psi \in \Lambda_c$ except the set B_c with probability under w_c^* upper bounded as in (3.10). A similar lower bound as in (3.13) is naturally true for both the maximin and minimax redundancies of Λ_c.

The lower bound in (3.9) can be used for almost all sources in the class if w_c^* is close to a uniform prior. However, if this is not the case, the vanishing probability of B_c w.r.t. w_c^* may not mean that the set B_c is negligible w.r.t. Λ_c. The second result in [13] is therefore more general for any w_c. We will review it here with slight modifications for our needs for the uniform prior.

Let $\mu_c(\cdot)$ be the uniform prior on the sources in Λ_c. Let $\Lambda_{c,\varepsilon} \subseteq \Lambda_c$ be a subclass of Λ_c, that satisfies a given condition and contains almost all sources in Λ_c, except for a set of sources $\overline{\Lambda}_{c,\varepsilon}$, that do not satisfy the condition and whose probability under the uniform prior over Λ_c is negligible, i.e.,

$$\mu_c\left(\overline{\Lambda}_{c,\varepsilon}\right) \to 0, \tag{3.14}$$

where we use $\mu_c(\Omega)$ for some set Ω to denote the probability of the set under the uniform distribution in Λ_c. Now, let us randomly draw a source from $\Lambda_{c,\varepsilon}$ in the following manner. First, select a random vector $\mathbf{\Phi} \overset{\triangle}{=} \left(\Psi^1, \Psi^2, \ldots, \Psi^{M_{\mathbf{\Phi}}}\right)$ of $M_{\mathbf{\Phi}} \geq M$ sources, under some distribution $w_c(\boldsymbol{\varphi} \mid \Lambda_{c,\varepsilon})$ over $\Lambda_{c,\varepsilon}$. Let $\boldsymbol{\varphi} = \left(\psi^1, \psi^2, \ldots, \psi^{M_\varphi}\right)$ denote its value, and assume that for every $\psi \in \Lambda_{c,\varepsilon}$ there is exactly one possible value $\boldsymbol{\varphi}$ of $\mathbf{\Phi}$ that contains ψ. Then, a source Ψ is drawn from the random set $\mathbf{\Phi}$, where for all i, $1 \leq i \leq M_{\mathbf{\Phi}}$, $\mu\left(\Psi = \psi^i \mid \mathbf{\Phi} = \boldsymbol{\varphi}\right) = 1/M_\varphi$. The distribution $w_c(\boldsymbol{\varphi} \mid \Lambda_{c,\varepsilon})$ is chosen in a manner that ensures that all $\psi \in \Lambda_{c,\varepsilon}$ are uniformly distributed in $\Lambda_{c,\varepsilon}$, i.e., $w_c(\psi \mid \Lambda_{c,\varepsilon}) \overset{\triangle}{=} w_c(\boldsymbol{\varphi} : \psi \in \boldsymbol{\varphi} \mid \Lambda_{c,\varepsilon})/M_\varphi = \mu_c(\psi)/\mu_c(\Lambda_{c,\varepsilon})$. By choosing in this manner we can control which sources are picked in the same set $\boldsymbol{\varphi}$, while still keeping the uniform prior, and also pick sets $\boldsymbol{\varphi}$ of different sizes M_φ.

For a random choice of the set of sources $\mathbf{\Phi} = \boldsymbol{\varphi}$, let us now assume that x^n may only be generated by a source from the set. We can use the definitions of the decision rule in (3.11) and the probability of error in (3.12) to compute the error probability $P_e(\boldsymbol{\varphi})$ for each $\boldsymbol{\varphi}$. Let $B_c(\boldsymbol{\varphi})$ denote the set of sources $\psi \in \boldsymbol{\varphi}$ for which

$$R_n(L,\psi) < (1-\varepsilon)\frac{\log M_\varphi}{n}. \tag{3.15}$$

Then, given $\mathbf{\Phi} = \boldsymbol{\varphi}$, the probability of $B_c(\boldsymbol{\varphi})$ can be upper bounded by (see [13])

$$\mu(B_c \mid \mathbf{\Phi} = \boldsymbol{\varphi}) = \frac{|B_c(\boldsymbol{\varphi})|}{M_\varphi} \leq \frac{P_e(\boldsymbol{\varphi})\log M_\varphi + 2}{\varepsilon \log M_\varphi} \leq \frac{P_e(\boldsymbol{\varphi})\log M + 2}{\varepsilon \log M}, \tag{3.16}$$

where $|\cdot|$ denotes the number of elements in a set. If $M \to \infty$, and if for every choice $\boldsymbol{\varphi}$ of $\mathbf{\Phi}$, we can upper bound the probability of error by

$$P_e(\boldsymbol{\varphi}) < \eta, \text{ where } \lim_{n\to\infty} \eta = 0, \tag{3.17}$$

then

$$\begin{aligned} \mu_c(B_c) &= \mu_c(\Lambda_{c,\varepsilon}) \cdot \mu_c(B_c \mid \Lambda_{c,\varepsilon}) \\ &\le 1 \cdot \int_{\boldsymbol{\varphi}} d\boldsymbol{\varphi} \cdot w_c(\mathbf{\Phi} = \boldsymbol{\varphi} \mid \Lambda_{c,\varepsilon}) \cdot \mu(B_c \mid \mathbf{\Phi} = \boldsymbol{\varphi}) \le \frac{\eta \log M + 2}{\varepsilon \log M} \to 0, \end{aligned} \tag{3.18}$$

and the redundancy of almost all sources $\psi \in \Lambda_c$, except the sources in the negligible probability sets B_c and $\overline{\Lambda}_{c,\varepsilon}$, is lower bounded by

$$R_n(L, \psi) \ge (1 - \varepsilon) \frac{\log M}{n}. \tag{3.19}$$

Note that this result holds regardless of the value of $w_c(\boldsymbol{\varphi} \mid \Lambda_{c,\varepsilon})$. Hence, although for any $\boldsymbol{\varphi}$ there must be a unique value of $w_c(\boldsymbol{\varphi} \mid \Lambda_{c,\varepsilon})$ to keep the distribution of ψ in $\Lambda_{c,\varepsilon}$ uniform, we need not worry about computing this value, as long as each source $\psi \in \Lambda_{c,\varepsilon}$ is contained in exactly one set $\boldsymbol{\varphi}$. Furthermore, this result implies, that the same lower bound can be obtained for most sources under non-uniform priors $w_c(\psi)$ designed in this manner.

Up to this point, we discussed the redundancy inside the subclass Λ_c, i.e., the *intra-subclass* redundancy. However, it is assumed that the unknown source is inside the class Λ, and can belong to any subclass Λ_c in this general PSS class, where each such subclass may have different capacity. It is obvious that additional redundancy is required for a universal code on the complete class Λ in order to enable the decoder to identify the subclass Λ_c. This additional redundancy is referred to as the *inter-subclass* redundancy (see [7]). We now upper bound the minimax redundancy of Λ_c in this case, and show the guidelines for designing this term of the redundancy to be negligible w.r.t. the intra-subclass redundancy. Similar analysis can be performed to bound the maximin redundancy.

Let L_c^* be the code that achieves the minimax redundancy $R_n^+(\Lambda_c)$ of Λ_c, i.e., for all $\psi \in \Lambda_c$,

$$R_n(L_c^*, \psi) \le \sup_{\psi' \in \Lambda_c} R_n(L_c^*, \psi') = R_n^+(\Lambda_c). \tag{3.20}$$

By Kraft's inequality, there must exist a PMF $P_c(x^n)$ that satisfies

$$L_c^*(x^n) \ge -\log P_c(x^n), \ \forall x^n. \tag{3.21}$$

Let us now define the PMF

$$P(x^n) \triangleq \sum_c w(c) P_c(x^n), \tag{3.22}$$

where $w(\cdot)$ is some weight function (prior). Let us choose the length function

$$L'(x^n) = \lceil -\log P(x^n) \rceil . \tag{3.23}$$

Then, the redundancy of L' for any $\psi \in \Lambda_c$ is obtained by

$$\begin{aligned} R_n(L', \psi) &= \frac{1}{n}\left[E_\psi L'(X^n) - E_\psi L_c^*(X^n)\right] + \frac{1}{n} E_\psi L_c^*(X^n) - H_\psi(X^n) \\ &\le \frac{1}{n}\left[E_\psi \log \frac{P_c(X^n)}{P(X^n)} + 1\right] + R_n(L_c^*, \psi) \\ &\le \frac{1}{n}\left[-\log w(c) + 1\right] + R_n(L_c^*, \psi) . \end{aligned} \tag{3.24}$$

The first inequality is obtained by definition of the length functions, and the second by the definition of $P(x^n)$ and by reducing the denominator of the logarithm. Let L^* denote the length function that achieves the best redundancy for the worst source $\psi^* \in \Lambda_c$, given the source is in Λ. By definition of L^*, for any other code including L' there must be a source $\psi' \in \Lambda_c$ (that may be identical to ψ^* or may be different) that achieves larger redundancy. Therefore,

$$\begin{aligned} R_n(L^*, \psi^*) \le R_n(L', \psi') \;&\le\; R_n(L_c^*, \psi') + \frac{1}{n}\left[-\log w(c) + 1\right] \\ &\le R_n^+(\Lambda_c) + \frac{1}{n}\left[-\log w(c) + 1\right] . \end{aligned} \tag{3.25}$$

The last two inequalities are obtained from (3.24) and (3.20), respectively.

The bound in (3.25) is essentially an upper bound on the minimax redundancy $R_n^+(\Lambda_c \mid \Lambda)$ of the subclass Λ_c given the source is in the class Λ. We conclude that if there exists a mixture code over the subclasses of Λ, for which the normalized negative logarithm of the weight assigned to Λ_c in the mixture is negligible w.r.t. $R_n^+(\Lambda_c)$ for every c, then $R_n^+(\Lambda_c \mid \Lambda)$ and $R_n^+(\Lambda_c)$ are asymptotically equal, and hence the additional redundancy required to determine the subclass of a source ψ is negligible.

4. The Lower Bound. The main result, which lower bounds the redundancy for almost all sources in Λ_c as in (1.3), is presented in this section. In particular, we simply lower bound the capacity and the random coding capacity of the channel defined by $\{P_\psi, \psi \in \Lambda_c\}$ for every $c = o(n)$ and use the results reviewed in the preceding section to lower bound the redundancy for most sources. First, we define a subclass $\Lambda_c' \subseteq \Lambda_c$, in which we determine a set of M sources, which we prove to be distinguishable, and then use (3.13) to lower bound the minimax and maximin redundancies and the redundancy for a group of sources with probability that goes to 1 when $n \to \infty$ under the capacity achieving prior. Then, we use the random coding results, described in (3.14)-(3.19), to obtain the stronger result which lower bounds the redundancy for almost all sources in Λ_c with the lower bound in (1.3), but at the cost of increasing ε w.r.t. the minimax lower bound. The difficulty in deriving the bounds, i.e. in determining the largest M for which the set of sources is still distinguishable, lies in the asymptotics that result from allowing $c \to \infty$ as $n \to \infty$. We begin with a formal regularity condition, which is assumed about the class $\{p_\theta\}$, and is satisfied for common parametric classes. The regularity condition, its analysis, and some of the analysis in the first part of the proof of the minimax bound resemble, in part, the respective presentations in [12], although they are not identical. The similar parts are presented here for the sake of completeness.

4.1. The Regularity Condition. The following regularity condition is assumed about the parametric family of segmental PMF's $\{p_\theta, \theta \in \boldsymbol{\Theta}\}$ throughout this paper.

Condition A. *Let the random vector X^l be governed by the parameter θ. Then, there exists an estimator $\hat{\theta} = f\left(X^l\right)$ of θ that satisfies*

$$p_\theta\left(\left\|\hat{\theta}-\theta\right\| \geq \frac{\lambda}{2 \cdot l^{0.5(1-\alpha)}}\right) \leq \frac{1}{l^{r\alpha}}, \tag{4.1}$$

for every $\theta \in \boldsymbol{\Theta}$, where l is sufficiently large, and $r > 0$, $\alpha > 0$, and $\lambda > 0$ are all constants.

Condition A is required to hold for any fixed value of r. In particular, if l is sufficiently large, it should hold if α is *fixed* arbitrarily small and if $r = \beta/\xi$, where β is a constant and $\xi > 0$ is *fixed* at some arbitrarily small value. The constant λ is used to normalize the resolution of a grid of points in $\boldsymbol{\Theta}$. It should typically satisfy $1 \leq \lambda < 2$.

Condition A is somewhat less demanding than the condition described in [12], that requires that for some positive r', there exists a constant $\kappa(r')$ such that for every $\theta \in \boldsymbol{\Theta}$, and all large enough l,

$$E_\theta\left(\left\|\hat{\theta}-\theta\right\|^{r'}\right) \leq \frac{\kappa(r')}{l^{r'/2}}. \tag{4.2}$$

It can be shown, that if for every r' we take $r < r'/2$, condition A results from the above condition.

Condition A holds for memoryless sources, Markov sources, finite-state sources, and other classes of practical interest as long as r is kept small enough w.r.t. l. However, if r is too large, the condition may no longer hold. This property should be carefully treated when using this condition. It results in the need for separate proof in the main theorem for the case where the number of segments is very large. The following lemma shows values of r for which condition A holds for memoryless sources.

Lemma 4.1. *Let*

$$r \leq \frac{\lambda^2}{8 \ln 2} \cdot \frac{l^\alpha}{\alpha \log l} - \frac{k \log(l+1)}{\alpha \log l}. \tag{4.3}$$

Then, condition A holds for any memoryless source over an alphabet of $k+1$ letters.

Proof. Let $\hat{\theta}$ be the estimator of θ, which uses the empirical relative frequency of a letter to estimate its probability. (Note that this is the maximum likelihood estimator). Then, the *divergence* between $p_{\hat{\theta}}$ and p_θ satisfies (see [2])

$$\begin{aligned}(4.4)\qquad D\left(p_{\hat{\theta}}||p_\theta\right) &\triangleq \sum_{x\in\Sigma} p_{\hat{\theta}}(x)\log\frac{p_{\hat{\theta}}(x)}{p_\theta(x)} \geq \frac{1}{2\ln 2}\left[\sum_{x\in\Sigma}\left|p_{\hat{\theta}}(x)-p_\theta(x)\right|\right]^2\\ &\geq \frac{1}{2\ln 2}\sum_{x\in\Sigma}\left|p_{\hat{\theta}}(x)-p_\theta(x)\right|^2 \geq \frac{1}{2\ln 2}\left\|\hat{\theta}-\theta\right\|^2,\end{aligned}$$

where the last inequality is obtained since the Euclidean distance between the parameter vectors is computed without adding the probability difference of the last letter in the alphabet.

Now, let A denote the event that $\left\|\hat{\theta}-\theta\right\| \geq 0.5\lambda l^{-0.5(1-\alpha)}$. Then, by typical sets analysis (see [2]),

$$\begin{aligned}(4.5)\qquad p_\theta(A) &\leq (l+1)^k \cdot 2^{-l\min_A D\left(p_{\hat{\theta}}||p_\theta\right)}\\ &\leq (l+1)^k \cdot 2^{-\frac{1}{2\ln 2}l\min_A\left\|\hat{\theta}-\theta\right\|^2} \leq \frac{1}{l^{r\alpha}},\end{aligned}$$

where the last inequality is obtained by lower bounding the minimum Euclidean distance of event A by $0.5\lambda l^{-0.5(1-\alpha)}$, and lower bounding the expression in the resulting negative exponent of 2 by $r\alpha\log l$ using the condition in (4.3). This concludes the proof of the Lemma 4.1.

From Lemma 4.1, we observe that if $r = 1/\alpha$, then condition A holds for memoryless sources. Additionally, if $l = m^\xi$ for a fixed arbitrarily small $\xi > 0$ for $m \to \infty$, it also holds if $r = \beta/\xi$ for some positive constant β. Note that ξ can even go to zero as $m \to \infty$, but at a slow rate satisfying $\xi > \beta(\log\log m)/(\log m)$ for some constant β. However, if r is larger than the bound in (4.3), condition A may no longer hold for memoryless sources, and in fact, it can be shown that this is the case, starting at some larger r. The same behavior is true for other classes of sources, although the particular different regions of r may vary.

The following two properties, used in proving the theorems presented in this section, result directly from condition A. Define $\mathbf{g} \triangleq \left(g^1, g^2, \ldots, g^{M_\theta}\right)$ as a grid of $M_\theta = l^{0.5k(1-\xi)}/\left(K\lambda_\theta^k\right)$ points in the k-dimensional space $\boldsymbol{\Theta}$, where $\xi > 0$ is fixed arbitrarily small. The constant $K \geq 1$ is the relation between the volumes of the spaces $[0,1]^k$ and $\boldsymbol{\Theta}$, i.e., since not all points in $[0,1]^k$ are necessarily valid values of segmental parameters, only $1/K$ of the volume of $[0,1]^k$ contains valid values of θ. (For example, for a memoryless source with 3 alphabet letters, the probability of the second letter β must not exceed $1 - p_\theta(\alpha)$, where α is the first letter). The constant λ_θ is a resolution factor as the one defined in Condition A. Each component of a grid point g^j is smaller than 1 and is a nonnegative integer multiple of $\lambda_\theta l^{-0.5(1-\xi)}$ with additional displacement of $0.5\lambda_\theta l^{-0.5(1-\xi)}$. Let $l = \beta m^{1-\xi'}$, where $m \to \infty$ is the mean segment length. Now, let X^l be generated by p_θ where $\theta = g^j$ for some index j, $1 \leq j \leq M_\theta$, and let $\hat{\theta} = f\left(X^l\right)$ be an estimator of θ that satisfies condition A. Define the *grid estimator* $\hat{\theta}^g \in \mathbf{g}$ of θ as the point $g^i \in \mathbf{g}$, which is the center of a k-dimensional cube with sides $\lambda_\theta l^{-0.5(1-\xi)}$, inside which the estimator $\hat{\theta}$ falls, (i.e., the grid point in $\mathbf{g}$ for which the Euclidean distance from $\hat{\theta}$ is the smallest). Then, by definition of the grid estimator and condition A,

$$p_\theta\left(\hat{\theta}^g \neq \theta\right) \leq p_\theta\left(\left\|\hat{\theta} - \theta\right\| \geq \frac{\lambda_\theta}{2l^{0.5(1-\xi)}}\right) \leq \frac{1}{l^{r\xi}} = \frac{1}{\beta^{r\xi} m^{r\xi(1-\xi')}}. \tag{4.6}$$

Now, let $l = \lambda_t m^\xi$, where λ_t, $1 \leq \lambda_t < 2$, is a normalizing factor later used to determine the number of grid points, and let x^l be generated by θ. Let θ' satisfy $\|\theta - \theta'\| > \lambda_t^{1/4} l^{-1/4} = m^{-\xi/4}$. Then, if $\hat{\theta}$ satisfies condition A,

$$\begin{aligned} p_\theta\left(\left\|\hat{\theta} - \theta\right\| \geq \left\|\hat{\theta} - \theta'\right\|\right) &\leq p_\theta\left(\left\|\hat{\theta} - \theta\right\| \geq 0.5\left\|\theta - \theta'\right\|\right) \\ &\leq p_\theta\left(\left\|\hat{\theta} - \theta\right\| \geq \frac{\lambda_t^{1/4}}{2l^{1/4}}\right) \leq \frac{1}{l^{r/2}} = \frac{1}{\lambda_t^{r/2} m^{\xi r/2}}, \end{aligned} \tag{4.7}$$

where the first inequality is obtained using the triangle inequality $\left\|\hat{\theta} - \theta'\right\| \geq \|\theta - \theta'\| - \left\|\hat{\theta} - \theta\right\|$, and the last one by using condition A with $\alpha = 0.5$ and $\lambda = \lambda_t^{1/4}$.

4.2. The Minimax Lower Bound. We now derive the lower bound for the minimax and maximin redundancies and for the redundancy for a set of sources $\psi \in \Lambda_c$ with probability that goes to 1 under w_c^*. The bound is obtained by selecting a subclass Λ_c' of the class Λ_c and showing that we can select a set of $M = m^{(1-\varepsilon)(0.5kq+c)}$ sources from Λ_c' that are distinguishable, i.e., with probability of error P_e as defined in (3.12) that goes to 0. Hence, by the redundancy capacity theorem, the minimax redundancy for $\psi \in \Lambda_c'$ is lower bounded by $(\log M)/n$, and by (3.13) the same result applies to the specified group of sources in Λ_c, and (1.3) immediately follows for this group. The cost of applying the lower bound to a large group of sources, which is reflected by the ε term in M, is shown to be larger as q becomes very large w.r.t. m because the exponential growth rate of the set of distinguishable sources is reduced as q grows. We now present the theorem and its proof.

THEOREM 4.1. *Let condition A hold, fix an arbitrarily small $\varepsilon > 0$, and let n be sufficiently large and c be of smaller order than n. Let w_c^* be the capacity achieving prior between $\Psi \in \Lambda_c$ and X^n. Then (1.3) holds for $R_n^-(\Lambda_c)$, $R_n^+(\Lambda_c)$, and for any uniquely decipherable code with length function $L(\cdot)$ and every $\psi \in \Lambda_c$ except a set of sources $B_c \subset \Lambda_c$, for which $w_c^*(B_c) \to 0$ as $n \to \infty$.*

Proof. We begin with the case in which there exists some constant ν such that as $n \to \infty$, $q \leq m^\nu$, i.e., the number of segments is not significantly larger than the mean segment length. This part of the proof will no longer hold if $q \gg m$, i.e., if for every constant ν, $q > m^\nu$ as $n \to \infty$. Let $\psi \in \Lambda_c'$ if and only if both of the following conditions hold. First, the ith transition satisfies $t_i \in [(i - 0.25)m, (i + 0.25)m)$, and second, the segmental parameters satisfy $\|\theta_i - \theta_{i-1}\| > m^{-\xi/4}$ for some fixed arbitrarily small $\xi > 0$. The first condition ensures that each transition occurs within an interval of length $0.5m$, and for each segment there exists an interval of length $0.5m$ entirely within that segment. The second condition ensures that the statistical transitions between segments are large enough.

Now, let us select a set $\boldsymbol{\varphi}$ of M sources from Λ'_c as follows. For the ith transition, define the jth element, $1 \le j < 0.5m^{1-\xi}/\lambda_t$, in the grid $\boldsymbol{\tau}_i$ as

$$(4.8) \qquad \tau_i^j \triangleq (i - 0.25)\, m + (j-1)\,\lambda_t m^{\xi} < (i + 0.25)\, m,$$

where λ_t, $1 \le \lambda_t \le 2$, is some constant. Thus for every i, the number of grid points in $\boldsymbol{\tau}_i$ is $M_{t_i} = M_t = 0.5m^{1-\xi}/\lambda_t$. Now, let $l = 0.5m$, and let the grid for the ith segmental parameter be defined as the grid $\mathbf{g}$ defined in Section 4.1, i.e., with $M_\theta = (m/2)^{0.5k(1-\xi)} / \left(K\lambda_\theta^k\right)$ grid points. Now, the set $\boldsymbol{\varphi}$ consists of all sources ψ such that $t_i \in \boldsymbol{\tau}_i$ and $\theta_i \in \mathbf{g} \;:\; \|\theta_i - \theta_{i-1}\| > m^{-\xi/4}$. The number M of sources in $\boldsymbol{\varphi}$ is thus bounded by

$$(4.9) \qquad M \ge M_t^c \cdot \zeta^{kq} M_\theta^q = 0.5^{c+0.5kq(1-\xi)} \cdot \left(\frac{\zeta}{\lambda_\theta}\right)^{kq} \cdot \frac{1}{K^q \lambda_t^c} \cdot m^{(1-\xi)(0.5kq+c)},$$

where the factor ζ, $1 > \zeta \ge 1 - 2m^{-\xi/4}$, results from the Euclidean distance condition between the segmental parameters of two adjacent segments. Taking the logarithm of the last equation and absorbing all the lower order terms in ε', we obtain

$$(4.10) \qquad \log M \ge (1 - \varepsilon')\,(0.5kq + c)\log m.$$

It is now left to show that for $\boldsymbol{\varphi}$ as defined, the probability of error P_e goes to zero under the assumption of the theorem that $m \to \infty$ (which results from $n \to \infty$ and $c = o(n)$). Once this is shown, (3.13) can be used to conclude the proof of the theorem, absorbing ε' in ε, where for any given ε, a smaller ξ can always be chosen. To show that the sources in $\psi \in \boldsymbol{\varphi}$ are distinguishable, let us define a sub-optimal decision rule, and show that the probability of error of this decision rule goes to zero.

Let the decision rule be defined as follows. Find the grid estimator $\hat{\theta}_i^g \in \mathbf{g}$ (as defined in Section 4.1) of θ_i using the data string $x_{(i-0.75)m+1}^{(i-0.25)m}$, which clearly is entirely generated by the segmental parameter θ_i. Let $\hat{\boldsymbol{\theta}}^g$ denote the estimate obtained for the extended vector $\boldsymbol{\theta}$ in this manner. Now, parse the ith interval $[(i-0.25)\,m, (i+0.25)\,m)$ into phrases of length $\lambda_t m^{\xi}$ each. Let $\hat{\theta}_{ij}$ denote the estimate of θ from the data in the jth phrase obtained by an estimator that satisfies condition A. Define $\hat{\theta}_{ij}^g$ as

$$(4.11) \qquad \hat{\theta}_{ij}^g \triangleq \begin{cases} \hat{\theta}_i^g, & \text{if} \left\|\hat{\theta}_{ij} - \hat{\theta}_i^g\right\| \le \left\|\hat{\theta}_{ij} - \hat{\theta}_{i+1}^g\right\| \\ \hat{\theta}_{i+1}^g, & \text{otherwise} \end{cases}.$$

If there is a single parsing point (between two phrases), for which $\hat{\theta}_{ij}^g = \hat{\theta}_i^g$ for all phrases in the interval to the left of this point, and $\hat{\theta}_{ij}^g = \hat{\theta}_{i+1}^g$ for all phrases in the interval to the right of this point, then define this point as $\hat{t}_i$. (This includes the edge points of the interval, for which there are no phrases on one side). If there is no such point, randomly pick any of the change points in the interval. Let $\hat{\mathbf{t}}$ denote the estimate of the transition path obtained in this manner, which is, by the process of estimating $\hat{\mathbf{t}}$, clearly a function of $\hat{\boldsymbol{\theta}}^g$.

Now, let X^n be generated by $\psi \in \boldsymbol{\varphi}$, and let Ω be the decision region of X^n for which we decide ψ. Then, the probability of error of the optimal decision rule can be upper bounded by

$$\begin{aligned}
(4.12) \qquad P_\psi\left\{\overline{\Omega}\right\} &\leq P_\psi\left\{\left[\hat{\boldsymbol{\theta}}^g \neq \boldsymbol{\theta}\right] \cup \left[\hat{\mathbf{t}}\left(\hat{\boldsymbol{\theta}}^g\right) \neq \mathbf{t}\right]\right\} \\
&= P_\psi\left\{\hat{\boldsymbol{\theta}}^g \neq \boldsymbol{\theta}\right\} + P_\psi\left\{\left[\hat{\mathbf{t}}\left(\hat{\boldsymbol{\theta}}^g\right) \neq \mathbf{t}\right] \cap \left[\hat{\boldsymbol{\theta}}^g = \boldsymbol{\theta}\right]\right\} \\
&= P_\psi\left\{\hat{\boldsymbol{\theta}}^g \neq \boldsymbol{\theta}\right\} + P_\psi\left\{\left[\hat{\mathbf{t}}\left(\boldsymbol{\theta}\right) \neq \mathbf{t}\right] \cap \left[\hat{\boldsymbol{\theta}}^g = \boldsymbol{\theta}\right]\right\} \\
&\leq P_\psi\left\{\hat{\boldsymbol{\theta}}^g \neq \boldsymbol{\theta}\right\} + P_\psi\left\{\hat{\mathbf{t}}\left(\boldsymbol{\theta}\right) \neq \mathbf{t}\right\} \\
&\leq \sum_{i=1}^{q} p_{\theta_i}\left\{\hat{\theta}_i^g \neq \theta_i\right\} + \sum_{i=1}^{c} P_\psi\left\{\hat{t}_i\left(\boldsymbol{\theta}\right) \neq t_i\right\} \\
&\leq q \cdot \frac{2^{r\xi}}{m^{r\xi}} + c \cdot \frac{0.5m^{1-\xi}}{\lambda_t} \cdot \frac{1}{\lambda_t^{r/2} m^{r\xi/2}}.
\end{aligned}$$

The first inequality is since we defined a sub-optimal decision rule. Then, we use the equality $\Pr(A \cup B) = \Pr(A) + \Pr\left(\overline{A} \cap B\right)$, and then replace $\hat{\mathbf{t}}\left(\hat{\boldsymbol{\theta}}^g\right)$ by $\hat{\mathbf{t}}(\boldsymbol{\theta})$ in the joint probability to obtain the next equality (which is true since $\Pr(A, B) = \Pr(A \mid B)\Pr(B)$). Next, we use the fact that $\Pr(A, B) \leq \Pr(A)$, and then the union bound is used to obtain the next inequality. Finally, the two properties of condition A described in (4.6) and (4.7) are used to obtain the last inequality, where the union bound is used in the second term w.r.t. the number of phrases. Since the bound is true for every $\psi \in \boldsymbol{\varphi}$, it is true for P_e. Taking $r = [1 + (\log q) / (\log m)] \cdot 2/\xi$ results in

$$(4.13) \qquad P_e \leq \frac{q 2^{r\xi}}{m^{r\xi}} + \frac{0.5c}{\lambda_t^{r/2+1}} \frac{m^{1-\xi}}{m^{r\xi/2}} \leq \frac{1}{m^\xi} \rightarrow 0,$$

as long as $\log q \leq \nu \log m$ for some constant ν, and this concludes the proof for this case. However, if this is not the case and for every constant ν, $\log q > \nu \log m$ as $n \rightarrow \infty$, condition A may no longer hold for the proposed choice of r, and thus the above bound on P_e does not hold. We next complete the proof for this case.

First, let us define the grids $\mathbf{g}$ and $\boldsymbol{\tau}_i$, such that $\zeta^k M_\theta$ and M_t are both integer powers of 2, and therefore the Galois fields $GF\left(\zeta^k M_\theta\right)$ and $GF\left(M_t\right)$ (see [11]) exist. This is done by choosing the constants λ_t and λ_θ^k accordingly between 1 and 2. Now, the first $(1-\eta)\, q$ segmental parameters and $(1-\eta)\, c$ transition times for some fixed arbitrarily small $\eta > 0$ are selected from the respective grids as in the first part of the proof and take any of the $\zeta^k M_\theta$ and M_t grid points, respectively. (Note that we can arbitrarily limit the number of possible grid points for θ_1, which is not subjected to the condition of sufficiently large Euclidean distance from another parameter, to $\zeta^k M_\theta$ by spacing them farther apart). Now, to select the last ηq segmental parameters and ηc transition times, let us construct systematic linear block codes $[q, (1-\eta) q]$ and $[c, (1-\eta) c]$ over $GF\left(\zeta^k M_\theta\right)$ and $GF\left(M_t\right)$, respectively, where the notation $[n, k]$ is used to denote block codes with information blocks of length k and codewords of length n over the respective Galois field (see [11]). The indices in the respective grids of the first $(1-\eta) q$ segmental parameters and of the first $(1-\eta) c$ transition times will serve as information blocks. Then, the indices in the grids of the last ηq segmental parameters and the last ηc transition times will be the parity check sets of the respective codes, and will determine the respective parameters.

Defining φ to contain only the sources ψ that make legal codewords in both codes, we decrease $\log M$ by a factor of $1-\eta$, i.e.,

$$\log M \geq (1-\eta)\left(1-\varepsilon'\right)(0.5kq + c) \log m. \tag{4.14}$$

Absorbing $\eta + \varepsilon' - \eta\varepsilon'$ in ε and using (3.13), we need to show that the probability of error in the new set φ now goes to zero in order to conclude the proof of the theorem for the case in which $q \gg m$ (i.e., when for every constant ν, $q > m^\nu$ as $n \to \infty$). To do so, let us design the codes with the largest possible *Hamming* distances d_θ and d_t. (The Hamming distance of a code is the smallest number of components that are different between any two distinct codewords in the code.) Hence, an uncorrectable error in estimating $\psi \in \varphi$ occurs only if more than $\lfloor (d_\theta - 1)/2 \rfloor$ segmental parameters or if more than $\lfloor (d_t - 1)/2 \rfloor$ transition times are estimated incorrectly. We begin with a lemma that shows existence of linear codes with Hamming distances proportional to ηq and ηc.

LEMMA 4.2. *There exist linear codes as defined above with Hamming distances $d_\theta = \alpha_\theta \eta q$ and $d_t = \alpha_t \eta c$, respectively, for every fixed α_θ and α_t, $0 \leq \alpha_\theta, \alpha_t < 1$, for which $\alpha_\theta \eta q$ and $\alpha_t \eta c$ are integers, as $m \to \infty$.*

Proof. We prove for the code over the transition times, but the proof is similar for the second code. For convenience, we omit the subscript t from the parameter α_t. The proof is based on the Gilbert-Varsharmov bound (see [9], [19]). The inequality

$$\sum_{i=0}^{\alpha\eta c-2} \binom{c-1}{i} \left(M_t - 1\right)^i \leq \sum_{i=0}^{\alpha\eta c-2} 2^{ch(i/c)} M_t^i \leq \alpha\eta c \cdot 2^c \cdot M_t^{\alpha\eta c} < M_t^{\eta c} \tag{4.15}$$

holds for every α, $0 \leq \alpha < 1$, and $m \to \infty$, where the first inequality is obtained by the inequality (see [2])

$$\frac{1}{j+1} 2^{jh(i/j)} \leq \binom{j}{i} \leq 2^{jh(i/j)}. \tag{4.16}$$

Therefore, by the Gilbert-Varsharmov bound there exists a linear code $[c, (1-\eta) c]$ over $GF\left(M_t\right)$ with Hamming distance $d_t = \alpha\eta c$, concluding the proof of the lemma.

Now, let us choose codes for which $\lfloor (d_\theta - 1)/2 \rfloor = \eta q/4$ and $\lfloor (d_t - 1)/2 \rfloor = \eta c/4$. Such codes can correct errors in up to $\eta c/4$ components of $\mathbf{t}$, and in up to at least $\eta q/8$ components of $\boldsymbol{\theta}$. This is due to the fact that an error in estimating θ_i may cause a second error in the representation of θ_{i+1} in the code, even if θ_{i+1} itself is estimated correctly. This is due to the differential nature of the code for $\boldsymbol{\theta}$ necessary to ensure no small transitions. Using the union bound for the worst source $\psi \in \boldsymbol{\varphi}$, equation (4.16), the fact that $m \to \infty$, and taking $r = 8/\xi$, we obtain

$$\begin{aligned}
P_e &\leq P_\psi \left\{ \hat{\boldsymbol{\theta}}^g \neq \boldsymbol{\theta} \right\} + P_\psi \left\{ \hat{\mathbf{t}}\left(\boldsymbol{\theta}\right) \neq \mathbf{t} \right\} \\
&\leq \begin{pmatrix} q \\ \eta q/8 \end{pmatrix} \cdot \left(\frac{2^{r\xi}}{m^{r\xi}} \right)^{\eta q/8} + \begin{pmatrix} c \\ \eta c/4 \end{pmatrix} \cdot \left(\frac{1}{2\lambda_t^{r/2+1} m^{\xi(r/2+1)-1}} \right)^{\eta c/4} \\
&\leq 2^{q\left[h\left(\frac{\eta}{8}\right) + \frac{r\xi\eta}{8}(1-\log m)\right]} + 2^{c\left[h\left(\frac{\eta}{4}\right) - \frac{\eta}{4}\left(1+\left(\frac{r}{2}+1\right)\log\lambda_t\right) - \frac{\eta}{4}\left(\xi\left(\frac{r}{2}+1\right)-1\right)\log m\right]} \\
&\to\ 0,
\end{aligned} \tag{4.17}$$

and conclude the proof of Theorem 4.1.

4.3. The Bound for Most Piecewise Stationary Sources. We now extend the result of Theorem 4.1 to hold for almost all sources $\psi \in \Lambda_c$ for any $c = o(n)$ and not only under the capacity achieving prior. This is done by using the uniform prior over all sources in the class and the results described in (3.14)-(3.19) in Section 3, and at the expense of increasing the value of ε even more, although it is shown that ε can still be made arbitrarily small. The analysis in [12] assumes that we can define a subset of Λ_c, that consists only of the sources with sufficiently long segments and sufficiently large transitions, for which the probability goes to 1 under the uniform prior, and thus the probability of the complementary set goes to zero. This assumption is no longer true if c is allowed to grow with n. However, it turns out that a similar assumption holds in this case for a different less strict definition of the subset of Λ_c, denoted by $\Lambda_{c,\varepsilon}$. If the subset $\Lambda_{c,\varepsilon}$ is defined to consist of sources that contain at most δq short segments and at most δc small transitions for an arbitrarily small fixed $\delta > 0$, it can be shown that its probability goes to 1 under the uniform prior. This is the first part of the proof of the lower bound for almost all sources. Then, we use the random coding redundancy-capacity theorem to complete the proof. We use the same techniques that were used in the proof of Theorem 4.1 to show that selected sets of sources are distinguishable, although the actual sets of sources are selected differently, since sources now must be selected randomly, allowing all sources from the subset $\Lambda_{c,\varepsilon}$ to be selected, while maintaining the uniform prior to allow use of the random coding redundancy-capacity theorem. We now present the main result in Theorem 4.2, and conclude this section with its proof.

THEOREM 4.2. *Let condition A hold, fix an arbitrarily small $\varepsilon > 0$, and let n be sufficiently large and c be of smaller order than n ($c = o(n)$). Then (1.3) holds for any uniquely decipherable code with length function $L(\cdot)$ and for every $\psi \in \Lambda_c$ except for a set of sources $\left(B_c \cup \overline{\Lambda}_{c,\varepsilon}\right) \subset \Lambda_c$, for which the probability under the uniform* *prior over Λ_c goes to zero as $n \to \infty$.*

Proof. Let $\delta > 0$ and $\xi > 0$ be fixed arbitrarily small, where their values are determined by ε as will be shown at the end of the proof. Define $\Lambda_{c,\varepsilon}$ for two different cases. If $c \le m^{\xi/8}$, let $\Lambda_{c,\varepsilon}$ be the set of sources $\psi \in \Lambda_c$ for which all transitions satisfy $\|\theta_i - \theta_{i+1}\| > m^{-\xi/4}$ and all segments satisfy $t_i - t_{i-1} > m^{1-\xi/2}$. Otherwise, define $\Lambda_{c,\varepsilon}$ as the set of sources $\psi \in \Lambda_c$ for which there are at most δc small transitions that do not satisfy the condition above and at most δq short segments that do not satisfy the condition. Now, let us first bound the probability of the complementary set $\overline{\Lambda}_{c,\varepsilon}$ and show that it goes to zero under the uniform prior.

Using the union bound, the probability of $\overline{\Lambda}_{c,\varepsilon}$ is bounded by

$$\mu_c\left(\overline{\Lambda}_{c,\varepsilon}\right) \le \mu_c\left(\boldsymbol{\theta} \in \overline{\Lambda}_{c,\varepsilon}\right) + \mu_c\left(\mathbf{t} \in \overline{\Lambda}_{c,\varepsilon}\right), \tag{4.18}$$

where $\boldsymbol{\theta} \in \overline{\Lambda}_{c,\varepsilon}$ denotes all sources ψ for which the segmental parameters' vector $\boldsymbol{\theta}$ does not satisfy the condition which defines $\Lambda_{c,\varepsilon}$. The set $\mathbf{t} \in \overline{\Lambda}_{c,\varepsilon}$ is defined similarly w.r.t. $\mathbf{t}$.

The first term in (4.18) is bounded as follows: For any given value of θ_i, θ_{i+1} can take any possible value in $\boldsymbol{\Theta}$, except for the k-dimensional ball whose center is in θ_i, in order to satisfy $\|\theta_i - \theta_{i+1}\| > m^{-\xi/4}$. The probability that this condition is not satisfied, is therefore, the relative portion of this ball from $\boldsymbol{\Theta}$, which is bounded by the relative portion of the containing cube. Hence,

$$\mu\left(\|\theta_i - \theta_{i+1}\| \le m^{-\xi/4}\right) \le 2^k \cdot K \cdot m^{-k\xi/4}. \tag{4.19}$$

For $c \le m^{\xi/8}$, the union bound on all transitions can be used, and we still obtain an expression that goes to zero, that upper bounds $\mu_c\left(\boldsymbol{\theta} \in \overline{\Lambda}_{c,\varepsilon}\right)$. Otherwise, the probability of $\boldsymbol{\theta} \in \overline{\Lambda}_{c,\varepsilon}$ is upper bounded by the probability that $\|\theta_i - \theta_{i+1}\| \le m^{-\xi/4}$ for any choice of δc or more segments from the last c segments (the segmental parameter of the first segment can take all values in $\boldsymbol{\Theta}$). This probability is upper bounded with the union bound by

$$\begin{aligned} \mu_c\left(\boldsymbol{\theta} \in \overline{\Lambda}_{c,\varepsilon}\right) &\le \begin{pmatrix} c \\ \delta c \end{pmatrix} \mu\left(\|\theta_i - \theta_{i+1}\| \le m^{-\xi/4}\right)^{\delta c} \\ &\le \begin{pmatrix} c \\ \delta c \end{pmatrix} \left(2^k \cdot K \cdot m^{-k\xi/4}\right)^{\delta c} \le 2^{c\left[h(\delta)+\delta\log\left(K2^k\right)-\frac{\xi}{4}\delta k\log m\right]}. \end{aligned} \tag{4.20}$$

The last inequality is obtained using (4.16). The right hand side of the last inequality goes to zero as $m \to \infty$ for any value of $c > m^{\xi/8}$.

The total number of paths with c transitions in an n-tuple is the combination of c elements selected from $n-1$ elements. Since the extra 1 has no effect on the results, we will neglect it in the analysis and absorb it in resulting constants. To bound the second term of (4.18) for $c \le m^{\xi/8}$, we lower bound $\mu_c\left(\mathbf{t} \in \Lambda_{c,\varepsilon}\right)$ by an expression that goes to 1. By selecting transitions one by one such that each transition eliminates the range of at most $2m^{1-\xi/2}$ points around it for the selection of the next transitions, we can bound the total number of transition paths $\mathbf{t}$ for which all segments are at least $m^{1-\xi/2}$ long by

$$|\mathbf{t} \in \Lambda_{c,\varepsilon}| \geq \frac{1}{c!} \prod_{i=0}^{c-1} \left(n - 2im^{1-\xi/2}\right). \tag{4.21}$$

Taking the logarithm of the product, with straightforward manipulations and substitutions, we obtain

$$\ln \prod_{i=0}^{c-1} \left(n - 2im^{1-\xi/2}\right) \geq \int_0^c \ln \left(n - 2xm^{1-\xi/2}\right) \cdot dx \geq c \ln \frac{n}{e^{2/m^{\xi/2}}}. \tag{4.22}$$

Using Stirling's approximation (see [2]), the first order approximation of $\ln(1-x) \approx -x$ for very small x, and with straightforward manipulations, we obtain

(4.23)

$$\ln \mu_c\left(\mathbf{t} \in \Lambda_{c,\varepsilon}\right) \geq c \ln \frac{n}{e^{2/m^{\xi/2}}} - \ln \frac{n!}{(n-c)!} \geq -\frac{2c}{m^{\xi/2}} + O\left(c^2/n\right) \geq -\frac{2}{m^{3\xi/8}} \to 0,$$

where the last inequality uses the assumption that $c \leq m^{\xi/8}$. This results in $\mu_c(\mathbf{t} \in \overline{\Lambda}_{c,\varepsilon}) \to 0$. Note that this will not be true for $c \geq m^{\xi/2}$ if we had defined $\Lambda_{c,\varepsilon}$ similarly for larger values of c.

For $c > m^{\xi/8}$, let us now construct the set $\mathbf{t} \in \overline{\Lambda}_{c,\varepsilon}$ (for a given $\boldsymbol{\theta}$). The first $(1-\delta)\,c$ transitions are randomly chosen with no restriction from the n points (where the order they are chosen is not their time order). Now, the last δc transitions are chosen only in the vicinity of previously chosen transitions (i.e., at a time distance not larger than $m^{1-\xi/2}$ from a previously chosen transition). This ensures that $\mathbf{t}$ is in $\overline{\Lambda}_{c,\varepsilon}$, and the group of all choices constitutes the set $\mathbf{t} \in \overline{\Lambda}_{c,\varepsilon}$. The number of total points from which the last δc transitions are picked can be upper bounded by $2cm^{1-\xi/2}$. Of course, some choices of $\mathbf{t}$ are repeated in this manner, and thus using (4.16), we have

$$\begin{aligned} \mu_c\left(\mathbf{t} \in \overline{\Lambda}_{c,\varepsilon}\right) &\leq n \cdot 2^{-nh\left(\frac{c}{n}\right)} \begin{pmatrix} n \\ (1-\delta)\,c \end{pmatrix} \begin{pmatrix} 2cm^{1-\xi/2} \\ \delta c \end{pmatrix} \\ &\leq 2^{-c[(\xi\delta/2-1/c)\log m - \kappa - (\log c)/c]} \to 0, \end{aligned} \tag{4.24}$$

where the second inequality is obtained by straightforward manipulations, using the approximation $\ln(1-x) \approx -x$ for small x, and (4.16) to bound the combination numbers, and κ is a constant that depends on δ, ξ, c and m. We therefore showed that $\mu_c\left(\overline{\Lambda}_{c,\varepsilon}\right) \to 0$ in all cases, concluding the first part of the proof.

We now select random sets $\boldsymbol{\Phi}$ of random sources $\boldsymbol{\Psi} \in \Lambda_{c,\varepsilon}$, show that the sources in each set are distinguishable, and lower bound the sizes of those sets. We begin with classes for which there exists a constant ν such that $q \leq m^{\nu}$. To make sure that every possible point $\psi \in \Lambda_{c,\varepsilon}$ can be selected in *exactly* one value φ of the random set $\boldsymbol{\Phi}$, the set $\boldsymbol{\Phi} \subset \Lambda_{c,\varepsilon}$ is randomly chosen in the following steps:

1. Parse the time interval into phrases of length $m^{1-\xi}$ each.
2. Randomly select the number of transitions α, $0 \le \alpha \le \delta c$, for which $\|\Theta_i - \Theta_{i+1}\| \le m^{-\xi/4}$, and the number of segments β, $0 \le \beta \le \delta q$, for which $T_j - T_{j-1} \le m^{1-\xi/2}$. The numbers α and β are precisely the *exact* numbers of small transitions and short segments, respectively, for *all* the sources $\mathbf{\Psi}$ that will be selected in the particular set $\mathbf{\Phi}$.
3. Randomly designate the indices i of the α small transitions and the indices j of the β short segments. For each small transition, randomly choose the values T_i, Θ_i, and Θ_{i+1}, such that $\|\Theta_i - \Theta_{i+1}\| \le m^{-\xi/4}$. For each short segment, randomly choose Θ_j, T_{j-1}, and T_j, such that $T_j - T_{j-1} \le m^{1-\xi/2}$. All the values selected in this step will be *identical* for *all* sources $\mathbf{\Psi} \in \mathbf{\Phi}$ for this particular choice of $\mathbf{\Phi}$. Denote the set of the indices of all transitions not determined in this step by $I_t(\mathbf{\Phi})$, and the set of indices of all the segments that were not determined in this step by $I_\theta(\mathbf{\Phi})$.
4. For every transition $j \in I_t(\mathbf{\Phi})$, randomly choose the index Π_j of the phrase from the parsing in step 1, in which the jth transition will occur for all sources $\mathbf{\Psi} \in \mathbf{\Phi}$. Choose Π_j such that all points in the phrase are more than $m^{1-\xi/2}$ points away from the nearest point in Π_{j-1} if $j-1 \in I_t(\mathbf{\Phi})$ or from T_{j-1} otherwise, and from the nearest point in Π_{j+1} if $j+1 \in I_t(\mathbf{\Phi})$ or from T_{j+1} otherwise.
5. Let $\mathbf{g}$ be a grid as defined in the discussion preceding (4.6) for $l = m^{1-\xi}$, where g^ρ is the ρth grid point. Then, for every $i \in I_\theta(\mathbf{\Phi})$ define the random grid $\mathbf{G}_i$, where $G_i^\rho = g^\rho + U_i$, and U_i is a k-dimensional random vector (identical for all ρ's) uniformly distributed inside the k-dimensional cube, whose center is at the zero vector and whose sides are all of length $\lambda_\theta l^{-0.5(1-\xi)}$.
6. For each $j \in I_t(\mathbf{\Phi})$ define the ρth point of the random grid $\mathcal{T}_j$ as

$$\mathcal{T}_j^\rho = (\Pi_j - 1)\, m^{1-\xi} + (\rho - 1)\, \lambda_t m^\xi + V_j < \Pi_j m^{1-\xi}, \tag{4.25}$$

   where V_j is a random variable (identical for all ρ's) uniformly distributed on the discrete values $\left(0, 1, 2, \ldots, \lambda_t m^\xi - 1\right)$, (and assuming $\lambda_t m^\xi$ divides $m^{1-\xi}$).
7. The set $\mathbf{\Phi}$ contains all sources $\mathbf{\Psi}$, such that all parameters selected in step 3 are identical to all $\mathbf{\Psi} \in \mathbf{\Phi}$, and then for $j \in I_t(\mathbf{\Phi})$, $T_j \in \mathcal{T}_j$, and for $i \in I_\theta(\mathbf{\Phi})$, $\Theta_i \in \mathbf{G}_i$, but cannot be any point in the set of points $\{G_i^\rho\} \subset \mathbf{G}_i$ for which $\|G_i^\rho - \Theta_{i-1}\| \le m^{-\xi/4}$.

The construction of the set $\mathbf{\Phi}$ ensures that at most a single transition, which is not identical to all $\mathbf{\Psi} \in \mathbf{\Phi}$ occurs in a phrase. It also limits the number of segmental parameters identical to all sources in $\mathbf{\Phi}$ or the number of transitions identical to all sources in $\mathbf{\Phi}$ to $3\delta q$ (for the latter, we will limit this number to $3\delta c$, absorbing the extra negligible term in the low order terms). Hence, for any choice φ of $\mathbf{\Phi}$, the number of sources M_φ in the set is lower bounded, similarly to (4.9), by

$$M_\varphi \geq \left(\frac{\zeta}{\lambda_\theta}\right)^{(1-3\delta)kq} \cdot \frac{1}{K^{(1-3\delta)q} \cdot \lambda_t^{(1-3\delta)c}} \cdot m^{(1-3\delta)\left[(1-\xi)^2 0.5kq+(1-2\xi)c\right]}. \tag{4.26}$$

Absorbing all low order terms in ε'', we obtain

$$\log M_\varphi \geq (1-\varepsilon'')\,(0.5kq+c)\log m,\ \forall \boldsymbol{\varphi}, \tag{4.27}$$

where $\varepsilon'' > \varepsilon'$, for ε' defined in (4.10). Hence, using (3.19) and absorbing all low order terms in ε, we can obtain the lower bound of (1.3). The values of δ and ξ can now be determined from ε using the last inequalities. The bound will hold for almost all sources in Λ_c if we prove that the sources in every choice of $\mathbf{\Phi}$ are distinguishable.

To upper bound the probability of error $P_e(\mathbf{\Phi})$ for the random set $\mathbf{\Phi}$, let us define the decision rule. The decision rule must determine only the segmental parameters Θ_i for $i \in I_\theta(\mathbf{\Phi})$, and the transition times T_j for $j \in I_t(\mathbf{\Phi})$, since all other parameters are identical for all sources in $\mathbf{\Phi}$. A similar sub-optimal decision rule to the one used to prove Theorem 4.1 can be used. Because all segments that are not identical to all sources in $\mathbf{\Phi}$ must be longer than $m^{1-\xi/2}$, each such segment must contain a phrase of length $m^{1-\xi}$ entirely inside the segment. Since transitions take place in the same phrases for all $\Psi \in \mathbf{\Phi}$, we can use such a phrase to estimate Θ_i, using the grid estimator of the grid $\mathbf{G}_i$ for every $i \in I_\theta(\mathbf{\Phi})$. Then, after we estimate all segmental parameters, we use the same estimator as in the proof of Theorem 4.1 but with separation points from the grid $\mathcal{T}_j$, to estimate the transition time T_j inside the phrase Π_j, for every $j \in I_t(\mathbf{\Phi})$.

Let $\Psi = (\mathbf{\Theta}, \mathbf{T}) \in \mathbf{\Phi}$ be the source for which the probability of the error region is the largest. Then, the probability of error for the optimal decision rule in the set $\mathbf{\Phi}$ is bounded as in (4.12) by

$$\begin{aligned} P_e(\mathbf{\Phi}) &\leq \sum_{i\in I_\theta(\mathbf{\Phi})} p_{\Theta_i}\left\{\hat{\Theta}_i^{G_i} \neq \Theta_i\right\} + \sum_{i\in I_t(\mathbf{\Phi})} P_\Psi\left\{\hat{T}_i(\Theta) \neq T_i\right\} \\ &\leq q\cdot\frac{1}{m^{r\xi(1-\xi)}} + c\cdot\frac{m^{1-2\xi}}{\lambda_t}\cdot\frac{1}{\lambda_t^{r/2} m^{r\xi/2}}. \end{aligned} \tag{4.28}$$

Taking $r = [1 + (\log q)/(\log m)]\cdot 2/\xi$, as long as there exists a constant ν such that $q \leq m^\nu$, results in decay to zero of the right hand side of the last inequality, and thus of the error probability. This results in a negligible set $B_c \subset \Lambda_{c,\varepsilon}$ of sources for which the lower bound of (1.3) does not hold. Since $\overline{\Lambda}_{c,\varepsilon}$ is also negligible, the proof of Theorem 4.2 is concluded by the redundancy capacity theorem for the case in which the number of segments is not too large.

If $q \gg m$, i.e., for every constant ν, $q > m^\nu$ as $n \to \infty$, the proof is concluded in a similar manner to the proof of Theorem 4.1 for this case, with three main differences. First, the codes are now constructed on smaller fields, denoted by $GF\left(\zeta^k M'_\theta\right)$ and $GF\left(M'_t\right)$, since the richness of the respective grids reduces because they are defined by intervals of length $m^{1-\xi}$ instead of $0.5m$. Additionally, the codeword lengths are reduced by a factor of at most $(1-3\delta)$ because of short segments and small transitions. Finally, since all sources must be selected in the random coding scenario, and the construction of the codes reduces the number of sources in a new set, which will be denoted by $\mathbf{\Phi}'$, only to the sources in both codes, there must also be random selections of other sources in the original set $\mathbf{\Phi}$. This is done by adding a last random selection which for every set $\mathbf{\Phi}$ uniformly selects one of the $\left(\zeta^k M'_\theta\right)^{\eta q'}$ cosets (see [11]) of the code over the segmental parameters and one of the $M_t'^{\eta c'}$ cosets of the code over the transition times, where q' and c' are the numbers of indices in $I_\theta\left(\mathbf{\Phi}'\right)$ and $I_t\left(\mathbf{\Phi}'\right)$, respectively. The new set $\mathbf{\Phi}'$ will contain only vectors from the selected coset (where if the coset with the zero vector leader is chosen, the actual code constitutes $\mathbf{\Phi}'$). Note that if $i+1 \notin I_\theta\left(\mathbf{\Phi}'\right)$, Θ_{i+1} is determined for all $\Psi \in \mathbf{\Phi}'$. This should restrict the choice of Θ_i, so that it is sufficiently distant from both Θ_{i-1} and Θ_{i+1}, reducing the number of valid choices in the grid $\mathbf{G}_i$. However, since both T_i and Θ_{i+1} are known for every $\Psi \in \mathbf{\Phi}'$, segmental parameters Θ_i insufficiently distant from Θ_{i+1} are allowed.

Finally, as in (4.14), the number of sources in the new $\mathbf{\Phi}'$ is bounded by

$$\log M_{\mathbf{\Phi}'} \geq (1-\eta)\left(1-\varepsilon''\right)(0.5kq+c)\log m. \tag{4.29}$$

Absorbing all low order terms in ε, whose value can be used to determine ξ, δ and η, the proof for this case is concluded by bounding the error probability as in (4.17), replacing the expressions in (4.17) by those for this case, and showing that the error probability still goes to zero. This concludes the proof of Theorem 4.2.

5. Tightness and Achievability. The lower bound of (1.3) was derived in Theorem 4.2 for the redundancy of almost all sources $\psi \in \Lambda_c$ for every c. However, when we employ a universal code on a PSS, we seldom know the actual number of transitions c in advance. Therefore, an additional redundancy term is required in order to relay to the decoder which subclass $\Lambda_c \subset \Lambda$ is most likely to have generated x^n, or in other words, the number of transitions in the sequence. This raises the question of whether the bound of (1.3) is tight or not. Particularly, this question is strengthened since it can be shown that there exists a subclass $\Lambda_\varepsilon \subseteq \Lambda$, such that $\Lambda_\varepsilon \stackrel{\triangle}{=} \bigcup_{c=o(n)} \Lambda_{c,\varepsilon}$, whose probability in Λ goes to 1, in which sets of sources are distinguishable, where the logarithm of the number of sources selected from each $\Lambda_{c,\varepsilon}$ is as in (4.29). Using the redundancy-capacity theorem, the capacity of the whole class Λ therefore takes a value larger than the capacity of Λ_c, with the largest possible c. However, we would like to benefit from the fact that the capacity of each subclass Λ_c is smaller than that of Λ, where in particular this is significant for small c. It turns out that the extra redundancy term that distinguishes between different values of c can always be designed to be negligible w.r.t. the lower bound, and hence is absorbed in the ε term of the bound of (1.3) (decreasing the value of ε), and thus the bound is tight. This is also explained by the fact that every subclass $\Lambda_c \subset \Lambda$, with

c significantly smaller than the maximum c allowed, is of negligible volume w.r.t. the whole class, and therefore, the sources from this subclass are contained in the set B for which the bound for the complete class Λ does not hold.

For a given c, one way to achieve the lower bound, which resembles the one proposed in [12] for a fixed c, is to look at all possible transition paths $\mathbf{t}$, and code each segment imposed by $\mathbf{t}$ using an optimal universal code for stationary sequences. For example, the Krichevsky-Trofimov mixture code (see [10]), which was extended to the context tree algorithm (see [22]) for finite state sources, can be used for such sources. Then, we choose the path for which the shortest code is obtained, relay this path to the decoder, and then, code each segment with the optimal universal code. Unlike [12], the path must be relayed to the decoder by coding the lengths of the segments using Elias' representation of the integers (see [6]), instead of the absolute transition times. The redundancy of this code, whose length function is denoted by $L_c^*(\cdot)$, is upper bounded by

$$\begin{aligned}(5.1)\qquad R_n\left(L_c^*,\psi\right) &\leq (1+\varepsilon)\left[\frac{k}{2n}\sum_{i=1}^{q}\log\left(t_i-t_{i-1}\right)+\frac{1}{n}\sum_{i=1}^{c}\log\left(t_i-t_{i-1}\right)\right]\\ &\leq (1+\varepsilon)\left(0.5kq+c\right)\frac{\log m}{n}.\end{aligned}$$

The first inequality is obtained by the properties of an optimal universal code and of Elias' representation of the integers, and the second by Jensen's inequality. The bound of the second inequality can be obtained if instead of using Elias' representation and a universal code for stationary segments, we perform a double mixture, where in the second mixture we weight all possible paths uniformly, with probability of approximately $2^{-nh(1/m)}$ for each path.

Let us now assume that $L_c^*(\cdot)$ is an optimal code, that achieves the lower bound for a given c. Taking the number of transitions c for which the code described above obtains the shortest representation for x^n, and relaying this number to the decoder using the $(1+\varepsilon')\log c$ bits of Elias' representation, or defining a mixture code as in (3.22), and taking $w(c) = \alpha c^{-(1+\varepsilon')}$, where the constant α is determined by the allowable range of c, results in a universal code for every c over the class Λ, for which the additional term of $(1+\varepsilon')\log c$ bits for representing c as in (3.25) is always negligible w.r.t. the lower bound.

Last, the techniques that were proposed in [20] and [18] for strongly sequential coding of memoryless PSS's use a state transition diagram based double mixture method, where a single state s is created at each time unit s, and any state $s \leq j$ at time j represents all paths $\mathbf{t}$ for which the time unit of state s is the most recent transition at the current time unit j. Each state is assigned a weight, which is the double mixture over the segmental parameters of all paths represented by the state. The first mixture over the segmental parameters is performed using the Krichevsky-Trofimov mixture probability, which is updated sequentially in each state, while the mixture over the transition paths is performed by assigning portions of old states s

to new states j at time j. If the portion of an old state allotted to a new state is given by $\varepsilon/\left(j-s\right)$, this technique results in redundancy of $(1+\varepsilon)\log\left(t_i - t_{i-1}\right)$ extra bits for the ith transition, and thus in an upper bound similar to that in (5.1). The complexity of these techniques can be reduces (see [17] and [16]), without affecting the asymptotic behavior of the redundancy by combining change point estimation techniques with the mixture methods.

Note that most methods described above, including the low complexity strongly sequential ones, if properly designed, are in fact optimal also in the sense that for the sources in the negligible sets $\Lambda_{c,\varepsilon}$ and B_c smaller redundancy than that of the lower bound is achieved. Finally, all the above techniques (except the one with reduced complexity) achieve the lower bound not only for the mean redundancy, but also for the pointwise redundancy, defined in (2.7).

6. Summary and Conclusions. In this paper, we derived a general lower bound for almost all sources in the class of piecewise stationary sources. The bound was first derived in the minimax sense, and applied to most sources in the class but only when sources are weighted under the prior that achieves the capacity between the class and the data sequence, using the strong version of the redundancy-capacity theorem. Then, it was shown that reducing the bound by a small negligible factor results in a bound that applies to almost all sources in the class. Unlike a previously known result, the new bound applies not only to the case where the number of transitions is fixed, but also if the number of transitions grows with the sequence length. Finally, the tightness and achievability of the lower bound were discussed, and the new lower bound confirmed the optimality of recently proposed schemes for memoryless PSS's.

Acknowledgment. The authors gratefully acknowledge Professor Neri Merhav for many stimulating discussions.

REFERENCES

[1] M. Burrows and D. J. Wheeler, *A block-sorting lossless data compression algorithm*, Digital Systems Research Center, Palo Alto, CA, May 10, 1994.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[3] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.*, Academic Press, New York, 1981.

[4] L. D. Davisson, *Minimax noiseless universal coding for Markov sources*, IEEE Trans. Inform. Theory, IT-29(1983), pp. 211-215.

[5] L. D. Davisson, *Universal noiseless coding*, IEEE Trans. Inform. Theory, IT-19:6(1973), pp. 783-795.

[6] P. Elias, *Universal codeword sets and representation of the integers*, IEEE Trans. Inform. Theory, IT-21:2(1975), pp. 194-203.

[7] M. Feder and N. Merhav, *Hierarchical universal coding*, IEEE Trans. Inform. Theory, 42:5(1996), pp. 1354-1364.

[8] R. G. Gallager, *Source coding with side information and universal coding*, unpublished manuscript, September 1976.

[9] E. N. Gilbert, *A comparison of signalling alphabets*, Bell System Tech. J., 31(1952), pp. 504-522.

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

[11] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[12] N. Merhav, *On the minimum description length principle for sources with piecewise constant parameters*, IEEE Trans. Inform. Theory, 39:6(1993), pp. 1962-1967.

[13] N. Merhav and M. Feder, *A strong version of the redundancy-capacity theorem of universal coding*, IEEE Trans. Inform. Theory, 41:3(1995), pp. 714-722.

[14] J. Rissanen, *Universal coding, information, prediction, and estimation*, IEEE Trans. Inform. Theory, IT-30:4(1984), pp. 629-636.

[15] J. Rissanen and G. G. Langdon, Jr., *Arithmetic coding*, IBM J. Res. Dev., 23:2(1979), pp. 149-162.

[16] G. I. Shamir and D. J. Costello, Jr., *Asymptotically optimal threshold based low complexity sequential lossless coding for piecewise stationary memoryless sources*, in: Proceedings of the 1999 IEEE Information Theory and Networking Workshop, pp. 32, Metsovo, Greece, June 27 - July 1, 1999.

[17] G. I. Shamir and D. J. Costello, Jr., *Asymptotically optimal low complexity sequential lossless coding for piecewise stationary memoryless sources - Part I: The regular case*, IEEE Trans. Inform. Theory, 46:7(2000), pp. 2444-2467.

[18] G. I. Shamir and N. Merhav, *Low complexity sequential lossless coding for piecewise stationary memoryless sources*, IEEE Trans. Inform. Theory, 45:5(1999), pp. 1498-1519.

[19] R. R. Varsharmov, *Estimate of the number of signals in error correcting codes*, Dokl. Akad. Nauk SSSR 117:5(1957), pp. 739-741.

[20] F. M. J. Willems, *Coding for a binary Independent Piecewise-Identically-Distributed source*, IEEE Trans. Inform. Theory, 42:6(1996), pp. 2210-2217.

[21] F. M. J. Willems and M. Krom, *Live-and-die coding for binary piecewise i.i.d. sources*, in: Proceedings of the 1997 IEEE International Symposium on Information Theory, pp. 68, Ulm, Germany, June 29 - July 4, 1997.

[22] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, *The Context-Tree weighting method: basic properties*, IEEE Trans. Inform. Theory, 41:3(1995), pp. 653-664.